Models and techniques

J. de Leeuw
Departments of Psychology and Mathematics UCLA
405 Hilgard Avenue
Los Angeles, CA 90024
USA

In the situations typically encountered in the social sciences the methodology of traditional statistics is neither a good description of data analysis practice nor a good prescription to arrive at satisfactory summaries of the data. On the other hand the traditional statistical techniques are excellent data analysis tools, and statistical models are useful devices that can be used to develop and evaluate data analysis techniques. These general considerations are applied in this paper to evaluate the usefulness of techniques such as multidimensional scaling and correspondence analysis.

Key Words & Phrases: data analysis, description vs inference, exploratory and confirmatory, correspondence analysis, multidimensional scaling.

1. Introduction

In this short paper we discuss the traditional statistical approach to data analysis, considered both in its descriptive and normative aspects. We shall try to find out how realistic the statistical description of data analysis practice is, and how appropriate the normative statistical rules are. Let us begin by defining this traditional scientific procedure that is still advocated in many statistics books. According to the prescriptions we must start by formulating a theory about the phenomenon we are studying. This theory is formulated on the basis of prior knowledge that we have. It is then translated into a statistical model, this model is used to generate predictions about aspects of our data, and the theory of testing hypothesis is then used to find out if these predictions come true.

This model for the process of data analysis has been enormously influential, for various reasons. It was related to the concept of critical experiments, borrowed from the prestigious physical sciences, and it was a straightforward practical translation of Popper's enormously influential conjectures-and-refutations philosophy of science. Moreover it was backed up with a great deal of powerful mathematical apparatus, and certainly the earlier statistical techniques such as correlation, t-testing, and the analysis of variance made a lot of sense intuitively.

In this short note I will argue that the statistical model for scientific behavior is quite wrong, or at least of limited applicability. This argument has

both normative and descriptive aspects. I shall apply the general discussion to the particular case of homogeneity analysis, and indicate in which way this techniques differs from the usual statistical multivariate analysis techniques.

2. Classical statistics as description of scientific behavior

In this section we will briefly review in how far statistics, as defined above, is successful as a descriptive model of the data analysis process. If we look at the social and behavioral sciences, it is clear that the statistical model is extremely popular, but that there is an almost universal violation of its basic rules. It is not true that people first formulate a model, then collect data, and then perform statistics. Especially in the case of complex multivariate models such as regression or factor analysis the actual procedures are much more intricate. With computer programs such as LISREL (Joreskog and Sorbom, 1984), which implement even more complicated linear structural models, they become even less tractable. The model gets adapted in the process, various modifications are tried and rejected, new parameters are introduced, and so on. According to the standard model this means that the form of the model that is ultimately used is a random variable with a distribution over the possible models, and that this distribution must be taken into account in subsequent calculations. This is practically impossible, however, because the decisions made by the scientist cannot be formalized before the data are collected.

There have been various recent attempts to formalize the process of model choice, and to take the possible choices into account while computing sampling distributions (Dijkstra, 1988, reviews some of these attempts). But this is horrendously complicated, and it merely shifts the problem to another location in the process. There is no guarantee that investigators will stick to the options provided by the statistical models, in fact it seems clear that all of the rules one can possibly think of will be violated again and again. And, as we have seen many times in the non-scientific world, rules and laws that continue to exists although nobody obeys them and takes them seriously merely lead to hypocrisy.

Another problem, which is related to the first one, is that the models typically proposed by statistics are not very realistic. Especially in multivariate situations, and especially in the social and behavioral sciences, the assumptions typically made in the standard statistical models do not make much sense. Data are usually not even approximately normally distributed, replications are often not independent, regressions are not linear, items are not Rasch, and so on. This is perhaps not very serious because it appears to be a problem that can, in principle, be repaired fairly easily. But for the current practice it does mean that following the prescriptions of classical statistics easily leads to hypocrisy. The confidence intervals and tests of hypotheses of statistics are valid only if the model is true. Because we know that the model is never true, not even for an idealized population, it is not clear what we must do with this statistical information. This does not mean, by the way, that the models and corresponding techniques are useless. On the contrary, most of the established

statistical techniques are also very useful data analysis techniques. Otherwise they would not have survived. We merely must interpret our use of them in a different way than we are used to.

