EXPLORATORY ANALYSES OF MULTIDIMENSIONAL CONTINGENCY TABLES

Nanny Wermuth

Institut für Medizinische Statistik und Dokumentation
Johannes Gutenberg-Universität, 65 Mainz, Federal Republic
of Germany

SUMMARY

During recent years it has become easy to estimate and test parameters in multidimensional contingency tables under different model assumptions. Nevertheless, there is a gap between the availability of computational results and the ability to explain model implications to the medical researcher. We suggest that checks for consistency and for plausibility of a given contingency table may narrow this gap. Thus, we propose to use simple comparisons and well-developed statistical methods in an exploratory manner.

1. INTRODUCTION

It is more difficult to understand the interrelations among many variables than among a few. This commonplace statement seems to be contradicted by recent developments in statistical methodology for analyzing multidimensional contingency tables (e.g. see Bishop, Holland and Fienberg [1975] , Goodman [1970], or Grizzle, Starmer and Koch [1969]). The ease with which computational results are available makes it appear as if there were only littly differences between analyzing the relations among ten or between three variables. If at all, this can be true only for contingency tables that are viewed as a set of numbers instead of as a set of observations. This distinction between numbers and data has been stressed by Finney [1975] :

In: Proceedings of the 9th Int. Biometric Conference, Boston 1976.
The Biometric Society: Raleigh.

> "Numbers chosen to illustrate a statistical method can have any magnitudes that do not contradict the mathematical model. Observational data are subject to many requirements of internal consistency and of plausibility in relation to previous information.... They "have real existency outside the statistician's file".

Whenever some data are viewed as numbers only, no long discussions are necessary about data collection, about missing responses or about outlying observations. This promotes and justifies the use of numbers to illustrate computational and formal aspects of a statistical method. Indeed, many more analyses of numbers than of observations appear in the statistical literature. As a consequence, too little is being taught about the analysis of actual observations.

The significance testing approach presents one important obstacle to discussing methods for data analysis. In order to keep a significance level at a prespecified level it cannot be permitted to look at a set of data in different ways or to analyze it repeatedly employing a variety of statistical methods. Therefore, whenever a fixed significance level is the most important guideline to a statistician, he can naturally not be concerned with the pecularities of a given set of data, with checking for consistency or with contemplating transformations of the data. When Tukey [1970] reinvented exploratory analysis he avoided a confrontation with classical statisticians by introducing new terminology and by postulating that tools for exploratory analyses are different from classical statistical techniques. The former are supposedly used to detect important aspects of a given set of data; the latter may be used to confirm hypotheses only. Inspite of this distinction. standard techniques have of course been employed in exploratory ways. In those instances the techniques tend to have one of the following qualifiers attached to them: heuristic, descriptive, or search procedure. These signal that one is concerned with data analysis instead of significance testing, even though test statistics and corresponding quantiles might be computed just like for a test of significance. This fact is a source of confusion to nonstatisticians and an offence to significance testing ideologists, but it does not hinder the fruitful use of these methods in data analysis.

Another, more serious obstacle to the successful teaching about analyses of actual observations lies in the nature of the task: the methods are likely to be unstructured, and they have to be tied closely to the specific set of data. Therefore, one given approach need not be useful for the next set of data.

This does apply to our suggestions in this paper as well. They originated from a particular kind of study: a large scale prospective observational study on pregnancy and child development that was started in 1964 in Germany. Since neither a sampling plan nor randomization was used to obtain the observations, many potential confounding factors exist. We are therefore hesitant to report interesting associations, immediately. Instead, we hope to detect effects of confounding factors by analyzing multidimensional contingency tables.

While the need for multivariate analyses is largest in the case of observational studies, they will have their place in controlled clinical trials or in experiments as well. With our suggestions we attempt to move away from mere formal analyses, so that eventually a feeling common among serious medical researchers is lessened, the feeling "that statistical wool is being pulled over their eyes" (Colton [1974]).

We investigate as an example the following question: is cigarette smoking during pregnancy related to an increased danger of perinatal mortality? We have available information on 452 classified variables for 7871 women, who entered the study during their first trimenon, returned for repeated check-ups and kept diaries on the course of their pregnancies and on the development of their children -if possible- up to an age of three years (compare Koller et al. [1974]).

