Working Paper Series No. 103

DATA ANALYSIS IN DEVELOPMENT RESEARCH AN ARGUMENT

Chandan Mukherjee
and
Marc Wuyts

June 1991

CONTENTS

Just as there is no best way to listen to a Tchaikovsky symphony, or to write a book , or to raise a child, there is no best way to investigate social reality. Yet methodology has a role to play in all this. (Caldwell, 1981, p. 252)

INTRODUCTION[1]

Hypothesis searching is a crucial preoccupation of any researcher engaged in the analysis of development issues, and indeed it often accounts for much of the creativity of such research. In this paper we focus on this process of hypothesis searching based on a dialogue between theoretical reflection and empirical investigation.

We are dealing, therefore, with the problems faced by a researcher who is about to embark on empirical investigation in an area of socioeconomic analysis. Typically, the initial research question can be quite broadly specified. A researcher may be interested, for example, in studying the effects of migration on rural livelihoods in a particular village or in coming to grips with the complex interplay between productivity, pay structure and work organisation in a production situation where different groups of workers with different skills cooperate or conflict. To be able to embark on empirical investigation our researcher will need a well structured working hypothesis but this is not the same as a fully specified statistical hypothesis. Instead, a working hypothesis essentially involves a theoretical mapping of possible avenues of explanation of the social reality under study.

Our concern in this paper is with the role of data in the process of search which follows the initial formulation of a working hypothesis. It is a process of searching for the best explanation in the light of the available evidence (the data). This explains

our interest in methods of exploratory data analysis and related techniques as tools in the process of hypothesis searching. In a sense, however, data exploration is a bit of a misnomer. Certainly, it involves playing around with data, but what is being explored are feasible explanations - possible hypotheses. The purpose of the exercise is a process of theorizing and it is this aspect which accounts for its critical role in social research. The outcome of this process is a preferred conjecture - an hypothesis - which now, hopefully, has got more flesh on its bones and can stand up to scrutiny. That is the point where the work of detection ends and the court of law - confirmatory analysis - can take over.

This process of hypothesis searching also implies a kind of inference which Harman attractively labelled as "inference to the best explanation" (Harman, [1965], in Brody and Grandy, 1989, pp. 323-328). Section 1 of this paper provides a brief introduction to this concept and contrasts it with other forms of inference.

Subsequently, section 2 turns to the methodology of data analysis in a context of inference to the best explanation. It argues that these exploratory methods are not merely descriptive in nature but involve abstraction. More specifically, they need to be rooted in statistical theory as the exploration of hypotheses also implies exercises in statistical modelling.

Finally, section 3 discusses the question of the width and the depth of the data we use in hypothesis searching. Moreover, we argue that data are not just mere facts. The categories in which they are cast and the variables we choose highlight selective aspects of the social reality under investigation.

This paper, therefore, makes a case for the recognition of specific methodological approaches in data analysis appropriate for hypothesis searching, or stated otherwise, for inference to the best explanation. We believe that this type of data analysis

should be clearly distinguished from the methods of formal statistical inference. This would help researchers who are often puzzled by the lack of relevance of courses in statistics to deal with what they perceive as the main task and indeed the most creative aspect of social analysis - constructing meaningful research hypotheses in the process of a dialogue between theoretical reflection and empirical analysis.

I. INFERENCE TO THE BEST EXPLANATION

Hypothesis searching versus statistical inference

Our argument in this paper is that the process of making conjectures is an important kind of inference in its own right which should not be confused or mingled up with inductive statistical inference, the cornerstone of statistics teaching.[2] In fact, often our data do not allow us to do more than making conjectures. In our opinion, this exercise is worth the effort in its own right.

To start with, a distinction can be made between three kinds of inference: deduction, induction and hypothesis (Hacking, 1990, p. 207)[3]. This distinction between different types of inferences, and particularly between statistical inductive inference and 'the method of hypothesis', is crucial to help us locate the role of data in the process of hypothesis searching. Our concern is with the latter type of inference: the process of conjecturing a preferred hypothesis (p. 207). In fact, the method of statistical inductive inference and the method of hypothesis are fundamentally different since "probability has nothing to do with hypothesis, while it has something to do with induction" (p. 207).

This latter issue (probability within statistical induction) is paramount in most statistics courses and textbooks. There it is assumed that our researcher is fully equipped with a well-

specified hypothesis to be tested against the data specifically sampled for that purpose. In this case, there is no confusion about what constitutes the population from which the sample is drawn and about the precise specification of the probabilistic model which hypothesizes the stochastic nature of that population. This is the realm of formal statistical inference "where some relatively precise specification or description of a manageable mathematical model allows us, at least apparently, to tie up our uncertainties in a neat package - for example, through confidence limits, significance tests, fiducial inference, posterior distributions, or likelihood functions" (Mosteller and Tukey, 1977, pp. 21-22).

Hence, formal inductive statistical inference deals with assessing uncertainty in terms of a probability model known but for the values of a few parameters. Probability then enters in the process of inductive inference through significance tests like the t and F tests, confidence limits, likelihood functions, and so on.[4]

'The method of hypothesis', however, is concerned with a different question altogether. Here we are dealing with a situation where the construction of an explanation is the name of the game. This, as we shall see, is the process of 'inference to the best explanation'.[5] While statistical inductive inference involves making probability statements within the exact boundaries of a well-specified hypothesis, in this case what is being infered is the hypothesis itself. The hypothesis thus arrived at may subsequently be subjected to confirmatory analysis (statistical inductive inference based on fresh data), but the way we arrive at the hypothesis does not involve probability statements of a nature common in statistical inference precisely because we are not equipped from the outset with a clear probability model. Hence, inference to the best explanation constitutes a preamble to subsequent confirmatory analysis (i.e. inductive inference), yet as a method it stands on its own and is distinct in its approach and its methods.

