twelve

ANALYSING QUANTITATIVE DATA

This chapter, perhaps more than any other, illustrates the point made early in the book in Chapter 1 (section 1.5) – that the emphasis throughout the book is on understanding the logic of the research process, rather than on the formulaic learning of research techniques. This chapter is of course about statistics – statistics is the field which has developed the techniques for analysing quantitative data. However, the fact that statistics is based on mathematics presents a problem for many students.

Underlying statistical techniques are mathematical symbols, equations and formulas. My experience is that, if statistics is taught emphasising equations, formulas, and so on (which I used to do), it puts very many students off. As a result, they feel that they are unable to do quantitative research.

However, much more important than knowing the equations and formulae is knowing the logic behind the various statistical techniques, and knowing how and when each can be used in a quantitative research situation. Thus, the objective of this chapter is to describe this logic of quantitative data analysis techniques, in ways that do not require sophisticated mathematics. For this reason, there are no equations or formulae in the chapter, even though it is entirely about statistics. It is not about the mathematics of the statistics. It is about the logic behind the statistical techniques.

It is an obvious point but I stress it in my teaching nonetheless, that every statistical procedure is based on a logical strategy. In the historical development of each statistical technique, this logical strategy preceded the mathematics. Indeed, the mathematics formalises the logical strategy. This applies to the simplest and most commonplace statistics – such as the mean and standard deviation – as well as to more complex and sophisticated techniques – such as factor analysis or multivariate analysis of variance.

I want students to understand, on a logical basis, the need for the different statistical techniques (what was the problem which gave rise to the development of this technique in the first place?) and the logical strategy behind the development of the technique. The equations and formulae which implement this logic, are, in my opinion, of less significance at this level of research. They no longer need to be memorised, since today they are freely available in books and computer programmes. I find that these days I often don't remember them myself. But I know where to look them up when I need them, and I know how and where they can be used in research. This is what I want my students to know.

Example: Take the simplest and most commonplace statistic with which everybody is familiar – the mean, more commonly known as the average. What is the problem which gave rise to its development and use? Simply put, the problem is that we need a way of summarising a set of numbers, scores or readings. When we have multiple data points, multiple scores or readings, we have difficulty comprehending and interpreting them. We need some sensible way of summarising the multiple pieces of information. In particular, we need to know what is called the 'central tendency' of the data.

How can we summarise a set of numbers? What measure of central tendency can we use? There are different possible ways – we could use the most commonly occurring number (technically called the mode) or we could use the number above and below which 50% of the numbers fall (technically called the median). But we very often use the mean, because it is the point in a distribution of numbers around which the sum of the squared deviations from those numbers is a minimum. We need to decode the technical phrase 'the sum of the squared deviations is a minimum'. What does it mean?

Well, wouldn't a good measure of central tendency be based on the deviation (or difference) of each number in the distribution from that central tendency measure? And wouldn't the best measure be where those deviations, taken together, are at a minimum? We can assess the deviation of each number from this measure of central tendency by simple subtraction (either each number from the mean or the mean from each number). But because this gives us both positive and negative numbers, which cancel each other out (mistakenly leading to the conclusion that there is no overall deviation!), we remove the positive and negative signs by squaring each deviation. That gives us squared deviations. Then we add them up – this is what 'taken together' means – giving us the sum of the squared deviations. And it turns out that the mean – the average, which we all know how to get by simply adding the scores and dividing by the number of scores – is the point in the distribution of numbers around which this 'sum of the squared deviations' is a minimum. No other point will give a smaller 'sum of squared deviations'.

Then I illustrate all of this by taking the simplest distribution of just three numbers – say 3, 4 and 5 – working through the above steps, showing that the sum of the squared deviations around the mean of 4 (which comes out to 2 when we work through the simple subtraction, squaring and adding operations) is smaller than for any other value we select. I try other values, working through the simple subtraction, squaring and adding operations, to show that all other sums of squared deviations are higher than 2.

Why spend so much time on this when we all know how to calculate the mean (average) and use it all the time? Because it is a powerful illustration of the main point I want to make throughout this chapter – that every statistical technique is based on a logical strategy, which can be described without reference to mathematics, or at least without reference to complicated mathematics. I want students to see this very clearly and I think the simple example, given above, helps to do this. In my experience, this approach is a very good way of reducing the fear of statistics, and therefore of quantitative research, that many students feel. And, to round off this section, I show students that we have just illustrated and implemented the 'least squares' approach to statistical analysis. This is one of the conceptual and mathematical foundations of modern statistical analysis.

