ORSA/TIMS

CENTRE HEC-ISA

ASSOCIATION FRANÇAISE DU MARKETING

MARKETING SCIENCE CONFERENCE

JUNE 24 - 27, 1987

CENTRE HEC-ISA

Jouy-en-Josas (France)

Data Analysis in Applied Socio-Economic Statistics, with Consideration of Correspondence Analysis

Edmond MALINVAUD, Director General, INSEE, France

Data analysis in applied socio-economic statistics, with consideration of correspondence analysis(1)

E. Malinvaud

Introduction

Although the collection of statistics still leaves much to be desired, research workers often have a wealth of data at their disposal. Traditional teaching of statistics and econometrics stresses one general methodology for the use of data, namely statistical inference within the framework of a well specified stochastic model. Often, however, researchers have good reasons to be reluctant to commit themselves to one particular model right away , they need to first proceed to some exploratory analysis, so as to detect the main features of their data ; no model is then so well established that it should be given full precedence. But it would be naïve to consider such data analysis as immune from difficulties.

Two reasons lead me to speak of this subject for the present occasion. Your conference is held in a country where devotion to data analysis is particularly strong and moreover uses well codified methods that are still little familiar elsewhere. On the other hand, being both an econometrician and the French official statistician, I have a particular motivation to make the bridge between what may appear as being two opposite positions, the one close to subject matter scientists looking for results of statistical inference, the other at the source of the data base that is being collected with no commitment to any particular theory.

(1) Given at the Marketing Science Conference, Jouy-en-Josas, June 1987

.../...

In order to set the stage for what follows, I should like to present from the start an example. Efficiency in marketing often requires a good knowledge of the exact localization of various kinds of consumers. A characterization of the socio-economic composition of localities provides a useful information in this respect. In statistically developed countries, data exist coming in particular from population censuses ; but they are so plentiful that they do not directly exhibit a typology of localities which could usefully provide the first categorization for many market studies. The common distinction between urban and rural areas, although significant, is clearly too rough. A recent data analysis performed on the 36 000 French communes provides such a typology (see N. Tabard, 1985).

The data base takes advantage of the existence in this country since thirty years of a classification of professions that intends to be suited for all types of socio-economic studies in which the notion of social classes matters. For the 1975 census a high level of aggregation of this classification gives the breakdown of the population of each commune into 26 main categories defined with respect to the profession of the household's main earner. The large table recording the 26 proportions for the 36 000 communes was processed by a kind of canonical correlation analysis for qualitative data, called "correspondence analysis", which will be studied later in this paper. Each commune can then be characterized by fewer numbers than the 26 proportions, namely by the values of the main "factors" that the correspondence analysis has detected. A clustering technique applied to this simpler characterization leads to the classification of communes into 35 types that can be further aggregated into four main groups. These 35 types are moreover found to be quite discriminating with respect to a number of other observations, such as the population growth rate of the commune.

.../...

Whereas this example gives an idea of the scope and significance of data analysis, methodologists have to wonder more generally on the role and limitations of the approach, as well as on the best techniques to be applied. In the first part of this paper, I shall consider the history of these questions. The second part will be devoted more specifically to a statistical theory suited for correspondence analysis.

I. The role of data analysis

I could start here with a general discussion. But I think it more illuminating to survey the evolution of ideas over the last sixty years, paying attention mainly to the field I know best, namely economics.

During the twentieth century economists have taken the habit of looking at statistics for making their theories or analyses more precise ; at the same time systems of economic statistics have been built almost everywhere. This may have been the major change in my science during this century. But it is only natural that the basic principles to be applied for drawing conclusions from data did not emerge right away.

At first, collection of data was the main concern. But attention was quickly brought to the definition of indices, to the use of graphics and more generally to methods of descriptive statistics. Truly econometric studies began when empirical economic laws were claimed to result from regressions of the observed values of one variable against those of another one, or a few other ones.

