Francia. Forschungen zur westeuropäischen Geschichte
Herausgegeben vom Deutschen Historischen Institut Paris
(Institut historique allemand)
Band 12 (1984)

DOI: 10.11588/fr.1984.0.51452

Zur Forschungsgeschichte und Methodendiskussion

Ferdinand Linthoe Næshagen

STATISTICS AND HISTORICAL RESEARCH*

1. Introduction: types of historical statements

The resentment many scholars feel against statistics [1], is not peculiar to history, it is also found in younger social sciences, such as psychology or sociology, and in philosophy[2]. As will be shown, the arguments against statistics in these sciences rest upon very shaky foundations and it may appear that the same is the case in history. It could well be that statistics in this case too will yield knowledge which is both more reliable and more intimate than that yielded by other methods.

It is often said that the insight gained cannot be any better than the theory as it is formulated. This is, I think, an overvaluation of the importance of theories. Rather than theories, we ought to see methods as the crucial factor in research, as Farrington seems to do when he says that »The true history of science – should be rather a history of method than of results, for the latter are often accidental and only seem impressive to later generations when they have been rediscovered by improved methods«[3]. Without suitable methods it is difficult to distinguish between good and bad theories, and progress will be slow and erratic, and if the methods are quite worthless, the theorising will be so too.

The point of departure should be the distinction between three kinds of statements, law, generalisation and singularisation: The law is a statement about a superpopulation, one which is not limited in time and space, such as ›the richer a country is, the greater the social inequalities‹. The generalisation is a statement about a population which is limited in time and space, such as ›sixth century Galloroman clergymen had a greater chance of becoming bishops than German clergymen had‹. The singularisation is a statement about one unit, such as ›the pretender Gundovald believed that he was the son of king Chlotachar‹.

2. Generalisations

No matter how it may appear to the philosopher, to the researcher generalisations are antecedent to laws and singularisations as they represent a bit of reality which is so small that he is able to investigate it, and yet so large that he can see the pattern in it.

These statements about limited populations all tell how one or more characteristics are

* I gratefully acknowledge the comments from and discussions with H. J. Eysenck, Baruch Fischhoff, Lars Hamre, Halvor Kjellberg, Hans Olav Egede Larssen, Geirr Leistad, Erik Næshagen, J. R. Ravetz, Sigurd Skirbekk and Finn Tschudi.

1 For instance Jacques Barzun, Clio and the Doctors, Chicago (The University of Chicago Press) 1974.

2 A list of suitable invectives is found in Paul E. Meehl, Clinical Versus Statistical Prediction, Minneapolis (The University of Minnesota Press) 1954, p. 4–6.

3 Benjamin Farrington, Science in Antiquity, Oxford (Oxford University Press) 1979, p. 31.

distributed over the units in the population, or, to put it in another way, how the units are distributed over the values of one or more variables. This is the case even if the statement is as imprecise as ›most Norwegians like to read about history‹ or ›the richer a present day Norwegian city, the greater the social inequalities‹. Unless such statements about units and variables are quite meaningless, they claim that the units are distributed in certain ways, and by that they deny that they are distributed in certain others. This means that they may be accepted if the distributions are such, and must be rejected if they are different.

As the way in which the units are distributed, will decide whether to accept or reject, a correct picture of the distribution is needed. This means that data have to be classified correctly, and they have to be collected and treated in such a way that we do not skip or lose data and for that reason end up with a biased picture. If, for instance, the researcher wants to test the statement about cities, wealth and social inequality, he will have to gather information about all Norwegian cities, or a representative sample of them, in order to see how many end up in cell one: »rich, great inequality«, cell two: »rich, small inequality«, cell three: »poor, great inequality« and cell four: »poor, small inequality«. Only when this is done, can he decide whether the majority is found in cell one and four, and therefore accept the statement, or cell two and three, and therefore reject it.

Statements about causes are no different, they can all be »translated« into distributions, such as when we restate »an unhappy childhood leads to alcoholism« as »there are more alcoholists among people who have had an unhappy childhood, than among those who have had a happy one«. That such correlations sometimes are spurious, meaning that the variables appear to be connected because both are determined by a third variable, makes no difference. On the contrary statistics is the best means to discover whether such correlations are spurious or not[4], and for instance enable us to reformulate the generalisation as »alcoholism as well as unhappiness in childhood are both due to a genetic disposition to mental instability« or »more alcoholics and people with an unhappy childhood are found among those who have a genetic disposition to mental instability than among those who have not«.

Whether the distribution really is such as the statement claims, may be investigated with as simple methods as the one described above, or more sophisticated ones, but nothing better can be done than collecting data about the units, counting and computing them, whether one does it in the head or with pencil and paper. No argumentation for or against a generalisation, for instance arguing from rational grounds that rich cities must show greater inequality, can carry any weight compared to an investigation of what the distribution really is like. It could of course be, as Smedslund says[5], that a logical analysis would show a generalisation to be self-evident, but whenever there is disagreement, it obviously is not, and then only a statistical investigation can be the referee[6].

In this respect there is no difference between the social sciences, including history, and the natural sciences. Habermas's claim, that the natural sciences count and compute while the social sciences ought to use ›Verstehen‹ or hermeneutics[7], cannot be taken seriously. Even if researchers had decided to study explicit norms only, not actual behaviour, they could hardly get along without the use of statistics. Written norms, laws, can only give a fragmentary picture

4 On these, see chapter 6.

5 Jan Smedslund, Analysing the Primary Code: From Empiricism to Apriorism, in: D: R. Olson (ed.): The Social Foundations of Language and Thought. Essays in Honor of Jerome S. Bruner, New York 1980, p. 47–73.

6 This is not to say that goodness of fit or accuracy as Kuhn calls it, is the only type of criterium which is used. In the essay »Objectivity, Value Judgment and Theory Choice« (p. 320–39 in Kuhn, The Essential Tension, Chicago and London 1977) he also mentions consistency, scope, simplicity and fruitfulness but he regards accuracy as »the most nearly decisive« (p. 323).

7 Jürgen Habermas, Technik und Wissenschaft als Ideologie, the essay »Erkenntnis und Interesse«, Frankfurt (Suhrkamp Verlag) 1968.

of a society as they only cover part of human behaviour, and the picture may very well be misleading as laws very well could have been empty declarations[8]. So may the unwritten norms, and these can, besides, only be charted through a statistical investigation which would show to what extent they were accepted in given regions, given periods and given social groups, and in many cases we would wish to know more than that, for instance how views of norms were interconnected and connected with the social characteristics of the proponents. This cannot be satisfactorily done without something very like an ordinary social science survey, no single person could give us sufficient and reliable information.

The next choice, between head statistics or paper-and-pencil-statistics and its more advanced descendants (both of which I shall just call statistics in the following), cannot be hard either, for estimates of the distribution, as a percentage, as a probability or as a statement with expressions like »many« or »most«, all presuppose some kind of memory. The researcher has to »remember« the units he has investigated, and as medieval scribes knew, human memory is unreliable and needs to be supported by writing[9]. That people often miss badly when estimating distributions, has now been so thoroughly proved by psychological laboratory investigations that the researchers have put this question on the shelf, and instead try to find out what kind of mistakes people make[10].

It is often said, both by science theorists and laymen, such as for instance the test subjects mentioned above, that the human brain is more competent than for instance the computer in dealing with more complicated, multivariate and non-linear distributions, as when for instance Liam Hudson says: »Although the prospect of handling more than four or five such variables in combination is statistically daunting, there is no reason why combinatory patterns should not be a standard feature of the mind's operation«[11]. The truth is the opposite, the »data processing« taking place in people's heads may be copied with fairly simple models; as a rule one indicator can explain forty percent of the variance and three indicators eighty percent[12]. This is simple stuff for a microcomputer[13], it could very well have been done with pencil and paper, and using such gadgets we would also avoid most of the systematic and unsystematic errors which creep in

8 We do know something about which types of norms are less likely to be broken or disregarded. Clearly specified norms are less likely to be broken than vague ones, those which apply (and therefore only can be broken) in the presence of a large number of people, are less likely to be broken than those which apply in the presence of a small number or none. For such reasons norms about legal and administrative procedure are less likely to be broken than others, and may therefore give the historian reliable background knowledge.

