1 Sociology as a population science: the central idea

Sociology should be understood as a population science in the sense of Neyman (1975).

On the occasion of the 500th anniversary of the birth of Nicholas Copernicus (1473–1543), the US National Academy of Sciences sponsored a collection of essays on 'quasi-Copernican' scientific revolutions. The volume was edited by the Polish-born statistician Jerzy Neyman, who supplied brief introductions to its several sections. In one such introduction, to a series of essays on 'The Study of Chance Mechanisms – A Quasi-Copernican Revolution in Science and Mathematics', Neyman (1975: 417) made the following observation:

> Beginning with the nineteenth century, and increasing in the twentieth, science brought about 'pluralistic' subjects of study, categories of entities satisfying certain definitions but varying in their individual properties. Technically, such categories are called 'populations'.

Neyman emphasised that populations in this technical sense could, substantively, be of quite different kinds. They could be human or other animal populations, but also populations of, say, molecules or galaxies. The common feature of such populations was that, while their individual elements were subject to considerable variability and might appear, at least in some respects, indeterminate in their states and behaviour, they could nonetheless *exhibit aggregate-level regularities of a probabilistic kind*.[1]

[1] I was first directed to Neyman's remarks by a reference to them in Duncan (1984: 96). As will become readily apparent, Dudley Duncan is an author to whom I am

The aims of a science dealing with such pluralistic subjects of study – or, that is, of what could be called a 'population science' – were then twofold. The initial aim was to investigate, and to establish, the probabilistic regularities that characterise a particular population, or its appropriately defined subpopulations. In this regard, Neyman saw the use of statistical methods of both data collection and analysis as being essential. And indeed, fifty years previously, R. A. Fisher (1925: 2) had already *defined* statistics as 'the study of populations, or aggregates of individuals', and had represented statistics as foundational for all sciences that were primarily concerned with the properties of aggregates rather than of their individual members. It may moreover be noted, in view of what is to follow, that Fisher then added the remark that 'Statistical methods are essential to social studies, and it is principally by the aid of such methods that these studies may be raised to the rank of science.'[2]

However, Neyman also made it clear that once population regularities had been empirically established, the further aim of a population science had to be that of determining the processes or 'mechanisms' which *in their operation at the individual level* actually produced these regularities. And since the regularities – the explananda of a population science – were probabilistic, the mechanisms that would need to be envisaged would be ones that, rather than being entirely grounded in deterministic laws, *incorporated chance*. A new form of scientific explanation was implied.

Neyman's claim that from the nineteenth into the twentieth century the increasing study of 'pluralistic' entities on a statistical

indebted in many other respects. He must be regarded as one of the great pioneers in conceptualising and practising sociology as a population science. Another who contributed significantly, although in a less explicit way, was my former teacher at the London School of Economics, David Glass – now shamefully disremembered in British sociology – under the influence of his own teacher, the extraordinary polymath Lancelot Hogben (see Hogben, 1938).

[2] Neyman and Fisher were of course the leading antagonists in what has been described as 'the widest cleft in statistics' over issues of hypothesis-testing. But, as Louçã (2008: 4) has observed, in their vision of statistics as the language for a new form of science, they were in fact 'quite close'.

basis marked a scientific revolution has been amply justified by later work in the history of science. What has in fact become known as the 'probabilistic revolution' (Krüger, Daston and Heidelberger, 1987; Krüger, Gigerenzer and Morgan, 1987) is now widely recognised as one of the most – if not the most – momentous intellectual developments of the period in question. 'In 1800', to quote Hacking (1987: 52), 'we are in the deterministic world so aptly characterised by Laplace. By 1936 we are firmly in a world that is ultimately indeterminate . . . Chance, which, for Hume, was "nothing real" was, for von Neumann, perhaps the only reality.'[3] However, as Hacking goes on to stress (see also Hacking, 1990), it is important to see that complementary to 'the erosion of determinism' was 'the taming of chance': that is, the process of making chance and its consequences intelligible and manageable on the twin bases of assemblages of numerical data and the application of probability theory.