I have tried to make this logical (rather than mathematical) approach saturate this chapter. If I were setting a formal examination on the material of this chapter, I would ask for a short paragraph description of the logic behind the main techniques – especially analysis of variance and correlation-regression. I would not be asking for equations and formulae, nor for exercises requiring the calculations inevitably involved in statistical work.

I don't wish to downplay the importance of equations, formulae and exercises, but I do want to see them in what can be called their 'proper pedagogical place'. I have taught this quantitative data analysis material, in different ways, for more than 40 years. This includes, many years ago, teaching the equations-and-formulas way. I don't think it leads to good outcomes in terms of the training of researchers. Take the example of teaching the analysis of variance. What I have found is that one of two unfortunate outcomes very frequently occurs when we teach the analysis of variance the 'mathematical way' rather than the 'logical way'. Either students have difficulty coping with the mathematics involved in the equations and formulae, and feel that the door to quantitative research is therefore closed to them. Or, they understand and embrace the equations and formulae, and now try to put every research question into an analysis of variance framework. I regard both outcomes as unfortunate.

In today's world – where many beginning researchers come to graduate study without a strong foundation in mathematics – I think it is better to concentrate on the logical approach. This is what I have tried to do throughout this chapter, including with the analysis of variance. Variance itself is a central analytic concept and I spend considerable time on it, discussing it conceptually and making sure students understand its significance. I stress the importance of the research strategy of finding out how something differs or varies (say between people) and then trying to 'explain' these differences or this variance. This of course feeds directly into the 'accounting for variance' research strategy. I stress that this is a very natural way of thinking – we use it all the time, including when we are not in a research context (here, I give examples of everyday use) – and that it is also of fundamental importance in qualitative research and analysis.

In the remainder of this chapter, I have tried to apply the logical approach described above to a description of some of the central techniques developed and used for the analysis of quantitative data, thus:

- cross-tabulations and contingency tables (with chi-square)
- one- and two-way analysis of variance, with interaction

- the analysis of covariance
- moving from univariate analysis to multivariate analysis
- simple correlation and regression
- multiple correlation and regression
- the analysis of complex survey data
- factor analysis
- multiple linear regression (MLR).

As sections 12.4.2 to 12.4.7 show, I place special emphasis on MLR, both as a powerful and flexible quantitative data analysis tool and as a general quantitative design strategy. I do this for several reasons:

- As both a research design and data analysis strategy, it has very wide applicability across many social science research areas (see my book, *Survey Research: The Basics*, p.370, for examples).
- It is conceptually easy for students to understand.
- It implements, in a very direct way, the 'accounting for variance' research strategy described and stressed throughout this book, both in terms of the proportion of variance in a dependent variable accounted for by a set of independent variables, and the order of importance among them in accounting for this variance.
- It is flexible, in the sense that researchers can build their own statistical models for sophisticated data analysis, including hypothesis testing; this makes them less reliant on previously developed formulae.

(Examples of building your own statistical models are testing for interactions between independent variables for their effects on a dependent variable, and testing for non-linearity and curvi-linearity in relationships among variables. This was illustrated long ago by Bottenberg and Ward in their book (see Bottenberg, R.A. and Ward, J.H. (1963) *Applied Multiple Linear Regression*. Texas: Airforce Systems Command; see also Kerlinger, F.N. and Pedhazur, E.J. (1973) *Multiple Regression in Behavioural Research*. New York: Holt, Rinehart and Winston), where they showed that six simple and easy to understand steps in formulating and testing regression models enable researchers to construct their own statistical models to test different hypotheses, and in so doing to conduct a very sophisticated level of analysis).

Factor analysis

Despite being controversial in the eyes of some research methodologists, factor analysis is widely used. Once again, emphasising its mathematical basis can be off-putting to many students, whereas its basic logic is not difficult to understand. The mathematical side of the technique is greatly complicated because there are different varieties of factor analysis (basically in answer to the questions of what we should place in the diagonal of the matrix of correlations and how we know when we have finished factoring). In section 12.6, I have ignored these complications, focusing on the need which gives rise to the technique, and on the logic underlying the technique. I have also stressed its role in the notion of different levels of abstraction and in the process of raising the level of abstraction in the data. As I have tried to show in Figure 12.6, there are remarkable similarities between quantitative and qualitative research on this very point. Factor analysis formalises the process in quantitative research, whereas the process is relatively unformalised in qualitative research.

Statistical inference

Here again, the topic of statistical inference and significance – the subject of much hand-wringing and hair-pulling when taught from its mathematical point of view – can be approached and understood on a purely logical basis, starting with the need or problem which gives rise to the concept in the first place. This is what I have tried to do in section 12.7.