6 Statistics and methods of data collection

In sociology as a population science, the foundational role played by statistics in establishing population regularities stems, in the first place, from the need for methods of data collection that are able to accommodate the degree of variability characteristic of human social life, in particular at the individual level, and that can thus provide an adequate basis for the analysis of regularities occurring within the variation that exists.

Over many decades now, the role of statistics in sociology has tended steadily to increase. Yet this tendency would seem to have attracted rather more in the way of criticism – for example, as expressing an unacceptable 'positivism' – than of attempts at explanation: that is, explanation of why it should be that two statistically informed methodologies, sample survey research and multivariate data analysis, should in fact, despite all opposition, have become so central to sociology.

As regards data collection, which is the concern of the present chapter, the following observation may be taken as a starting point. The editors of a leading text on social survey research, Wright and Marsden (2010: 4), make the scarcely disputable claim that 'the sample survey has emerged as the principal means of obtaining information on modern human populations', but they then have little, if anything, to say *about why this should be so*. What can, however, be shown, and has in the present context to be emphasised, is that the sample survey did not achieve its present prominence in some more or less fortuitous way. It did so because it represented the eventual solution to two (closely related) problems that persisted in social research from the mid-nineteenth through to the mid-twentieth century.

The first of these problems was that the data on human populations obtained from the censuses and registration procedures that were developed in Western societies from the later eighteenth

century onwards, while for many purposes invaluable, had, if only on grounds of cost, to be quite restricted in their scope. Other methods of data collection were therefore needed through which population coverage could be traded off against the possibility of obtaining information of a wider-ranging yet also more detailed kind. However, the second problem then arose. If 'complete enumeration' was to be supplemented by such 'partial studies' – to use the language of the day – some appropriate methodology had to be developed for moving 'from part to whole' on a reliable basis.

An attempt at meeting the first problem came with the 'monographic' approach to social research, as advocated and practised most notably by Frédéric Le Play and his followers. Le Play proposed, and sought to implement, a methodology which entailed the first-hand and protracted study of individuals in the context of their families and communities – in effect, an early form of ethnographic case study (Zonabend, 1992). Information was to be collected on the economic conditions under which individuals lived, but in greater depth than in official statistics, together with information on their primary social relations, their life histories, their aspirations for the future, and their moral beliefs and values. Researchers needed to 'speak the same language' as the men and women they studied and to 'enter into their minds' (Silver, 1982: 41–75, 171–83).

Such research was thus designed – to use now modern terminology – to be intensive rather than extensive in its nature, with the emphasis being placed on the qualitative characteristics of the material obtained in each case, rather than on the number of cases studied. Thus, in the first edition of Le Play's own major work, he presented monographs relating to just thirty-six working-class families spread across different European countries; he extended this to fifty-seven families in the second edition (Le Play, 1877–79). A number of his followers produced similar collections, also dealing mainly with working-class living conditions and family relations.

However, while the Leplaysians introduced a way of collecting social data that could go beyond what was possible with complete

enumeration, and that did indeed produce data of a kind hitherto little available – in particular, on family forms and family budgets – their approach ran into serious difficulties in dealing with the second problem previously noted: that of moving from part to whole. The Leplaysians clearly wished to take the findings of their monographs as a basis for advancing general propositions about European working classes; and, in order to justify this, the claim they made was that their cases were so selected as to be 'typical' – at least, say, of workers in particular occupations, industries or regions. But they then failed to provide any consistent or compelling account of how they actually achieved such selection for typicality. Le Play himself maintained that adequate guidance in this respect could be obtained from 'local authorities', such as civil servants, clergy or doctors, while some of his followers proposed that cases could be chosen that were shown by official statistics to be in various respects 'average' – making then an appeal to the Queteletian idea (see Chapters 1 and 5) that the average would represent the socially significant type, with variation around the average being merely contingent and thus of little scientific interest.

