Source: © Getty Images

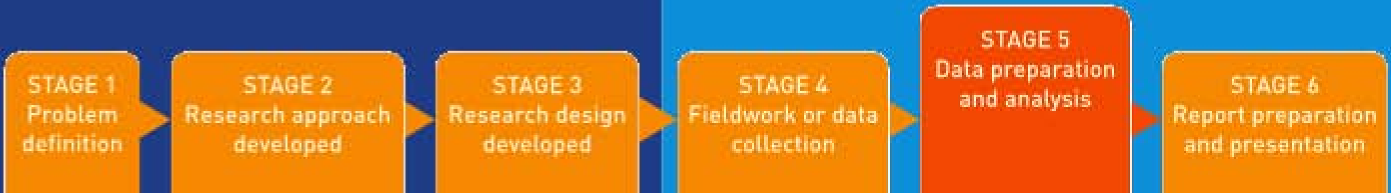
17 Data preparation

> *Perhaps the most neglected series of activities in the marketing research process. Handled with care, data preparation can substantially enhance the quality of statistical results.*

Objectives

After reading this chapter, you should be able to:

1. discuss the nature and scope of data preparation and the data preparation process;
2. explain questionnaire checking and editing and the treatment of unsatisfactory responses by returning to the field, assigning missing values and discarding unsatisfactory responses;
3. describe the guidelines for coding questionnaires, including the coding of structured and unstructured questions;
4. discuss the data cleaning process and the methods used to treat missing responses: substitution of a neutral value, imputed response, casewise deletion and pairwise deletion;
5. state the reasons for and methods of statistically adjusting data: weighting, variable respecification and scale transformation;
6. describe the procedure for selecting a data analysis strategy and the factors influencing the process;
7. classify statistical techniques and give a detailed classification of univariate techniques as well as a classification of multivariate techniques;
8. understand the intra-cultural, pan-cultural and cross-cultural approaches to data analysis in international marketing research;
9. identify the ethical issues related to data processing, particularly the discarding of unsatisfactory responses, violation of the assumptions underlying the data analysis techniques, and evaluation and interpretation of results and paired samples.



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

Decisions related to data preparation and analysis should not take place after data have been collected. Before the raw data contained in the questionnaires can be subjected to statistical analysis, they must be converted into a form suitable for analysis. The suitable form and the means of analysis should be considered as a research design is developed. This ensures that the output of the analyses will satisfy the research objectives set for a particular project.

The care exercised in the data preparation phase has a direct effect upon the quality of statistical results and ultimately the support offered to marketing decision-makers. Paying inadequate attention to data preparation can seriously compromise statistical results, leading to biased findings and incorrect interpretation.

This chapter describes the data collection process, which begins with checking the questionnaires for completeness. Then we discuss the editing of data and provide guidelines for handling illegible, incomplete, inconsistent, ambiguous or otherwise unsatisfactory responses. We also describe coding, transcribing and data cleaning, emphasising the treatment of missing responses and statistical adjustment of data. We discuss the selection of a data analysis strategy and classify statistical techniques. The intra-cultural, pan-cultural and cross-cultural approaches to data analysis in international marketing research are explained. Finally, the ethical issues related to data processing are identified with emphasis on discarding unsatisfactory responses, violation of the assumptions underlying the data analysis techniques, and evaluation and interpretation of results.

We begin with an illustration of the data preparation process set in the context of Formula One Racetrack Project.

Focus on

Sports Marketing Surveys

Data preparation

In the Formula One Racetrack Project, the data were obtained by face-to-face and telephone interviews. As the questionnaire was developed and finalised a preliminary plan was drawn up of how the findings could be analysed. The questionnaires were edited by a supervisor as they were being returned from Australia, Brazil, France, Germany, Italy, Japan, Spain and the UK. The questionnaires were checked for incomplete, inconsistent and ambiguous responses. Questionnaires with problematic responses were queried with supervisors in each of the eight countries. In some circumstances, the supervisors were asked to recontact the respondents to clarify certain issues. Thirty-five questionnaires were discarded because the proportion of unsatisfactory responses rendered them too poor to use. This resulted in a final sample size of 2,050.

A codebook was developed for coding the questionnaires; this was done automatically as the questionnaire was designed using the SNAP software (fully integrated survey software with on-screen questionnaire design, data collection and analysis for all types of surveys, **www.snapsurveys.com**). The data were transcribed by being directly keyed in as telephone interviews were conducted and onto personal digital assistants (PDAs) as face-to-face interviews were conducted. The SNAP software has a built-in error check that identifies out-of-range responses. About 10% of the data were verified for other data entry errors. The data were cleaned by identifying logically inconsistent responses. Most of the rating information was obtained using five-point scales, so responses of 0, 6 and 7 were considered

→

out of range and a code of 9 was assigned to missing responses. If an out-of-range response was keyed in, the SNAP software package did not allow any continuation of data entry; an audible warning was made. New variables that were composites of original variables were created. Finally, a data analysis strategy was developed.

The Formula One Racetrack example describes the various phases of the data preparation process. Note that the process is initiated while the fieldwork is still in progress. A systematic description of the data preparation process follows.

The data preparation process

Editing
A review of the questionnaires with the objective of increasing accuracy and precision.

Coding
Assigning a code to represent a specific response to a specific question along with the data record and column position that the code will occupy.

The data preparation process is shown in Figure 17.1. The entire process is guided by the preliminary plan of data analysis that was formulated in the research design phase. The first step is to check for acceptable questionnaires. This is followed by **editing**, **coding** and transcribing the data. The data are cleaned and a treatment for missing responses is prescribed. Often, after the stage of sample validation, statistical adjustment of the data may be necessary to make them representative of the population of interest. The researcher should then select an appropriate data analysis strategy. The final data analysis strategy differs from the preliminary plan of data analysis due to the information and insights gained since the preliminary plan was formulated. Data preparation should begin as soon as the first batch of questionnaires is received from the field, while the fieldwork is still going on. Thus, if any problems are detected, the fieldwork can be modified to incorporate corrective action.

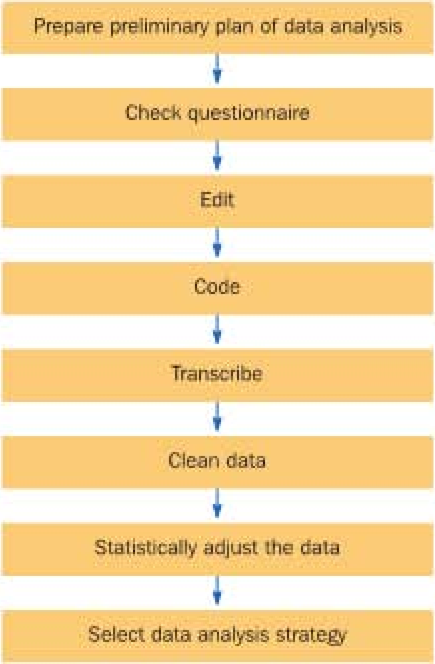

Prepare preliminary plan of data analysis
Check questionnaire
Edit
Code
Transcribe
Clean data
Statistically adjust the data
Select data analysis strategy

Figure 17.1
Data preparation process

Checking the questionnaire

The initial step in questionnaire checking involves reviewing all questionnaires for completeness and interviewing or completion quality as illustrated in the following example.

Example

Custom cleaning[1]

According to Johan Harristhal of Gfk Custom Research (**www.cresearch.com**), completed questionnaires from the field often have many small errors because of the inconsistent quality of interviewing. For example, qualifying responses are not circled, or skip patterns are not followed accurately. These small errors can be costly. When responses from such questionnaires are put into a computer, Custom Research runs a cleaning program that checks for completedness and logic. Discrepancies are identified on a computer printout, which is checked by the tabulation supervisors. Once the errors are identified, appropriate corrective action is taken before data analysis is carried out. Custom research has found that this procedure substantially increases the quality of statistical results.

Researchers do not just depend upon error checks at the data entry stage; checks should be made whilst fieldwork is still under way. If the fieldwork was contracted to a data collection agency, the researcher should make an independent check after it is over. A questionnaire returned from the field may be unacceptable for several reasons:

1. Parts of the questionnaire may be incomplete.
2. The pattern of responses may indicate that the respondent did not understand or follow the instructions. For example, filter questions may not have been followed.
3. The responses show little variance. For example, a respondent has ticked only 4s on a series of seven-point rating scales.
4. The returned questionnaire is physically incomplete: one or more pages is missing.
5. The questionnaire is received after the pre-established cut-off date.
6. The questionnaire is answered by someone who does not qualify for participation.

If quotas or cell group sizes have been imposed, the acceptable questionnaires should be classified and counted accordingly. Any problems in meeting the sampling requirements should be identified, and corrective action, such as conducting additional interviews in the under-represented cells, should be taken where this is possible, before the data are edited.

Editing

Editing is the review of the questionnaires with the objective of increasing accuracy and precision. It consists of screening questionnaires to identify illegible, incomplete, inconsistent or ambiguous responses. Responses may be illegible if they have been poorly recorded. This is particularly common in questionnaires with a large number of unstructured questions. The data must be legible if they are to be properly coded. Likewise, questionnaires may be incomplete to varying degrees. A few or many questions may be unanswered.

At this stage, the researcher makes a preliminary check for consistency. Certain obvious inconsistencies can be easily detected. For example, respondents may have answered a whole series of questions relating to their perceptions of a particular bank, yet in other questions may have indicated that they have not used that particular bank or even heard of it.

Responses to unstructured questions may be ambiguous and difficult to interpret clearly. The answer may be abbreviated, or some ambiguous words may have been used. For structured questions, more than one response may be marked for a question designed to elicit a single response. Suppose that a respondent circles 2 and 3 on a five-point rating scale. Does this mean that 2.5 was intended? To complicate matters further, the coding procedure may allow for only a single-digit response.

Treatment of unsatisfactory responses

Unsatisfactory responses are commonly handled by returning to the field to get better data, assigning missing values, and discarding unsatisfactory respondents.

