MEASURING EVIDENCE

Statistical Handbook
for value for money
and performance audit


NAO
NATIONAL AUDIT OFFICE


AUDIT SCOTLAND

MEASURING EVIDENCE

Statistical Handbook
for value for money
and performance audit

Preface

Teams working on performance audits and value for money studies have to draw on a range of methodologies and techniques to help them form robust conclusions. Using statistical methods to interrogate quantitative evidence is one of the specialist methods that can add considerable value to audit work. Statistical techniques can be used to provide clear measures of costs, benefits and performance. They can also help to assess the strength of evidence when investigating audit questions, and to ensure that conclusions and recommendations are well-founded and evidence-based.

Government departments, agencies and NDPBs are increasingly producing complex and comprehensive quantitative information. The ability to use such data intelligently and effectively is an important skill for staff engaged on performance audit and value for money work. This handbook, a joint publication by the National Audit Office and Audit Scotland, is a resource that can be used to develop and enhance that skill.

The handbook is a guide on how statistical methods can be applied in performance audit and value for money work. It can be used as a source of reference on particular techniques; to provide background when dealing with specialist statistical consultants; or as a textbook on applied quantitative methodology. It is illustrated throughout with real examples based on actual audit data, and gives examples of how to interpret basic statistical information, as well as providing an introduction to the principles underlying statistically valid reasoning on the basis of such information.

We hope you find the handbook useful. If you have any questions on the material presented here, or would like further advice on applying statistical techniques to the analysis of data in a performance audit or value for money study, we encourage you to contact the report's authors, Alex Scharaschkin and Caroline Jackson at the NAO and Mik Wisniewski at Audit Scotland.

SIR JOHN BOURN
Comptroller and Auditor General
National Audit Office

BOB BLACK
Auditor General
Audit Scotland

Table of Contents

CHAPTER 1

INTRODUCTION

A habit of basing convictions upon evidence, and of giving to them only the degree of certainty which the evidence warrants, would, if it became general, cure most of the ills from which the world suffers.

Bertrand Russell

Chapter 1
Introduction

Purpose of this handbook

1.1 This handbook has been developed jointly by the National Audit Office and Audit Scotland to support the work of auditors and performance audit staff. It is both a source of guidance and a reference manual on the use of statistical techniques in value for money (VFM) and performance audit work.

1.2 Quantitative analysis is one of the most powerful audit tools for developing robust, evidence-based conclusions. Using statistical techniques appropriately to analyse quantitative data can provide clear measures of costs, benefits and performance. Statistical methods also help in properly substantiating conclusions, by providing assessments of the strength of evidence for hypotheses concerning the audit issues. Although the overall conclusions and recommendations of a performance audit or VFM study will always be based on results from a number of qualitative and quantitative sources of information, quantitative measures will often form the basis of headline messages about savings, benefits or effectiveness. For this reason it is important that all staff undertaking performance audit and VFM work should have an understanding of the main ways in which quantitative information can be summarised and analysed, and the basic principles underlying statistically valid reasoning on the basis of such information.

1.3 The handbook is not intended as a substitute for expert advice and guidance from members of your statistical advisory team. Nor is it a guide to the specific skill of using software packages such as Excel and SPSS to analyse data, although examples of output from these packages, and explanations of how to interpret it, have been given throughout the text. Rather, this handbook

- provides information on underlying statistical concepts and techniques that are particularly relevant for performance audit and VFM work;
- gives the necessary background to understanding the results of statistical analyses (whether carried out in-house using relevant software, or contracted out); and
- illustrates the types of questions that can be effectively addressed using statistical methods, with demonstrations of how the results of appropriate analyses can support audit conclusions.

It also serves as a companion text to NAO training courses on statistical reasoning and quantitative analysis.

Using statistical methods for evidence-based conclusions

1.4 In financial audit work, assurance is provided by adherence to an explicit set of accounting and auditing standards. There are certain elements that have to be checked to support an opinion on an account, and the opinion is substantiated by reference to these. The nature of VFM and performance audit work is such that detailed standards of this type are not easily formulated in such a way as to be relevant across the entire field of potential subjects for investigation. Instead, more general **principles** govern good quality VFM studies and performance audits (for a statement of the NAO VFM principles

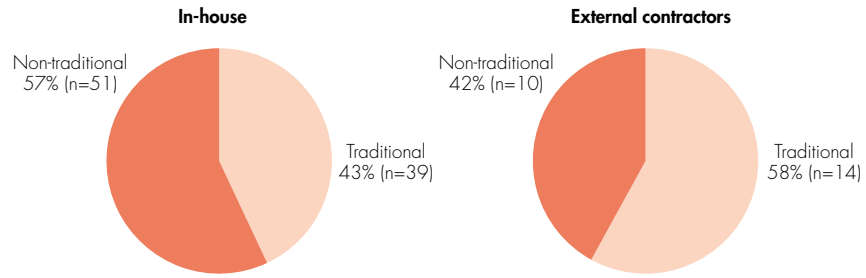
see the NAO *VFM Handbook*). An overarching principle is that VFM studies and performance audits should use an appropriate **research methodology**.

1.5 The key elements of good research methodology are: **clear diagnosis** of the situation being investigated, to come up with clear questions whose answers will provide a conclusion on the issue under investigation; and **appropriate analysis** of the evidence, to answer the audit questions. This handbook is concerned with carrying out appropriate analysis when the evidence consists of quantitative data. The examples below show how using appropriate statistical techniques can help avoid erroneous conclusions, and ensure that inferences are evidence-based.

Example 1: Avoiding the risk of unsubstantiated conclusions

A performance audit team at Audit Scotland collected data on the provision of domestic services in hospitals. A question of interest in the audit was the type of work carried out by domestic staff, classified as either 'traditional' or 'non-traditional' (the latter including a wider range of duties than traditionally associated with domestic service staff). Having collected some data via a survey, the team summarised the results using the figure below.

Remit of domestic staff by type of provider



Source: Survey (n=114)

The team was interested in the question of whether in-house domestic services were more likely to have non-traditional remits than those provided by external contractors.

Since the results shown in the pie charts are based on sample data, the team had to check whether the apparent difference between the two groups of hospitals had arisen by chance, in this particular sample, or whether it represented a real difference in the population of all hospitals. Chapter 5 explains how such questions can be addressed by testing for the 'statistical significance' of the difference. Carrying out a significance test for the difference in this case, as demonstrated in §5.23, shows that it is not significant at the 5% level (see §§4.33–4.34 for an explanation of this terminology).

Hence, the study team concluded that the survey results alone did not provide sufficiently robust evidence to conclude that, across all hospitals, in-house domestic services were more likely to have non-traditional results than those provided by external contractors. In order to substantiate such a conclusion they would have to collect additional information, either by increasing their sample size, or from other sources.

Example 2: Drawing evidence-based conclusions

The NAO study team working on the report Giving Domestic Customers a Choice of Electricity Supplier (HC 85, 2000-01) carried out a large scale survey of electricity customers to investigate the factors that, following the introduction of competition in the market, influenced whether customers switched electricity provider or not. They found that switching rates varied according to a number of factors, as indicated in the list below, which shows the proportion of customers with different characteristics who switched suppliers.

Low income	20%	High income	30%
Rural customer	12%	Non-rural	24%
Pay by prepayment meter	15%	Pay by direct debit	31%
Live in Scotland	15%	Live in England/Wales	25%

There were different market and regulatory arrangements in place in Scotland from those in England and Wales, and the study team was interested in whether there was evidence

to suggest that these had led to Scottish customers having relatively less chance of switching supplier. However, as can be seen from the list, there were other factors associated with low levels of switching, such as being a rural customer; and there are relatively more rural customers in Scotland. In order to control for these factors and establish whether, after having taken them into account, there was still a difference between switching rates in Scotland and switching rates in England and Wales, the study team used a logistic regression model (see Chapter 6). This enabled them to conclude that the rates in Scotland were still significantly lower, even after allowing for the effects of other factors, and helped to substantiate their conclusion about the impact of the regulatory regime in Scotland.

Structure of the handbook

1.6 The remaining chapters of this handbook detail basic statistical techniques and concepts, and illustrate them with examples from performance audit and VFM work.

Chapter 2, *Data sources and management,* deals with collecting and managing quantitative data files. It points to the need to make best use of existing data sources, sets out good practice in setting up datasets before carrying out analysis, and indicates the points to check in assessing data quality.

Chapter 3, *Basic data analysis,* outlines the principles of exploratory data analysis. It sets out different ways of summarising variables and examining variation. It explains the different measures that are appropriate for different types of data, and indicates various classes of questions that can be investigated by graphical and numerical methods.

Chapter 4, *Inference,* is an introduction to the theory that underpins generalising conclusions robustly from sample data. It explains the concepts of sampling error, confidence intervals and statistical significance testing.

Chapter 5, *Statistical testing,* discusses in more detail the process of testing hypotheses, and explains how to carry out some common statistical tests. It identifies which tests are appropriate for different sorts of audit questions.

Chapter 6, *Relationships in data,* introduces simple and multiple linear regression models, and shows how to interpret the results of such models through examples. It also includes a discussion of some extensions to the basic regression model, and lists some other relevant techniques, such as factor and cluster analysis.

1.7 If you would like any further information on any of the techniques covered in this guide, or require assistance in applying them in an audit or VFM study, please contact your statistical advisory team. We hope you find the handbook relevant and useful as a guide and reference manual.

CHAPTER 2

DATA SOURCES AND MANAGEMENT

This chapter deals with collecting and managing quantitative data files. It points to the need to make best use of existing data sources, sets out good practice in setting up data sets before carrying out analysis, and indicates the points to check in assessing data quality.

Chapter 2
Data sources and data management

Introduction

2.1 Well-designed VFM examinations or performance audits are underpinned by clear audit questions. One of the most important tasks for study teams is deciding how to collect the necessary evidence to arrive at robust answers to these questions. In most cases, at least some of the evidence will consist of quantitative data.

2.2 Some of the main ways to interrogate quantitative data and draw valid conclusions are set out in the following chapters of this handbook. But before any analysis can begin, the datasets that will be used have to be assembled and set up appropriately. It is also important to consider data quality and completeness. This chapter introduces some of the key points to bear in mind when collecting data and setting up data files for analysis.

Sources of data

2.3 VFM studies frequently incorporate surveys or censuses of audited bodies. A well-designed survey can elicit vital information which has not previously been collected or analysed, and hence add value to any existing evaluative work in the area under consideration. However it is always worth spending some time investigating existing sources of quantitative data both within the

audited body and outside it, before embarking on any new data collection exercise. Not only does this help to maintain good relations with audited bodies, it also reduces the risk of wasting time in compiling information that is already available, and helps to make questionnaire surveys shorter, more focussed and more efficient. A powerful technique is to combine new survey data with information from existing departmental databases, for instance to examine whether there is evidence of any relationships between quantifiable aspects of business practices and key performance indicators.

Using existing data sources

2.4 A useful source for much up to date quantitative information is the Office for National Statistics (ONS), which is continually updating the statistical information published on its website at www.statistics.gov.uk. A comparable source of information specifically for Scotland is the Scottish Executive website at www.scotland.gov.uk/stats. The ONS concentrates especially on producing economic and population statistics. Many VFM studies and performance audits make use of the economic indicators produced by ONS – for example the GDP deflator index that is used to adjust for inflation when comparing sums of money over time. But there are many other ONS products that can be extremely useful as sources of information for audit purposes. The annual publication *Social Trends*, for instance, uses census data and a number of other large scale surveys carried out by ONS to summarise key UK population and socio-economic information.

Using official statistics

The NAO report *Widening Participation in Higher Education in England* (HC 485, 2001-02) combined background population data on ethnicity and gender from ONS with data on university entry from the Higher Education Funding Council for England to analyse the participation rates of young people in higher education by ethnicity and gender.

2.5 Most ONS products published either on the website or in hardcopy also have a named statistician contact who can provide more information about the underlying data from which figures have been compiled. If you think that a particular set of statistics may be relevant for answering a study question, but are unsure about precise matters of definition or coverage, it is always worth getting in touch with the relevant statistician. It is sometimes the case that ONS holds other, unpublished, analyses based on the data that are exactly what you are looking for. It is also possible that ONS may be able to provide you with a customised analysis of the data.

2.6 Most government departments and agencies produce statistical information about all aspects of their business. When discussing data requirements with departments, you should speak to representatives of their analytical services divisions or technical teams as well as contacts in the substantive area in which you are interested. As well as providing expert advice on precisely what data they hold and any caveats relevant to its use, departmental statisticians may also be able to indicate analyses that, if undertaken, would add value to their own work.

Building on existing data

The NAO team on the report *The New Deal for Young People* (HC 639, 2001-02) was able to extend analyses undertaken by the Analytical Services Division of the Department for Work and Pensions. They used multiple regression techniques to construct context-adjusted performance indicators for the local offices delivering the programme.

2.7 If you are conducting a study or audit in an area in which there are a number of existing data sources and evaluations, it is worth commissioning or carrying out a literature review of the available information. In some cases it may be possible to undertake a **systematic review**, or **meta-analysis** of a number of quantitative studies. Such analyses are common in medical research, where they are used to combine the results of several studies to obtain robust information on, for example, the effectiveness of particular drugs or treatments. The move towards more explicitly evidence-based policy making in the UK has led to the creation of databases of social research such as the Campbell Collaboration (www.campbellcollaboration.org), that are intended to be used to facilitate similar analyses for researchers, evaluators and policy makers. Carrying out a quantitative meta-analysis requires that the sources of information used satisfy certain criteria of rigour, and you should seek expert advice for this type of work.

Collecting new data

2.8 The NAO guide *Taking a Survey* contains information on setting up and administering questionnaire and telephone surveys. It provides guidance on sampling, sample sizes, questionnaire construction and basic analysis and presentation.

2.9 The important topics of sampling methodology and questionnaire design for effective data collection are therefore not discussed in this handbook. The focus of this guide is on how to use the data collected to answer the questions of interest in the audit or study. For further information on designing surveys or other data collection exercises you should refer to the latest edition of *Taking a Survey*, and consult your technical advisory team.

Types of data

2.10 There are a number of ways of classifying data into different types. It is important to be aware of these classifications, as the statistical and analytical techniques introduced later in this guide are organised according to the types of data to which they can be applied.

Categorical and continuous data

2.11 One way of describing data is as either **categorical** or **continuous**. **Categorical** data consists of descriptions or labels used to identify attributes of a subject. A small number of distinct categories contains all the cases, including a separate category for the cases with missing values. For example, respondents to a survey could be categorised as male or female, or as living in the North, South, East or West of the country.

2.12 **Continuous** data can, at least in theory, take on all possible numerical values in a given interval. For example, measurements of length or weight provide continuous data. Actually, nearly all social and economic data of the sort that underpins performance audit and VFM work is, strictly speaking, categorical. For instance sums of money are usually only recorded to the nearest unit that is sensible in the context in which they occur – this might be the nearest penny, or

pound, or even the nearest hundred or thousand pounds. Nevertheless, data of this sort, consisting of a large number of finely divided categories, can be treated as continuous for use in statistical analyses that require continuous data.

2.13 For the purposes of working effectively with datasets and running analyses using software packages it is useful to assign numeric **codes** to categorical data. So you might code data on respondents' gender using the assignment 'female'=1, 'male'=2. This would enable you to set up a dataset (see §2.16) in which the 'gender' column consisted of a series of 1s and 2s. Note that the assignment of numeric codes to categorical variables is entirely arbitrary, and does not imply that it is valid to carry out arithmetical operations on such data. For instance, one could calculate the average of all the numbers in the 'gender' column of a dataset set up as previously described, to obtain a number between 1 and 2. But the concept of 'average gender' is not meaningful.

Ordinal scales

2.14 Categorical data can be divided further into **nominal** and **ordinal** data. As the terminology implies, nominal data are measurements or pieces of information about attributes that do no more than assign subjects to classes. Thus, data on survey respondents' gender is classed as nominal data. If there is a logical ordering on the categories, the data is referred to as ordinal. So for instance data collected from a questionnaire survey in which respondents were asked to indicate whether they 'agree strongly', 'agree', 'disagree' or 'disagree strongly' with a statement would be classed as ordinal. Here there is an underlying scale of agreement, such that 'agree strongly' is higher on the scale than 'agree' or 'disagree'.

2.15 There is some disagreement in the social science literature as to how to treat the data that arise from responses to questionnaire surveys where respondents are asked to indicate their opinions on a scale with a small number of discrete points, such as the scale of agreement mentioned in the previous paragraph. Such questions (known as 'Likert scale questions') are very common in social and market research, and if you use a consultant to carry out such research, you should check how they propose to analyse them. Technically, the data from these questions are ordinal categorical data, and should be analysed using the techniques discussed as appropriate for such data in Chapters 3 and 5 of this guide. In practice, however, some analysts treat them as continuous measures, so they can use a range of other techniques, for example to examine whether there are significant differences in 'average satisfaction' between different groups. There is probably less justification for doing this these days, when most software for statistical analysis is capable of carrying out powerful analyses (such as logistic regression – see Chapter 6) on categorical data. However, treating the data as continuous can sometimes provide results that are approximately correct, and that may be easier to interpret in the context of interest. If you intend to treat the responses to Likert-scale questions as continuous in an audit, you should check the robustness of this approach with members of your technical advice team.

Creating data files

Using spreadsheets

2.16 Data that have been collected via a survey must be entered onto a spreadsheet or similar package before they can be analysed. The standard way in which this is done is to set up a matrix in which each row corresponds to an individual case or survey respondent, and each column

corresponds to a different measurement or piece of data. For example, if you were investigating grants awarded to a sample of 100 small businesses you might have collected information on the size of each grant, when it was made, and the number of employees at the business in question. You should also set up a unique identification number for each grant. This would produce a **dataset** consisting of 100 rows and 4 columns, which would look something like the table below.

Grant_id	Size	Date	No_empl
1	13,500	06-FEB-02	27
2	5,000	19-MAY-02	6
3	4,000	17-AUG-02	11

2.17 All standard software packages for statistical data analysis can import data set up in this type of matrix format. Most packages, such as SPSS and S-plus, have their own inbuilt spreadsheets to enable data to be entered directly into the package. Alternatively, they can read in data from sources such as Excel spreadsheets.

2.18 Often, you will carry out basic analysis in Excel. You may, however, need to use a specialist package for more detailed investigation. In order to ensure that your data sets can be imported easily into other software packages, you should observe the following points of good practice:

- The first row of the spreadsheet should contain only the column names. For maximum compatability with other software packages, these should be no more than eight characters long, be composed only of letters, numbers and underscores, and start with a letter.

- All rows apart from the first should contain only data. There should be no embedded formulae, or notes or comments.
- No text should be entered in a column intended for numbers. This includes entries such as 'N/A' and 'missing'. If text characters are present in a column of numbers, the analysis software may interpret all the numeric information as character strings. To indicate missing data, either leave the cell blank, or use specific codes as explained in §2.31.
- Every case in the spreadsheet should contain a unique identifier. If you create multiple datasets during the study (for example, from initial and follow-up surveys), this identifier must be consistent across datasets. This enables the data to be merged.
- The dataset should be documented. This means recording, in a separate document, what the column names refer to in full, and how any categorical variables have been coded (see §2.13)

Web-based data collection

2.19 If you use a web-based survey tool to collect your data, the data will automatically be entered into some form of database, eliminating any errors due to mis-keying information. Using the web-survey modules of statistics packages such as SNAP or SPSS will result in the dataset being set up in the appropriate format for analysis by the relevant package. [For simple web surveys it is possible to gather data using the 'forms' capability in standard (html) web-page programming, and to store this data in a spreadsheet set up in accordance with the principles set out in the preceding paragraph, so that it can be analysed using any standard package.] See the NAO guide *Netting Results: The Case for Web-based Surveys* for more information on carrying out surveys over the web, and consult your technical advisory team for assistance if you are thinking of carrying out or commissioning such a survey.

Contracting out data entry

2.20 In general, it is usually sensible to contract out large data entry jobs to a data entry bureau. This is usually better value for money than using staff time for an essentially routine process, and you can also specify measures to help ensure that data is entered accurately and reliably. The question of data quality is discussed further in §§2.26–2.32. One mechanism that is often used by bureaux to reduce the risk of entry errors is double keying, whereby the data is independently entered twice. The resulting files are compared, and any entries which differ between them are checked against the original sources and corrected as necessary. It is worth having data double-keyed for any sizeable data entry job. In one Audit Scotland study where this was done, an initial error rate of 5% was noted. That is, 5% of the data in the two files did not match. Double-keying enabled these inevitable data entry errors to be identified and corrected.