Such arguments, perhaps not surprisingly, met with a good deal of contemporary scepticism – extending in some instances to charges that the selection of cases was in fact biased so as to lend support to Le Play's socially conservative views.[1] And in the light of subsequent research, it has indeed become evident enough that generalisations of a quite mistaken kind were advanced – the most notable, perhaps, being that leading to what is now known by historical demographers as 'the myth of the extended family'. That is, Le Play's attempt to

[1] It was pointed out that where 'local authorities' guided the selection of cases, they might be expected to pick out families who were known to be supportive of the status quo rather than those who were in some way dissident; and suspicions of conservative bias were only reinforced where the assumption was made that atypicality implied not only statistical but also social deviance. Several re-studies carried out in localities covered by Le Play's work did in fact claim to show less harmony in community and workplace relations and less satisfaction with the prevailing order than he had indicated (see Lazarsfeld, 1961; Silver, 1982: 54–75).

represent the extended family as being prevalent in pre-industrial Western as well as Eastern Europe (Laslett, 1972).

More importantly, though, as well as doubts being raised about the monographers' selection of cases, an objection to their approach of a yet more basic kind was made by a number of statisticians, even as they too recognised the need to move beyond complete enumeration in the collection of social data. A leading figure in this regard was Anders Kiaer, Director General of the Norwegian Bureau of Statistics from 1877 to 1913. What Kiaer (1895–96, 1903) and others argued was that, in designing partial studies, the monographers' quest for typicality, even if it could be realised, was still mistaken in principle. This was so because human populations had to be studied not in terms simply of social types but in such a way as to take account, as Kiaer put it (1895–96: 181), of 'all the variation of cases that one finds in life'. Thus, the aim in partial studies should not be to achieve typicality but rather to move from part to whole in a quite different way: that is, by selecting a *representative sample* of the population under investigation in the sense of one that would provide 'a true miniature of it' in respect of the full degree of variation existing among its members in all attributes of research interest.[2] In other words, what Kiaer was in effect urging was a shift from typological to population thinking in the methodology of data collection parallel to that which was occurring in the methodology of data analysis with the move from the Queteletian statistics of the average to the Galtonian statistics of variation.

Kiaer himself pioneered the method of what became known as 'purposive' (or sometimes as 'judgemental') sampling. With such sampling, census and other aggregate statistics were initially drawn on in order to select – in modern terminology – primary and secondary sampling units so that these would give an overall 'match' with the

[2] A strong echo of this argument is found in the study of working-class family budgets made by Halbwachs (1912), who criticises Le Play for his concentration on supposedly typical cases to the neglect of the full range of variation that could be shown to exist.

target population. Fieldworkers were then required to follow certain routes within the secondary units and to select for interview not individuals who were 'typical' but rather those who would, in the light of the fieldworker's own knowledge, best represent the whole range of social variation existing within the unit. When the survey was completed, its degree of success in producing 'a true miniature' could be gauged, Kiaer (1903) believed, by comparing the distributions of respondents on various 'control variables', such as age, marital status and occupation, with established census distributions.

However, while population sampling as developed by Kiaer marked a clear advance in partial studies, his approach still did not provide a final solution to the problem of generalising from part to whole: that is, from sample to population. Statisticians more versed in the emerging probability theory of the day than was Kiaer pointed out that he did not consider whether discrepancies revealed by his checks through control variables were or were not greater than could be expected by chance; and, further, that even if such checks appeared satisfactory in regard to univariate distributions, this result would not necessarily extend to joint distributions (see further Kruskal and Mosteller, 1980; Lie, 2002).[3]

The solution to the part–whole problem in the context of sample surveys came in fact only with what could be regarded as a further, still more significant advance in population thinking. This was the development, in the place of purposive sampling, of probabilistic, or 'random', sampling. With this approach, the key requirement was that every individual in the target population should be given

[3] This last point was in fact made by von Bortkiewicz at a meeting of the International Statistical Institute in 1901 (Kruskal and Mosteller, 1980), and again later in a critique of Max Weber's research on industrial workers in Germany (Verein für Sozialpolitik, 1912). This study by Weber (1908), and another he earlier made of agrarian workers east of the Elbe (Weber, 1892), could be regarded, along with Charles Booth's (1889–1903) study of poverty in London, as ones that sought to bridge the gap between censuses and monographs, but without setting out any underlying rationale for moving from part to whole – or at least not one that could be generally applied.

an equal probability of being selected in the sample or – in later, more sophisticated formulations – a probability that was calculable and not zero. An important pioneer of probabilistic sampling was A. L. Bowley (1906; see also Bowley and Burnett-Hurst, 1915), and chiefly from his work a new conception of representativeness in a sample emerged. Under this new conception, the aim in sampling was not directly to produce 'a true miniature' of the target population, as Kiaer had sought to do, but rather to produce, through probabilistic methods, a sample that was representative or 'fair' in the sense of being *unbiased*, along with some estimation of the error likely to result in making inferences from a probabilistic sample even in the absence of bias.

