Part 1

Numbers, Data and Analysis

1

Contents

This chapter introduces the basic terminology of statistical data analysis. After reading it, you should understand:

- the stages of analysis,
- what variables and cases are
- the idea of concepts and their indicators
- the four levels of measurement

and become familiar with the two datasets used for examples throughout the book: the World Development Indicators (WDI) and the General Household Survey (GHS) datasets.

'More women go out to work … Then do chores too' said a *Daily Mirror* headline[1]. To back up the point, it quoted a report from the Office of National Statistics which showed that 72 per cent of women have jobs (up from 57 per cent in 1971), and that women with full-time jobs spend about two hours a day doing housework, compared with an average of 32 minutes for men. Nevertheless, 62 per cent of couples think that household tasks should be shared equally.

Being at ease with figures and charts is an essential part of becoming a good social scientist. This book is about how to calculate and interpret percentages and other similar statistics. In this introductory chapter, we shall consider where quantitative data come from, how they differ from qualitative data and how quantitative data can be organised and compared.

Throughout the book, we shall be using examples from two social surveys. The General Household Survey (GHS) is run by the Office of National Statistics (part of the UK Government) and questions a large sample of British households every year. The World Development Indicators (WDI) dataset is collected by the World Bank and consists of data on 208 countries of the world. In this chapter, we shall also briefly introduce these two datasets.

Collecting data

Most social science data is collected by interviewing people. Sociologists either ask a set of pre-set questions (a structured interview) or conduct an interview which is more like a conversation, often recording the answers on a tape recorder for later transcription and analysis. In the former case, every respondent is asked essentially the same set of questions and the answers can be quantified relatively easily. A large number of interviews can be carried out quickly. In the latter case, the interviews often last much longer and are much harder to compare one with one another. Nevertheless, the insights these qualitative interviews give can be much deeper than those offered by standardised quantitative interviews.

There are other ways of gathering data. You have probably filled out questionnaires which arrived through the post, where you tick boxes according to how you want to answer

[1] *Daily Mirror*, Page 6, October 22, 1998

and post back the form (a 'mail survey'). This type of data is also easy to quantify. Other social scientists gather data through observation of social settings, or through examining administrative records or documents. In each of these cases, there is a choice about whether to quantify the data – that is, express it in terms of numbers – or leave it as qualitative data, using terms such as 'larger', 'more frequent' and so on, without putting a numerical value on them. Often, a sociologist will want to express some data quantitatively and some qualitatively, because this is best for the topic at hand. Neither quantitative nor qualitative data are intrinsically better than the other. It all depends on what you are trying to achieve. And often a mixture is better than either alone.

Not all sociologists find that they need to collect data themselves. Often it is quicker, cheaper and better to analyse data that has already been collected. For example, if you wanted to find out about the proportion of women who are now working full-time, it would be a waste of time and money to undertake a survey yourself. The job has already been done by the General Household Survey (and by other large Government surveys) at considerable expense and with great attention to the accuracy of the data collected. Fortunately, the data can be obtained at little or no cost from a Data Archive, a type of library which holds data from previous surveys.

Analysing data that was collected for some other reason or by some other organisation is called **secondary analysis**, as contrasted with the **primary analysis** which you carry out on data you collect yourself. Secondary analysis as a form of research is increasingly popular as more and more high quality datasets covering a very wide range of topics become available.

This book is only about the analysis of quantitative data. Analysing qualitative data requires somewhat different techniques and tools and is therefore left to other texts (e.g. Atkinson et al, 2001; May, 2002; Silverman, 2004). It also focuses specifically on the analysis of the data, saying little about how to collect it. Again, other texts will help you with data collection (e.g. de Vaus, 2002; Robson, 2002). Guidance on the overall process of research, of which statistical analysis is but one part, can be found in Gilbert (2001).

Analysis

The numbers which one collects from a survey tell you very little by themselves. Sociologists are much more interested in patterns and regularities: the features which are common to groups of people in different contexts and situations. To find these patterns, you need to engage in analysis. For example, the government survey which the *Daily Mirror* quoted from and which we summarised at the beginning of this chapter asked several thousand people throughout Britain about how much time they spent doing domestic tasks. Their actual answers, taken one by one, are not of much interest to a sociologist. Put the answers from all the men and all the women together and it becomes plain that women say that they still do many more of the household chores than men.