2.21 It is very important, if you contract out data entry work, to ensure that the contractors return the data to you in a fully documented form. They should provide a list of all the variables (column names) in the dataset, and a description of how any categorical variables have been represented numerically. It may be worth specifying in advance how you would like categorical variables to be coded. Although most data entry firms will set up codes for you, they may not do so in the most useful or appropriate way for the sorts of analyses you want to perform, and being clear from the outset about how you want the data to be set up can avoid time-consuming re-organisation of the dataset later.

2.22 When designing survey questionnaires, it is useful to think about how the question responses will translate into the matrix format described in §2.16. The number of columns needed in the matrix also provides a more realistic

measure of the amount of information being requested in the questionnaire than the number of questions it contains. If each questionnaire response is going to generate a very large number of columns of data, it may be worth re-checking whether all the questions are essential, and clearly related to the audit issues. Further guidance on questionnaire design in general is available in the NAO guide *Taking a Survey*.

In-house data entry

2.23 It is recommended that you seek advice from your technical advisory team before starting data entry in-house, especially if you are not experienced in setting up datasets. They will be able to recommend appropriate software packages, and also suggest ways of coding variables that are most appropriate for the sorts of analyses you need to carry out.

2.24 If you are going to carry out any relatively large scale data entry in-house, you might want to consider using a 'front end' such as a Microsoft Access form, rather than entering data directly onto a spreadsheet. You can set up forms to mimic the look of questionnaire pages, which makes it easier to enter data accurately. The software package SNAP is another tool that can be used both for designing questionnaires and data entry. It also provides the facility for some exploratory data analysis and statistical testing.

2.25 It may also be worth liasing with your technical advisory team if you are planning on receiving data files from audited bodies. In practice, large data matrices are not usually saved as spreadsheets, but in other formats such as delimited ASCII files. These days, the particular format used is not usually vital, as most analysis packages can interpret a variety of standard file formats. The most important point is that, whatever format is used to encode the data, full documentation is available. As long as it is clear precisely what

information is in the file, and what any numerical codes mean, a member of your statistical or technical team should be able to help you to set it up in an appropriate form for analysis.

Data quality and completeness

Checking data

2.26 Before undertaking any analysis of a set of quantitative data, you should do as much as you can to assure yourself that it has been accurately and reliably compiled. You should check with the providers of the data (e.g., a data entry bureau or a government department) what quality assurance mechanisms they have used to reduce the risk of errors. For example

- have any consistency checks been carried out? It is possible to check whether logical relationships in the data hold (e.g., the age entered for a questionnaire respondent's biological child would have to be smaller than the age entered for the respondent). Any cases where the results are not as expected should be re-checked against the sources.
- Have inadmissible values been excluded? If the dataset has been properly documented, you will know what categories are permissible, for categorical data. An entry of '50' for a data item supposedly restricted to the values 1, 2, 3, or 4 must be an error.

2.27 Using the exploratory data analysis techniques described in Chapter 3 will enable you to detect unusual or outlying values that may represent data entry errors. The EXPLORE command in SPSS, or the DESCRIPTIVE STATISTICS option in the Excel DATA ANALYSIS toolpack are useful for checking whether values seem credible. Sorting data spreadsheets by columns, or using the FREQUENCIES

command in SPSS to produce a frequency distribution, will show all the distinct values that have been recorded, and will enable you to check for any inadmissible values.

Missing data

2.28 **Missing data** can be a problem for data collected via a survey. Although response rates for NAO and Audit Scotland surveys of departments and agencies are generally high, the rates for surveys of the public, or of specific groups such as members of a profession or users of a service, are usually much lower. Low response rates can lead to result **bias**, if the missing data are missing in a systematic fashion. For example, the service users who take the time to respond to a survey might be more likely to include those who wish to express their dissatisfaction with the service than those who are reasonably satisfied. If so, indicators of satisfaction levels calculated from the data are likely to under-estimate the population satisfaction levels.

2.29 A small amount of data missing *at random* is unlikely to bias results (though it can result in having less power to draw robust conclusions, because of the smaller sample size it entails). One way of assessing how randomly data are missing is to compare any indicators for which information is available on the missing cases with the same indicators on the *non-missing* cases. For instance, if the survey has been carried out by post, you will have the addresses of the survey recipients. If comparing the regional distributions for respondents and non-respondents shows that there is a higher proportion of inhabitants of rural areas among the non-respondents than among the respondents, then the results could be biased because of a systematic under-representation of respondents from rural areas.

2.30 If you have concerns about response bias in data resulting from a survey, you should consult your technical advisory team. There are statistical methods known as **imputation** techniques that can be used to adjust for the effects of missing data. Sometimes, however, there is no alternative but to make a concerted effort to increase the response rate, if conclusions based on the data are to be soundly based.

2.31 Missing data can also cause problems for both data input and analysis. Consideration needs to be given at the data input stage as to how to code missing values. These can occur in different ways. One type of missing data will occur when data was expected but not actually provided, for example the respondent was expected to answer a particular question but failed to do so. Another type, however, may be quite legitimately "missing". For example, we may be asking respondents to indicate whether they were satisfied with the service provided by an audited body allowing responses of "Yes" and "No". If "No" we have a follow-up question asking for the main sources of dissatisfaction. Clearly, for those answering "Yes", the follow-up question will not be answered and we will have "genuine" missing values. The coding of the data file will need to distinguish between these.

2.32 Missing values also need to be considered at the analysis stage in terms of which cases will be included in the analysis and which will not. Packages like SPSS typically offer the user a choice between leaving missing values out of the analysis either on a "listwise" or "pairwise" basis. Listwise omission means that each individual variable used in the analysis has any missing values taken out of the dataset for that variable. Pairwise means that cases are included in the analysis only when they have no missing data for any of the variables included in the analysis. The statistical results will differ depending which option is used.

This is illustrated in the table below. Case #4 has a missing value for Size and case #5 has a missing value for No. empl. Analysing these two variables on a listwise basis (for example calculating the mean of each), SPSS would use the existing 4 data values for Size and also the existing 4 values for No. empl. On a pairwise basis, however, neither case #4 or case #5 would be included.

Grant_id	Size	Date	No_empl
1	13,500	06-FEB-02	27
2	5,000	19-MAY-02	6
3	4,000	17-AUG-02	11
4		08-JUNE-02	5
5	7,400	23-MARCH-02	

Summary:
Key points from Chapter 2

- Investigate existing sources of data thoroughly before collecting your own. Can data providers such as the Office for National Statistics help in providing more detailed or customised analyses?
- Discuss your data requirements with representatives of analytical services divisions, or technical teams within departments and audited bodies. Ask for expert advice, if necessary, when agreeing on what will be provided.
- When creating your own data files for subsequent analysis, follow the good practice points in §2.18. If you use consultants for data entry, make sure they provide full documentation on how they have set up datasets. You may wish to provide them with detailed instructions on how to label spreadsheet columns, assign numerical values to data, and code any missing data. Again, if in doubt, seek expert advice earlier rather than later.
- Check survey data for obvious keying errors and impossible values. If you use consultants or a data entry bureau to enter data, check if they double-key the data.
- Check missing data in survey responses. Is there any evidence of systematically missing data? If so, you should seek expert advice on whether there is likely to be any bias in the results.

CHAPTER 3

BASIC DATA ANALYSIS

This chapter outlines the principles of exploratory data analysis. It sets out different ways of summarising variables and examining variation. It explains the different measures that are appropriate for different types of data, and indicates various classes of questions that can be investigated by graphical and numerical methods.

Chapter 3
Basic data analysis

Introduction

3.1 Data analysis starts with clearly formulated questions. The point of collecting and examining any sort of data at all in a performance audit or value for money study is to *address the questions* being investigated. This may seem obvious, but it is surprising how tempting it can be simply to collect a lot of data that seems to be more or less relevant to the general issues under consideration, with the hope that subjecting this to subsequent 'analysis' will then automatically reveal patterns, trends, and even conclusions for the report. Properly *directed* data analysis-that is, analysis undertaken to investigate specific questions-is one of the most powerful techniques in the auditor's toolkit for producing quantified, evidence-based conclusions. On the other hand, just gathering a large quantity of administrative or survey data, and then taking this to a statistical or technical expert with the request to 'analyse the results, and identify the key features and trends' is usually at best inefficient, and at worst ineffective.

3.2 Although it is not always possible to know in advance all the questions which will turn out to be of interest, good study design requires clear thinking about what the key questions for the study will be. Formal issue analysis is a technique that can be very helpful for generating these sorts of specific, focussed questions. As the fieldwork proceeds, and information and data become available, the questions need to be revisited, modified and refined

on the basis of one's increasing knowledge of the realities, subtleties and practicalities of the situation. The case study in this chapter, based on actual data from a recent NAO VFM study, shows in a simplified form how this process works in practice. It also demonstrates how data analysis should not be something that occurs as a discrete task towards the end of the fieldwork. Rather, it is an ongoing process that helps to assess the evidence for possible answers to the audit questions as information is gathered. It may also suggest further questions that should be addressed to forestall potential criticism of conclusions which do not take all the relevant information into account, or which are unduly influenced by unusual or unrepresentative cases.

3.3 This chapter deals with what is known in statistics texts as *exploratory data analysis* (EDA). Most EDA techniques do not require application of the more theoretical statistical tests and methods discussed in the following chapters of this Handbook. The purpose of EDA is to investigate *how key measures and indicators vary*, and the sorts of conclusions we might draw from these patterns of variation. Having arrived at some tentative results, the next stage is usually to ask whether we can *generalise* (for example, from a sample to a wider population), and how we can assess the *strength of evidence* for potential conclusions. These questions can be investigated using the techniques of statistical inference, testing and modelling dealt with in Chapters 4 to 6.

Some important terminology

3.4 We shall illustrate some of the basic procedures of exploratory data analysis using some of the data collected for the NAO report on *Improving Student Achievement in Higher Education* (HC 486, 2001-02) as a case study. Before doing so, however, we need to define some key terms. In data

analysis we refer to the measures, indicators or quantities in which we are interested as **variables**. Usually, we collect information on variables in the form of a spreadsheet of data, sometimes called a **dataset** or a data matrix. Typically the columns represent variables, and the rows show the different **cases** (or 'units', 'observations' or 'records') for which we have information about the variables.

3.5 For example, suppose you were investigating the procurement of printing services by government departments. The following very small dataset is an extract of the sort of data you might have available. It shows the costs of six printing jobs, and the department (*'A'* or *'B'*) that commissioned each job. This defines two variables: cost and department. A third variable-job ID number-has been added to the spreadsheet to make it easier to identify cases. As mentioned in Chapter 2, having a unique case identifier is good practice in data management.

job_ID	dept	Cost
1	A	25,200
2	A	13,278
3	B	18,600
4	A	52,450
5	B	27,386
6	B	22,646

3.6 In carrying out EDA relevant to a study question, we usually distinguish between **response** and **explanatory** variables. Response variables are those that measure or represent the key quantities of interest, while explanatory variables represent factors that we think may affect or influence the responses. (Response variables are also called **dependent** variables, and explanatory

variables are also known as **independent** variables.) For instance, you might have collected the small dataset above to investigate the question '*is department A paying too much for printing services?*' In this case, 'cost' would be the response variable, and 'department' would be an explanatory variable. You would be interested in finding out how cost varies, *conditional,* or *dependent*, on department.

3.7 A general strategy in investigating questions of this sort is first to *describe* the values of the response variable in some way, and then to see whether, using that description, we can identify *differences* broken down by values of the explanatory variable. One simple way of describing, or summarising, a variable is by giving its average. The average cost of the six jobs in this dataset is £26,593. If we now calculate averages separately for the department A jobs and the department *B* jobs-in other words, work out the **conditional** averages for cost, given department-we find that the average cost paid for jobs by department A is £30,309, while the average cost paid for jobs by department B is £22,877. So perhaps there is some evidence to suggest that department A is paying too much (at least in comparison with department *B*'s costs).

3.8 Without further work, clearly, this can be no more than a suggestion. In order to assess the strength of the evidence, we should have to consider how *representative* this very small sample is of the printing costs of all jobs commissioned by the two departments, and consequently how much confidence we can have that, if there really is a difference in average costs, it is of about this size. We should also need to examine whether any apparent difference persists *conditional on other relevant explanatory variables*. The dataset given provides no information on the types of jobs whose costs are quoted. Obviously we should expect differences in costs for jobs of different sizes (e.g. number of pages of copy printed) and complexity

(e.g. colour *vs* black and white printing). These are examples of other explanatory variables that we should have to look at to provide a robust answer to our question. If it turns out to be the case that, conditional on type of job (classified in some sensible manner), the average cost for department A is still higher than that for department B, then we should have a stronger case for concluding that the answer to the question 'is department A paying too much for printing services?' is 'yes'.

3.9 These themes of *description/summary, comparison/conditioning and representativeness/confidence* are central to reasoning with quantitative data, and we shall see them running through many of the examples in this book. The techniques presented in this and the following chapters provide ways of addressing these themes explicitly, and hence helping to support well-argued and properly substantiated conclusions in VFM studies and performance audits.

An overview of exploratory data analysis in audit work

3.10 The box below contains an outline of how EDA usually proceeds in VFM/performance audit work. Each stage of this process is illustrated in the case study example that follows.

The process of exploratory data analysis

1. **Formulate each question to be investigated**

   These should be stated as clearly as possible. If you use the Issue Analysis approach to study design, the questions you will be addressing are those at the lowest level of sub-issue for each main issue, and should, in principle, be capable of being answered 'yes' or 'no'.

2. **What measures capture the variables of interest?**

   You will need to consider whether existing indicators and measures are appropriate for addressing the question, or whether you need to construct your own. Will you be able to obtain the necessary data from administrative or departmental sources, or will you need to collect information yourself (for example, through a survey)? Refer to Chapter 2 for more information about data sources.

   Lack of data may mean you have to revise your question. You might, for instance, narrow its scope to deal with a measure for which data is available. But ***you must not let data limitations lead you into investigating the wrong issues****. If it is simply impossible to obtain relevant quantitative data for a particular issue, perhaps you could obtain other sorts of qualitative evidence. If you have identified an issue whose examination is important to answer questions of economy, efficiency or effectiveness, and for which very limited information is available from the audited body, then that, of course, is a conclusion in itself. Even so, it may be worth trying to collect your own data to address at least part of the issue, to try to gain an idea of the scale of the problem.*

3. Identify response and explanatory variables

*Use a software package such as Excel or SPSS to **depict variables graphically**. You should look at graphical summaries of the variation in each variable, such as bar charts or histograms. This can also help to identify anomalies and errors in the data, such as apparently impossible values, outliers and missing cases. It is important to investigate these as soon as possible, since even the most sophisticated statistical analysis is useless if it is based on unreliable, erroneous or inappropriate data.*

4. Consider the following key questions for each variable

(i) What is its typical value?

(ii) How does it vary?

*These are the key questions concerning the **description**, or **summary**, of each variable. Some of the ways of answering these graphically and by means of numerical summaries are discussed in this chapter.*

5. Consider the following questions for the response variable

(i) How does it vary, *given* (or *broken down by*) the explanatory variables?

(ii) What do you conclude about possible relationships in the data? Does this suggest an answer to the question, or does it suggest modifications of the question followed by further exploration of the data?

*These are the key questions concerning **conditional variation** in the response variable. On the basis of these we can start to build up evidence about possible causal relationships between variables.*

*Note that occasionally you may not have any explanatory variables (if the aim of the study is purely descriptive, for example). In such cases you will not be interested in examining conditional variation. But even purely descriptive investigations may throw up questions of comparison. Often the proper response to a numerical description of a phenomenon (such as levels of recorded crime for the current year) is to ask '**compared with what?**' (e.g last year? Five years ago? Other countries?...) Meaningful descriptions require a context, and that is why comparison, as well as description, is a fundamental part of most EDA.*

6. Can we generalise? How strong is the evidence?

Having come up with some descriptive information, and some indications of possible relationships in the data, the final stage is to examine the extent to which this evidence is robust enough to answer the question of interest. This will often involve quantifying the strength of relationships, and the precision of results calculated from samples. Techniques for answering these types of questions are part of the topics of statistical inference, testing and modelling, and are treated in Chapters 4 to 6.

Case study: Dropout from Higher Education Institutions

3.11 One of the issues of interest to the study team working on the NAO study on *Improving Student Achievement in English Higher Education* (HC 486, 2001-02) was the extent to which students who start a course at a higher education institution drop out before completion. Clearly, funding places for students who drop out wastes money, and denies places to students who might have gone on to obtain a qualification. Finding ways of improving

retention and improvement rates could consequently help to improve the value for money achieved by the Higher Education Funding Council, the body that allocates funding to universities and colleges.

3.12 As part of the investigative work carried out in this study, the team wanted to investigate the question '*Does teaching quality affect dropout rate*?' The box below shows how the general EDA process applies in this case.

Example of exploratory data analysis

1. Formulate the question

The study team had a clearly formulated, quite explicit question: 'Does teaching quality affect dropout rate?'

2. What measures capture the variables of interest?

There are two variables here: dropout rate and teaching quality. Investigation of the available data at the Funding Council, the Higher Education Statistics Agency, the Universities and Colleges Central Admissions Service and the Times Good University Guide showed that an average teaching assessment score, given by the Higher Education Quality Assurance Agency, was available for a large sample of institutions. This provides an indicator of teaching quality that can be used in initial data analyses.

The most comprehensive data available on drop-out rates was for full-time, first degree students, so it was decided to focus initial analyses on this group. This narrows the scope of the original question somewhat. It also means that in examining the relationship between assessed teaching quality and drop-out it may be necessary to consider separately institutions whose student population does not consist mainly of full-time undergraduates.

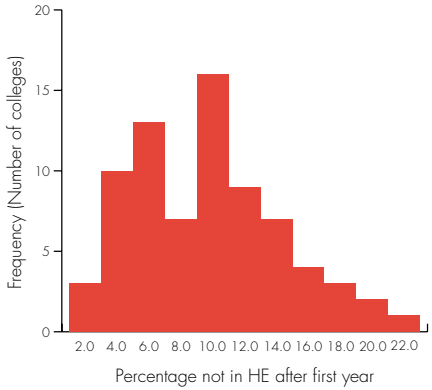
The Funding Council compiles performance indicators for universities and colleges, and included in these is information on drop-out rates. This is actually expressed as a 'continuation rate'. For example if 10% of students drop out of higher education after their first year in a particular college, that college is recorded as having a 90% first-year

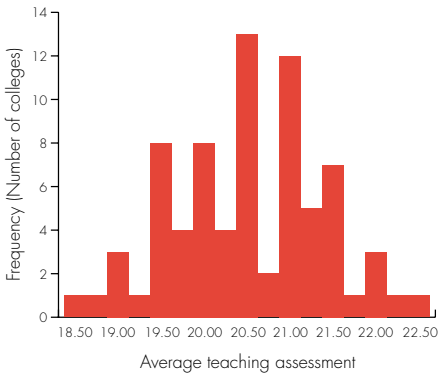
continuation rate. Currently, reliable data is only available for first year continuation rates (although the Council expects to produce indicators for later years as the data become available). However, it appears that most undergraduates who are going to drop out do so during their first year. So it was decided that further narrowing the scope of the question to concentrate on first-year drop-out, at least in the first instance, would be reasonable.

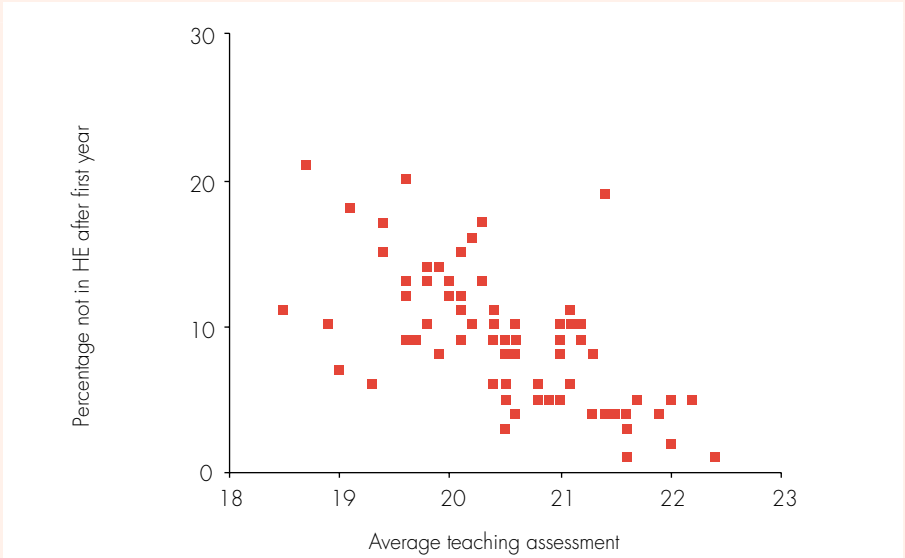
3. Identify response and explanatory variables

The response variable for this piece of analysis is 'first year drop-out rate for full-time, first degree students'. This was obtained from the published performance indicators by subtracting the quoted percentage continuation rate from 100%. The only explanatory variable on beginning the analysis is 'average teaching assessment score'.