Returning to the field. Questionnaires with unsatisfactory responses may be returned to the field, where the interviewers recontact the respondents. This approach is particularly attractive for business and industrial marketing surveys, where the sample sizes are small and the respondents are easily identifiable. The data obtained the second time, however, may be different from those obtained during the original survey. These differences may be attributed to changes over time or differences in the mode of questionnaire administration (e.g. telephone versus in-person interview).

Assigning missing values. If returning the questionnaires to the field is not feasible, the editor may assign missing values to unsatisfactory responses. This approach may be desirable if (1) the number of respondents with unsatisfactory responses is small; (2) the proportion of unsatisfactory responses for each of these respondents is small; or (3) the variables with unsatisfactory responses are not the key variables.

Discarding unsatisfactory respondents. In another approach, the respondents with unsatisfactory responses are simply discarded. This approach may have merit when (1) the proportion of unsatisfactory respondents is small (less than 10%); (2) the sample size is large; (3) the unsatisfactory respondents do not differ from satisfactory respondents in obvious ways (e.g. demographics, product usage characteristics); (4) the proportion of unsatisfactory responses for each of these respondents is large; or (5) responses on key variables are missing. Unsatisfactory respondents may differ from satisfactory respondents in systematic ways, however, and the decision to designate a respondent as unsatisfactory may be subjective. Both these factors bias the results. If the researcher decides to discard unsatisfactory respondents, the procedure adopted to identify these respondents and their number should be reported, as in the following example.

Example

Declaring 'discards'[2]

In a cross-cultural survey of marketing managers from English-speaking African countries, questionnaires were posted to 565 firms. A total of 192 completed questionnaires were returned, of which four were discarded because respondents suggested that they were not in charge of overall marketing decisions. The decision to discard the four questionnaires was based on the consideration that the sample size was sufficiently large and the proportion of unsatisfactory respondents was small.

Coding

Many questionnaire design and data entry software packages code data automatically. Examples of the options available will be presented in the Internet and computer applications section and on the Companion Website. Learning how to use such packages or even using spreadsheet packages means that the process of coding is now a much simpler task for the marketing researcher. Many of the principles of coding are based on the days of data processing using 'punched cards' or even, much more recently, DOS files. Whilst there may be many data analysts who could present coherent cases for the use of original forms of data entry, the greater majority of researchers enjoy the benefits of a simpler, speedier and less error-prone form of data entry, using proprietary software packages such as SNAP (**www.snapsurveys.com**) or Keypoint2 (**www.camsp.com**). The nature and importance of coding for qualitative data was introduced in Chapters 6 and 9. For quantitative data, which can include coping with open-ended responses or responses to 'Other – Please State…', it is still important to understand the principles of coding, as reference to the process is made by so many in the marketing research industry. The examples of coding presented are based on the Formula One Racetrack Project.

Coding means assigning a code, usually a number, to each possible answer to each question. For example, a question on the gender of respondents may be assigned a code of 1 for females and 2 for males. For every individual question in a questionnaire, the researcher decides which codes should be assigned to all its possible answers.

If the question posed has only two possible answers, the codes assigned of 1 or 2 take up one digit space. If the question posed had 25 possible answers such as *'Apart from Formula One, what other sports do you follow on TV or through any other media?'*, the possible answers and assigned codes of 1 to 25 would take up two digit spaces. The reason for focusing upon the digit spaces required for any particular question relates to the convention in marketing research to record the answers from individual questionnaire respondents in 'flat ASCII files'. Such files were typically 80 columns wide. The columns would be set out into 'fields', i.e. assigned columns that relate to specific questions. Thus the task for the researcher after assigning codes to individual question responses was to set out a consecutive series of fields or columns. These fields would represent where the answers to particular questions would be positioned in the ASCII file. In each row of a computer file would be the coded responses from individual questionnaire respondents. Each row is termed a 'record', i.e. all the fields that make up the response from one respondent. All the attitudinal, behavioural, demographic and other classification characteristics of a respondent may be contained in a single record.

Table 17.1 shows an extract from the Formula One Racetrack questionnaire and Table 17.2 illustrates the answers to these questions from a selection of respondents as set out in codes, fields and records. The classification questions set out on Table 17.1 were placed at the start of the questionnaire, forming the 'Screening' part of the questionnaire. It is followed by the first question in the section on 'Attitudes and Opinions towards F1'.

Question 2 has two possible answers, coded 1 to 2, that take up one digit space (there could be a space for 'Refused' to cope with the very rare occasions where the respondents refuse to state their gender and the interviewer cannot discern their gender. It is usually more prevalent in postal questionnaires. Question 3 has eight possible answers, coded 1 to 8, that take up one digit space. Question 4 has seven possible answers, coded 1 to 7, that take up one digit space. Note that if the actual number of Grands Prix viewed were entered (rather than a category) two digit spaces would be needed. Question 5 has 12 possible answers which are coded 01 to 12, taking up two digit spaces. Note that at the end of each question is a small number in parentheses. These numbers represent the first field positions of each question as illustrated in Table 17.2.

Table 17.1 Classification questions from the Formula One Racetrack survey

Question 1 – Have you watched at least one hour of a Formula One race on television in the 2003 season?			(5)
1	□	Yes	
2	□	No (terminate interview)	
Question 2 – Please enter gender			(6)
1	□	Male	
2	□	Female	
3	□	Refused	
Question 3 – Please could you tell me how old you are?			(7)
1	□	18–24	
2	□	25–29	
3	□	30–34	
4	□	35–39	
5	□	40–44	
6	□	45–49	
7	□	50–55	
8	□	Refused	
Question 4 – Out of 16 Formula One Grands Prix held in the 2003 season, how many have you watched on television?			(8)
1	□	1–2	
2	□	3–4	
3	□	5–6	
4	□	7–9	
5	□	10–12	
6	□	13–15	
7	□	16	

Question 5 – Which one Formula One team do you support? (DO NOT READ OUT AND TICK ONLY ONE.) (9–10)

01	□	Don't have a favourite team
02	□	Don't know
03	□	BAR (Honda)
04	□	Ferrari
05	□	Jaguar (Cosworth)
06	□	Jordan (Ford)
07	□	McLaren (Mercedes)
08	□	European Minardi
09	□	Renault
10	□	Sauber
11	□	Toyota
12	□	Williams (BMW)

Table 17.2 **Illustrative computer file held on a flat ASCII file**

	Fields					
Records	*1–4*	*5*	*6*	*7*	*8*	*9–10*
Record #1	0001	1	1	2	2	01
Record #11	0011	1	1	3	1	03
Record #21	0021	1	2	7	6	04
Record #2050	2050	1	1	5	4	11

In Table 17.2, the columns represent the fields and the rows represent the records of each respondent. The field space 1–4 is used to record an assigned number to each respondent. Table 17.3 illustrates how the same data may be entered using a spreadsheet. Each row represents an individual respondent and each column represents the fields required to hold the response to an individual question. Note that there is a column that identifies a specific number attached to each record. Many survey analysis packages record a unique ID for each record so that, as the answers to an individual questionnaire are entered, the ID is automatically updated. However, if a unique ID is attached to each questionnaire before it is sent out (e.g. in a postal survey), the ID may be entered as a distinct field (see column A).

Table 17.3 Example of computer file held on a spreadsheet program

Records in rows	*Individual fields in columns*					
	A	*B*	*C*	*D*	*E*	*F*
	ID	*Question 1*	*Question 2*	*Question 3*	*Question 4*	*Question 5*
2	1	1	1	2	2	01
12	11	1	1	3	1	03
22	21	1	2	7	6	04
2051	2050	1	1	5	4	11

Coding is still required to identify the individual responses to individual questions. Spreadsheets are normally wide enough to allow an individual record to be recorded on one line, and they can be set up so that whoever is entering the data can clearly keep track of which questions relate to which columns. Spreadsheets can be used as a format to analyse data in a wide variety of data analysis packages and so are very versatile. They do, however, have shortcomings. The next paragraph will go on to illustrate these.

In many surveys, multiple-choice questions are widely used. An example of a multiple-choice question is shown in Table 17.4, a question from the Formula One Racetrack survey which examines respondents' perceptions of Formula One.

Table 17.4 Multiple-choice question from the Formula One Racetrack survey

Question 8 – Please tell me which of the following words you think best describe Formula One. (READ OUT AND TICK ALL THAT APPLY.)

01	□	Boring
02	☑	Competitive
03	□	Dangerous
04	□	Dynamic
05	☑	Elitist
06	☑	Exciting
07	□	Expensive
08	□	Extravagant
09	☑	Innovative
10	□	Inspirational
11	□	Prestigious
12	□	Safe
13	□	Sexy
14	□	Too technical

In essence, each of the options presented in question 8 is an individual 'yes' or 'no' question. In the example shown, the respondent has replied 'yes' to the second, fifth, sixth and ninth variables. Using a spreadsheet, this question would be coded as shown in Table 17.5 where the response in Table 17.4 is represented as 'Record 1'. The 'ticks' above have been coded as a '1' to represent 'yes' and '0' as 'no'.

Table 17.5 Formula One Racetrack, question 8 spreadsheet presentation

Records	*Individual fields in columns*								
	Question 8								
	AJ *Q8(1)*	*AK* *Q8(2)*	*AL* *Q8(3)*	*AM* *Q8(4)*	*AN* *Q8(5)*	*AO* *Q8(6)*	*AP* *Q8(7)*	*AQ* *Q8(8)*	*AU* *Q8(9)*
1	1	0	0	0	1	1	0	0	1
2	1	1	0	0	0	1	1	0	0
3	0	0	1	1	1	0	0	0	0
4	1	1	0	0	1	0	1	1	1
... *n*	0	1	1	0	0	1	1	0	0

Entering data for multiple-choice questions is a simple task on a spreadsheet, provided that there are not so many of these question types in a survey. In the Formula One Racetrack questionnaire there was one question related to which cars 'belonged' (owned, leased or a company car) to particular households. Had the open-ended question been closed into a multiple-choice question of 38 different car manufacturers, this would have meant a spreadsheet of 38 columns. Had the question gone beyond the manufacturer, e.g. Renault, to the type of Renault, e.g. Clio, the number of columns could have run to thousands. If respondents had indicated that they had a Renault, this would mean finding the precise column to enter a '1' and then 37 '0s'. This is a lengthy and potentially error-prone task. This is where proprietary questionnaire design and survey packages really hold many advantages, i.e. they make the data entry task very simple for multiple-choice questions of any length and check for errors. Again, the use of proprietary packages will be outlined at the end of this chapter.