The method of hypothesis

Harman ([1965], 1989) explains this kind of inference as follows:

> In making this inference one infers, from the fact that a certain hypothesis would explain the evidence, to the truth of that hypothesis. In general, there will be several hypotheses which might explain the evidence, so one must be able to reject all such alternative hypotheses before one is warranted in making the inference. Thus one infers, from the premise that a given hypothesis would provide a 'better' explanation for the evidence than would any other hypothesis, to the conclusion that the given hypothesis is true.
>
> There is, of course, a problem about how one is to judge that one hypothesis is sufficiently better than another hypothesis. Presumably such a judgement will be based on considerations such as which hypothesis is simpler, which is more plausible, which explains more, which is less ad hoc, and so forth. (p.324).

The issue at hand, therefore, concerns the choice among a range of possible explanations on the basis of the available evidence.

To enable us to start a dialogue between theory and the data it is necessary at the outset that we combine our general theoretical knowledge of the subject matter with some hard theoretical thinking on the issue under study into a working hypothesis: a research agenda which sets out the different questions to explore, maps out different routes of feasible explanations, and delineates the boundaries of our search. Consequently, it shapes what we consider to be relevant evidence to the research question at hand. That is, it sets the boundaries of the evidence considered relevant in the process of searching for the best explanation. This is important since the process of inference to the best explanation needs to be based on a consideration of the total available evidence in order to verify whether such inference is indeed warranted or not (Harman, 1989, p.325)[6]. This has implications for the design of an observational programme in particular, and of broader data

gathering in general; a question we shall return to in section III of this paper.

In effect, a working hypothesis is nothing else but an hypothesis about the search for hypotheses. It is akin, therefore, to throwing a fishing net into the sea with the fish as yet uncaught but with a good knowledge about the location of the likely fishing waters.

Obviously, it goes without saying that a given working hypothesis is never complete in some absolute sense. Indeed, any working hypothesis is always located within one or another theoretical tradition or school of thought which shapes the researcher's approach in analysis. Furthermore, even within a particular tradition, much depends on proper homework, creativity and the particular approach of a researcher within a broader school of thought. Not much more can be said in this paper on this issue since it obviously depends on the subject matter concerned and the specificities of the research area.

We argued that a working hypothesis in a context of inference to the best explanation has little in common with a statistical hypothesis as used in statistical inductive inference. This has important consequences for the nature of data analysis in the search for the best explanation.

Firstly, as pointed out above, this method of conjecturing a preferred hypothesis has nothing to do with statistical inductive inference. Hence, it does not involve any probability statements of the type implicit in confidence intervals or significance tests such as the t or F tests, and so on. That is, it is not within the realm of confirmatory analysis since it does not start with a clearly specified probability model about the behaviour of the data.

Secondly, it relies on methods of data exploration and related techniques (as discussed in section II) which aim to pick up

hints and clues from the data so as to be able to narrow down the search and to deepen the analysis. This process is 'theory driven' but involves a dialogue between theory and the data. As the process proceeds some avenues of explanation may turn out to be dead end streets in the light of the available evidence, while others may look promising and invite further questions which provide a firmer grip on the issue under study. It is in this sense that exploratory data analysis is a bit of a misnomer; we look at data but what is being explored are feasible explanations.

In this respect, hunting out interesting indications from the data, or hint-searching (Mosteller and Tukey, 1977, p. 29), is essential to the process of constructing hypotheses, but the art of the matter consists in assessing the total available evidence in order to work one's way towards the explanation which proves to be more plausible, simpler and more coherent. Selectively picking and choosing bits and pieces of individual indications and elevating these to the status of serious hypotheses may satisfy the researcher bend on data mining, but it certainly does not constitute a valid method of inference to the best explanation.

Finally, data exploration in the process of conjecturing a preferred hypothesis requires that we also come to terms with the patterns of variation in the data. In fact, as we shall argue in the next section, this involves statistical modelling and not just descriptive analysis. Numerical data just do not give us only quantitative information but also qualitative insights. For example, the shape of a graph conveys quality as well as quantity. This is often forgotten in the context of statistical inductive inference since here the qualitative aspects are often hidden away in the assumptions on the character of the probability model. As such, the analysis tends to center on estimating values of the parameters of the model (the mean, the variance, and so on) and issues concerning the qualitative nature of variations in the data are pushed in the

background. But, for example, the assumption that a variable is distributed normally (or, for that matter, lognormal, exponential, or gamma) is in fact a much stronger qualitative statement about the data than merely giving its mean.

Hence, data analysis in the context of hypothesis searching is as much concerned with quality as with quantity. It is a fallacy to think that quantitative analysis only gives us numbers to think about. It is however a common error. Hill (1986, p.21), for example, points out that many development economists put too much emphasis on averages (typical values) when in fact they ought to have drawn a graph.

Conclusion

This section attempted to locate the process of hypothesis searching as a valid, and indeed essential, form of inference in its own right. Its approach and its method should not be confused with that of statistical inductive inference proper. Often we can do no more than making conjectures without moving on to comprehensive, thorough and systematic hypothesis testing. This type of exercise, however, is an important, valid, and indeed creative part of the research process.