Having identified response and explanatory variables, the first task that should be undertaken is a graphical examination of each of them, to gain a feel for their typical magnitudes and ranges of variation. Then we should examine how they vary *together*, to see whether there is any evidence of a relationship between them. Both variables in this case are continuous, so histograms and a scatterplot are appropriate ways of summarising their variation and covariation (the types of graphical summaries appropriate in different situations are discussed further in §3.14). The figures below show these graphical summaries, as produced using the data analysis package SPSS.







The immediate messages we gain from these graphics are:

- There is quite a spread of drop-out rates across institutions. The average rate looks to be around 10%, but there are institutions with rates as low as 1% and as high as 21%.
- The distribution of 'average teaching assessment' exhibits a number of peaks and troughs. This pattern sometimes arises in measures which (as here) are composites, or functions, of other indicators. Each institution's score (assessed on a scale with a maximum of 25) is an average of a number of assessments, and the process of computation, and the rounding rules applied may make certain scores more likely to be obtained than others. These appear more or less apparent in a histogram summary of the data depending on the number of distinct intervals used along the x-axis.

- There is only a small absolute range of variation for average teaching scores (from about 18 to about 22), illustrating the need to use figures quoted to at least one decimal place in accuracy to obtain sufficient discrimination between institutions. This suggests a danger of possible spurious precision in using this indicator to rank institutions in terms of teaching quality. It may be worth revisiting the data available to see whether more disaggregated indicators are available. In judging the validity of the conclusions from any data analysis it is always necessary to consider the extent to which proxy measures (such as 'average teaching assessment') capture the underlying factors of interest (such as 'teaching quality'). The measure of teaching quality used here is possibly the best single summary measure available-and is a recognised part of the higher education quality assurance process- though clearly a single numeric measure cannot completely summarise all aspects of teaching quality.
- Bearing this reservation in mind, there nevertheless appears to be a fairly strong relationship between teaching quality and drop-out, as illustrated in the scatterplot. The better the teaching quality, on the whole, the lower the drop-out rate.

4. Describe typical values and variation

Although histograms and scatterplots show graphically how measures of interest vary, it is also useful to have some quantitative indicators of 'typical' values. Appropriate indicators to use, for different types of data, and their interpretation, are discussed in §§3.21–3.33 below. For the continuous variables under consideration here, information on typical values is captured by the **mean** and **standard deviation**. The mean (the ordinary average, calculated by summing values and dividing by the number of cases) gives an indication of the 'central tendency' of the data, and the standard deviation gives an idea of how spread out around the mean the values of the variable are. Another indicator of spread that is useful when comparing two or more variables is the **coefficient of variation**. This is just the standard deviation divided by the mean, expressed as a percentage. It enables a direct comparison of variation, even when the variables concerned have different absolute magnitudes.

The table below gives the means, standard deviations and coefficients of variation for the two variables of interest in this example (once again, produced using the SPSS analysis package).

	Mean	Standard deviation	Coeff. of variation
Dropout rate (per cent)	9.2	4.6	50%
Teaching quality (score)	20.5	0.9	4%

Number of cases in the sample: 75

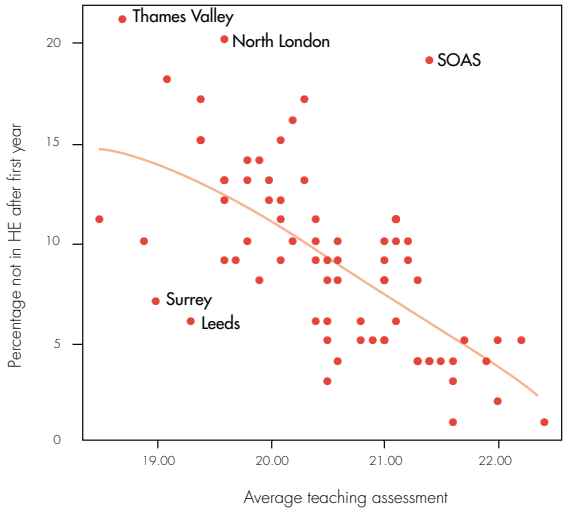
From these summary statistics it is apparent that the average dropout rate is 9.2%, but there is considerable variation around this average across the institutions in the dataset. On the other hand, the average summary measure of teaching quality is quite high (20.5 on a scale with a maximum of 25), and there is little variation around this. This suggests that it may be worth investigating in more detail the way in which the 'average teaching quality' measure is derived. It is clear that only a small part of the 25-point scale is actually being used for this measure, and hence its value as a discriminator between institutions is less than optimal. The reasons for this could relate to the construction of the indicator itself (averaging several sub-measures will tend to reduce overall variation). They might also have to do with the way in which assessors actually mark quality of teaching (they may tend to use only the higher end of the scale, perhaps because of the way the assessment criteria are defined). It is possible that following this up might lead to a conclusion about part of the higher education performance measurement regime. If you were undertaking this analysis as part of the study team, you would have to decide how relevant this was to the scope of the study.

5. Examine the relationship between response and explanatory variables

The first step in appraising the evidence for a relationship between two continuous variables should always be to examine a scatterplot. If you use a statistical analysis package such as SPSS to create such a plot from a dataset, you can use it not only to assess the extent to which a trend is present, but also to identify cases which appear to deviate from such a trend. The latter can often be of particular interest in VFM work, as

they can provide instances of unexpectedly good or poor performance, and hence may be useful as examples of good or poor practice (or at least provide some pointers as to where to start looking for 'best practice' examples).

The figure below reproduces the scatterplot of dropout rate and teaching quality, but with a trendline added, and with some of the more 'atypical' cases labelled. This was created using the Interactive Graphs feature in SPSS, which enables identification of individual cases on such a display. The trendline is created by the software using a 'local regression' algorithm. You can think of this as providing an average for the response variable, conditional on the value of the explanatory variable at each point on the line. Here, the trend is for the average dropout to decrease, as teaching quality increases.



A common way of quantifying the strength of the relationship between two continuous variables is by means of the (Pearson) **correlation coefficient**, discussed further in §3.40. This is appropriate when (as here) the relationship seems to be reasonably approximated by a straight line that is, we can approximate the curved trend line in the diagram with a simple straight line.

The value of the correlation between average teaching assessment and dropout rate for this dataset is –0.64. The negative sign indicates that the two variables are negatively related: institutions with higher average teaching assessments tend to have lower rates of dropout of first-year students. The correlation is reasonably strong in magnitude. As noted in §3.41, the square of the **correlation coefficient** indicates the extent to which the explanatory variable 'explains' the variation in the response variable. Here, $-0.64^2=0.41$, so average teaching assessment accounts for 41% of the variation in drop-out rates.

Some of the points in the scatterplot that represent institutions which appear to differ from the general trend have been labelled. Surrey and Leeds have relatively low scores on the teaching assessment measure, but also have low rates of dropout. Thames Valley and North London, on the other hand, have higher than average rates of drop out given their teaching assesment scores. However, it can be seen from the plot that these four institutions fall in a relatively sparsely populated part of the cloud of points in the scatterplot. Using a statistical package to obtain a 'five number summary' of average teaching assessment (see §3.27), shows that the lower quartile of this variable is 19.9 (in other words, that three-quarters of institutions have teaching scores better than 19.9). There is less evidence of a definite trend for institutions with relatively low teaching scores, as relatively few institutions fall into this category, and those that do vary quite considerably in their dropout rates. So one should be cautious in concluding that the four institutions mentioned are definitely 'atypical'.

On the other hand, SOAS (the School of Oriental and African Studies, part of the University of London) does appear to be a special case (often referred to as an outlier in this type of analysis). It has a high teaching assessment, but also a high dropout rate. On investigation, it became clear that this institution differs in many respects from most of the others in the sample. It has a high proportion of overseas students, and one of the highest proportion of mature age students. It offers courses in highly specialised subjects such as ancient and modern oriental languages, many of which are completely new to students on entry. It therefore seems appropriate to consider SOAS separately from the bulk of institutions in the sample when examining the relationship between teaching quality and dropout rates.

Removing SOAS from the sample, and recalculating the correlation coefficient for the remaining cases, gives a value of –0.70, indicating that, SOAS aside, there is evidence of a moderately strong relationship.

6. Can we generalise? How strong is the evidence?

The analyses shown above indicate that there is reasonable evidence to suggest a relationship between teaching quality (at least, those aspects of teaching quality captured by the official 'average teaching quality' indicator) and the dropout rate for first year, full-time undergraduates. On average, the better the teaching score, the lower the rate of dropout. This suggests it would be worthwhile to look more closely at specific aspects of teaching quality and practice, to see if there are any potential recommendations that might help colleges reduce their dropout rates. This was in fact done in the study, using interviews and a questionnaire survey.

Clearly, there are many other factors that might help to explain the variation in dropout rates between institutions. An obvious one is the prior qualifications of the students at each institution. The average A-level score on entry for the institutions in the sample was also available from the Universities and Colleges Central Admissions Service. It turns out that the correlation between average entry score and dropout rate (for institutions excluding SOAS) is -0.89. So student prior achievement is a better predictor of dropout than is teaching quality. However, these two explanatory variables are not statistically independent of each other. The correlation between them is 0.71, indicating that colleges with high assessed teaching quality also tend to have students with high prior qualifications. To provide a more sophisticated account of the relationship between teaching quality and dropout, we need to look at the extent to which teaching quality affects dropout rates, *controlling* for students' prior attainment. The techniques for addressing these sorts of questions are introduced in Chapter 6. You will probably need to seek expert advice to help you construct and interpret the sorts of regression models discussed there. But the task of doing this is considerably facilitated if you have spent some time carefully exploring your data using the sorts of EDA techniques outlined in this chapter. You will then be in a position to have some tentative answers to your research questions, to be able to discuss possible limitations in the data, to note interesting individual cases that might be followed up to illustrate particular points, and to be clear about precisely what kind of more sophisticated analytical work could add most value to the study.

Key tools and methods: graphical examination

Choosing a method

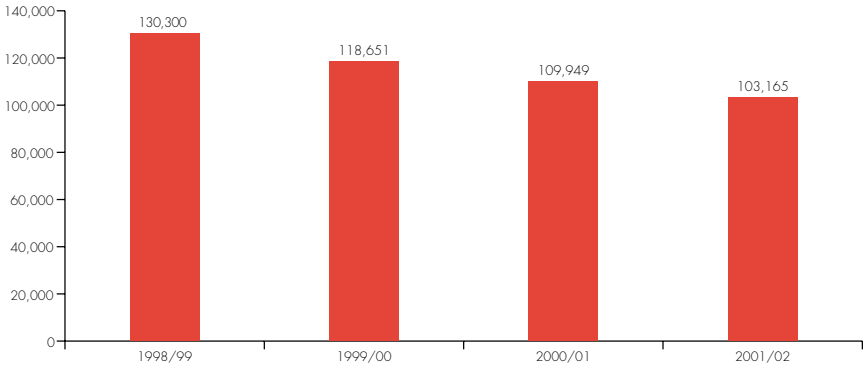
3.13 The case study above illustrated some of the main ways of exploring data **graphically** (through histograms and scatterplots, in this case), and through **numerical summaries** (e.g. averages, standard deviations, correlation coefficients). There are a number of standard graphical methods, and summary measures, that can be used in most applications of EDA. Their appropriateness depends on

- the *type of variables* being examined: whether they are **categorical** or **continuous** (see Chapter 2 to remind yourself of the definitions of these terms); and
- the *type of question* being asked: whether it is a **summarising** (descriptive) question about a variable, or a **conditioning** question asking about the effect of an explanatory variable, or variables, on a response.

Summarising variation

3.14 For a **summarising** question, it is usually best to start with producing **bar charts** for categorical variables, and **histograms** or **boxplots** for continuous variables. Examples of these types of graphical displays are shown below.

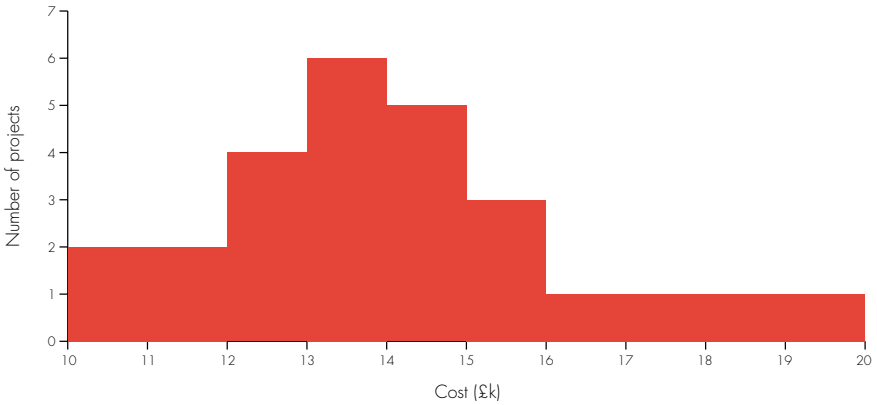
Example: bar chart



Features of bar charts

- Usually show numbers or proportions in each category
- Only the *height* of the bars is relevant, not their width
- Line charts can be used as equivalents: in these, bars are replaced with points at their tops and connected by lines
- Bars for each category do not touch

Example: histogram



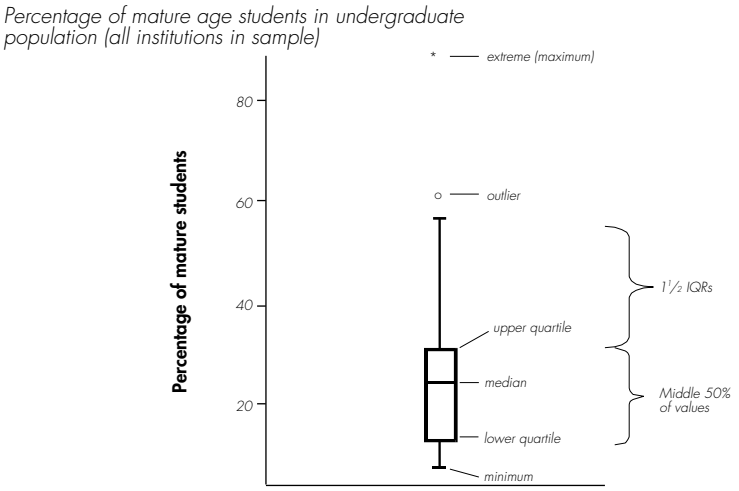
Features of histograms

- Bars touch
- Can have bars of different width, but need to adjust their height accordingly
- The *area* of each bar represents the number of cases.

Data for example here:

Cost of project	Number
£1000–£1199	4
£1200–£1299	4
£1300–£1399	6
£1400–£1499	5
£1500–£1599	3
£1600–£1999	4

Example: boxplot



Features of boxplots

- A way of quickly assessing variation and skew.
- The box shows the interquartile range (IQR) – the middle 50% of values.
- Other elements of a typical boxplot are marked on the example. The line below the box extends to the minimum value, unless this is more than 1.5 IQRs below the lower quartile, in which case the line extends 1.5 IQRs. Similarly for the line above the box.
- Values more than 1.5 IQRs away from the upper or lower quartiles are called outliers, and shown separately.

Examining conditional variation

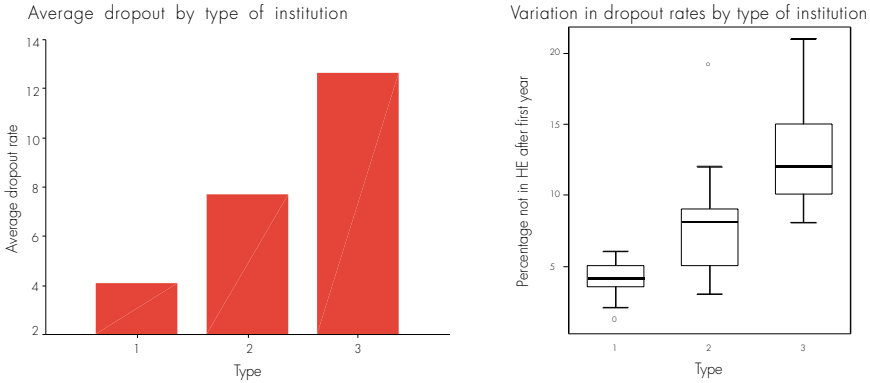
3.15 For a **conditioning** question, the table below shows some graphical displays that can be used.

Displaying conditional variation		
	Categorical response	**Continuous response**
Categorical explanatory	■ Bar or line chart of response proportions in each category	■ Bar or line chart of response averages in each category; ■ Grouped box plot
Continuous explanatory	■ Scatterplot of response proportions in each category	■ Scatterplot; ■ Smoothed trendline/local regression line

3.16 Basic graphics such as bar charts can be adapted to deal with conditioning questions by using *summary functions*, such as proportions or averages, as indicated in this table.

3.17 For instance, in the NAO study of higher education retention and achievement already discussed in the case study earlier, a question that arose was 'how do dropout rates vary between types of institutions?' There is a standard three-way classification of higher education institutions into 'Russell Group' universities (the older institutions such as Oxford, LSE and UCL), 'pre-1992' and 'post-1992' institutions. The classification of each institution in the sample was available in a column in the dataset headed 'type', with '1' indicating 'Russell Group', '2' indicating 'pre-1992', and '3' indicating 'post-1992'. Thus the question of interest here was to examine the effect of the categorical explanatory variable 'type' on the continuous response variable

'dropout rate'. The table above shows that possible ways of examining this graphically are a bar chart showing the average dropout by type of institution, or a grouped box plot, showing a box plot for each type of institution. These two possibilities are illustrated below



3.18 Although the bar chart showing average dropout by institution type is possibly a more familiar type of graphic, it does not convey as much information as the grouped box plot. The box plot, as well as showing the differences in average dropout rates between types of institution, also summarises how dropout rates vary within each type. It shows that the ranges of variation for types 1 and 3 do not overlap, while the ranges for types 2 and 3 do, to some extent. Although the average dropout is higher for type 3 than for type 2, there is an institution of type 2 with a dropout rate higher than the upper quartile for type 3.

3.19 As shown in Chapter 5, this sort of information is very useful when assessing whether apparent differences between groups that arise from a consideration of sample data (as is the case here) are likely to reflect real differences in the whole population from which the sample was taken. If the variation within

groups is less than the variation *between* groups (as is the case in comparing type 1 and type 3 institutions), this provides evidence to suggest that the apparent difference is more likely to be 'real'. If, on the other hand, the variability within groups is reasonably large, in comparison with the difference between them, on some summary measure such as 'average value', then there is less evidence that there is a real population difference. In order to assess the strength of evidence formally it is necessary to use the kinds of techniques discussed in Chapter 5. Graphical examination of the data, however, provides an important visual summary of conditional variation when investigating relationships between variables.

3.20 In summary, a vital first stage of any exploratory data analysis is a graphical examination of the data, directed by the questions of interest. Software packages such as Microsoft Excel have some facilities for doing this. Specialist statistical analysis packages such as SPSS and S-plus are perhaps more useful for graphical EDA (the figures in this chapter were all created in SPSS). The 'Interactive graphs' feature of SPSS can be used to create two and three dimensional graphical displays. It is possible to view displays from different angles and to edit them to incorporate different variables or summary functions. Similar facilities exist in S-plus, which also has a larger standard 'gallery' of graphics, and the potential for more customised output. Detailed instructions on how to use these packages are beyond the scope of this guide. It is, however, easy to create graphics in either of them, and you should explore such packages further, if you need to undertake exploratory analysis of quantitative data as part of a performance audit or value for money study.

Key tools and methods: numerical summaries

3.21 Because it is easier to compare figures than lists of numbers, graphical methods should be used to gain an initial impression of how quantities of interest vary when investigating an issue. In order to firm up conclusions and enable more formal testing of hypotheses, however, it is necessary to use numeric summaries to capture key features of the data. This section presents the main summary measures that are used in EDA. Many of these also form the basis of the statistical testing techniques described in Chapter 5.

Summarising variables numerically

3.22 The table below shows measures that can be used to answer **summarising** questions. Measures for conditioning questions are discussed in §§3.34–3.43. In the table, categorical variables have been subdivided into two classes: those in which there is no particular ordering among the categories, and those in which such an ordering exists. An example of a **non-ordered categorical** variable is 'gender'. If one has data on the gender of respondents to a questionnaire survey, for example, this just provides a way of dividing the respondents into categories (male and female): there is no inherent ordering of these categories. On the other hand, a variable such as 'degree class' is an **ordered categorical variable**. A First is better than a 2:1, which in turn is better than a 2:2, etc. This is not a continuous variable, and there is no sense in which a First is 'three times as good as' a Third, for example; but the categories are clearly ordered nevertheless, and it is possible to make use of this fact in analysing data of this kind.