Codebook

Whether, the researcher uses DOS-based systems or a spreadsheet, a summary of the whole questionnaire, showing the position of the fields and the key to all the codes, should be produced. With proprietary software packages, this summary is created automatically as the questionnaire is designed. Such a summary is called a codebook. Table 17.6 shows an extract from the Formula One Racetrack codebook. The codebook shown is based upon using a spreadsheet to enter the data. Depending upon which type of data entry is used, the codebook style will change, but in essence the type of information recorded is the same.

Codebook
A book containing coding instructions and the necessary information about the questions and potential answers in a survey.

A **codebook** contains instructions and the necessary information about the questions and potential answers in a survey. A codebook guides the 'coders' in their work and helps the researcher identify and locate the questions properly. Even if the questionnaire has been pre-coded, it is helpful to prepare a formal codebook. As illustrated in Table 17.6, a codebook generally contains the following information: (1) column identifier, (2) question name, (3) question number, and (4) coding instructions.

Table 17.6 Extract from the Formula One Racetrack survey codebook

Column identifier	*Question name*	*Question number*	*Coding instructions*
A	Respondent ID		Enter handwritten number from top right-hand corner of the questionnaire

B	Watched Formula One	1	Yes = '1'
			No = '2'

C	Gender	2	Male = '1'
			Female = '2'

D	Age band	3	Enter number as seen alongside ticked box: 18–24 = '1' 25–29 = '2' 30–34 = '3' 35–39 = '4' 40–44 = '5' 45–49 = '6' 50–55 = '7' Refused = '8'
E	Viewing frequency	4	Enter number as seen alongside ticked box: 1–2 = '1' 3–4 = '2' 5–6 = '3' 7–9 = '4' 10–12 = '5' 13–15 = '6' 16 = '7'

Coding open-ended questions

The coding of structured questions, be they single or multiple choice, is relatively simple because the response options are predetermined. The researcher assigns a code for each response to each question and specifies the appropriate field or column in which it will appear; this is termed 'pre-coding'.[3] The coding of unstructured or open-ended questions is more complex; this is termed 'post-coding'. Respondents' verbatim responses are recorded on the questionnaire. One option the researcher has is to go through all the completed questionnaires, list the verbatim responses and then develop and assign codes to these responses. Another option that is allowed on some data entry packages is to enter the verbatim responses directly onto the computer, allowing a print-off of the collective responses and codes to be assigned before all of the questionnaires have been entered. The coding process here is similar to the process of assigning codes in the analysis of qualitative data as described in Chapter 9. The verbatim responses to 1,000 questionnaires may generate 1,000 different answers. The words may be different but the essence of the response may mean that 20 issues have been addressed. The researcher decides what those 20 issues are, names the issues and assigns codes from 1–20, and then goes through all the 1,000 questionnaires to enter the code alongside the verbatim response.

The following guidelines are suggested for coding unstructured questions and questionnaires in general.[4] Category codes should be mutually exclusive and collectively exhaustive. Categories are mutually exclusive if each response fits into one and only one category code. Categories should not overlap. Categories are collectively exhaustive if every response fits into one of the assigned category codes. This can be achieved by adding an additional category code of 'other' or 'none of the above'. An absolute maximum of 10% of responses should fall into the 'other' category; the researcher should strive to assign all responses into meaningful categories.

Category codes should be assigned for critical issues even if no one has mentioned them. It may be important to know that no one has mentioned a particular response. For example, a car manufacturer may be concerned about its new web page design. In a question 'How did you learn about the new Renault Clio?', the Web should be included as a distinct category, even if no respondents gave this as an answer.

Transcribing

Transcribing data involves keying the coded data from the collected questionnaires into computers. If the data have been collected via the Internet, CATI or CAPI, this step is unnecessary because the data are entered directly into the computer as they are collected. Besides the direct keying of data, they can be transferred by using mark sense forms, optical scanning or computerised sensory analysis (see Figure 17.2). Mark sense forms require responses to be recorded in a pre-designated area coded for that response, and the data can then be read by a machine. Optical scanning involves direct machine reading of the codes and simultaneous transcription. A familiar example of optical scanning is the transcription of universal product code (UPC) data, scanned at supermarket checkout counters. Technological advances have resulted in computerised sensory analysis systems, which automate the data collection process. The questions appear on a computerised gridpad, and responses are recorded directly into the computer using a sensing device.

Except for CATI and CAPI, an original record exists which can be compared with what was either automatically read or keyed. Errors can occur in an automatic read or as data are keyed and it is necessary to verify the dataset, or at least a portion of it, for these errors.

A second operator re-punches the data from the coded questionnaires. The transcribed data from the two operators are compared record by record. Any discrepancy between the

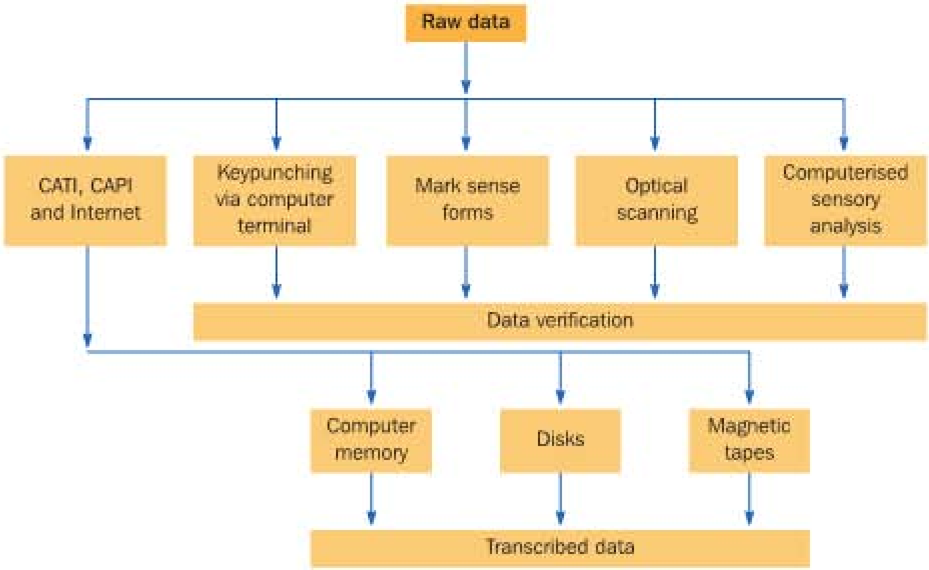


Figure 17.2
Data transcription

two sets of transcribed data is investigated to identify and correct for data keyed in error. Verification of the entire dataset will double the time and cost of data transcription. Given the time and cost constraints, and that experienced operators who key data are quite accurate, it is sufficient to verify 10–25% of the data. With automatically read data, the completed dataset that has been read can be compared with original records. Again, a percentage may be selected and checks made to see what may have caused differences between the original record and the read data (e.g. respondents entering two ticks when only one was requested).

When CATI, CAPI or the Internet are employed, data are verified as they are collected. In the case of inadmissible responses, the computer will prompt the interviewer or respondent. In the case of admissible responses, the interviewer or the respondent can see the recorded response on the screen and verify it before proceeding.

The selection of a data transcription method is guided by the type of interviewing method used and the availability of equipment. If CATI, CAPI or the Internet are used, the data are entered directly into the computer. Keypunching via a computer terminal is most frequently used for ordinary telephone, in-home and street interviewing and traditional mail interviews. The use of computerised sensory analysis systems in personal interviews is increasing with the growing use of handheld computers. Optical scanning can be used in structured and repetitive surveys, and mark sense forms are used in special cases.[5] The following example illustrates the use of scanned data, and, in this circumstance, the advantage it has over keypunching.

Example

Scanning the seas[6]

Princess Cruises (**www.princess.com**) operates large cruise ships around the world. It is the third largest cruise line, carrying around 700,000 passengers annually with 27 ships. Princess wished to know what passengers thought of the cruise experience, but wanted to determine this information in a cost-effective way. A scannable questionnaire was developed that allowed the cruise line to transcribe the data quickly from thousands of responses, thus expediting data preparation and analysis. This questionnaire is distributed to measure customer satisfaction on all voyages. In addition to saving time as compared with keypunching, scanning has also increased the accuracy of survey results. The Senior Marketing Researcher for Princess Cruises, Jamie Goldfarb, commented, 'When we compared the data files from the two methods, we found that although the scanned system occasionally missed marks because they had not been filled in properly, the scanned data was still more accurate than the keypunched file.'

Cleaning the data

Data cleaning
Thorough and extensive checks for consistency and treatment of missing responses.

Data cleaning includes consistency checks and treatment of missing responses. Even though preliminary consistency checks have been made during editing, the checks at this stage are more thorough and extensive, because they are made by computer.

Consistency checks

Consistency checks
A part of the data cleaning process that identifies data that are out of range, logically inconsistent or have extreme values. Data with values not defined by the coding scheme are inadmissible.

Consistency checks identify data that are out of range or logically inconsistent or have extreme values. Out-of-range data values are inadmissible and must be corrected. For example, respondents were asked to express their degree of agreement with a series of lifestyle statements on a 1–5 scale. Assuming that 9 has been designated for missing values, data values of 0, 6, 7 and 8 would be out of range. Computer packages can be programmed to identify out-of-range values for each variable and will not progress to another variable within a record until a value in the set range is entered. Other packages can be programmed to print out the respondent code, variable code, variable name, record number, column number and out-of-range value.[7] This makes it easy to check each variable systematically for out-of-range values. The correct responses can be determined by going back to the edited and coded questionnaire.