In the previous paragraphs we discussed in how far statistical models are realistic in a 'within-paradigm' sort of way. Given the general way of looking at such models, it turns out that they are not realistic in many important respects. But in fact the situation is much more serious. The whole framework of classical frequentist statistics is based on independent replications of experiments. Statistical statements are not about the data that we have observed, but they are about a hypothetical series of replications under exactly identical conditions. It seems to me that such statements are not interesting for many social and behavioral science situations, because the idea of independent replications is irrelevant. Different individuals or societies or historical periods are not replications from some sampling universe, they are essentially unique. There is no need to generalize to a hypothetical population. All we can require in situations like these is an appropriate description or summarization of the data which illustrates the points the scientist wants to make and which documents the choices that have been made.

3. Statistical priesthood

It pays to take a somewhat closer look at the normative aspects of statistics. Statistics, especially applied statistics, provides scientists with a number of clear cut rules of behavior, and tells them what they should not do in their data analyses. In these dark and uncertain times it is useful and soothing to have a number of clear cut rules to live by. This created the image of the statistical priesthood, borrowing a term introduced in a related context by Van Dantzig (1957), an image which is still alive, although it is slowly losing its force.

There used to be a time when statisticians and their cronies, the methodologists, always complained that they were consulted too late. Scientists only arrived at their offices after the data had been collected, i.e. after the damage was done. The implication was that a much better study would have resulted if the statistician had been consulted earlier. A rather safe statement, because it was obviously impossible to verify it. Another implication was that the data in the present study were all but worthless. With visible distaste the statistician fed them into SPSS. The client left with feelings of guilt, and with feelings of intellectual and moral inferiority.

Of course I am exaggerating here. Or, to put it somewhat differently, I am using a model. It could be true that this model is mainly relevant in the social and behavioral sciences, but I doubt that it is. It could also be that this is all in the past, and that I am flogging the the proverbial dead horse, but again my experience tells me otherwise. The situation in the social and behavioral sciences is not really worse than that in agriculture or biology. It is merely the case that in the social sciences the prescriptions of classical statistics make less sense.

Of course if the rules of statistics are prescriptive, in the sense that they tell a scientist what he should and should not do, then these rules cannot be wrong, but they can be impractical and often they are impossible to obey. Trying to impose the use of such rules where they do no apply is clearly not productive. It is true, by the way, that often the rules are prescriptive in a conditional way: if you assume A, then you must do B. Much of mathematical statistics and decision theory is formulated in this way. This is relatively harmless, it is sufficient to reply that you do not assume A, and consequently you do not have to do B. The statistical priesthood on the other hand has two counter arguments. The first one is that you must assume something, otherwise you can do nothing. The appropriate answer here is that this is nonsense. I can compute a mean value and I can draw a straight line through a cloud of points without assuming anything. I can also cross the street without first formulating a probabilistic model, and computing the probability that I will arrive at the other side in one piece.

The second 'priesthood' argument is that you do B, therefore you must have assumed A. If you use unweighted least squares you are a Bayesian with a flat prior, if you compute the mean you assume that your errors are normally distributed, if you sum the correct items you assume that the Rasch model is true, and so on. This is, of course, faulty and quite silly logic, based on the fact that the person assumes that nothing exists outside his own universe of discourse, a familiar trick in fundamentalist reasoning.

4. Stability and generalizability

Originally, of course, statistics was descriptive. This is not only true for the older demographic forms, but also for most of the work of Galton and Pearson. Although the notion of a probabilistic model is already very strong in Fisher's work, the emphasis on inference and cookbook forms of instant rationalism becomes prominent with the Neyman-Pearson, decision-theoretical, and Bayesian schools. In the meantime scientists continue to use statistical techniques for description, of course, although they are often forced to add some ritual statements about significance of the results and although they are forced to hide the most interesting part of their data analysis, finding the form in which they have eventually presented their results.

In order to prevent possible misunderstandings, we emphasize that the information that $z = 1.96$ is a useful descriptive statement, often more useful than the statement that the difference in means is 4.89. We could add the additional information that if the data are sampled from two identical normal distributions, and we repeat this sampling experiment an infinite number of times, then a value of $z = 1.96$ or higher only occurs in approximate 2.5% of the cases. It is not clear at all how relevant this additional information is, although it does provide some sort of scale on which the results of different experiments can be more easily compared.

It is often argued that the results of an experiment as such do not mean very much. Such results must be stable and generalizable. I agree, of course.