In the early stages of the probabilistic revolution, the social sciences did in fact play a leading part. In particular, Quetelet's application (1835/1842, 1846, 1869) of the Gaussian 'error curve' – or the normal distribution – to the display of regularities in the 'moral statistics' of marriage, illegitimacy, suicide and crime represented a pioneering attempt to show how a higher-level probabilistic order could emerge from out of individual actions that were generally supposed to be non-deterministic in character, or, that is, to express individual will and choice (Porter, 1986: chs 2, 6 esp.). And the notable development then was that the influence of Quetelet's work extended *from* the social *into* the natural sciences – somewhat ironically, given his great ambition to create a 'social physics'. As Krüger (1987: 80) has observed, at this point 'the familiar hierarchy of the disciplines' was inverted.

[3] Hacking is here referring to von Neumann's mathematical formulation of quantum theory. This aimed to preclude the possibility of 'hidden variables' that, if identified, would allow for phenomena that otherwise appeared probabilistic to be understood as deterministic – so that particles possessed a definite position and velocity at all times. For an accessible account, see Kumar (2008: ch. 14).

Most notably, Quetelet's use of the error curve provided a model for James Clerk Maxwell in his development of the kinetic theory of gases (Gillispie, 1963; Porter, 1982). Within a gas, the lower-level processes of colliding molecules were, in principle, subject to deterministic Newtonian laws; but the vast numbers of molecules involved meant that, in practice, a probabilistic treatment – 'statistical physics' – was required. In work carried out in the later 1860s, Maxwell took a version of the error curve to represent the *distribution* of molecular velocities within an ideal gas, so that, while nothing could be said about individual molecules, it became possible to calculate the *proportion* of molecules with velocities within a given range at any given temperature. Maxwell was generous in his acknowledgement of his borrowing in this regard from Quetelet and his followers. In speaking to the British Association for the Advancement of Science, he referred to physicists adopting a method of analysis new to them but which 'has long been in use in the section of Statistics' (cited in Gigerenzer et al., 1989: 62; see also Mahon, 2003: ch. 6).[4]

Subsequently, Fisher (1922), in seeking to integrate Mendelism into Darwin's evolutionary theory, adopted a model closely analogous to that which Maxwell had taken over from Quetelet, with biological populations corresponding to the populations of molecules. Under this model, natural selection could be seen as operating amid a multiplicity of random causes – any of which might have a predominant influence at the level of a particular individual – while, however, the probabilistic processes of natural selection remained the key determinants of the evolution of the population as a whole (Morrison, 2002).

In association with such developments, evolutionary biology, as Ernst Mayr (2001; see also 1982: ch. 2) has described, became the field in which the most explicit development of 'population thinking' occurred. In a scientific world dominated by physics and chemistry, what Mayr characterises as 'typological thinking' had prevailed,

[4] Ludwig Boltzmann, another pioneer of statistical physics, was also influenced by the work of Quetelet and his followers and expositors (Porter, 1986: 125–8).

centred on the properties of – and the deterministic laws applying to – entities of a supposedly homogenous rather than a 'pluralistic' kind, such as nuclear particles or chemical elements. But in evolutionary biology, increasing recognition came to be given to *the variation existing within* the entities under study – that is, to variation among the individuals making up a population – while at the same time interest focused on the probabilistic regularities that were still discernible amid this variation and on the processes, or mechanisms, through which these regularities were created.[5]

In contrast, in the social sciences, despite their influential role in the origins of the probabilistic revolution, there was a failure to exploit the possibilities that it opened up in both research and theory. Sociology at least (see Goldthorpe, 2007: vol. 2, chs 8 and 9) can still be regarded as not having a fully resolved relationship with this revolution and with the new ways of scientific thinking that it prompted.[6] Few sociologists today would believe that they should aim at formulating deterministic laws in the manner attempted by Comte, Spencer or Marx, intended to provide a comprehensive understanding of the structure, functioning and development of human societies. But for those who would still wish to maintain the idea of sociology as being, at least potentially, a science of some kind, the issue remains – and has been remarkably little addressed – of just what kind of science this might be. More specifically, if the search for deterministic laws in sociology is misconceived, then one may ask: to what objectives *is* sociological enquiry to be orientated and how is one to understand the rationale of the research activities that are carried out in their pursuit? As I indicated in the Introduction, the proposition that sociology should be understood as a population science is directed towards

[5] I am indebted to Yu Xie for drawing my attention to Mayr's remarkable work and to its relevance – as will later emerge – to current issues in sociology. See further Xie (2005).