3.23 The table is hierarchical in the sense that measures for categorical variables can also be used for ordered categorical variables and for continuous

variables, and measures for ordered categorical variables can also be used for continuous variables.

Summarising variables	
Type of variable	**Measures**
Categorical: non-ordered	• Proportions in categories
	• Mode
Categorical: ordered	• Median
	• Range
	• 5 number summary
Continuous	• Mean
	• Variance; standard deviation
	• Trimmed mean
	• Coefficient of variation
	• z-score

3.24 The simplest way of summarising a categorical variable is simply to give the **proportions of cases** that fall into each category. For example, it might be the case that 63% of the respondents to a survey carried out as part of an audit were male, and 37% were female. The **mode** is the category containing the highest proportion of cases (male, in this example). Variables can be **multi-modal**, if there are several categories containing equal highest proportions.

3.25 For variables whose values can be ordered, a useful indicator of the 'central' value is the **median**, or **50th percentile**. This is the value or category such that (roughly) half the cases in the dataset fall in a higher category on this variable, and (roughly) half the cases fall in a lower category. For continuous variables it is possible to calculate the median such that it divides the data exactly in half (computer packages such as Excel-using the 'data analysis'

tools-or SPSS can be used to do this). For ordered categorical variables with a relatively small number of categories, the median category is chosen as the one that gives the closest to a 50:50 split.

3.26 As well as describing the central value of a variable, it is important to be able to describe quantitatively the extent of variation around that central value. This provides an idea of how 'typical' the central value is, and, as noted in §§3.18–3.19, is particularly important when comparing variables across groups. A simple measure of variation for numeric-valued variables is the **range**, the difference between the highest and lowest values. Clearly, however, this may give a misleading picture of the extent of variation if most values are not close to these two extremes. Another measure which is often used is the **inter-quartile range (IQR)** or **mid-range**. This is the range within which the middle 50% of values of the variable fall. It extends from the **lower** to the **upper quartile** of the variable. (**Quartiles** are defined analogously to the median. The lower quartile, or **25th percentile**, is a value such that 25% of the cases in the dataset fall in a lower category on this variable, and 75% of the cases fall in a higher category. The upper quartile, or **75th percentile,** is a value such that 75% of the cases fall in a lower category, and 25% fall in a higher category, on the variable.) The box on a box plot shows the interquartile range. The bottom of the box is drawn at the lower quartile and the top of the box at the upper quartile. Cases that are more than 1½ IQRs away from either the lower quartile or the upper quartile are often called **outliers**. They are 'atypical' cases that may require separate examination.

3.27 Computer packages such as Excel, SPSS and S-plus will calculate the median and upper and lower quartiles for a variable in a dataset. A common way of providing a quick description of variation in numeric-valued or continuous quantities is by means of the **five number summary**. This consists of the minimum, lower quartile, median, upper quartile and maximum values of the

variable in question. From this summary you can also see the range (difference between the minimum and maximum) and the IQR (difference between the lower and upper quartiles). You can gain some idea of the shape of the distribution of values of the variable. For example, if the difference between the upper quartile and the maximum is much greater than the difference between the minimum and the lower quartile, then the values of the variable are more 'stretched', or **skewed**, to the right. In general, however, it is easy to assess shape and skew in variables by looking at plots such as histograms.

3.28 The most common measure of 'central tendency' for a continuous variable is its **mean** (the usual 'average' calculated by adding up all the values and dividing by the number of cases). Although the concept of 'average value' is sufficiently well understood to make the mean a useful summary quantity to quote in reports, you should be aware that it is affected by extreme values in the data, and it may be an inappropriate summary of the 'central' value for data that is highly skewed. This sometimes happens with datasets recording payments of money. For example, consider the following data, which represents the salaries of nine staff members in a department:

£15,950	£31,150
£19,901	£31,150
£24,640	£37,000
£24,642	£85,800
£26,850	

The mean salary here is £33,009. Note, however, that seven out of the nine staff members in the department have salaries lower than this. This is due to the influence of the single atypically large salary of £85,800. In summarising this data it would be better either to calculate the mean for the data excluding

this atypical value (and report the outlier separately), or to use the median to indicate central tendency. The data are listed in order here, so the median-the middle value – is £26,850.

3.29 Many statistics packages, as well as providing the mean value of a continuous variable, will also provide a trimmed mean. This is the mean of all the data excluding the extremes. Often, a 95% **trimmed mean** is used, in which the most extreme 5% of the data are discarded when calculating the mean. This is in accordance with the reasoning of the preceding paragraph. If you notice, in examining output from a data analysis package, that the 95% trimmed mean differs considerably from the mean, this probably indicates the presence of extremes in the data, and you may wish to consider those cases separately (or check that they are not data entry errors).

3.30 The measure of dispersion or variation around the average that is most commonly quoted for continuous variables is the **standard deviation**. This is a very convenient measure mathematically, and has certain properties that make it ideal for use in formal statistical testing, as described in Chapter 5. The formula for calculating the standard deviation is slightly complicated, but given below for reference. Essentially, it is calculated by working out the average squared distance of each data point from the mean, and then taking the square root of the result (so that the result is expressed in the same units as the original data).

$$s = \sqrt{\sum_{i=1}^{n} \frac{(\bar{x} - x_i)^2}{n-1}},$$

where the sample consists of n data points x_i, and the sample mean is $\bar{x}$.

In practice, you can obtain the standard deviation of any variable in a dataset from a package such as Excel. The square of the standard deviation is called the **variance**.

3.31 How should the standard deviation be interpreted? One useful fact is that for a variable which is *approximately normally distributed* (see Chapter 4), about two-thirds of the cases are within one standard deviation of the mean, and about 95% of cases are within two standard deviations of the mean. So if you know that a variable has a mean value of 100, say, and a standard deviation of 15, then (if the distribution of values of the variable is roughly normal), you would expect the majority of the values of the variable to lie between 85 and 115, and nearly all the values to be included in the interval 70 to 130. A value more than two standard deviations away from the mean would be unusual. The term ***z*-score** is used to express how many standard deviations away from the mean a particular value is. In this example, the value 115 is one standard deviation greater than the mean, so its *z*-score is +1.0. The value 77.5 is one and a half standard deviations less than the mean, so its *z*-score is -1.5.

3.32 You may also see the standard deviation being used as an indication of **risk** in project appraisal. The Treasury publication *Appraisal and Evaluation in Central Government* (the 'Green Book') sets out the cost-benefit framework within which departments are expected to appraise policy options. As explained in the NAO guide *Measuring Costs and Benefits: A Guide on Cost Benefit and Cost Effectiveness Analysis*, the results of such appraisals are usually summarised as net present values. In general, the financial models used in appraising large scale projects are complex, and attempt to take account of uncertainties in the information upon which they depend. Increasingly, therefore, the results of such modelling present more than just a single number (the expected net present value) to summarise the net benefits or costs of a project. Sometimes, a distribution of possible values will be

presented (see §§3.36–3.34 of *Measuring Costs and Benefits*). Some of the information in this distribution can be captured by giving its mean and standard deviation. The mean is the 'expected' net present value, and the standard deviation provides a measure of the risk that, in fact, the actual costs/benefits of the project may be different from this expected value. A small standard deviation suggests that most possible outcomes cluster quite closely around the expected value, so there is less risk of an unexpectedly low (or high) net present value for the project than there would be if the standard deviation were higher.

3.33 Because the standard deviation is a measure of dispersion *around the mean*, it is difficult to compare directly the standard deviations of variables with different means. The **coefficient of variation** (CV) is defined as the standard deviation divided by the mean (generally expressed as a percentage). CVs can be compared between variables: see the example given in the case study earlier in this chapter.

Quantifying conditional variation

3.34 The final set of numeric summaries used in exploratory data analysis are those for examining **conditioning**, rather than summarising, questions. The table below is analogous to the one presented earlier in the discussion of graphical methods. It shows some of the methods that can be used to explore the effects of categorical or continuous explanatory variables on categorical or continuous response variables.

3.35 Summarising conditional variation numerically brings us close to the topic of statistical testing. The square-bracketed entries in the table show some of the methods than can be used to explore each type of question in more detail.

Summarising conditional variation		
	Categorical response	**Continuous response**
Categorical explanatory	■ Response proportions by category (look at differences/relative risks) ■ Response medians by category (for ordered data) ■ [contingency table/ chi-squared]	■ Response means by category (look at differences/effect sizes) ■ [*t*-test/analysis of variance]
Continuous explanatory	■ Response proportions for grouped values of explanatory variable ■ [logistic regression: odds ratios]	■ Correlation coefficient * [regression: regression coefficients]

3.36 The terms in square brackets are explained in Chapters 5 and 6. The other new terms in the table are 'relative risk', 'effect size' and 'correlation coefficient'. These concepts are discussed below.

3.37 The **relative risk** is a measure of how a categorical response variable depends on a categorical explanatory variable. It is simply the ratio of response proportions between two categories of interest of the explanatory variable. It is frequently used in health and health economics contexts. For example, the table below presents the results of a trial of the use of beta-blocker drugs as prevention against heart attacks. In this experiment (a randomised controlled trial), 1,000 middle-aged male patients were given beta-blocker treatment, while an equivalent group of 1,000 other patients

was not treated. The table shows the number of cardiac events that occurred within one year in each case.

No. of cardiac events	
Control group (not treated)	10
Trial group (treated)	7

The relative risk of a cardiac event, for treated patients compared with non-treated patients, is (7/1,000)/(10/1,000) = 0.7. In other words, those who took the beta-blockers had only 70% of the risk of a cardiac event of those who did not. Note however that relative risks do not provide any information about the *absolute* likelihood of an event. For both groups in this trial, the vast majority of patients did not suffer a cardiac event. The absolute reduction in proportions is only (10/1,000)-(7/1,000) = 0.3%. If you do calculate relative risks, or if a consultant provides information on relative risks as part of a report for you, it is always worth checking the absolute numbers as well.

3.38 A simple measure of how a continuous response variable differs between the categories of a categorical explanatory variable is just the differences in mean values. The example using a small dataset on costs for printing jobs in two government departments in §3.5 took this approach. Here, the question of interest is whether the response variable (cost) is affected by the explanatory categorical variable (department A or department B). The table below (produced using the Excel AVERAGE and STDEV functions on the dataset) gives the mean and standard deviations of costs for departments A and B separately.

	Dept A	Dept B
	25,200	18,600
	13,278	27,386
	52,450	22,646
mean	**30,309**	**22,877**
St. dev.	**20,079**	**4,397**

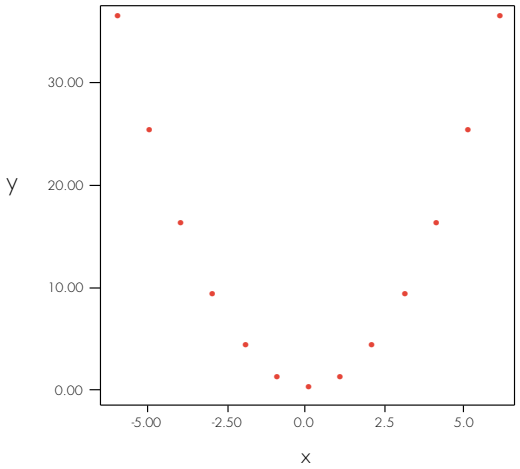
The difference in means between the departments is £7,432. But note also that there is more variation within department A than within department B. How important is a difference of around £7,000, given the range of magnitudes across all the cases? One way of assessing this is to look at the effect size the difference represents. This essentially expresses the difference in standard deviation units, rather than in absolute terms. The **effect size** in this case is 0.5, i.e., the difference between the groups is about half a standard deviation. The effect size is calculated by dividing the difference in means by the 'pooled' standard deviation of the two groups. You can think of this as an average of the standard deviations across both groups. (In fact, it is calculated as $\sqrt{(n_1 s_1^2 + n_2 s_2^2)/(n_1 + n_2 - 2)}$, where n_1 and n_2 are the number of cases in each group, and s_1 and s_2 are the standard deviations.)

3.39 An effect size of 0.5, as in this example, is generally taken to be moderate. As a rough rule of thumb, effects of more than one standard deviation are often regarded as reasonably substantial. However, judgement, based on contextual knowledge, should guide the interpretation of an effect size for any given situation.

3.40 A measure of the strength of the (linear) relationship between two continuous variables is provided by the (Pearson) **correlation coefficient**. This is an index that

varies between −1 and 1, where −1 indicates 'perfect negative relationship', 0 indicates 'no relationship' and 1 indicates 'perfect positive relationship'.

3.41 Note that the correlation coefficient only measures the strength of *linear* (straight line) relationships. The diagram below shows a perfect relationship between two variables (in fact here $y=x^2$), but one for which the correlation is zero. This illustrates once again the importance of always exploring data graphically. In fact, it is possible to use correlational techniques when the relationship is not a straight line one, but this requires transformation of the variables, a topic which is touched on in Chapter 6.



3.42 It is often useful to calculate the square of a correlation coefficient. This represents the 'proportion of variation in one variable explained by the other variable'. For example, a perfect positive relationship would have a correlation coefficient of 1.00. In this case, all the points in the scatterplot would lie on a straight line, and knowing the value of one variable would be tantamount to knowing the value of the other. In other words, 100% of the variation in the response variable would be explained by the explanatory variable (and *vice versa*). This is reflected in the fact that the square of the

correlation coefficient in this case is 1.00. A correlation of 0.8, for instance, would mean that 64% of the variation in the response variable is accounted for by variation in the explanatory variable.

3.43 Finally, it is worth reiterating the well-known saying that 'correlation does not imply causation'. Simply finding a significant correlation between two variables does not imply on its own that one causes the other. For example, examining Australian datasets reveals a positive correlation between ice cream sales and the number of shark attacks on swimmers. This is due to an underlying factor common to both variables: ice cream sales and shark attacks both increase during the summer. Supporting conclusions about causality requires more than purely statistical data. Nevertheless, quantitative information can be an important part of the evidence used to arrive at causal conclusions.

A final word on graphics

3.44 This chapter has emphasised the importance of exploring data graphically to examine patterns of variation and trends. It should be noted, however, that while software packages such as Excel and SPSS can produce graphical output that is extremely useful for exploratory data analysis, such output is not always sufficient on its own for use in VFM or performance audit reports. There are many other concerns, beyond purely technical ones, in designing graphics that get messages across clearly while remaining faithful to the underlying data. One of the best introductions to some of the issues involved doing this effectively is Edward Tufte's *The Visual Display of Quantitative Data* (Connecticut: Graphics Press, 1983). You might also want to look at Tufte's website (www.tufte.com), which contains a number of other links to resources on *information design*, as this topic is called.

3.45 Graphical presentation and information design per se are beyond the scope of this guide. You are encouraged to use software packages to explore graphically, as a matter of course, the datasets you collect in VFM or performance audit work. You may wish, however, to consult with colleagues in your central advisory team when designing graphics based on quantitative data that you intend to use in reports for publication.

Summary: Key points in Chapter 3

- Data analysis is driven by questions. Formulate your questions as clearly as possible.
- Consider what variables should be investigated, based on your questions. Which are response variables and which are explanatory?
- Explore the data graphically, using the suggested types of graphics outlined in this chapter. Watch out for outliers and unusual values-what do they tell you?
- Consider the key questions set out on page 39. Your aim is to use the data to arrive at plausible tentative answers to questions, that can be tested further, if necessary, using the methods of the later chapters of this guide, or by comparing with other qualitative and quantitative evidence.

CHAPTER 4

INFERENCE

This chapter is an introduction to the theory that underpins generalising conclusions robustly from sample data. It explains the concepts of sampling error, confidence intervals and statistical significance testing.

Chapter 4
Inference

Generalising from sample data

4.1 The previous chapter summarised a number of the basic techniques that can be used to explore datasets. Often, the data under consideration is a sample taken from a wider population, and a fundamental question of interest is the extent to which results obtained from the sample data provide evidence for conclusions about the population as a whole. This chapter introduces some of the basic theory that underpins **inferential** or inductive reasoning from sample to population data.

4.2 A theoretical result of fundamental importance for statistical inference is the **central limit theorem**, discussed briefly in §4.21. It is this result that makes inferential techniques such as opinion polling possible. Although there is always uncertainty in generalising from a sample to a population, the central limit theorem enables this uncertainty to be quantified (using tools such as confidence intervals, discussed later in this chapter). It therefore provides a way of assessing the strength of evidence for conclusions based on quantitative data. For example, an opinion survey conducted on a random sample of 1,000 people has a 'margin of error' of around 3% (techniques based on the central limit theorem allow this margin of error to be calculated). If the poll reveals that 59% of the sample support a particular statement, there is good evidence to suggest that there really is majority (more than 50%) support for that statement in the population as a whole. Although the actual

population figure will probably not be 59%, it is very likely that it will be in the range 56% to 62%.

4.3 Conversely, using tools such as confidence intervals (or margins of error) can help assess whether there is insufficient quantitative evidential support to assert a conclusion defensibly. An example of this occurred in the 2000 presidential elections in the USA, where the margins of error in the opinion polls were larger than the difference between the two main candidates. Consequently it was impossible to make a defensible evidence-based prediction as to who would win on the basis of the poll results.

4.4 Although it is not necessary to read the section dealing with the theory behind using the central limit theorem in §§4.20–4.46 to understand the following chapters of this guide, the material is included for reference to demonstrate how important concepts such as confidence intervals and p-values are derived. These concepts play a vital rôle in later chapters, and anyone undertaking performance audit or value for money work should have at least a basic understanding of them.

Some terminology

4.5 The term **population** is used to refer to the entire collection of units or individuals which is of interest in a performance audit or VFM study. For example, you might want to draw conclusions about populations such as 'Scottish secondary schools', 'NHS trusts', 'UK Benefits Agency offices', 'users of NHS Direct in England in 2002' or 'customers of electricity companies'. Many of these are very large, so you would often be reliant, at least in part, on **sample** data for your evidence. A **sample** is any subset of a population. There are many ways of choosing samples (see the NAO guides *Taking a*

Survey and *Sampling* for a fuller discussion), but the simplest way of obtaining a **representative** sample of a population is to take a **simple random sample**, in which each element of the population has an equal chance of inclusion in the sample.

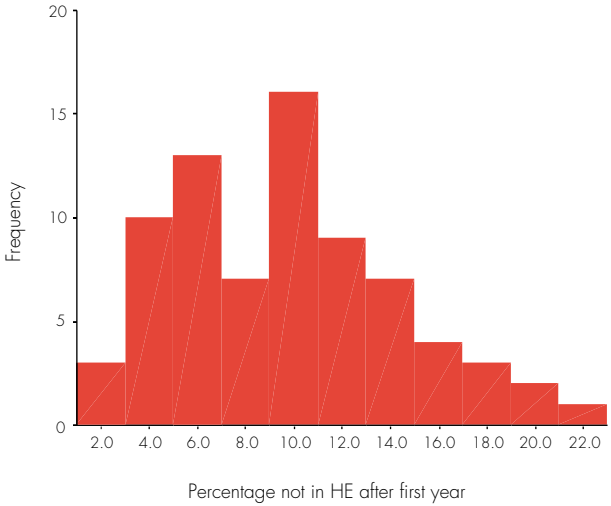
4.6 A **parameter** is a numeric summary measure or description for the population as a whole. For example, any of the measures described in Chapter 3, such as the median, the mean, the standard deviation or the proportion with a given attribute, could be parameters of interest for the purpose of addressing a particular audit question. You might want to know the proportion of electricity customers who have switched supplier in the last 12 months, or the average length of time a caller to NHS Direct has to wait before being dealt with.

4.7 A **statistic** is an estimate of a parameter, based on sample data. Whereas population parameters are usually the quantities of interest in a study, the information actually available is often about sample statistics. The process of statistical **inference** is a way of using these statistics to make statements about underlying population parameters.

Distributions

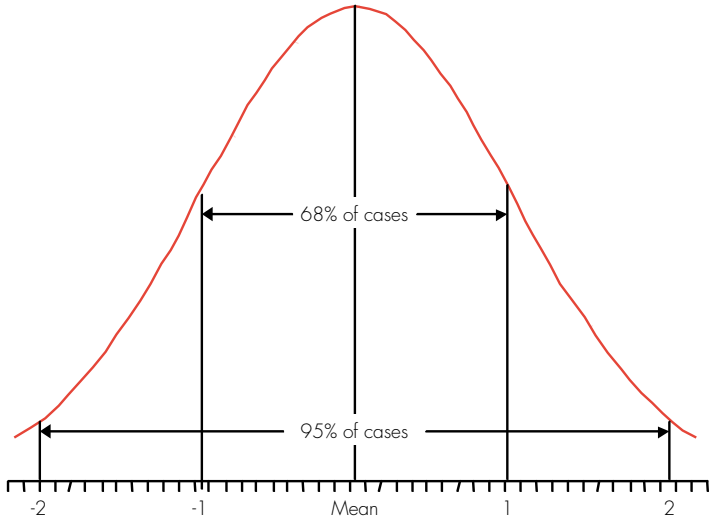
4.8 Chapter 3 emphasised the importance of identifying the response and explanatory variables that are relevant for each audit question. The **distribution** of each variable is a list of all the possible values it can take, together with their frequencies of occurrence. The **population distribution** shows the frequencies with which the different possible values of the variable occur for the whole population. The **sample distribution** gives the same information for a sample.

