METHODS FOR THE ANALYSIS OF CONTINGENCY TABLES IN ROAD SAFETY RESEARCH

Contribution to NATO Advanced Study Institute on Contingency table analysis for road safety studies, Sogesta Conference Center, Urbino, Italy, 18-29 June 1979.
In: Fleischer, G.A. (ed.). Contingency table analysis for road safety studies; Proceedings of the above mentioned conference, Part I.B., pp. 35-46. Sijthoff & Noordhoff, Alphen aan den Rijn, 1981.

R-79-24
S. Oppe
Voorburg, 1979
Institute for Road Safety Research SWOV, The Netherlands

INTRODUCTION

A course on contingency table analysis for road safety studies seems to be rather specialistic. To investigate whether or not this subject matter is worth to be selected as a subject for an ASI-meeting, it is important to know the nature of the data in road safety research. Moreover it is necessary to know the power of the techniques for the analysis of contingency tables (ct's) and their limitations. This lecture is an attempt to give an outline of both factors.

A ct, sometimes called a cross-table, is a table of counts. Observations are classified according to one or more characteristics. After the classification a table is achieved, with numbers of observations in each cell. In the analysis of such a table, one is mainly concerned with the distribution of the observations over the cells. The dominant aim in such an analysis is the separation of the systematic effects and the random fluctuations that together resulted in the observed numbers. Therefore a model is needed to describe the systematic effects and an error theory that deals with the deviations of the observed data from the data that are expected according to the model.

The investigator is primarily interested in the systematic effects. He wants to check assumptions with regard to the underlying structure or model, which process is called "hypothesis testing", or he wants to specify the underlying model structure, which is called "parameter estimation". In both cases the random component is important. The research worker has to investigate to what extend random fluctuations may have influenced his results. Therefore he needs a theory that copes both with the underlying structure and the nature of the random fluctuations.

As to this last aspect, much work has been done in the past. With regard to the model specification, recent research has resulted in many improvements that account for an increased attention to the analysis of ct's. The result of this research is also of great importance for the analysis of road safety data. Road safety is measured by road-traffic accidents. Sometimes alternative measures

are used such as the number of road-traffic conflicts, when no accident data is available. What follows regards to accident data and conflict data, because both consist of counts. Even when accident rates are used (accidents per head of the population, per road length, per vehicle mile travelled etc.) a measure of road safety results which is in fact based on counts.
If one concentrates on the fact that these measures result from counts, then the analysis will tend to an explanation of that number of counts in that particular cell of the table. This leads to questions like: "what is the probability of that number of observations in that particular cell?" A completely different approach is got if one stresses the point that he deals with a measure of safety, a safety score. The relevant question then seems to be: "what makes the score for that particular cell so high or so low?" Both types of analysis do appear in road safety research.
The analysis of ct's is primarily concerned with the distribution of observations over cells. The second approach leads in most cases to an analysis in the context of linear models, such as linear regression or the analysis of variance. The approaches however lead to completely different model assumptions. In the analysis of variance approach the expected value for the score in cell <i,j> of a two-way table, is generally supposed to result from two additive components, according to the row and column position of the cell:

$$E(x_{ij}) = r_i + c_j$$

In the analysis of ct's a multiplicative model is generally assumed, although additive models do exist. In many practical cases the choice of the model structure depends less on the nature of the data and more on factors such as the statistical complexity of the model, or the knowledge about or availability of the techniques. From a statistical point of view, the linear model has many advantages, resulting from the applicability of the theory of linear vector spaces on statistics and the statistical properties related with linear transformations. The presentation of the multiplicative

model for the analysis of ct's as a linear model for the log-counts resulted in the same advantages and largely accounts for the increased interest in the ct-analysis.
Before we go into more detail with regard to the model structure it seems good to recollect some of the basic theories of ct-analysis.

THE MULTINOMIAL MODEL

In probability theory there are a number of so called "ideal experiments" that may be used as a model for the analysis of ct's. One of these theoretical experiments leads to the multinomial model. The experiment consists of a number of repeated independent draws of a sample element from a population. Each draw resulting in one of several possible outcomes. To apply this model, the following assumptions are needed:

- the probability that a particular observation is classified into a certain cell is independent from the classification of other observations;
- the probability is the same for each observation;
- the observations are classified in one and only one cell, in other words the events are mutually exclusive and exhaustive;
- no assumptions are made about how the events occur. The occurrence of events is taken for granted.