Reflection about what was so achieved occurred about at the time when the Econometric Society was founded. Ragnar Frisch then wrote his "Pitfalls in the construction of statistical demand curves" (1933). But the way out of the pitfalls was not immediately obvious. Actually Ragnar Frisch's solution differed from the one that was adopted in the fifties by the maturing econometric movement. His "Confluence analysis" (1934) belonged to the realm of descriptive statistics, even though it was inspired by the concern of protecting the econometrician against effect of the errors that disturb his observations.

Descriptive statistics was also at the time the purpose of various research groups throughout the world. In particular in New York the National Bureau of Economic Research launched a major project concerning the empirical study of business cycles. What then appeared as masses of data were gathered, graphed and processed by ad hoc methods, the aim being to know the complex pattern of business fluctuations. The main outcome of this work was the heavy book of A. Burns and W. Mitchell (1946).

This research, and others of a similar kind, however raised issues with an inference content, such as : how to detect true turning points of a time series ? What were the useful leading indicators ?

Faced with these questions, with the queries about the meaning of fitted demand curves, with the discussions raised by J. Tinbergen (1939) first attempts at the macro-modelling of economies, a group of econometricians working mainly in Chicago at the Cowles Commission for Research in Economics came to the conclusion that the only sound way of solving these issues was by what they called "the probability approach" (see in particular T. Haavelmo, 1944).

This approach is simply adhesion to the principle that, in order to learn from a sample of observations, one must proceed within a prespecified stochastic model that correctly represents the generation of the data. This appraoch was bound to interest one day or another methodologists working on economic data since during the forties, fifties and sixties it dominated and stimulated other branches of applied statistics, its expression been already found in R. Fisher (1925) widely used book. But the young econometricians were often emphatic and even dogmatic in pleading for it. In T. Koopmans (1950), T. Haavelmo wrote : "The purely geometric properties of a set of points in the

sample space are insufficient as a basis for statistical inference. In fact, a sample of observations is just a set of cold, uninteresting numbers unless we have a theory concerning the stochastic mechanism that has produced them" (p.265).

An interesting methodological conflict occurred with a group of the National Bureau of Economic Research, when T. Koopmans (1947) wrote a twenty page review of the book by A. Burns and W. Mitchell (1946) under the title "Measurement without theory". (The article, as well as the following interchange about it between R. Vining and T. Koopmans was reprinted for instance in R.A. Gordon and L.R. Klein, 1965). The review concentrated on discussing the appropriate methodology for the analysis of business fluctuations ; it opposed the National Bureau empiricism, qualified as belonging to "the Kepler stage", to the structural equation approach, claimed to belong to "the Newton Stage". Nedless to say, considering when it was written, the review was definitely a plea for the Cowles Commission approach and a critic of the then most better known National Bureau methodology.

A long interchange between R. Vining, coming in defence of the Bureau, and T. Koopmans was published two years later about this review. A substantial part was taken by a discussion about the potentialities of a structural system built by aggregation of individual demand and supply equations derived from maximizing behavior. But the most interesting parts concerned the respective merits of the empirical approach and "the probability approach".

Actually R. Vining was very moderate in his comments on Koopman's review. He wrote : "The discussion seems somewhat strained to me, and... I believe that one might raise the possibility that Koopman's argument contains a misleading emphasis if not an error... The work of Burns and Mitchell that

.../...

is being criticized purports to be a work of discovery and hypothesis-seeking, and it is not clear at all what the meaning of "efficiency" should be in this context. Statistical efficiency is an attribute of an estimation and testing procedure rather than of a procedure of search". He then, insisted at length on the facts that the book of Burns and Mitchell was explictly designed to outline certain methods adopted in an explorative study of economic variations, that factual systematized knowledge in economics was meager and that it was perhaps more important for economists to concentrate on problem 1, "the searching for regularities and interrelations of regularities and the feeling around for interesting theoretical models", rather than on the subsequent problems of testing, estimation and prediction. Although not giving up in the least his main line, Koopmans granded in his reply".. there remains scope for doubt whether all hypothesis-seeking activity can be described and formalized as a choice from a preassigned range of alternatives", i.e. whether all hypothesis-seeking activity can apply the probability approach.