4.9 For example, a response variable in the case study in Chapter 3 was 'the proportion of English full-time first year undergraduates who dropped out of higher education in 1999-2000'. Information about this variable was available for a sample of 75 higher education institutions. The sample distribution of the variable was depicted in Chapter 3 by means of the histogram below.



4.10 The list of values and frequencies which makes up this distribution can be obtained from SPSS using the FREQUENCIES command.

4.11 There are a number of standard types of distribution that occur regularly in many applications of statistics. The most important of these is the **normal distribution**, the shape of which is depicted below. A continuous variable is said to be **normally distributed** if its distribution is of this form.



4.12 The following are properties of a normally distributed variable:

- the values of the variable are symmetrically distributed about their mean;
- 68% of all the cases of the variable occur within one standard deviation of the mean;
- 95% of all the cases of the variable occur within 1.96 standard deviations of the mean.

These properties are important for the discussion of how the central limit theorem can be used to derive confidence intervals in §4.23.

Sampling error

4.13 Given a sample of data, and faced with the problem of drawing a conclusion about a population parameter, your first thought would probably be to calculate the corresponding sample statistic, and use that as an estimate of the parameter. For example, suppose you were examining the uptake of a particular public service by the elderly. You might be interested in the average age of service users. If you were able to interview a sample of users to obtain their ages, you could work out the average age of all the users in the sample. Clearly, however, it is possible that the sample average might not be a very good estimate of the population average. It might be that the particular sample you interviewed happened to be, simply by chance, not very representative of the population as a whole.

4.14 The risk of selecting an unrepresentative sample is reduced by using an appropriate sampling methodology, such as simple random sampling. But even with such a methodology the estimate of the population average calculated in this way is clearly sample dependent. If a slightly different random sample of the same size had been drawn, the estimate would have been slightly different.

4.15 This sample dependence is called **sampling error**. In order to assess the strength of evidence for a conclusion based on sample data, we need to be able to evaluate how great the sampling error is likely to be.

4.16 The sampling error in a sample statistic reflects the *variability in the calculated values of the statistic, across all different possible samples of the same size taken from the given population*. In theory, it would be possible actually to draw each of these possible samples, and to calculate the desired statistic for each one. This would provide a new distribution, not of sample data, but of

different possible values of the sample statistic itself. Such a distribution is called the **sampling distribution of the statistic**.

4.17 The 'thought experiment' of constructing a sampling distribution provides the link between making a statement about a summary measure calculated on a single sample, and making a statement about the whole population. The summary measure can be thought of as falling somewhere in a sampling distribution. Statistical theory enables us to describe the properties of this distribution, and hence to draw conclusions about how likely it is that the population parameter is 'close to' the estimate given by the calculated statistic.

4.18 Recall from Chapter 3 that a way of measuring variability is by means of the standard deviation. The standard deviation of the sampling distribution of a statistic is called the standard error of the statistic. So the standard error provides a measure of sampling error. It quantifies the extent to which the statistic is likely to vary, depending on the particular sample that has been drawn. If the **standard error** of a statistic is small, it means that any other random sample of the same size as that actually drawn would be likely to give a similar result. So the statistic is likely to be a good estimate of the parameter. If the standard error is large, the statistic is not so likely to be a good estimate.

4.19 To summarise: We are often interested in population parameters (such as the average age of people using a particular public service), but only have sample data (such as the ages of 200 people using the service). We can calculate statistics based on the sample (e.g. the average age for the 200 people for whom we have data); but these are liable to a certain amount of error as estimates of overall population values. The extent of the error – the precision of our estimate – is quantified by the standard error of the statistic.

Most computer software packages will quote standard errors, at least for common statistics such as averages, which are accurate *provided that the sample is a simple random one*. If you have not used simple random sampling to gather your data the standard errors quoted by most packages will need adjusting, and you should consult a statistical expert.

Some underlying theory

(may be omitted without loss of continuity – go to §4.27)

4.20 The material in this section is slightly more theoretical, and is not required for understanding the rest of the chapter. It is included to demonstrate how statistical theory can be used to formalise the link mentioned in §4.17 between information from a sample and more general statements about the population as a whole.

4.21 Consider a variable x, whose mean value, on some population, is μ, and whose standard deviation, on that population, is σ. Suppose all possible random samples of size n are drawn from the population, and the sample mean of x is calculated for each. A result from statistical theory called the **central limit theorem** says that, provided n is reasonably large (in practice, more than about 30), these sample means are approximately normally distributed, with mean μ and standard deviation $\sigma / \sqrt{n}$. In other words, for random samples of more than about 30 cases, the sampling distribution of the mean is normal, with mean μ and standard deviation $\sigma / \sqrt{n}$.

4.22 Note that this result applies irrespective of the shape of the distribution of x. No matter how skew or non-normal the distribution of the variable x is on the population of interest, the *sampling distribution of the possible mean values of* x is normal.

4.23 This means we can use the properties of the normal distribution listed in §4.12, with reference to the *sampling distribution*. In particular, 95% of the possible mean values of x must fall within two standard errors of the mean μ of the sampling distribution. Expressed algebraically, this means that

$$-1.96\frac{\sigma}{\sqrt{n}} \le \mu - \bar{x} \le 1.96\frac{\sigma}{\sqrt{n}} \quad \text{95\% of the time,}$$

where $\bar{x}$ is the mean of x calculated for the particular sample of data we have collected.

4.24 Rearranging the inequality in §4.23 gives

$$\bar{x} - 1.96\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x} + 1.96\frac{\sigma}{\sqrt{n}} \quad \text{95\% of the time.}$$

Note that this now gives a *likely range of values for the (unknown) population mean* μ. It indicates that 95% of the time, the population mean will be between the two values listed (i.e. 95% of intervals constructed in this way, from different possible random samples from the population, will contain μ).

4.25 The only problem with the formula in §4.24 is that it contains the unknown population standard deviation σ. However, if n is large, a good estimate of the population standard deviation is just the sample standard deviation s calculated for the sample of data we have collected. Let $SE = s/\sqrt{n}$; then SE is an estimate of the standard error of the mean that only uses sample information, and we have

$$\bar{x} - 1.96SE \le \mu \le \bar{x} + 1.96SE \quad \text{95\% of the time.}$$

4.26 The range of values given in §4.25 is called a '95% confidence interval' for the mean. It is often interpreted as indicating that we can be '95% confident' that the true population mean is within 1.96 standard errors of the sample mean calculated for a single random sample from the population.

Confidence intervals

4.27 Standard errors for a number of common summary measures, such as means and proportions, can be calculated from statistical theory, and are available from most data analysis packages. A very important concept, often in practice derived from the standard error of a statistic, is the notion of a **confidence interval**.

4.28 A confidence interval for a population parameter is a *range of values within which the parameter is likely to fall*. Confidence intervals allow us to assess, on the basis of sample data, the strength of evidence for assertions about the wider population.

4.29 The way in which a confidence interval for a population parameter is calculated depends on the form of the sampling distribution for the corresponding sample statistic. For many common parameters, such as means and proportions, the sampling distribution is normal, provided that we take a random sample of more than about 30 cases. For such parameters, a **95% confidence interval** is calculated by taking the sample estimate, and adding and subtracting 1.96 times the standard error:

95% confidence interval = sample estimate ± (1.96 x standard error).

The discussion in §§4.20–4.26 gives some background on how this formula is derived.

4.30 What does a 95% confidence interval tell us? It indicates that there is a 95% chance that the population quantity we are interested in lies within the given range of values, in the following sense. If we were to take repeated random samples from the population, and construct such an interval for each individual sample, then 95% of these intervals would contain the population

quantity. Of course, we only have one such sample. But, because using the formula in §4.29 will give a range of values which contains the population quantity in 95% of all possible samples, we can say that we are '95% confident' that an interval constructed in this way from the sample we have collected will contain the population value.

Example: Calculating a confidence interval for a mean

In the NAO VFM report *Managing Sickness Absence in the Prison Service* (HC 372, 1998-99), the study team examined a random sample of records from pay databases and personnel data management systems to ascertain the number of working days lost to sickness absence for Prison Service staff. From their sample, they used a software package such as Excel to find that the average number of days lost for staff at female prisons over a year was 25.6, with a standard error of 1.9.

A 95% confidence interval is found, in accordance with the formula in §4.30, by adding and subtracting 1.96 times the standard error from the sample estimate of the mean. Thus:

Lower bound: 25.6 – (1.96 (1.9)) = 21.9;

Upper bound: 25.6 + (1.96 (1.9)) = 29.3.

So the study team was able to conclude that it could be 95% confident that the true average number of calendar days lost through sickness absence at all women's prisons in the year in question was between 21.9 and 29.3 days.

4.31 The formula in §4.29 shows how to calculate a 95% confidence interval, but the choice of 95% as a suitable 'level of confidence' is actually entirely arbitrary. Confidence intervals can be calculated for any level of confidence. The 95% level is the most common one, and has become conventional in most social science research. But other levels are sometimes used, and for parameters such as means, where the sampling distribution is normal, the

same approach can be used as in §4.29 to derive a confidence interval. All that needs changing is the number by which the standard error has to be multiplied. The table below gives the relevant multiples for some common levels of confidence. It demonstrates how there is a trade-off between increasing the certainty that the parameter lies within a certain range of values, and obtaining a range that is small enough to provide sufficiently precise information to be useful.

Confidence level	Interpretation	Multiple
90%	90% certainty that the interval contains the population parameter	1.64
95%	95% certainty that the interval contains the population parameter	1.96
99%	99% certainty that the interval contains the population parameter	2.33

For instance, in the example above on sickness absence, a 90% confidence interval for the mean number of days lost through absence is 25.6 ± (1.64 x 1.9), i.e., 22.5 to 28.7 days.

Significance tests and hypothesis testing

4.32 A confidence interval gives a range of values within which a parameter is likely to occur. There are often occasions when we want to assess whether a specific numeric value (rather than a range of values) for the parameter is likely. To do this we need to carry out a formal hypothesis test. The formula in §4.29, which applies in many situations where a population quantity is estimated from a random sample of data, can be rewritten as follows:

95% of the time,

$-(1.96 \times \textit{standard error}) \leq (\textit{sample estimate} - \textit{population value}) \leq (1.96 \times \textit{standard error})$

Another way of expressing this is to say that 95% of the time,

$$-1.96 \leq \frac{\textit{sample estimate} - \textit{population value}}{\textit{standard error}} \leq 1.96\,.$$

Suppose you wanted to test whether the population value were equal to some value of interest, say v. You could calculate the value of

$$t = \frac{(\textit{sample value}) - v}{\textit{standard error}},$$

which is obtained by inserting the hypothesised population value (v) in the middle part of the preceding inequality. If you then found that $-1.96 \leq t \leq 1.96$, you could conclude that the sample data are consistent with a population parameter value of v, because t has the sort of value that would be expected when the population value is v. In the terminology of statistical hypothesis testing, you would 'fail to reject the hypothesis that the population value is v'. But if you found that, say, t=−2.8, or t=3.4, you would reject the hypothesis that the population value is v because your sample has generated a value of t that would be very unusual (would only occur 5% of the time) if the hypothesis were true.

4.33 This demonstrates the logic of **statistical hypothesis testing**. We hypothesise a value for a population parameter, and determine values of a **test statistic** (in §4.32, the quantity t) that would cause us to reject the hypothesis-values that would only rarely (say 5% of the time) be realised if the hypothesis were true. We then calculate the value of the test statistic for our sample data, and either reject the hypothesis or not, depending on the value of the test statistic.

Since occasionally (5% of the time) extreme values of the test statistic will be realised even when the hypothesis is true, there is a 5% chance that we shall wrongly reject the hypothesis. The chance of doing so is known as the **significance level** or ***p*-value** of the test. As noted, hypotheses are usually tested at the 5% level of significance, but this is only a convention, and sometimes other levels such as 1% or 10% are used. The underlying idea is that, if you reject a hypothesis on the basis of a test with a low significance level, such as 5% or 1%, you are unlikely to be wrong. So an assertion based on such a test has a good evidence base.

4.34 The sorts of hypotheses that can be tested in this way are not restricted to simple statements about whether or not a population parameter is likely to have a certain value. A common scenario of interest is to test whether or not an observed difference between two groups, based on sample data, reflects an underlying population difference. Carrying out a hypothesis test can provide the evidence to support an appropriate conclusion. If the result of the test suggests that it is unlikely (at the given significance level) that the difference between the groups is zero, we say that there is a **statistically significant difference** between the groups.

4.35 Different sorts of hypotheses give rise to different test statistics. Some of the main types of statistical tests that are useful in performance audit and VFM work are set out, with examples, in Chapter 5. Although the tests have a variety of different names, and give rise to different sorts of output from common analysis packages such as SPSS, they all rely fundamentally on the logic outlined in this chapter.

Summary:
Key points from Chapter 4

- The values of population parameters are often estimated using sample statistics. The standard error of a statistic is a measure of its precision as an estimator. The standard error takes account of sampling error.
- A confidence interval for a parameter is a likely range of values for that parameter. Confidence intervals can often be calculated from standard errors. Conventionally, a 95% level of confidence is used. Confidence intervals can be used to assess the strength of evidence for a conclusion.
- Statistical hypothesis testing is a formal process that provides an indication of how consistent observed sample data are with a hypothesised statement. A test statistic is calculated to determine the likelihood of the hypothesis being true, given the observed data. If the value of the test statistic is sufficiently extreme, the hypothesis can be rejected.

CHAPTER 5

STATISTICAL TESTING

This chapter deals with the process of testing hypotheses, and explains how to carry out some common statistical tests. It identifies which tests are appropriate for different sorts of audit questions.

Chapter 5
Statistical testing

Basic principles of hypothesis testing

5.1 A basic outline of statistical hypothesis testing was given in §4.33. Essentially, it is a process of deciding, on the basis of sample data, whether or not to reject a hypothesis (called the **null hypothesis**) about a wider population. The decision is taken on the basis of how likely it is that the appropriate **test statistic** for the question under investigation would attain the value that it actually has attained – on the basis of the sample data – if the null hypothesis were true. Suppose this test statistic has a probability p of achieving a value as extreme as has been observed. If p is small enough (usually, less than 0.05), the conclusion is that the data are not consistent with the null hypothesis, and the hypothesis is rejected. The probability p is called the ***p*-value** or **significance level** of the test statistic. (The details of how *p*-values are calculated are beyond the scope of this guide. In practice, however, all statistical software packages will provide p values for all test statistics they produce.)

5.2 Since hypothesis testing is framed around the idea of checking whether there is sufficient evidence to *reject* a hypothesis, null hypotheses in social science and medical research are usually set up to be propositions that the researcher hopes to be able to reject. Thus, null hypotheses generally propose an *absence* of differences or effects, For example, a test of a new drug intended to relieve the symptoms of the common cold might be carried out by randomly

assigning cold sufferers to treatment either with the new drug or with a placebo, and recording the proportions in each group whose symptoms were alleviated after treatment. To arrive at an evidence-based conclusion about the efficacy of the drug, a medical researcher would set up as the null hypothesis the proposition 'there is no difference in the population proportion of patients whose symptoms are relieved, between the two groups'. If a statistical test showed that this hypothesis could be rejected, the researcher would conclude that there was, indeed, evidence that the drug had had an effect – that an observed difference between the groups was not just due to sampling error.

5.3 The situation in social science research and policy analysis is complicated by the fact that it is usually impossible to assign subjects randomly to 'control' and 'treatment' groups when, for instance, piloting a new programme. Generally, evaluators and auditors are faced with a situation in which a programme has been implemented (for example, a scheme to improve literacy skills of primary school pupils). Some performance measures for the programme will probably be available (such as the test results of pupils in the schools where the scheme was adopted). To assess how effective the programme has been, an evaluator needs to make some form of comparison; in other words, to investigate a difference between groups – e.g. between the test results for schools which did, and which did not, adopt the literacy scheme. (Sometimes the comparison may be between an observed outcome and a *counterfactual* alternative: an assessment of how the outcome, after implementation of the programme, differs from what *would have* been the case had the programme not been introduced.)

5.4 In the medical example of §5.2, evidence about the extent to which any observed difference is likely to reflect a real effect, and not just sampling error, can be assessed by setting up a formal hypothesis test. This is because any other possible causes of the difference have been *controlled for*, by

randomly assigning subjects to the different possible treatments. In the more general case where we do not have random assignment to groups, we need to have some other mechanism for controlling for other possible factors that could influence the size of the difference, before arriving at a conclusion about whether it is statistically significant. For instance, in the school example, as well as possibly reflecting some effect of the literacy programme, test results will probably be influenced by the academic ability of the pupils in the schools in each group, their levels of parental support, the quality of teaching, etc. A statistical method of controlling for these additional factors is to set up a **regression model**, in which the response variable is the difference in performance measures under investigation. One of the explanatory variables in the model would be a categorical variable that records, for each school, whether it was, or was not, included in the programme. The other explanatory variables would be measures of the other factors that we wish to control for. Such a model enables the effect of the literacy programme to be tested, controlling for other possible factors which might have influenced literacy outcomes. Regression models are discussed in Chapter 6 of this handbook. They are powerful tools for obtaining quantitative evidence in VFM studies or performance audits that focus on assessing effectiveness.

5.5 The important point to bear in mind when reading about, or using, the statistical testing techniques discussed in this chapter is that testing should not be a process that is carried out mechanically. It is very easy to feed a dataset into a statistical package and come up with a lot of test statistics and *p*-values. But you cannot conclude necessarily that, whenever a *p*-value is less than 0.05, you have demonstrated a substantive result. You need to consider whether, in analysing the comparison or difference that is being tested, it is necessary to control for other factors. If so, you may need to use the techniques discussed in Chapter 6. And even if the result is *statistically*

significant, you also need to consider whether it is *substantively* significant. **Do not confuse statistical significance and practical importance.**

5.6 With large sample sizes, even very small absolute differences between groups may turn out to be statistically significant. This does not necessarily mean that they are of practical importance. Suppose a social researcher discovered, on the basis of the results of a large survey, a statistically significant difference, between Edinburgh and London, in the proportion of adults who support the legalisation of cannabis She would not conclude from this that there was evidence of an important difference in attitudes between inhabitants of the two cities, if the actual proportions supporting legalisation were 73.8% and 71.9% respectively. Although the difference in proportions is statistically significant, it is not substantively important.

5.7 There is a difference here between performance audit and social research. Whereas social researchers are usually trying to find significant and substantively important effects, auditors may also be interested to discover evidence that an effect is likely to be small. If a programme has cost a great deal to implement, and analysis shows that the effect of the programme, as measured by a confidence interval for an appropriate comparative measure, is actually probably rather small, then the auditor is in a good position to question the value for money that has been achieved.

5.8 The rest of this chapter covers some common statistical tests. Although different tests are appropriate for addressing different types of questions, the basic process for all statistical tests is as follows.

1. Set up the null hypothesis.

2. Obtain the relevant data.

3. Select the relevant test procedure, and (use a software package to) obtain a *p*-value for the test statistic.

4. If the *p*-value is small enough (normally, less than 0.05), reject the null hypothesis. If the *p*-value is large (normally, greater than 0.05), do not reject the null hypothesis.

5.9 It must be emphasised that this prescription should only be followed as part of the broader study process. It relies on clear audit questions having been developed, and appropriate data having been collected. Using significance tests, and appropriate confidence intervals, to assess evidence for conclusions is only part of forming overall judgements. Nevertheless, they are an important and useful part, and a way of maximising the information provided by quantitative data.

5.10 The table overleaf summarises the tests that are discussed in this chapter, and indicates the types of question for which each is appropriate. The rest of the chapter provides reference information on how to carry out the different types of tests.