These assumptions lead to the following probability distribution: The probability that from a number n of observations, x_i observations are classified in class i, is equal to:

$$P\ (X_1 = x_1\ \Lambda \ldots \Lambda X_i = x_i\ \Lambda \ldots \Lambda X_m = x_m) = n! \prod_i^m \frac{1}{x_i!}\ p_i^{x_i}$$

where p_i is the probability for each observation to be classified in class i. The expected number of observations in each class is equal to n.p; the variance is equal to $n.p_i.(1-p_i)$ and the covariance between the observations in class i and class j is equal to $-n.p_i.p_j$.

For road safety research it may be concluded that the model can be used for the analysis of ct's that consist of accident numbers. However, if the observations are numbers of cars involved in accidents or numbers of persons injured in accidents, then the model is not applicable because many observations are collected in groups and therefore not independent anymore.

For the analysis of ct's it is important to realise that the events are in fact composite events. The classification of an observation in cell <i,j> means a classification in row i and column j. This

results in a special restriction of the model. It is often assumed that the classification of an observation in row i is independent from the column where the observation is in and vice versa. Two events A and B are said to be statistically independent if the probability $P(A \cap B)$ that both events occur is equal to the product P(A). P(B) of the probabilities that each of the events occurs. The independence of an observation in row i from being in column j is therefore expressed in the model with the restrictrion that the probability p_{ij} of an observation in cell <i,j> is the product $p_i . p_j$ of the probabilities of an observation in row i and in column j. Thus the hypothesis of no interaction between rows and columns leads to a multiplicative model for the cell counts. The common used Chi-square test of no interaction results from this model. Here the probabilities p_i and p_j are estimated from the marginal row and column distributions of the table.

THE POISSON MODEL

Apart from the multinomial model there is another model that is in use in the analysis of ct's with road safety data. This is the Poisson model.
For the application of this model all the assumptions of the multinomial model are needed together with one extra assumption regarding the occurrence of observations. Here the observations are not taken for granted, but it is supposed that the occurrences are Poisson distributed with parameter λ. The probability distribution is written as follows:

$$P(k;\lambda) = e^{-\lambda}.\lambda^{k}/k!$$

which means that the probability of exactly k observations, given the Poisson parameter λ, is equal to the expression at the right side of the equation. If we make this assumption and further assume that the observations are multinomially distributed over the cells of the table with probabilities p_{ij}, then it is proved that the distribution of the number of observations in each cell is Poisson distributed with probability λp_{ij}. Otherwise it can be proved that if the number of observations in each cell is Poisson distributed with Poisson parameter λp_{ij}, then the total number of observations is also Poisson distributed with parameter $\lambda = \Sigma\lambda p_{ij}$.
Moreover it is easily proved that the conditional distribution of the observations over the cells, given the total number of observations, is a multinomial distribution with probabilities p_{ij}. In formula:

$$P\ (X_1 = x_1\ \Lambda \ldots \Lambda X_m = x_m \Big| \sum_i^m x_i = k) =$$

$$= \prod_i^m \left\{ e^{-\lambda p_i}\ (\lambda p_i)^{x_i} / x_i! \right\} \Big/ \left\{ e^{-\Sigma\lambda p_i}\ (\Sigma\lambda p_i)^{k} / k! \right\} =$$

$$= k! \prod_i^m \frac{1}{x_i!}\ p_i^{x_i}$$

Thus, in this context, the multinomial distribution can be regarded as a restricted case of the Poisson model.
To investigate in which cases the Poisson assumption is satisfied, it is useful to start with the most common interpretation of the Poisson model.
If we apply the before mentioned theoretical experiment to the occurrence of observations, then a trial can be regarded as a unit period of time during which an event may or may not occur with probability p (more than one event may occur). If the time period is divided in n equal parts then we assume for each period of time that the probability of an event is $p_n = p/n$. This results in n trials each with probability p_n of an event to occur. The expected number of total events remains the same, being equal to $\lambda = n.p_n$.
If n tends to infinity, then the probability of more than one event becomes negligible and the trials may be interpreted as independent binomial trials.
It is proved that the limit distribution of the total number of events in this case is equal to the Poisson distribution.
The mean and variance of the Poisson distribution both equal λ.
The essential assumptions are that the occurrence of an event does not depend on the history of previous trials, or in other words that the trials are independent and that the probability of an event is equal for each trial, or that the events occur homogeneous in time.
The first assumption does not find much resistance in road safety research: accidents are rare events and seldom one accident causes another. In most cases where this however is true, one often regards such a chain of accidents as one (complicated) accident. The second assumption however troubles many investigators. Traffic flow changes rapidly over time and the accident rate is supposed to change with it. However, the homogeneity assumption needs not to hold over a long period of time. From the fact that the sum of Poisson distributed variables is again Poisson distributed, it follows that it is enough that the homogeneity assumption holds for a short period.
In many cases support is found for the Poisson assumptions. In cases where these assumptions does not hold, one sometimes assumes a mixed or compound Poisson process. In these cases the variance

exceeds the mean, which is true for instance with the negative binomial distribution. Here it is assumed that the sampling is from Poisson distributions with different parameters. The statistical properties of these compound distributions lead to serious complications as far as the analysis of ct's is concerned and therefore do not lead to practical alternatives.

