Soc. Sci. Med. Vol. 29, No. 2, pp. 253–258, 1989
Printed in Great Britain

0277-9536/89 $3.00 + 0.00
Pergamon Press plc

STATISTICAL ANALYSIS OF COMPLEX HEALTH AND SOCIAL DATA

SVEND KREINER

Statistical Research Unit, Faculty of Medicine, University of Copenhagen, DK-2200 Copenhagen, Denmark

Abstract—The selection of appropriate statistical models is a vital aspect of the analysis of complex health and social data. This paper examines issues related to model selection. It is concluded that the range of methods now available makes old controversies less relevant. Researchers may now look for strategies that combine the advantages of non-parametric and parametric approaches without automatically accepting the limitations of either. Examples of new directions are provided.

Key words—parametric methods, non-parametric methods, normal distribution, distribution free, graphical models

INTRODUCTION: AN OLD CONTROVERSY

As I understand it, the goal of the paper which I have been asked to write is to address methodological issues in research involving complex health data. The focus of the discussion should be on the selection of statistical models for the type of multidimensional data analysed in social science investigations in the health field. The paper should be geared toward researchers trained in social science/public health rather than toward statisticians and take up the subject of appropriate use of parametric and non-parametric statistical procedures. The latter issue is an old controversy that many statisticians find surprising.

Statistical methods have developed a great deal over the past 30 years, partly due to the advent of and improvements in high speed computers. Many sophisticated and innovative techniques have been made widely available in statistical packages like SPSS, BMDP, SAS, LISREL, GLIM, GENSTAT and many more. However, when one examines articles presenting results from survey research concerned with health issues, educational studies or indeed social science studies in the more general sense, one becomes aware that the available potential often is not utilized.

It is not that the statistical packages are not used. They are indeed very widely used. What is surprising is that the age-old controversy of choosing between so called non-parametric and so called parametric methods (usually interpreted as methods and models based on the normal distribution) is still with us today. It is perhaps the widespread, sometimes exclusive familiarity with and often uncritical use of normal distribution models that are the real sources of continuing discussions carried out in the name of the old controversy. Normal distribution procedures account for a quite limited part of the range of parametric methods which includes options based on other distributions as well as methods for testing non-linear relationships. Therefore I will review the old controversy on parametric versus non-parametric methods primarily form the standpoint of questioning its relevance and then try to look a little bit ahead, pointing to developments which may eliminate this controversy and bring about a unification of models, techniques and ideas, which today seem to represent completely different approaches to the statistical analysis of data.

The issue of choosing statistical models is discussed here in the context of routine analyses of interaction/association conducted on data from survey investigations concerned with health or various types of health services research where the findings may be completely distorted by the choice of an inappropriate statistical model or method.

PARAMETRIC VERSUS NON-PARAMETRIC METHODS

Looking back to the fifties two different points of view dominated discussions of statistical analyses of multidimensional data sets. In the *non-parametric* point of view it was stressed that a precise relationship should exist between the level of measurement and the permissible statistical operations (stemming from the seminal paper of Stevens [1]; see also Stevens [2]). The qualitative nature (nominal or ordinal) of a large part of the observations and measurements appearing in the social and behavioural sciences would then lead to reject models presuming interval scales, e.g. the normal distribution.

In the *parametric* point of view, equating 'parametric' with 'normally distributed', it was argued that we get the same "objective, impartial, and neutral judgement whether one resorts to the use of the parametric or non-parametric tests of significance" [3]. Others argued that treating ordinal variables as if they were interval, implicitly assuming that normal distribution assumptions could be applied, allows one to employ more powerful statistical techniques [4]. (It should be emphasized again that the predominance of the normal distribution has nothing to do with the fact that it is parametric. Many different parametric models exist for measurements giving observations among non-negative real numbers, among real numbers in intervals with well defined upper or lower bounds, or for measurements on the discrete set of integers.)