9 The memory formula in diplomatics, for instance archbishop Aslak Bolt in 1429: *Sakar pes at væroldin ok all værolzlikin ping æru forgangelik ok menniskionna minne er brioskelikt af þy at alt minnaz ok eingte forgløma er Gudz signadha nadh ok ey menniskionna snieldheit* (Because the world and all worldly things are transitory and human memory is weak, because to remember all and forget nothing belongs to God's grace and not to human cleverness, in: Diplomatarium Norvegicum V no. 586).

10 A concise introduction to this field, decision theory, in Paul Slovic, Baruch Fischhoff and Sarah Lichtenstein, Behavioral Decision Theory, in: Annual Review of Psychology 28 (1977) p. 1–39.

11 Liam Hudson, Human Beings, St. Albans (Triad/Paladin Books) 1975, p. 29–30.

12 Paul Slovic and Sarah Lichtenstein, Comparison of Bayesian and Regression Approaches to the Study of Information Processing in Judgment, in Organizational Behavior and Human Performance 6 (1971) p. 649–744. About the overvaluation of human ability Shepard says: »Possibly our feeling that we can take account of a host of different factors comes about because, although we remember that at some time or other we have attended to each of the different factors, we fail to notice that it is seldom more than one or two that we consider at any one time« (Roger N. Shepard, On subjectively optimum selection among multiattribute alternatives, in: Maynard W. Shelly II and Glenn L. Bryan (eds.), Human Judgment and Optimality, New York, Wiley, 1964, p. 266).

13 George L. Haller states that in comparison to a computer man can repair himself, he can program himself, he can adapt his program to unexpected information, his memory capacity is many orders of magnitude greater than the computer's, his logical sophistication is many orders of magnitude greater

when doing it in the head. There is no call for non-linearity either, for a simple linear model will do the task better than anyone can do in his head[14].

In cases where the researcher's data seem subtle and hard to grasp, many theorists and laymen would probably advise him not to use pencil and paper. This is what Ottar Dahl must have had in mind when he claimed that »Statistical comparison comes in when one has a large number of simple and quantitatively comparable phenomena such as for instance births or deaths, emigration, incomes, prices et cetera«[15]. This may also be one of the reasons for Aubert's proposed soft data sociology[16], but if so, he is just as wrong as the others, laymen as well as theorists, for »One type of error in self-insight has emerged in all these studies. Judges (i. e. test subjects) strongly overestimate the importance they place on minor cues (i. e. their subjective weights greatly exceed the computed weights for these cues) and they underestimate their reliance on a few major variables«[17]. The belief that human thought is all that subtle, is in other words due to a lack of insight into one's own thought, and it is also wrong that statistical methods are incapable of dealing with such data. Contrary to what Fischer seems to think[18], and many others as well, it has proved possible to collect valid and reliable data, even when they seem quite subtle and exalted, for instance about human values, happiness, love and beauty[19]. The conclusion to be drawn is therefore the opposite of what many theorists and laymen have thought, for if subtle and vague data are used without the aid of statistics, another source of error has been added to that of the data. There is certainly no reason to believe that one source of error cancels the other.

If, however, the data represent a biased sample of the population we want to generalise about, as often is the case in historical research, statistical methods will be less reliable and might even be inferior to for instance conclusions from rational grounds, unless methods can be found which make it possible to correct for such bias. This, which I see as the historian's main problem, deserves separate treatment and will be dealt with below[20].

than its, he has a variety of input-output devices. On the other hand, man is subject to fatigue and distraction, he requires motivation, his access to his memory is unreliable, his logical processes are slow and notoriously unreliable, he is unable to reproduce on demand most of the logical steps in his processing of information because he is unaware of them, his readin-readout processes are several orders of magnitude slower than his logical processes and his input-output devices (specifically language, the most important) are inexact, and therefore subject to misinterpretation. (George L. Haller, Our State of Mind in 2012 A. D., in: Proceedings of the I. R. E. 1962, p. 624–27). It ought not to be forgotten that computers today can play poker, play chess with an Elo-rating of 2100, beat the world backgammon champion, have qualified as a specialist in internal medicine and can diagnose mental illness better than the average psychiatrist.

14 Robyn M. Dawes and Bernard Corrigan, Linear Models in Decision Making, in: Psychological Bulletin 81 (1974) p. 95–106.

15 Ottar Dahl, Grunntrekk i historieforskningens metodelære, Oslo (Universitetsforlaget) 1967, p. 117.

16 Vilhelm Aubert, Det skjulte samfunn, Oslo (Pax) 1969. The essay in question »Om metoder og teori i sosiologien« (p. 192–224) is not printed in the English original version »The Hidden Society«. Another reason is the insight he believes to be found in everyday thinking (on this, see chapter 3). Aubert, however, is no enemy of statistics, but he has, it seems to me, been led by his customary tolerance into some quite uncalled for bridge building between useful and useless methods.

17 Slovic and Lichtenstein (ut supra n. 12) p. 684.

18 David Hackett Fischer, Historians' Fallacies, New York (Harper & Row) 1970, p. 90–94.

19 Values: Milton Rokeach, The Nature of Human Values, New York (The Free Press) 1973. Happiness: Frank M. Andrews and Stephen B. Withey, Social Indicators of Well-Being, New York (Plenum Press) 1976. Love: Glenn D. Wilson et David Nias, Le Charme a ses raisons, Paris (Tchou) 1977 (The English original version, not available to me, is entitled: Love's Mysteries). Beauty: Hans J. Eysenck, Aesthetic Preferences and Individual Differences, in: David O'Hare (ed.), Psychology and the Arts, Sussex 1981, p. 76–101.

20 Chapter 5.

That we are not justified in speaking of objective data[21], is irrelevant to the question of statistics as it concerns all data without exception. What is important to the researcher, is that if he proceeds in one way, his data will be fairly intra- and intersubjective, meaning that he will classify the same data in the same way each time, and that other researchers also will classify them in the same way. If he proceeds in another way, if he for instance does what Skjervheim[22] and many other philosophers want him to, and uses ›Verstehen‹ instead of pencil and paper and a written coding instruction, his data will not[23], and then data will not even satisfy the minimum requirements. For the same reason the objection we find in Habermas's not very clear exposition, that the researcher builds on a previous understanding[24], is irrelevant; he probably does so, but the distributions have to be checked in any case.

It is also worth noticing that when the researcher uses statistics, he is led to giving a more detailed description of what he does. This is of considerable importance, for as Fischhoff has said[25], informal models give considerable free scope for confusion, contradiction and seeing things which are not there, compared to formal models which enable the researcher and his critics to discover such sources of error. In this way informal models come to be much easier to save from collisions with contrary reality, or, in other words, much less falsifiable, and the more so because the head statistics they often go together with, contribute to this latitude[26].

It is also worth noticing that pencil-and-paper statistics or more modern procedures are used more and more in applied social science, that is for tasks where criteria for failure or success often are obvious and where it therefore is important to be right rather than being thought right[27]. Contrary to what Skjervheim says, criteria such as control and, in particular, prediction must carry considerable weight. He may be right in saying that »it seems just as important simply to understand man as being able to control him«[28], but when one set of methods enable a researcher to give far better predictions than others, it is difficult to avoid concluding that they also enable him to understand better. This will be seen even clearer from the next chapter.

