

## Sampling: final and initial sample size determination

Source: © Getty Images



*Making a sample too big wastes resources, making it too small diminishes the value of findings – a dilemma resolved only with the judicious use of sampling theory.*



### Objectives

- 1 define key concepts and symbols pertinent to sampling;
- 2 understand the concepts of the sampling distribution, statistical inference and standard error;
- 3 discuss the statistical approach to determining sample size based on simple random sampling and the construction of confidence intervals;
- 4 derive the formulae to determine statistically the sample size for estimating means and proportions;
- 5 discuss the non-response issues in sampling and the procedures for improving response rates and adjusting for non-response;
- 6 understand the difficulty of statistically determining the sample size in international marketing research;
- 7 identify the ethical issues related to sample size determination, particularly the estimation of population variance.

STAGE 1  
Problem  
definition

STAGE 2  
Research approach  
developed

STAGE 3  
Research design  
developed

STAGE 4  
Fieldwork or data  
collection

STAGE 5  
Data preparation  
and analysis

STAGE 6  
Report preparation  
and presentation

## Overview

This chapter focuses on the determination of sample size in simple random sampling. We define various concepts and symbols and discuss the properties of the sampling distribution. Additionally, we describe statistical approaches to sample size determination based on confidence intervals. We present the formulae for calculating the sample size with these approaches and illustrate their use. We briefly discuss the extension to determining sample size in other probability sampling designs. The sample size determined statistically is the final or net sample size; that is, it represents the completed number of interviews or observations. To obtain this final sample size, however, a much larger number of potential respondents have to be initially contacted. We describe the adjustments that need to be made to the statistically determined sample size to account for incidence and completion rates and calculate the initial sample size. We also cover the non-response issues in sampling, with a focus on improving response rates and adjusting for non-response. We discuss the difficulty of statistically determining the sample size in international marketing research and identify the relevant ethical issues.

Statistical determination of sample size requires knowledge of the normal distribution and the use of normal probability tables. The normal distribution is bell shaped and symmetrical. Its mean, median and mode are identical (see Chapter 18). Information on the normal distribution and the use of normal probability tables is presented in the appendix to this chapter. The following example illustrates the statistical aspects of sampling.

### Example Has there been a shift in opinion?

The sample size used in opinion polls commissioned and published by most national newspapers is influenced by statistical considerations. The allowance for sampling error may be limited to around three percentage points.

The table that follows can be used to determine the allowances that should be made for sampling error. These intervals indicate the range (plus or minus the figure shown) within which the results of repeated samplings in the same time period could be expected to vary, 95% of the time, assuming that the sample procedure, survey execution and questionnaire used were the same.

#### Recommended allowance for sampling error of a percentage

<i>In percentage points (at 95% confidence level for a sample size of 385)</i>	
Percentage near 10	3
Percentage near 20	4
Percentage near 30	4
Percentage near 40	5
Percentage near 50	5
Percentage near 60	5
Percentage near 70	4
Percentage near 80	4
Percentage near 90	3

The table should be used as follows. If a reported percentage is 43 (e.g. 43% of Norwegian chief executives believe their company will suffer from staff shortages in the next 12 months),

look at the row labelled 'percentages near 40'. The number in this row is 5, so the 43% obtained in the sample is subject to a sampling error of 5 percentage points. Another way of saying this is that very probably (95 times out of 100) the average of repeated samplings would be somewhere between 38% and 48%. The reader can be 95% confident that in the total population of Norwegian chief executives between 38% and 48% believe their company will suffer from staff shortages in the next 12 months, with the most likely figure being 43%.

The fortunes of political parties measured through opinion polls are regularly reported in newspapers throughout Europe. The next time that you read a report of a political opinion poll, examine the sample size used, the confidence level assumed and the stated margin of error. When comparing the results of a poll with a previous poll, consider whether a particular political party or politician has *really* grown or slumped in popularity. If there is a news item that the popularity of Party X has grown or the approval of President Y has diminished, is there really anything to report? Can the reported change be accounted for within the set margin of error as summarised in this example?

## Definitions and symbols



Confidence intervals and other statistical concepts that play a central role in sample size determination are defined in the following list:

- *Parameter.* A parameter is a summary description of a fixed characteristic or measure of the target population. A parameter denotes the true value that would be obtained if a census rather than a sample was undertaken.
- *Statistic.* A statistic is a summary description of a characteristic or measure of the sample. The sample statistic is used as an estimate of the population parameter.
- *Finite population correction.* The finite population correction (fpc) is a correction for overestimation of the variance of a population parameter – for example, a mean or proportion – when the sample size is 10% or more of the population size.
- *Precision level.* When estimating a population parameter by using a sample statistic, the precision level is the desired size of the estimating interval. This is the maximum permissible difference between the sample statistic and the population parameter.

**Table 15.1** Symbols for population and sample variables

Variable	Population	Sample
Mean	$\mu$	$\bar{X}$
Proportion	$\pi$	$p$
Variance	$\sigma^2$	$s^2$
Standard deviation	$\sigma$	$s$
Size	$N$	$n$
Standard error of the mean	$\sigma_{\bar{X}}$	$S_{\bar{X}}$
Standard error of the proportion	$\sigma_p$	$S_p$
Standardised variate ( $z$ )	$\frac{X - \mu}{\sigma}$	$\frac{X - \bar{X}}{S}$
Coefficient of variation ( $C$ )	$\frac{\sigma}{\mu}$	$\frac{S}{\bar{X}}$

- *Confidence interval.* The confidence interval is the range into which the true population parameter will fall, assuming a given level of confidence.
- *Confidence level.* The confidence level is the probability that a confidence interval will include the population parameter.

The symbols used in statistical notation for describing population and sample characteristics are summarised in Table 15.1.

## The sampling distribution

### Sampling distribution

The distribution of the values of a sample statistic computed for each possible sample that could be drawn from the target population under a specified sampling plan.

The **sampling distribution** is the distribution of the values of a sample statistic computed for each possible sample that could be drawn from the target population under a specified sampling plan.<sup>1</sup> Suppose that a simple random sample of five sponsors is to be drawn from a population of 20 Formula One sponsors. There are  $(20 \times 19 \times 18 \times 17 \times 16)/(1 \times 2 \times 3 \times 4 \times 5)$ , or 15,504 different samples of size 5 that can be drawn. The relative frequency distribution of the values of the mean of these 15,504 different samples would specify the sampling distribution of the mean.

An important task in marketing research is to calculate statistics, such as the sample mean and sample proportion, and use them to estimate the corresponding true population values. This process of generalising the sample results to a target population is referred to as **statistical inference**. In practice, a single sample of predetermined size is selected, and the sample statistics (such as mean and proportion) are computed. Theoretically, to estimate the population parameter from the sample statistic, every possible sample that could have been drawn should be examined. If all possible samples were actually to be drawn, the distribution of the statistic would be the sampling distribution. Although in practice only one sample is actually drawn, the concept of a sampling distribution is still relevant. It enables us to use probability theory to make inferences about the population values.

The important properties of the sampling distribution of the mean, and the corresponding properties for the proportion, for large samples (30 or more) are as follows:

- 1 The sampling distribution of the mean is a normal distribution (see the appendix at the end of the chapter). Strictly speaking, the sampling distribution of a proportion is a binomial. For large samples ( $n = 30$  or more), however, it can be approximated by the normal distribution.
- 2 The mean of the sampling distribution of the mean

$$\bar{X} = \left( \sum_{i=1}^n X_i / n \right)$$

or of the proportion ( $p$ ) equals the corresponding population parameter value,  $\mu$  or  $\pi$ , respectively.

- 3 The standard deviation is called the **standard error** of the mean or the proportion to indicate that it refers to a sampling distribution of the mean or the proportion and not to a sample or a population. The formulae are

$$\begin{array}{l} \text{Mean} \\ \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \end{array}$$

$$\begin{array}{l} \text{Proportion} \\ \sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}} \end{array}$$

### Statistical inference

The process of generalising the sample results to a target population.

### Standard error

The standard deviation of the sampling distribution of the mean or proportion.

- 4 Often the population standard deviation,  $\sigma$ , is not known. In these cases, it can be estimated from the sample by using the following formula:

$$s = \sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}}$$

or

$$s = \sqrt{\frac{\sum_{i=1}^n X_i^2 - \frac{\left(\sum_{i=1}^n X_i\right)^2}{n}}{n-1}}$$

In cases where  $\sigma$  is estimated by  $s$ , the standard error of the mean becomes

$$\text{est. } \sigma_{\bar{X}} = \frac{s}{\sqrt{n}}$$

where 'est.' denotes that  $s$  has been used as an estimate of  $\sigma$ .

Assuming no measurement error, the reliability of an estimate of a population parameter can be assessed in terms of its standard error.

- 5 Likewise, the standard error of the proportion can be estimated by using the sample proportion  $p$  as an estimator of the population proportion,  $\pi$ , as

$$\text{est. } S_p = \sqrt{\frac{p(1-p)}{n}}$$

- 6 The area under the sampling distribution between any two points can be calculated in terms of **z value**. The  $z$  value for a point is the number of standard errors a point is away from the mean. The  $z$  values may be computed as follows:

$$z = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}}$$

For example, the areas under one side of the curve between the mean and points that have  $z$  values of 1.0, 2.0 and 3.0 are, respectively, 0.3413, 0.4772 and 0.4986. (See Table 2 in the Appendix of statistical tables.) In the case of proportion, the computation of  $z$  values is similar.

- 7 When the sample size is 10% or more of the population size, the standard error formulae will overestimate the standard deviation of the population mean or proportion. Hence, these should be adjusted by a finite population correction factor defined by

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

In this case,

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$

#### z value

The number of standard errors a point is away from the mean.