II. DATA ANALYSIS IN HYPOTHESIS SEARCHING

Exploration involves abstraction

Data analysis in which confirmatory procedures of statistical inference is not applied is generally decried a 'mere exercise in descriptive statistics'. This is understandable. In most textbooks, the classical descriptive methods are presented in isolation of formal statistical theory. At best it is a prelude to formal analysis; at worst it is a brief introduction to a few formulae for calculating a mean or a standard deviation. In contrast, statistical inference is presented as firmly rooted in

probability theory and mathematical statistics. That is, a statistical model or hypothesis always specifies clearly the probabilistic assumptions about the population from which a sample is drawn randomly to enable proper inferences to be made. In comparison with statistical inference, therefore, descriptive statistics does indeed appear rather elementary, devoid of theory, and definitely unexciting. In this section, we challenge this view and argue that descriptive statistics is in fact a misnomer when looked upon from the perspective of hypothesis searching as an activity in its own right.

The exploration of data, however descriptive it may appear, cannot be done in isolation from formal statistical theory. This point is generally overlooked, particularly in statistics textbooks. In fact, the history and practice of statistics always involved a two-ways avenue between empirical statistical distributions and their formal mathematical counterparts in probability theory. The latter were modelled on the former, and, in turn, provide us with theoretical handles to be able to recognise patterns within the former.

When we look at the data using analytical graphs or tables, we do not just see patterns in the data; more often we recognise them, at least the dominant ones. But, this implies that we explore data with a priori abstractions in our mind from which we select an appropriate fit. To do this we always have statistical theory and models at the back of our minds. It follows that, in the dialogue between theory and data, 'theory' does not just refer to substantive theory (economics, sociology etc.) relating to the research topic in question, but also to statistical theory.

The question of tools

There are now a number of excellent books available which describe a range of extremely useful techniques, particularly

visual displays, in Exploratory Data Analysis [EDA] (Tukey, 1977; Erickson and Nosanchuk, 1979; Chambers et al. (1983)). The guiding principles behind these exploratory techniques are healthy scepticism and openness (Hartwig and Dearing, 1979, p. 9). These principles are, in general, the basis of any creative research. Healthy scepticism has led to the choice of resistant statistics like quantiles adapted from classical descriptive methods and newly developed summaries. Openness as a principle enables the researcher to look freely at the data from different points of view, to look deeper into the data, and to find out what the data appear to show. What these techniques demonstrate is that it is possible to extract a wealth of 'indications' from rather messy data. To go beyond 'indication' to assess uncertainty is of course desirable but its feasibility depends on the way the data were obtained.[7] But indication is a necessary first step forward.

The most powerful tools available in this literature for data exploration are graphical displays[8] (Chambers et. al., 1983). It is undoubtedly the best medium to look for patterns and relationships, to check how reasonable are the assumptions which underlie statistical modelling, and, most importantly, to discover the unexpected. Much research has been done in this area, mainly in three directions: graphs which allow researchers to look at and compare distributions (e.g. quantile plots; Q-Q plots; normal probability plots; and so on), graphs which plot raw data in one or more dimensions, and more complex graphs which plot quantities derived from the data through modelling (e.g. various types of plots of residuals and of fitted values).

Another set of techniques developed recently concern diagnostic analysis (a comprehensive study is provided in Belsley et.al, 1980). These techniques - influence diagnostics and residual analysis - evolved out of the practice with classical regression methods. Though these techniques are mostly presented in the context of confirmatory analysis they constitute, in our view, an extremely useful set of tools for exploring hypotheses with

data. In fact, the use of diagnostic analysis is generally restricted to the pathology of estimation (Gilbert, 1990, p. 280). We argue that these techniques can be made to play a creative role in exploration of hypothesis.

A final set of techniques which have shown their worth in data analysis are transformations (for example, a good treatment of this area can be found in Atkinson, 1985). The point about these is not so much the mathematical operations they entail, but the fact that they alter the distributional characteristics of the data. This way, data which are heavily skewed can be moulded back into shape.

Our argument is that a synthesis of these three sets of techniques - exploratory data analysis, diagnostic analysis, and transformations - provides powerful statistical tools in the process of inference to the best explanation. The ingredients for such a synthesis can already be found, besides in the specialised literature, in more recent advanced textbooks (for example, Myers, 1990; Rawlings, 1988) and also some recent basic texts (for example, Hamilton, 1990; Moore and McCabe, 1989).

The methodology of searching

The search for hypotheses through data exploration is a process of trial and error in modelling patterns within the data. To do this, we use abstractions derived from statistical theory. Statistical models and procedures are based on distributional assumptions. Thus, the modelling process consists in the examination of distributions exhibited by the data and using properties of theoretical distributions to abstract the patterns from the data:

> Dempster (1983) suggests that the difference between EDA (Exploratory Data Analysis) and statistical modelling is often exaggerated. Although statistical model building makes use of formal probability calculations, the probabilities usually have no sharply defined interpretation (either frequentist or Bayesian), and the

> whole model-building process is simply a form of exploratory analysis. Dempster notes that both EDA and statistical modelling cycle back and forth between fitting curves and looking at residuals. EDA generally starts with summaries and displays, whereas the modelling approach starts by fitting a curve. (Diaconis, 1985, pp.27).

The study of relationships is the primary concern of data analysis in social science. Regression analysis plays the most important role in this context. Mosteller and Tukey (1977) have shown how EDA techniques can be used to study relationships with regression. Furthermore, texts like Myers (1990) and Rawlings (1988) show how diagnostic analysis can be incorporated within the framework of formal regression techniques.

In the EDA literature, the trial and error process of modelling is schematically presented as:

DATA = FIT + RESIDUALS

The exploration is envisaged in an alternating sequence of finding a FIT (pattern) and examining the RESIDUALs for further fitting. The process ends when there is no more discernible pattern left in the RESIDUALs. In our perspective, the FIT embodies the substantive modelling of how things are related, how they work. RESIDUALS reflect uncertainties about the substantive model empirically arrived at through exploration. To arrive at a statistical model we need to take account of both the FIT and the RESIDUALs.