21 Hans Skjervheim, Objectivism and the Study of Man, Oslo (Universitetsforlaget) 1959.
22 Ut supra.
23 See below, chapter 3, in particular the paragraph on bootstrapping.
24 Ut supra (n. 7). Habermas may be regarded as typical of the opponents of statistics: No research is discussed, he relies entirely on (a very verbose) argument.
25 Baruch Fischhoff, For those condemned to study the Past: Reflections on Historical Judgment (draft, reprinted in a somewhat revised form in: New Directions for Methodology of Social and Behavioral Science 1980, no 4 [Fallible Judgment in Behavioral Research], Jossey-Bass, San Francisco). Fischhoff here (in the draft) refers to James S. Coleman, The Mathematical Study of Small Groups, in: Herbert Solomon (ed.), Mathematical thinking on Measurement of Behavior, Glencoe, Illinois 1960, p. 1–149, and to R. J. Harris, The Uncertain Connection between Verbal Theories and Research Hypotheses in Social Psychology, in: Journal of Experimental Social Psychology 12 (1976) p. 210–19.
26 I have the impression that there is less outright cheating in non-quantitative research, and take the reason to be that the latitude given by non-quantitative procedures enables the researcher to have it his own way without cheating, that the falsifiability, in other words, is low. Thus if I play patience, for instance syveren (the sevens), and permit myself to put any card, not just kings, in a vacant row, the game will come out nine times out of ten and there is no reason to cheat. In more exact sciences, on the other hand, dirty work has been going on ever since Galileo and Newton tried their hands at it. Fabrications may however, like those of Sir Cyril Burt or Gregor Mendel, be exposed through a statistical analysis. On this and related subjects, see Michael Mahoney, Scientist as Subject: The Psychological Imperative, Cambridge, Mass. (Ballinger) 1976.
27 As for instance in the military. On this, see Peter Watson. War on the Mind, Harmondsworth (Penguin Books) 1980. Michael Inbar's argument for extensive use of computers in bureaucratic decision making (Inbar, Bureaucratic Decision Making, Beverly Hills/London 1979) is based on similar considerations.
28 Ut supra (n. 21) p. 31.

3. Singularisations

Statements about single cases are by no means peculiar to history, practitioners of the younger social sciences, and psychologists in particular, often make them when they for instance predict whether a psychotic patient can be let out of the asylum, a prisoner can be let out on parole or an applicant will do well in a university study or a job. So do many laymen when for instance a stockbroker decides whether to buy or sell, a physician whether to operate or to medicate or an employer whether to hire a prospective employee or not.

It is not hard to find a suitable test for one kind of singularisations, namely predictions. If the prediction comes true, it is good, and if it does not, it is bad. This, in its turn, gives us a suitable test for the methods which may be used: thus, the methods which give the greatest number of predictions coming true, hits, are obviously the best methods.

It is easy to see one way in which predictions and postdictions or explanations differ; in the first case the causal connections are drawn forwards in time, and in the second backwards. It is, however, less easy to see how this can matter, how it can make it necessary to use different methods (and indeed why conclusions about predictive methods should not be true about postdictive ones). What rather seems to be the important difference, is that those who postdict already have all the data, and for that reason lack such an obviously valid and efficient criterium for choosing between methods as those have who predict. To shelve the question of experimental testing of methods for postdictions would nevertheless be a mistake, showing insufficient knowledge of the problems modern social scientists, in particular psychologists, can solve. There are more criteria for validation than just prediction, and psychologists have cracked some very hard experimental nuts[29]. Just to mention some of the possibilities, studies of wisdom after the event, following Fischhoff's lead, will obviously throw light upon an important problem in postdictions[30], and the question of psychological insight might be answered by studies such as Smith's in which test subjects are asked to fit disconnected bits of biographical information into its true context[31]. Just to study the reliability, to see whether researchers using one set of methods arrive at the same results, would be even simpler, and might in some cases be enough, for if not, the methods do not even satisfy the minimum requirements. This is so because the validity must be low whenever the reliability is, precisely as in target practice where the number of hits must be low whenever the spread is great.

It would have been easier to give an answer to the question if we had known more about the way historians think, and in particular what kind of criteria they use when deciding whether to accept or reject a hypothesis. Quite possibly they sometimes use one set, at other times another, but it is far from easy to distinguish between them. The theorists are satisfied with describing the procedures in the abstract, and historians say little about this aspect of their work.

As may be seen from the theoretical discussion[32], the criteria seem to be of two kinds, they are either taken from logic or from experience. Thus, drawing conclusions from what is rational, like for instance Dray[33] in a procedure which seems to have much in common with Smedslund's

29 On criteria for validation, see n. 72. For hard experimental nuts, see for instance S. E. Asch, Effects of Group Pressure upon the Modification and Distortion of Judgment, in: Harold Guetzkow (ed.), Groups, Leadership and Men, Pittsburgh 1951, p. 170–90; S. Milgram, Some Conditions of Obedience and Disobedience to Authority, in: I. D. Steiner and M. Fishbein (eds.), Current Studies in Social Psychology, New York 1965, p. 243–62.

30 Baruch Fischhoff, The Silly Certainty of Hindsight, in: Psychology Today, 1975, p. 71–76.

31 Henry Clay Smith, Sensitivity Training: The Scientific Understanding of Individuals, New York (McGraw-Hill) 1973.

32 For instance Patrick Gardiner (ed.), The Philosophy of History, Oxford (Oxford University Press) 1974.

33 William Dray, The Historical Explanation of Actions Reconsidered, in: Gardiner (ut supra) p. 66–89.

apriorism[34], historians may conclude, as Lars Hamre does (in a discussion), that archbishop Olav Engelbriktsson's purpose with the troops sent to Bergenhus and Akershus in 1536 cannot have been to conquer these castles, no rational man would have tried that with so small forces, but to establish bridgeheads, and that can only have been because he expected the forces of the Count Palatine. It is easy to see how this could have been put in a syllogistic form (which to cover all alternatives would have to be more developed than this sketch). Using the other set of criteria, the same conclusion might also have been reached by way of a law or generalisation based on experience[35], stating that prelates hardly ever take military action unless they are fairly sure of having superior forces, that the archbishop's own forces obviously were inferior, and for that reason reinforcements (from the Count Palatine) must have been expected.

There are certainly cases, like that above, when aprioristic conclusions, in which other cases are not taken into account, seem more credible than empiristic ones, in which they are. All the same, rational man, which at first sight might appear to be a concept of the same kind as economic man, is in fact charged with many more tasks, and is for that reason unlikely to serve well enough. It is, for one thing, obvious that the concept cannot always be used; though it seems unlikely, the archbishop may have sent off these expeditions as an irrational reaction to Christian III's threat against his faith and his fatherland, which means that the researcher has to choose between these two alternatives, and needs criteria to do so. Furthermore, when action is not wholly (or partly) irrational, it can hardly be modelled without a weighing of chosen values (motives) and sizing up of perceived probabilities which presupposes more knowledge than the historian usually has[36]. Thus the archbishop may for instance have felt that his honour required him to fight no matter what the outcome would be, or he may have had reason to hope for traitors in the garrisons. Logic alone cannot show what weight to give these alternatives, and then there is nothing to go by except a feeling that one alternative must be the right one, an aha-experience which very well could have been due to some prejudice[37].

In order to decide whether our singularisation is acceptable (as well as to decide whether our methods are), we shall need some sort of feedback, and it is difficult to see how we can have this except by a statistical investigation. In this respect there is no difference between postdictions and predictions, because the hypothesis that John Doe will commit larceny because of his unhappy childhood, is not shown to be acceptable just by the fact that he does commit larceny. Neither this nor any other of his many characteristics come out with the label »Cause of Larceny«, and if we know nothing more, any one of them, his asthma, his nagging wife, the loss of his driver's licence or the resentment he feels against the world in general, could almost[38] equally well have been the cause, and we could equally well have made out a logical case for any one of them. In other words, if we take into consideration just this one case, it is hard to see how

34 Ut supra n. 5. But Smedslund, who is one of the pioneers in decision research (see his »The Concept of Correlation in Adults«, in: Scandinavian Journal of Psychology 4, 1963, p. 165–73) probably does not wish to grant his apriorism such a wide field.

35 Carl G. Hempel's view. See for instance his »Reasons and Covering Laws in Historical Explanation«, in: GARDINER (ut supra n. 32) p. 90–105.

36 While the historian, for instance the medievist, probably needs more knowledge than the student of contemporary society does, seeing that the latter has a lead in his fuller and more detailed background knowledge (see n. 96).

37 What Finn Tschudi calls the interocular traumatic test (the truth of it is felt as a blow between the eyes).

38 But not quite. There is at least some little difference between the hypothesis which has passed this very simple test, and those which have not. Those who postdict have no such test and therefore seem to be exposed to the same temptations as those who sell elastic band by the yard, a task which according to the Danish Storm P. takes very great strength of character. For empiric studies of the postdictor's temptations, see Baruch FISCHHOFF (ut supra n. 25) and Baruch FISCHHOFF and R. BEYTH, ›I Knew It Would Happen‹ – Remembered Probabilities of Once-Future Things, in: Organizational Behavior and Human Performance 13 (1975) p. 1–16.

any of the alternative explanations can be falsified. As the psychological studies referred to below bear out, whether we predict or postdict, only a statistical pattern, like the iron filings showing magnetic lines, can tell us what we need to know about the connections between larceny and the alternative explanations, and enable us to choose the most likely to be true.