Analysis consists partly of constructing generalisations: for example, the generalisation that women do more household work than men. Another important element of analysis is explanation. As sociologists, we not only want to know about the social world, but also about why it is like that. So, for example, we might come to the data believing that UK society remains patriarchal, that is, with men still dominating women. We might therefore not be surprised to find that women are still doing the majority of the housework. Alternatively, we might think that as the proportion of women in employment increases, the distribution of housework in dual-earner couples would tend to equalise and therefore be surprised that it is still so unequal. In either case, we would be approaching the data with a prior theory and then testing that theory against generalisations derived from the data. The data either support or cast doubt on the theory.

Induction and deduction

Where does such theory come from? There are two sources: theory can be generated from comparing lots of examples and finding where they have features in common, a process called **induction**. Or theory can come from deriving consequences from a more wide-ranging, 'grander' theory, a process called **deduction**.

For example, suppose that you noticed that among couples you knew, the women were doing a lot more of the cooking and cleaning and men much more of the repairs and decorating even though the men said that they were in favour of equality in doing domestic chores. You might wonder about this disparity between what your friends say and what they actually do. In a small way you are building a theory by induction. You have made an observation and generalised it. The next step would be to see whether what you observed among your friends is true within a wider circle of people. One of the Government social surveys might provide data for this purpose. A national survey could give you information about a cross-section of couples, without the danger that your generalisation would be checked only on your own friends.

If you had access to national data you could also **elaborate** your theory to take account of other factors that might affect the division of domestic labour. For example, you could see whether it was as true for young couples as for old, for the rich as well as the poor and for the well educated as well as those with only basic educational qualifications only. In this way you could enrich the theory to take account of different factors.

Clearly the theory about the division of domestic labour we have been developing here is not the same kind of thing as the grand theories of classical sociologists such as Marx and Weber. More modest theories which describe small parts of social life are often called **middle range** theories. Many middle range theories have been deduced from grander theories. For example, Talcott Parsons (Parsons and Bales, 1956) argued that the family has the function of socialisation, educating children in the ways of the society in which they live. In order to carry out this function, the family needs to have the roles of mother and father clearly differentiated, with the mother maintaining the home and the father providing an income. From this grand theoretical statement, we could derive a middle range theory about the differentiation of household tasks within families. Of course, Parsons' functionalist theory has been much criticised (see, for example, Skidmore, 1975 for a summary) and other theories have emphasised the very unequal division of labour between the sexes and the power relations which help to maintain this (Charles, 2002). Feminist theories of the family can also be used to deduce generalisations that can then be compared with data from surveys.

Deduction and induction are therefore opposite sides to a coin: in carrying out induction, we start with some examples and develop a theory which covers them all. Deduction involves explaining a theory, a generalisation, or particular cases with reference to a more over-arching theory. Thus, induction is a method for generating theories and deduction for applying them. However, in practice, the two tend to get intertwined: one first develops a theory through induction, then tests it against some data using deduction, finds that the theory is not quite right and amends the theory, tests it again and so on.

Neither induction nor deduction is foolproof as a method of doing social science. A survey sample might not be exactly typical of the wider population and so theories based on it could be misleading. With luck, checking the theories using additional and more representative datasets should reveal this and prevent you drawing invalid conclusions. It is wise always to be aware a theory could be wrong, even if it seems to be backed by a wide range of data. A good strategy when testing theories is to aim at **falsification**, that is, try to find data which might show the theory to be false. For example, if the domestic division of labour tends to be particularly egalitarian among those who have university level education, the way to test the theory that men and women have unequal loads within the household is to seek out data about those who are well-educated. If the theory is true even for university students, it may be true for all; if it fails for university students then we know at once it cannot be a valid theory for the population as a whole. Falsification can be summed up as a strategy for testing theories which goes for the most difficult cases first.

Variables

Variables and cases

A survey consists of the same set of questions asked of a large number of respondents. When the pile of completed questionnaires returns to the survey office, the data need to be put into the computer for analysis. For ease of calculation, survey analysis programs generally assume that the data are numerical. This means that the survey questionnaires (which will show the answers as ticks in boxes or as written verbatim responses) need to be **coded** to convert the answers into numerical form.