For some time after the First World War, the old and the new forms of sampling did in fact remain in a state of uneasy coexistence. But the decisive contribution eventually came from none other than Neyman, with whom this book began. In a classic paper, Neyman (1934) gave a compelling demonstration, on both theoretical and empirical grounds, of the dangers of purposive sampling and of the advantages of probabilistic sampling and the calculation of 'confidence intervals' in moving from sample to population. Also important was Neyman's demonstration of how prior knowledge of the target population, which had played a large role in purposive sampling, could be properly brought into sample design: that is, in informing not the selection of sampling units but rather the initial 'stratification' of the target population into subpopulations believed to differ in ways relevant to the purposes of the survey, each of which could then be sampled probabilistically, and with, if desired, differing sampling fractions.[4]

[4] Its apparent disregard for all prior population knowledge was something that, at an intuitive level, would appear to have told against probabilistic sampling in the debates of the interwar years. Neyman (1952: 122) at a later point revealed that he himself had wondered 'how would this random sampling work in practice'. He was reassured by its trial application, under his guidance, in a study of the structure of

Post-Neyman, it could be said, the need for the probabilistic sampling of populations in social research became progressively accepted. Proponents of all forms of non-probabilistic sampling have been forced into increasingly defensive positions (Smith, 1997). For example, forms of quota sampling, though still used in market research and political polling (despite well-known disasters, such as in the British General Election of 1992 and again in that of 2015), would now rarely be thought appropriate for serious social scientific research. The basic consideration is that if a sample is *not* selected probabilistically, the non-negligible possibility always exists that the actual procedures involved will themselves be a source of bias: that is, because they will in some way 'tap into' social regularities existing within the target population, so that information is more likely to be obtained from members of this population who possess certain characteristics than from others.[5]

Desrosières (1991, 1993: ch. 7 esp.), in reviewing much the same history as in the foregoing, has represented monographs – or, in modern terminology, case studies – and sample surveys as two different methods of social data collection, each with its own inherent 'logic' of moving from part to whole, which reflect different ways of

the Polish working class undertaken by Jan Pieckalkiewicz (1934). Neyman's (1952) *Lectures and Conferences on Mathematical Statistics and Probability* is dedicated to the memory of Pieckalkiewicz, murdered by the Gestapo in 1943, and to Neyman's other former colleagues in Warsaw who died in the Second World War.

[5] Thus, with quota samples, an initial problem is that of how well the configuration of the sample – the 'quotas' – matches that of the population on the control variables selected. But a further problem is that of how far, in meeting the quotas, the practice of taking substitutes for those individuals who refuse to be interviewed – which is often up to half of those approached – creates an 'availability bias'. Such a bias would appear to have been a major factor in the failure of the polls at the 1992 British General Election. In some cases, it should be added, the very nature of the research problem being addressed may mean that the probabilistic sampling of a target population is not practical and other methods have therefore to be used: as, for example, with the 'snowball' sampling of populations that are 'hidden' because, say, of their members' deviant or subversive activities (Salganik and Heckathorn, 2004) or with the sampling required in social network research where the aim is to go beyond 'ego-centered' to 'complete' networks. But what is important is that in such cases the attempt is then made to evaluate the sample obtained against the 'gold standard' that probabilistic sampling provides.

envisaging human societies: in fact, those expressed in the holistic and the individualistic paradigms, as discussed in Chapters 2 and 3. The increasing dominance of survey methodology, Desrosières then maintains, has to be understood as reflecting macro-social changes, such as the emergence of popular democracies and mass consumer markets. However, such an 'externalist' account of the scientific developments in question, apart from depending on a large measure of quite conjectural history, is seriously deficient in neglecting what would be the focus of an 'internalist' account: that is, the processes through which successive problems were recognised, addressed and overcome. If two different 'logics' of moving from part to whole were indeed involved, then one – that deriving from the individualistic paradigm – was shown, on the basis of evidence and analysis, to be superior to the other – that deriving from the holistic paradigm. And, to revert to the starting point of this chapter, this could in itself be taken as a sufficient explanation of why surveys, with probabilistic sample selection, have become 'the principle means of obtaining information' on human populations: that is, simply because they are the way of undertaking partial studies of these populations, in all their heterogeneity, that can provide the most cogent rationale for moving from part to whole. Such an internalist understanding of the dominance of survey methodology in social research does, moreover, carry wider implications in at least two respects.