In terms of applied routine statistics the normal distribution has dominated the field, not because of the inherent type of measurement adhered to, but rather because it is a distribution where everything seems to work very easily. The ease of working with the model arises from the possibility for conducting data reductions in terms of means, variances and covariances, thus allowing direct calculations, without having to resort to cumbersome iterative procedures, which in practice are impossible without computers.

The choice of the normal, as *the* parametric model, was (and is) then not a question of the right model, but rather a (not irrelevant) question of convenience.

ON THE INEFFICIENCY OF NON-PARAMETRIC METHODS

Convenience when facing computationally intractable situations plays an important role for the non-parametric approach as well, but in quite another way. For the non-parametric statistician the problems which may be solved by the statistical analysis have to be very simple problems indeed. Problems of marginal association between two variables, or in some cases problems of partial association controlling for a third variable, are in reality the only kind of problems which are within reach from the field of non-parametric statistics.

Both approaches, dominated by convenience, have their own very different problems and are certainly open to criticism. Proponents of the parametric approach would criticize the non-parametric methods for being *inefficient*:

—Using available non-parametric methods one can very rarely use all of the information available on the measurement. Interval scales are treated as continuous ordinal scales, and qualitative ordinal scales are not infrequently treated as nominal scales.

—If a *correct* parametric model could be found (not necessarily normal) the power of the tests inferred by this model would most certainly be greater than the corresponding non-parametric tests.

—It is virtually impossible to take the confounding effect of more than one other variable into account within a non-parametric frame of reference. Conclusions concerning associations between variables will not take all available information into account, which may lead to directly misleading results. Simpson's paradox [5] is one well known but often disregarded problem. The often encountered cross-tabular behaviour, producing hundreds of two-way tables is the typical non-parametric way of handling multivariate problems.

Non-parametric methods are therefore *inefficient*. They seldom take all available information into account and they (almost) always have to reduce the problem to something which is so simple that the original research questions sometimes become almost unrecognizable. Or even worse, the research problems may be stated in such simple terms that they do not really make sense any more. As can be seen, severe limitations must be accepted to accommodate the statistical method.

Seen in this light the parametric methods based on the normal distribution seem enchanting. They are efficient. Given normality assumptions, they produce the most powerful tests. The best estimates can usually be derived and easily calculated. With parametric procedures models can be constructed for a large number of variables with complex recursive or causal structure, thereby making sure that all of the data have been taken into account when findings regarding associations between variables are put forward.

PROBLEMS WITH THE NORMAL DISTRIBUTION

Problems with the parametric statistical methods based on the normal distribution must be taken up in greater depth because they are both more complex and less obvious. Two levels of problems should be considered. One level relates to violating assumptions which constitute the scientific foundation of the theory and model used. The second level of problems focus on the restrictiveness of the methods and the resulting consequences in analysis of the type of data with which we are concerned.

Assumptions and robustness

A major problem with procedures based on the normal distribution is that in many cases they only look efficient. Criticizing this approach one would immediately refer to problems with the assumptions of linearity and normality on which the theory of the normal distribution is based. Methods based on the normal parametric model are just as restrictive as the non-parametric models—only in a less obvious and therefore more dangerous way.

One must remember that the statistical model is a mathematical logical system, and that *everything*—from derivation of test statistics and estimates to the final conclusions—are part of that system. The logical foundation of a model is the justification for using the model and any statistical techniques derived from the model. Knowingly violating the axioms of the system (the basic assumptions of the model) and proceeding as if nothing has happened goes against the very foundation of mathematical and logical thinking. The meaningfulness and validity of the analysis, and thereby the conclusions of the study, depend on the extent to which the mathematical conditions are met.

Putting it in diplomatic terms Cox [6] in his now classic book on analysis of binary data, discussing the use of linear models for binary data, remarks, "that the use of a model, the nature of whose limitations can be foreseen, is not wise, except for very limited purposes".

Thus it can be seen that procedures based on the normal distribution cannot be used in the analysis of data from typical survey research studies without presenting serious problems for judging the soundness of the results. The majority of variables are qualitative: nominal, binary or ordinal variables defined by a limited number of categories. Some index scales based on calculating numbers of responses of a given type on a set of separate questions (variables which might be characterized as discrete interval or ratio scales) may also be included.