The Gaussian distribution and linear models are central to the theory of statistical relationship. The visual aspects of linear statistical relationship have many attractive points from the point of view of exploration and modelling. It is simple, easy to grasp and interpret, and it is easy to examine the variation around the relationship. In practice, most often, the theory of linear models provide the abstractions which are used to explore relationships in the data. Moreover, transformations allow us to extend the scope of these linear models.

Residual analysis and diagnostic plots play a major role in this analysis. They serve two purposes. First, besides providing clues and hints about the nature of relationships (e.g. non-linearity) between the variables under investigation, they force us to see the unexpected, and thereby enriches the substantive analysis. In fact, outliers and influential points often tell us more than the FIT itself. Secondly, they tell us about the nature of uncertainty regarding the relationships under study.

Two further points need to be made with respect to the utility of diagnostic plots of residuals. First, the role of t-statistics and the overall F-statistics in confirmatory analysis runs parallel to 'partial regression plots' and plots of 'predicted versus observed' values in exploratory analysis. In fact, t- and F-statistics can be regarded as convenient summary statistics in an exploratory perspective. Indeed, in confirmatory analysis, these statistics serve the purpose of inference about a pre-specified population, whereas, in exploratory analysis the plots tell us how far the data at hand conform to the specific fit we made. They also show up outliers which do not agree with the dominant relationship (a possible basis for further exploration of one's hypothesis) and influential points which affect the nature of our fit.

Second, diagnostic plots help in assessing how <u>reasonable</u> the distributional assumptions are which underlie the statistical abstractions we have used. We stress the word 'reasonable' because we do not pretend that any statistical model we arrive at through exploration is anywhere near as neat and clean as their theoretical counterpart. Abstractions are to be <u>used</u> and not to be <u>believed</u>; data never conform to the abstractions we use to analyze them. What matters is whether the assumptions are approximately satisfied.[9]

The investigation of distributional assumptions is necessary if

we are to come up with a statistically testable hypothesis. More importantly, departures of the distribution of the residuals from its theoretical counterpart can in turn provide clues about the nature of the variation in the data and invite us to look deeper into the data or to try an alternative abstraction. Investigating distributional features matters greatly because it gives us valuable qualitative insights about the underlying social processes.

As such, the analysis of residuals can play an important creative role in exploring hypotheses rather than merely being a tool to identify pathological problems in confirmatory analysis.

Conclusion

We argued in this section that exploratory data analysis as the statistical tool of inference to the best explanation runs parallel to the statistical methods of inductive inference. Both are rooted in distribution theory, draw from it, and enrich it. Both attempt to deal with similar problems but from different perspectives just as the task of a detective differs from that of a lawyer or a judge. But it is useful to keep in mind that nobody was ever convicted for a crime which went undetected in the first place.

III. DATA IN RESEARCH: WHICH POPULATION AND WHICH FACTS?

Meaningful categories and summaries

Data are always organised in categories. For example, as Hacking (1990, p.3) puts it beautifully, historically "categories had to be invented into which people could conveniently fall in order to be counted." Categories matter because they structure what we perceive to be relevant

populations for which we try to obtain meaningful indications. Categories, therefore, are essential tools in the process of theorizing with numbers.

The question of meaningful categories constituted an important conceptual barrier to the spread of statistical methods to the social sciences in the nineteenth-century (Stigler, 1986 and 1987); a problem which still remains alive up to date. Indeed, the problem of categories was important in two respects: (1) the categorisation of data into homogeneous groups on the one hand, and (2) the character of the indications based on such categories, on the other hand. Furthermore, both issues are closely related with one another.

It is useful to discuss both issues and their interrelation in somewhat more depth. The questions which they raise are not just of interest to the historian or the philosopher of science, but they address fundamental issues which concern the applied researcher every step of the way in the process of theorizing.

The problem can best be illustrated by looking at the early history of the contrasting developments of statistical applications in astronomy on the one hand, and social sciences (particularly economics and sociology) on the other (Stigler, 1987; Hacking, 1990, chapter 13).[10]

The problem consisted in the types of uncertainty confronted in each of these sciences. In astronomy, the issue at hand was the problem of measurement errors. In its simplest form, the issue was that of estimating a single astronomical quantity, real but unknown, on the basis of repeated observations subject to measurement errors (Hacking, 1990, p.108). In this case data can be classified a priori, disciplined by accepted mathematical theory and what were seen as uniform conditions of observations (Stigler, 1987, pp.289-290). The problem was that of 'the combination of observations' to arrive at the best estimate of this real but unknown astronomical quantity. Uncertainty sprang

from measurement errors only. This problem propelled the remarkable developments of the method of least squares, the Gaussian normal curve and Laplace's limit theorem which provided substance to the behaviour of the errors around the constant mean (Stigler, 1987; Hacking, 1990, ch.13).

In social sciences the problem of categorising data in homogeneous groups - that is, "groups for which the major influential factors could be considered constant and residual variation was seen as haphazard accidental causes" (Stigler, 1987, p.289) - was far more complex. There was much less accepted theory to classify data a priori, and the inability to do so forced social scientists to search for it a posteriori - a much more difficult problem (Ibid, p.290). In fact, in 1844, the Belgian statistician Quetelet tried to do just that with his concept of "average man" and his method of fitting normal curves to the data to check whether any particular grouping could be seen as a homogeneous category. It was essentially an "attempt to categorize data as homogeneous based upon analyses internal to the data" (Ibid, p.290, emphasis added).