In any case, if we take a look at what historians do, we hardly ever find them using just one of these methods in an explicit formal model. What they do, whether in an article or a book length monograph, seems to be what philosophers call hermeneutics[39], going into the subject step by step, sometimes referring to a priori reasoning, sometimes to empiric regularities. The counterpart of this in psychological investigations would be the unstructured interview in which the interviewer probes and follows up significant indications by further questions.

Such procedures seem to offer the researcher considerable freedom; he is not bound to follow a preconceived model which may make him overlook important data which do not fit into it. On the other hand such procedures will make it more difficult to discover non sequiturs and internal contradictions which are unlikely to be less frequent in historical research than in sociology or psychology where in fact »several recent case studies have shown that when their verbally stated assumptions are formalized, some of our popular and accepted theories can be shown to contain internal contradictions«[40].

Furthermore, while historians' reasoning often seems to be aprioristic, it also often refers to or seems to imply regularities, and these can hardly ever have been investigated with statistical methods, but with the head statistics I mentioned above[41].

The historian is, in fact, facing the same choice as the psychologist is when he has to choose between what he calls statistical and clinical methods. This choice is an important one to the psychologist who knows very well that a wrong bit of advice, based on a wrong prediction, may lead to considerable unpleasantness and sometimes even misery for those concerned. The first to deal with this subject was Paul E. Meehl, and his review of some twenty studies did not show a single case where psychologists using informal, clinical methods were better at predicting than those using formal, statistical ones[42]. Sawyer has later reviewed forty-five more studies covering a great variety of tasks, and came to the same unequivocal conclusion[43], and Dawes sums up the present situation by saying that »nothing published in the standard journals after Sawyer's review has demonstrated that a context exists in which clinical prediction is superior«[44]. As for the method preferred by clinical psychologists, and which most resembles those used by historians, the unstructured interview, it has, in spite of (or perhaps because of)[45] the large quantity of information it yields, a »near total lack of validity«[46].

39 For instance Hans-Georg Gadamer, Wahrheit und Methode, Tübingen (Mohr/Paul Siebeck) 1975. Hermeneutic validation may appear similar to Cronbach and Meehl's construct validation (see n. 72), but is basically different in that the network is established not by statistical distributions but by the aha-experiences mentioned above.

40 Fischhoff (ut supra n. 25).

41 Chapter 2.

42 Ut supra (n. 2).

43 Jack Sawyer, Measurement and Prediction, Clinical and Statistical, in: Psychological Bulletin 66 (1966) p. 178–200.

44 Robyn M. Dawes, »You Can't Systematize Judgment: Dyslexia«, in: Fallible Judgment (ut supra n. 25) p. 67–78.

45 There are strong indications of this: »Clinical combination actually predicts less well with data collected by both modes (mechanical [i. e. statistical] and clinical) than with only mechanically collected data, and clinical combination that includes a mechanical prediction is inferior to the mechanical composite alone« (Sawyer, ut supra n. 43, p. 192).

46 Dawes (ut supra n. 44), referring to E. L. Kelley's still unchallenged article »Evaluation of the Interview as a Selection Technique«, in: Proceedings of the 1953 International Conference on Testing Problems,

The studies of psychologists predicting are by no means the only ones which strongly indicate the futility of the alternatives to statistics. Information-processing deficiencies, getting less, and less reliable, information out of the data than a statistical model can give, have been shown time and again when persons from many professions or trades have been studied on the job; this has caused considerable concern, and some would now put these deficiencies »on a par with motivational conflicts as causes of the ills that plague humanity«[47]. In another remarkable kind of studies, the so called bootstrapping, persons from several groups, say physicians or businessmen, have been studied, and by statistics a model of their decisions is made. In this model we shall of course still find the systematic errors (since it copies the person in question), but the shortcomings as regards reliability, the random errors or »noise«, due to absentmindedness, lack of consequence and similar causes, will be eliminated, and by that means alone a fairly simple model will do a better job of predicting and making decisions than the person it copies[48].

There are, however, some cases where head statistics undoubtedly works, cases, that is, in which the statistics is quite obvious. Thus, when we conclude that the Samson Filippusson met with on Lista in 1421, is the same as Sir Eindride Erlendsson's bailiff in Ryfylke in 1440, we do so from a, possibly unconscious, computation of probabilities, as can be seen if we figure out what we would have concluded if the man's name had been the more ordinary Sigurd Olavsson[49]. Conclusions based upon language are of the same kind, such as when verbal similarities make us conclude that there must be some connection between the privilege conceded to the Hanseatic towns by king Magnus Håkonsson of Norway in 1278 and the privilege conceded to Wismar by king Kristoffer of Denmark in 1323[50]. Nevertheless, when names, for instance, have not been as striking and unusual as that of Samson Filippusson, researchers have often overlooked possible linkages or made wrong linkages. Besides, while conclusions of this kind are indispensable to research, they do not often answer what researchers think of as the important questions.

The psychological studies mentioned above seem to be based on the assumption that there is just one alternative to statistics, and what the psychologists certainly have in mind, is head statistics[51]. For that reason we do not know to what extent the inferior performance of the clinicians is due to logical shortcomings and to what extent it is due to biased estimates. If we had tried to prevent the former by using formal models, for instance the procedure Loh describes in

Princeton 1954 (not available to me). – This does not prevent the use of unstructured interviews, nor have investigations showing the uselessness of various other approaches, like Freudian psychology (on this see EYSENCK and WILSON, ut supra n. 102), put a stop to them. About this Meehl says: »Personally, I find the cultural lag between what the published research shows and what clinicians persist in claiming to do with their favourite devices even more disheartening than the adverse evidence itself« (Paul E. MEEHL, The cognitive activity of the clinician, in: American Psychologist 15, 1960, p. 19–27). One reason for this may be the strength of folk science (for this very fruitful concept, see J. R. RAVETZ, Scientific Knowledge and its Social Problems, Oxford 1971. But in any case, how many will say: I am sorry, but what I have been doing before, is just crap.

47 SLOVIC, FISCHHOFF and LICHTENSTEIN (ut supra n. 10) p. 3–4.

48 SLOVIC and LICHTENSTEIN (ut supra n. 12) p. 721–24. For a theory of bootstrapping see Finn TSCHUDI, Brunswik's lens model and multidimensional scaling, mimeograph, the Institute of Psychology, University of Oslo 1974.

49 Diplomatarium Norvegicum vol. III no. 656 and IV no. 877.

50 Diplomatarium Norvegicum vol. V no. 10 and Diplomatarium Danicum 2. series, vol. IX no. 5. In the opinion of my colleague Odd Sandaaker, who mentioned this, this may not be anything more exciting than a receiver's dictation.

51 It is for this reason that psychologists' studies of generalisations throw light upon singularisations and vice versa.

his very longwinded book[52], or Smedslund's theorem system[53], the performance would probably improve, but until it has been investigated, it is impossible to tell how great the gain would be.

Other methods have also been mentioned, such as ›Verstehen‹ or empathy which is what Sir Francis Galton practised so successfully when he entered so deeply into the mind of the paranoiac that he, after a short walk, felt that even the cabhorses spied on him, or when he hung up a picture of Punch, worshipped it, and gradually came to feel for it what the barbarian feels for his idol[54]. Empathy is a nice word, but it is hard to tell what it really stands for. It carries a hint of an extrasensory creeping inside somebody else's skin, but I doubt that many would confess to such an interpretation, and if it is not that, it must be either aprioristic reasoning or reference to experience or both.

In any case the high regard in which many hold this method, seems to stem from the high regard in which they hold everyday thinking, believing that the social scientist at his best is merely a little ahead of educated common sense[55]. This view has no foundation in reality. To be sure, people do learn and make use of their learning[56], but while their conceptual apparatus and way of thinking enables them to deal with their own everyday affairs, after a fashion, it cannot, as a rule, give more than a fragmentary and often biased view of relationships. The faith people often have in their own or somebody else's insight, is nothing much to go by. The studies referred to above, as well as Smith's monograph on the subject[57], show that as a rule this is an overestimation. For this not only their self-assertion but reality itself may be responsible, for instance when their self-fulfilling prophecies come true[58], and mistakes are covered up in many ways[59], for oneself as well as others. Nor do the prospects for training in understanding people seem very bright, at least with the methods in general use today; laymen who have taken part in various schemes, such as the fashionable T-groups, have no more insight after than they had

52 Werner Loh, Kombinatorische Systemtheorie: Evolution, Geschichte und logisch-mathematischer Grundlagenstreit, Frankfurt (Campus Verlag) 1980.

