7 Statistics and methods of data analysis

In sociology as a population science the foundational role played by statistics in establishing population regularities stems, in the second place, from the need for methods of data analysis that are able to demonstrate the presence and the form of the population regularities that are emergent from the variability of human social life.

In this chapter, I move on from the role played by statistics in informing methods of data collection in sociology as a population science – that is, through sample survey research – to the role statistics plays in informing data analysis – that is, through what has become known as 'multivariate analysis'. In fact, close links exist in the social sciences between sample survey research and multivariate data analysis; they have to a large extent evolved together.

In order to use surveys to capture the degree of variability in human social life or, in other words, population heterogeneity, the nature of this heterogeneity, or of such part of it as is of research interest, must be specified: that is to say, variables must be envisaged. This entails, to revert to the discussion of Chapter 5, the formation of appropriate concepts and then the development of classifications or scales through which these concepts can be made operational as variables with an adequate degree of reliability and validity. In turn, data from survey research expressed in variable form become the material to which methods of multivariate analysis can be applied in order to bring out – to make visible – the perhaps quite complex relations existing among variables. And it is through the statistical modelling of social data in this way that, in the manner suggested by Stigler, the objects of study of sociology are formed: the population regularities emergent from individual variability that constitute appropriate sociological explananda.

The term 'variable sociology' has often been used pejoratively in attacks on quantitative sociology, whether from 'anti-positivist'

or other positions. It is, however, important to note that two quite different lines of criticism arise, although they are not infrequently confounded.

One objection (see e.g. Abbott, 1992; Esser, 1996; Sørensen, 1998) is concerned with the way in which, in sociological analysis conducted in terms of relations among variables, the action and interaction of individuals that underlie these relations can be, and often are, lost from sight. Thus, not only the description of social regularities *but also their explanation* is given at the level of variables: that is to say, explanation takes the form simply of showing how far the dependent variable of an analysis can be statistically 'accounted for' by those variables deemed to be independent. It is, in other words, variables rather than individuals that 'do the acting'. This argument is one with which I am in essential agreement, and I return to it in Chapter 8.

A second objection is of a more radical but, I believe, far less compelling kind. This objection goes back at least to Blumer (1956) and is to the effect that thinking in terms of variables is inadequate in that much of what is important in human social life cannot be 'reduced' to variable form, even for purposes of description. In part, this critique relies simply on examples of what could be accepted as bad practice in variable sociology: inadequate conceptualisation, deficiencies in the way concepts are made operational and so on. But insofar as it is to be regarded as a critique in principle, it suffers from one major weakness: namely, that its proponents have been unable to offer any alternative to the language of variables as a means of describing features of human social life. It is indeed difficult to envisage any alternative, and this point is underlined by the fact that in qualitative just as in quantitative sociological work, one finds that the language of variables is quite routinely, if only implicitly, resorted to, and further that multivariate analysis is in effect often attempted, albeit at only a verbal level.[1]

[1] Insofar as in qualitative work concepts are not translated explicitly into variables with due concern for reliability and validity, as is a requirement in quantitative work (see Chapter 5), two further consequences may be noted. First, the absence of

'Variable sociology' has then to be regarded as the best, if not the only, way available of producing descriptions of probabilistic population regularities; and its consequent importance is well brought out in a late paper by Robert Merton (1987: 2–6) under the rubric of 'establishing the phenomena'. What Merton is concerned with is the need, before proceeding further in sociological analysis, to ensure that two requirements are met.

The first requirement is that it should be clear that some social regularity does indeed exist – or, as Merton puts it, that events of a certain kind have 'enough of a regularity to *require* and *allow* explanation' (1987: 2–6; italics added). Both words that I have emphasised are significant. To go back to the position I took up in Chapter 4, events that do not display regularity do not call for the sociologist's attention: they are not appropriate sociological explananda, and seeking to explain them sociologically will not be rewarding.

The second requirement then is that every effort should be made to ensure that the form of the regularity in question is properly understood. What at first sight appears to be a fairly straightforward regularity may well, on closer examination, turn out to be a more complex one.

Merton develops his argument by giving various examples of supposed social regularities that, in the light of further research, proved to be non-existent or to have been misconstrued, and he takes the occasion to reassert a view he had expressed two decades previously: that a concern with establishing the phenomena should not be dismissed as 'mere empiricism' since 'pseudo-facts have a way of inducing pseudo-problems, which cannot be solved because matters are not what they purport to be' (Merton, 1959: xv).[2]

this discipline makes it much easier for slippage in meanings and usages to occur. Second, it becomes difficult, if not impossible, for empirical findings to be placed in the public domain in a form that would allow others to reanalyse them – in the way that data-sets resulting from quantitative research are now placed in data archives, together with appropriate documentation, as a matter of course.