The levels of measurement behind most of the variables are formally sufficient reason to reject the normal distribution as a realistic description of the random variation of most variables in a research study. Normal distribution theory presumes that we have continuous variables with a possible range from minus to plus infinity. Should one, for convenience reasons, want to use the normal distribution anyway, there are at least two other problems related to the model's assumptions. Many variables in social science and health studies are very heavily skewed with a large percentage of observations lying close to either the minimum or the maximum of the range of the variables. Consider for instance variables describing health, either perceived health measured qualitatively by a set of ordinal categories, e.g. 'very good', 'good', 'poor' and 'very poor', or quantitatively by e.g. the number of symptoms experienced within the last year. Distributions of this kind of variables are usually skewed with the majority of the population having relatively few symptoms and experiencing relatively good health. The distributions of these variables, and this goes for typical measurements of attitudes and abilities as well, are certainly very far from symmetric, one of the most important properties of the normal distribution. Secondly, we should not disregard the fact that qualitative categories are very often quite *arbitrary*. A different number of answer categories phrased in slightly different ways usually could have been used without changing what is being measured by the variable. Still, using any kind of linear model, e.g. the normal, requires that we are able to represent the categories by scores measuring the exact differences between categories. If equidistant scores are used—which in most cases happens quite automatically, we are in fact equating differences between 'very good' and 'good' and differences between 'poor' and 'very poor', which on the face of it seems quite problematic.

Should the question of allocating scores to qualitative categories seem minor, one has only to consider the arbitrariness of the choice of categories, e.g. the following three possible sets of categories for a question of experienced health:

Scale A:	'Very good'	= 2
	'Good'	= 1
	'Neutral'	= 0
	'Bad'	= −1
	'Very bad'	= −2
Scale B:	'Good'	= 2
	'Fairly good'	= 1
	'Neutral'	= 0
	'Fairly bad'	= −1
	'Bad'	= −2
Scale C:	'Very good'	= 3
	'Good'	= 2
	'Fairly good'	= 1
	'Neutral'	= 0
	'Fairly bad'	= −1
	'Bad'	= −2
	'Very bad'	= −3.

The scorings seem natural for all three scales and are, as long as we only use them as ordinal numbers, quite unproblematic. The moment we try to use them in a linear frame of reference we are in trouble. Scales A and B just might work at the same time, but one should not forget, that the 'good' of scale B probably includes the 'very good' of scale A, and that the semantic content of the categories therefore are not the same for scale A and B.

It is, however, impossible that scale C should work within a linear frame of reference at the same time as scale A and B as the distance between 'neutral' and 'good' in one case is double the distance from 'good' to 'very good', while the distance is the same in another case. Linearity then depends on the choice of categories. When this choice is arbitrary, it follows that the linearity must be arbitrary as well. Even if one of the three alternative scales did fit into a linear frame of reference we could never be sure that we had chosen the right one. There is more to the arguments than this, but it should be obvious that linearity should not be taken as a general paradigm for analysing qualitative data or arbitrary data cast in quantitative frames.

In defence of the parametric normal some argue, hoping against hope, that everything eventually will turn out right, that variables should be almost normally distributed and that methods based on the (multivariate) normal anyway are 'robust' in the sense that they will give the right end results even though the foundation of the analysis is shaky. They often even require that anyone questioning these methods should *prove* that the methods are *not* robust. This turns around the order of things. The burden of proof must *always* lie with the researcher who wants to violate the assumptions of the models or theories on which the methods to be used are based. This is really a heavy burden as no generally applicable results concerning robustness exist. We have results concerning robustness in very special cases and other results demonstrating no robustness in other special cases, but no general results. The reader is referred to Miller [7] providing a review of results for relatively simple designs (avoiding, however, multidimensional analyses and problems with binary or categorical variables) and Boomsma [8] reporting on a large scale investigation of the type of covariance models usually referred to as LISREL models. The results were certainly not generally in favour of the normal approach: LISREL was not robust with a sample size smaller than 100. It was recommended not to use a sample size smaller than 200. And LISREL was not robust "against discrete variables having a skewed distribution". Notice the emphasis on the problem of skewed distributions which, as pointed out above, occurs in many if not most distributions of health and attitude measurements.