Responses can be logically inconsistent in various ways. For example, respondents may indicate that they charge long-distance calls to a calling card from a credit card company, although they do not have such a credit card. Or respondents report both unfamiliarity with and frequent usage of the same product. The necessary information (respondent code, variable code, variable name, record number, column number and inconsistent values) can be printed to locate these responses and to take corrective action.

Finally, extreme values should be closely examined. Not all extreme values result from errors, but they may point to problems with the data. For example, in the Formula One Racetrack survey, respondents were asked to name the manufacturer of the cars that 'belonged' to their households. Certain respondents recorded 10 or more cars. In these circumstances the extreme values can be identified and the actual figure validated in many cases by recontacting the respondent.

Treatment of missing responses

Missing responses
Values of a variable that are unknown because the respondents concerned provided ambiguous answers to the question or because their answers were not properly recorded.

Missing responses represent values of a variable that are unknown either because respondents provided ambiguous answers or because their answers were not properly recorded. Treatment of missing responses poses problems, particularly if the proportion of missing responses is more than 10%. The following options are available for the treatment of missing responses.[8]

Substitute a neutral value. A neutral value, typically the mean response to the variable, is substituted for the missing responses. Thus, the mean of the variable remains unchanged, and other statistics such as correlations are not affected much. Although this approach has some merit, the logic of substituting a mean value (say 4) for respondents who, if they had answered, might have used either high ratings (6 or 7) or low ratings (1 or 2) is questionable.[9]

Substitute an imputed response. The respondents' pattern of responses to other questions is used to impute or calculate a suitable response to the missing questions. The researcher attempts to infer from the available data the responses the individuals would have given if they had answered the questions. This can be done statistically by determining the relationship of the variable in question to other variables based on the available data. For example, product usage could be related to household size for respondents who have provided data on both variables. Given that respondent's household size, the missing product usage response for a respondent could then be calculated. This approach, however, requires considerable effort and can introduce serious bias. Sophisticated statistical procedures have been developed to calculate imputed values for missing responses.[10]

Casewise deletion
A method for handling missing responses in which cases or respondents with any missing responses are discarded from the analysis.

Casewise deletion. In **casewise deletion**, cases or respondents with any missing responses are discarded from the analysis. Because many respondents may have some missing responses, this approach could result in a small sample. Throwing away large amounts of data is undesirable because it is costly and time consuming to collect data. Furthermore, respondents with missing responses could differ from respondents with complete responses in systematic ways. If so, casewise deletion could seriously bias the results.

Pairwise deletion
A method for handling missing responses in which all cases or respondents with any missing responses are not automatically discarded; rather, for each calculation, only the cases or respondents with complete responses are considered.

Pairwise deletion. In **pairwise deletion**, instead of discarding all cases with any missing responses, the researcher uses only the cases or respondents with complete responses for each calculation. As a result, different calculations in an analysis may be based on different sample sizes. This procedure may be appropriate when (1) the sample size is large, (2) there are few missing responses, and (3) the variables are not highly related. However, this procedure can produce unappealing or even infeasible results.

The different procedures for the treatment of missing responses may yield different results, particularly when the responses are not missing at random and the variables are related. Hence, missing responses should be kept to a minimum. The researcher should carefully consider the implications of the various procedures before selecting a particular method for the treatment of non-response.

Statistically adjusting the data

Procedures for statistically adjusting the data consist of weighting, variable respecification and scale transformation. These adjustments are not always necessary but can enhance the quality of data analysis.

Weighting

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

In **weighting**, each case or respondent in the database is assigned a weight to reflect its importance relative to other cases or respondents. The value 1.0 represents the unweighted case. The effect of weighting is to increase or decrease the number of cases in the sample that possess certain characteristics. (See Chapter 15, which discussed the use of weighting to adjust for non-response bias.)

Weighting is most widely used to make the sample data more representative of a target population on specific characteristics. For example, it may be used to give greater importance to cases or respondents with higher quality data. Yet another use of weighting is to adjust the sample so that greater importance is attached to respondents with certain characteristics. If a study is conducted to determine what modifications should be made to an existing product, the researcher might want to attach greater weight to the opinions of heavy users of the product. This could be accomplished by assigning weights of 3.0 to heavy users, 2.0 to medium users and 1.0 to light users and non-users. Because it destroys the self-weighting nature of the sample design, weighting should be applied with caution. If used, the weighting procedure should be documented and made a part of the project report.[11]

Example

Determining the weight of community centre users

A mail survey was conducted in the Scottish city of Edinburgh to determine the patronage of a community centre. The resulting sample composition differed in age structure from the area population distribution as compiled from recent census data. Therefore, the sample was weighted to make it representative in terms of age structure. The weights applied were determined by dividing the population percentage by the corresponding sample percentage. The distribution of age structure for the sample and population, as well as the weights applied, are given in the following table.

Source: © Alamy

Age group percentage	*Sample percentage*	*Population*	*Weight*
13–18	4.32	6.13	1.42
19–24	5.89	7.45	1.26
25–34	12.23	13.98	1.14
35–44	17.54	17.68	1.01
45–54	14.66	15.59	1.06
55–64	13.88	13.65	0.98
65–74	15.67	13.65	0.87
75 plus	15.81	11.87	0.75
Totals	100.00	100.00	

Age groups under-represented in the sample received higher weights, whereas over-represented age groups received lower weights. Thus, the data for a respondent aged 13–18 would be overweighted by multiplying by 1.42, whereas the data for a respondent aged 75 plus would be underweighted by multiplying by 0.75.

Variable respecification

Variable respecification
The transformation of data to create new variables or the modification of existing variables so that they are more consistent with the objectives of the study.

Variable respecification involves the transformation of data to create new variables or to modify existing variables. The purpose of respecification is to create variables that are consistent with the objectives of the study. For example, suppose that the original variable was product usage, with 10 response categories. These might be collapsed into four categories: heavy, medium, light and non-user. Or the researcher may create new variables that are composites of several other variables. For example, the researcher may create an Index of Information Search (ISS), which is the sum of information new car customers seek from dealers, promotional sources, the Internet and other independent sources. Likewise, one may take the ratio of variables. If the amount of purchases at a clothes shop (X_1) and the amount of purchases where a credit card was used (X_2) have been measured, the proportion of purchases charged to a credit card can be a new variable, created by taking the ratio of the two (X_2/X_1). Other respecifications of variables include square root and log transformations, which are often applied to improve the fit of the model being estimated.

Dummy variables
A respecification procedure using variables that take on only two values, usually 0 or 1.

An important respecification procedure involves the use of **dummy variables** for respecifying categorical variables. Dummy variables are also called binary, dichotomous, instrumental or qualitative variables. They are variables that may take on only two values, such as 0 or 1. The general rule is that to respecify a categorical variable with K categories, $K - 1$ dummy variables are needed. The reason for having $K - 1$, rather than K, dummy variables is that only $K - 1$ categories are independent. Given the sample data, information about the Kth category can be derived from information about the other $K - 1$ categories. Consider gender, a variable having two categories. Only one dummy variable is needed. Information on the number or percentage of males in the sample can be readily derived from the number or percentage of females. The following example further illustrates the concept of dummy variables.

Example

'Frozen' consumers treated as dummies

In a survey of consumer preferences for frozen foods, the respondents were classified as heavy users, medium users, light users and non-users, and they were originally assigned codes of 4, 3, 2 and 1, respectively. This coding was not meaningful for several statistical analyses. To conduct these analyses, product usage was represented by three dummy variables, X_1, X_2 and X_3, as shown.

Product usage category	*Original variable code*	*Dummy variable code*		
		X_1	X_2	X_3
Non-users	1	1	0	0
Light users	2	0	1	0
Medium users	3	0	0	1
Heavy users	4	0	0	0

Note that X_1 = 1 for non-users and 0 for all others. Likewise, X_2 = 1 for light users and 0 for all others, and X_3 = 1 for medium users and 0 for all others. In analysing the data, X_1, X_2, and X_3 are used to represent all user/non-user groups.

Scale transformation

Scale transformation
A manipulation of scale values to ensure compatibility with other scales or otherwise to make the data suitable for analysis.

Scale transformation involves a manipulation of scale values to ensure comparability with other scales or otherwise to make the data suitable for analysis. Frequently, different scales are employed for measuring different variables. For example, image variables may be measured on a seven-point semantic differential scale, attitude variables on a continuous rating scale, and lifestyle variables on a five-point Likert scale. Therefore, it would not be meaningful to make comparisons across the measurement scales for any respondent. To compare attitudinal scores with lifestyle or image scores, it would be necessary to transform the various scales. Even if the same scale is employed for all the variables, different respondents may use the scale differently. For example, some respondents consistently use the upper end of a rating scale whereas others consistently use the lower end. These differences can be corrected by appropriately transforming the data.

Example

Health care services: transforming consumers[12]

In a study examining preference segmentation of health care services, respondents were asked to rate the importance of 18 factors affecting preferences for hospitals on a three-point scale (very, somewhat, or not important). Before analysing the data, each individual's ratings were transformed. For each individual, preference responses were averaged across all 18 items. Then this mean $\bar{X}$ was subtracted from each item rating X_i, and a constant C was added to the difference. Thus, the transformed data, X_t, were obtained by

$$X_t = X_i - \bar{X} + C$$

Subtraction of the mean value corrected for any uneven use of the importance scale. The constant C was added to make all the transformed values positive, since negative importance ratings are not meaningful conceptually. This transformation was desirable because some respondents, especially those with low incomes, had rated almost all the preference items as very important. Others, high-income respondents in particular, had assigned the very important rating to only a few preference items. Thus, subtraction of the mean value provided a more accurate idea of the relative importance of the factors.

Standardisation
The process of correcting data to reduce them to the same scale by subtracting the sample mean and dividing by the standard deviation.