First of all, it underlines the fact that the difficulty experienced by the Leplaysians of how to demonstrate the typicality of monographs, or case studies, as a basis for generalising from them has never been resolved. And it is in turn difficult not to see this as the main factor underlying the declining popularity in sociology today of case studies, at least as a means of characterising the populations within which they are situated.

As an illustration here, one may take the rather rapid fall-off that occurred in the number of 'community studies' undertaken in British sociology following a period from the 1930s through to the 1960s, in which they were – as also in the US and elsewhere – among

the most prominent forms of social research. What could be regarded as a turning-point in Britain came with the attempt made by Ronald Frankenberg, a disciple of Gluckman (see p. 29), to draw on a selection of community studies in order to 'generalize grandly' about British society 'as a whole' (Frankenberg, 1966: 11–12). Frankenberg's integrative idea was that of a 'morphological' – in effect a rural–urban – continuum of types of community that related primarily to the degree of role differentiation among their inhabitants. But, as well as being of questionable relevance to many of the communities to which Frankenberg sought to apply it, this idea clearly failed to provide a convincing grounding for any wider synthesis. In a subsequent collection of papers on community studies, in part querying their future (Bell and Newby, 1974), little reference was made to Frankenberg's work, and then only critically. What is in this way pointed up is – to echo Kiaer's criticism of the monographers – the error of supposing that typological thinking can be adequate to capture the actual range of population heterogeneity. What undermined Frankenberg's generalising ambitions was not only the increasing variation expressed in the steady emergence of new types of community in post-war Britain – for example, 'bimodal' villages in part colonised by urban commuters, inner-city localities characterised by ethnic divisions and conflict, and 'gentrified' former working-class districts. Far more serious was the quite overlooked variation represented in the social lives of the large numbers of individuals resident in urban and suburban areas in which the very existence of communities of the spatially well-defined kind on which Frankenberg's typology depended would have to be seen as highly problematic.

It may in this connection be further noted that Yin, the author of what is perhaps now the leading text on case-study methodology, explicitly states, in some contrast to positions taken up by earlier authors, that case studies should *not* be regarded as being generalisable in a statistical sense: that is, to populations (Yin, 2003: 10). And, in similar vein, Morgan (2014: 298) acknowledges that there are no systematic rules available, analogous to those used in statistical

work, 'for inferring – or transporting – findings beyond the single case study (or even beyond two or three such case studies that suggest the same results)'. Yin, it should be added, goes on to argue that case studies *can* be generalised, if not to populations, then to 'theoretical propositions' (2003: 10). What this means is not entirely clear. But if what is being claimed is that case studies can take on wider significance where they serve as a means of illustrating or, better, of *testing* theoretical propositions *in relation to which their selection has been specifically made* – for example, as in some sense 'deviant' or 'critical' cases – then the argument clearly carries force. It is at all events with this kind of purpose in mind that, in the context of sociology as a population science, case studies could be most appropriately and usefully undertaken (see further Chapter 9).[6]

The second implication of an internalist understanding of why survey research has become dominant in sociology is the following. Any methodology that is to prove capable of superseding, or even supplementing, survey research must itself incorporate the demonstrated advantages of such research in the study of human populations. This point is relevant in regard to arguments now sometimes put forward, usually from externalist positions, to suggest that the age of social surveys is passing, and in particular as a result of the growing possibilities offered to social science by 'big data'.