Returning to the question of variables being approximated by a normal distribution, the rationale provided in some cases is the central limit theorem. What sometimes seems to be forgotten is that the problem of choosing the right distribution is a problem of sampling distribution, while the central limit theorem (and other limit theorems) deal with the asymptotic distribution of the mean (and other well-behaved *functions* of observations) as the number of observations increase to infinity. The central limit theorem then may result in robustness of estimates of certain parameters even in cases of non-normality, "while non-normality in certain not very special cases may lead the analysis into catastrophic errors" [9] due to problems with confidence intervals and distributions of test statistics.

As we in social research are not dealing with functions of several realizations of the same variable, but with sampling-distributions of separate variables, the central limit theorem gives us no assurance that the normal distribution always must be an appropriate distribution. Said another way, if a research question involves a simple regression problem with one single regressor, many different methods including both general linear models (GLMs) and generalized linear models (GLIMs) are available. The Central Limit Theorem (CLT) can then be used to show that means are asymptotically normal, and that certain simple tests (e.g. the t-test) will work even though the sampling distributions are not normal. However, we cannot solve complex multidimensional problems by calculating means and performing t-tests, so the CLT cannot solve our problem. Assuming that sampling distributions are multidimensionally normal goes beyond the CLT which cannot be used as a justification for using normal distribution procedures for all kinds of analyses. (Take the not uncommon case where variables such as gender are implicitly assumed to be normally distributed. Ridiculous, of course, so it seems necessary to say explicitly that the CLT cannot be used inappropriately.)

The matter may be different in econometrics when one is dealing with macro level phenomena and variables which are summary measures, like mean expenditure per family per year, mean income per family per year etc. But we are not dealing with econometrics. The variables encountered in health research are quite a different type of variables.

The restrictiveness of the normal distribution

Any kind of statistical model permits certain problems to be raised, certain questions to be asked, which at the same time make it impossible to address other problems. This is certainly true of the non-parametric methods, but it is also the case with the linear normal models. One should therefore always, when considering a specific statistical model, think about which types of problems the model implicitly disregards in relation to the research problem at hand.

The most serious deficiency of the linear normal distribution in this respect seems to be the restrictiveness concerning the nature of the associations we are able to investigate and describe. It is obvious that when one decides only to consider linear relations one will never discover non-linear relationships.

If our only goal were prediction, then optimal linear least square predictors (which may be determined analysing the covariance structure of the variables independently of assumptions of normality), may suffice. Even when relationships are non-linear, the linear least square predictors may from a pragmatic point of view be considered 'good enough'. The non-linear relations between variables may provide us with even better predictors, but it is—again from a purely pragmatic point of view—not always obvious that it would be worth the bother to determine them.

The point, however, is that we are usually interested in much more than prediction. The primary goal of scientific research is *understanding*, which usually goes far beyond prediction (although proper understanding would ultimately lead to optimal predictors, of course). And when understanding is what we are looking for, then routinely describing non-linear relationships as linear is not good enough.

The assumption of linearity is only one problem. One can anticipate that this specific problem may be solved by an appropriate transformation of data when relations are not linear, at least if relationships are monotonous. What is sometimes forgotten, however, is that interactions are always assumed to be invariant across differing values of control variables. One can only include first order interactions in a linear normal model. Higher order interactions between variables distributed according to the multidimensional normal distribution cannot be included, cannot even be formulated as a hypothesis to be tested. That is, once one moves beyond simple regression problems with single regressors, normal distribution models describing the random variation of multidimensional sets of variables are tested in terms of means and covariance matrices and therefore, do not permit higher order interactions among regressors.

Anyone used to working with survey and other social and health data using the so-called log linear models will know that higher order interactions are very much with us. They are in fact the rule rather than the exception. Furthermore, interactions are often peculiar—sometimes pointing in one direction, often disappearing given certain values of background variables and appearing given other values of the same variables.