[6] For interesting discussion of the – eventual – accommodation of economics to the probabilistic revolution, see the papers in Krüger, Gigerenzer and Morgan (1987: part III).

answering these questions, and at the same time towards situating sociology more securely within the probabilistic revolution.

What is implied, to put the matter at its broadest, is the following. The concern of sociology should be with populations or subpopulations of *Homo sapiens* (or better, perhaps – see Chapter 2 – of *Homo sapiens sapiens*) in their location in place and time; and the goal of sociological enquiry should be an understanding, not of the states and behaviour of the particular individual members of such populations in all their variability, but rather of the regularities that are the properties of these populations themselves, even though they are emergent only from the behaviour or – more precisely, as will later be argued – from the *actions* of their individual members.

To spell out more fully what is entailed by this primary proposition will be the task of the chapters that follow, each starting from its own subsidiary proposition. In conclusion of the present chapter, I add some further preliminary remarks concerning regularities in human populations and their determination and explanation. These may help to provide a context for the course of the subsequent argument and to signpost a number of major issues that will arise.

The regularities that can be identified in human populations, and more specifically in human social life, are diverse in their range and in their complexity. The regularities in 'moral statistics' to which Quetelet initially drew attention were relatively simple ones relating to the stability over time of rates of different kinds of individual actions and of their outcomes in national or regional populations. But Quetelet himself was forced eventually to recognise not only differences in such rates across these populations, but also significant differences *among their various subpopulations*; that is, among different groupings of individuals as defined in terms of age, gender, ethnicity, occupation and so on. And in this latter regard, he was then led to move on from essentially bivariate analyses to what can be recognised as early attempts at the multivariate analysis of social regularities of the kind that would be standard in present-day research (see esp. Quetelet, 1835/1842: part 3 with reference to crime rates).

In sociology today, the complexity of the regularities on which attention centres is of course often much further increased. For example, the concern may be not only with regularities expressed in the stability of particular forms of individual action and their outcomes within populations or in prevailing differences among populations or their subpopulations, but also with *regularities in changes* in these respects over time – and where time may be treated with reference to historical periods, the succession of birth cohorts or the individual life-course. Or the focus may be on regularities that exist between patterns of individual action and the locations of individuals in micro-, meso- or macro-level social contexts, as represented, say, by primary groups, social networks, associations and organisations, or by institutional and other variable aspects of the wider social structure. Or again, interest may lie in seeking regularities entirely at the supra-individual level: for example, among the structural features of 'total' – national or state – societies.

However, there are two further aspects of regularities in human social life that, while associated with their degree of complexity, are, for present purposes, of more direct relevance: what could be called their *visibility* and their *transparency*.

Consider the following case. There is a marked regularity in the number of individuals who drive their cars past my house between the hours of 7 and 9 a.m. on weekdays, and a regular and substantial decline in the numbers who do so on Saturdays and Sundays. These regularities would be apparent in their general form to any casual observer, and a standard traffic count would serve to establish them with some precision. Further, one could in this case readily construct a simple – although, as will later be seen, what could still be regarded as a paradigmatic – account of how these regularities are brought about. That is, a causal narrative couched in terms of individuals' ends – on weekdays, typically those of getting to work or ferrying children to school – and of the courses of action through which they then seek to achieve these ends, given the various constraints and opportunities that define their conditions of action. In short, the regularities in question could be regarded as being both highly visible *and* transparent. It is relatively easy both to see them and to 'see through' them; that is, to see through them to the social processes through which they are generated and sustained.