53 Smedslund (ut supra n. 5).

54 Derek William Forrest, Francis Galton. The life and work of a Victorian genius, London 1974.

55 Skjervheim (ut supra n. 21) p. 76; Aubert (ut supra n. 16) p. 220.

56 There are certainly situations where feedback and redundancy make efficient learning possible (see Robin M. Hogarth, Beyond Discrete Biases: Functional and Dysfunctional Aspects of Judgmental Heuristics, in: Psychological Bulletin 90, 1981, p. 197–217). Though true for, for instance, hedge clipping, it is unlikely to be so for research as well as many everyday activities (see for instance the paragraph about bootstrapping).

57 Ut supra, n. 31.

58 Hillel J. Einhorn, Overconfidence in Judgment; in: Fallible Judgment (ut supra n. 25) p. 1–16.

59 This description of methods for covering up is taken from psychological practice but will mutatis mutandis be valid for historians as well. They are: Selectiveness, remembering what fits in with one's beliefs and forgetting what does not. Universality, making statements likely to be right for almost everyone, saying for instance that »you get depressed from time to time«, just like fortune tellers often do with much success. Reversibility, making statements with a reservation so that they will be true no matter what the outcome is, saying for instance »this man has a tendency to act out, but will probably be held in check by his dependency needs«. Partly cloudyness, making statements so imprecise that they will cover almost all eventualities, for instance saying »the weather will be partly cloudy« which will be untrue only if the sky is completely blue or not a blue spot is seen all day. Fuzziness, using concepts which cannot be checked because they have no definite meaning, saying for instance that someone is »sincere« or »phony« or »a latent homosexual«. Smith (ut supra n. 31), p. 214–15, summing up from L. A. Rorer, The proper domain of prediction, unpublished paper, Oregon Research Institute 1965 (not available to me). – The statements described above are in fact ways of saying very little while appearing to say very much. So are statements like »Some A's are B«, which do not tell us anything of the

before, although they generally believe so[60], and clinical psychologists are no better than undergraduates or other professionals[61]. It seems that the natural way of thinking, into which researchers as well as laymen easily slip back unless they make an effort to follow some sort of formal statistical procedure, is something very like magical or mythical thinking[62]. Just like in medieval humoral pathology connections are not established by what goes together or follows after one another but by what Shweder and D'Andrade call conceptual affiliations[63], like similarity and supraordination, as the medieval physician would reason that a fever, an excess of warmth, could be cured by a medicine made from a cold substance like henbane or mandrake root. Thus many of the connections laymen as well as many researchers take for granted, for instance that somebody with a low self esteem cannot be a leader, or that children who often seek attention from their mother, also are more likely to cling to her apron strings, are simply wrong; »broad, empirically homogeneous multi-item traits or syndromes (for example, extravert: likes parties, feels at ease talking before a group, introduces himself to strangers) are difficult to induce from behavior observational evidence –. Instead, one discovers that minor changes in context, task or response mode produce major changes in individual difference ranking –«[64]. In other words, the criteria which make historians accept or reject psychological explanations, may have much more in common with literary conventions than with psychological research.

It also happens that expressions such as »my experience tells me« or »in my experience« are used. This can only refer to some sort of statistic, and the relevant question would be: »Has your experience been processed with head statistics or paper-and-pencil statistics?«

4. Laws

There is some reason to believe that in the philosophy of history there is altogether too much talk of statements about populations which are not limited in time and space, historical laws[65]. Statements of this kind can hardly be anything else but statements about units we know about, plus a prognosis that the same will be the case with units we know nothing about. In this respect both past and future present a crux to the researcher; the future because we know nothing about it (except through guesses based on the assumption that we shall find the same constants and trends), and the past because we only have knowledge about bits of it, and these bits are unlikely to make up an unbiased picture.

relationship between A and B. To do that, the statements would have to be like »Unusually many (or few) A's are B«. To look for such statements in generalising historiography ought to be an enlightening and rewarding experience.

60 Smith (ut supra n. 31) p. 31–32.

61 Smith (ut supra n. 31) p. 34–36.

62 Richard A. Shweder, Likeness and likelihood in everyday thought: magical thinking in judgments about personality, in: P. N. Johnson-Laird and P. C. Wason (eds.), Thinking. Readings in Cognitive Science, Cambridge (University Press) 1977, p. 446–67.

63 Richard A. Shweder and Roy G. D'Andrade, The Systematic Distortion Hypothesis, in: Fallible Judgment (ut supra n. 25) p. 37–58.

64 Shweder and D'Andrade in: Fallible Judgment (ut supra n. 25) p. 47–48. This gives good reason to suspect the kind of historiography where psychological explanations play a large part. Unless the matter is quite trivial, how likely are historians to find tenable explanations when both their data and their methods are inferior to those of psychologists?

65 This may be seen from almost any anthology. In the philosophy of history it is, in fact, what a musician would have called one of the golden oldies.

Reasoning from statistical theory, we have to conclude that all sciences of experience are geographical and historical, as we only in this way can be sure that all types of units have an equal chance of being represented in the sample. When many natural sciences cut such statistical corners and disregard this claim, when merely the thought of studies such as »The Reaction between Oxygen and Hydrogen Atoms in Sixteenth Century Guatemala« makes us grin, this is not due to any theoretical reasoning but the fact that long experience has shown natural scientists that their material can be purified so that »le cas pur« can be studied (but insufficient purification often has given odd results[66], and may have shown geographical variation as for instance the sulphur content of copper ore varies from one place to another).

Like the proverbial master craftsman, the natural sciences may cut corners in this way and yet arrive at laws[67]. The social sciences can hardly do so; they cannot for instance purify a priest with fire and sulphur until he becomes just priest and nothing more, and the state of the sources makes it impossible to replicate them the way they should have been[68]. This, in its turn, means that it may be impossible ever to arrive at any conclusions which may be called social science laws, or historical laws if you prefer that term. This may not, however, be such a serious loss as it would appear from the great fuss philosophers of history make about laws[69]. The predictions social scientists make are merely based upon generalisations, such as when psychologists predict about a modern Norwegian without the slightest knowledge about Vikings or Bronze Age Chinese, and the actions of a medieval archbishop may equally well be explained from studies of other medieval archbishops without the slightest knowledge about archbishops in ancient Paphlagonia or modern England. There is no reason to think that the predictions or explanations would have hit the target more often if they had been based on laws.

5. Generalisations and problematic data

It happens to most researchers, at one time or another, that they are unable to find the kind of data they would have preferred to work with, but to those who study the distant past, historians, archeologists and others, it happens so often that it may be regarded as the ordinary state of affairs. These then have to use whatever data can be found, and the best of them have shown a lively inventiveness, not inferior to the best of any other group, in squeezing information out of data which apparently have no information to give.

It does of course happen that they neither can find data which certainly have a bearing upon their subject, nor data which possibly may have it, but there is no reason to give up after the first try. As they ought to know, something may turn up which makes it possible to use data which otherwise could not have been used, sometimes a method but just as often a theory which throws light on connections which have not previously been seen, shows us new valid indicators in other words. Thus Asgaut Steinnes has seen that farm names may carry information on

66 Arne Næss, The Pluralist and Possibilist Aspect of the Scientific Enterprise, Oslo/London (Universitetsforlaget/Allen & Unwin) 1972, p. 15.

67 On this, see Ravetz (ut supra n. 46).

68 It is only when the replications are random (see the chapter on sampling in any text book on statistics) that we can have full certainty for their representativeness. Practical reasons in most cases compel us to settle for less, like replications from what appears to be the most dissimilar societies or groups.

69 Exceptions are Olaf Helmer and Nicholas Rescher, On the Epistemology of the Inexact Sciences, in: Management Science 6 (1960) p. 25–40; Carey B. Joynt and Nicholas Rescher, On Explanation in History, in: Mind 1959, p. 383–88; Carey B. Joynt and Nicholas Rescher, The Problem of Uniqueness in History, in: History and Theory 1 (1961) p. 150–62.