Often, questionnaires are **pre-coded**, with every possible answer to a question assigned a number which is printed on the survey form. An example is the question reproduced in Exhibit 1.1 from the General Household Survey. If there is no answer at all, for example if the question is not relevant to a respondent or no answer was given, there still needs to be a numeric code provided. In this case, a special code is assigned, often 9 or –9, to indicate *missing data*. Chapter 2 describes coding procedures in more detail.

Once coding has been completed, the data can be entered into the computer. To help with keeping the data organised, the numbers are arranged in a regular form called a **data matrix** (or data array). Imagine a grid of rows and columns — one row for each respondent, and one column for each question in the survey. The numeric codes for every answer can be entered in the body of the grid. The grid is like a map: to find the answer code for a particular respondent on a particular question, look along the row for that respondent until you find the column corresponding to that question. Every cell in the grid will have a number in it, although some, where no answer was given, will contain the missing data code.

There will be one row for every respondent (or, as it is sometimes called, *case*). The order of cases in the grid does not matter, although often the rows will be arranged in order of the identification numbers assigned to respondents. Each column contains the coded answers for one survey question, that is, for one *variable*. For example, one of the columns of the GHS data matrix will hold data about the variable, 'whether the respondent has consulted a doctor in the last two weeks', derived from answers to the question shown in Exhibit 1.1. Generally speaking, there will be one variable for every question in the survey, although some complicated questions including sub-questions may generate more than one variable.

Ask all

During the 2 weeks ending yesterday, apart from any visit to a hospital, did you talk to a doctor for any reason at all, either in person or by telephone?

EXCLUDE: CONSULTATIONS MADE ON BEHALF OF CHILDREN UNDER 16 AND PERSONS OUTSIDE THE HOUSEHOLD.

Yes 1
No 2

Exhibit 1.1 Question showing pre-coding from the GHS
Source: ONS, 2003d

Thus the raw material of quantitative data analysis is a rectangular data matrix of rows, one per respondent or case, and columns, one per variable. The main job of the survey analysis program is to store this matrix of numbers, so permitting statistics about the variables to be calculated.

Variable centred analysis

Almost all statistical analysis in the social sciences is focused on finding relationships between the columns (variables) of the matrix and is therefore called **variable-centred** analysis. Suppose we take two variables such as sex and occupation from a data matrix constructed from a survey of the general population. We might be interested in whether there is some link between the two variables: do women tend to have different occupations from men? Another way of expressing the question is to ask whether there is a **bivariate relationship** between the two variables (bivariate meaning about two variables). We shall be examining the idea of bivariate relationships and how they can be assessed in much more detail in later chapters. For the moment, let us assume that we can detect a relationship between sex and occupation in the data. We can use this conclusion in two ways. We might make a **prediction**: for example, that women graduating from university are likely to go into a different range of jobs than men. Having made this prediction, we might go on to make policy recommendations about how inequalities in occupational selection might be reduced. Alternatively, we might use the finding as the basis for an **explanation**. Suppose we had found that on average women in full-time jobs earn less than men. One explanation could be that women are paid less than men for doing the same job, but another explanation might rely on the relationship between gender and occupation to suggest that women earn less than men on average because they tend to be in lower paid occupations. (Of course, both these explanations might be true at the same time).

Variable-centred analysis dominates sociological research using quantitative data, but it is not the only possible approach. One alternative is to use data to contribute to a history or narrative account of events. In this approach, generalisation tends to be down-played in favour of understanding the particularities of the event under study (Abell, 1987). Another possibility is to build computers models which reproduce some of the patterns in the data, an approach now coming to called social simulation (Gilbert and Troitzsch, 2005).

The data sets

Throughout this book, we shall be illustrating the use of statistical methods and tools by using data from two datasets: data assembled by the World Bank about conditions in most countries of the world; and data from the General Household Survey, a survey of households in Great Britain that gathers factual data about household members and the homes they live in. These datasets are commonly used by social scientists for secondary analysis. In the examples in the following chapters, we have focused on aspects of women's position from a number of points of view, using the two datasets in complimentary ways. In doing this, however, we have only scratched the surface of the information available in the datasets. There are many more variables in each of them than we have mentioned in this book and many other important issues that can be explored using them.