Everybody agrees. But the only appropriate way to find out if a result is stable is to replicate the experiment. If it is impossible to replicate the experiment, then perhaps the idea of repeated experiments does not make sense either. What statistics does is to provide information about stability under replications without actually carrying out the replications. It does this by substituting a mathematical model for actual empirical operations. It seems to me that there are many dangers involved in this practice, because so much hinges upon the appropriateness of the models. In the social sciences people have computed a lot of probabilities on the basis of statistical models, and these probabilities indicated that their results were significant, where the suggestive terminology merely meant that they were stable under replications. In the rare cases that replications have been carried out this often proved to be a rather optimistic assessment. It is not really necessary to illustrate this, everybody familiar with the history of the social sciences in his own field can think of hundreds of examples. In fact it seems to be the case that the social sciences clearly illustrate that there is nothing inherently cumulative and self-correcting about the development of any one particular science.

But more seriously, in many cases replications were not carried out at all, either because the fashions had changed and the topic was not interesting any more, or because replication was not possible because of the nature of the subject. But what do statements of significance mean if they cannot be verified, and are only based on embedding the actual data in a strange and unattractive framework that seems to have very little to do with these data. It is a truism that statistics cannot establish causality of relationship. It is quite incredible, by the way, that most people who quote this result are engaged on the very same page in trying to accomplish what they have just declared to be impossible. But in the same way statistics cannot prove the stability of a relationship or an effect either. Causality and stability come from careful experiment manipulation and replication within each of the empirical sciences, not from mathematical formalisms.

Our conclusions so far, on the basis of the above, can be summarized quite briefly. The task of statistics is to describe the results of empirical investigations and experiments in such a way that the investigator can more easily make his predictions and generalizations. Thus it is not the task of statistics to make generalizations. Statistical inference, whatever it is, is not useful for empirical science. Many statistical procedures are very useful, many statistical measures of fit provide convenient scales of comparison, and many statistical models provide interesting theoretical illustrations and gauges with which we can compare our actual data. But generalization, prediction, and control are outside of statistics, and inside the various sciences. Statistics has given us many useful tools and scales, but it has not given us a methodology to make the appropriate inductions.

5. SCALING

As a consequence of the above general discussion we can now formulate our point of view regarding scaling techniques, and so-called optimal scaling techniques such as homogeneity analysis or correspondence analysis in particular. A detailed description of these techniques can be found in GIFI (1981) or in VAN RIJCKEVORSEL and DE LEEUW (1988).

In the first place there are many ways in which these techniques can be introduced. It is quite possible, for instance, to think of a model for which homogeneity analysis provides consistent estimates. If we assume for instance that a scaling of the variables exists which linearizes all bivariate regressions, then homogeneity analysis finds this scaling. Thus if we combine HOMALS and LISREL, for instance, we have a technique which is consistent for a model in which we do not assume that the regressions are linear but merely that the regressions can be linearized. If somebody says that these optimal scaling techniques are not useful because they are not linked to a statistical model, and they cannot be tested for their truth, then there are two things we can do. If we have the time and the opportunity to do so, we can carefully try to explain to this person that the criteria he or she uses to evaluate the usefulness of data analysis techniques do not belong to the field of science but are more appropriate for religious gatherings. If there is no such opportunity we can point out that we accept his/her criteria, but they have not been applied correctly, because all we are trying to do is to estimate the parameters of a well defined statistical model.

A second way of deriving homogeneity analysis and related techniques is the idea of optimal scaling. In a generalized form it amounts to the following. Suppose we compare correlation matrices in terms of some numerical criterion, which can be the determinant, the eigenvalues, multiple or canonical correlations, or whatever. If transformations of the variables are allowed, then the correlation matrix and the numerical criteria based on the correlation matrix obviously are functions of these transformations. Given a criterion we can now look for the transformations or scores which maximize or minimize it. Homogeneity analysis for instance maximizes the largest eigenvalue of the correlation matrix, analysis of variance and regression techniques maximize the multiple correlation between the dependent and independent variables, and covariance structure techniques with transformation maximize the multinormal likelihood. There are innumerable variations of these optimal scaling techniques, some of them old, some of them new. For all of them it is true that they can be useful as descriptive tools or as predictive instruments, whether they are consistent with some model or not. Again most of the variations of these optimal scaling techniques are consistent with the model in which all regressions can be linearized.

The third way to introduce homogeneity analysis is to use ideas derived from multidimensional scaling (DE LEEUW and HEISER, 1980). We want to make a low-dimensional picture of the data in such a way that objects or individuals which have a lot in common in terms of the original variables are

relatively close together, while others are relatively far apart. Basically we look for points in the plane such that the within category distance of individuals is small and the between category distance is large, and this for all variables simultaneously. There is no need to think of a probabilistic model in this context, and there is also no need to mention optimal scaling of the variables. We merely have a plotting technique here, which tries to show the most salient characteristics of the data in a low-dimensional projection. The question whether we merely see capitalization on change is not very relevant here, because the concept can not even be defined without assuming the framework of classical statistics and without assuming some true model. This last way of introducing homogeneity analysis is the most honest one, the most interesting one, and certainly historically the most authentic one.