Norway's ancient political history[70], and by an even more remarkable inspiration J. C. Russell has seen that the compass alignments of graves may carry information on early medieval plague mortality in Britain[71]. Such connections need not remain something which is justified by plausible argument only, in many cases they make it possible to see yet other connections, tying together many hypotheses and their data in a lattice or a nomological network[72] which gives us more reason to believe in them than any argument could.

Another problem, often found together with the first, is that data can only be found about a sample of the population; the researcher may find satisfactory data about some of the units, quite unsatisfactory ones or none at all about the rest. Since the sample is taken by historical processes, not by the historian, he cannot disregard the possibility that it may be biased, and that the conclusion he draws may be invalid.

It cannot be said that historians' practice shows any great awareness of this danger; they often seem to live in a statistical state of innocence, being worried about things there is no reason to worry about, and quite unworried when there is good reason to worry. They do, for instance, often make a fuss about the size of the sample, while there in fact is no use in ten thousand units if the sample is biased while ten might be enough if the sample is representative[73]. Quite often they also just compute from such samples, without either arguing for its representativeness or trying to correct for bias. At other times again they simply make a statement without saying how they have arrived at it, whether by apriorism or, as one quite often may suspect, by some sort of head statistics.

As for the latter method, it is obviously inferior to paper-and-pencil statistics any time, but the former need not be so when this statistics only has biased samples to work with, nor may a third method in which the hypothesis is tested against what contemporary observers directly or indirectly have said about the distributions.

Using this third method for instance to test a hypothesis about the previous religious status of early Merovingian bishops (whether laymen or clergymen) we try to find what contemporary observers such as Gregory of Tours or pope Gregory I have said about the subject[74]. If we want to discover the origin of the first Cathar missionaries, we might likewise try to find what is implied (indirectly said) when the term *bulgarus* is used meaning Cathar. Neither in these nor most similar cases can we have anything like a reasonable certainty that our informants have been able to make a correct estimate of the distribution; they seldom had anything to go by except head statistics. Thus a few dramatic cases may very well have shocked the two Gregories into believing that laymen often became bishops, and the accidental arrival of Bulgar missionaries in one district may have made people believe that this was the case elsewhere too, just as they may have believed that a great many of the Cathars (or Bulgarians) were sodomites, as we see from the English »bugger« (which is derived from *bulgarus*). Furthermore, while direct statements of the kind we want, are few (many are of the useless kind »the Goths fled as

70 Asgaut Steinnes, Husebyar, Oslo (Den norske historiske forening) 1955.

71 Josiah Cox Russell, The Earlier Medieval Plague in the British Isles, in: Viator 7 (1976) p. 65–78.

72 A more detailed description of that very fruitful concept is found in Lee Cronbach and Paul E. Meehl, Construct Validity in Psychological Tests, in: Psychological Bulletin 52 (1955) p. 281–302. Ravetz (ut supra, n. 46) is particularly interested in the diachronic aspect, but the synchronism of the nomological network would easily fit into his lattice.

73 See for instance Sidney Siegel, Nonparametric Statistics for the Behavioral Sciences, New York (McGraw-Hill) 1956, p. 96–104 on the Fisher test. The drawback of small samples is not that we see significant relationships which are not there, but that often we do not see significant relationships which are there. Furthermore, it may be impossible to discover whether these relationships are due to a third variable (spurious) or not.

74 Gregory of Tours' History of the Franks IV 46, and pope Gregory I's Registers vol. 7, part I: 368–71.

they usually do« as Gregory of Tours tells us[75]) it may be difficult to get a firm grip of indirect statements, which often are studied with a linguistic analysis rather more complicated than my example above. It must have appeared so to most historians too, for except for special purposes, like the study of ideology, such linguistic-historical analyses are mostly used when other sources are lacking, such as when studying German society outside the Roman Empire or the original Indo-European society[76]. In any case a source critical discussion of such statements must pay great attention to the question »what is the likelihood that the author of the statement has made a reasonably correct estimate of the distribution?«.

Data are needed whether the methods are aprioristic or statistical, and whenever there is reason to suspect biased data, the credibility of both decline. It could be that the credibility of statistics which follows data more closely, will decline more[77], but on the other hand statistics offers remedial measures which may not be found in aprioristic research:

One may, in the first place, try to discover whether in fact such a bias is to be found, for instance checking that there are no obvious discrepancies between proportions found in the sample and the proportions one could expect in the population. Thus sociologists for instance will check as a matter of routine that there is roughly the same number of men and women in their sample, and not more than twelve or thirteen percent with a university level education. Medievists, of course, know less about the proportions to be expected, but they do have some knowledge, for instance about the proportion of clergymen to be expected, or the proportions of bishops from various provinces, and will thus be able to correct for bias, the possible over-representation of clergymen or bishops from certain provinces, multiplying either by an average value or two limit values.

When the proportions in the population are unknown, it may still, in some cases, be possible to form a hypothesis about them, and test the hypothesis, much in the same way as historians habitually discuss source critical hypotheses together with the substantive hypotheses. Thus, if we want to find out whether the church was better able to adapt to the late medieval recession than other groups, we shall have to make a statistical investigation of who sold land to whom, where and when. But when data have been found, showing by districts and periods how much was bought and sold by churchmen, noblemen, townsmen and farmers, the true proportions in the population can only be inferred by also taking into account how the channels of transmission have worked. Information about sales has come to us through registers, letter books, seventeenth century court rolls and diplomas preserved in church archives, farm archives and others, sometimes through two channels and even three. Because of the many channels and the cases of double transmission, a hypothesis about the working of the channels may be tested, although, to be sure, such tests will have to build on more assumptions than tests of present day affairs would[78]. For other types of sources, other ways may be used: In the study of a medieval historian, one may, for instance, assume that a medieval historian will regard one kind of units, say persons or events, as more newsworthy than another, and test it by seeing whether the proportion of the latter kind decreases more quickly with increasing distance from his location in space and time, as one would expect from press research. When several sources

75 Gregory of Tours' History of the Franks II 37.

76 For instance Walter Schlesinger, Das Heerkönigtum, in: Das Königtum, seine geistigen und rechtlichen Grundlagen, Sigmaringen (Jan Thorbecke) 1956, p. 105–41, and Émile Benveniste, Le vocabulaire des institutions indo-européennes I–II, Paris 1969.

77 Though I do not think so. See n. 96.

78 But although considerable, the difference is one of degree, not of kind, for biased samples are far from unknown in research on present day affairs. Furthermore, with the building up of a nomological network (see n. 72), a theory may grow pretty strong.

can be found, it may also be possible to use methods like data quality control or meta-analysis[79] in which statements about the phenomenon studied are taken as dependent variables, and as the independent variables those qualities in the sources likely to make for good or bad reporting. In the long view systematic studies of the transmission of information will also yield useful information, as Tord Høivik has said[80]. Such studies might develop into generalised knowledge[81] as an historical auxiliary discipline, call it biaslogy[82] if you like.

Multiple testing will also give greater certainty than we could have by one test alone, just as we will have from replications[83], as for instance in studies of basic personality characteristics replicated in a number of countries[84]. Inventiveness as well as knowledge of the subject may show how to derive other testable hypotheses from those we want to test. In this way Zvi Razi has studied the population trends in the Halesowen district, not just by the number of men mentioned in his court rolls, but also by the number of women, taken separately, the replacement rates and the mortality[85]. He has thus shown that his hypotheses about the trend and about the representativeness are simpler and therefore more acceptable than the contradictory ones which could not claim that the trend was different without also claiming that the sources are biased, and would find it difficult to give a credible and simple explanation of the bias which also fitted the data.

Even more than in disciplines more fortunate in their data[86] a clear and consistent reasoning is needed to make up for the flaws, and there is no better way of ensuring this than by making the reasoning explicit and formal so as to deny non sequiturs and contradictions a hiding place.

All the same it cannot be denied that the historian studying, say, the middle ages seems to be in much the same situation as a certain gambler who was reproached by a friend because he had gambled and lost, and answered: »I knew that the game was crooked, but it was the only game in town.« Unless the medievist restricts himself to those few cases in which there is information about almost all units, say those districts from which church registers are preserved, or well documented groups as the aristocracy, there will always be a danger that the sources are biased, over-representing some groups and thus distorting the distributions. Remedial measures may be taken, but the danger will always be there.