## Statistical approaches to determining sample size

Several qualitative factors should also be taken into consideration when determining the sample size (see Chapter 14). These include the importance of the decision, the nature of the research, the number of variables, the nature of the analysis, sample sizes used in similar studies, incidence rates (the occurrence of behaviour or characteristics in a population), completion rates and resource constraints. The statistically determined sample size is the net or final sample size: the sample remaining after eliminating potential respondents who do not qualify or who do not complete the interview. Depending on incidence and completion rates, the size of the initial sample may have to be much larger. In commercial marketing research, limits on time, money and expert resources can exert an overriding influence on sample size determination. In the Formula One Racetrack Project, the sample size was determined based on these considerations.

The statistical approach to determining sample size that we consider is based on traditional statistical inference.<sup>2</sup> In this approach the precision level is specified in advance. This approach is based on the construction of confidence levels around sample means or proportions.

## The confidence interval approach

The confidence interval approach to sample size determination is based on the construction of confidence intervals around the sample means or proportions using the standard error formula. As an example, suppose that a researcher has taken a simple random sample of 300 households to estimate the monthly amount invested in savings schemes and found that the mean household monthly investment for the sample is €182. Past studies indicate that the population standard deviation  $\sigma$  can be assumed to be €55.

We want to find an interval within which a fixed proportion of the sample means would fall. Suppose that we want to determine an interval around the population mean that will include 95% of the sample means, based on samples of 300 households. The 95% could be divided into two equal parts, half below and half above the mean, as shown in Figure 15.1.

Calculation of the confidence interval involves determining a distance below ( $\bar{X}_L$ ) and above ( $\bar{X}_U$ ) the population mean ( $\bar{X}$ ), which contains a specified area of the normal curve.

The  $z$  values corresponding to  $\bar{X}_L$  and  $\bar{X}_U$  may be calculated as

$$z_L = \frac{\bar{X}_L - \mu}{\sigma_{\bar{X}}}$$

$$z_U = \frac{\bar{X}_U - \mu}{\sigma_{\bar{X}}}$$

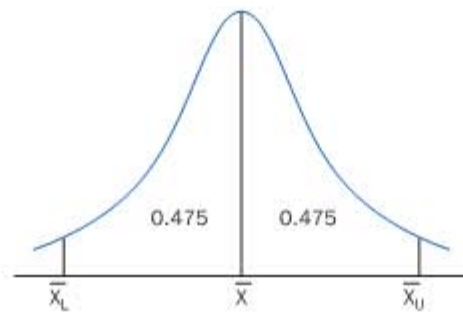
where  $z_L = -z$  and  $z_U = +z$ . Therefore, the lower value of  $\bar{X}$  is

$$\bar{X}_L = \mu - z\sigma_{\bar{X}}$$

and the upper value of  $\bar{X}$  is

$$\bar{X}_U = \mu + z\sigma_{\bar{X}}$$

**Figure 15.1**  
The 95% confidence interval



Note that  $\mu$  is estimated by  $\bar{X}$ . The confidence interval is given by

$$\bar{X} \pm z\sigma_{\bar{X}}$$

We can now set a 95% confidence interval around the sample mean of €182. As a first step, we compute the standard error of the mean:

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{55}{\sqrt{300}} = 3.18$$

From Table 2 in the Appendix of statistical tables, it can be seen that the central 95% of the normal distribution lies within  $\pm 1.96$   $z$  values. The 95% confidence interval is given by

$$\begin{aligned}\bar{X} \pm 1.96\sigma_{\bar{X}} &= 182.00 \pm 1.96(3.18) \\ &= 182.00 \pm 6.23\end{aligned}$$

Thus, the 95% confidence interval ranges from €175.77 to €188.23. The probability of finding the true population mean to be within €175.77 and €188.23 is 95%.

### Sample size determination: means

The approach used here to construct a confidence interval can be adapted to determine the sample size that will result in a desired confidence interval.<sup>3</sup> Suppose that the researcher wants to estimate the monthly household savings investment more precisely so that the estimate will be within  $\pm €5.00$  of the true population value. What should be the size of the sample? The following steps, summarised in Table 15.2, will lead to an answer.

- 1 Specify the level of precision. This is the maximum permissible difference ( $D$ ) between the sample mean and the population mean. In our example,  $D = \pm €5.00$ .
- 2 Specify the level of confidence. Suppose that a 95% confidence level is desired.
- 3 Determine the  $z$  value associated with the confidence level using Table 2 in the Appendix of statistical tables. For a 95% confidence level, the probability that the population mean will fall outside one end of the interval is 0.025 (0.05/2). The associated  $z$  value is 1.96.
- 4 Determine the standard deviation of the population. This may be known from secondary sources. If not, it might be estimated by conducting a pilot study. Alternatively, it might be estimated on the basis of the researcher's judgement. For example, the range of a normally distributed variable is approximately equal to  $\pm 3$  standard deviations, and one can thus estimate the standard deviation by dividing the range by 6. The researcher can often estimate the range based on knowledge of the phenomenon.

**Table 15.2** Summary of sample size determination for means and proportions

Steps	Means	Proportions
1 Specify the level of precision.	$D = \pm\epsilon 5.00$	$D = p - \pi = \pm 0.05$
2 Specify the confidence level (CL).	CL = 95%	CL = 95%
3 Determine the z value associated with the CL.	z value is 1.96	z value is 1.96
4 Determine the standard deviation of the population.	Estimate $\sigma$ : $\sigma = 55$	Estimate $\pi$ : $\pi = 0.64$
5 Determine the sample size using the formula for the standard error.	$n = \frac{\sigma^2 z^2}{D^2}$ $n = \frac{55^2 (1.96)^2}{5^2}$ $= 465$	$n = \frac{\pi(1-\pi) z^2}{D^2}$ $n = \frac{0.64 (1 - 0.64) (1.96)^2}{(0.05)^2}$ $= 355$
6 If the sample size represents 10% of the population, apply the finite factor popular correction (fpc)	$n_c = \frac{nN}{N + n - 1}$ $= \bar{X} + z s_{\bar{X}}$	$n_c = \frac{nN}{N + n - 1}$ $p \pm z s_p$
7 If necessary, re-estimate the confidence interval by employing s to estimate $\sigma$ .	$D = R\mu$	$D = R\pi$
8 If precision is specified in relative rather than absolute terms, determine the sample size by substituting for D.	$n = \frac{C^2 z^2}{R^2}$	$n = \frac{z^2 (1 - \pi)}{R^2 \pi}$

5 Determine the sample size using the formula for the standard error of the mean:

$$z = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}}$$

$$= \frac{D}{\sigma_{\bar{X}}}$$

or

$$\sigma_{\bar{X}} = \frac{D}{z}$$

or

$$\frac{\sigma}{\sqrt{n}} = \frac{D}{z}$$

or

$$n = \frac{\sigma^2 z^2}{D^2}$$

In our example,

$$n = \frac{55^2 (1.96)^2}{5^2}$$

$$= 464.83$$

$$= 465 \text{ (rounded to the next highest integer)}$$



It can be seen from the formula for sample size that sample size increases with an increase in the population variability, the degree of confidence and the precision level required of the estimate. Because the sample size is directly proportional to  $\sigma^2$ , the larger the population variability, the larger the sample size. Likewise, a higher degree of confidence implies a larger value of  $z$ , and thus a larger sample size. Both  $\sigma^2$  and  $z$  appear in the numerator. Greater precision means a smaller value of  $D$ , and thus a larger sample size because  $D$  appears in the denominator.

- 6 If the resulting sample size represents 10% or more of the population, the finite population correction (fpc) should be applied.<sup>4</sup> The required sample size should then be calculated from the formula

$$n_c = \frac{nN}{(N + n - 1)}$$

where  $n$  = sample size without fpc

$n_c$  = sample size with fpc

- 7 If the population standard deviation,  $\sigma$ , is unknown and an estimate is used, it should be re-estimated once the sample has been drawn. The sample standard deviation,  $s$ , is used as an estimate of  $\sigma$ . A revised confidence interval should then be calculated to determine the precision level actually obtained.

Suppose that the value of 55.00 used for  $\sigma$  was an estimate because the true value was unknown. A sample of  $n = 465$  is drawn, and these observations generate a mean  $\bar{X}$  of 180.00 and a sample standard deviation  $s$  of 50.00. The revised confidence interval is then

$$\begin{aligned}\bar{X} \pm z_{\bar{X}} &= 180.00 \pm 1.96 \times \left( \frac{50.0}{\sqrt{465}} \right) \\ &= 180.00 \pm 4.55\end{aligned}$$

or  $175.45 \leq \mu \leq 184.55$ . Note that the confidence interval obtained is narrower than planned, because the population standard deviation was overestimated, as judged by the sample standard deviation.

- 8 In some cases, precision is specified in relative rather than absolute terms. In other words, it may be specified that the estimate be within plus or minus  $R$  percentage points of the mean. Symbolically,

$$D = R\mu$$

In these cases, the sample size may be determined by

$$\begin{aligned}n &= \frac{\sigma^2 z^2}{D^2} \\ &= \frac{C^2 z^2}{R^2}\end{aligned}$$

where the coefficient of variation  $C = \sigma/\mu$  would have to be estimated.

The population size,  $N$ , does not directly affect the size of the sample, except when the fpc factor has to be applied. Although this may be counter-intuitive, upon reflection it makes sense. For example, if all the population elements are identical on the characteristics of interest, then a sample size of one will be sufficient to estimate the mean perfectly. This is true whether there are 50, 500, 5,000 or 50,000 elements in the population. What directly affects the sample size is the variability of the characteristic in the population. This variability enters into the sample size calculation by way of population variance  $\sigma^2$  or sample variance  $s^2$ .