But the latter soft expression of a doubt does not seem to have received much weight in the thoughts of econometricians during the following years. The previous quotation of T. Haavelmo shows a complete confidence on the necessity and universality of the probability approach. It is typical of a position consistently taken in econometrics during twenty years. The force of the movement was too strong for the rather cautious but thoughtful words of R. Vining to have changed its course.

I do not need to insist here on the enormous benefits that economics gained from adhesion of econometricians to the probability approach. Without it, most of the many quantitative results and tests that are currently reported in journals and books would not have been obtained. A good understanding of the approach must find its place toward the beginning of any

teaching on econometrics. Inference on economic phenomena will always be mainly based on it. However, some of R. Vining's doubts have been recently echoed within the econometric literature. Some have argued that current econometric practice is often weak at the specification stage.

At the beginning of his book, E.E. Leamer (1978) correctly states that "specification searching" often occurs, based on the same data to be later used for estimation, that this activity is not recognized by present teaching and that it is worthy of a systematic study intended at improving its methodology.

More particularly pointed to macroeconomic modelling, the critique raised by C. Sims (1980) also concerns the specification stage, which is said to often involve "incredible" hypotheses that uduly simplify what are known to be complex phenomena. One would overstate the critique if one would forget that, in all fields of science, useful hypotheses often are simplications. But the critique deserves serious consideration.

The positive proposals of C. Sims are also interesting in the present discussion. They suggest that econometricians should avoid introducing a priori restrictions and should concentrate their effort on a descriptive unconstrained study of the multidimensional stochastic process ruling the evolution of the main economic variables. Thus these proposals recommend an approach that has much in common with the National Bureau empiricism and with R. Frisch's attempts at describing "geometric properties of sets of points in the sample space", attempts that were considered as rather uninteresting by T. Haavelmo.

Some of the writings of C. Sims and other econometricians working with him seem to argue for a complete replacement of the traditional macroeconometric methods by the new multidimensional time series analysis they are promoting. Accepting to go as far would be tantamount to rejecting the probability approach. I had occasion to explain elsewhere why the arguments in favor of such a revolution cannot be accepted (E. Malinvaud, 1984). But, seen as providing a complement to present practices, the proposed analyses are quite valuable.

C. Sims' writings must properly be understood as a plea for a more conscious exploratory analysis of the data, before any model is specified. They then transpose to econometrics recommendations made for all fields of application by some mathematical statisticians who, following J.-P. Benzécri (1973) and J. Tukey (1977), now promote all kinds of unconstrained data analysis. As long as they are not understood as a negation of the probability approach but as stressing the importance of a well conceived first exploratory phase in any analysis of data sets, these recommendations are healthy.

One should clearly distinguish this interest for data analysis from a different concern, pointing to the lack of robustness of some of the most widely used techniques of statistical inference. Indeed, the very notion of robustness requires existence of a stochastic model. This model is less specific than the one in which the properties of the technique under discussion are usually studied ; it exists however as soon one speaks of the probability distribution of the statistics commonly computed and as soon as one looks for alternative statistics whose distributions would be less sensitive to departures from some of the classical hypotheses.

.../...

Let me sum up. When dealing with large data sets and before even considering making a true inference, statisticians often need to know, or are required to say, something about the structure of the data, what Haavelmo called "the geometric properties of the set of points in the sample space". Exploratory methods of data analysis provide an answer to this requirement ; hence, they become all the more useful as more bulky and more complex data become available, on which nothing can be seen at first sight.

But this analysis is subject to a very strong limitation : as long as it is not embeded within a stochastic specification, not even a very weak one, it cannot serve for any founded statement about the phenomenon that has been observed. Users of the data often want to hear or make such statements. Hence, one has to somehow resort to "the probability approach".

It is my view also, but this is more debatable, that, in teaching, the probability approach offers the advantage of clearly showing to the students some possible purpose of the analysis, on which he can concentrate his attention. But the drawback, namely that the student might assume the stochastic model to hold under all circumstances, has to be seriously kept in mind. After this warning, I hope the rest of my talk will not be found to be biased.

II. Correspondence analysis

The question of how to best balance the uses of the descriptive exploratory approach and of the inference approach is well illustrated by reference to the French experience with correspondence analysis.