In contrast, though, the regularities that would be more typically the concern of sociology as a population science would be ones that are neither so readily visible nor, moreover, so transparent even when made visible; and wide-ranging implications then follow for the practice of sociology if understood in the way in question. Thus, to fulfil the first aim of a population science – that is, to empirically establish population regularities – will generally require, in the case of human societies, considerable effort in data collection and analysis. What is entailed is the design and application of research procedures capable of revealing aggregate-level regularities that were previously perhaps only vaguely sensed, if not entirely unrecognised, within the societies in which they operate. For example, to revert to the regularities that preoccupied Quetelet and his followers in rates of marriage, illegitimacy, suicide and crime, the possibility of establishing these regularities on any reliable basis only arose once national governments began to develop the apparatus of what would now be called 'official statistics', including population censuses and various registration systems.[7]

And to move on to the present day, one could say that the main scientific achievement of sociological research, as based on population surveys of differing design and the analysis of the data

[7] It was lack of data of the kind in question that can be seen as chiefly impeding the efforts of the British 'political arithmeticians' of the seventeenth and earlier eighteenth centuries, such as John Graunt, William Petty, Gregory King and Edmond Halley. In their pioneering efforts in the field of demography, broadly conceived, they were forced to work – although typically with great resource and ingenuity – on a miscellany of limited and frequently faulty data drawn from land surveys, tax returns, parish records of births and deaths, bills of mortality and so on. I learnt much about these early population scientists from the research of David Glass (1973) and later from that of Richard Stone (1997), being privileged to be a member in the 1960s of the Department of Applied Economics at Cambridge, which was largely Stone's creation.

produced, has so far been its demonstrated capacity to reveal population regularities in the more complex forms earlier referred to – regularities which, without the methodology in question, powerfully reinforced by increasing computing power for the purposes of both the storage and the analysis of data, simply could not have been accessed.

To illustrate here from my own field of research, although many other examples could readily be provided, a large number of studies have by now been undertaken into patterns and trends of intergenerational social mobility – studies characterised by growing conceptual sophistication. In particular, a crucial distinction has been made between absolute and relative mobility rates, with the former referring to mobility as actually experienced by individuals and the latter comparing the chances of individuals of different social origins attaining different class destinations (see e.g. Grusky and Hauser, 1984; Goldthorpe, 1987; Erikson and Goldthorpe, 1992; Breen, 2004; Ishida, 2008). The extensive work of data collection and conceptually informed statistical modelling involved has brought out both population regularities in, and also historically specific features of, the societies studied *of a kind that could not otherwise have been observed* – and certainly not by the 'lay members' of these societies in the course of their everyday lives, despite the close connection in fact existing between the regularities and specificities in question and their own life-chances and life-choices.[8]

[8] It is of course the case that in social mobility research, as in other fields, some amount of disagreement can arise over exactly what are the regularities in evidence: for example, over whether a long-term tendency can be observed for relative mobility rates to become more equal rather than fluctuating in a trendless fashion. But while such disagreements may figure rather prominently in the current research literature, this should not be allowed to detract from the important degree of consensus that is often in other respects established: for example, in the case of mobility research, on the fact that change in absolute rates is overwhelmingly driven by structural effects rather than by change in relative rates; or on the fact that where changes in relative rates do occur, whether directional or otherwise, they are usually very slow, in that they tend to result from cohort replacement effects far more than from period effects.

However, to return to my earlier distinction, making population regularities visible does not imply making them transparent; that is, it does not imply fulfilling the second aim of a population science: that of determining the processes – or, one could say, the causal mechanisms – through which regularities established at the aggregate level are produced at the individual level. In the case of sociology, this must mean demonstrating how these regularities derive ultimately from individual action and interaction. And it has to be acknowledged that, if sociology can by now claim some genuine success as a population science so far as revealing population regularities is concerned, its achievements to date in making these regularities transparent – that is, in accounting for them in the way indicated – have been a good deal less impressive. Regularities that may have been described in a quite detailed form often remain more or less opaque. Social mobility research would, unfortunately, again provide a good illustration of the point.[9]

Distinguishing between the dual tasks of a population science aimed at making population regularities first visible and then transparent – the one essentially a task of description, the other of explanation – is of key importance. This will become increasingly apparent as the argument of subsequent chapters proceeds.

[9] My own initial attempt at remedying the situation can be found in Goldthorpe, 2007: vol. 2, ch. 7, on which I hope to build in the course of research in which I am currently engaged.