Which statistical test to use

	Type of question			
	Test on single group	**Difference between groups**	**Difference between two conditions on the same group**	**Relationship and association**
Continuous response variable	• 1-sample *t*-test (§5.12)	• 2-sample *t*-test (§5.21) • Mann-Whitney Test (§5.27) • ANOVA (§5.34) • Kruskal-Wallis Test (§5.43)	• Paired *t*-testing (§5.29) • Wilcoxon signed rank test (§5.33)	• Correlation coefficient (§5.51)
Categorical response variable	• Goodness of fit test (chi-squared) (§5.44) • Test of proportion (§5.17)	• Test of difference between proportions (§5.28)		• Chi-Squared test (§5.47)

Having decided which test is appropriate for your question, refer to the relevant section of this chapter for details.

Testing on a single group

5.11 Statistical testing on a single group is appropriate when you wish to assess whether a statistic calculated from sample data differs significantly from a given value. For example, a department might have a target of processing 70% of a certain type of request within three days of receipt. You might have a sample of 150 requests, of which 63% were processed within three days. You could use a test of proportions (see §§5.17–5.18) to assess whether this sample value is significantly different from 70%.

Single sample t*-test*

5.12 The single sample *t*-test is used to test whether an observed mean from a single sample differs from a given value.

5.13 The NAO study on sickness absence in the Prison Service, discussed in the example in Chapter 4, established, using sample data, that staff at female prisons have on average 25.6 days absence, with a standard error of 1.9. Suppose that the Prison Service's target for sickness absence is set at 19 days. We want to know whether the level of absence for staff at female prisons differs significantly from this, and so carry out a single sample *t*-test. The null hypothesis is that there is no difference between the average number of days of sickness absence, calculated on the basis of the sample, and the target parameter. In other words, the null hypothesis is that the difference between the sample statistic of 25.6 days, and the target parameter of 19 days, is caused simply by sampling error.

5.14 The SPSS output below shows the results of the test. The value of the test statistic, t, is 3.210. The column headed 'Sig. (2 tailed)' provides the p-value,

or significance level. The difference in means, together with a 95% confidence interval for the difference, is also provided.

One-Sample Test

Test Value = 19.0

	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval for the difference	
					Lower	Upper
sickness days	321	29	0.00162	6.60	2.39	9.89

5.15 The *p*-value in this case is 0.00162. This means that there is only a 0.16% chance of observing a test statistic of 3.21 or more, if the null hypothesis, of no difference between average staff sickness days and the target level, is true. Since the *p*-value is less than 0.05, the usual cut off for significance, we conclude that the data are inconsistent with the null hypothesis, and that the level of sickness absence amongst staff is higher than the target level.

5.16 If you want to carry out a single sample *t*-test by hand, you can use the following formula:

$$t = \frac{(\bar{x} - v)\sqrt{n}}{s},$$

where $\bar{x}$ is the sample mean, v is the value you are comparing it with, n is the sample size and s is the sample standard deviation. If the value of t is more than 2, or less than −2, and your sample size is more than about 30, you can conclude that the difference between $\bar{x}$ and v is significant at the 5% level.

Testing a proportion

5.17 A proportion can be thought of as a special case of a mean. Given a categorical variable, the proportion of cases in a particular category can be found by setting up a variable (called a **dummy variable**) that takes the value 1 for all cases in the dataset that are in the category of interest, and 0 for all other cases. The mean value of this variable is equal to the proportion of cases in the given category. You can then run a single sample *t*-test using the dummy variable to test whether the proportion differs significantly from a target value.

5.18 Alternatively, to carry out the test by hand, you can use the following formula, where p is the proportion calculated on a sample of size n, and v is the value you are comparing the proportion with:

$$t = \frac{(p - v)\sqrt{n}}{\sqrt{p(1-p)}}.$$

In order to use this formula, proportions must be expressed as decimal fractions (not as percentages), and you need a sample size of more than about 30. If the value of t that results is greater than 2, or less than −2, you can conclude that the difference between p and v is significant, at the 5% level.

Example: Testing a proportion

The study team investigating sickness absence in the Prison Service examined a sample of 50 recorded absences. They found that 86% of these had also been recorded on the organisation's pay database. The organisation had a target of recording at least 92% of all such absences on the pay database. The team wanted to know whether they had sufficient evidence to suggest that the target was not being met.

This can be tested using the formula in §5.18. For the given data, this gives

$$t = \frac{(p - v)\sqrt{n}}{\sqrt{p(1-p)}} = \frac{(0.86 - 0.92)\sqrt{50}}{\sqrt{0.86(1-0.86)}} = -1.22.$$

Since this is not less than −2, we conclude that the difference is not significant, at the 5% level. In other words, there is insufficient evidence to conclude from this sample that the organisation is not meeting its target. In order to draw a firm conclusion, the team would either have to use other evidence in addition to the sample data, or to increase the sample size.

Exact tests

5.19 The formulae in §5.16 and §5.18 should only be used for samples of more than 30 cases. If you have a smaller sample size, slightly different approaches, known as *exact tests*, are required. You should consult a statistical expert before carrying out significance testing on a small sample.

Testing differences between two groups

5.20 A frequent question of interest is whether there is evidence of a difference, on some (continuous) measure, between two distinct groups. This is the case when, in the terminology of Chapter 3, the response variable being investigated is continuous, and its variation conditional on a categorical

explanatory variable is being examined. If the explanatory variable has only two categories (such as 'male' and 'female', or 'before implementing program', 'after implementing programme'), the methods outlined in this section can be used.

t-test for independent groups (two-sample *t*-test)

5.21 The *t*-test for independent samples tests whether there is a difference between the means of two *independent* groups. The division of cases into two groups must be such that the cases in one group are not related to the cases in the other group.

5.22 To continue with the example on examining sickness absence in the Prison Service discussed earlier, a question of interest to the study team was whether there was a significant difference in levels of absence between staff at male and female prisons. This can be answered by carrying out an independent samples *t*-test.

5.23 The null hypothesis in this case is that there is no difference in number of sickness days between female and male prisons. Using SPSS to calculate the test statistic and significance level gives the following output.

Group statistics

sample	N	Mean	Std. Error (Mean)
female	30	25.5	1.913
male	40	20.99	0.480

Independent Samples Test

t	df	Sig. (2-tailed)	Mean Difference	St. Error Difference	t-test for Equality of Means 95% Confidence Interval of the Difference	
					Lower	Upper
2.3123	68	0.0119	4.56	1.972	0.694	8.43

5.24 The output shows that for the first sample there are 30 prisons, with a mean sickness absence rate of 25.55 days per year and a standard error of 1.91 days. The second sample has 40 prisons and a mean rate of 20.99 days with standard error of 0.48 days. The value of t, the test statistic is given in the second table, along with its p-value in the column headed 'Sig. (2-tailed)'. (The 'df' column provides the number of 'degrees of freedom' for the test, a concept which relates to the sampling distribution of the test statistic, but is not important for interpreting the output).

5.25 The output also shows the difference in sickness absence rates between the two samples (in the 'Mean Difference' column). The standard error of this difference and a 95% confidence interval for it are also given.

5.26 The p-value, or significance level, in this example is 0.012. Therefore, there is a 1.2% chance of gaining a test statistic of the magnitude calculated if the null hypothesis of no difference is true. This is strong evidence against the null hypothesis, suggesting that sickness absence rates do differ between male and female prisons.

Wilcoxon or Mann-Whitney test

5.27 A more general test for comparing two groups is the Wilcoxon rank-sum test (or an equivalent alternative known as the Mann-Whitney test). This evaluates

the null hypothesis that the two groups were sampled from identical populations. This is broader than the null hypothesis tested by the independent samples t-test, which deals specifically with means. If the two populations are assumed to have the same shape and dispersion, the Wilcoxon test will actually test for differences in their medians. This test is available in SPSS, under the ANALYSE>NON PARAMETRIC TESTS>2 INDEPENDENT SAMPLES menu option.

Testing for differences between two proportions

5.28 As mentioned in §5.17, statements about proportions can be converted into statements about means by setting up appropriate dummy variables in the dataset. While it is possible to derive formulae specifically for testing the significance of the difference between two proportions, in practice it is often easiest to create the relevant dummy variables, and then use a *t*-test to compare their means. You can use the =IF formula in an Excel spreadsheet, or the TRANSFORM>RECODE menu option in SPSS, to create dummy variables quickly.

Paired sample *t*-test

5.29 The paired sample *t*-test is a way of examining differences between two samples of data that are related or paired (for instance, if the data consist of pairs of measurements on the same subjects). It works by looking at the mean of the difference between the paired observations.

5.30 To revisit the example of investigating rates of sickness absence once again, suppose you wanted to assess whether there had been a significant change in absentee rates between 1995 and 2000. You might have the following data:

Prison	1995 Absence rate	2000 Absence rate	Difference
1	20.3	25.2	–4.9
2	29.2	25.7	3.5
3	28	27.4	0.6
4	22.5	19.3	3.2
5	19.7	25.5	–5.8
6	24	29.1	–5.1
7	26.4	22.7	3.7
8	23.7	27.2	–3.5
9	26.1	25	1.1
10	27.1	26.1	1

In this case the two sets of absence data, for 1995 and 2000, are not independent, as the prisons are the same in both years. The null hypothesis is that the mean difference is zero (that is, that there is no difference between absence rates in 1995 and 2000).

5.31 The output that results from requesting a paired samples *t*-test in SPSS for these data is as shown below.

Paired Samples Test

	Mean difference	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference		t	df	sig
				Lower	Upper			
1995–2000	–0.62	3.81	1.21	–3.35	2.11	–0.514	9	0.619

5.32 This output shows that the mean difference in absentee rates between the two years is –0.62 days. A 95% confidence interval for this difference is –3.35 to 2.11. The significance level (*p*-value) of the test statistic is 0.619. As this is considerably greater than 0.05, we cannot reject the null hypothesis that the mean difference is zero. You would conclude that there is insufficient evidence, on the basis of these data, to suggest that absentee rates have changed between 1995 and 2000.

Wilcoxon signed rank test

5.33 The Wilcoxon signed rank test is an alternative to the paired samples *t*-test that tests the broader null hypothesis that the paired samples were drawn from either identical populations, or from symmetric populations with the same mean. It is available in SPSS under the ANALYSE>NON PARAMETRIC TESTS>2 RELATED SAMPLES menu option

Testing differences between more than two groups

Analysis of variance

5.34 **Analysis of variance** (often abbreviated to **ANOVA**) is a way of testing if two or more groups have different means. It is method that can be used when examining the variation of a continuous response variable conditional on a categorical explanatory variable. Whereas methods such as the *t*-test address this question in the special case of a dichotomous explanatory variable, ANOVA can be used when the explanatory variable has more than two categories. (ANOVA gives equivalent results to the *t*-test in the two-category case). It can also be used when there is more than one categorical explanatory variable. It is often used in medical research, in analysing the results of drug trials.

5.35 ANOVA tests the null hypothesis that the mean value of the response variable is the same for all the groups. Although it is a test for differences in means, it actually works by comparing *variances* (see §3.30). This is because, in order to conclude with confidence that two groups are different from one another with respect to some characteristic, it must be established that the differences *between* them must substantially exceed the differences *within* them, with respect to that characteristic. ANOVA partitions the overall variability in the response variable into variation within and between groups, and uses this breakdown to help test whether there is a significant difference between group means. The test statistic used to do this is called the *F* statistic, and is provided by all statistical analysis software. If the *F* value resulting from an ANOVA test yields a sufficiently small *p*-value (this is also provided in all computer output), the null hypothesis of no difference between the groups can be rejected.

5.36 ANOVA can be extended in many ways, and you may wish to seek expert advice before carrying out this type of testing. The simple example that follows illustrates a few of the basic principles, and demonstrates some typical SPSS output.

A simple example

5.37 Suppose we have a sample of mathematics test results for five pupils at three different schools, and we want to examine whether there is evidence of a difference in results between schools. The data are as below:

	School A	School B	School C
	79	65	70
	79	75	80
	65	59	65
	74	76	89
	60	65	70
Mean	**71.4**	**68.0**	**74.8**
Variance	**73.3**	**53.0**	**92.7**
Standard deviation	**8.6**	**7.3**	**9.6**

5.38 Using the COMPARE MEANS option in SPSS to obtain an ANOVA on these data gives the output shown below.

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	115.6	2	57.8	0.791781	0.475341
Within Groups	876	12	73		
Total	991.6	14			

5.39 The 'sum of squares' and 'mean squares' column give information on how the total variation in examination results is partitioned between and within schools. The test statistic (*F*) and its significance level are used to decide whether or not to reject the null hypothesis of 'no difference between schools'. In this case, the significance level (0.48) is considerably higher than 0.05, so there is insufficient evidence to suggest a difference between the means of the groups.

5.40 This example shows what happens when there is only one explanatory variable – school, in this case. It is possible to include more than one explanatory factor in an ANOVA. For instance, we might have information on each pupil's gender as well as test results, as shown below.

	School A	School B	School C
	79 (m)	65 (f)	70 (m)
	79 (m)	75 (m)	80 (m)
	65 (f)	59 (f)	65 (f)
	74 (f)	76 (m)	89 (m)
	60 (f)	65 (f)	70 (f)
Average (male)	**79**	**75.5**	**79.7**
Average (female)	**66.3**	**63.0**	**67.5**
Overall average (male)	**78.1**		
Overall average (female)	**65.6**		

5.41 Running an ANOVA on this dataset provides separate results on the way in which the overall variation in scores can be divided up into variation due to gender and variation due to school. The SPSS output is as shown below.

	Sum of Squares	df	Mean Square	F	Sig.
Between group (gender)	234.39	1	234.39	3.501	0.0882
Between group (school)	20.92	2	10.46	0.16	0.854
Within Groups	736.29	11	66.94		
Total	991.6	14			

5.42 The output shows test statistics (*F* values) for the effects of gender and school separately. The significance level of the test statistic for gender is 0.088, which is quite small (although if we were testing at a fixed level of 0.05 we would still not reject the null hypothesis of 'no difference by gender'). The significance level for school differences is much higher (0.85), indicating that the data do not provide evidence to suggest a difference in schools, once gender has been allowed for.

Kruskal-Wallis test

5.43 The Kruskall-Wallis test is a generalisation of the Wilcoxon rank-sum test discussed in §5.27 to the case of three or more independent groups. It is analogous to analysis of variance, but tests the broader hypothesis that all samples were drawn from identical populations. It is particularly sensitive to differences in medians. It cannot be extended to examine the effects of more than one explanatory variable. It can be obtained in SPSS under the ANALYSE>NON PARAMETRIC TESTS>K INDEPENDENT SAMPLES menu option.

Testing goodness of fit

5.44 Sometimes a question that arises in performance audit work is whether the distribution of a categorical variable differs from an expected or target breakdown into categories. For example, you might want to examine whether the allocation of staff to different aspects of a department's business is in accordance with a target profile. The **chi-squared** test is a way of addressing this question. It assesses the *goodness of fit* of an observed sample distribution to a given, 'expected' distribution. The term chi-squared comes from the traditional symbol for the test statistic used, the Greek letter χ (chi).

5.45 The chi-squared test for goodness of fit is available in SPSS under ANALYSE>NON PARAMETRIC TESTS>CHI-SQUARE. It is also quite easy to work out the chi-squared test statistic by hand, if you only have a small number of categories to consider. For each category, work out the squared difference between the observed (*O*) and expected (*E*) number of cases. Divide this by the expected number of cases for that category. Do this for each category, and sum the results. In symbols,

$$\chi^2 = \sum \frac{(O-E)^2}{E^2}$$

5.46 To calculate the p-value associated with a chi-squared statistic, you can use the Excel CHIDIST function. Typing =CHIDIST(*c*,*d*) into any cell on a worksheet, where c is the value of the test statistic calculated as in §5.45, and d is one less than the number of categories, will return the appropriate p-value.

Example: Using the chi-squared test to assess goodness of fit

120 staff with salaries in the range £35,000 to £45,000 were asked to indicate on a five-point scale what their salary was. The results were as shown in the table below.

Salary	£35000–£36999	£37000–£38999	£39000–£40999	£41000 £42999	£43000–£45000
Expected no. of	31	15	21	20	33

The expectation was that equal numbers of staff would be in each category. This means that, for this sample, the expected number in each category would be 24. A chi-squared test can be used to test this null hypothesis. Using the formula in §5.45 with E=24 for each category, and the values of O for each category as given in the table yields.

$$\chi^2 = \frac{(31-24)^2}{24} + \frac{(15-24)^2}{24} + \frac{(21-24)^2}{24} + \frac{(20-24)^2}{24} + \frac{(33-24)^2}{24} = 9.83.$$

Entering =CHIDIST(9.83,4) into a cell on an Excel spreadsheet gives a *p*-value of 0.04. Thus, there only a 4% chance of obtaining such a large χ^2 value as observed in the sample if the difference between the observed and expected numbers of staff in each category was simply due to sampling error. So there is good evidence to suggest that the population breakdown of salaries is not uniform between categories.

Testing for association

Using the chi-squared test

5.47 The chi-squared test can also be used to test whether there is an association between two categorical variables. When investigating whether two such variables are likely to be related, the first step is to produce a **contingency table**, or crosstabulation, showing the number of cases in the sample in each combination of categories of the variables. For example, the table below was produced (using SPSS) from a dataset of 582 responses to a questionnaire asking users of a particular government service to categorise their satisfaction with the service. A question of interest was whether satisfaction levels were related to users' ages. Information on age was collected as a categorical variable, as shown in the table. The table is a contingency table showing how satisfaction levels are broken down by age.

Age Category		Service satisfaction: Strongly Negative	Somewhat Negative	Neutral	Somewhat Positive	Strongly Positive	Total
18–24	Count	12	5	14	11	4	46
	% within Age category	26.1	10.9	30.4	23.9	8.7	100
25–34	Count	20	21	33	20	33	127
	% within Age category	15.7	16.5	26.0	15.7	26.0	100
35–49	Count	37	43	50	48	52	230
	% within Age category	16.1	18.7	21.7	20.9	22.6	100
50–64	Count	20	31	51	27	18	147
	% within Age category	13.6	21.1	34.7	18.4	12.2	100
64+	Count	4	5	9	6	8	32
	% within Age category	12.5	15.6	28.1	18.8	25.0	100
	Count	93	105	157	112	115	582
	% within Age category	16.0	18.0	27.0	19.2	19.8	100

5.48 If there were no relationship between the two variables, the number of cases in each cell of the table would be completely determined by the row and column totals for the table. It is a simple exercise to show that, if the variables in such a table are independent of each other, then the expected number of cases that would occur in each cell is

$$\frac{\text{row total x column total}}{\text{overall total}}.$$

5.49 So, a chi-squared test can be used to compare the observed numbers of cases in each cell of the table with the number that would be expected

(calculated using the formula in §5.48) if there was no relationship between the variables. This can be done by hand, as described in §5.45, or using a software package such as SPSS. The only difference in doing the calculation is in obtaining the *p*-value of the test statistic. It can still be obtained using the CHIDIST function in Excel, but in the form =CHIDIST(s, (*r*-1)*(*c*-1)), where *r* is the number of rows in the table, and c is the number of columns in the table. (The product (*r*-1)(*c*-1) is known as the number of degrees of freedom of the statistic.)

5.50 Using SPSS to produce a chi-squared analysis for the table in §5.47 gives a chi-squared value of 24.6, with 16 degrees of freedom. The *p*-value for this is 0.078. Thus, using a 5% significance level, we do not reject the null hypothesis of no association between satisfaction level and age. Although a glance at the figures suggests that there appears to be some relation between satisfaction and age, with younger service users being more dissatisfied, in fact the difference between younger and older users is not statistically significant.

Correlation

5.51 A measure of the (linear) relationship between two continuous variables is given by the **correlation coefficient** between them, already discussed in §§3.40–3.43. Software packages will return significance levels for correlation coefficients, but they are not usually very useful. They can be used to test whether the correlation between two variables is significantly different from zero; but this is usually not of great interest. The size of the correlation coefficient is more important as a practical indication of the strength of the relationship. A way of quantifying the size of the effect of one continuous variable on another is by means of an appropriate regression coefficient. This is discussed in Chapter 6.

Summary:
Key points from Chapter 5

- Statistical tests can help to assess the strength of evidence for audit conclusions. All tests involve calculating a test statistic and a significance level for the test statistic. The significance level shows how likely it would be to obtain a test statistic with a value as extreme as that observed, if the null hypothesis were true. If the significance level is small, normally less than 0.05, the null hypothesis can be rejected.
- Different test statistics are appropriate for different types of questions. The table in §5.10 shows which tests can be used to address which types of questions. Examples of how to interpret the results of these tests are given throughout the chapter.
- It is important not to confuse statistical significance with practical importance. Effects can be statistically significant, but substantively small. Calculating confidence intervals for quantities of interest, such as differences in mean values between groups, can help to establish likely ranges of variation and hence inform evidence-based judgements of their substantive significance.
- If you are uncertain about how to carry out or interpret any of the tests detailed in this chapter, consult your technical advisory team.