The method is intended to be used for a descriptive analysis of large contingency tables. The research leading to its definition had the same spirit as the one of R. Frish (1934) proposing "confluence analysis" refered to above, except that it applies mainly to qualitative and not quantitative data, and that it is appropriate for the computer age. This research developed an approach already envisaged in the 1930's ; it was performed mainly by J.-P. Benzécri and a number of his students ; it led to a well codified methodology (see for instance Benzécri et al., 1973, or Lebart et al., 1984). The method was then widely diffused in France and applied on a wealth of data, often even mechanically ; many articles, published in French scientific journals concerning specific fields, use it for reporting the main content of the data set being discussed.

Interest for correspondence analysis is now spreading outside of France. It is being examined by statisticians who have long worked on qualitative data, such as L. Goodman, and progressively a more balanced view on its usefulness and limitations is emerging. It is a proper time for considering it in international gatherings.

I shall attack the subject from an angle that is opposite to the one commonly chosen in the data analysis literature. I am indeed going to concentrate attention on one particular stochastic specification, as if inference on it was fully appropriate. Proceeding in this way is illuminating, as long as one keeps in mind that the specification may be very remote from reality in some cases, even to the point of being misleading. The situation is

here the same as with the relationship between linear estimation theory and the practice of regression analysis. This angle of attack, however, neglects the presentation of many technicalities that the reader ought to know for an efficient application of descriptive correspondence analysis (on those, see for instance Lebart et al., 1984). Finally, I shall only give a sketch of the theory. Since this presentation is close to the one adopted by L. Goodman, I shall use his notation, rather than the one common among French experts.

In the standard case, to which I shall limit attention here, the data come from the observed counts of units classified according to two criteria. Except for the total number of observations N, they boil down to an I x J cross-classification table giving the frequencies p_{ij} of observations falling in the i-th row and j-th the column of the table (i = 1,2...I ; j = 1,2...J). The row and column sums are denoted respectively as $p_{i.}$ and $p_{.j}$.

It is proposed to deal with these data as if they where coming from N independent drawings from a multinomial distribution with IJ events of respective probabilities P_{ij}. In some cases this stochastic hypothesis may be quite close to the actual selection process of observations. In other cases a known sampling design may have been applied, with for instance stratification and clustering ; it is then possible to judge, at least heuristically, which results of the multinomial hypothesis are biased and how. In other cases still the hypothesis may appear as reflecting some from of subjective probability on the phenomenon under study, in the same way as the regression hypothesis often does.

Moreover, the probabilities P_{ij} have a special form that generalizes the independences case ($P_{ij} = P_{i.} P_{.j}$), namely they may be written as :

.../...

$$(1) \qquad P_{ij} = P_{i.}\, P_{.j}\left(1 + \sum_{m=1}^{M} \lambda_m\, x_{im}\, y_{jm}\right)$$

in which the parameters are the $P_{i.}$, $P_{.j}$, λ_m , x_{im} , y_{jm} and possibly also M. For identification of all these parameters, the following restrictions are stated, without loss of generality :

$$(2) \qquad \sum_{i=1}^{I} x_{im}\, P_{i.} = \sum_{j=1}^{J} y_{jm}\, P_{.j} = 0$$

$$(3) \qquad \sum_{i} x_{im}^{2}\, P_{i.} = \sum_{j} y_{jm}^{2}\, P_{.j} = 1$$

$$(4) \qquad \sum_{i} x_{im}\, x_{in}\, P_{i.} = \sum_{j} y_{jm}\, y_{jn}\, P_{.j} = 0 \qquad \text{if } n \neq m$$

$$(5) \qquad \lambda_m > 0$$

Conditions (2) imply that $P_{i.}$ is the row sum of the P_{ij} , i.e. the marginal probability of the observation falling in one of the J cells (i,j) for j = 1,2...J. The specification (1) implies that the rank of the matrix P is the minimum of the three numbers M+1, I and J. Without loss of generality we shall assume :

$$(6) \qquad I \underset{=}{<} J$$

Moreover we shall concentrate attention on the case in which M+1 < I ; the case M + 1 = I imposes no restriction on the matrix P, which is then simply reparametrized by the right hand side of (1) (the model is then said to be "saturated" by L. Goodman).