79 Raoul Naroll, Data Quality Control – A New Research Technique, New York (The Free Press) 1962, and Idem, Some Thoughts on Comparative Method in Cultural Anthropology, in: Hubert M. Blalock and Ann B. Blalock (eds.), Methodology in Social Research, New York 1968, p. 236–77; Gene V. Glass, Barry McGaw and Mary Lee Smith, Meta-analysis in social research, Beverly Hills/London (Sage Publications) 1981.

80 In »Uvitenskap og metode«, in: Historisk tidsskrift (Oslɔ) 47 (1968) p. 213–19. But one kind of source does not just yield one kind of information but many, depending upon what we look for and what our point of departure is (see the beginning of this chapter). For that reason such a large scale charting as Høivik has in mind, will be impossible, but minor studies are not, nor is a rather more formal, external, description.

81 We for instance already know quite a bit about what medieval historians included and what they left out, and their principles seem to have been not very different from those of modern reporters.

82 As professor Henning Mørland tells me, bias is in the last instance derived from epikarsios, but very few would see the connection between bias and epikarsiology.

83 But see n. 68.

84 For instance Hans J. Eysenck and Glenn D. Wilson (eds.), The Psychological Basis of Ideology, Lancaster (MTP Press) 1978.

85 Zvi Razi, Life, Marriage and Death in a Medieval Parish, Cambridge (Cambridge University Press) 1980.

86 Fortunate in the sense of having more, and more representative, data, but many will probably feel that the medievist's problematic data make his task all the more exciting.

6. Statistics and the historical research tradition

Better methods for testing theories will of course also influence the research tradition itself. Some theorists, like Næss and Feyerabend[87], regard it as a sign of health that many schools of thought are found within a research tradition[88]. Others, like the sociologists Peter Rossi and Leonard Gordon, regard it as an unhealthy sign[89], and they seem to be right. To be sure, it cannot be denied that a theory which almost everybody has rejected, may turn out to be tenable, while a theory which almost everybody has accepted, may turn out to be untenable. If we, however, consider a shorter period of time than eternity, the picture becomes a different one, for it is clear that sometimes many man-labour years are spent on theories which nevertheless remain as sterile as they first were. Thus, while they ultimately may turn out to be fruitful, it is clear that they will have less to show for the efforts they consume than other theories. Many sudden, and, as it seems, unmotivated changes which we often find in the social sciences, should probably be seen as signs of such a wasteful state of affairs, showing that the theory testing mechanisms are unable to weed out those theories which are likely to remain sterile for the next fifty years or so from those which are not. It should not be thought that we cannot find criteria which enable us to discover some of these sterile theories. Ravetz, for instance, has mentioned several characteristics of one sterile school of thought[90] and others may no doubt be found. The use of such criteria may account for the differences found within any given research tradition, such as the sociological one, for in each case the statistical school of thought can show considerably greater constance than the other, a sign that its theory testing mechanisms work better.

One reason for the superiority of the statistical school of thought is fairly obvious. Although exaggerated precision merely seems pretentious, the low precision given by head statistics, as a rule expressed as »many« or »most«, will make it difficult to distinguish between spurious correlations and those which are genuine. For this at least trivariate analyses are necessary, which in a pencil-and-paper analysis would mean eight cells or more in the table, and could not be any simpler in a head analysis. In view of what we already know about man's capacity for this kind of task, it is inconceivable that the head statistician in this way can keep count of the proportions in so many states or »cells in the head«. For this reason head statisticians are probably far more likely to go off on a wild goose chase, mislead by correlations which later turn out to be spurious. The head statistician would also be far more likely to overlook differences or similarities which call for an explanation and thus show new aspects of the subject investigated, for as Vilhelm Aubert says statistical methods sometimes »have created phenomena which otherwise would not have existed«[91].

It is furthermore wrong to say, as historians sometimes do, that statistical methods lead to a false feeling of security, except in so far as a method which works well also will lead to a greater

87 Arne Næss (ut supra n. 66); Paul Feyerabend, Against Method, London (NLB) 1975.

88 Kuhn is often cited in support of scientific nihilism – one theory is as good as the other – a point of view often favoured by creationists and others pushing inferior theories. To this Kuhn would probably declare that he is no Kuhnist, confer the amplification and clarification of his views in Thomas S. Kuhn, The Essential Tension, Chicago and London (The University of Chicago Press) 1977; see also note 3 in his »Reflections on my Critics« (in: Imre Lakatos and Alan Musgrave, Criticism and the Growth of Knowledge, Cambridge 1972, p. 263).

89 Footnotes (published by the American Sociological Association) vol. 8 (1980) no. 6.

90 A »folk-science related to a ›romantic‹ philosophy of nature, combining stress on craftsman's manipulation, a personal involvement in the work, a democracy of participation, and a distrust of abstract or mathematical reasoning«. Such movements are all »doomed to end in tragic failure, since they are incapable of producing worthwhile scientific knowledge« (Ravetz, ut supra n. 46, p. 392–93).

91 Aubert (ut supra n. 16) p. 168.

feeling of security. On the contrary it seems that a precise statement is taken as a declaration of war by other researchers, as for instance J. C. Russell's daring statistical ventures have been: »Russell is, in fact, a pioneer more than anyone previously mentioned, except Graunt, and must be expected to make mistakes. – Russell's chief virtue, in fact, is that he gives others something to refute. All his figures may be altered eventually, but the debt to him will remain«[92].

It is also of some importance that when the researcher uses paper-and-pencil statistics, he is led to give a more detailed description of what he does, and for that reason his procedures may more easily be analysed step by step. This not only makes it easier to judge the merit of his work, but also to study the procedures and discover which ones to receive into the developing methodological canon of the research tradition. Researchers using other methods, on the other hand, such as softdata sociologists, humanist psychologists and others, are as a rule vague about what they do, and it is therefore difficult to distinguish the wheat from the tares.

It should not be thought that there is any basic contradiction between what is said in the chapters above and the historical research tradition. Although the evidence comes from the psychologists, the point of departure, from which I believe the rest can be derived, was three assumptions which most historians probably would agree with: a) that man remembers badly, b) that man is liable to make logical short circuits and c) that researchers are not very different from other men in these respects.

Thus the conclusion that experts are not half as expert as they are thought to be[93] ought not to be as surprising as it might seem at first sight. Nor does the view that it is methods and not any individual wisdom which give research the capacity to produce worthwhile knowledge, conflict with the scepticism towards hero-worship a thorough knowledge of history seems to lead to. In this view it might even seem possible that progress in research is less due to the intervention of insightful researchers than to exposure, by way of statistics, to what Monod calls chance and necessity[94], a process in which blind trial and error inevitably leads to molecules evolving into elephants and men or something resembling them[95].

7. Conclusions and recommendations

The conclusions drawn from what is said above, will in the last instance depend upon the answers given to the two following questions:

There is in the first place a question about the propriety of drawing conclusions about history from studies of other trades and professions. As historians have not yet been studied, they could turn out to be the black swans psychologists are looking for, people who get more reliable and intimate knowledge with ›Verstehen‹, hermeneutics, empathy and similar methods, than with

92 T. H. Hollingsworth, Historical Demography, London (The Sources of History/Hodder and Stoughton) 1969, p. 58. – On reading this H. O. Egede Larssen quoted Kumbel's grook: »Det du formulerer klart/ modbevises i en fart,/ det du mæler i det dunkle/ vil som visdom længe funkle.« (What you formulate clearly, will quickly be disproved. What you say obscurely, will long shine as wisdom).