The problem with the normal distribution is in fact not only that it does not recognize the possibility of complex interactions but that it has shaped people's thinking to such an extent that many often proceed directly to models including only two-factor interactions even when analysing qualitative data by log linear or logistic/logit models.

The normal model not only restricts the statistical analysis in many crucial ways, it also has restricted the thinking of people applying statistical methods to the extent that they sometimes seem to believe that no other models exist: "Because most of the variables that interest us as researchers are continuous at the conceptual level and are reasonably close to normally distributed in the population of interest..." [10].

It may be concluded that the normal distribution as a paradigm for parametric models is based on information which is usually not available or may even be known to be false, so that it only looks efficient. The choice between non-parametric and normal distribution procedures therefore seems to be a choice between inefficiency and pseudo-efficiency. It is not surprising that the controversy was never resolved.

THE DEVELOPMENT OF STATISTICAL METHODS SINCE THE LATE FIFTIES

Although differences of opinion on these issues continue to emerge, the range of options now available for statistical analyses of health data make the old controversy irrelevant. In the wake of the computer revolution, a whole range of models and techniques which were almost unthinkable possibilities a few decades ago are widely available. Parametric options include:

- the log-linear models for multidimensional contingency tables, providing the same kind of support for multidimensional qualitative data as the multidimensional normal does for quantitative data (limited only by the fact that standard programmes in practice seldom handle more than 6–8 variables at one time);
- the logistic regression model, a simple regression model with a binary dependent variable which is just as flexible as normal distribtuion models for simple regression problems with a continuous dependent variable;
- the Cox regression model for censored survival data, providing the means for survival analysis in terms of regression analysis with a large number of independent risk factors;
- the different latent trait models and item response models for validating psychological tests and index scales in general;
- the LISREL models describing—within a normal distribution framework—complex covariance structures.

The progress with regard to normal distribution procedures has been impressive. Thus it is surprising that routine statistical work often seems restricted to non-parametric and parametric modelling based on simplification of problems in the sense that (e.g.) complex regression problems with multidimensional response variables are turned into simple regression problems with only one response variable. Parametric procedures based on the normal distribution model are often still used regardless of the kind of data analysed.

One reason for the continuation of these tendencies may be the atomization of statistical methods into techniques with seemingly little in common. For example, both ordinary (linear) regression and logistic regression basically addresses the same type of problem, but they are still treated in most textbooks in applied statistics as procedures for two quite different types of problems. As long as such subjects are treated separately the choice between them seems to be very fundamental, and one may tend to familiarize oneself with one (the most promising) and not the other, and thereafter not really consider a choice for future analyses.

Not only may researchers now choose among a wide array of parametric options for addressing statistical problems, it is also possible to anticipate the kind of unification of the different statistical models and techniques which allows for the advantages of both approaches without automatically accepting the limitations of either.

GRAPHICAL MODELS: PARAMETRIC AND NON-PARAMETRIC MODELS

Recent research concerning a class of models termed 'graphical models' or rather 'block recursive graphical models' seems to provide a step in this direction, unifying most (all) known normal models for multidimensional data, models for contingency tables, logistic regressions and many more, while at the same time applying very weak assumptions satisfying the most cautious non-parametric statistician, and providing tools, for practical (computationally simple) problem solving.

The basic assumptions for the graphical models are assumptions concerning conditional independence of pairs of variables given *the rest* of the variables of the study (excluding of course variables following after the variables of interest if a recursive or causal structure is assumed). This very simple set of assumptions, in principle as simple as the assumptions characteristic of non-parametric methods, implies several useful results concerning collapsibility of the data on marginal distributions. Results which may be read directly off an 'independence graph' representing variables by vertices and conditional independence by missing edges.

Within this general class of models, association or interaction is defined as conditional dependence. The general class of graphical models may be specified further by assumptions concerning the type of variables and distributions. If all variables are qualitative we get a set of models defined as a subset of the class of log linear models for contingency tables [11–13]. Assuming all variables to be continuous and normally distributed leads to the so-called class of covariance selection models [14].