.../...

Correspondence analysis may be said to estimate the parameters, for a fixed M, by minimization of :

$$(7) \qquad U = \sum_{ij} \left[p_{i.}\, p_{.j} \right]^{-1} \left[p_{ij} - P_{ij} \right]^2$$

The I estimated M-vectors x^*_i then somehow exhibit the departure from independence of the i modality when it is confronted with the J-classification. In actual correspondence analysis one often considers the case M = 2 ; on a plane a graph of the points x^*_i shows how close are two distinct modalities i and h of the I - classification ; this is then seen much more easily than on the table of the p_{ij} , especially if J is large. Similar comments can be made on the graph of the M - vectors y^*_j . I have not the space to dwell here on the interest of such graphs, which is well recognized and actually explains the success of correspondence analysis (on this, see L. Lebart et al., 1984).

At this stage I hasten to add that my presentation deviates on a number of points from an exact description of the method, as it is actually proposed and used. I chose this deviation for simplicity, considering the purpose of this paper.

Minimization of (7) leads to estimate $P_{i.}$ and $P_{.j}$ respectively by $p_{i.}$ and $p_{.j}$, while the λ_m , x_{im} and y_{jm} are estimated by M solutions of the following system, those corresponding to the M largest λ after $\lambda = 1$ (except for marginal cases, the system has I solutions satisfying (5)) :

.../...

(8) $$\sum_j y_j p_{ij} = \lambda x_i p_{i.} \qquad i = 1,2...I$$

(9) $$\sum_i x_i p_{ij} = \lambda y_j p_{.j} \qquad j = 1,2...J$$

(10) $$\sum_i x_i^2 p_{i.} + \sum_j y_j^2 p_{.j} = 2$$

(It is easy to check that the solutions of this system fulfil the conditions (2) to (4)).

The system (8) - (10) may be written in matrix form. Let u be the (I+J) vector with components x_i and y_j , p be the IJ-matrix of the p_{ij} , A be the block-diagonal matrix :

(11) $$A = \begin{bmatrix} 0 & p \\ p' & 0 \end{bmatrix}$$

B be the diagonal matrix with I+J rows and columns, the diagonal elements being the $p_{i.}$ and $p_{.j}$. Equations (8) to (10) can then be written as :

(12) $$Au = \lambda Bu$$

(13) $$u'Bu = 2$$

or equivalently :

(14) $$Cu = \lambda u$$

(15) $$u'Bu = 2$$

.../...

where the matrix $C = B^{-1} A$ has the same structure as matrix A with, in the NE block, the conditional frequencies $p_{j/i} = p_{ij}/p_{i.}$ and in the SW block, the other conditional frequencies $p_{i/j} = p_{ij}/p_{.j}$. The form (14)-(15) is convenient for a theoretical study of the estimators.

One may show that, when the stochastic model holds and the number N of observations indefinitely increases, the estimators are consistent. One can even derive from (14)-(15) formulas permiting calculation of the asymptotic covariance matrix of the estimators. Since these formulas are a bit complex, I shall refrain from displaying them here. But I think they ought to be found in the technical literature upon which I am touching here.

Indeed, this is precisely where the inference approach begins to be useful. Any statistician gifted with a critical mind and often confronted with results and graphs of correspondence analysis comes, one day or another, to wonder whether some of the disparities exhibited by these results are significant and worth remembering. It was my case on a number of occasions and I gather that the motivation for L. Goodman to write his 1986 article had to do also with this feeling of uneasiness. But, before I proceed, I must warn the reader that this comment does not mean to be destructive. On the contrary, when proceeding to statistical testing, I found thus far that a large majority of the results I had been questioning appeared significant, at least when judged with the multinomial hypothesis, which may sometimes be too generous.