**Sample size determination: proportions**

If the statistic of interest is a proportion rather than a mean, the approach to sample size determination is similar. Suppose that the researcher is interested in estimating the proportion of households that own a dog as a pet. The following steps should be followed:<sup>5</sup>

- 1 Specify the level of precision. Suppose that the desired precision is such that the allowable interval is set as  $D = p - \pi = \pm 0.05$ .
- 2 Specify the level of confidence. Suppose that a 95% confidence level is desired.
- 3 Determine the  $z$  value associated with the confidence level. As explained in the case of estimating the mean, this will be  $z = 1.96$ .
- 4 Estimate the population proportion  $\pi$ . As explained earlier, the population proportion may be estimated from secondary sources, or from a pilot study, or may be based on the judgement of the researcher. Suppose that based on secondary data the researcher estimates that 64% of the households in the target population own a dog. Hence,  $\pi = 0.64$ .
- 5 Determine the sample size using the formula for the standard error of the proportion:

$$\begin{aligned}\sigma_p &= \frac{p - \pi}{z} \\ &= \frac{D}{z} \\ &= \sqrt{\frac{\pi(1 - \pi)}{n}}\end{aligned}$$

or

$$n = \frac{\pi(1 - \pi)z^2}{D^2}$$

In our example,

$$\begin{aligned}n &= \frac{0.64(1 - 0.64)(1.96)^2}{(0.05)^2} \\ &= 354.04 \\ &= 355 \text{ (rounded to the next highest integer)}\end{aligned}$$

- 6 If the resulting sample size represents 10% or more of the population, the fpc should be applied. The required sample size should then be calculated from the formula

$$n_c = \frac{nN}{(N + n - 1)}$$

where  $n$  = sample size without fpc

$n_c$  = sample size with fpc

- 7 If the estimate of  $\pi$  turns out to be poor, the confidence interval will be more or less precise than desired. Suppose that after the sample has been taken, the proportion  $p$  is calculated to have a value of 0.55. The confidence interval is then re-estimated by employing  $s_p$  to estimate the unknown  $\sigma_p$  as

$$p \pm z s_p$$

where

$$S_p = \sqrt{\frac{p(1 - p)}{n}}$$

In our example

$$S_p = \sqrt{\frac{0.55(1-0.55)}{355}} \\ = 0.0264$$

The confidence interval, then, is

$$0.55 \pm 1.96(0.0264) = 0.55 \pm 0.052$$

which is wider than that specified. This is because the sample standard deviation based on  $p = 0.55$  was larger than the estimate of the population standard deviation based on  $\pi = 0.64$ .

If a wider interval than specified is unacceptable, the sample size can be determined to reflect the maximum possible variation in the population. This occurs when the product  $\pi(1 - \pi)$  is the greatest, which happens when  $\pi$  is set at 0.5. This result can also be seen intuitively. Since one half of the population has one value of the characteristic and the other half the other value, more evidence would be required to obtain a valid inference than if the situation was more clear cut and the majority had one particular value. In our example, this leads to a sample size of

$$n = \frac{0.05(0.5)(1.96)^2}{(0.05)^2} \\ = 384.16 \\ = 385 \text{ (rounded to the next higher integer)}$$

- 8 Sometimes, precision is specified in relative rather than absolute terms. In other words, it may be specified that the estimate be within plus or minus  $R$  percentage points of the population proportion. Symbolically,

$$D = R\pi$$

In such a case, the sample size may be determined by

$$n = \frac{z^2(1-\pi)}{R^2\pi}$$

### Example

#### Statistical sampling: not always an emergency<sup>6</sup>

The city of Los Angeles wished to evaluate demand for non-emergency city services and service usage patterns. The aim was to implement a new system that would alleviate some of the strain on the city's 911 emergency phone system. A telephone survey of 1,800 randomly selected city residents was conducted.

The survey was stratified into two groups of 900 each: residents who had requested a service in the past six months and a group of other residents. The sample size was determined by using a 95% confidence interval and a margin of error of 3.5%. At this confidence level, one would expect that if all the residents of Los Angeles were asked the same survey, the responses to the survey would change by no more than  $\pm 3.5\%$ .

To confirm that the sample size of 900 was adequate, calculations for sample size determination by proportions were made as follows, using the maximum population variation ( $\pi = 0.05$ ). The precision of  $D$  in this study is 0.035 for a 95% confidence level.

$$n = \pi(1 - \pi)z^2/D^2 \\ n = [(0.05)(1 - 0.05)(1.96^2)]/(0.035)^2 = 784$$

Therefore, the 900 sample size was more than sufficient.

## Multiple characteristics and parameters

In the preceding examples, we focused on the estimation of a single parameter. In most marketing research projects, several characteristics, not just one, are of interest. The researcher is required to estimate several parameters, not just one. The calculation of sample size in these cases should be based on a consideration of all the parameters that must be estimated.

For example, suppose that in addition to the mean household spend at a supermarket, it was decided to estimate the mean household spend on clothes and on gifts. The sample sizes needed to estimate each of the three mean monthly expenses are given in Table 15.3 and are 465 for supermarket shopping, 246 for clothes and 217 for gifts. If all three variables were equally important, the most conservative approach would be to select the largest value of  $n = 465$  to determine the sample size. This will lead to each variable being estimated at least as precisely as specified. If the researcher was most concerned with the mean household monthly expense on clothes, however, a sample size of  $n = 246$  could be selected.

**Table 15.3** Sample size for estimating multiple parameters

	Variable Monthly household spend on:		
	Supermarket	Clothes	Gifts
Confidence level	95%	95%	95%
z value	1.96	1.96	1.96
Precision level ( $D$ )	€5	€5	€4
Standard deviation of the population ( $\sigma$ )	€55	€40	€30
Required sample size ( $n$ )	465	246	217

## Other probability sampling techniques

So far, the discussion of sample size determination has been based on the methods of traditional statistical inference and has assumed simple random sampling. Next, we discuss the determination of sample size when other sampling techniques are used. The determination of sample size for other probability sampling techniques is based on the same underlying principles. The researcher must specify the level of precision and the degree of confidence and estimate the sampling distribution of the test statistic.

In simple random sampling, cost does not enter directly into the calculation of sample size. In the case of stratified or cluster sampling, however, cost has an important influence. The cost per observation varies by strata or cluster, and the researcher needs some initial estimates of these costs.<sup>7</sup> In addition, the researcher must take into account within-strata variability or within- and between-cluster variability. Once the overall sample size is determined, the sample is apportioned among strata or clusters. This increases the complexity of the sample size formulae. The interested reader is referred to standard works on sampling theory for more information.<sup>8</sup> In general, to provide the same reliability as simple random sampling, sample sizes are the same for systematic sampling, smaller for stratified sampling and larger for cluster sampling.

## Adjusting the statistically determined sample size



Sample size determined statistically represents the final sample size that must be achieved to ensure that the parameters are estimated with the desired degree of precision and set level of confidence. In surveys, this represents the number of completed interviews or questionnaires. To achieve this final sample size, a much greater number of potential respondents have to be contacted. In other words, the initial sample size has to be much larger because typically the incidence rates and completion rates are less than 100%.<sup>9</sup>

### Incidence rate

The rate of occurrence of persons eligible to participate in the study expressed as a percentage.

**Incidence rate** refers to the rate of occurrence or the percentage of persons eligible to participate in the study. Incidence rate determines how many contacts need to be screened for a given sample size requirement.<sup>10</sup> For example, suppose that a study of pet ownership targets a sample of households. Of the households that might be approached to see if they qualify, approximately 75% own a pet. This means that, on average, 1.33 households would be approached to obtain one qualified respondent. Additional criteria for qualifying respondents (e.g. product usage behaviour) will further increase the number of contacts. Suppose that an added eligibility requirement is that the household should have bought a toy for their pet during the last two months. It is estimated that 60% of the households contacted would meet this criterion. Then the incidence rate is  $0.75 \times 0.6 = 0.45$ . Thus the final sample size will have to be increased by a factor of  $(1/0.45)$  or 2.22.

### Completion rate

The percentage of qualified respondents who complete the interview. It enables researchers to take into account anticipated refusals by people who qualify.

Similarly, the determination of sample size must take into account anticipated refusals by people who qualify. The **completion rate** denotes the percentage of qualified respondents who complete the interview. If, for example, the researcher expects an interview completion rate of 80% of eligible respondents, the number of contacts should be increased by a factor of 1.25. The incidence rate and the completion rate together imply that the number of potential respondents contacted – that is, the initial sample size – should be  $2.22 \times 1.25$  or 2.78 times the sample size required. In general, if there are  $c$  qualifying factors with an incidence of  $Q_1 \times Q_2 \times Q_3 \dots \times Q_c$  each expressed as a proportion, the following are true:

$$\text{Incidence rate} = Q_1 \times Q_2 \times Q_3 \times \dots \times Q_c$$

$$\text{Initial sample size} = \frac{\text{final sample size}}{\text{incidence rate} \times \text{completion rate}}$$

The number of units that will have to be sampled will be determined by the initial sample size. Often a number of variables are used for qualifying potential respondents, thereby decreasing the incidence rate. Completion rates are affected by non-response issues. These issues deserve particular attention and are detailed in the next section.