[2] I recall once hearing much the same warning being given, in more colourful fashion, by Bill Sewell: 'before you come up with some smart explanation of how the pig got into the tree, just be sure that it *is* the pig that is in the tree'.

Seeking to meet Merton's requirements through multivariate data analysis – that is, using such analysis in the attempt to demonstrate the presence of associations or correlations among variables and to express these relationships in a valid rather than a 'spurious' form – has in fact for long been a central concern of quantitative sociology. In the period after the Second World War, for example, such a concern figured prominently in the work of the group around Paul Lazarsfeld – a close colleague of Merton – at the Bureau of Applied Social Research at Columbia (see e.g. Kendall and Lazarsfeld, 1950; Lazarsfeld, 1955). In analysing contingency tables derived primarily from social survey data, Lazarsfeld and his associates typically began with bivariate relationships and then sought to 'elaborate' these through the introduction of third and further variables, whether in the role of antecedent, mediating or possibly confounding variables.

Lazarsfeld himself saw such elaboration, especially when linked with the time-ordering of variables, as being directed towards demonstrating causation, or at least potential causation. However, his procedures could be better taken as having an essentially descriptive value: that is, as a means of reliably establishing explananda rather than of providing explanations in causal terms. And this did in fact become increasingly evident by the 1970s, as the Lazarsfeldian approach to the analysis of contingency tables was developed and superseded by more formal and powerful loglinear modelling and related methods, notably on the basis of the work of Leo Goodman.[3] Such modelling

[3] A further valuable descriptive technique for sociologists, latent class analysis – in effect the categorical counterpart of factor analysis for continuous variables – was also pioneered by Lazarsfeld (see esp. Lazarsfeld and Henry, 1968); and this too can be understood as closely related to loglinear modelling (McCutcheon and Mills, 1998). In view of criticisms earlier made of typological thinking in sociology, it should be added here that, where typologies are constructed on the basis of latent class analysis – or of optimal matching techniques, as discussed later in this chapter – they can be regarded, in contrast to *a priori* or 'ideal' types, simply as empirical findings: that is, as in themselves a form of revealed population regularity. For it is an important feature of the techniques in question that, where properly used, they may well lead to a negative conclusion: that is, indicate that no regularity in the form of a manageable typology is to be found.

explicitly focuses on revealing patterns of association – and including, perhaps, quite complex interactions – among variables in multi-way contingency tables without any causal implications being claimed and indeed without any need arising for variables to be distinguished as dependent or independent. As Goodman (2007a: 16) has put it in a retrospective paper, what is in this way chiefly contributed is the possibility of bringing out – making visible – for further study regularities of a hitherto unrecognised kind: that is, regularities that were previously 'hidden, embedded in a block of dense data'.

To give a specific illustration of this latter point, I may turn to what was, for a time, a controversial issue in British electoral sociology, in which I had myself some passing involvement: that is, the issue of the role played by gender in political party support. Polling data collected prior to British General Elections from 1945 through at least to the 1980s consistently revealed that the percentage-point difference in Conservative versus Labour support favoured the Conservatives to a greater extent among women than among men. Commentators, especially from the left – a notable example being Hart (1989) – were then led to argue that this 'gender gap' revealed Labour's lack of concern with women's interests and a preoccupation with inequalities and exploitation associated with social class rather than with gender. If Labour had appealed to women to the same extent as to men, it was claimed, the party would have won all elections in the period in question. In other words, the gender gap was explained by a specifically gender effect: that stemming from what Hart (1989) refers to as Labour politicians' 'masculinist blinkers'.

However, while a gender gap in voting could indeed be regularly observed, the nature of the regularity was not in fact well understood until more detailed survey data than those provided by the polling agencies became available. On this basis, the bivariate relationship between gender and vote could be 'elaborated' through multivariate analyses that brought in the further factors of class and of age. And what these analyses then indicated was that the gender gap was a regularity far more complex in its form than had initially appeared. It