In fact, Quetelet's attempt was not really successful since "the mere appearance of normality is not at all sufficient to conclude homogeneity" (Ibid, p.290). Indeed, "normal curves can and do arise in countless ways from the compounding of other normal curves" (Ibid., p.290). To solve this problem required more complex techniques (such as regression analysis, analysis of variance, and related linear models) to enable a researcher to unravel the effects of various factors in cross-classified data (Ibid., p.291). But this, once more, raises the issue of the various relevant categories to be taken into consideration and the various ways these can be combined in cross-classification to analyze the effects of different factors.

But there was a further dimension which arose when applying techniques borrowed from astronomy to the social sciences. In astronomy, the purpose was to estimate a real but unknown

physical quantity, but what was being estimated when computing an average with social data? Hacking (1990) answers this question as follows:

> It was Quetelet's less noticed next step, of 1844, that counted for more than the average man. He transformed the theory of measuring unknown physical quantities, with a definite probable error, into the theory of measuring ideal or abstract properties of a population. Because these could be subjected to the same formal techniques they became real quantities. This is a crucial step in the taming of chance. It began to turn statistical laws that were merely descriptive of large-scale regularities into laws of nature and society that dealt in underlying truths and causes. (p.108).

This conceptual advance had far-reaching repercussions on scientific practice - it consisted in "the subtle but profound reinterpretation of statistical averages and the implications of their existence" (Krüger, 1987, p.80). Furthermore, it also had a profound influence on what could be called policy analysis. Quetelet lived in a time of social reform, particularly with respect to public health and elementary education. In fact, "statistics became the favoured instrument of observation and reasoning of the liberal social reformer, who was determined to leave alone all things individual, i.e. man's free agency, and to battle only against these collective contingencies that threatened an average man's health and vitality, contingencies that were phenomena of large numbers" (Metz, 1987, p.342). For this to be possible an average ideal or abstract property of a population had to become a real quantity as a property of the whole body that has not meaning for its elementary parts. Average life expectancy, for example, is such a concept.

This brief digression on the history of statistics highlighted the importance of the way we categorize socioeconomic data as well as of the meaning of the indications based on these categories. The main point we wanted to make is that the choice of categories shapes our conception of what we consider to be the relevant population and how we would go about modelling the

characteristics of a population.

Internal and supplementary uncertainty

In classical statistical inference as taught in most statistics courses or textbooks, no confusion can arise, in principle, from the character of the data and of the population from which they are sampled. Indeed, the research hypothesis clearly defines the relevant population in question - i.e. the data can be classified a priori and sampled in accordance with this classification - and, furthermore, specifies a probabilistic model, known up to a few parameters, of the stochastic nature of this population. Appropriate sampling design, whether through experimentation or through an observational programme, secures that the sample data allow us to make the appropriate inferences given the assumptions of the model. There is little ambiguity in this context: the hypothesis clearly defines the population and sets the stage for proper sampling procedures. Statistics books take this for granted.

However, in the process of inference to the best explanation, the problem is more complex. We do not start from a cast-iron hypothesis before we engage in sampling our field data (if, at all, we do rely on own data). Obviously, in designing an observational programme to gather field data we are guided by our working hypothesis: an umbrella-type, somewhat vague, but broad- spectrum research agenda which allows us to explore various avenues of inquiry.

This brings us back to the problem that, in social analysis, we cannot fully rely on theory to give us precise a priori categories for our data as well as an exact set of all factors to be taken into consideration. The inability to do so, forces us to widen the search by starting off with multiple handles on the problem. This point is well argued in the following quotation:

> To go beyond indication, we need to assess the uncertainty

> of our indications. Although precision of assessment has value, reality of assessment is more basic, because we can be easily misled by variables not represented or recognized in a study.
> We assign contributions to uncertainty to two sources: those that might be judged from the data at hand - internal uncertainty; and those that come from causes whose effects are not revealed by the data - supplementary uncertainty. Thus internal and supplementary uncertainty are two vague concepts intended to aid our understanding of uncertainty , variation, and stability. Failure to attend to both sources can lead to serious underestimates of uncertainty and consequent overoptimism about the stability of the indication. To avoid these traps, we need to choose a satisfactory error term from the data, and we need to allow for sources of variation that are present but not made visible by the data gathering process.
> Good design in observational programs and experiments can reduce the impact of all kinds of variation upon the uncertainty of our results. Design can be especially valuable in helping to make sure that major sources of variation are introduced into the investigation. It is often wise to "broaden the base" of a narrowly focused investigation so that the internal uncertainty can properly represent the real variation and the supplementary uncertainty can be reduced. (Mosteller and Tukey, 1977, p.119)

In a way, this argument for a broad based design is nothing but a concrete application of our principle of taking account of the total evidence when engaging in a process of inference to the best explanation. As our purpose is to search for hypotheses, our research design should allow us to explore various avenues of possible explanations.

Data as facts?

But the problem is often more complex. Not uncommonly, development researchers have to work with the data <u>at hand</u> - mostly <u>official statistics</u>, but also data produced by other researchers or institutions. For example, development economists (particularly, macro economists) and demographers typically tend to rely almost exclusively on official data to explore or test their theories. In contrast, various types of fieldworkers (anthropologists, population scientists, sociologists or

economists engaged in field research, etc.) usually work with own data obtained through fieldwork.