## Non-response issues in sampling



The two major non-response issues in sampling are improving response rates and adjusting for non-response. Non-response error arises when some of the potential respondents included in the sample do not respond (see Chapter 3). This is one of the most significant problems in survey research. Non-respondents may differ from respondents in terms of demographic, psychographic, personality, attitudinal, motivational and behavioural variables.<sup>11</sup> Evaluating these differences was detailed in Chapter 14 in the process of sample validation. For a given study, if the non-respondents differ from the respondents on the characteristics of interest, the sample estimates can be seriously biased. Higher response rates, in general, imply lower rates of non-response bias, yet response rate may not be an adequate indicator of non-response bias. Response rates themselves do not indicate whether



Source: © Getty Images

the respondents are representative of the original sample.<sup>12</sup> Increasing the response rate may not reduce non-response bias if the additional respondents are no different from those who have already responded but do differ from those who still do not respond. As low response rates increase the probability of non-response bias, an attempt should be made to improve the response rate.<sup>13</sup> This is not an issue that should be considered after a survey approach has been decided and a questionnaire designed. Factors that improve response rates are integral to survey and questionnaire design. As detailed in Chapter 13, the marketing researchers should build up an awareness of what motivates their target respondents to participate in a research study. The researchers should ask themselves

what their target respondents get in return for spending time and effort, answering set questions in a full and honest manner. The following sub-section details the techniques involved in improving response rates and adjusting for non-response.

### Improving the response rates

The primary causes of low response rates are refusals and not-at-homes, as shown in Figure 15.2.

**Refusals.** Refusals, which result from the unwillingness or inability of people included in the sample to participate, result in lower response rates and increased potential for non-response bias. The following examples illustrate some of the factors that drive refusals, terminations and completed interviews.

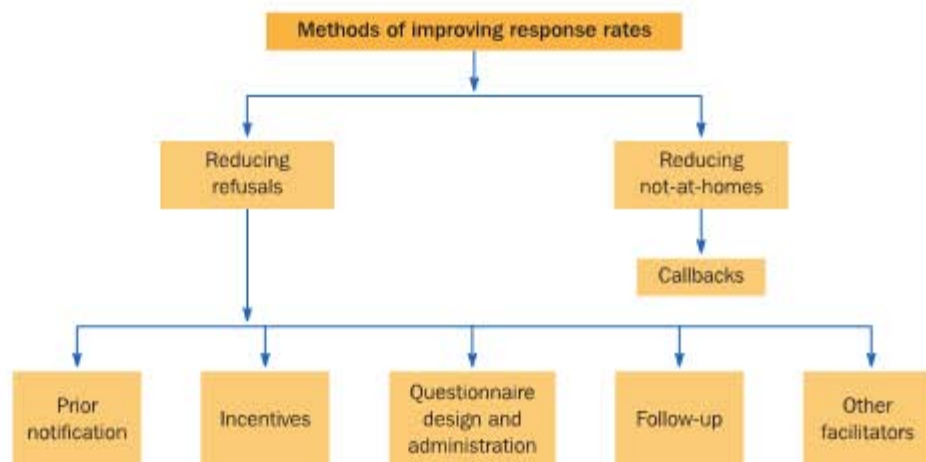


Figure 15.2 Improving response rates

#### Example Reasons for refusal<sup>14</sup>

In a study investigating the refusal problem in telephone surveys, telephone interviews were conducted with responders and non-responders to a previous survey, using quotas of 100 for each subsample. The results are presented in the following table:

## Refusals, terminations and completed interviews

Property	Total sample	Responders	Non-responders
Number of refusals (1)	224	31	193
Number of terminations (2)	100	33	67
Number of completed interviews (3)	203	102	101
Total number of contacts (1 + 2 + 3)*	527	166	361
Refusal rate [1/(1 + 2 + 3)]	42.5%	18.7%	53.5%
Termination rate [2/(1 + 2 + 3)]	19.0%	19.9%	18.5%
Completion rate [3/(1 + 2 + 3)] †	38.5%	61.4%	28.0%

\* A total of 1,388 attempts was required to make these contacts: the 166 responder contacts required 406 attempts (with one call-back per respondent), and the 361 non-responder contacts required 982 attempts (with two call-backs per respondent). The sampling frame contained 965 phone numbers – 313 respondents and 652 non-responders.

† Responder/non-responder differences were significant at  $\alpha = 0.05$  (two-tail test).

The study found that people who are likely to participate in a telephone survey (responders) differ from those who are likely to refuse (non-responders) in the following ways:

- 1 Confidence in survey research
- 2 Confidence in the research organisation
- 3 Demographic characteristics
- 4 Beliefs and attitudes about telephone surveys.

A 2002 study by the Council for Marketing and Opinion Research (CMOR) in the USA showed that 18% of respondents say that the Internet is their first choice when doing a survey, up from 10% in 1999. On respondent cooperation, only 7% of respondents called were willing to answer questions. The situation in the USA reflects conditions in other developed markets such as Europe and Japan. Although average response rates in Japan are significantly higher than in the USA and Europe, they show a continuous downward trend, mainly by the rise of refusals.

**Example****'No time' main reason for refusal<sup>15</sup>**

Research conducted by Paul Iavarakas, Vice President of Nielsen Media Research ([www.nielsenmedia.com](http://www.nielsenmedia.com)), showed that there were four basic 'causes' for non-cooperation. These are 'no time or interest', 'privacy and confidentiality concerns', 'not enjoying the survey experience' and 'not believing one's opinions are important'. Paul says it is known that mentioning the survey's topic can significantly affect response rates:

*Response rates for surveys on topics that are of higher interest to respondents can be up to 40% higher. Health issue surveys for instance, usually get high response rates. But is it acceptable to mention the topic in the introduction to a survey? One professor at the Ohio State University Survey Center insisted that we not mention the survey topic because of the risk of non-response due to only people being interested in the survey being willing to be interviewed. Balancing effective measures to improve response rates and concerns about non-response errors will always be a very difficult trade off.*

Given the potential differences between respondents and non-respondents, researchers should attempt to lower refusal rates. This can be done by prior notification, incentives, good questionnaire design and administration, follow-up and other facilitators:

- *Prior notification.* In prior notification, potential respondents are telephoned, sent a letter or email notifying them of the imminent mail, telephone personal or Internet survey. Prior notification increases response rates, as the respondent's attention is drawn to the purpose of a study and the potential benefits, without the apparent 'chore' of the questionnaire. With the potential respondent's attention focused upon the purpose and benefits, the chances increase for a greater reception when approached actually to complete a survey.<sup>16</sup> A key element of notifying potential respondents lies in how surveys are introduced as illustrated in the following example.

### **Example** Only incentives increase response rates?<sup>17</sup>

In a study organised by a New Zealand marketing research agency that conducted omnibus surveys using CATI, a measure of the effectiveness of 'introductions' upon response rates was made. Four introduction elements were tested: an incentive (prize draw for a weekend holiday); an assurance that the survey was not a sales pitch; an assurance of confidentiality; and a short versus longer description of the survey topic. Overall, only the incentive significantly increased the response rate. In combination, the best result and the only one to achieve a significantly higher response rate was the use of the incentive coupled with a 'no-sales' assurance. The use of the incentive did not appear to encourage people to lie about their eligibility as respondents.

- *Incentives.* Response rates can be increased by offering monetary as well as non-monetary incentives to potential respondents. Monetary incentives can be prepaid or promised. The prepaid incentive is included with the survey or questionnaire. The promised incentive is sent to only those respondents who complete the survey. The most commonly used non-monetary incentives are premiums and rewards, such as pens, pencils, books and offers of survey results.<sup>18</sup> Prepaid incentives have been shown to increase response rates to a greater extent than promised incentives. The amount of incentive can vary from trivial amounts to tens of euros. The amount of incentive has a positive relationship with response rate, but the cost of large monetary incentives may outweigh the value and quality of additional information obtained.
- *Questionnaire design and administration.* A well-designed questionnaire can decrease the overall refusal rate as well as refusals to specific questions (see Chapter 13). If the questionnaire and experience of answering the questions are interesting for the respondent, using words and logic that are meaningful to them, the response rate can improve. Likewise, the skill used to administer the questionnaire in telephone and personal interviews can increase the response rate. Trained interviewers are skilled in refusal conversion or persuasion. They do not accept a no response without an additional plea. The additional plea might emphasise the brevity of the questionnaire or importance of the respondent's opinion. Interviewing procedures are discussed in more detail in Chapter 16.
- *Follow-up.* Follow-up, or contacting the non-respondents periodically after the initial contact, is particularly effective in decreasing refusals in mail and Internet surveys. The researcher might send a reminder to non-respondents to complete and return the questionnaire. Two or three mailings may be needed in addition to the original one. With proper follow-up, the response rate in mail surveys can be increased to 80% or more.<sup>19</sup> Follow-ups can be done by postcard, letter, telephone, email or personal contacts.
- *Other facilitators.* Personalisation, or sending letters addressed to specific individuals, is effective in increasing response rates, especially when practised in conjunction with prior notification.<sup>20</sup> The only downside of such an approach is actually obtaining the names and contact details of those to whom a questionnaire should be sent. The



following example illustrates the two approaches to increasing response rates. The first is a procedure used by *Bicycling* magazine specific to its postal survey, the second, used by the research agency Arbitron, is designed to encompass all of its survey work.

**Example** *Bicycling* magazine's procedure for increasing response to traditional mail surveys<sup>21</sup>

*Bicycling* magazine ([www.bicycling.com](http://www.bicycling.com)) conducts a semi-annual survey of individual bicycle dealers. The following procedure is used to increase the response to the survey:

- 1 An 'alert' letter is sent to advise the respondent that a questionnaire is coming.
- 2 A questionnaire package is posted five days after the 'alert' letter. The package contains a cover letter, a five-page questionnaire, a new \$1 bill, and a stamped addressed envelope.
- 3 A second package containing a reminder letter, a questionnaire and a stamped return envelope is posted five days after the first package.
- 4 A follow-up postcard is sent a week after the second package.
- 5 A second follow-up postcard is sent a week after the first.