The Poisson distribution is used in many investigations of road safety research. Recently the multiplicative Poisson model has been used for the analysis of ct's with regard to accident data.

Rasch (1973) applied the model to accidents classified according to road categories and days. Furthermore he used the model to estimate parameters for accident proneness of different drivers and to test whether or not these parameters changed with time. For these test he used the Chi-square test, based on the conditional Poisson distributions.

Hamerslag (1977) uses the multiplicative Poisson model to estimate the parameters for different classes of several accident characteristics jointly, under the hypothesis of independence between the characteristics. Which is a rather strong assumption, that only can be released by combining variables in one new variable.

De Leeuw & Oppe (1976) used a weighted version of the Poisson model. It has been applied for instance to ct's with accident numbers collected over different areas, or periods of time. This model is rather similar to the Multiplicative Poisson model with unequal cell rates of Andersen (1977).

RECENT DEVELOPMENTS IN CT-ANALYSIS

We shall now come to a more systematic description of the recent developments.

Most of the applications of an analysis of ct's are instances of testing the hypothesis of no interaction in two-way tables. Probabilities of row and column classification are estimated from the row and column marginals.

From these values the expected number E_{ij} of observations in each cell is computed as $n.p_{i}.p_{j}$ and compared to the observed number of counts O_{ij}.

A measure of discrepancy X^2 between both series of values called Chi-square is defined as:

$$X^2 = \sum_{i,j} (O_{ij} - E_{ij})^2 \Big/ E_{ij}$$

Under the assumption that the hypothesis of no interaction is true, the value of X^2 depends only on random fluctuations. The distribution X^2 is therefore assumed to be based on the properties of the multinomial distribution only.

For each total number of observations and each model specified in terms of cell-probabilities, there is a discrete set of possible values of X^2, the probability of each value depending on the probability of the corresponding set of cell-observations, given that specified multinomial model. Computation of these exact probabilities is rather cumbersome. Only in very restriced cases tests based on these exact distributions of X^2 are of practical use. Fishers exact test for 2x2 tables is an instance of these tests. Therefore, in practice additional assumptions are made in order to arrive at the distribution of X^2-values. In fact it is assumed then that the value of X^2 is distributed as the sum of a number of squared standard normal variates. This results from a well known limit theorem for the multinomial distribution. Practical values of this theoretical distribution denoted with χ^2 and known as the Chi-square distribution, can easily be found from tables if the number of squared independent standard normal variates is known. This number is often

called the degrees of freedom (df). The use of these tables is only warranted when relatively large numbers of observations are present. Therefore one often speaks of large sample tests.
Not much is known about the usefullness of the χ^2-distribution in small samples, other than with 2 * 2 tables. There are a number of handrules for usefullness in some situations. An overview can be found in Cochran (1952).
Oppe (forthcoming) investigates the behaviour of maximum likelihood and modified minimum Chi-square estimates for log-linear parameters and related X^2-distributions for a number of tables with small expected cell counts. This is done by means of the Monte Carlo method.
The neccessity of large samples often brings investigators in a rather difficult position. In many cases there are only small numbers of accidents on which statistical analysis can be based. This is one of the reasons why the Chi-square test did not get much attention for a long time. There is another reason. The Chi-squared test as described above is used as a test of no interaction. If this hypothesis has to be rejected, then the test does not tell us in what way the model fails to describe the data, but only that it fails to do so. In other words, the Chi-square analysis is a poor instrument for theory building.
Another problem often mentioned in hypothesis testing, using the Chi-squared test, is the fact that small and uninteresting deviances from the null-hypothesis will lead to a rejection of that hypothesis when a very large number of data has been collected. Thus besides the question of significance, there is the question of relevance. This problem stresses the need for parameter estimation and computation of their confidence regions as information additional to hypothesis testing.
The reason why the interest in the analysis of ct's is increased specially since 1960 is not because the problem of small numbers is solved but mainly because of a more detailed model testing procedure and the increased possibilities for parameter estimation.
There are four main developments that must be mentioned:
1. The application of the Chi-square test as a test for the hypothesis of no interaction, is generalised to ct's with more than

two ways of classification. Foldvary & Lane (1974) in measuring the effect of compulsory wearing of seat belts, used the method of partitioning the total Chi-square in Chi-square values for first, second and third order interactions in a four-way table.