The GHS data set

The General Household Survey (GHS) is carried out by the Office of National Statistics (ONS), a government department that provides statistical services, in order to provide background information for policy making. The survey was started in 1971 and runs continuously – data are gathered throughout the year and then collected together and distributed annually. The respondents are people who live in private households in Great Britain. They are approached by ONS interviewers in their own homes to take part in the survey. Interviews are sought with all the household members aged 16 and over.

The proportion of the households approached that respond averages around 70 per cent, very high for a survey as complex as the GHS. Of all the households approached, around

21 per cent do not wish to take part and 3 per cent could not be contacted (these percentages vary slightly from year to year).

There is a set of core questions which are always asked, and in addition new sections are introduced each year and others dropped, while others are repeated every few years. The core questions include those on housing, employment, education, health and household membership. Other topics covered in some years include informal caring for other members of the household, burglary, dental health, pension schemes, drinking alcohol, fertility, long-distance travel, smoking, voluntary work and many others. The GHS consists almost entirely of factual questions about the household and its members; there are very few questions asking for statements of attitude or belief. Full details about the GHS can be found in the annual reports (e.g. ONS, 2004).

In this book, we have mainly used a subset of the data collected by the GHS from April 2002 to March 2003. The same subset may be accessed on the World Wide Web if you would like to repeat these analyses or do your own (see Appendix A). We have also used some data collected in 1991. The data in computer readable form for all recent years of the survey are available for academic research from the Data Archive at the University of Essex. The Archive also holds copies of many other datasets of interest to sociologists.

The WDI data set

The World Development Indicators (WDI) dataset was put together by the World Bank from a wide range of variables, called social indicators, which were themselves developed and measured by other international agencies such as the World Health Organisation (WHO), the United Nations Children's Fund (UNICEF), UNESCO and so on. The indicators measure such matters as health, education, nutrition, and economic activity in most countries of the world.

For this dataset, therefore, the rows of the data matrix represent countries, rather than individuals, and the columns are the social indicators. The indicators are published in a book of tables (World Bank, 2004) and also as a data file on disk. Having them as a data file means that all manner of comparisons and summaries can be done on them by social scientists. This dataset is also available on the World Wide Web (see Appendix A).

In contrast with the GHS, the WDI dataset does not record information about individuals, but about countries. Collecting data about countries can be even more difficult than getting data from individuals. The World Bank relies on statistical information published by others, often by the countries themselves. There are three major difficulties with this which need to be borne in mind when using the data. First, the countries' definitions of terms may differ. For example, if one is interested in the proportion of the population that has completed secondary education in different countries, one has first to define exactly what is meant by 'secondary education' in a way which makes sense when applied to the education systems of all the countries involved. With some indicators, including education, this can sometimes be hard or impossible to achieve. Second, when a suitable definition of the indicator has been agreed, one has to find data which have been measured according to that definition in all the countries. Sometimes this may mean making estimates from other, more easily available data; sometimes the data is simply not available in the form required (in which case the data will be coded and recorded as 'missing' for that country and that variable). Third, ideally the indicators should be measured at the same time – or at least in the same year – for all countries. Again, in practice, this is hard to achieve. Most of the data in the dataset used in this book relates to measurements made in the late 1990s, but sometimes earlier measurements have had to be used because nothing later is available.

Concepts and indicators

The problem of finding good indicators for the WDI data is an example of a general problem of sociological research: the relationships between indicators and the concepts that they are intended to measure. **Concepts** are the building blocks of theory: ideas such as occupation, gender, education, status, and income are all concepts. Most quantitative theories express some relationship between concepts, for example, that income is related to level of education. However, concepts themselves are not measurable directly. Instead, for each concept in the theory there must be a corresponding **indicator**. An indicator is a method of measurement that aims to measure the concept accurately. For example, an indicator for income might be the sum of money that respondents report when asked about their total weekly earnings. An indicator for education might be the highest educational qualification obtained by the respondent. There are often a variety of possible indicators for a given concept, some of which are more accurate and others more easily measured (for example, the indicator of income could be improved, but made more onerous for the respondent by including interest from savings accounts and dividends from shares; the sums reported may need to be averaged over a number of weeks; and the respondent might need to be asked to produce pay slips, rather than relying on memory). One of the skills of a good researcher is to devise good indicators for the theories being investigated.