One important question to test is whether the model should not be applied with a smaller value of M than the one chosen for the fit. On a particular 4x6 tables L. Goodman (1986) found that M = 1 was enough while a correspondence analysis had been applied on the satured model and graphs corresponding to M = 2 had been presented ; in other words, λ_2^*, λ_3^*, the $x^*_{i_2}, x^*_{i_3}$, $y^*_{i_2}$ and $y^*_{j_3}$ were jointly insignificant (in the discussion published with the article it is, however, pointed out that correspondence

.../...

analysis is usually applied to much larger tables for which the hypothesis M < 2 is definitely rejected by the test).

A simple test, asymptotically satisfactory under the multinomial model, can be applied to this type of hypothesis. More precisely the χ^2 goodness of fit test is well-known : it amounts to testing model (1) with a given M against the satured model. The statistics is :

$$(16) \qquad D_M = N \sum_{ij} \frac{\left(p_{ij} - P^*_{ij}\right)^2}{P^*_{ij}}$$

to be compared with a χ^2 having (I-M-1)(J-M-1) degrees of freedom. Similarly, a test of a given M against a larger M' is obtained by comparison of $D - D$ with a χ^2 having (M'-M)(I+J-M-M'-2) degrees of freedom.

The difference between (16) and the criterion (7) minimized by the correspondence analysis estimates is noteworthy. The minimum value U^*_M of (7) differs from (16) by the denominators that are $P^*_{i.} P^*_{.j}$ instead of P^*_{ij}. The magnitude and even the sign of the difference $D - NU^*_M$ are difficult to judge a priori. This is why one should recommend systematic computation of the statistics D_M in the correspondence analysis programs. Lacking this computation one ought, however, to look at the value of the U^*_M since a small value of for instance $U^*_M - U^*_{M'}$ with M < M' will reveal likely insignificance of the components beyond those corresponding to M. This examination is very simple because it is easy to show that :

$$(17) \qquad U^*_M = \sum_{ij} \frac{\left(p_{ij} - p_{i.}p_{.j}\right)^2}{p_{i.}p_{.j}} - \sum_{m=1}^{M} \lambda_m^2$$

.../...

Hence, as soon as $N\lambda_M^2$ is small (by comparison with a χ^2 having I+J-2M -1 degrees of freedom), one should refrain from paying attention to the M-th components of the correspondence analysis. It happened to me to find cases were this simple rule was not applied. (Remember, however, that what is called λ_m^2 here is usually denoted λ_m in the technical correspondence analysis literature).

Statistical theory also permits to study whether better estimates than the correspondence analysis ones exist for model (1) with the multinomial hypothesis. Indeed, L. Goodman (1985) has shown that the asymptotically efficient maximum likelihood estimates somewhat differ from those of correspondence analysis (this was already shown in J.-C. Deville and E. Malinvaud, 1983). The marginal probabilities $P_{i.}$ and $P_{.j}$ are still estimated by the corresponding frequencies $p_{i.}$ and $p_{.j}$. Writing model (1) more simply as :

$$P_{ij} = P_{i.} P_{.j} (1 + z_{ij}) \tag{18}$$

one can moreover find the following maximum likelihood equations :

$$\sum_j \frac{y_j p_{ij}}{1 + z_{ij}} = 0 \qquad \sum_i \frac{x_i p_{ij}}{1 + z_{ij}} = 0 \tag{19}$$

these equations replacing (8) and (9) and having enough solutions with $0 < \lambda < 1$ for the selection of the M that jointly give the maximum value to the likelihood.

.../...

The nature of the difference between maximum likelihood and correspondence analysis can perhaps be seen through the following argument (it can also be seen by looking at the maximization criterion, as was done in J.-C. Deville and E. Malinvaud,1983). Since the $|z_{ij}|$ are usually small, much smaller than 1, let us accept the approximation of replacing $(1+z_{ij})^{-1}$ by $1-z_{ij}$. The first one of equations (19) can then be written as :

$$(20) \qquad \sum_j y_j \, p_{ij} = \sum_m \lambda_m \, x_{im} \, p_{i.} \sum_j y_j p_{j/i} \, y_{jm}$$

This would give exactly equation (8) if, of the M summations in j appearing in the right hand member, M-1 would be equal to zero, while the one corresponding to the value of m of the particular solution considered would be equal to 1 ; but this condition is not very different from the identification restrictions (3) and (4).