93 As can be seen from chapters 2 and 3 and their references.

94 Jacques Monod, Le hasard et la nécessité, Paris 1970.

95 The expression is taken from P. W. Atkins, The Creation, Oxford and San Francisco (Freeman) 1981, p. 3. There is, in any case, no denying that chance plays an important part in scientific discovery, as can be seen from René Taton, Reason and Chance in Scientific Discovery, London (Hutchinson) 1957. A similar view lies behind the physicist John Archibald Wheeler's often quoted (in an abbreviated form): »Through all (research, FLN) one sees the spirit of catch as catch can, trial and error, progress by making almost all possible mistakes, the great point being only to make them as quickly as possible and to learn from them« (Id., A Septet of Sibyls: Aids in the Search for Truth, in: The American Scientist 44, 1956, p. 360–77).

statistics. This might even appear likely, seeing that the historian's problems are different, but, as a closer look will show, this is true for most other groups too; there cannot be many of them whose problems do not somehow differ from the others. This means that the relevant question is whether the problems of the historian are different in any essential respect, whether they have any bearing on the choice between statistics or other methods[96]. Such discussions often become interminable, since it is almost always possible to pick out some difference and claim that it is an essential one[97]. In the last instance only empirical studies, showing whether the differences really matter, can decide the point. As it is, considering how many dissimilar groups which have been tested, and how strong and unambiguous the circumstantial evidence is[98], the conclusion that ›Verstehen‹ and similar methods are inferior to statistics, should be preferred until empirical studies of historians might show the opposite.

There is in the second place the more general question of the propriety of allowing empirical evidence which contradicts philosophical arguments. Thus, in the case of the present anomaly where almost all the empirical studies point one way, and most of the philosophical argument another, ought we not to let ourselves be guided by philosophy and throw out the empirical studies? I think not, and for the following reason: The questions treated in the studies referred to above, may once have been philosophical questions, but are now beyond doubt empirical ones, questions, in fact, about distributions. For answering such questions counting and computing is indispensable, and merely to argue about them is to confuse the categories. This is the case even when the questions are quite general, such as when it for instance is claimed that one method, say hermeneutics, gives more insight than another. If this insight then has factual consequences, for instance for hitting the mark when guessing what went before, what came together with or what followed (for unravelling questions about causality in other words), we have in the percentages of hits a criterium to measure the methods with, and then tests like those used by Fischhoff or Smith[99] will be appropriate while a philosophical argument will not. Only if it is claimed that insight has no factual consequences, does not enable one to predict or postdict, can the question be reserved for philosophy. Other questions too, about which philosophy has little or nothing to say, nevertheless concern subjects which are fundamental to methodology, such as the reliability and validity of data. Thus the assessment of validity, whether data in fact tell us what we assume them to tell, has its fundamental rationale not from philosophers but from the psychologists Lee Cronbach and Paul Meehl[100]. As for reliability, which is assessed as a matter of routine in the younger social sciences, philosophers have, as far

96 I suspect that much of the reason for history's appearing different is found in the historian's background knowledge, the trivial but indispensable knowledge of details like administrative procedures, metrology, diplomatics and language (knowledge which usually functions as tacit assumptions but in principle may be tested). Students of present day society also need a comparable background knowledge, but it is acquired in a different way, merely by growing up one might say, and learning is made much easier by plentiful feedback. Actually the role this knowledge plays in empathy, hermeneutics and such, make such methods less likely to work well when studying less well documented societies, as may be seen if we ask whether we would have grinned as broadly if Galton (see p. 500) had tried to enter into the mind of a nineteenth century Londoner instead of a savage.

97 In, for instance, the delaying action fought by Christianity it has variously been claimed that Christianity differed from other religions in being monotheistic, in being based on revelation, in preaching high ethical standards and a number of other things besides. And why not? Sooner or later the immunisation will succeed.

98 Circumstantial evidence may carry considerable weight, even to the most prudent. When I mentioned this to a British friend, he said: »I know, I have hanged six men on circumstantial evidence.«

99 See note 30 and 31.

100 See note 72.

as I know, nothing to say about it although it may be an essential consideration when choosing between methods, seeing that validity must be low whenever reliability is[101].

This leads to a third question, a question of what place philosophy and empirical research, respectively, should have in the theory of history. There is, as I have tried to show above, some reason to think that a too great reliance on philosophy has made it difficult for history theorists to see that some theoretical questions indeed are, or can be made, empirical questions, and they have therefore committed the cardinal error of treating questions of fact as philosophical questions. Because of this history lacks much of the tested information necessary to choose between methods, and must either borrow or develop its own or, more practically, both. This may also have saved methods which might not stand up to testing, from prying empiricists, rather like in the odd kind of scientific discussion found in other social sciences where some schools hardly ever meet empirical studies, such as the ones I have referred to above, with other empirical studies, but either ignore them or counter them with some sort of philosophical or quasiphilosophical argument, and in that way ensure a long life for their theories[102]. For such reasons a renewed staking out of frontiers between philosophy and research seems to be called for, and together with that a more vigorous effort to analyse and test history's methodological assumptions, particularly as regards biased samples and low validity data, following the lead of the younger social sciences.

At this point the working historian will probably look up from the pages and ask what consequences this has for his work. I shall probably hem and haw, seeing that this is a subject on which passions run high and where statements are likely to be interpreted in ways surprising to the author, as Kuhn found out. Nevertheless, what is said above, had its origin in real research problems, and there is little that is ambiguous about the empirical evidence. The main points concerning generalisations can be summed up as follows:

1. In view of what is known about the danger of logical short circuits (p. 495), covering up of mistakes (p. 500, n. 59) or immunisation by way of »privatisation« (p. 507) and what is known about confirmation bias[103], the reasoning should be spelt out in formal models of the alternative explanations. From these testable, numerical, consequences may be derived, the more the better (p. 503).

2. In view of what is known about the erratic and imprecise working of human memory (ut supra, passim) head statistics should be avoided. The alternatives to pencil-and-paper statistics are study of observers' statements (p. 503–504) or shelving the question.

3. In view of what is known about the danger of bias (chapter 5) any sample which is not taken by the researcher from the whole population should be considered suspect. Such samples should be tested for presence or absence of bias with the methods sketched out in that chapter or with other methods which no doubt will be developed when historians give more thought to this crucial problem. Replications and analogies should be dealt with in a corresponding way (p. 502).

101 For reliability (inter- and intrasubjectivity), see page 495, for the relationship between reliability and validity, see page 496, for reliability and bootstrapping see page 499. More specific treatment of the subject can be found in text books on method such as Klaus KRIPPENDORF, Content Analysis: An Introduction to Its Methodology, Beverly Hills and London (Sage Publications) 1980, or monographs on the subject, such as Edward G. CARMINES and Richard A. ZELLER, Reliability and Validity Assessment, Beverly Hills and London (Sage Publications) 1979.

102 Such expedients have kept psychoanalysis alive for almost ninety years. Some empirical studies do exist, but their left hand character is clearly seen in H. J. EYSENCK and G. D. WILSON, The Experimental Study of Freudian Theories, London (Methuen & Co.) 1973.

103 C. R. MYNATT, M. E. DOHERTY and R. D. TWENEY, Confirmation bias in a simulated research environment: an experimental study of scientific inference, in: P. N. JOHNSON-LAIRD and P. C. WASON, Thinking (ut supra, note 62) p. 315–25.

4. In view of the experiences of other social scientists, historians' views of data suitable for statistics seem too restrictive (p. 494). What lies behind these views, can hardly be anything else than considerations of reliability. This, however, needs not be a matter of inference, reliability can be tested by fairly simple methods (n. 101).

5. As many Norwegian methodologists like to say, simple common sense statistics will carry a long way. What this implies can be seen from for instance Ottar Hellevik's text books[104].

6. Nothing which is said above, leads to a depreciation of traditional historical craftsmanship and background knowledge. What probably will appear less valuable in view of what is said above, is Grand Theory, unsupported by proven methods as it generally is.

Concerning singularisations the most important lesson is that our understanding of others is much more fallible than we take it to be (p. 498–501). When we also take into account the ambiguous feedback we have in such cases (p. 497–498), there seems good reason to show caution, using model thinking of the Smedslund-Loh pattern and basing the explanations on tested generalisations, the best point of departure according to Smith's study[105]. There is also reason to note that singularising studies may not be as valuable as is usually thought, for many singularising studies do not add up to a trustworthy generalisation unless they have been chosen the way a representative sample should.

104 Forskningsmetode i sosiologi og statsvitenskap, Oslo (Universitetsforlaget) 1971, and: Kausalanalyse av krysstabeller, Oslo (Universitetsforlaget) 1980. The title of the English translation will probably be: Introduction to Causal Analysis (George Allen & Unwin) 1984(?).

105 Ut supra (note 31).