How good an indicator is, is a question about its validity and reliability. **Validity** concerns the extent to which an indicator accurately measures the concept. **Reliability** concerns the consistency of the measurement. For example, an indicator that asked respondents how much they earned last week in order to assess their income will probably not be very valid: the answer will depend on the respondent's memory and is likely to omit some components of income. It is also likely to be rather unreliable: if the respondent is paid weekly and sometimes does overtime, the answer will depend on which week of the year the question happened to have been asked. However valid and reliable the indicators we choose to use are, a degree of measurement error is likely to creep in and this needs to be remembered when analysing sociological data.

Kinds of data

There are a number of kinds of quantitative data. It is important to distinguish between them because the kind of data affects the types of analysis that can be done. For example, while it is perfectly acceptable to find the average age of a group of people, it would not be sensible to calculate their average religious affiliation because the result would be meaningless. Unfortunately, finding the average of a set of numerical codes representing respondents' religions is just as easy as finding the average of their ages: in either case, one simply has a batch of numbers to average. Thus, when doing statistical analysis it is important to keep your wits about you, so that you do not unreflectively apply statistical procedures to obtain meaningless results.

There are three ways in which data can be classified that have consequences for the way in which they should be analysed: whether the data are about individuals or aggregates; whether the data are measuring a continuous or discrete variable; and the level of measurement of the variable. Let us consider each of these in turn.

Individual and aggregate data

Every individual person is different. It is therefore hard to predict with any accuracy how a particular individual will behave. Even if we have quite detailed knowledge about a person, we can still be surprised by what they do. For example while, as we saw earlier, women with full-time jobs do an average of two hours housework a day, some full-time working women do much less than this, and some do more. The actual number of hours per day is quite unpredictable, however much data you have about individuals, and is quite likely to vary from one day to the next.

In contrast, if we have data about aggregates, we expect to be able to make much more accurate predictions. For example, if the aggregate amount of housework per day in the UK among full-time working women was two hours in 1998, we can be reasonably confident that it is not going to be different by more than a few minutes in 1999. This is because the figure has been obtained by averaging over very many people: individual idiosyncrasies tend to be smoothed out in the process.

Another example of data for an aggregate variable is shown in Exhibit 1.2. The data concerns the rates of infant mortality (number of deaths per annum of infants under one year of age per 1,000 live births). The variable describes the aggregate infant mortality rate in each of the 208 countries covered by the WDI dataset. Exhibit 1.2 shows how many countries have infant mortality rates in each of the given bands.

About 40 per cent (82) of the 208 countries have rates of infant mortality of 20 deaths per 1,000 or lower. About 35 per cent (72) have more than 40 deaths per 1,000. Not surprisingly, it is the less developed countries that have the higher infant mortality rates. Later in the book, we shall be examining some of the factors that influence countries' rates of infant mortality, such as the number of doctors, the level of literacy and the rate of population growth.

In general, statistical analyses tend to give more reliable results when they depend on aggregate data than when they use individual data. You will find that the analyses in this book based on the WDI dataset – which consists of variables aggregated over whole countries – often give 'better' results than those based on the GHS, which consists of data about individuals. Of course, this is not a reason for using only aggregate data. The unit of

Range of infant mortality rate*	Number of countries
0–20	82
21–40	35
41–60	15
61–80	17
81–100	15
101–120	14
121–140	6
141–160	3
161–180	2
No data	19
Total	208

*Number of deaths per annum of infants under one year of age per 1000 live births

Exhibit 1.2 The variation in infant mortality around the world

Source: World Bank, 2004

analysis (e.g. the individual or the country) needs to be chosen with a view to the problem or topic that you want to study.

Continuous and discrete

The second way in which data varies is according to whether what is being measured comes naturally in 'lumps' or not. For example, children come in units of one child. There is no such thing as 0.5 of a child. We call a variable measuring the number of children in a family a **discrete** variable. On the other hand, the distance of someone's home from their place of work could be any number of miles; for example, we would have no difficulty with the idea that the distance is 1.237 miles. Variables measuring such quantities are known as *continuous* variables.