2. THE OBSERVATIONS AND OUTSIDE INFORMATION

Except for the main two variables we believed at least seven additional ones (Tables 1 and 2) to be closely related. As a first step we present the rate of missing observations and relative frequencies for each of our nine variables. If we were to form a nine-dimensional table, only those cases with complete information on

TABLE 1

MISSING VALUES AND RELATIVE FREQUENCIES FOR SIX VARIABLES

Variables and missing values (m.v.)	Original categories and frequencies for all observation (n=7871 - m.v.)		Redefined categories	Frequencies		
				all observations	contingency table 1 (n=5945)	contingeny table 2 (n=6924)
1. Outcome of pregnancy (141=1,8%)	1. Child alive	87,7	1.=1	97,1	97,8	97,8
	2. Child dead	2,6	2.=2	2,9	2,2	2,2
	3. Abortion	9,7				
2. Length of gestation in days (139=1,8%)	1. less than 197	10,1	1.=2+3	9,4	9,0	9,2
	2. 197 to 250	4,5	2.=4	12,0	12,0	12,0
	3. 251 to 260	3,9	3.=5	78,6	79,0	78,8
	4. 261 to 270	10,8				
	5. 271 and more	70,7				
3. Mother's age (0)	1. less than 20	5,2	1.=2+3	31,0	32,1	31,8
	2. 20 to 24	25,8	2.=3	40,4	40,8	40,9
	3. 25 to 29	40,4	3.=4+5	28,6	27,1	27,3
	4. 30 to 34	20,9				
	5. 35 and more	7,7				
4. Living area (35=0,4%)	1. country or small town	56,8	same	same	57,4	
	2. larger town (100.000 inhabitants)	43,2			42,6	
5. Cigarette smoking during pregnancy (332=4,2%)	1. no	69,9	1.=1	69,9	70,2	
	2. occasional	13,3	2.=2+3	20,3	20,2	
	3. 1 to 5 per day	7,0	3.=4+5	9,8	9,6	
	4. 6 to 10per day	5,7				
	5. 11 and more "	4,1				
6. Mother's Schall-index (995=12,6%)	1. less than 0,95	26,2	1.=1	26,2	26,0	
	2. 0,95 to 1,15	54,4	2.=2+3	73,8	74,0	
	3. 1,16 and more	19,4				

TABLE 2

MISSING VALUES AND RELATIVE FREQUENCIES FOR THREE VARIABLES

Variables and missing values (m.v.)	Original categories and frequencies (n=7871 - m.v.)		Redefined variable and its categories	Frequencies in contingenc table 2
7. Previous	1. none	34,6	1.=none	35,7
pregnancies	2. one	31,9	2.=one: child alive	23,0
(7=0,01%)	3. two	18,6	3.=one: stillbirth or abortion	9,4
	4. three and more	14,9	4.=two and more: children alive	10,3
			5.=two and more: some stillbirths or abortions	21,6
8. Previous	1. none	95,2		
stillbirths	2. one or more	4,8		
(0)				
9. Previous	1. none	70,7		
abortions	2. one	19,9		
(0)	3. two and more	9,4		

all nine variables would be included. Summing up the missing value rates, we see that we were to loose at worst 20,3% of our original cases, an amount that could still be acceptable. But, since the total number of children that were born dead or that died within seven days after birth is 202 or 2,9%, and since only about 10% of all women smoked heavily the interesting events are too rare to analyze all nine variables simultaneously in a meaningful way.

The display of the relative frequencies also shows in which sense the observations represent a selected material. In our case, the relative frequencies indicate why the observations are not representative for all pregnant women in Germany: we have a relatively high percentage of women with their first pregnancies and of women with previous complications in pregnancies. They are overrepresented, presumably because they were looking for better than average care and because they expected to get this by participating in the study.

Generally, cause and effect relations may well be studied with observations that are not representative for a whole population. The only essential requirement is that the groups with the hypothesized causal factor absent and present do not differ with respect to confounding variables. Of course, there is little hope to obtain comparable groups in the face of too many confounding factors. Only if the potential confounding factors tend to be associated in a simple way, one might succeed in reducing the number necessary for simultaneous analysis.

It is wise to base information about potential confounding factors on outside information, and not only on associations in simple two-way tables from the own data. The reason is that a marginal association or the lack thereof may be produced by different associations in subgroups of all observations, that is by different partial associations. Just a reminder is that each multidimensional table itself can be regarded as the marginal table of some higher-dimensional one and that therefore it can share the deficiency just described for the two-way table.

For our question, we decided to begin with two six-dimensional contingency tables, the first one to learn more about the variables expected to interrelate mainly with cigarette smoking (variables 1, 2, 3, 4, 5, 6 from Table 1), the second one to give insight into the effects in our data of well-known determinants of perinatal mortality (variables 1, 2, 3, 7, 8, 9 from Tables 1 and 2).