For example, Savage and Burrows (2007) have claimed that, with the generation and accumulation of vast quantities of 'transactional' data, especially within the private, commercial sector, the privileged role of the sample survey as a means of obtaining information on human populations is being called into question. In comparison with transactional data-gathering, the sample survey is, in their view, 'a

[6] As a graduate student in sociology, I was given as a prime example of deviant case analysis the study of democracy within the International Typographical Union by Lipset, Trow and Coleman (1956) – that is, in relation to Michels' 'iron law of oligarchy'. This work was then an influence on the design of the Affluent Worker study in which I was later involved, which took relatively well-paid workers in the rapidly growing industrial town of Luton as providing a critical case for testing the thesis of progressive working-class *embourgeoisement* (Goldthorpe et al., 1969).

very poor instrument', and is in fact unlikely to remain 'a particularly important research tool' (Savage and Burrows, 2007: 891–2).[7]

However, what is here rather remarkably ignored is the extent of the deficiencies that, from a social science standpoint, are apparent in transactional data, as indeed in most other forms of big data, whether resulting from commercial or other – for example, social media – activity (Couper, 2013). To begin with, problems of sample selection bias and thus of representativeness must be expected to arise from the very processes through which big data are generated; and where claims are made that sampling issues do not arise since 'all' cases are covered, further problems are then often apparent regarding the 'all'. That is to say, the population reference lacks clarity or is of doubtful social science relevance. Still more seriously, though, big data-sets usually include only a rather limited range of variables, and then ones relating to the concerns underlying the data-creation process and only coincidentally to sociological concepts or theory. And, in turn, the analysis of such data is typically aimed, as big data enthusiasts do indeed emphasise (see e.g. Mayer-Schönberger and Cukier, 2013: chs 1–4), at making relative short-term *predictions* in regard to some particular outcome, on the basis of entirely inductive, correlational pattern-seeking, with little regard for the need to proceed from the empirically established regularities to causal *explanations*. The dangers in this approach, even for the limited purposes in question, should be evident enough; and, indeed, some initially much-publicised 'successes', such as the Google Flu Trends project, have turned out, on closer examination, to be seriously flawed (Lazer et al., 2014).

Whatever value big data may have for 'knowing capitalism', its value to social science has, therefore, for the present at least, to

[7] In direct criticism of survey research, Savage and Burrows put forward only one pertinent point: that it currently faces a problem of declining, and possibly increasingly biased, response rates. However, they then say nothing of the significant advances that have recently been made in addressing this problem through methods of weighting for non-response or for the multiple imputation of missing data.

remain very much open to question. Sociologists should, of course, be ready to use data of whatever provenance if their own purposes can in this way be well served. Some relatively early forms of big data, such as the national registers in Nordic countries, which provide comprehensive information on individuals' incomes or education, are extremely valuable resources that are already widely exploited. But what authors such as Savage and Burrows fail to see is that little will be achieved in drawing on data that fall short of standards that have been established for good scientific reasons. Issues of the quality of data and their fitness for purpose cannot be overlooked. As Cox and Donnelly (2011: 3) aptly put it, 'A large amount of data is in no way synonymous with a large amount of information'. And, indeed, where large data-sets of low quality are analysed, and especially by inductive, 'data-dredging' methods, the risk of *negative* outcomes is high: that is, of noise being mistaken for signal (Silver, 2012) and of essentially arbitrary and thus quite misleading results being produced.[8]

Finally here, it should be noted that, far from there being any actual indications of sample survey research entering into a period of decline, what is at the present time striking is the increasing amount of such research being undertaken and the growing sophistication of survey design and implementation. Moreover, among the most important advances being made are ones that in effect undermine standard criticism of survey research, especially as put forward from holistic positions, to the effect that the conception of society that such research entails is unduly atemporal and atomistic.[9]

[8] It is in fact on essentially these lines that cogent criticism (Mills, 2014) has been directed against efforts led by Savage (Savage et al., 2013) to construct a 'new social class map' for Britain on the basis of highly biased big data from self-selected respondents to an internet survey – supplemented by data from a quota-sample survey whose degree of representativeness is not open to any reliable estimation.

[9] For a review of, and a powerful response to, such criticisms, written at the height of the 'reaction against positivism', see the courageous book of Cathie Marsh (1982), whose tragically early death robbed survey research in Britain of one of its rising stars.