The importance of the distinction between discrete and continuous variables is that different kinds of statistical methods are appropriate for each. In addition, the distinction can sometimes be a source of confusion. For example, in 2002 the average size of a household in Great Britain was 2.8 people. Obviously, this does not mean that there are households with 2.8 people in them, because people only come in units of a whole person. An average is a measure which is continuous, taking any value, while the variable on which this average has been based, the number of people in households, is discrete.

Another example of potential confusion arises from the practice of grouping categories for the purpose of collecting data. While age is in principle a continuous variable, a survey question will certainly not ask for age in terms of so many years, months, weeks, days, hours, minutes and seconds. Instead you might be asked how old you are, expecting an answer in years, or just in which age group you fall, given a list of five or six groups from which to choose. Once the data from this question have been processed, the result will look to the analyst as though the age variable is discrete even though the underlying variable is continuous.

An example is shown in Exhibit 1.3.

Age	**% in pension scheme**
16–19	1.2
20–29	17.0
30–39	27.9
40–49	29.7
50–59	22.0
60–69	2.2
Total	100

Exhibit 1.3 Whether a member of a pension scheme by age, for women in employment, Great Britain 2002

Source: ONS, 2003b

The table shows age divided into ten year age groups (except the first, which begins at 16). The right hand column shows the percentage of employed women who are members of a pension scheme in each age group. The proportion in the youngest age group, ages 16 to 19, is only 1.2 per cent. It rises to 17 per cent in the twenty to twenty-nine age group and to nearly thirty per cent among the thirty and forty year olds. It then falls again to only twenty-two per cent among those aged 50 to 49, while a mere 2 per cent are members among those aged sixty and over. Dividing the sample into age groups in this way is convenient because it shows at a glance the trends in pension scheme membership by age. If the table had one row for each different age, 16, 17, 18, 19 and so on, it would become a much larger table and harder to make sense of. The Exhibit shows that older women are less likely to join pension schemes and that young women, only just in the labour market, take some time to decide to join a scheme.

Levels of measurement

Earlier in the chapter, we remarked that the quantitative data with which this book is concerned is often obtained from respondents' answers to survey questions. Usually the answers themselves are not numerical, but to include them in a dataset such as the GHS and to analyse them statistically, we need to convert them into numbers using a coding procedure. In doing this, we are converting from qualitatively different answers (for example, 'disagree', 'neutral' or 'agree' in reply to a question about building out-of-town supermarkets) to quantitatively different codes (e.g. 1, 2 or 3). While this is an acceptable procedure, it can give rise to difficulties because the numbers have different properties from the original answers provided by the respondent. For example, if we had just the numerical codes, we might be misled into assuming that we could calculate the average view about out-of-town supermarkets from these codes. But this would be incorrect: since we don't know how strongly individual respondents agreed or disagreed with the idea, we can't pool all the responses to find a meaningful average.

The problem stems from the **level of measurement** of the variable. Conventionally, social scientists distinguish four levels of measurement, although in practice they only use three of them.

If a variable measures only differences between cases, it is measuring at the **nominal** level. For instance, a variable measuring political preference is measured at the nominal level. If one person is recorded as voting Labour and another Conservative, this means only that they vote differently. This remains true even if, for the purpose of analysis, Labour voters are coded as '1' and Conservative voters as '2'. Although the Tory voters are coded 2, one cannot conclude that they are twice as political as Labour voters coded 1. The numbers must be used only as indications that the respondents are in different categories.

In some questions the answers are clearly ordered in terms of bigger and smaller, or better and worse. For example, we might ask respondents about their level of education by finding out about their highest educational qualification, suggesting categories such as GCSE exam passes only, A-levels, or a degree, coded as 1, 2 and 3 respectively. Someone who has a degree is clearly better educated than someone who only has GCSEs. However, it would not be right to say that someone who has a degree is three times better educated than someone with GCSEs, even though the numeric code for a degree is 3 and that for GCSEs is 1. The categories for the respondents' answers are different, as with the nominal level of measurement, but in this case they are also ordered from low to high educational achievement. We therefore say that this variable is measured at the **ordinal** level of measurement. While the categories of an ordinal variable can be ordered (or 'ranked'), the amount of difference between categories is not available. Variables measured at both the nominal and ordinal levels are sometimes referred to as **categorical** variables.