**Example** Arbitron's response to low response rates<sup>22</sup>

Arbitron ([www.arbitron.com](http://www.arbitron.com)), a major marketing research agency, set about trying to improve response rates in order to improve the quality of its surveys. It created a cross-functional team of employees to evaluate the problem and to make recommendations. Their method was named the 'breakthrough method' and resulted in a complete change in the Arbitron approach to surveys. The team recommended six major strategies to improve response rates:

- 1 Maximise the effectiveness of placement/follow-up calls.
- 2 Make materials more appealing and easier to complete.
- 3 Increase Arbitron name awareness.
- 4 Improve survey participant rewards.
- 5 Optimise the arrival of respondent materials.
- 6 Increase usability of returned diaries.

Eighty initiatives were launched to implement these six strategies. As a result, response rates improved significantly. However, in spite of encouraging results, marketing researchers at Arbitron remain cautious. They appreciate that there is a daily challenge to maintain high response rates that encompasses everything they do in designing and implementing surveys.

**Not-at-homes.** The second major cause of low response rates is not-at-homes. In telephone and in-home personal interviews, low response rates can result if the potential respondents are not at home when contact is attempted. A study analysing 182 commercial telephone surveys involving a total sample of over 1 million consumers revealed that a large percentage of potential respondents were never contacted. The median non-contact rate was 40%. In nearly 40% of the surveys, only a single attempt was made to contact potential respondents. The results of 259,088 first-call attempts using a sophisticated random-digit dialling system show that less than 10% of the calls resulted in completed interviews, and 14.3% of those contacted refused to participate.<sup>23</sup>

The likelihood that potential respondents will not be at home varies with several factors. People with small children are more likely to be at home. Consumers are more likely to be at home at weekends than on weekdays and in the evening as opposed to during the afternoon. Pre-notification and appointments increase the likelihood that the respondent will be at home when contact is attempted.

The percentage of not-at-homes can be substantially reduced by employing a series of call-backs, or periodic follow-up attempts to contact non-respondents. The decision about the number of call-backs should weigh the benefits of reducing non-response bias against the additional costs. As call-backs are completed, the call-back respondents should be compared with those who have already responded to determine the usefulness of making further call-backs. In most consumer surveys, three or four call-backs may be desirable. Although the first call yields the most responses, the second and third calls have a higher response per call. It is important that call-backs be made and controlled according to a prescribed plan.

### Adjusting for non-response

Low response rates increase the probability that non-response bias will be substantial. Response rates should always be reported, and, whenever possible, the effects of non-response should be estimated. This can be done by linking the non-response rate to estimated differences between respondents and non-respondents. Information on differences between the two groups may be obtained from the sample itself. For example, differences found through call-backs could be extrapolated, or a concentrated follow-up could be conducted on a subsample of the non-respondents. Alternatively, it may be possible to estimate these differences from other sources.<sup>24</sup> To illustrate, in a survey of owners of vacuum cleaners, demographic and other information may be obtained for respondents and non-respondents from their guarantee cards. For a mail panel, a wide variety of information is available for both groups from syndicate organisations. If the sample is supposed to be representative of the general population, then comparisons can be made with census figures. Even if it is not feasible to estimate the effects of non-response, some adjustments can still be made during data analysis and interpretation.<sup>25</sup> The strategies available to adjust for non-response error include subsampling of non-respondents, replacement, substitution, subjective estimates, trend analysis, simple weighting and imputation.

**Subsampling of non-respondents.** Subsampling of non-respondents, particularly in the case of mail surveys, can be effective in adjusting for non-response bias. In this technique, the researcher contacts a subsample of the non-respondents, usually by means of telephone or personal interviews. This often results in a high response rate within that subsample. The values obtained for the subsample are then projected to all the non-respondents, and the survey results are adjusted to account for non-response. This method can estimate the effect of non-response on the characteristic of interest.

**Replacement.** In replacement, the non-respondents in the current survey are replaced with non-respondents from an earlier, similar survey. The researcher attempts to contact these non-respondents from the earlier survey and administer the current survey questionnaire to them, possibly by offering a suitable incentive. It is important that the nature of non-response in the current survey be similar to that of the earlier survey. The two surveys should use similar kinds of respondents, and the time interval between them should be short. As an example, as the Formula One Racetrack Project is repeated each year, the non-respondents in the present survey may be replaced by the non-respondents in the original survey.

#### Substitution

A procedure that substitutes for non-respondents other elements from the sampling frame who are expected to respond.

**Substitution.** In **substitution**, the researcher substitutes for non-respondents other elements from the sampling frame who are expected to respond. The sampling frame is divided into subgroups that are internally homogeneous in terms of respondent characteristics but heterogeneous in terms of response rates. These subgroups are then used to

identify substitutes who are similar to particular non-respondents but dissimilar to respondents already in the sample. Note that this approach would not reduce non-response bias if the substitutes are similar to respondents already in the sample.

**Subjective estimates.** When it is no longer feasible to increase the response rate by subsampling, replacement or substitution, it may be possible to arrive at subjective estimates of the nature and effect of non-response bias. This involves evaluating the likely effects of non-response based on experience and available information. For example, married adults with young children are more likely to be at home than single or divorced adults or than married adults with no children. This information provides a basis for evaluating the effects of non-response due to not-at-homes in personal or telephone surveys.

#### Trend analysis

A method of adjusting for non-response in which the researcher tries to discern a trend between early and late respondents. This trend is projected to non-respondents to estimate their characteristic of interest.

**Trend analysis.** Trend analysis is an attempt to discern a trend between early and late respondents. This trend is projected to non-respondents to estimate where they stand on the characteristic of interest. For example, Table 15.4 presents the results of several waves of a mail survey. The characteristic of interest is money spent on shopping in supermarkets during the last two months. The known value of the characteristic for the total sample is given at the bottom of the table. The value for each successive wave of respondents becomes closer to the value for non-respondents. For example, those responding to the second mailing spent 79% of the amount spent by those who responded to the first mailing. Those responding to the third mailing spent 85% of the amount spent by those who responded to the second mailing. Continuing this trend, one might estimate that those who did not respond spent 91% [ $85 + (85 - 79)$ ] of the amount spent by those who responded to the third mailing. This results in an estimate of €252 ( $277 \times 0.91$ ) spent by non-respondents and an estimate of €288 [ $(0.12 \times 412) + (0.18 \times 325) + (0.13 \times 277) + (0.57 \times 252)$ ] for the average amount spent in shopping at supermarkets during the last two months for the overall sample. Note that the actual amount spent by the respondents was €230 rather than €252, and that the actual sample average was €275 rather than the €288 estimated by trend analysis. Although the trend estimates are wrong, the error is smaller than the error that would have resulted from ignoring the non-respondents. Had the non-respondents been ignored, the average amount spent would have been estimated at €335 [ $(0.12 \times 412) + (0.18 \times 325) + (0.13 \times 277)$ ]/ $(0.12 + 0.18 + 0.13)$  for the sample.

**Table 15.4** Use of trend analysis in adjusting for non-response

	Percentage response	Average euro expenditure	Percentage of previous wave's response
First mailing	12	412	–
Second mailing	18	325	79
Third mailing	13	277	85
Non-response	(57)	(230)	91
Total	100	275	

#### Weighting

A statistical procedure that attempts to account for non-response by assigning differential weights to the data depending on the response rates.

**Weighting.** Weighting attempts to account for non-response by assigning differential weights to the data depending on the response rates.<sup>26</sup> For example, in a survey on PC, the sample was stratified according to income. The response rates were 85%, 70% and 40%, respectively, for the high-, medium- and low-income groups. In analysing the data, these subgroups are assigned weights inversely proportional to their response rates. That is, the weights assigned would be 100/85, 100/70 and 100/40, respectively, for the high-, medium- and low-income groups. Although weighting can correct for the

differential effects of non-response, it destroys the self-weighting nature of the sampling design and can introduce complications. Weighting is further discussed in Chapter 17 on data preparation.

**Imputation**

A method to adjust for non-response by assigning the characteristic of interest to the non-respondents based on the similarity of the variables available for both non-respondents and respondents.

**Imputation.** **Imputation** involves imputing, or assigning, the characteristic of interest to the non-respondents based on the similarity of the variables available for both non-respondents and respondents.<sup>27</sup> For example, a respondent who does not report brand usage may be imputed based on the usage of a respondent with similar demographic characteristics. Often there is a high correlation between the characteristic of interest and some other variables. In such cases, this correlation can be used to predict the value of the characteristic for the non-respondents (see Chapter 14).



## International marketing research

When conducting marketing research in international environments, statistical estimation of sample size may be difficult because estimates of the population variance in particular countries may be unavailable. Hence, the sample size is often determined by qualitative considerations, as discussed in Chapter 14: (1) the importance of the decision, (2) the nature of the research, (3) the number of variables, (4) the nature of the analysis, (5) sample sizes used in similar studies, (6) incidence rates, (7) completion rates, and (8) resource constraints. If statistical estimation of sample size is at all attempted, it should be realised that the estimates of the population variance may vary from country to country. For example, in measuring consumer preferences, a greater degree of heterogeneity may be encountered in countries where consumer preferences are not well developed. Thus, it may be a mistake to assume that the population variance is the same or to use the same sample size in different countries as illustrated in the following example.

**Example**

### The Chinese take to the sky<sup>28</sup>

The airline industry seems to have a strong and promising market potential in China. The airline market in China is growing rapidly. With the equivalent of billions of euros spent, China is trying to satisfy surging demand and to catch up with the rest of the world. The domestic airline traffic is growing at a rate of up to 30% a year. Strong economic growth, surging foreign trade, heavy investment in new airports, and a desire to travel overseas after many years of restriction, have helped to fuel the boom. China is making rapid progress in increasing its fleet and training pilots. For millions of Chinese, air travel is a relatively new experience and many more millions have never flown. Hence, Chinese preferences for air travel are likely to exhibit much more variability compared with Europeans. In a survey to compare attitudes towards air travel in China and European countries, the sample size of the Chinese survey would have to be larger than the European survey in order for the two survey estimates to have comparable precision.