Data, however, are not neutral, neither with respect to which data are collected, nor how they are structured. Official data, however, acquire that particular quality of being looked upon as hard facts. Indeed, the routine collection of official statistical data forges "a domain of the factual" (Hopwood, 1984, p. 168). These data both reflect and, in turn, influence the emphases that are given in public debates (Ibid, p. 169). For example, Hopwood argues that "much of the apparatus of national income accounting emerged to aid the management of a constrained wartime economy (Seers (1976), Tomlinson (1981, pp. 129-132)). [But] in the postwar discussions on the desirability of economic growth, however, GNP started to take on many of the attributes of being an end in itself and, in the process, had a more widespread influence on other governmental policies" (1984, p. 169).[11]

Data, therefore, are partial records which provide us with a selective visibility (Ibid., p. 170) of a society and an economy. This reality of selective visibility and its impact in terms of forging the domain of the factual is not only true for economic data, but also for data to do with people and the way we perceive ourselves and others.[12]

This avalanche of 'facts', therefore, has a profound influence on debates within the public sphere and also, in various degrees, on social scientists rapped up in social research and policy analysis. Not surprisingly, macro economists and demographers are more caught up in this particular view of reality, and, consequently, they tend to be more affirmative in their analysis and to see their conclusions as being more objective than those of their fellow social scientists. Theirs is a world of hard facts. In contrast, social and economic anthropologists, population scientists, and economists or sociologists engaged in field work are often more hesitant in

their approach, more aware of the complex nature of variation and uncertainty in social data, less inclined to generalise too quickly, and also more tentative in their conclusions.[13]

This contrast is further compounded by the fact that official data - census data, national income accounts, household budget survey data, and so on - are mostly aggregated and structured into formal accounting frameworks. But aggregation hides the internal variability of the data within categories and, furthermore, accounting techniques often 'resolve' conflicts among the data through formal procedures. Both these elements - aggregation and formal accounting procedures - are obviously important and relevant, but they also strengthen our perception of the objective nature of the facts presented this way.

Conclusion

In this section we argued that the categories in which we cast our data structure what we perceive to be relevant populations and meaningful indications. In statistical inference the definition of the relevant population is not a problem. The hypothesis being tested clearly defines the population under study and defines a priori and comprehensively the categories in which our data are sampled. In contrast, in the process of hypothesis searching the question as to which data to collect and how to organize them is of paramount importance.

Indeed, we argued that in the process of hypothesis searching we do best to explore different possible avenues of feasible explanations of the issue under study, and, hence, to seek to organize the data in ways which gives us multiple handles to explore different explanations. In terms of research design, this means that we try to make sure to incorporate the various possible sources of variation so as to reduce the supplementary uncertainty inherent in our search for feasible explanations. This is an application of our earlier principle of taking

account of the total evidence when working towards hypothesis formulation.

Finally, we argued that official data are important sources of evidence to be taken into consideration, but we should never forget the pervasive influence they exert on the way we interpret social reality and analyze it. They often confront us as hard facts - an appearance which is furthermore compounded by the aggregate nature of such data and by the formal consistency imposed on the data by the accounting procedures used.

CONCLUSION

The argument put forwards in this paper is as yet very tentative in nature. In fact, it is no more than an exploration on our part based on our experience as applied data analysts and teachers in statistics. This exploration was prompted by a question which puzzled us greatly - namely, that the problems we encountered in day to day research contrasted sharply with the usual emphases around which traditional statistics books are structured. For us, the main problem in development research is that of using data to make conjectures about the issue under study; statistics books start from the premise that we have a well-specified hypothesis at hand and data are collected to test it, not to find it.

Our concern, therefore, is with data in the process of hypothesis searching. This explains our interest in exploratory data analysis and related techniques. These techniques, however, are often presented as sophisticated elaborations of descriptive statistics: a 'box of tools' to get more mileage out of data. We argue differently for three reasons.

Firstly, Exploratory data analysis and related techniques involve looking at the data but what is being explored are possible hypotheses, not the data per se. These methods involve

getting hints and clues from the data so as to be able to pursue different avenues of explanations of the question under study. In this sense, we argued that these methods are appropriate for a distinct type of inference - namely, inference to the best explanation.

Secondly, we argued that exploratory data analysis, seen in this way, is not descriptive in nature. Rather, it involves abstraction in two distinct ways. Firstly, the search for meaningful hypotheses is itself an exercise in abstraction. In fact, in most cases, it constitutes the most creative part of the research process. But, secondly, data exploration involves that we employ abstractions to make sense of the data themselves. We do not just see patterns in the data; we also recognise them. To do this, we rely on abstractions provided by statistical theory (particularly, distribution theory). Residuals - which play a central role in these techniques - only assume meaning once a model is tried out on the data.

Finally, we argued that the process of hypothesis testing also forces us to think hard about the theoretical categories we employ to structure and collect our data and about the meaning of the indications derived from them or from the way we combine them. Textbooks in statistics take for granted that we know exactly what the relevant population is from which the data have to be sampled. The definition of the population and of the corresponding categories to structure the observed data is therefore considered unproblematic and totally determined from the outset. In contrast, our experience tells us that this is often the hardest part of the research process.

The art of making conjectures - the process of inference to the best explanation - involves a complex but flexible dialogue between theoretical reflection and empirical analysis. There is no best way to do this but, nevertheless, it requires a clear perception of the methodological issues involved. Our argument is that the methods of data analysis appropriate for this type

of inference should be recognized in their own right - distinct from, but parallel to the methods of statistical inductive inference.

NOTES

1. This paper was produced under the co-operation between the Population & Development Training Programmes at Centre for Development Studies (Trivandrum, India) and Institute of Social Studies (The Hague, The Netherlands) which form part of the UNFPA Global Training Programme in Population and Development. The authors thank the UNFPA for its assistance in making this joint effort possible.

   We are thankful to Prof.Sarathi Acharya of the Tata Institute of Social Studies (Bombay, India) for his stimulating comments on an earlier draft.