On the one hand, the development over recent decades both of repeated cross-sectional surveys of the same population and of longitudinal or panel surveys of the same samples of individuals has become of prime importance *to the understanding of processes of social change*. In particular, longitudinal surveys are crucial to the complex task of separating out the influence on individual life-courses exerted by period, birth-cohort and age or life-cycle effects; or, in other words, to meeting the requirement of Wright Mills (1959) – a one-time strident adversary of survey research – that a key focus of sociological inquiry should be on the intersection of history and biography. And a general feature of analyses based on survey data of the kind in question is their demonstration of how, through these differing effects, a remarkable diversity in individuals' life-courses is created (see e.g. Ferri, Bynner and Wadsworth, 2003) – fully justifying the argument of Wrong (see p. 24) that such diversity is always likely to represent a powerful countervailing force against the homogenising tendencies of enculturation and socialisation.

On the other hand, hierarchical survey designs, in which supra-individual entities are first sampled, and then individuals within these entities, specifically allow for the operation of 'contextual' effects on individuals' life-chances and life-choices: that is, the effects of the social composition and structure of the groups, networks, organisations, associations, communities and so forth in which they are involved. And, in turn, such designs make it possible for the importance of contextual effects to be assessed in comparison with those of individuals' own, variable, characteristics.

In case studies of holistic inspiration, it is often simply assumed that contextual effects are pervasive. For example, such an assumption underpinned much of the work of the Institute of Community Studies in London in the 1950s and 1960s. As Platt (1971: 75–7, 96–8 esp.) has observed, the social class composition of local communities was represented as in various respects shaping the lives of their individual members – but without the question even being considered of whether, or to what extent, such contextual effects could

actually be demonstrated independently of the effects of individuals' own class positions.[10] However, in research based on hierarchical survey designs, while contextual effects are usually shown up to some extent in relation to outcomes of interest, they are most often found to be clearly *less* important than are those of individual-level variables – as, say, in the case of school effects on children's academic performance; or, otherwise, contextual effects prove to be difficult to separate out from individual *selection* effects – as, say, in the case of constituency or neighbourhood effects on voting behaviour. In this latter connection, it is also relevant to note that, of late, the inadequate treatment of selection effects has become the focus of criticism (see e.g. Lyons, 2011) of the extreme claims made by some social network analysts that networks exert 'amazing power' over individuals' lives and operate as 'a kind of human superorganism' (Christakis and Fowler, 2010: xii). The crucial question that arises is that of how far their networks influence individuals or how far individuals choose and thus influence their networks.[11]

The main concern of this chapter has been to show that probabilistic sample surveys represent a statistically informed methodology that is foundational for sociology as a population science: that is, because such surveys constitute the best means so far devised of moving from part to whole when trading off population coverage against informational content in the collection of data from human

[10] More recently, unsupported assumptions on similar lines can be found in literature in which the individual risk of poverty or 'social exclusion' is associated with the contextual effects of living in inner-city districts or 'sink' estates.

[11] Lyons' particular concern is with the claim made by Christakis and Fowler that the influence of social network membership increases the risk of obesity. A more general issue of the existence or strength of contextual effects in regard to health and well-being also arises in debates over the work of Wilkinson and Pickett (2010): that is, over their argument that the adverse effects of economic inequality operate not only at the individual level but also, and more importantly, at the societal level. Research reviews indicating that the evidence for such contextual effects is, at best, patchy (e.g. Lynch et al., 2004; Leigh, Jencks and Smeeding, 2009) have received no serious response from Wilkinson and Pickett, and the bivariate scatterplots on which they largely rely are of little help in supporting their case: analyses based on appropriate hierarchical survey designs and multilevel modelling are called for.

populations. In the light of the foregoing, it can, however, also be said that the advance of survey methodology has *in itself* significantly aided the development of population thinking, as opposed to typological thinking, in sociology. In particular, awareness has been increased of the degree of individual variation, especially where a life-course perspective is taken, and of the limits on the extent to which this variation is modified by individuals' involvement in supra-individual entities. It is in this regard of interest to note a comment made by the author of a leading text on the modelling of data from more sophisticated survey designs. Hox (2010: 8) observes that, while Durkheim's conception of sociology was as a science 'that focuses primarily on the constraints that a society can put on its members', there are now good grounds for some reversal of perspective: that is, for a focus on the extent to which the features of sociocultural entities are shaped by the actions of the individuals by whom they are, or once were, populated.