The maximum likelihood equations make a system of many non-linear equations. An iterative method for solving them was proposed in L. Goodman (1985). No doubt, one could imagine other iterative methods. Alternatively, one could use an asymptotically equivalent three stage estimate that was suggested to me by A. Monfort (personal communication) : the first stage applies a standard correspondence analysis ; the second stage minimizes :

$$(21) \qquad \sum_{ij} \frac{\left(p_{ij} - P_{ij}\right)^2}{p_{ij}}$$

as if the P_{ij} were constrained by a linear model, more precisely the linear model that is tangent to model (1) at the point found by the correspondence analysis estimate ; this linear model does not exactly fulfil the constraints of model (1) ; this is why a third stage occurs in which the fitted values of P_{ij} obtained at the second stage are processed through the correspondence analysis equations, with of course $P_{i.} = p_{i.}$ and $P_{.j} = p_{.j}$.

.../...

Once the maximum likelihood estimates, or their asymptotic equivalents, have been obtained, tests are easily defined. One can consider maximum likelihood ratio tests ; but χ^2 tests based on the statistics D_M defined by (16) should usually give just about the same answer.

This second part of my paper has sketched a theory of estimation and test for model (1), when the multinomial hypothesis applies. This theory ought to be extended by consideration of other sampling hypotheses, which does not seem to raise insuperable difficulties in principle ; but of course adequate estimates and tests may then become more complex, and this may prevent frequent application on the large tables with which correspondence analysis typically deals. This route would nevertheless be worth exploring.

References

J.-P. Benzécri et al. (1973), L'Analyse des Données, Dunod, Paris.

A.F. Burns and W.C. Mitchell (1946), Measuring Business Cycles, National Bureau of Economic Research, New York.

J.-C. Deville and E. Malinvaud (1983), "Data analysis in official socio-economic statistics", Journal of the Royal Statistical Society, Series A, vol 146, part 4.

R.A. Fisher (1925), Statistical Methods for Research Workers, Oliver and Boyd, Edinburgh.

R. Frisch (1933), "Pitfalls in the construction of statistical demand curves", Frankfurter Gesellshaft für Konjoncturforschung, Heft 5, Leipzig.

R. Frisch (1934), Statistical Confluence Analysis by Means of Complete Regression Systems, Okonomiste Institute, Oslo.

L. Goodman (1985), "The analysis of cross-classified data having ordered and/or unordered categories : association models, correlation models, and asymmetry models for contingency tables with or without missing entries, Annals of Statistics, p. 10-69.

L. Goodman (1986), "Some useful extensions of the usual correspondence analysis approach and the usual log-linear models approach in the analysis of contingency tables", with discussion, International Statistical Review, Dec. 1986.

R.A. Gordon and L.R. Klein, ed. (1965), Readings in Business Cycles, Irwin, Homewood, Illinois.

T. Haavelmo (1944), "The probability approach in econometrics", Econometrica, Supplement, July.

T. Koopmans, (1947), "Measurement without theory", Review of Economic Statistics, p. 161-172.

T. Koopmans, ed. (1950), Statistical Inference in Dynamic Economic Models, Cowles Commission Monograph N° 10, Wiley, New York.

E. E. Leamer (1978), Specification Searches - Ad hoc Inference with Non experimental Data, Wiley New York.

L.Lebart, A. Morineau and K.M. Warwick (1984), Multivariate Descriptive Statistical Analysis, Wiley, New York.

E. Malinvaud (1984), Comment on T. Doan, R. Litterman and C. Sims, "Forecasting and conditional projection using realistic prior distributions", Econometric Reviews, Spring 1984.

C. Sims (1980), "Macroeconomics and reality", Econometrica, January.

N. Tabard (1985), "Structure économique des communes, reproduction, consommation", Consommation, N° 1

J. Tinbergen (1939), Statistical Testing of Business-cycle Theories, League of Nations, Geneva.

J. Tukey (1977), Exploratory Data Analysis, Addison Wesley.