2. This point is well-stated by Diaconis in the following quote: "Suitable theories of inference and the mathematics of probability underlie many, nowadays routine, applications of statistics. These include randomized clinical trials, sample surveys, least-squares fits to underlying models, quality control, and many other examples. Yet, none of the classical theories of statistics comes close to capturing what a real scientist does when exploring new data in a real scientific problem. All formal theories - Neyman-Pearson, decision-theoretic, Bayesian, Fisherian, and others - work with prespecified probability models. Typically, independent and identically distributed observations come from a distribution supposed known up to a few parameters. In practice of course, hypotheses often emerge after the data have been examined; patterns seen in the data combine with subject-matter knowledge in a mix that has so far defied description." (1985, p.22).

3. The distinction between deductive and inductive inference is well known. Swijtink (1987) puts it as follows:

   > "An argument is a piece of discourse, consisting of a set of premises, and a single conclusion. Premises and conclusion express propositions: that is, what they say is true or false. An argument is deductively valid if its conclusion follows strictly from its premises, in the sense that its premises could not be true and its conclusion false. Logic tries to find tools with which one can establish deductive validity, and the philosophy of logic tries to explain what could possibly be meant by the 'could not'. Similarly, an argument is inductively strong if it is improbable, given that the premises are true, that the conclusion is false" (p. 274).

   Note that inductive reasoning does not mean to say that the conclusion has a probability of such and such. Rather, it states that the conclusion is reached by an argument that, with such and such a probability, gives true conclusions from true premises (Hacking, 1990, p. 209).

4. In statistical inference, estimation and hypothesis testing evolve around a few unknown parameters of an otherwise fully specified statistical model, possibly supplemented by some diagnostic testing as to the validity of the assumptions of the model. In actual practice, however, the process is not as clean cut as all that. Data often play a part in arriving at a 'better' specification; a procedure which

undoubtedly entails the danger of data mining. In econometrics, for example, the discrepancy between textbook teaching and actual practice gave rise to an intense debate on the validity of specification searches based on data exploration and diagnostic testing (Granger, 1990; Spanos, 1990, Pagan, 1990). This debate mostly revolves around model specification of a particular econometric function (a consumption function, a production function, an Engel curve, and so on) or a set of simultaneous equations. The question then arises whether the data should be allowed to play a role alongside economic theory in determining which variables to include in the equation, its functional form, and, in the case of time series, the nature of the lag structures involved.

In this paper we do not address this problem of specification searches in the context of modelling a particular function or a set of equations suggested by economic (or social) theory. Our focus is in fact quite different.

5. The concept is also known under different terms as Harman explains: "The inference to the best explanation corresponds approximately to what others have called "abduction," "the method of hypothesis," "hypothetic inference," "the method of elimination," "eliminative induction," and "theoretical inference". I prefer my own terminology because I believe that it avoids most of the misleading suggestions of the alternative terminologies." ([1965], 1990, pp. 323-324).

6. Note that we are not considering the issue of choosing between contesting theories through formal statistical inductive inference. The point we are making here concerns the process through which we arrive at the construction of a particular conjecture.

7. "Naturally, we all desire an adequate assessment of both indication and their uncertainties, but we shouldn't refuse good cake only because we can't have frosting too". (Mosteller and Tukey, 1977, 27).

8. Chambers et al. (1983) put this as follows: "There is no single statistical tool that is as powerful as a well chosen graph. Our eye-brain system is the most sophisticated information processor ever developed, and through graphical displays we can put this system to good use to obtain deep insight into the structure of data. A enormous amount of quantitative information can be conveyed by graphs; our eye-brain system can summarize vast information quickly and extract salient features, but it is also capable of focusing on detail. Even for small sets of data, there are many patterns and relationships that are considerably easier to discern in graphical displays than by any other data analytic method. (p.1).

9. See, for example, Chambers et. al. who argue that: "When we use a phrase like "data generated from a normal distribution" it is important to keep in mind that we do not mean it in a precise sense. Real data can never come from a genuine

normal distribution, for that would require data with infinite precision and with the possibility of including arbitrarily large and small values. In practical terms, all data are discrete, since they have limited accuracy, and they are bounded above and below. When we ask, "Are the data normally distributed?" we are really asking "Is the empirical distribution of data sufficiently well-approximated by a normal distribution for the purposes we have in mind?"" (Chambers et.al., 1983, pp.192).

10. "Two different "sciences", one of the sky and one of society, were "started" in the 1600s. Astronomy developed and used statistical methods of assessing uncertainty rather early; social sciences knew well of astronomy's triumph yet lagged by nearly one century" (Stigler, 1987, p.289).

11. The argument in full runs as follows:

"Consequently, the economic has undoubtedly been made more visible than the social and the political, in many spheres of public life. Now, however, the imperatives of those economic 'facts' can be contrasted with the more questionable dictates of political ideology and social preference, illustrating, in the process, the powerful influence that the recording tools of the public domain can have.
In fact, organisational accounts laid down to orientate management action to wider ends can themselves come to serve as statements of the ends to be achieved (Ridgeway, 1956). Much of the apparatus of national income accounting emerged to aid the management of a constrained wartime economy (Seers (1976), Tomlinson (1981, pp. 129-132)). In the postwar discussions on the desirability of economic growth, however, GNP started to take on many of the attributes of being an end in itself and, in the process, had a more widespread influence on other governmental policies. The imagery of profitability and self-sufficiency has been subject to the same transformation. And, in more recent times, we have seen how actions have come to be taken in the name of such indicators as the Public Sector Borrowing Requirement - a complex and ambiguous indicator that emerged in the context of attempts to manage a constrained economy." (Hopwood, 1984, p. 169)

12. "The systematic collection of data about people has affected not only the ways in which we conceive of a society, but also the ways in which we describe our neighbour. It has profoundly transformed what we choose to do, who we try to be, and what we think of ourselves. Marx read the minutiae of official statistics, the factory reports from the factory inspectorate and the like. One can ask: who had more effect on class consciousness, Marx or the authors of the official reports which created the classifications into which people came to recognize themselves? These are examples of questions about what I call 'making up people'" (Hacking, 1990, p. 3)

13. Hill's book (1986) provides an interesting challenge to development economists who are all too ready to depend exclusively on official statistics and to rely heavily on generalisations which do not stand up to scrutiny if tested in serious fieldwork.