3. THE CONTINGENCY TABLE AND THE ORIGINAL OBSERVATIONS

Even if we form only a six-dimensional contingency table we have to keep the number of categories per variable small in order to obtain as few zero cell observations as possible. This means that we have to compromise between the interest of the medical researcher, who wishes to obtain very detailed information, and the parsimony of the available data. We summarize neighbouring categories with low relative frequencies, for instance length of gestation 197 to 250 days with 251 to 260 days, and mother's age less than 20 years with 20 to 24 years. We combine categories the distinction of which is unimportant to our question: since cigarette smoking is expected to be related only to underweight, we redefine normal and overweight as one new category. Finally, we delete those categories and their corresponding observations that are not relevant to our question: abortions and length of gestation less than 197 days (compare Table 1). We form into a single variable those three variables that contain information about previous pregnancies (Table 2).

In a next step we try to reassure ourselves that we do not -as a result of this data manipulation- misrepresent the information contained in the original observations. A high missing value rate may be the reason why the subgroup of observations contained in the contingency table may be biased. This should show up in different relative frequencies as derived from all observations and from the contingency table. In our case (Table 1), we observe a pretty good agreement. The slightly lower rate in perinatal mortality is due to the fact that some stillbirths with a length of gestation under 197 days had not been classified as abortions.

This could easily be reconstructed from the routine comparison of all two-way tables before and after deleting and combining categories. We can also detect whether we have inadvertently changed the kind or the strength of an association by redefining the categories. In the case of a large missing value rate we prefer to compare the marginal associations of the multi-dimensional contingency table with the two-way tables obtained from all observations by taking two variables at a time. These simple checks for consistency produce as a byproduct a reasonable familiarity with the data at hand. Discussions about the plausibility of the marginal associations tend to follow naturally, e.g.: the marginal association between perinatal mortality and cigarette smoking in our contingency tables is mainly due to a smallest rate of perinatal mortality for women who smoke only occasionally or less than five cigarettes per day: a rather unlikely effect.

4. MULTIPLICATIVE MODELS AND COLLAPSING

For a multidimensional contingency table itself the main concern should be with the likelihood that its content can be reproduced in other studies. Some information on this can be derived from the number of cells with zero observations and from the comparison of each pair's marginal with its partial association given all other variables (see Table 3). Whenever differences among marginal associations appear levelled among the partial associations, then the observations are too sparse. On the other hand, a clear

TABLE 3

MEASURES FOR MARGINAL AND PARTIAL ASSOCIATIONS IN CONTINGENCY TABLE 1

Variable	Marginal			Partial		
pair	LR-χ^2	d.f.	p	LR-χ^2	d.f.	p
(1,2)	295.62	2	0.00	345.63	72	0.00
(1,3)	5.84	2	0.05	70.75	72	0.52
(1,4)	0.00	1	0.97	45.45	54	0.79
(1,5)	8.54	2	0.01	59.81	72	0.84
(1,6)	0.54	1	0.38	36.18	54	0.97
(2,3)	14.70	4	0.01	93.84	96	0.54
(2,4)	0.77	2	0.68	61.04	72	0.82
(2,5)	1.40	4	0.84	76.94	96	0.92
(2,6)	2.20	2	0.33	55.78	72	0.92
(3,4)	6.35	2	0.04	75.31	72	0.37
(3,5)	88.14	4	0.00	163.23	96	0.00
(3,6)	8.11	2	0.02	61.39	72	0.81
(4,5)	60.55	2	0.00	112.90	72	0.00
(4,6)	8.64	1	0.00	52.29	54	0.54
(5,6)	0.22	2	0.89	55.78	72	0.92

LR-χ^2 = Likelihood-ratio chi-square statistic
d.f. = degrees of freedom
p = quantile, corresponding to LR-χ^2 with d.f.

indication that the multidimensional analysis is appropriate are pronounced reversals in the strengths of associations. Especially interesting are situations in which some pairs appear marginally as unrelated but partially as strongly related. In our data there is a clear need of further condensation of the observations.

Thus, we can either combine further categories for some variables or we can try to reduce the dimension of the contingency table by collapsing (Bishop [1971]). To the latter end we use a model search procedure among multiplicative models that we have described previously (Wermuth [1976a,b], Wermuth, Wehner and Gönner [1976]). For our first contingency table we find a model denoted as 1235/456 to be well-fitting, and we apply to it the rules given by Bishop [1971] for collapsing variables. Thus, we may sum over variables 4 and 5 without introducing changes in the associations among the remaining variables. The plausibility of this step may be verified by looking at the measures for association in Table 3: variables 4 and 6 interrelate strongly only with each other and with one further variable with variable 5. We now can decide that these two variables do not represent important confounding factors.

For our second contingency table we cannot find a truly well-fitting multiplicative model. Therefore, it is not advisable to reduce the dimension of this table by collapsing. This leaves us with two four-dimensional tables that have three variables in common: perinatal mortality, length of gestation and mother's age, the fourth variable being cigarette smoking in the first table and information on previous pregnancies in the second table.