In this example, the scale transformation is corrected only for the mean response. A more common transformation procedure is **standardisation**. To standardise a scale X_i, we first subtract the mean, $\bar{X}$, from each score and then divide by the standard deviation, s_x. Thus, the standardised scale will have a mean of 0 and a standard deviation of 1. This is essentially the same as the calculation of *z* scores (see Chapter 15). Standardisation allows the researcher to compare variables that have been measured using different types of scales.[13] Mathematically, standardised scores, z_i, may be obtained as

$$z_i = \frac{(X_i - \bar{X})}{s_x}$$

Selecting a data analysis strategy

The process of selecting a data analysis strategy is described in Figure 17.3. The selection of a data analysis strategy should be based on the earlier steps of the marketing research process, known characteristics of the data, properties of statistical techniques and the background and philosophy of the researcher.

Data analysis is not an end in itself. Its purpose is to produce information that will help address the problem at hand. The selection of a data analysis strategy must begin with a consideration of the earlier steps in the process: problem definition (step 1), development of an approach (step 2) and research design (step 3). The preliminary plan of data analysis prepared as part of the research design should be used as a springboard. Changes may be necessary in the light of additional information generated in subsequent stages of the research process.

The next step is to consider the known characteristics of the data. The measurement scales used exert a strong influence on the choice of statistical techniques (see Chapter 12). In addition, the research design may favour certain techniques. For example, analysis of variance (see Chapter 19) is suited for analysing experimental data from causal designs. The insights into the data obtained during data preparation can be valuable for selecting a strategy for analysis.

It is also important to take into account the properties of the statistical techniques, particularly their purpose and underlying assumptions. Some statistical techniques are appropriate for examining differences in variables, others for assessing the magnitudes of the relationships between variables, and still others for making predictions. The techniques also involve different assumptions, and some techniques can withstand violations of the underlying assumptions better than others. A classification of statistical techniques is presented below.

Finally, the researcher's background and philosophy affect the choice of a data analysis strategy. The experienced, statistically trained researcher will employ a range of techniques, including advanced statistical methods. Researchers differ in their willingness to make

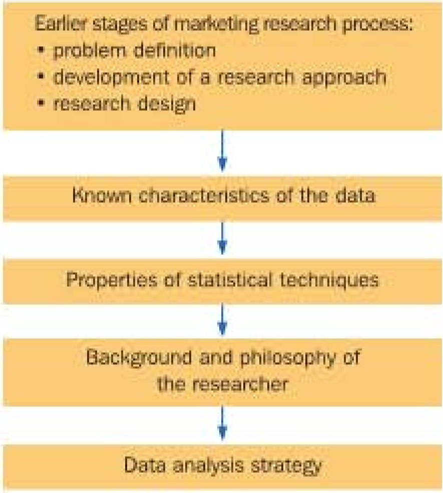

Earlier stages of marketing research process:
• problem definition
• development of a research approach
• research design
Known characteristics of the data
Properties of statistical techniques
Background and philosophy of the researcher
Data analysis strategy

Figure 17.3 Selecting a data analysis strategy

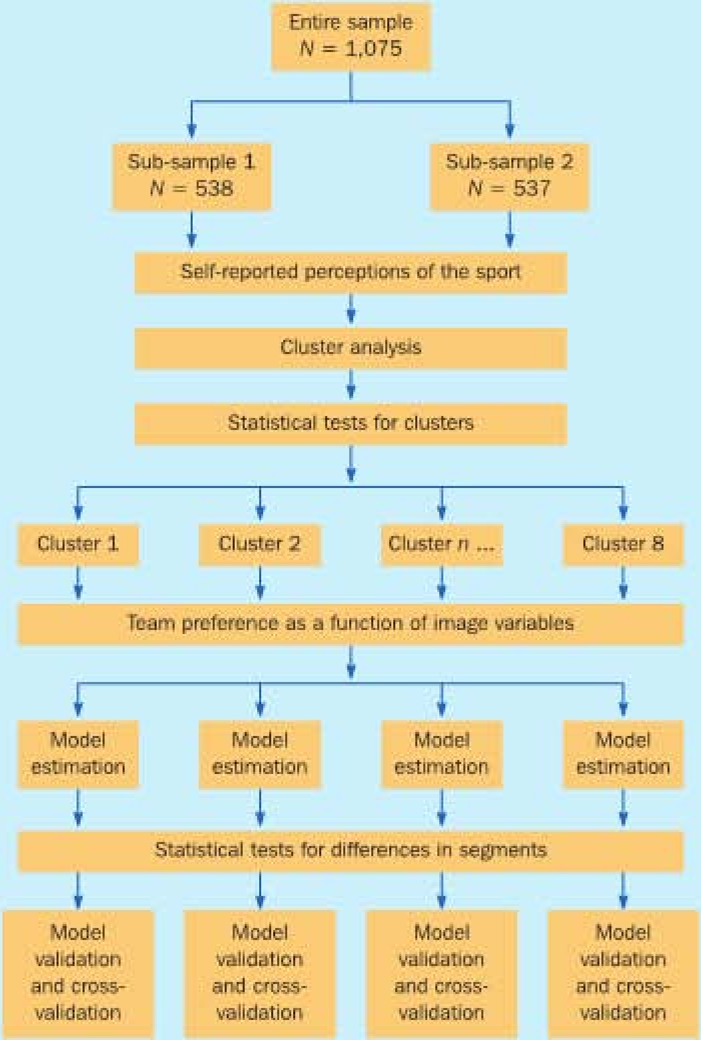
assumptions about the variables and their underlying populations. Researchers who are conservative about making assumptions will limit their choice of techniques to distribution-free methods. In general, several techniques may be appropriate for analysing the data from a given project. We use the Formula One Racetrack Project to illustrate how a data analysis strategy can be developed.

Focus on

Formula One Racetrack Project

Data analysis strategy[14]

As part of the analysis conducted in the Formula One Racetrack Project, Formula One Image was modelled in terms of the most important issues that shaped perceptions of the sport. The sample was split into halves. The respondents in each half were clustered on the basis of their perceptions. Statistical tests for clusters were conducted, and eight segments were identified. Formula One image was modelled in terms of the evaluations of the race teams. The model was estimated separately for each segment. Differences between segment preference functions were statistically tested. Finally, model verification and cross-validation were conducted for each segment. The data analysis strategy adopted is depicted in this figure:


Entire sample
N = 1,075
Sub-sample 1
N = 538
Sub-sample 2
N = 537
Self-reported perceptions of the sport
Cluster analysis
Statistical tests for clusters
Cluster 1
Cluster 2
Cluster n ...
Cluster 8
Team preference as a function of image variables
Model estimation
Model estimation
Model estimation
Model estimation
Statistical tests for differences in segments
Model validation and cross-validation
Model validation and cross-validation
Model validation and cross-validation
Model validation and cross-validation

A classification of statistical techniques

Univariate techniques
Statistical techniques appropriate for analysing data when there is a single measurement of each element in the sample, or, if there are several measurements on each element, when each variable is analysed in isolation.

Multivariate techniques
Statistical techniques suitable for analysing data when there are two or more measurements on each element and the variables are analysed simultaneously. Multivariate techniques are concerned with the simultaneous relationships among two or more phenomena.

Metric data
Data that are interval or ratio in nature.

Non-metric data
Data derived from a nominal or ordinal scale.

Independent samples
The samples are independent if they are drawn randomly from different populations.

Paired samples
The samples are paired when the data for the two samples relate to the same group of respondents.

Statistical techniques can be classified as univariate or multivariate. **Univariate techniques** are appropriate when there is a single measurement of each element in the sample or when there are several measurements of each element but each variable is analysed in isolation. **Multivariate techniques**, on the other hand, are suitable for analysing data when there are two or more measurements of each element and the variables are analysed simultaneously. Multivariate techniques are concerned with the simultaneous relationships among two or more phenomena. Multivariate techniques differ from univariate techniques in that they shift the focus away from the levels (averages) and distributions (variances) of the phenomena, concentrating instead on the degree of relationships (correlations or covariances) among these phenomena.[15] The univariate and multivariate techniques are described in detail in Chapters 18–24, but here we show how the various techniques relate to each other in an overall scheme of classification.

Univariate techniques can be further classified based on whether the data are metric or non-metric (as introduced in Chapter 12). **Metric data** are measured on an interval or ratio scale, whereas **non-metric data** are measured on a nominal or ordinal scale. These techniques can be further classified based on whether one, two or more samples are involved. It should be noted that the number of samples is determined based on how the data are treated for the purpose of analysis, not based on how the data were collected. For example, the data for males and females may well have been collected as a single sample, but if the analysis involves an examination of gender differences, two samples will be used. The samples are **independent** if they are drawn randomly from different populations. For the purpose of analysis, data pertaining to different groups of respondents, e.g. males and females, are generally treated as independent samples. On the other hand, the samples are **paired** when the data for the two samples relate to the same group of respondents.

For metric data, when there is only one sample, the *z* test and the *t* test can be used. When there are two or more independent samples, the *z* test and *t* test can be used for two samples, and one-way analysis of variance (one-way ANOVA) can be used for more than two samples. In the case of two or more related samples, the paired *t* test can be used. For non-metric data involving a single sample, frequency distribution, chi-square, Kolmogorov–Smirnov (K–S), runs and binomial tests can be used. For two independent samples with non-metric data, the chi-square, Mann–Whitney, median, K–S and Kruskal–Wallis one-way analysis of variance (K–W ANOVA) can be used. In contrast, when there are two or more related samples, the sign, Wilcoxon, McNemar and chi-square tests should be used (see Figure 17.4).

Dependence techniques
Multivariate techniques appropriate when one or more of the variables can be identified as dependent variables and the remaining ones as independent variables.

Interdependence techniques
Multivariate statistical techniques that attempt to group data based on underlying similarity and thus allow for interpretation of the data structures. No distinction is made as to which variables are dependent and which are independent.