was in fact largely an epiphenomenon of *other regularities in which gender-linked voting was not involved*. In responding to a further paper by Hart (1994), I investigated patterns of association in a four-way contingency table of sex × age × class × vote at the 1964 General Election by applying a series of loglinear models (Goldthorpe, 1994) and was able to show that this table was well fitted by a model proposing two three-way associations: that is, of sex, age and class and of age, class and vote. When a further association between sex and vote was added to this model, the improvement in fit was not significant. In other words, the gender gap could be seen as the outcome, on the one hand, of the tendency of women to live longer than men, and especially women in more advantaged classes; and, on the other hand, of the tendency for older people, and especially older people in more advantaged classes, to be more likely to vote Conservative than Labour. A focus on the simple bivariate association that was most immediately in evidence could therefore be seriously misleading. Although over the period in question women *were* more likely than men to favour the Conservatives, multivariate analysis revealed that to seek to explain this regularity in terms of a gender effect – such as Labour's 'masculinist blinkers' – was to grapple with a Mertonian 'pseudo-problem'.[4]

It may, however, be added that my own and others' analyses of what exactly was involved in the gender gap in voting did at the same time serve to reveal one further, quite genuine problem: that is, one concerning the effect of age on voting and the implications of this for the gender gap. The question clearly emerged of whether the age effect was to be understood in life-cycle terms – people tend to become politically more conservative as they get older – or rather in birth-cohort, or 'political generation' terms. Subsequent research, based on repeated surveys of the British electorate, has in fact given strong support to the latter interpretation. And, consistently with this, as

[4] This is not, of course, to say that no such blinkers existed – only that, if they did exist, they were of little relevance in explaining the existing gender gap in voting.

individuals born in the earlier twentieth century have died out, so too has the gender gap in party support that previously existed – and with, if anything, a gap opening up in the reverse direction. The question does of course again arise of whether any such reversal itself results specifically from a gender effect or from other factors. But the need to go beyond the simple bivariate association to establish the precise form of the explanandum is now well appreciated by researchers in the field (see e.g. Inglehart and Norris, 2003: ch. 4).

As, then, contingency table analysis has evolved, its prime importance in sociology as a means of providing descriptions – although perhaps quite sophisticated descriptions – rather than explanations of population regularities has become generally recognised. Loglinear modelling and related methods are now routinely applied in this way in many areas of sociological research: for example, apart from electoral sociology, in the study of social mobility, of social class, gender and ethnic inequalities in educational attainment, and of patterns of homogamy and heterogamy.

However, what has then also to be recognised is that, in the case of regression analysis – the most widely used form of multivariate analysis in sociology – a tendency has of late become evident for this likewise to be seen not as a method of obtaining causal explanations of social phenomena but, again, as one that best serves to establish and describe them. Thus, in a current text on regression analysis, intended primarily for social scientists, one can in fact find the author explicitly stating – in marked contrast to what would have been expected over preceding decades – that regression should be understood as 'inherently a descriptive tool' (Berk, 2004: 206). And what is in turn of further interest in the present context is that it is possible to trace out a reasoned connection between this development and the understanding of sociology as a population science.

For early proponents of regression analysis in sociology, such as Blalock (1961), it was its apparent potential as a means of moving 'from association to causation' (see Freedman, 1997) in research fields largely reliant on observational rather than experimental data that

was its prime attraction. In Chapter 8, I am concerned to show how this understanding of the use of regression has subsequently met with mounting criticism from both statisticians and sociologists alike, and would seem by now to have rather few – overt – supporters. Here, though, I wish to bring out the more positive side of the situation: that is, the way in which an alternative and far more sustainable view of the role of regression in sociology has emerged, and one that would today seem to be very widely adopted in research, at all events *de facto*.

The difference between the two approaches to the use of regression that are in question here is illuminatingly set out in a paper by Xie (2007). Xie emphasises that, with both approaches, regression is the same statistical operation – but that crucial differences arise in the objectives pursued, in underlying assumptions and in the interpretation of the results that are obtained. Blalock and those following him, Xie argues, adhered to what may be called a 'Gaussian' conception of regression. In this case, the aim is to establish a law-like causal relationship between what are taken to be the independent and the dependent variables of the analysis, and the deviation of individual observations from this relationship is then in effect treated as measurement error: that is, as simply undesirable noise. Blalock can thus be regarded as a quantitative analyst much in the style of Quetelet, with, as it were, the least-squares solution of the regression equation replacing the average as the focus of scientific interest; or, in other words, as Xie suggests, Blalock was essentially a 'typological thinker' (Xie, 2007).