It is important to realise that survey response rates can vary widely across countries. In a 2000 business mail survey conducted in 22 countries, the response rates varied from a low of 7.1% in Hong Kong to a high of 42.1% in Denmark with the overall response rate being 20%. The study also analysed factors to help explain the differences in response rates. Factors included the cultural and geographical distance from the Netherlands (where the survey was mailed), the level of foreign sales, number of employees and size of corporation.<sup>29</sup>



## Ethics in marketing research

As discussed in this chapter, statistical methods can be used to determine the sample size and, therefore, have an impact on the cost of the project. While this is usually an objective way of determining the sample size, it is, nonetheless, susceptible to fraud. The sample size is heavily dependent on the standard deviation of the variable and there is no way of knowing the standard deviation until the data have been collected. To resolve this paradox, the computation of the sample size must be performed using an estimate of the standard deviation. This estimate is derived based on secondary data, judgement or a small pilot study. By inflating the standard deviation, it is possible to increase the sample size and thus the project revenue. Using the sample size formula, it can be seen that increasing the standard deviation by 20%, for example, will increase the sample size by 44%. But this is clearly unethical.

Ethical dilemmas can arise even when the standard deviation is estimated honestly. It is possible, indeed common, that the standard deviation in the actual study is different from that estimated initially. When the standard deviation is larger than initially estimated, the confidence interval will also be larger than desired. When such a situation arises, the researcher has the responsibility to disclose this to the client and jointly decide on a course of action. The ethical ramifications of mis-communicating the confidence intervals of survey estimates based on statistical samples are underscored in political polling.

### Example

#### Surveys serve up elections<sup>30</sup>

The dissemination of some survey results has been strongly criticised as manipulative and unethical. In particular, the ethics of releasing political poll results before and during the election have been questioned. Opponents of such surveys claim that voters are misled by these results. First, before the election, voters are influenced by whom the polls predict will win. If they see that the candidate they favour is trailing, they may decide not to vote; they assume that there is no way their candidate can win. The attempt to predict the election results while the election is in progress has come under even harsher criticism. Opponents of this practice feel that this predisposes voters to vote for the projected winner or that it may even discourage voters from voting, even though the polls have not closed, because the media projects that there is already a winner. Furthermore, not only are the effects of these projections questionable, but frequently the accuracy of the projections is questionable as well. Although voters may be told a candidate has a certain percentage of the votes within 1 per cent, the confidence interval may be much larger, depending on the sample size.

Researchers also have the ethical responsibility to investigate the possibility of non-response bias, and make a reasonable effort to adjust for non-response. The methodology adopted and the extent of non-response bias found should be clearly communicated.



## Internet and computer applications

The main use of the Internet in sample size calculations is to track down potential sampling frames that could be used to define and classify a population. With different sampling frames collected and 'cleaned' in a database package, the ultimate population size can be determined. If there is a finite size to a population, the Internet can play a vital role in tracking down all elements of that population.

Using database packages to record the identity of survey respondents, the researcher can keep track of non-respondents. The database can help to determine whether there are particular geographic locations or types of non-respondent that are problematic. The rapid identification of non-respondents enables researchers to develop tactics to encourage a response.

The use of the Internet itself is thought to increase response rates for certain types of respondent. Primarily this is because of the ease of the survey process. Because the length of a survey on the Internet is hidden from respondents, they are less likely not to respond due to survey length. The inclusion of electronic skip patterns in the questionnaire design also helps to promote higher response rates by making it easier for the respondent to manoeuvre through the survey. Internet surveys can also use attractive design and plug-ins such as music and video to make the process more interesting.

For electronic mail surveys, prior emails sent by the sponsor to announce the survey are strongly suggested by many of the Internet marketing research firms. These firms also send reminder emails to non-respondents a few days after the initial survey has been sent. Another way to motivate respondents to complete surveys is to tell them whether and the extent to which a service will be enhanced. Another motivator is to provide incentives for completion of surveys. A Research Information Chat Board request on how to recruit respondents for web-based interviewing drew a note that a respondent had participated in a survey because he would be entered into a 'cash-draw' for participating. Another technique cited was to give points redeemable for merchandise or discounts to survey respondents.

There are a number of websites on the Internet that offer free use of sample size and confidence interval calculators. A simple web search will reveal a great array of sources from software designers such as [www.surveysystem.com](http://www.surveysystem.com) (in Research Aids) to statistics faculties in universities. You will be expected to specify a desired precision level, confidence level and population variance. By incorporating the cost of each sampling unit, the sample size can be adjusted based upon budget considerations.

## Summary



The statistical approaches to determining sample size are based on confidence intervals. These approaches may involve the estimation of the mean or proportion. When estimating the mean, determination of sample size using the confidence interval approach requires the specification of precision level, confidence level and population standard deviation. In the case of proportion, the precision level, confidence level and an estimate of the population proportion must be specified. The sample size determined statistically represents the final or net sample size that must be achieved. To achieve this final sample size, a much greater number of potential respondents have to be contacted to account for reduction in response due to incidence rates and completion rates.

Non-response error arises when some of the potential respondents included in the sample do not respond. The primary causes of low response rates are refusals and not-at-homes. Refusal rates may be reduced by prior notification, incentives, excellent questionnaire design and administration, and follow-up. The percentage of not-at-homes can be substantially reduced by call-backs. Adjustments for non-response can be made by subsampling non-respondents, replacement, substitution, subjective estimates, trend analysis, simple weighting and imputation.

The statistical estimation of sample size is even more complicated in international marketing research because the population variance may differ from one country to the next. The preliminary estimation of population variance for the purpose of determining the sample size also has ethical ramifications.

## Questions



- 1 Define:
  - a the sampling distribution
  - b finite population correction
  - c confidence intervals.
- 2 What is the standard error of the mean?
- 3 What is the procedure for constructing a confidence interval around a mean?
- 4 Describe the difference between absolute precision and relative precision when estimating a population mean.
- 5 How do the degree of confidence and the degree of precision differ?
- 6 Describe the procedure for determining the sample size necessary to estimate a population mean, given the degree of precision and confidence and a known population variance. After the sample is selected, how is the confidence interval generated?
- 7 Describe the procedure for determining the sample size necessary to estimate a population mean, given the degree of precision and confidence but where the population variance is unknown. After the sample is selected, how is the confidence interval generated?
- 8 How is the sample size affected when the absolute precision with which a population mean is estimated is doubled?
- 9 How is the sample size affected when the degree of confidence with which a population mean is estimated is increased from 95% to 99%?
- 10 Define what is meant by absolute precision and relative precision when estimating a population proportion.
- 11 Describe the procedure for determining the sample size necessary to estimate a population proportion given the degree of precision and confidence. After the sample is selected, how is the confidence interval generated?
- 12 How can the researcher ensure that the generated confidence interval will be no larger than the desired interval when estimating a population proportion?
- 13 When several parameters are being estimated, what is the procedure for determining the sample size?
- 14 Define incidence rate and completion rate. How do these rates affect the determination of the final sample size?
- 15 What strategies are available for adjusting for non-response?

## Exercises

- Using a spreadsheet (e.g. Excel), program the formulae for determining the sample size under the various approaches described in this chapter.
- Using the website of a major newspaper in your country, search for reports of three recent major surveys. Write a report on the sample sizes used and the extent to which details of precision, confidence levels and any other factors affecting the sample were reported. Note any reporting that sensationalises statistical changes over time that are within the margin of error tolerances.
- A major electric utility company would like to determine the average amount spent per household for air-conditioning during summer months. From its own records it knows how much electricity is consumed per household but not how much is spent on particular appliances and the attitudes towards the use of those appliances. Thus management believe that a survey should be conducted. What procedure would you recommend for determining the sample size?
- You work as the marketing research manager for a chain of theme restaurants. A new menu has been developed based upon organic and fair-trade produce. Before the new menu is introduced, management are concerned about how existing and potential customers will react. How would you approach the sample size calculations for this task? Present your plan to a group of students representing the board of the chain.
- In a small group discuss the following issues: 'Statistical considerations are more important than administrative considerations in determining sample size' and 'The real determinant of sample size is what managers feel confident with; it has little to do with statistical confidence.'

### Video Case Exercise: Subaru

Evaluate the reasons for the high response rates to Subaru surveys. What lessons of Subaru's success can be generalised to other survey designs.



download from  
www.pearsoned.co.uk/  
malhotra\_euro

## Appendix: The normal distribution

In this appendix, we provide a brief overview of the normal distribution and the use of the normal distribution table. The normal distribution is used in calculating the sample size, and it serves as the basis for classical statistical inference. Many continuous phenomena follow the normal distribution or can be approximated by it. The normal distribution can, likewise, be used to approximate many discrete probability distributions.<sup>31</sup>

The normal distribution has some important theoretical properties. It is bell shaped and symmetrical in appearance. Its measures of central tendency (mean, median and mode) are all identical. Its associated random variable has an infinite range ( $-\infty < x < +\infty$ ).