With both discrete variables and continuous variables we get a class of mixed models containing several standard models (regression models, models for analysis of variance etc.). Lauritzen [15] reviews the present state of the art of mixed graphical models and provides further references on the subject. Advantages of the graphical models are both conceptual and technical including:

(1) A unification of seemingly very different association models within a common framework.

(2) The possibility of describing—within the same formal framework—interactions in both *quantitative* (partial or conditional correlations for continuous data, cross-product ratios or odds ratios for qualitative data) and *qualitative* terms distinguishing between direct, indirect and spurious interactions, local and global interactions (global interactions being interactions which disappear for certain but not all values of conditioning variables), hidden vs marginal interactions, uniform vs non-uniform interactions (positive conditional correlations for some values, negative for other) etc.

(3) Models based on very basic assumptions (conditional independence and invariance of interactions).

(4) Independence graphs which work not only as nice graphical illustrations of the structure of the model (as in models for path analysis—which by the way fits comfortably within the framework of graphical models), but also as an analytical tool which may be analysed mathematically. The graphs thereby thus provide important information for shaping the direction along which the statistical analysis should proceed as well as information which should be used in the interpretation of the results.

Although graphical models have only been known for the past 10 yr and until now have only been used extensively in connection with purely qualitative data, there is no doubt that they will provide a very fruitful background for both unification of existing models for association and development of new ones. This includes non-parametric models which will be able to handle not only very simple problems but also complex problems with many variables for analysis of recursive models with a large number of qualitative variables, strategies which may be easily generalized to mixed models with many variables.

CONCLUSION

It is suggested that researchers analysing social science and health data should no longer accept limitations and restrictions which belong to a quite different era. What was good in the past is no longer good enough. This is not to downgrade the excellent work done with traditional methods, but rather to maintain that the compromises we have to accept (and we will never get beyond compromises) should be defined on today's terms and not in terms of what was once necessary.

REFERENCES

1. Stevens S. S. On the theory of scales of measurement. *Science* **103,** 677–680, 1946.
2. Stevens S. S. Measurement, statistics and the schemapiric view. *Science* **161,** 849–856, 1968.
3. Boneau C. A. A note on measurement scales and statistical tests. *Am. Psychol.* **16,** 260–261, 1961.
4. Bohnstedt G. W. and Carter T. M. Robustness in regression analysis. In *Sociological Methodology* (Edited by Costner A.), pp. 118–146. Jossey Bass, San Francisco, Calif., 1971.
5. Louis T. A. Analysis of categorical data: exact tests and long-linear models. In *Statistics in Medical Research* (Edited by Mike V. and Stantey K. E.), pp. 402–431. Wiley, New York, 1982.
6. Cox D. D. *The Analysis of Binary Data.* Methuen, London, 1970.
7. Miller R. G. Jr. *Beyond Anova, Basics of Applied Statistics.* Wiley, New York, 1986.
8. Boomsma A. On the robustness of LISREL (maximum likelihood estimation) against small sample size and non-normality. Rijks-universiteit te Groningen, 1983.
9. Borgatta E. F. and Bohnstedt G. W. (Eds) Level of measurement: once over again. In *Social Measurement. Current Issues,* pp. 23–38. Sage, Beverly Hills, Calif., 1981.
10. Darroch J. N., Lauritzen S. and Speed T. Markov-Fields and log linear models for contingency tables. *Ann. Statist.* **8,** 522–539, 1980.
11. Wermuth N. and Lauritzen S. Graphical and recursive models for contingency tables. *Biometrika* **70,** 537–552, 1983.
12. Edwards D. and Kreiner S. The analysis of contingency tables by graphical models. *Biometrika* **70,** 553–562, 1983.
13. Wermuth N. Linear recursive equations, covariance selection and path analysis. *J. Am. Statist. Ass.* **75,** 963–972, 1980.
14. Lauritzen S. Mixed graphical models. *Scand. J. Statist.* In press.