The third level of measurement is the **interval** level. Variables measured at this level have categories that are not only ordered, but for which we can also calculate the relative difference between any two categories. There are very few examples of interval level variables in social science. The fourth level is the **ratio** level. These are variables for which there is a meaningful concept of a zero amount. It follows that it is possible to calculate ratios (such as two to one) for variables measured at the ratio level. For example, income is a ratio level variable: there is no problem about the meaning of zero income and we can say that someone's income is twice as much as someone else's. Like all ratio level variables, income also has the interval property: the difference between someone who earns £10,000 and someone who earns £20,000 is simple to calculate. In fact, social scientists commonly speak of interval level variables when they ought to be referring to ratio level variables. The difference between the two levels of measurement is usually not important for statistical analyses.

These levels of measurement can be arranged in order. Variables at all the levels of measurement share the feature of having distinguishable categories. Ratio, interval and ordinal variables have categories that are ordered. Ratio and interval variables have meaningful differences between their categories. And only ratio variables have a true zero. We shall see in later chapters how the level of measurement of a variable affects the types of statistical analysis that can be performed. In general, ratio and interval level variables allow more complex and more informative analyses to be carried out because variables measured at these levels contain more information (e.g. about the numerical difference between categories) than variables measured at the ordinal or categorical levels.

Overview of book

Most sociological research involves more data than can conveniently and easily be analysed by hand. Computers are therefore now an integral part of the process of analysis. In the next chapter, we introduce the best known computer program for managing sociological data. There are examples and instructions throughout the book of how to use this package. In the following four chapters (Chapters 3 to 6), we consider how to calculate, analyse and display variables from a data matrix. Chapter 7 considers the importance of the normal distribution. This leads us on to considering bivariate relationships between two variables in chapters 8 to 9.

Chapters 10 and 11 discuss the idea of sampling from large populations and the accuracy of the conclusions that you can derive from samples. Finally, in the last chapter, we consider forms of analysis that examine the relationship between more than two variables.

Summary

This book is about how to analyse quantitative data, either collected by the researcher specifically for the purpose (primary analysis) or by others, possibly for other purposes (secondary analysis). Analysis consists of constructing and testing theories. Theories are usually composed of concepts linked by relationships. They are developed through a process of induction from data and tested by means of deducing consequences and then comparing the expected consequences with data: a process of deduction. In order to subject theories to the most stretching test, one tries to find situations where the theory is at risk of being falsified, rather than collecting confirming instances.

Quantitative data consists of measurements made on a number of variables about a set of respondents or cases. The data (for example, ticks on a postal questionnaire or answers to an interviewer's questions) have to be coded into numerical form before they can be analysed. The numbers can most easily be analysed if they are arranged in a data matrix of variables by cases. Variables are measured using indicators, whose accuracy is assessed in terms of their validity and reliability.

Depending on the topic and the source of the data, quantitative data may be collected about individuals or aggregates such as countries. What is measured may come in discrete unit, such as children, or may take on any value, as with age and income. Data may be measured at any one of four levels of measurement: categorical, ordinal, interval and ratio, each successive level representing additional information that has been built into the variable. Arithmetic operations can be carried out on interval and ration variables, but not on categorical and ordinal ones.

You should now understand

- the different kinds of quantitative data
- the ideas of induction and deduction
- how quantitative data can be arranged in a matrix for analysis.

Exercises

1. For each of the concepts listed below (a to d):
   - describe an indicator of that concept;
   - suggest the level at which the concept is being measured (categorical, ordinal or interval/ratio);
   - list one or more problems which might threaten the validity of the indicator's measurement of the concept;
   - assess the likely reliability of the indicator.

     a. Level of education
     b. Alcohol consumption
     c. Opinion about recycling domestic waste for environmental reasons
     d. Actual behaviour in recycling domestic waste
2. Construct a plausible hypothesis involving at least two of the four concepts listed in question 1.
3. Locate an article in a recent issue of an academic journal (e.g. *Sociology*, *Sociological Research Online*, *British Journal of Sociology*, or the *American Journal of Sociology*) that uses quantitative data. See whether you can find examples within the article of a:
   - Hypothesis
   - Concept
   - Indicator
   - Variable
   - Case

Did the research follow a deductive or inductive research strategy? Is it based on primary or secondary analysis?