Multivariate statistical techniques can be classified as dependence techniques or interdependence techniques (see Figure 17.5). **Dependence techniques** are appropriate when one or more variables can be identified as dependent variables and the remaining ones as independent variables. When there is only one dependent variable, cross-tabulation, analysis of variance and covariance, multiple regression, two-group discriminant analysis and conjoint analysis can be used. If there is more than one dependent variable, however, the appropriate techniques are multivariate analysis of variance and covariance, canonical correlation and multiple discriminant analysis. In **interdependence techniques,** the variables are not classified as dependent or independent; rather, the whole set of interdependent relationships is examined. These techniques focus on either variable interdependence or interobject similarity. The major technique for examining variable interdependence is factor analysis. Analysis of inter-object similarity can be conducted by cluster analysis and multidimensional scaling.[16]

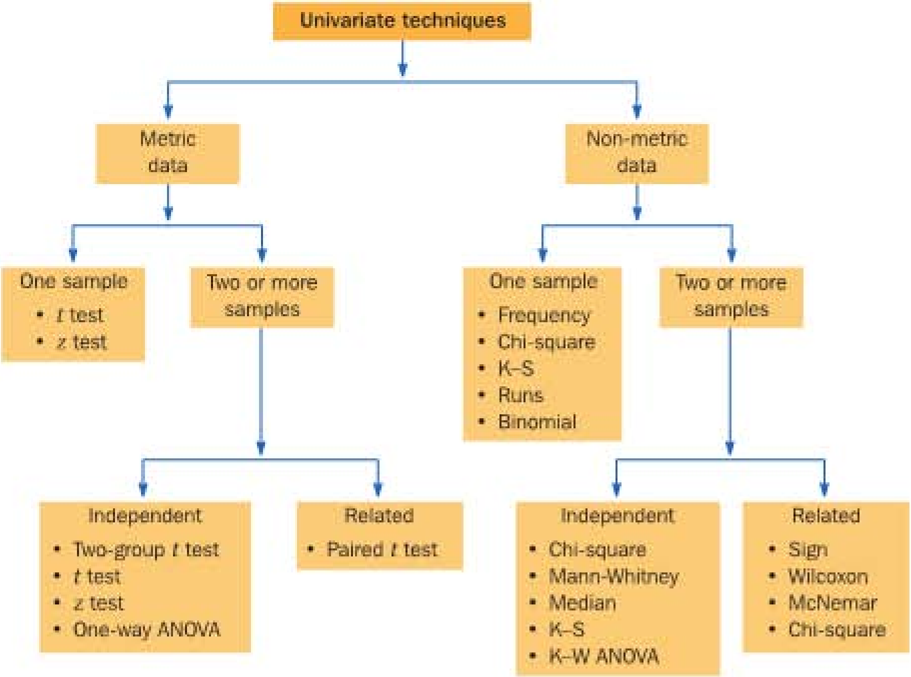


Figure 17.4
A classification of univariate techniques

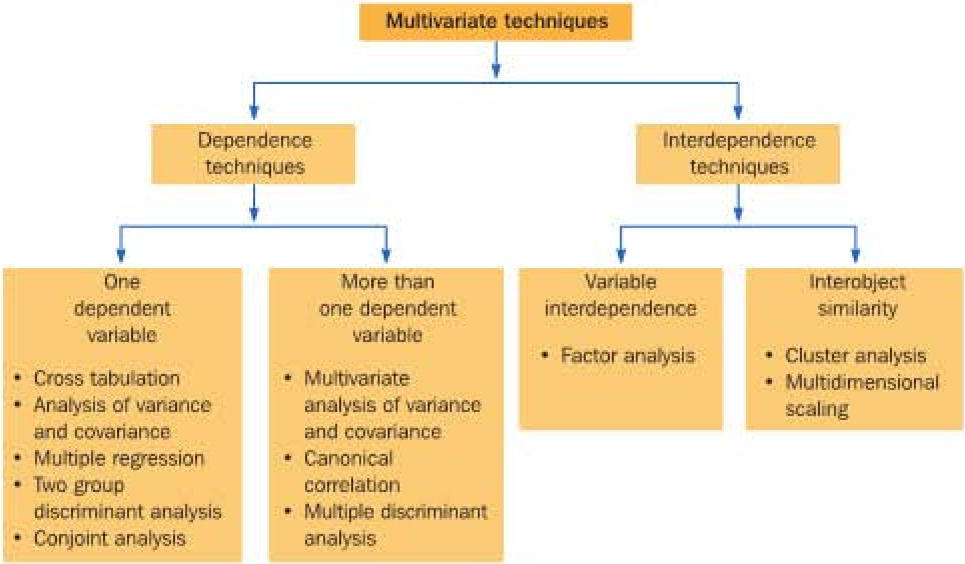

Multivariate techniques
Dependence techniques
Interdependence techniques
One dependent variable
• Cross tabulation
• Analysis of variance and covariance
• Multiple regression
• Two group discriminant analysis
• Conjoint analysis
More than one dependent variable
• Multivariate analysis of variance and covariance
• Canonical correlation
• Multiple discriminant analysis
Variable interdependence
• Factor analysis
Interobject similarity
• Cluster analysis
• Multidimensional scaling

Figure 17.5
A classification of multivariate techniques

International marketing research

Before analysing the data, the researcher should ensure that the units of measurement are comparable across countries or cultural units. For example, the data may have to be adjusted to establish currency equivalents or metric equivalents. Furthermore, standardisation or normalisation of the data may be necessary to make meaningful comparisons and achieve consistent results.

Example

A worldwide scream for ice cream[17]

Over half the sales of Häagen-Dazs (**www.haagen-dazs.com**), the US ice cream manufacturer, come from markets outside the USA. Its sales in Asia, the UK, France and Germany are increasing at a phenomenal rate. In December 2001, General Mills, Inc. sold its 50% stake in Häagen-Dazs to the Swiss company Nestlé for $641 million (£500 million). How has Häagen-Dazs achieved this situation? Marketing research conducted in several European countries (e.g. the UK, France and Germany) and several Asian countries (e.g. Japan, Singapore and Taiwan) revealed that consumers were hungry for a high-quality ice cream with a high-quality image and were willing to pay a premium price for it. These consistent findings emerged after the price of ice cream in each country was standardised to have a mean of zero and a standard deviation of unity. Standardisation was desirable because the prices were specified in different local currencies and a common basis was needed for comparison across countries. Also, in each country, the premium price had to be defined in relation to the prices of competing brands. Standardisation accomplished both of these objectives.

Based on these findings, Häagen-Dazs first introduced the brand at a few high-end retailers; it then built company-owned stores in high-traffic areas; and finally it rolled into convenience stores and supermarkets. It maintained the premium quality brand name by starting first with a few high-end retailers. It also supplied free freezers to retailers. Hungry for quality products, consumers in the new markets paid double or triple the price of home brands. In the USA, Häagen-Dazs remains popular, although faced with intense competition and a health-conscious market. This added to the impetus to enter international markets.

Data analysis could be conducted at three levels: (1) individual, (2) within country or cultural unit, and (3) across countries or cultural units. Individual-level analysis requires that the data from each respondent be analysed separately. For example, one might compute a correlation coefficient or run a regression analysis for each respondent. This means that enough data must be obtained from each individual to allow analysis at the individual level, which is often not feasible. Yet it has been argued that, in international marketing or cross-cultural research, the researcher should possess a sound knowledge of the consumer in each culture. This can best be accomplished by individual-level analysis.[18]

Intra-cultural analysis Within-country analysis of international data.

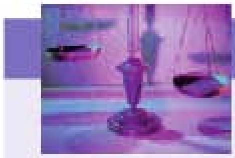
In within-country or cultural unit analysis, the data are analysed separately for each country or cultural unit. This is also referred to as **intra-cultural analysis**. This level of analysis is quite similar to that conducted in domestic marketing research. The objective is to gain an understanding of the relationships and patterns existing in each country or cultural unit.

Pan-cultural analysis Across-countries analysis in which the data for all respondents from all the countries are pooled and analysed.

Cross-cultural analysis A type of across-countries analysis in which the data could be aggregated for each country and these aggregate statistics analysed.

In across-countries analysis, the data from all the countries are analysed simultaneously. Two approaches to this method are possible. The data for all respondents from all the countries can be pooled and analysed. This is referred to as **pan-cultural analysis**. Alternatively, the data can be aggregated for each country, and then these aggregate statistics can be analysed. For example, one could compute means of variables for each country, and then compute correlations on these means. This is referred to as **cross-cultural analysis**. The objective of this level of analysis is to assess the comparability of findings from one country to another. The similarities as well as the differences between countries should be investigated. When examining differences, not only differences in means but also differences in variance and distribution should be assessed.

All the statistical techniques that have been discussed in this book can be applied to within-country or across-country analysis and, subject to the amount of data available, to individual-level analysis as well.[19]



Ethics in marketing research

Ethical issues that arise during the data preparation and analysis step of the marketing research process pertain mainly to the researcher. While checking, editing, coding, transcribing and cleaning, researchers should try to get some idea about the quality of the data. An attempt should be made to identify respondents who have provided data of questionable quality. Consider, for example, a respondent who ticks the '7' response to all the 20 items measuring attitude to spectator sports on a 1–7 Likert scale. Apparently, this respondent did not realise that some of the statements were negative whereas others were positive. Thus, this respondent indicates an extremely favourable attitude towards spectator sports on all the positive statements and an extremely negative attitude on the statements that were reversed. Decisions on whether such respondents should be discarded, i.e. not included in the analysis, can raise ethical concerns. A good rule of thumb is to make such decisions during the data preparation phase before conducting any analysis.

In contrast, suppose that the researcher conducted the analysis without first attempting to identify unsatisfactory responses. The analysis, however, does not reveal the expected relationship; the analysis does not show that attitude towards spectator sports influences attendance at spectator sports. The researcher then decides to examine the quality of data obtained. In checking the questionnaires, a few respondents with unsatisfactory data are identified. In addition to the type of unsatisfactory responses mentioned earlier, there were responses as '4', the 'neither agree nor disagree' response, to all the 20 items measuring attitude towards spectator sports. When these respondents are eliminated and the reduced dataset is analysed, the expected results are obtained showing a positive influence of attitude on attendance of spectator sports. Discarding respondents after analysing the data raises ethical concerns, particularly if the report does not state that the initial analysis was inconclusive. Moreover, the procedure used to identify unsatisfactory respondents and the number of respondents discarded should be clearly disclosed, as in the following example.