REFERENCES

Atkinson, A.C. (1985), Plots, Transformations and Regressions, Clarendon Press, Oxford.

Belsley, D.A., Kuh, E., Welsch, R.E. (1980), Regression Diagnostics: Identifying Influential Data and Sources of Collinearity, John Wiley & Sons, New York.

Caldwell, B. (1982), Beyond Positivism: Economic Methodology in the Twentieth Century, George Allen & Unwin, London.

Chambers, J.M., Cleveland, W.S., Kleiner, B., Tukey, P.A. (1983), Graphical Methods for Data Analysis, Wadsworth & Brooks/Cole Publishing Company, California.

Demptster, A.P. (1983), 'Purposes and Limitations of Data Analysis', pp. 117-133, in Box, G.E.P., Leonard, T, and Wu, C. -F. (Eds) (1983), Scientific Inference, Data Analysis, and Robustness, Academic Press, New York.

Diaconis, P. (1985), 'Theories of Data Analysis: From Magical Thinking Through Classical Statistics', pp. 1-36, in Hoaglin et al. (1985), op. cit.

Erickson, B.H. and Nosanchuk, T.A. (1979), Understanding Data, The Open University Press, Milton Keynes.

Gilbert, C.L. (1990) 'Professor Hendry's Econometric Methodology', pp. 179-303, in Granger (1990), op.cit.

Granger, C.W.J., ed., (1990), Modelling Economic Time Series, Clarendon Press, Oxford.

Granger, C.W.J. (1990a) 'General Introduction', pp. 1-28, in Granger (1990), op. cit.

Hacking, I. (1990), The Taming of Chance, Cambridge University Press, London.

Hamilton, L.C. (1990), Modern Data Analysis: A First Course in Applied Statistics, Brooks/Cole Publishing Company, California.

Harman, H.H. ([1965], 1989), 'The Inference to the Best Explanation', pp. 323-327, in Brody, B.A. and Grandy, R.E. (1989), Readings in the Philosophy of Science, Prentice Hall, New Jersey.

Hartwig, F. with Dearing, B.E. (1979), Exploratory Data Analysis, Sage Publications, London.

Hill, P. (1986), Development Economics on Trial, Cambridge University Press, London.

Hoaglin, D.C., Mosteller, F., Tukey, J.W. (1985), Exploring Data

Tables, Trends, and Shapes, John Wiley & Sons, New York.

Hopwood, A. (1984), 'Accounting and the Pursuit of Efficiency', in Hopwood, A and Tomkins, C (1984), pp. 167-187, Issues in Public Sector Accounting, Philip Allan Publishers Limited, Oxford.

Krüger, L. (1987), 'The Slow Rise of Probabilism: Philosophical Arguments in the Nineteenth Century', pp. 59-90, in Krüger et al. (1987), op. cit.

Krüger L., Daston, L.J., Heidelberger, M. (1987), eds., The Probabilistic Revolution, Vol.1, The MIT Press, London.

Myers, R.H. (1990), Classical and Modern Regression with Applications, 2nd edition, PWS-KENT Publishing Company, Boston.

Metz, K.H. (1987), 'Paupers and Numbers: The Statistical Argument for Social Reform in Britain during the Period of Industrialization', pp. 337-350, in Krüger et al. (1987), op. cit.

Moore, D.S. and McCabe, G.P. (1989), Introduction to the Practice of Statistics, W.H.Freeman and Company, New York.

Mosteller, F. and Tukey, J.W. (1977), Data Analysis and Regression: A Second Course in Statistics, Addison-Wesley Publishing Company, London.

Pagan, A.R. (1990), 'Three Econometric Methodologies: A Critical Appraisal', pp. 97-120, in Granger (1990), op.cit.

Rawlings, J.O. (1988), Applied Regression Analysis: A Research Tool, Wadsworth & Brooks/Cole Publishing Company, California.

Ridgeway, V.F. (1956), 'Dysfunctional Consequences of Performance Measurements', Administrative Science Quarterly, Vol.1, pp. 240-247.

Seers, D. (1976), 'The Political Economy of National Accounting', in Caircross, A. and Pur, M. (eds), Employment, Income Distribution and Development Strategy, Macmillan.

Spanos, A. (1990), 'Towards a Unifying Methodological Framework for Econometric Modelling', pp. 335-364, in Granger (1990), op.cit.

Stigler, S. (1986), The History of Statistics: The Measurement of Uncertainty before 1900, Harvard University Press, London.

Stigler, S. (1987), 'The Measurement of Uncertainty in Nineteenth-Century Social Science', pp. 287-294, in Krüger

et al. (1987), op. cit.

Swijtink, Z.G. (1987), 'The Objectification of Observation: Measurement and Statistical Methods in the Nineteenth Century', in Krüger et al. (1987), op. cit..

Tomlinson, J. (1981), Problems of British Economic Policy 1870-1945, Methuen.

Tukey, J.W. (1977), Exploratory Data Analysis, Addison-Wisley Publishing Company, London.