In contrast, Xie identifies a 'Galtonian' conception of regression, which he associates primarily with the work of Dudley Duncan, pre-eminently a 'population thinker' (Xie, 2007). In this case, the aim of regression is not to determine causal relationships but rather, through the coefficients returned, to provide a parsimonious description of population variability in regard to the outcome with which the analysis is concerned. The focus is on the systematic component

of this variability: that is, on the variability that occurs among the groups of sociological interest that are defined by the independent variables of the analysis. But it is at the same time understood that the error term of the equation will reflect real *within-group* variability, apart from measurement error *stricto sensu*. And, while of course attempts always can – and should – be made to elaborate the model so as to increase that part of the variability that can be treated as socially systematic, it has to be accepted that within-group, individual-level variability will always remain substantial.[5]

What may then be noted here is the affinity that exists, in regard to data analysis, between regression in this Galtonian sense and the individualistic paradigm in sociology – an affinity that runs parallel to that previously discussed, in regard to data collection, between sample survey research and this paradigm. In both cases alike, the source of the affinity lies in an awareness of the high degree of variability that exists in human social life at the individual level – that is, of population heterogeneity – and of the need for research methods that can be fully responsive to this variability while at the same time allowing the demonstration of such regularities as may exist within it.

It is in this perspective that one should in turn understand the point that is several times made by Duncan that no great importance can attach to the absolute size of the R^2s that are returned by

[5] Given Duncan's pioneering work in 'causal path' analysis in sociology and his later text on structural equation modelling (Duncan, 1975), it might be thought strange that he should be represented as standing in opposition to Blalock's position. However, as Xie (2007) documents, Duncan always emphasised the limitations of such techniques, especially in regard to the demonstration of causation. Xie also reports that Duncan informed him of the difficulties he had in correspondence with Blalock in getting across his views on sociology as a population science (Xie, 2007: 146). Duncan's correspondence – in this case with David Freedman – is of further interest as the apparent source of the distinction between Gaussian and Galtonian conceptions of regression (Xie, 2007: 145, 147). The shifting influence of the methodological work of Lazarsfeld, Duncan and Goodman on American sociology is hilariously captured in the 'anonymous document' reprinted in Goodman (2007b: 137), which should be introductory reading for all sociology students following courses in data analysis.

regression analyses in sociology – these being, with any sensible model, rarely much above 0.3.[6] While, under the holistic paradigm, as already remarked, the expectation would be that far more of population variance than this should be capable of being systematically accounted for, under the individualistic paradigm what is perhaps most remarkable is that regression analyses are usually able to show up *some* systematic effects – despite the fact that the data being analysed will, as Achen (1982: 13) has commented, derive from 'a hopeless jumble of human actors' all engaging to some degree in 'idiosyncratic behaviour as a function of numberless distinctive features of their histories and personalities'. Duncan himself observes, with apparent reference to unmet 'holistic' expectations, that, while the institutional and other structural features of a society may well serve to modify variability in a number of individual characteristics, this will still be 'not nearly so large as the number on which individuals actually differ' (Duncan, 1975: 166–7). And he then goes on pointedly to ask – in what might be regarded as a Malinowskian spirit (see pp. 26–7) – if those sociologists who despair of their low R^2s would 'care to live in the society so structured' that their particular collection of variables 'accounts for 90% instead of 32% of the variance in Y' (Duncan, 1975: 167).

If, then, it is the case that, in sociology as a population science, regression analysis should be seen as serving to establish probabilistic population regularities – in essentially the same way as explicitly descriptive techniques such as loglinear modelling and its derivatives – one further question rather directly arises and needs, in conclusion, to be addressed. That is, the question of whether, in their descriptive work, sociologists should make greater use than they

[6] By 'sensible' here is meant a model that does not include an independent variable that is so 'close' to the dependent variable as to make the analysis essentially uninformative. For example, in a regression model with individuals' social class position at time t as the dependent variable, a high R^2 could of course be achieved by including as an independent variable class position at $t - 1$ week.

presently do of statistical methods that, rather than being based on an explicit probabilistic model of some kind, rely on purely algorithmic modelling – in effect on 'machine learning' from the data under analysis through the application of various techniques of pattern search. The highly controversial issues that arise here in a general statistical context are well brought out in a paper by Breiman (2001) and subsequent discussion (see esp. Cox, 2001). However, care is needed in moving from this general context to the possibilities for algorithmic modelling in sociology specifically.