6. More on stability

In connection with the multidimensional scaling version of homogeneity analysis we can again discuss the problem of stability (or generalizability). The general idea of stability is very important in science, independent of its use in statistics. We usually do not want a small and uninteresting perturbation of our data to have a large effect on the results of our technique. But here are many ways to formalize the notion of stability mathematically without using probabilistic or statistical ideas. Continuity and differentiability, for instance, are also stability notions in this sense. With sufficient smoothness we can compute the derivatives of techniques, and we can look at the size of these derivatives. This is in most cases a useful undertaking, although there can very well be situations in science in which we want to model or detect instability.

One way of summarizing the size of derivatives is to use the techniques of statistical large-sample theory. If we assume that the individuals are a simple random sample from a population, then the size of the derivatives can be translated directly into the asymptotic standard errors computed by the so-called delta method. Other statistical frameworks lead to slightly different measures for the size of the derivatives (Wesselman, 1987). This means that we look at the usual confidence interval information in LISREL, Rasch, and so on, as quantities that summarize the stability of the techniques, once again given on a convenient scale with a sampling interpretation. It is well known that basically the same information is provided by resampling methods such as the Bootstrap and the Jackknife. In case this is not yet well known, these measures of stability are also available for homogeneity analysis, and they are fairly simple to compute (Van der Burg and De Leeuw, 1987).

The stability techniques above provide us with a convenient scale to assess the stability of the representations computed by techniques. In some cases a convenient scale for the goodness-of-fit statistics is also required. This corresponds with testing significance in classical statistics. Methods analogous to chi squared methods can be developed using derivatives, and resampling methods based on random permutations of the data are also possible. Again such techniques are also available for homogeneity analysis and related

techniques, although they are not incorporated in the current computer programs (DE LEEUW and VAN DER BURG, 1986).

7. CONCLUSION

Basing techniques on statistical models is an extremely useful heuristic device. Many other useful heuristic devices exist, for example those based on graphs and pictures. The statistical methodology 'behind the techniques' that is usually taught to harmless and unsuspecting scientists is a confusing and quite nonsensical collection of rituals. Many of the techniques work, quite beautifully, but this is despite of and certainly independent of this peculiar philosophy of statistics. And of course the fact that they work is not surprising, because the techniques of statistics are simply identical with the techniques of quantitative data analysis that have always been used in the sciences. Statistics *is* data analysis. This does not mean that we want to replace the academic discipline 'statistics' by the academic discipline 'data analysis', it merely means that statistics has always been data analysis.

REFERENCES

DE LEEUW, J. and W. HEISER (1982), Theory of multidimensional scaling, in: P.R. Krishnaiah and L. Kanal (eds.), *Handbook of statistics II,* North Holland Publishing Company, Amsterdam.

DE LEEUW, J., and E. VAN DER BURG (1986), The permutational limit distribution of generalized canonical correlations, in: E. Diday (ed.), *Data analysis and informatics IV,* North Holland Publishing Company, Amsterdam.

DIJKSTRA, T. (ed.) (1988), *On model uncertainty and its statistical implications,* Springer Verlag, Berlin (forthcoming).

GIFI, A. (1981), *Nonlinear multivariate analysis,* DSWO-Press, Leiden.

JÖRESKOG, K.G., and D. SORBÖM (1984), *LISREL VI; Analysis of linear structural relationships by maximum likelihood and least squares methods,* Scientific Software, Chicago.

VAN DANTZIG, D. (1957), Statistical priesthood I and II, *Statistica Neerlandica 11,* 1-16, 185-200.

VAN DER BURG, E., and J. DE LEEUW (1987), Use of the multinomial Jackknife and Bootstrap in generalized nonlinear canonical correlation analysis. University of Twente, department of Education, Research Report 87-7. (submitted for publication).

VAN RIJCKEVORSEL, J. and J. DE LEEUW (1988), *Developments in components and correspondence analysis,* Wiley, New York.

WESSELMAN, A.M. (1987), The population-sample decomposition method: A distribution-free estimation technique for minimum distance parameters, Dissertation, Erasmus University Rotterdam.

Received December 1987, Revised January 1988.