Example

Elimination of decision-makers unwilling to be ethical[20]

In a study of MBA graduates' responses to marketing ethics dilemmas, respondents were required to respond to 14 questions regarding ethically ambiguous scenarios by writing a simple sentence on what action they would take if they were the manager. The responses were then analysed to determine whether the respondent's answer was indicative of ethical behaviour. However, in the data preparation phase, six respondents out of the 561 total respondents were eliminated from further analysis because their responses indicated that they did not follow the directions which told them to state clearly their choice of action. This is an example of ethical editing of the data. The criterion for unsatisfactory responses is clearly stated, the unsatisfactory respondents are identified before the analysis, and the number of respondents eliminated is disclosed.

While analysing the data, the researcher may also have to deal with ethical issues. The assumptions underlying the statistical techniques used to analyse the data must be satisfied to obtain meaningful results. Any departure from these assumptions should be critically examined to determine the appropriateness of the technique for analysing the data at hand. The researcher has the responsibility of justifying the statistical techniques used for analysis. When this is not done, ethical questions can be raised. Moreover, there should be no intentional or deliberate misrepresentation of research methods or results.

Similarly, ethical issues can arise in interpreting the results, drawing conclusions, making recommendations, and in implementation. For example, the error terms in bivariate regression must be normally distributed about zero, with a constant variance, and be uncorrelated (Chapter 20). The researcher has the responsibility to test these assumptions and take appropriate corrective actions if necessary. Although interpretations, conclusions, recommendations and implementations necessarily involve subjective judgement, this judgement must be exercised honestly, free from personal biases or agendas of the researcher or the client.

Internet and computer applications

In evaluating software that can help the marketing researcher with data preparation tasks, a distinction should be made between survey design and analysis packages such as SNAP (**www.snapsurveys.com**), Keypoint2 (**www.camsp.com**) and Sphinxsurvey (**www.scolari.co.uk**), and statistical packages such as SPSS (**www.spss.com**), SAS (**www.sas.com**) and Minitab (**www.minitab.com**), and office products such as Microsoft Excel (**www.microsoft.com**). It is advised that these websites are visited to view the demonstration versions of the packages to gain a feel for their capabilities and applications.

Survey design and analysis packages such as SNAP, Keypoint2 and Sphinx perform a much broader range of tasks to help with the array of data preparation tasks than either the statistical packages or the more generic office products. Primarily, they allow the physical format of a questionnaire to be designed. The designed questionnaire can be printed out and used for traditional mail surveys, formatted for interviewer-led surveys using CATI or CAPI, or formatted for self-completion Internet surveys. As the questionnaire is designed on screen, the underlying structure of the questionnaire is automatically created. This results in the coding of question replies and the associated field positions automatically being worked out. It also means that any alterations in the design of the questionnaire automatically maintain the structure and integrity of the data being collected, ensuring accuracy in the later analysis stages.

Data can be entered directly by respondents as an Internet survey progresses, or entered by an interviewer on a CATI or CAPI survey. Alternatively, the data can be keyed directly from a paper questionnaire onto the computer. The packages have built-in checks for 'out-of-range' responses which halt the progress of data entry; for example, if a five-point Likert scale is used and a '7' is entered, there is an audible warning and the data entry halts until the error is corrected.

Multiple-choice questions typically appear as tick boxes and can be set as either a single response or a multiple response, and it is a simple task to highlight one or more of the boxes. Alternatively, numbers, dates or verbatim responses can be entered. The verbatim data can be transferred to qualitative data analysis packages if needed, or can be post-coded (coded later) and a new variable established for subsequent analysis.

Data preparation effort is considerably reduced when replies are entered directly onto the computer as each question is asked. For paper-based surveys this is not practical and optical scanning can be a cost-effective solution, particularly when there are a large proportion of multiple-response questions. The pages of the questionnaire are scanned and the systems search for marks within boxes, and even allow for situations where a reply has been altered or removed. The following example illustrates conditions where scanning is beneficial and an example of SPSS software to support the process.

→

Example

How Walcheren Hospital perked up its patients[21]

In 1996, the Netherlands passed the Quality Act for health care institutions. This placed a requirement on all health care organisations in the country to evaluate the quality of their services based on patient feedback. The hospital of Walcheren admits 120,000 patients each year, which creates a huge amount of data to analyse. Walcheren chose the SPSS suite of market research tools called Dimensions. Using the software helped the hospital to design a survey that could be scanned to capture the data, but eliminated the need for it to go through the time-consuming process of actually setting up the questionnaire for scanning. The survey has significantly contributed to patient well-being and satisfaction. Many patients noted that they did not receive enough details about their illnesses and treatment. These complaints led to hospital staff making changes to their communication processes.

Scanning becomes less cost effective when the proportion of open questions increases, as there is often a high level of manual intervention to clean and code the responses. Scanning systems currently recognise and interpret numbers and hand-printed text, but are not yet capable of accurately recognising handwritten text. Some systems can be 'trained' to recognise an individual style of handwriting, but such a task is inappropriate for normal marketing research surveys, particularly if they are self-completion. For certain types of surveys, scanning is a great time saver. For others, the scanning, coding and cleaning are no faster than manually entering the data onto the computer.

Verification is the process of manually re-entering a proportion of the questionnaires to check the accuracy of data entry, and this facility is regularly used for paper-based surveys. Options are available to set a percentage level (perhaps 5–10%) and the programs then randomly select questionnaires to be re-entered. The programs also allow for validation checks between questions to be set up, completing a comprehensive error avoidance and checking ability, which in total means that the full array of data cleaning tasks can be performed.

Once the dataset has been created, checked and cleaned, SNAP and Keypoint2 allow statistical adjustment of the data through the creation of weights, scale transformation and the ability to create new variables or modify existing ones. Univariate and basic multivariate statistical analyses can be performed and presented in an array of styles of tables and graphs. Before questionnaires are sent out, the forms of analysis and the resulting tables and graphs as dummy tables can be designed and the instructions for these forms stored. This means that, as questionnaire responses start to build up in any survey format, a full set of tables and graphs as an interim analysis can be performed at any time by the researcher.

In essence, these packages perform the complete array of tasks faced by the marketing researcher, from designing a questionnaire and ensuring a sound dataset is built up to analysing the data and presenting the findings. There are limitations to the extent and array of multivariate statistical analyses that can be performed by these packages. However, they do allow for links to statistical analysis packages. For example, SNAP has the option to transform the files that describe the question and answer structure, and the raw data of a survey, into a fully labelled and coded SPSS SAV file, ready to perform any of the multivariate analysis tasks as detailed in Chapters 18–24. Alternatively, SNAP can both export and import triple s files (www.triple-s.org), a standard used by over 50 survey-related software companies worldwide to represent both the definitions of the survey and the associated data.

Summary

Data preparation begins with a preliminary check of all questionnaires for completeness and interviewing quality. Then more thorough editing takes place. Editing consists of screening questionnaires to identify illegible, incomplete, inconsistent or ambiguous responses. Such responses may be handled by returning questionnaires to the field, assigning missing values or discarding unsatisfactory respondents.

The next step is coding. A numeric or alphanumeric code is assigned to represent a specific response to a specific question along with the column position or field that code will occupy. It is often helpful to prepare a codebook containing the coding instructions and the necessary information about the variables in the dataset. The coded data are transcribed onto discs or magnetic tapes or entered directly into a data analysis package. Mark sense forms, optical scanning or computerised sensory analysis may also be used. Good survey desgin software packages completely automate the coding process.

Cleaning the data requires consistency checks and treatment of missing responses. Options available for treating missing responses include substitution of a neutral value such as a mean, substitution of an imputed response, casewise deletion and pairwise deletion. Statistical adjustments such as weighting, variable respecification and scale transformations often enhance the quality of data analysis. The selection of a data analysis strategy should be based on the earlier steps of the marketing research process, known characteristics of the data, properties of statistical techniques, and the background and philosophy of the researcher. Statistical techniques may be classified as univariate or multivariate.

Before analysing the data in international marketing research, the researcher should ensure that the units of measurement are comparable across countries or cultural units. The data analysis could be conducted at three levels: (1) individual, (2) within-country or cultural unit (intra-cultural analysis), and (3) across countries or cultural units (pan-cultural or cross-cultural analysis). Several ethical issues are related to data processing, particularly the discarding of unsatisfactory responses, violation of the assumptions underlying the data analysis techniques, and evaluation and interpretation of results.

Questions

1 Describe the data preparation process. Why is this process needed?

2 What activities are involved in the preliminary checking of questionnaires that have been returned from the field?

3 What is meant by editing a questionnaire?

4 How are unsatisfactory responses that are discovered in editing treated?

5 What is the difference between pre-coding and post-coding?

6 Describe the guidelines for the coding of unstructured questions.

7 What does transcribing the data involve?

8 What kinds of consistency checks are made in cleaning the data?

9 What options are available for the treatment of missing data?

10 What kinds of statistical adjustments are sometimes made to the data?

11 Describe the weighting process. What are the reasons for weighting?

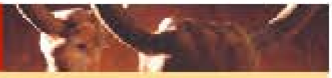
12 What are dummy variables? Why are such variables created?

13 Explain why scale transformations are made.

14 Which scale transformation procedure is most commonly used? Briefly describe this procedure.

15 What considerations are involved in selecting a data analysis strategy?

Exercises



1 Using the SNAP software package that accompanies this text (you are able to design nine questions of your survey and administer these to up to 25 respondents), design a short questionnaire on any topic and write out five potential sets of responses to the questions you have set. Enter these responses using SNAP's full range of data entry methods. Write a report on the pros and cons of the different data entry methods and the error checks that are built into the software.