What is important to note is that the research contexts in which algorithmic modelling has so far been most effectively applied often differ significantly from those most likely to obtain in sociology. Either they are ones in which the aim is to provide short-term predictions of practical importance from given, although often big, data (see the discussion in Chapter 6) – as, say, regarding daily ozone levels or risks of motorway congestion – and where, thus, predictive accuracy takes clear precedence over the interpretability of the results produced. Or they are ones in which it is possible for the algorithmic search to be given strong theoretical guidance – as, say, in the analysis of DNA sequences through optimal matching (OM) techniques. In sociology, by contrast, there is usually the possibility of designing data collection, through survey methods, with specific research problems in mind, and the main aim is in any event not individual-level prediction – that is, high R^2s – but rather to establish the correct form of such population-level regularities as may be present. But, at the same time, strong theory through which pattern search could be informed may well be lacking. While, therefore, there is no reason for sociologists to reject algorithmic modelling out of hand, it would, at least for the present, seem wise to resort to it only on the basis of a detailed evaluation of its likely advantages and disadvantages in particular cases.

For purposes of illustration here, one could take the use of OM (as pioneered in sociology by Abbott: see esp. Abbott and Tsay,

2000), and specifically its use in various aspects of life-course research (as insightfully discussed by Billari, 2005). In treating regularities in events within the life-course, such as occur in the formation and dissolution of partnerships and families or in entry into and exit from employment, and in particular in analysing the correlates of the occurrence and timing of such events, panel regression and event history modelling have come to play a central role. However, as Billari observes, it is difficult through such methods to envisage life-course events as forming 'total' sequences, or trajectories, rather than as being stochastically generated from one time-point to another. And a total view could well be desirable, not least insofar as individuals may themselves envisage their life-courses in this way and pursue long-term life-course strategies, albeit ones subject to various constraints and the operation of essential chance. In this regard, then, sequence analysis, as through OM techniques, has obvious attractions. The series of different states constituting some aspect of the life-course can be systematically 'matched', individual by individual, in terms of the extent and kind of changes that would be necessary to make one sequence of states identical to another, and, on this basis, a matrix of distances between all pairs of sequences can be algorithmically generated. In turn, this matrix can serve as input to some further clustering or multidimensional scaling algorithm that can – or at all events, may – yield a manageable empirical set of sequences.

However, as is in fact widely recognised, the major problem that arises here, as with most sociological applications of OM, is that of setting the 'costs' of transforming one sequence of life-course states into another: that is, the costs to be attributed to the required substitution, insertion or deletion of states. Since it is these costs that determine the distances between sequences, via the algorithmic modelling, they are fundamental to the entire OM analysis; and unless they can be given a convincing rationale, the resulting matrix of distances and any typologies of sequences derived from it are open to the charge that, rather than reflecting some actually existing social regularities,

they may be quite artefactual (Levine, 2000; Wu, 2000). Various ways of dealing with this problem – in effect, the problem of the validity of OM analyses – have been proposed (for a useful review, see Aisenbrey and Fasang, 2010), but what would seem of key importance is that the treatment of transformation costs should have as clear a theoretical basis as possible. Thus, some of the more persuasive OM analyses have been ones of individuals' worklife and, more specifically, social-class histories, in which transformation costs are derived from the theory underlying well-validated class schemata, such as those referred to in Chapter 5 (e.g. Halpin and Chan, 1998). And in such cases, it may then also be possible, by means, say, of regression analyses, for the emergent types of class histories to be related in theoretically coherent ways to antecedent variables such as social origins, cognitive ability and educational attainment (e.g. Bukodi et al., 2015). However, in many sociological applications of OM, including in life-course research, it has to be said that the choice of transformation costs does still appear arbitrary to a rather disturbing degree.

OM, and likewise other algorithmic modelling methods, could therefore best be regarded as ones of *potential* value to sociologists in their efforts to establish the population regularities that form their basic objects of study – but methods that still need to be used very selectively, and with full awareness of the pitfalls to which they may lead. They should, moreover, be seen as ones that are complementary, or indeed ancillary, to probabilistic modelling, rather than as representing some radical and comprehensive alternative. Claims to the effect that algorithmic modelling, and in particular sequence analysis, marks the end of 'the variables revolution' in empirical sociology and the emergence of a new paradigm that focuses on 'events in context' rather than on 'entities with variable attributes' (Abbott, 1995: 93; Abbott and Tsay, 2000: 24; Aisenbrey and Fasang, 2010: 422) would seem little more than rhetorical flourishes – comparable to those noted previously proclaiming the end of social surveys in the era of

big data. Data-driven 'computational' sociology (see Lazer et al., 2009) may well become more prominent. But, far from being superseded, variable sociology in the sense in which it has been understood in this chapter and its expression through probabilistic statistical modelling are, like social survey research, features of sociology that are here for the long term.