CHAPTER 6

RELATIONSHIPS IN DATA

This chapter introduces simple and multiple linear regression models, and shows how to interpret the results of such models through examples. It also includes a discussion of some extensions to the basic regression model, and lists some other relevant techniques, such as factor and cluster analysis.

Chapter 6
Relationships in data

Introduction

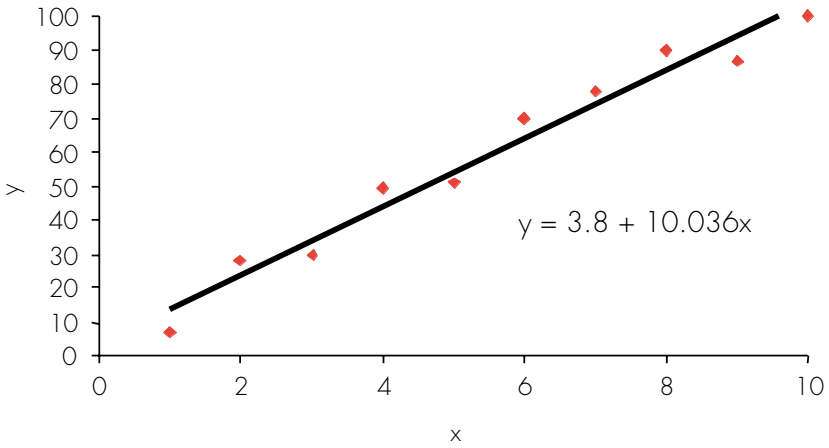
6.1 Ways of testing for associations between variables were discussed in Chapter 5. Often, we want to go beyond merely establishing whether two or more variables are related to looking more closely at the form of their relationship. For example, we might want to make predictions of plausible results in a particular context, or to see how changes in one aspect of a system could affect the overall picture. Questions of this sort can be addressed using forms of **regression analysis**.

6.2 Regression analysis can be a powerful analytical technique for VFM and performance audit. It predicts a single response variable from one or more explanatory variables. This chapter explains what regression is, how to carry out regression on a computer, and what to look out for when using it.

Simple linear Regression

6.3 Simple linear regression is the most basic form of regression analysis. It assumes a straight line relationship between two continuous variables and uses the data to produce a **line of best fit** for this relationship.

6.4 If the response variable is denoted by y and the explanatory variable is denoted by x, then a regression analysis of y on x provides the equation of the line of best fit, in the form $y = a + bx$. Here a is the intercept (the point at which the line crosses the y-axis), and b measures the slope of the line. The values of a and b provided by fitting a simple linear regression model to the data are called the **coefficients** of the model. The model allows you to predict values of y, for given values of x.



How regression coefficients are calculated

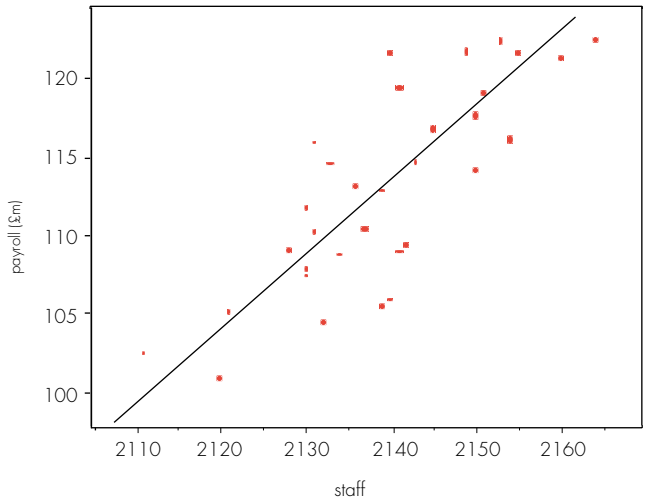
6.5 The regression line is calculated using a method called the *least squares principle*. Unless the response variable and the explanatory variable are perfectly correlated, a straight line drawn on a scatterplot of their values will not go through every point in the scatter. The difference, at each value of the explanatory variable x, between the value of the response y that is observed, and the value that is estimated by the line, is known as the **residual** at that point. The least squares principle chooses the line of best fit to the data by

minimising the sum of the squares of the residuals. It can be shown that the values of a and b required to do this are given by

$$b = \sum \frac{(x_i - \bar{x})(y_i - \bar{y})}{(x_i - \bar{x})^2},$$

$$a = \bar{y} - b\bar{x},$$

where the data being analysed consists of pairs (x_i, y_i) of values of the explanatory and response variables, and $\bar{x}$ and $\bar{y}$ are the mean values of x and y respectively.



ANOVA in Regression Analysis

6.6 Analysis of variance (ANOVA) was introduced in §§5.34–5.36. The ANOVA technique works by dividing the variation in the response variable into variation due to differences within groups and variation due to differences between groups. A similar process can be performed for regression analysis, whereby the total variation is broken into variation that is explained, or predicted, by the regression model, and the remaining variation in the data. An ANOVA table is produced in the SPSS output for a regression analysis. As in the examples in Chapter 5, this includes a test statistic, called the F statistic, that can be used to test a null hypothesis. In this case, the null hypothesis is that there is no relationship between the response and explanatory variables. A significant F statistic means that this hypothesis can be rejected, and that there is evidence to support the assertion that a definite relationship exists.

Prediction and Goodness of Fit

6.7 A regression equation can be used to predict a value for the response variable y from given values of the explanatory variables. A way of assessing the **goodness of fit** of the regression model is to examine how the predicted values of y compare with the actual values observed, for those cases where we have data. One way of summarising the strength of the relationship between the predicted and actual values is by means of the correlation coefficient between them. In regression analysis this is denoted by R. As mentioned in §3.42, it is generally instructive to calculate the square of a correlation coefficient, as this can be directly interpreted in terms of the proportion accountable variation. Thus, the R^2 (*R-squared*) measure, or **coefficient of determination**, which is reported in every regression output, represents the proportion of the variation in the response variable that is

accounted for by the explanatory variables. It can be calculated from the ANOVA table as the ratio of the 'regression sum of squares' to the 'total sum of squares'. When the regression accounts for all the variation $R^2 = 1.0$, indicating a perfect linear fit. When $R^2 = 0$ there is no linear relationship between response and explanatory variables. Usually, R^2 falls between these two values. An R^2 of 0.8, say, means that the explanatory variables account for 80% of the variation in the response variable.

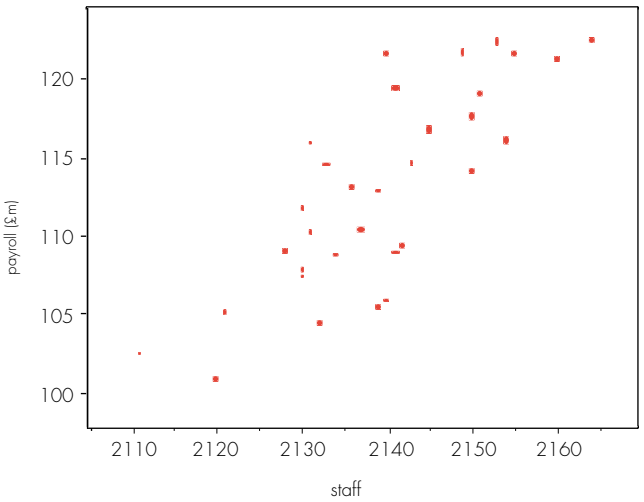
How to carry out simple linear regression

1. Identify response and explanatory variables.
2. Draw scatter diagrams to get an overview of the relationship. Does it appear linear? If not, it may still be possible to use regression analysis, but you may need to use a *transformation of variables* (see §6.22)
3. Use a package such as SPSS to run a regression of the response variable on the explanatory variable. In SPSS, go to ANALYSE > REGRESSION > LINEAR. Put the response variable in the box marked 'dependent variable', and the explanatory variable in the box marked 'independent variable', and press OK. This will produce the basic output.
4. Interpret the output as shown below.

Example: Interpreting SPSS output for simple linear regression

A study team needed some information on the expected payroll sizes for organisations of a certain type in a developing country, for benchmarking purposes. A sample of data on the size of payrolls for a number of such organisations was available. The team used simple linear regression on this sample to predict expected payroll size, given the number of staff in the organisations of interest.

In this case the response variable is payroll, and the explanatory variable is number of staff. A simple scatter graph of the data is shown below. It shows that the relationship between staff numbers and payroll size does appear approximately linear.



SPSS was used to run a linear regression, and produced the following output.

The **Model Summary** table provides the value of R^2 for the regression. In this case, $R^2=0.64$, meaning that staff numbers account for 64% of the variation in payroll figures.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.800	.640	.628	.38876

a Predictors: (Constant), STAFF

The **ANOVA** table provides an F statistic which tests the null hypothesis that there is no relationship between response and explanatory variables. Here, the significance level quoted for the F value is '.000', which in SPSS output means 'less than 0.0005'. It is certainly much less than 0.05, so there is a significant relationship in this case.

ANOVA

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	8.049	1	8.049	53.256	.000
	Residual	4.534	30	.151		
	Total	12.583	31			

a Predictors: (Constant), STAFF
b Dependent Variable: payroll figures (£m)

Finally, a table of the regression **coefficients** is provided. The 'unstandardized coefficients' values provide the a and b values in the regression equation $y=a+bx$.

Coefficients

Model		Unstandardized Coefficients		Unstandardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-80.424	12.568		-6.399	.000
	STAFF	4.287E-02	.006	.800	7.298	.000

a Dependent Variable: payroll figures (£m)

The value labelled '(Constant)' is the intercept (a) value, and the value labelled 'STAFF' is the coefficient for the STAFF variable (i.e., the ***b*** value).

So in this case the regression equation is

PAYROLL = -80.424+0.04287 (STAFF)

The study team wanted a 'ballpark' estimate of the expected payroll size for an organisation similar to those in the sample with 2,125 members of staff. This can be obtained from the equation, by setting STAFF=2,125:

PAYROLL = -80.424 + 0.04287 (2125) = £10.7m.

Thus, for this organisation, the predicted payroll is £10.67m.

Multiple Regression

6.8 **Multiple linear regression** extends the model of simple linear regression with one response and one explanatory variable to consider the effects of several explanatory variables simultaneously on a response. Multiple regression models can be used to assess the extent to which the response is influenced

by an explanatory variable, having allowed for the effects of other explanatory factors.

6.9 In a multiple regression analysis with k explanatory variables, coefficients b_0, b_1, b_2, ..., b_k are estimated using the principle of least squares to obtain an equation of best fit for the response variable y in the form

$$y = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_k x_k .$$

It is harder to interpret such an equation graphically than is the case for simple linear regression. But the regression coefficients can be interpreted as follows. The intercept, or constant, term b_0 represents a kind of 'baseline' value for y, around which it varies due to the effects of the explanatory variables x_i. The coefficient associated with each explanatory variable represents the average marginal change in y associated with a unit change in that explanatory variable. In other words, if the variable x_1 is increased by one unit, and all other variables are held constant, then the regression model predicts that y would change by b_1 units, on average.

6.10 Thus, regression coefficients are measures of the **sizes of the effects** of the explanatory variables on the response variable. A large regression coefficient indicates that a small change in the explanatory variable has on average a large impact on the response variable, other things being equal.

How to carry out Multiple Regression

1. Identify the response variable and the set of explanatory variables. You should decide on appropriate explanatory variables on the basis of the question under consideration. (It is generally not good practice to 'try out' different variables in the hope of improving the R^2 value of the model. The model should derive clearly from the substantive question you are investigating.)
2. Use a software package to run a multiple linear regression of the response variable on the explanatory variables. In SPSS, use the same commands as for simple linear regression, but include all the explanatory variables in the 'independent variables' box.
3. Interpret the output as shown below.

Note that SPSS includes options for how the explanatory variables are 'entered' into the model. The default setting is simply 'Enter', which runs the model as specified. Other settings such as 'Forward' and 'Backward' are sometimes used for exploratory regression analysis. They work by automatically selecting variables in such a way as to maximise R^2 values. You should consult the statistics team before attempting to use such settings. As noted above, regression analysis should normally be driven by substantive hypotheses and questions of interest, not by attempts to obtain a 'good' value of a particular summary statistic.

Example: Interpreting SPSS output for multiple regression

The NAO study team working on the report on *Improving Student Achievement in Higher Education* (HC 486, 2001–02), discussed in the case study in Chapter 3, were interested in the question of whether the level of mature age enrolment at higher education institutions had an effect on dropout rate. Exploratory data analysis had shown that institutions with higher proportions of mature age students tended to have higher dropout rates; but such institutions also tended to have lower average entry scores, and it was known that entry scores are a good predictor of dropout. The team therefore wanted to examine the effect of the proportion of mature age students on dropout rate, controlling for average entry score. This can be done by running a regression model in SPSS. The output is shown here.

The **model summary** table gives an R^2 value for the model. For models with several explanatory variables, it is technically better to use the 'Adjusted R-square' value for this (for technical reasons to do with the way the R^2 values are calculated). In this case, the adjusted R^2 value is 0.79, indicating that the model explains 79% of the variation in the response variable.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.894	.799	.793	2.057

a Predictors: (Constant), MATURE, ENTRY
b Dependent Variable: Percentage not in HE after first year

The **ANOVA** table is interpreted as for simple linear regression. Here the F statistic is highly significant (the notation '.000' meaning that the significance level is less than 0.0005), so it is likely that there are real effects of the explanatory variables on the response.

ANOVA

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1174.702	2	587.351	138.819	.000
	Residual	296.174	70	4.231		
	Total	1470.877	72			

a Predictors: (Constant), MATURE, ENTRY
b Dependent Variable: Percentage not in HE after first year

The regression coefficients are given in the table below. The 'unstandardized coefficients' quantify the effects of entry score and proportion of mature age students on dropout rates.

Coefficients

		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
Model		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	15.632	1.623		9.629	.000	12.394	18.869
	ENTRY	-.517	.061	-.603	-8.423	.000	-.639	-.394
	MATURE	.129	.025	.372	5.202	.000	.080	.179

a Dependent Variable: Percentage not in HE after first year

The analysis shows that there is a negative effect of entry score (the higher the average entry, the lower the dropout rate tends to be). The coefficient of -0.5 indicates that, on average, each increase of a point in average entry score is associated with a decline of about 0.5% in dropout rate. On the other hand, an increase of 1% in the percentage of mature age students enrolled is associated with an increase of about 0.1% in dropout; or in other words, each increase of 10% in mature age enrolment tends corresponds to a 1% increase in dropout rates, on average.

The significance levels for the regression coefficients are given, and in this example all are less than 0.05. This suggests that, even after controlling for entry score, there is a statistically significant effect of mature age enrolment on dropout. However, the size of the effect is not large. Institutions' mature age enrolments would have to differ by 10% to 20% for their dropout rates to differ noticeably.

Using regression analysis in benchmarking and comparing performance

6.11 Benchmarking is a commonly used technique in VFM and performance audit work. It relies fundamentally on making comparisons, and in order for these to be valid, they must be made on a like-with-like basis. Sometimes it is difficult to assess the extent to which different contexts might have affected the outcomes of interest in a benchmarking exercise. In such instances it may be possible to use a regression model to provide predicted or expected values of key performance measures for an organisation of interest that are based on data from comparable bodies or organisations.

6.12 Similarly, in studies of variations in performance, regression models can be used to produce 'context adjusted' or 'value-added' performance measures. These enable more valid comparisons of performance to be made than is possible on the basis of crude 'league tables' of indicators. For example, hospital mortality rates are obviously affected by the mix of cases presented by patients, and this should be taken account of in any comparison of mortality statistics between hospitals. Research shows that school examination results are probably more influenced by levels of pupil ability and relative social deprivation than by teaching quality. In order to make valid statements about comparative performance in such instances, therefore, it is necessary to make allowance for factors outside the control of the bodies being

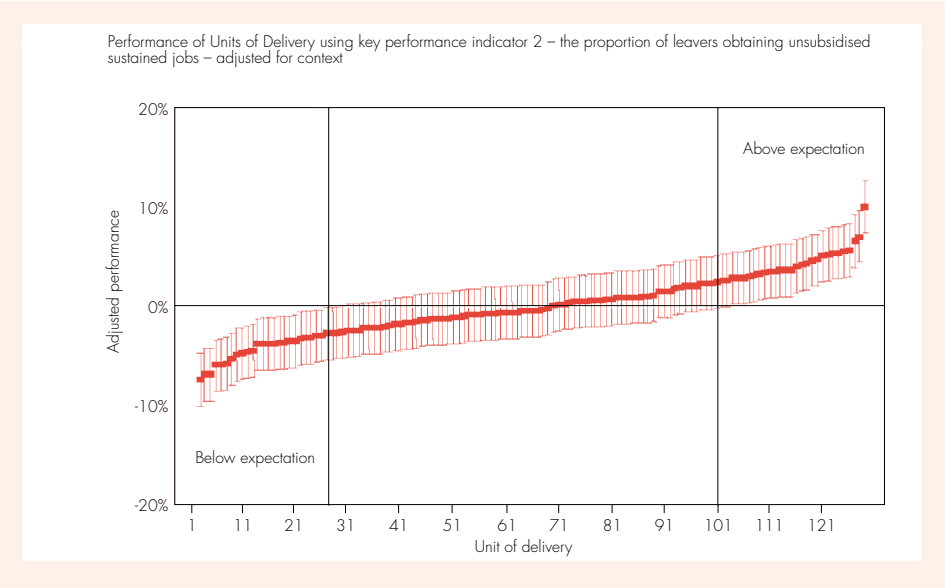
investigated before comparing their performance. Regression techniques can be used to do this.

Example: Using regression analysis to make context-adjusted comparisons

Part of the NAO study on *The New Deal for Young People* (HC 639, 2001-02) was an examination of the comparative performance of the regional offices (the so-called 'Units of Delivery' or 'UoDs') through which this government programme is delivered. The aim of the programme is to get more long term unemployed young people into jobs. Its effectiveness is assessed in part by means of 'key performance indicators' which record, for each Unit of Delivery, the proportions of programme participants who go on to get jobs.

The Units of Delivery are distributed across the country, under different local labour market conditions and with differing intakes of participants. In order to compare performance across UoDs, the study team used data on labour market conditions and characteristics of participants such as their education level and length of time unemployed to construct regression models for the key performance indicators. Comparing expected and actual indicators at each office provided an estimate of context-adjusted performance for each UoD. Those offices doing better than expected have positive values, and those performing worse than expected have negative values. The results for one indicator are shown in the graph below.

The graph shows adjusted performance measures for each UoD. Confidence intervals have been added to enable assessments of statistical significance. Those UoDs for which the intervals do not overlap zero (the horizontal line on the graph) are performing significantly better, or significantly worse, than expected.



Regression with dummy variables: analysis of covariance

6.13 Explanatory variables in regression models can be categorical as well as continuous. The effect of a continuous explanatory variable is quantified by its regression coefficient, as explained in §6.8. Categorical variables have to be converted into **dummy variables** (see §5.17) before they can produce interpretable coefficients.

6.14 Dummy variables take the value 1 to indicate the presence of an attribute and 0 to indicate its absence. Suppose you wanted to include a variable indicating 'region' as an explanatory factor in a regression analysis. It might be coded

North	1
East	2
West	3
South	4

6.15 Instead of representing this variable by a single column in the dataset, containing four possible values, you would need to set up four dummy variables, called, say '*N*', '*E*', '*W*' and '*S*'. The first would take the value 1 whenever region=1, the second would take the value 1 whenever region=2, etc. A portion of the expanded dataset including the dummy variables would look like this:

ID	region	N	E	W	S	...
1	3	0	0	1	0	...
2	4	0	0	0	1	...
3	1	1	0	0	0	...
4	1	1	0	0	0	...
...	...	...	...	...	...	...

6.16 Instead of using the variable 'region' in the regression equation, you would use the individual dummy variables. The regression coefficient associated with the '*N*' variable would then measure the effect of being in the North, the coefficient for the '*E*' variable would measure the effect of being in the East, and so on. In other words, effects are measured *for each category of a categorical variable separately*, rather than for the variable as a whole.

6.17 Because information about all but one of the dummy variables determines the value of the last one, one less dummy than the number of categories must be used to represent a categorical variable in a regression equation. In the example above, given the values of *N*, *E* and *W* for any case in the dataset, the value of *S* for that case is determined (since precisely one of the four dummies must take the value 1 for each case). It is an underlying assumption of regression analysis that no two explanatory variables are perfectly associated, so it would be impossible to enter all four variables into the regression equation. However, in order to get information about all the categories of the 'region' variable, it is sufficient to enter only three. This can be seen as follows.

6.18 Suppose the results of the regression indicated that the equation could be estimated as

$y = 6.2 + 1.5N - 3.6E + 4.9W.$

For the Northern region, $N=1$, and $E=W=0$, so to predict y for the North we have $y = 6.2 + 1.5 = 7.7$.