2 To a sample of five male and five female fellow students, pose the question 'Which of the following sports have you participated in over the last 12 months?' Create a list of what you think may be the top 10 sports that fellow students may have participated in, with space for 'Others – please specify'. After conducting the 10 interviews, how would you cope with the coding of the 'Others'? How would you rewrite this question and potential list of sports if you were to repeat the exercise?

3 You are the Marketing Research manager for AGA (**www.AGA.com**). AGA has developed a luxury refrigerator and matching freezer at €4,000 each. A European survey was conducted to determine consumer response to the proposed models. The data were obtained by conducting interviews at shopping malls in 10 European capital cites. Although the resulting sample of 2,500 is fairly representative on all other demographic variables, it under-represents the upper income households. The marketing research analyst who reports to you feels that weighting is not necessary. Discuss this question with the analyst (a student in your class).

4 You are the project manager for a data analysis firm. You are supervising the data preparation process for a large survey on personal hygiene issues. The data were collected via a postal survey and 1,823 questionnaires have been returned. The response rate was excellent, partially due to an excellent prize draw that accompanied the survey. However, you are suspicious about the quality of many responses and 290 questionnaires have missing responses. The data analyst preparing the data has not seen such a level of missing data and does not know how to cope with this survey. Explain to the data analyst how the missing responses and checks on the quality of data to be analysed should be performed.

5 In a small group discuss the following issues: 'Data processing is tedious, time consuming and costly; it should be circumvented whenever possible' and 'Conducting robust sampling is tedious, time consuming and costly; any structural problems in a sample can be simply resolved through weighting.'

Notes

[1] Trembly, A.C., 'Poor data quality: A $600 billion issue', *National Underwriter* 106 (11) (18 March 18 2002), 48; Higgins, K.T., 'Never ending journey', *Marketing Management* 6 (1) (Spring 1997), 4–7; Harristhal, J., 'Interviewer tips', *Applied Marketing Research* 28 (Fall 1988), 42–45.

[2] Keillor, B., Owens, D. and Pettijohn, C., 'A cross-cultural/cross-national study of influencing factors and socially desirable response biases', *International Jornal of Market Research* 43 (1) (First Quarter 2001), 63–84; Dadzie, K.Q., 'Demarketing strategy in shortage marketing environment', *Journal of the Academy of Marketing Science* (Spring 1989), 157–165. See also Nishisato, S., *Measurement and multivariate analysis* (New York: Springer-Verlag, 2002).

[3] Jenkins, S., 'Automating questionnaire design and construction', *Journal of the Market Research Society* 42 (1) (Winter 1999–2000), 79–85; Fink, A., *How to analyze survey data* (Thousand Oaks, CA: Sage, 1995); Alreck, P.L. and Settle, R.B., *The Survey Research Handbook* (Homewood, IL: Irwin, 1985), 254–286; Sidel, P.S., 'Coding', in Ferber, R. (ed.), *Handbook of Marketing Research* (New York: McGraw-Hill, 1974), 2-178–2-199.

[4] Kearney, I., 'Measuring consumer brand confusion to comply with legal guidelines', *International Journal of Market Research* 43 (1) (First Quarter 2001), 85–91; Luyens, S., 'Coding verbatims by computer', *Marketing Research: A Magazine of Management and Applications* 7(2) (Spring 1995), 20–25.

[5] Studt, T., 'Exclusive survey reveals move to high-tech solutions', *Research & Development* 43 (3) (March 2001), 37–38; Frendberg, N., 'Scanning questionnaires efficiently', *Marketing Research: A Magazine of Management and Applications* 5 (2) (Spring 1993), 38–42.

[6] Rydholm, J., 'Scanning the seas', *Marketing Research Review* (May 1993), www.princes.com (23 May 2002).

[7] Cronk, B.C., *How to use SPSS: A step by step guide to analysis and interpretation* (Los Angeles, CA: Pyrczak, 2002); Aster, R., *Professional SAS programming shortcuts: Over 1000 ways to improve your SAS programs* (Phoenixville, PA: Breakfast Communications, 2002); Sincich, T., Levine, D.M., Stephan, D. and Berenson, M., *Practical statistics by example using Microsoft Excel and Minitab* (Paramus, NJ: Prentice Hall, 2002); Middleton, M.R., *Data analysis using Microsoft Excel: Updated for Office XP* (Pacific Grove, CA: Duxbury, 2002).

8 Allison, P.D., *Missing data* (Thousand Oaks, CA: Sage, 2001); Lee, B.-J., 'Sample selection bias correction for missing response observations', *Oxford Bulletin of Economics and Statistics* 62 (2) (May 2000), 305; Freedman, V.A. and Wolf, D.A., 'A case study on the use of multiple imputation', *Demography* 32 (3) (August 1995), 459–470; Malhotra, N.K., 'Analysing marketing research data with incomplete information on the dependent variable', *Journal of Marketing Research* 24 (February 1987), 74–84.

9 A meaningful and practical value should be imputed. The value imputed should be a legitimate response code. For example, a mean of 3.86 may not be practical if only single-digit response codes have been developed. In such cases, the mean should be rounded to the nearest integer.

10 Murphy, K.M., 'Estimation and inference in two-step econometric models', *Journal of Business & Economic Statistics* 20 (1) (January 2002), 88–97; Kara, A., Nielsen, C., Sahay, S. and Sivasubramaniam, N., 'Latent information in the pattern of missing observations in global mail surveys', *Journal of Global Marketing* 7 (4) (1994), 103–126.

11 Some weighting procedures require adjustments in subsequent data analysis techniques. See Batholomew, D.J., *The analysis and interpretation of multivariate data for social scientists* (Boca Raton, FL: CRC Press, 2002); Yaniv, L., 'Weighting and trimming: heuristics for aggregating judgments under uncertainty', *Organizational Behaviour & Human Decision Processes* 69 (3) (March 1997), 237–249; Taylor, H., 'The very different methods used to conduct telephone surveys of the public', *Journal of the Market Research Society* 39 (3) (July 1997), 421–432.

12 Bradford, M., 'Health care access services for expats gain in popularity', *Business Insurance* 36 (1) (7 January 2002), 19–20; Woodside, A.G., Nielsen, R.L., Walters, F. and Muller, G.D., 'Preference segmentation of health care services: the old-fashioneds, value conscious, affluents, and professional want-it-alls', *Journal of Health Care Marketing* (June 1988), 14–24. See also Jayanti, R., 'Affective responses toward service providers: implications for service encounters', *Health Marketing Quarterly* 14 (1) (1996), 49–65.

13 See Johnson, R.A. and Wichern, D.W., *Applied Multivariate Statistical Analysis* (Paramus, NJ: Prentice Hall, 2001); Swift, B., 'Preparing numerical data', in Sapsford, R. and Jupp, V. (eds), *Data Collection and Analysis* (Thousand Oaks, CA: Sage, 1996); Frank, R.E., 'Use of transformations', *Journal of Marketing Research* 3 (August 1966), 247–253, for specific transformations frequently used in marketing research.

14 Vesset, D., 'Trends in the market for analytic applications', *KM World* 11 (4) (April 2002), 14. For a similar data analysis strategy, see Malhotra, N.K., 'Modelling store choice based on censored preference data', *Journal of Retailing* (Summer 1986), 128–144; Birks, D.F. and Birts, A.N., 'Service quality in domestic banks', in Birks, D.F. (ed.), *Global Cash Management in Europe* (Basingstoke: Macmillan, 1998), 175–205.

15 Bivariate techniques have been included here with multivariate techniques. Although bivariate techniques are concerned with pairwise relationships, multivariate techniques examine more complex simultaneous relationships among phenomena. See Tacq, J., *Multivariate Analysis Techniques in Social Science Research Analysis* (Thousand Oaks, CA: Sage ,1996).

16 DeSarbo, W.S., 'The joint spatial representation of multiple variable batteries collected in marketing research', *Journal of Marketing Research* 38 (2) (May 2001), 244–253; Carroll, J.D. and Green, P.E., 'Psychometric methods in marketing research: Part ii: Multidimensional scaling', *Journal of Marketing Research* 34 (2) (May 1997), 193–204.

17 Anon., 'For a scoop of their own', *Businessline* (17 January 2002), 1; Kilburn, D., 'Häagen-Dazs is flavor of month', *Marketing Week* 20 (23) (4 September 1997), 30; Maremont, M., 'They're all screaming for Häagen-Dazs', *Business Week* (14 October 1991).

18 McDonald, G., 'Cross-cultural methodological issues in ethical research', *Journal of Business Ethics* 27 (1/2) (September 2000), 89–104; Alasuutari, P., *Researching Culture* (Thousand Oaks, CA: Sage, 1995); Tan, C.T., McCullough, J. and Teoh, J., 'An individual analysis approach to cross-cultural research', in Wallendorf, M. and Anderson, P. (eds), *Advances in Consumer Research* 14 (Provo, UT: Association for Consumer Research, 1987), 394–397.

19 See, for example, Tian, R.G., 'Cross-cultural issues in Internet marketing', *Journal of American Academy of Business* 1 (2) (March 2002) 217–224; Spiller, L.D. and Campbell, A.J., 'The use of international direct marketing by small businesses in Canada, Mexico, and the United States: a comparative analysis', *Journal of Direct Marketing* 8 (Winter 1994), 7–16; Nyaw, M.K. and Ng, I., 'A comparative analysis of ethical beliefs: a four country study', *Journal of Business Ethics* 13 (July 1994), 543–556.

20 Newman, D.L. and Brown, R.D., *Applied Ethics for Program Evaluation Analysis* (Thousand Oaks, CA: Sage, 1996); Zinkhan, G.M., Bisesi, M. and Saxton, M.J., 'MBAs' changing attitudes toward marketing dilemmas: 1981–1987', *Journal of Business Ethics* 8 (1989), 963–974.

21 Anon, 'How Walcheren hospital perked up their patients', *Research in Business* (March 2004), 15.

Visit the *Marketing Research* Companion Website at www.pearsoned.co.uk/malhotra_euro for additional learning resources including annotated weblinks, an online glossary and a suite of downloadable video cases.