The normal distribution is defined by the population mean  $\mu$  and population standard deviation  $\sigma$ . Since an infinite number of combinations of  $\mu$  and  $\sigma$  exist, an infinite number of normal distributions exist and an infinite number of tables would be required. By standardising the data, however, we need only one table, such as Table 2 in the Appendix of statistical tables. Any normal random variable  $X$  can be converted to a standardised normal random variable  $z$  by the formula

$$z = \frac{X - \mu}{\sigma}$$

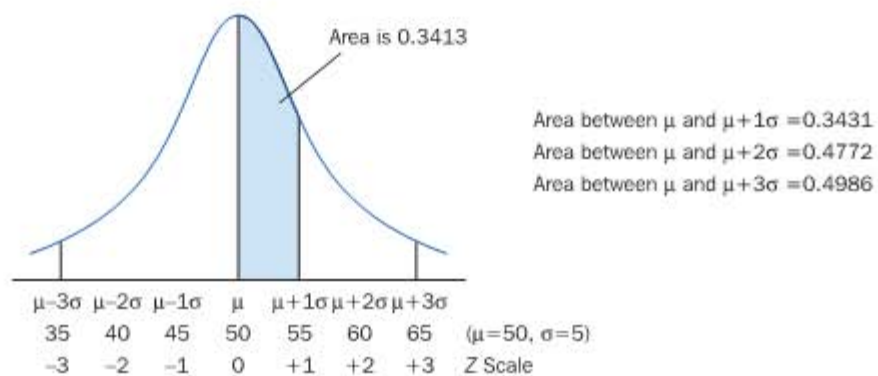


Note that the random variable  $z$  is always normally distributed with a mean of 0 and a standard deviation of 1. The normal probability tables are generally used for two purposes: (1) finding probabilities corresponding to known values of  $X$  or  $z$ , and (2) finding values of  $X$  or  $z$  corresponding to known probabilities. Each of these uses is discussed.

### Finding probabilities corresponding to known values

Suppose that Figure 15A.1 represents the distribution of the number of engineering contracts received per year by an engineering firm. Because the data span the entire history of the firm, Figure 15A.1 represents the population. Therefore, the probabilities or proportion of area under the curve must add up to 1.0. The marketing director wishes to determine the probability that the number of contracts received next year will be between 50 and 55. The answer can be determined by using Table 2 of the Appendix of statistical tables.

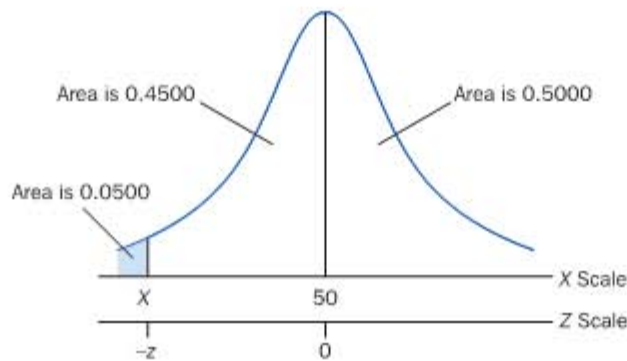
Table 2 gives the probability or area under the standardised normal curve from the mean (zero) to the standardised value of interest,  $z$ . Only positive entries of  $z$  are listed in the table. For a symmetrical distribution with zero mean, the area from the mean to  $+z$  (i.e.  $z$  standard deviations above the mean) is identical to the area from the mean to  $-z$  ( $z$  standard deviations below the mean).



**Figure 15A.1**  
Finding probability corresponding to a known value

Note that the difference between 50 and 55 corresponds to a  $z$  value of 1.00. Note that, to use Table 2, all  $z$  values must be recorded to two decimal places. To read the probability or area under the curve from the mean to  $z = +1.00$ , scan down the  $z$  column of Table 2 until the  $z$  value of interest (in tenths) is located. In this case, stop in the row  $z = 1.00$ . Then read across this row until you intersect the column containing the hundredths place of the  $z$  value. Thus, in Table 2, the tabulated probability for  $z = 1.00$  corresponds to the intersection of the row  $z = 1.0$  with the column  $z = 0.00$ . This probability is 0.3413. As shown in Figure 15A.1, the probability is 0.3413 that the number of contracts received by the firm next year will be between 50 and 55. It can also be concluded that the probability is 0.6826 ( $2 \times 0.3413$ ) that the number of contracts received next year will be between 45 and 55.

This result could be generalised to show that for any normal distribution the probability is 0.6826 that a randomly selected item will fall within 1 standard deviation above or below the mean. Also, it can be verified from Table 2 that there is a 0.9544 probability that any randomly selected, normally distributed observation will fall within  $\pm 2$  standard deviations above or below the mean, and a 0.9973 probability that the observation will fall within  $\pm 3$  standard deviations above or below the mean.



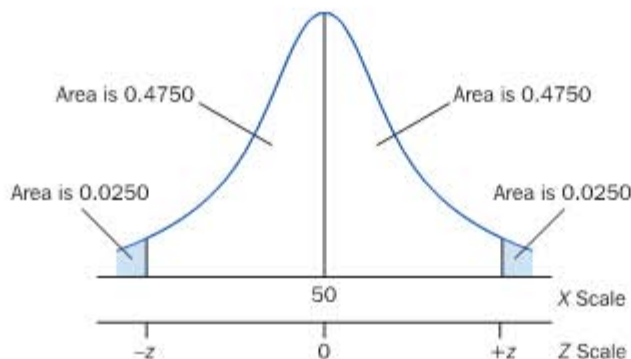
**Figure 15A.2**  
Finding values  
corresponding to  
known probabilities

### Finding values corresponding to known properties values

Suppose that the marketing director wishes to determine how many contracts must come in so that 5% of the contracts for the year have come in. If 5% of the contracts have come in, 95% of the contracts have yet to come. As shown in Figure 15A.2, this 95% can be broken down into two parts: contracts above the mean (i.e. 50%) and contracts between the mean and the desired  $z$  value (i.e. 45%). The desired  $z$  value can be determined from Table 2, since the area under the normal curve from the standardised mean, 0, to this  $z$  must be 0.4500. From Table 2, we search for the area or probability 0.4500. The closest value is 0.4495 or 0.4505. For 0.4495, we see that the  $z$  value corresponding to the particular  $z$  row (1.6) and  $z$  column (0.04) is 1.64. The  $z$  value, however, must be recorded as negative (i.e.  $z = -1.64$ ), since it is below the standardised mean of 0. Similarly, the  $z$  value corresponding to the area of 0.4505 is  $-1.65$ . Since 0.4500 is midway between 0.4495 and 0.4505, the appropriate  $z$  value could be midway between the two  $z$  values and estimated as  $-1.645$ . The corresponding  $X$  value can then be calculated from the standardisation formula, as follows:

$$\begin{aligned} X &= \mu + z\sigma \\ &= 50 + (-1.645)5 \\ &= 41.775 \end{aligned}$$

Suppose that the marketing director wanted to determine the interval in which 95% of the contracts for the next year are expected to lie. As can be seen from Figure 15A.3, the corresponding  $z$  values are  $\pm 1.96$ . This corresponds to  $X$  values of  $50 \pm (1.96)5$ , or 40.2 and 59.8. this range represents the 95% confidence interval.