We wish now to compare the observed differences in perinatal mortality for the various situations. Percentages are best suited for this purpose, but can barely be trusted if they are computed from one hundred percent values of less than twenty observations. Therefore, we combine some more categories to present the data. Also, for the first contingency table, we increase the number of observations by going back to the original observations. Since we have deleted living area and weight index (variables 4 and 6) from the list of important confounding factors, we need no longer exclude cases with missing values for these two variables.

Tables 4 and 5 represent, then, our summary of the information in our data on perinatal mortality and cigarette smoking and on some confounding factors. One may use these Tables 4 and 5 in deciding on whether the question should be further investigated.

4. DISCUSSION

We have arrived at two sets of data that may be further studied. Thus, we end at a point where usually the

TABLE 4

CIGARETTE SMOKING AND PERINATAL MORTALITY DEPENDENT ON MOTHER'S AGE AND LENGTH OF GESTATION

Length of gestation in days	Mother's age in years	More than five cigarettes per day	Perinatal mortality		Number of observations
			absolute	relative	
197 to 260	less than 30	no	50	13,7	365
		yes	9	18,4	49
	30 and more	no	41	21,8	188
		yes	4	(26,7)	15
261 and more	less than 30	no	24	0,6	4036
		yes	6	1,2	465
	30 and more	no	14	0,9	1508
		yes	1	1,0	125

TABLE 5

PERINATAL MORTALITY AND INFORMATION ON PREVIOUS PREGNANCIES DEPENDEND ON MOTHER'S AGE AND LENGTH OF GESTATION

Length of gestation in days	Mother's age in years	Previous pregnancies	Previous stillbirths or abortions	Perinatal mortality		Number of observations
				absolute	relative	
197 to 260	less than 30	no	-	23	15,4	149
		yes	no	18	15,1	119
		yes	yes	19	12,2	156
	30 and more	no	-	6	23,1	26
		yes	no	9	13,2	68
		yes	yes	34	29,3	116
261 and more	less than 30	no	-	10	0,5	1912
		yes	no	11	0,8	1412
		yes	yes	9	0,8	1184
	30 and more	no	-	6	2,1	282
		yes	no	5	0,7	705
		yes	yes	4	0,6	1682

illustration of methods for contingency table analysis just begins. As tools to explore our observations we employed simple comparisons of relative frequencies, comparisons of measures for marginal and partial associations, as well as a model search procedure and rules for collapsing variables in multidimensional contingency tables; subjective judgement was, of course, another important tool.

This kind of detective work is especially needed if effects of confounding factors are suspected. While this will mainly be the case in observational studies, it may easily happen in experimental situations as well: be it that randomization does not achieve what it is expected to do, be it that a high nonresponse or missing value rate destroys the original sampling plan and produces selection effects.

We hope to have stressed that an extensive dialogue between statistician and medical researcher is necessary -and possible- whenever many qualitative or classified qualitative variables are to be analyzed simultaneously.

ACKNOWLEDGEMENT

I wish to thank S. Koller for his suggestion to look at this particular set of variables. For technical and computational assistance my special thanks go to H. Nitsch and H. Gönner and for the careful typing of the manuscript to H. Bianco.

REFERENCES

Bishop, Y.M.M. (1971): Effects of collapsing multidimensional contingency tables. Biometrics 27, 545-62

Bishop, Y.M.M., Holland, P.W. and Fienberg, S.E. (1975): Discrete multivariate analysis: theory and practice. M.I.T. press, Cambridge

Colton, T. (1974): Statistics in Medicine. Little, Brown, Boston

Finney, D.J. (1975): Numbers and data. Biometrics 31, 375-86

Goodman, L.A. (1970): The multivariate analysis of qualitative data: interactions among multiple classifications. J. Amer. Statist. Assoc. 65, 225-56.

Grizzle, J.E., Starmer, C.F. and Koch G.G. (1969): Analysis of categorial data by linear models. Biometrics 25, 489-504

Koller, S. (1974): Schwangerschaftsverlauf und Kindesentwicklung. Unpublished intermediate report.

Tukey , J. (1970): Exploratory data analysis (limited preliminary edition). Wesley, Reading

Wermuth, N. (1976a): Analogies between multiplicative models in contingency tables and in covariance selection. Biometrics 32 (to appear: March)

Wermuth, N. (1976b): Model search among multiplicative models. Biometrics 32 (to appear: June)

Wermuth, N., Wehner, T. and Gönner, H. (1976): Finding condensed descriptions for multidimensional data. Computer Programs in Biomedicine (to appear; June)