Similarly for the Eastern region, $y = 6.2 - 3.6 = 2.6$; and for the Western region, $y = 6.2 + 4.9 = 11.1$.

For the South, we have $N=E=W=0$, so, simply, $y = 6.2$

The omitted category ('South', in this example), is called the **contrast** category.

6.19 When a multiple regression model includes both a set of dummy variables and one or more continuous explanatory variables, it is called an **analysis of covariance** (ANCOVA). Such models originated in experimental research, and the dummy variables are sometimes referred to as the **treatment levels**. ANCOVA models are used in medical research to assess the effectiveness of different treatments (encoded by dummy variables), while controlling for continuous contextual variables such as weight or blood pressure (called the **covariates**).

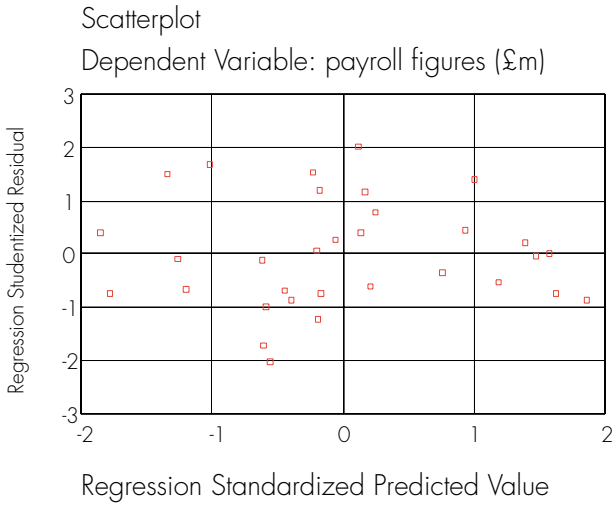
Assumptions underlying regression analysis

6.20 In order for the coefficients and statistics estimated in a regression analysis to be valid, certain assumptions must be met. These relate to the distribution of the **residual** terms-the differences between the predicted and observed values of the response variable. In particular, it is assumed that the residuals are normally distributed with constant variance and a mean of zero. The more the distribution of residuals departs from these conditions, the more risk there is of bias in the regression estimates. An important part of interpreting regression outputs, therefore, is examining **residual plots**.

6.21 These are constructed by plotting the predicted y values against the **standardised residuals** for y (available within the PLOTS option in LINEAR REGRESSION in SPSS). If the regression line is suitable for describing the data, most of the points (normally 95% of residuals) will be between ±2. **Outliers** are those cases with standardised residuals greater than 2. If a number are outside this region, it may be necessary to add additional terms to the regression model.

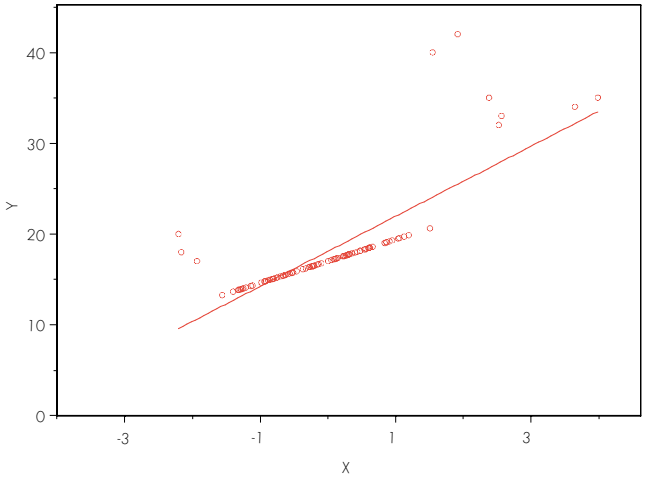
6.22 There should also be no pattern visible in the residual scatter plot. Evidence of trends or patterns may indicate the need for changing the form of the model, or **transforming** the variables so that the regression assumptions hold and unbiased estimates can be obtained for the transformed data. Common transformations include taking logarithms or square roots. A statistician will be able to advise you how to proceed in such cases, and also how to interpret regression coefficients for transformed variables.

6.23 The residual plot shown here demonstrates an acceptable scatter, with no obvious pattern.



6.24 Another factor to watch out for are cases which have a strong influence on your data, or **high leverage**. For example the graph opposite shows the scatter plot and regression line for a simple linear regression. The majority of cases follow a straight line, but there are a number of points which lie above the main group, and as a result, the regression line, rather than running

directly through the main set of points, has been 'levered up'. When this is the case, you often do not get a true picture of the relationship between the response and explanatory variable and so it is useful to run the analysis both including and excluding these observations. The main test statistic used for looking for influential observations is called **Cooke's distance**. Statistics packages such as SPSS and S-plus will list cases with large Cooke's distances.



6.25 It is assumed in regression analysis that no two explanatory variables are perfectly correlated with each other. A high degree of correlation between explanatory variables is referred to as **collinearity**, and can lead to estimates for regression coefficients that are unstable, in the sense of being highly sensitive to small changes in the data or model. Moderate collinearity is not a problem in the technical sense, but often means that the regression coefficients will have quite large standard errors and hence high p-values, which can make it difficult to draw strong conclusions from the analysis. In

general it is best not to include two highly correlated explanatory variables in a regression equation.

Extensions of regression techniques

Modelling non-linear relationships

6.26 The term 'linear' in 'linear regression' refers technically to the fact that the regression equation given in §6.8 is linear in the *regression coefficients*, not in the explanatory variables. If the relationship between explanatory and response variables appears to be better described by a curve than by a straight line, multiple regression models can still be used, incorporating *functions* of the explanatory variables. For instance, a **quadratic** relationship between x and y could be estimated using a regression equation of the form $y = a + bx^2$. The coefficients a and b would be calculated in the usual way, but the interpretation of b is now slightly more complicated, as it relates to marginal change in y with respect to x^2, rather than with respect to x. A statistician can advise you how to set up and interpret such models.

The general linear model

6.27 The **general linear model** is a very general and powerful statistical formulation that includes multiple linear regression as a special case, as well as providing extensions to many other situations in which the assumptions of multiple regression are not met (for example, cases where the response variable is not continuous). Statistical software packages usually provide the facility for setting up general linear models, and you may encounter the results of such modelling in work you commission from academic or technical experts.

Logistic regression

6.28 A special case of the general linear model that is used frequently in social science and economic research is **logistic regression**. This is an appropriate technique to use when investigating questions in which the response variable is a dichotomous variable (a categorical variable with two categories, such as 'yes'/'no', or 'male'/'female'). It works by predicting, for any given combination of values of the explanatory variables, the probability that the response variable will be in one or other of the two categories. The regression coefficients in logistic models still quantify the effects of the explanatory variables, but instead of being interpretable as measures of marginal change, they provide odds ratios. These assess the odds of the response being in one category rather than the other. An **odds ratio** of 1 means there is an equal chance of being in both categories. If odds ratios differ significantly from 1, they provide evidence of positive or negative effects of the explanatory variables on the response.

Using logistic regression to support conclusions

The study team working on the NAO report *Giving Domestic Customers a Choice of Electricity Supplier* (HC 85, 2000–01) used logistic regression methods to assess the influence of various factors on the likelihood of electricity customers' switching to a new electricity provider. They commissioned an omnibus telephone survey of electricity customers to gain information on factors such as income, location, size of bill and payment method. Using logistic regression they were able to quantify the impacts of these factors as odds ratios for whether or not customers switched supplier, and to conclude that there were significant differences in switching behaviour between customers in Scotland and England, once contextual effects had been controlled for. This provided evidence for the conclusion that differences in market structures and regulatory regimes between England and Scotland were affecting customer

Multilevel models

6.29 Very commonly the datasets that are relevant for performance audit and VFM examinations are **hierarchical** or **nested** in structure. For example, you might be interested in the performance of hospitals within NHS trusts within regions, of students within colleges within education authorities, or of staff within regional offices. A powerful extension of regression modelling that enables the sizes of effects at different **levels** within the data to be assessed is known as **multilevel modelling**.

6.30 Multilevel models can be used to construct value-added performance measures, by attributing effects explicitly to different levels in the data. For instance, the (pupil level) effects of general ability and level of parental support can be separated from the (school level) effects of teaching quality and management practices, in assessing school performance measures based on examination results. They can also be used to analyse **longitudinal** data, where follow-up data on subjects or organisations is collected over time. Longitudinal datasets can be thought of as nested structures, in which all the measurements on a given subject are nested within the case that represents that subject. Multilevel models are being increasingly used within departments and agencies to model performance data. You should consult the statistics team if you have collected a hierarchical dataset, to check whether multilevel methods of analysis might be appropriate.

More advanced techniques

6.31 The aim of this handbook is to provide a reference text on the more important and commonly used statistical techniques that can be used to assess evidence in support of conclusions. Some of these techniques, such as the methods

of exploratory data analysis discussed in Chapter 3, should be part of the standard toolkit of performance audit and VFM study teams. It is also important for auditors to be aware of how statistical testing and modelling techniques can be used to provide a secure evidence-base for conclusions, so they can act as 'intelligent consumers' of such analyses.

6.32 Although it is unlikely that you will carry out the sorts of advanced regression analyses described in the previous section yourself, you should be aware that the methods exist, and can be applied to investigate many of the sorts of real-world, complex questions that arise in VFM and performance audit work. It is also important to be aware of some additional types of analyses that you may wish to commission from experts, as they can help you to extract the maximum value and information from the quantitative data you collect during study fieldwork.

6.33 An important set of methods that have not been discussed at all in this guide are known as **multivariate methods**. In all the examples in this handbook, we have considered a *single* response variable at a time. Multiple regression models enable inferences about such a response to be drawn on the basis of multiple explanatory variables. However, multiple regression is usually classed by statisticians as a **univariate** technique, as it still only considers a single response.

6.34 Multivariate methods, such as factor and cluster analysis, multivariate analysis of variance and multivariate regression, consider several response variables simultaneously. **Factor** and **cluster** analysis are **data-reduction** techniques. They enable complex datasets to be simplified, or reduced, to illustrate underlying structure.

6.35 Factor analysis is often used with questionnaire response data, or with multiple performance measurements, to create a small number of **indices** or **scales** which capture most of the information in a large number of variables. For example, an organisation may have a great deal of management information about many aspects of its performance. It is likely that many of the indicators are correlated to some degree. Factor analysis enables the set of indicators to be reduced to a smaller set of core measures that are constructed from appropriate combinations of the indicators, and which represent the major underlying dimensions of performance.

6.36 Cluster analysis is a similar grouping technique, but is used to class cases, rather than variables, together. For instance, many market research companies use classifications of the public into broadly similar socio-economic categories (the ACORN classification system is a well-known categorisation). These categories are arrived at by cluster analysis of sets of response variables, such as income, type of job, leisure activities, collected through surveys and from administrative data. Each respondent is classed with other respondents of a 'most similar' type, depending on the values of the response measures.

6.37 Multivariate regression, and multivariate analysis of variance, are extensions of the regression and ANOVA techniques discussed in this guide to deal with multiple response variables. An extension of the analysis of the contingency table analysis discussed in §5.47-5.48 is known as **loglinear analysis**. It enables crosstabulations of multiple variables to be analysed. It is widely applied in analysing questionnaire data, and has also been used in studies of social mobility.

6.38 **Path analysis** is a statistical method for analysing quantitative data that yields empirical estimates of the effects of variables in a hypothesised causal system. Such systems usually involve both observed and postulated (or **latent**) variables. If sufficient assumptions are made about these variables, it is possible to use **structural equation modelling**, an application of multiple regression, to obtain numerical estimates of the causal relationships between variables.

6.39 All of these techniques are well-known and extensively used in social science and econometric research and in wider quantitative evaluation work. Many could add value to performance audit and VFM work also-for instance, factor analysis to assess key dimensions of performance, or loglinear modelling to examine relationships in respondents' answers to questionnaire survey. You are encouraged to consult the statistics team if you think any of these may be appropriate for an audit you are working on.

Summary:
Key points from Chapter 6

- Regression is a useful technique for investigating relationships between a response variable and one or more explanatory variables.
- Regression coefficients quantify the effects of explanatory variables on the response.
- Categorical explanatory variables are included in regression analyses through the use of dummy variables.
- Regression does not have to rely on linear relationships. It is sometimes more appropriate to use a quadratic or other term in the explanatory variables to explain the relationship better.
- Always check regression diagnostics, such as residual plots and leverage values, when interpreting regression output, and consult a statistician if you think the regression assumptions may not be met.
- Regression models can be extended in many ways. Logistic regression can be used with dichotomous response variables. Multilevel models are appropriate for hierarchical or nested datasets.
- Regression models can be used to control for contextual effects when comparing performance measures.
- Multivariate techniques, such as factor analysis, examine several response variables simultaneously. They can be used to simplify complex performance data into a smaller number of key indicators.

GLOSSARY

GLOSSARY

Glossary

This glossary contains the main terms and concepts introduced in the text. More information about each entry can be found in the section indicated in parentheses. *Italicised* terms are defined in other glossary entries.

Analysis of Variance (§5.34) – or ANOVA, is a *hypothesis test* to examine whether two or more groups have the same mean.

Bar chart (§3.14) – a graphical representation of *categorical* data showing frequency distributions for each category.

Boxplot (§3.14) – graphical representation of the average and spread of quantitative data, showing the *median*, *upper* and *lower quartiles* and *outliers*.

Categorical data (§2.11) – data which consists of descriptions or labels used to identify attributes of a subject in a small number of distinct categories.

Central Limit Theorem (§4.21) – A theoretical result that states that as sample size increases, the *sampling distribution* of the *mean* of a given set of data becomes approximately *Normal*. This result allows statistical tests to be carried out on the data.

Chi-Squared test of association (§5.47) – used on *contingency tables* as a statistical test to investigate whether there is an association between two *categorical variables*.

Chi Squared test of goodness of fit (§5.44) – a *hypothesis test* to assess how similar a set of observations are to their expected *distribution*.

Coefficient of Determination (§6.7) – a measure of proportion of variation in a *response variable* which is accounted for by the *explanatory variables*. Usually symbolised by R^2.

Coefficient of variation (§3.33) – a measure of variation calculated as the *standard deviation* divided by the *mean*. It allows variation for different variables to be compared on the same scale.

Confidence interval (§4.27) – a range of values within which a *parameter* is likely to lie. For a 95% confidence interval, there is a 95% chance that the true value of the parameter lies within the given range.

Contingency table (§5.47) – a table showing the breakdown of one *categorical variable* by the values of another *categorical variable*.

Continuous data (§2.12) – numeric data which can, in principle, take all possible values within a given interval.

Correlation Coefficient (§3.40) – a measure of the linear relationship between two *continuous variables*. Usually symbolised by *r*.

Dataset (§3.4) – a spreadsheet or matrix of data. Each row represents a case (e.g. a respondent to a survey), and each column represents a measurement or observation (e.g. the response to each question on a survey).

Distribution (§4.8) – A list of all the possible values a *variable* can take, together with their frequencies of occurrence.

Dummy variable (§5.17; §6.13) – a variable used to aid the analysis of *categorical* data. It takes the value 1 for cases which fall into the category of interest, and 0 otherwise.

Explanatory variables (§3.6) – variables which represent factors that may influence the measures, or *responses*, of interest.

Histogram (§3.14) – A graphical presentation of the frequency *distribution* of a numeric *variable*.

Hypothesis test (§5.1) – a test to see how consistent the observed data are with a hypothesised statement.

Inter-quartile range (§3.26) – a measure of variability calculated as the difference between the *upper* and *lower quartiles*.

Kruskal-Wallis test (§5.43) – a method for comparing groups similar to an *ANOVA* test.

Logistic regression (§6.28) – a type of *regression analysis* used when the *response variable* has only two possible outcomes, e.g yes/no, or success/failure.

Lower quartile (§3.26) – or 25th percentile, the value of a variable such that 25% of the cases fall below it and 75% fall above it.

Mann-Whitney test (§5.27) – see *Wilcoxon test*.

Mean (§3.12) – a measure of average (central location) for a numeric *variable*, calculated by summing all the values together and dividing by the number of cases.

Median (§3.25) – a measure of central location for an *ordinal* or numeric *variable*. When data are placed in ascending order the median, or 50th percentile, is the value in the middle.

Meta-analysis (§2.7) – see *systematic review*.

Mode (§3.24) – a measure of central location for a *variable*, calculated as the category containing the highest proportion of cases (or the most common value of a numeric *variable*).

Nominal data (§2.14) – Measurements of pieces of information about attributes, e.g. gender, eye colour.

Normal Distribution (§4.11) – the most important of the standard *distributions* used in statistics. A variable is said to be normally distributed if the values are symmetrically distributed in a 'bell shaped' curve around its *mean*.

Null hypothesis (§5.1) – in hypothesis testing, a hypothesis that there is no effect present, or that the observed data are no more different from some postulated values than would be likely just because of chance effects. The question of interest in a hypothesis test is usually whether it is reasonable to reject the null hypothesis or not.

Ordinal variable (§2.14) – a categorical *variable* which has a natural ordering. For example a classification of costs as 'high', 'medium' or 'low'.

Outlier (§3.14) – a value at the extremes of a distribution, an observation within a set of data with an unusually high or unusually low value.

Paired sample *t*-test (§5.29) – a test for assessing the difference in *means* between two related samples.

Parameter (§4.6) – a numeric summary measure or description for a *population*. Often *statistics*, calculated on the basis of sample data, are used to estimate parameters.

Population (§4.5) – the entire collection of units or individuals which is of interest for a particular research question in a performance audit or VFM study.

Proportion (§5.17) – the number of counts in a particular category divided by the total number in the dataset.

***p*-value (§4.33)** – information that results from a hypothesis test. The *p*-value, or **significance level** gives the probability of obtaining a test statistic as extreme or more extreme than that observed if the *null hypothesis* is true. The smaller the *p*-value, the stronger the evidence against the null hypothesis.

R^2 (§6.7) – see *coefficient of determination*.

Regression analysis (§6.1) – a statistical method for analysing the effect of a particular *explanatory variable* on a *response variable* whilst adjusting for other factors at the same time. It also allows the prediction of values of the response variables for different combinations of the explanatory variables.

Relative Risk (§3.37) – a measure of how a *categorical response variable* depends upon a *categorical explanatory variable*. It compares the response proportions between two categories of interest of the explanatory variable.

Residual (§6.5) – sometimes known as 'error', the difference between an observed and a predicted value in *regression analysis*.

Response variables (§3.6) – *variables* that measure the key quantities of interest for the purposes of investigating a specific research question.

Sample (§4.5) – any subset of a *population* of interest.

Sampling distribution (§4.16) – a hypothetical distribution of all possible values of a *statistic*, calculated for each possible *sample* of a given size from a given *population*.

Sampling error (§4.13) – the variability of the estimated values across all possible different *samples* of the same size from the given *population*.

Significance level (§4.33) – see *p-value*.

Single sample *t*-test (§5.12) – a test used to see if an observed *mean* from a single *sample* differs from a given value.

Skewness (§3.27) – An assessment of shape of a *distribution*. If a *histogram* has a long tail on one side then it is said to be skewed.

Standard deviation (§3.30) – a measure of the variation of a variable around its *mean* value. The standard deviation is the square root of the *variance*.

Standard error (§4.18) – the *standard deviation* of the *sampling distribution* of a *statistic*: a measure of the precision of a *statistic* as an estimate of the corresponding *population parameter*.

Systematic review (§2.7) – also known as **meta-analysis**, used to combine the results of several studies to obtain robust information on the measure of interest. Sometimes the term is used in a broader (non-statistical) sense to describe any review of existing studies or evaluations undertaken with reference to a defined set of criteria.

Test of proportions (§5.17) – an analogous test to the *t-test* when dealing with *proportions* rather than *means*.

Two sample *t*-test (§5.21) – a test to investigate whether the *means* of two independent *samples* are different from each other.

Upper quartile (§3.26) – or *75*th percentile – the value of a variable such that *75%* of the cases fall below it and *25%* above it.

Variables (§3.4) – the measures, indicators or quantities we are interested in when investigating a particular issue. In a spreadsheet *dataset*, the variables are the columns of data.

Variance (§3.30) – a measure of variability based on the squared deviations of the data values about the *mean*.

Wilcoxon signed rank test (§5.33) – a method for comparing related groups similar to the *paired sample t-test*.

Wilcoxon test (§5.27) – or **Mann-Whitney Test** – a test for comparing two independent groups similar to the *two sample t-test*.

***z*-score (§3.31)** – a measure which expresses, in units of *standard deviations*, how far away from the *mean* a particular value of a *variable* is.

Designed by Stairway Communications Ltd., London.


NAO
NATIONAL AUDIT OFFICE


AUDIT SCOTLAND