**Figure 15A.3**  
Finding values  
corresponding to  
known probabilities:  
confidence interval

## Notes

- 1 A discussion of the sampling distribution may be found in any basic statistics textbook. For example, see Berenson, M.L., Levine, D.M. and Krehbiel, T., *Basic Business Statistics: Concepts and Applications*, 8th edn (Englewood Cliffs, NJ: Prentice Hall, 2002).
- 2 Other statistical approaches are also available. A discussion of these is beyond the scope of this book, however. The interested reader is referred to Reynolds, M.R. Jr., 'EWMA control charts with variable sample sizes and variable sampling intervals', *IEE Transactions* 33 (6) (June 2001), 511–530; Sampath, S., *Sampling theory and methods* (Boca Raton, FL: CRC Press, 2000); Yeh, L. and Van, L.C., 'Bayesian double-sampling plans with normal distributions', *Statistician* 46 (2) (1997), 193–207; Blyth, W.G. and Marchant, L.J., 'A self-weighting random sampling technique', *Journal of the Market Research Society* 38 (4) (October 1996), 473–479; Nowell, C. and Stanley, L.R., 'Length-biased sampling in mall intercept surveys', *Journal of Marketing Research* 28 (November 1991), 475–479; Gillett, R., 'Confidence interval construction by Stein's method: a practical and economical approach to sample size determination', *Journal of Marketing Research* 26 (May 1989), 237.
- 3 Thompson, S.K., *Sampling* (New York: Wiley, 2002); Wall, M.M., 'An effective confidence interval for the mean with samples of size one and two', *American Statistician* (May 2001), 102–105; Chow, S.L., *Statistical Significance* (Thousand Oaks, CA: Sage, 1996).
- 4 Valliant, R.L., Dorfman, A.H. and Royall, R.M., *Finite Population Sampling and Inference: A Prediction Approach* (New York: Wiley, 2000).
- 5 Joseph, L. and Wolfson, D.B., 'Interval based versus decision-theoretic criteria for the choice of a sample size', *Statistician* 46 (2) (1997), 145–149; Frankel, M., 'Sampling theory', in Rossi, P.H., Wright, J.D. and Anderson, A.B. (eds), *Handbook of Survey Research* (New York: Academic Press, 1983), 21–67.
- 6 'City of Los Angeles Internet Services Project: Market Analysis and Best Practices Report', *e-Government Services Project Reports* (29 October 1999).
- 7 For a discussion of estimating sample costs, see Kish, L., *Survey Sampling* (New York: Wiley, 1965); Sudman, S., *Applied Sampling* (New York: Academic Press, 1976).
- 8 See, for example, Sampath, S., *Sampling theory and methods* (Boca Raton, FL: CRC Press, 2000); Bradley, N., 'Sampling for Internet surveys: an examination of respondent selection for internet research', *Journal of the Market Research Society* 41 (4) (October 1999), 387–395; Adcock, C.J., 'Sample size determination a review', *Statistician* 46 (2) (1997), 261–283; Sudman, S., 'Applied sampling', in Rossi, P.H., Wright, J.D. and Anderson, A.B. (eds), *Handbook of Survey Research* (New York: Academic Press, 1983), 145–194.
- 9 Adjusting for incidence and completion rates is discussed in Rao, P.S., *Sampling Methodologies with Applications* (Boca Raton, FL: CRC Press, 2001); Bickart, B., 'The distribution of survey contact and participation in the United States: constructing a survey based estimate', *Journal of Marketing Research* 36 (2) (May 1999), 286–294; Dillman, D.A., Singer, E., Clark, J.R. and Treat, J.B., 'Effects of benefits appeals, mandatory appeals, and variations in statements of confidentiality on completion rates for census questionnaires', *Public Opinion Quarterly* 60 (3) (Fall 1996), 376–389; Pol, L.G. and Pak, S., 'The use of two stage survey design in collecting data from those who have attended periodic or special events', *Journal of the Market Research Society* 36 (October 1994), 315–326.
- 10 Lee, K.G., 'Incidence is a key element', *Marketing News* (13 September 1985), 50.
- 11 Van Kenhove, P., 'The influence of topic involvement on mail-survey response behaviour', *Psychology & Marketing* 19 (3) (March 2002), 293; Fisher, M.R., 'Estimating the effect of nonresponse bias on angler surveys', *Transactions of the American Fisheries Society* 125 (1) (January 1996), 118–126; Martin, C., 'The impact of topic interest on mail survey response behaviour', *Journal of the Market Research Society* 36 (October 1994), 327–338.
- 12 Cummings, S.M., 'Reported response rates to mailed physician questionnaires', *Health Services Research* 35 (6) (February 2001), 1347–1355; Hill, A., Roberts, J., Ewings, P. and Gunnell, D., 'Nonresponse bias in a lifestyle survey', *Journal of Public Health Medicine* 19 (2) (June 1997), 203–207; McDaniel, S.W., Madden, C.S. and Verille, P., 'Do topic differences affect survey non-response?', *Journal of the Market Research Society* (January 1987), 55–66.
- 13 For minimising the incidence of non-response and adjusting for its effects, see Columbo, R., 'A model for diagnosing and reducing non-response bias', *Journal of Advertising Research* 40 (1/2) (January/April 2000), 85–93; Chen, H.C., 'Direction, magnitude, and implications of nonresponse bias in mail surveys', *Journal of the Market Research Society* 38 (3) (July 1996), 267–276; Brown, M., 'What price response?', *Journal of the Market Research Society* 36 (July 1994), 227–244.
- 14 Jarvis, S., 'CMOR finds survey refusal rate still rising', *Marketing News* 36 (3) (4 February 2002), 4; Baldauf, A., 'Examining motivations to refuse in industrial mail surveys', *Journal of the Market Research Society* 41 (3) (July 1999), 345–353; Baker, R., 'Nobody's talking', *Marketing Research: A magazine of management and applications* 8 (1) (Spring 1996), 22–24; Struebbe, J.M., Kernan, J.B. and Grogan, T.J., 'The refusal problem in telephone surveys', *Journal of Advertising Research* (June/July 1986), 29–38.
- 15 Havermans, J., 'Strict rules for telemarketers may help research', *Research World* (June 2004), 14.
- 16 Van Kenhove, P., 'The influence of topic involvement on mail-survey response behaviour', *Psychology & Marketing* 19 (3) (March 2002), 293; Groves, R.M., 'Leverage-saliency theory of survey participation: description and an illustration', *Public Opinion Quarterly* 64 (3) (Fall 2000), 299–308; Everett, S.A., Price, J.H., Bedell, A.W. and Telljohann, S.K., 'The effect of a monetary incentive in increasing the return rate of a survey of family physicians', *Evaluation and the Health Professions* 20 (2) (June 1997), 207–214; Armstrong, J.S. and Lusk, E.J., 'Return postage in mail surveys: a meta-analysis', *Public Opinion Quarterly* (Summer 1987), 233–248; Yu, J. and Cooper, H., 'A quantitative review of research design effects on response rates to questionnaires', *Journal of Marketing Research* 20 (February 1983), 36–44.
- 17 Brennan, M., Benson, S. and Kearns, Z., 'The effect of introductions on telephone survey participation rates', *International Journal of Market Research* 47 (1) (January 2005) 65–74.

- 18 Shaw, M.J., 'The use of monetary incentives in a community survey: impact on response rates, data, quality and cost', *Health Services Research* 35 (6) (February 2001), 1339–1346; Wayman, S., 'The buck stops here when it comes to dollar incentives', *Marketing News* 31 (1) (6 January 1997), 9; Biner, P.M. and Kidd, H.J., 'The interactive effects of monetary incentive justification and questionnaire length on mail survey response rates', *Psychology and Marketing* 11 (5) (September/October 1994), 483–492.
- 19 Zafer Erdogan, B., 'Increasing mail survey response rates from an industrial population: a cost effectiveness analysis of four follow-up techniques', *Industrial Marketing Management* 31 (1) (January 2002), 65.
- 20 Byrom, J., 'The effect of personalization on mailed questionnaire response rates', *International Journal of Market Research* (Summer 2000), 357–359; Dillman, D.A., Singer, E., Clark, J.R. and Treat, J. LB., 'Effects of benefits appeals, mandatory appeals, and variations in statements of confidentiality on completion rates for census questionnaires', *Public Opinion Quarterly* 60 (3) (Fall 1996), 376–389; Gendall, P., Hoek, J. and Esslemont, D., 'The effect of appeal, complexity and tone in a mail survey covering letter', *Journal of the Market Research Society* 37 (3) (July 1995), 251–268; Greer, T.V. and Lohtia, R., 'Effects of source and paper color on response rates in mail surveys', *Industrial Marketing Management* 23 (February 1994), 47–54.
- 21 *Bicycling Magazine's 1997 Semi-annual Study of US Retail Bicycle Stores* (September 1997).
- 22 Smith, J., 'How to boost DM response rates quickly', *Marketing News* 35 (9) (23 April 2001), 5; Peacock, J.D., 'Yes, you can raise response rates', *Journal of Advertising Research* 36 (1) (January 1996), RC7–RC10.
- 23 Keeter, S., 'Consequences of reducing non-response in a national telephone survey', *Public Opinion Quarterly*, 64 (2) (Summer 2000), 125–148; Bowen, G.L., 'Estimating the reduction in nonresponse bias from using a mail survey as a backup for nonrespondents to a telephone interview survey', *Research on Social Work Practice* 4 (1) (January 1994), 115–128; Kerin, R.A. and Peterson, R.A., 'Scheduling telephone interviews', *Journal of Advertising Research* (May 1983), 44.
- 24 Columbo, R., 'A model for diagnosing and reducing nonresponse bias', *Journal of Advertising Research*, (January/April 2000), 85–93; Rowland, M.L. and Forthofer, R.N., 'Adjusting for nonresponse bias in a health examination survey', *Public Health Reports* 108 (3) (May–June 1993), 380–386.
- 25 Larsen, M.D., 'The psychology of survey response', *Journal of the American Statistical Association* 97 (457) (March 2002), 358–359; Dey, E.L., 'Working with low survey response rates the efficacy of weighting adjustments', *Research in Higher Education* 38 (2) (April 1997), 215–227.
- 26 Qin, J., 'Estimation with survey data under non-ignorable non-response or informative sampling', *Journal of the American Statistical Association* 97 (457) (March 2002), 193–200; Kessler, R.C., Little, R.J. and Grover, R.M., 'Advances in strategies for minimising and adjusting for survey nonresponse', *Epidemiologic Reviews* 17 (1) (1995), 192–204; Ward, J.C., Russick, B. and Rudelius, W., 'A test of reducing call-backs and not-at-home bias in personal interviews by weighting at-home respondents', *Journal of Marketing Research* 2 (February 1985), 66–73.
- 27 Brewer, K., *Design and estimation in survey sampling* (London; Edward Arnold, 2001); Sao, J., 'Variance estimation for survey data with composite imputation and non-negligible sampling fractions', *Journal of American Statistical Association* (March, 1999), 254–265; Drane, J.W., Richter, D. and Stoskopf, C., 'Improved imputation of nonresponse to mailback questionnaires', *Statistics in Medicine* 12 (34) (February 1993), 283–288.
- 28 Dolven, B., 'The best little airline in China', *Far Eastern Economic Review* 165 (2) (17 January 2002), 32–35; Tse, A., 'Estimating the design factor for surveys in Hong Kong', *Marketing Intelligence and Planning* 13 (9) (1995), 28–29.
- 29 Harzing, A.-W., 'Cross-national industrial mail surveys: why do response rates differ between countries?', *Industrial Marketing Management* 29 (3) (May 2000), 243–254.
- 30 Taylor, H., 'Using Internet polling to forecast the 2000 elections', *Marketing Research* 13 (1) (Spring 2001), 26–30; Morwitz, V.G. and Pluzinski, C., 'Do polls reflect opinions or do opinions reflect polls? The impact of political polling on voters' expectations, preferences and behaviour', *Journal of Consumer Research* 23 (1) (June 1996), 53–67.
- 31 This material is drawn from Berenson, M.L., Levine, D.M. and Krehbiel, T., *Basic Business Statistics: Concepts and Applications* 8th edn (Upper Saddle River, NJ: Prentice Hall, 2002).

Visit the *Marketing Research Companion Website* at [www.pearsoned.co.uk/mathotra\\_euro](http://www.pearsoned.co.uk/mathotra_euro) for additional learning resources including annotated weblinks, an online glossary and a suite of downloadable video cases.