2. Test are not restricted to overall effects, but Chi-square values are decomposed with regard to subhypotheses of the model. Not only with regard to the different levels of interaction between variables as used by Foldvary & Lane, but also within for instance the two-way table to partition the total Chi-square in Chi-squares with respect to (subgroups of) single classes. The advantage over earlier methods where the subtables are analysed as such, is that the partitioning of the total Chi-square is exact and results in independent Chi-squares for subhypotheses.

3. As a result of this last mentioned development one has to put more attention to parameter estimation also. Subhypotheses as mentioned are derived from constraints on the parameters in the model, such as linear constraints, quadratic constraints on the row-parameters etc.

4. The unit of analysis has been widened from counts to weighted counts, where the counts are weighted before analysis. The hypothesis testing of main effects is sometimes possible with weighted numbers, for instance if one wants to test whether or not there is a significant difference between the number of fatalities in different countries per head of the population.

However, the most important reason for using these extentions of the Chi-square method comes from the presentation of the theory in terms of linear models. The linear model is assumed to exist for the log-counts. Therefore it has been called a log-linear model. The log-linear model states that the logarithm of the expected value of the cell counts can be decomposed as follows:

$$\ln(E(x_{ij})) = \mu + \alpha_i + \beta_j + \gamma_{ij}$$

If the parameters are not known, and must be estimated from the data, then it is always possible to find a perfect solution for the parameters of the above stated (saturated) model. Testing in-

dependence of rows and columns means that the hypothesis $\gamma_{ij} = 0$ $\forall_{i,j}$ is tested.

Stricter models such as $\ln(E(x_{ij})) = \mu + \alpha_i$ can be tested within the former model. It is also possible to test restrictions on parameters, such as linearity restrictions for instance on the α's. Moreover, the generalisation to higher order tables is going straight forward and tests of restricted models are conceptually clear. We shall not go into more detail now. A comprehensive description is given in Bishop, Fienberg & Holland (1975).

Sometimes the log-linear model is confused with the log-normal model. The log-normal model is used in the analysis of variance to stabilise the variance of the observations for all kinds of measurement. The strong resemblance between both techniques is best illustrated in Nelder & Wedderburn (1972). In their computerprogram GLIM they incorporated the log-linear model for the analysis of counts as a special case of the linear model. Only the log-transformation of the data is needed to apply the linear analysis.

This leads us back to the question about the additive and multiplicative interaction in ct's. Darroch (1974), who compares both models from a statistical point of view, gives the following representation of no interaction in a three-way table:

For the multiplicative model: $p_{ijk} = \alpha_{ij}.\beta_{ik}.\gamma_{jk}$ and for the additive model: $p_{ijk} \,/\, (p_i.p_j.p_k) = \alpha_{ij} + \beta_{ik} + \gamma_{jk}$.

This additive model was first introduced by Lancaster (1951). He used the model to partition the total Chi-square (that can be regarded as a measure of residuals) according to the α- and β- and γ-values. In this way the residuals are tested further according to an additive model for row and column and layer effects. This kind of a compromise between interaction seen as statistical independence and a decomposition of the residuals according to an additive interaction model leads to different results in tables of higher order than two-way tables. Despite some advantages of the additive interaction model, Darroch concludes to a slight preference of the multiplicative model based on the statistical properties only.

Oppe (forthcoming) applies a generalised linear model to analise

a table of accident rates according to wet pavement conditions and hourly traffic volume classes. He uses an additive model: $E(a^{*}_{ij}) = r_i + c_j$.

This model however was not applied to the datapoints a_{ij} themselves. A monotone transformation a^{*}_{ij} of the data was used that led to the best fit of the model. A description of this kind of monotone regression models is found in Kruskal (1965).

The transformation turned out to be logarithmic, suggesting a multiplicative model rather then an additive model. Therefore there are, besides the logical attractiveness of the multiplicative model and some statistical advantages, also emperical results that support the multiplicative model for this kind of accident data. In many cases these tests of the model are ignored and a poor fit, resulting from the use of an incorrect model is much too lightly interpreted as just random error. For instance, in linear regression models one seldom is interested in the magnitude of the error component. In log-linear tests, however, this is not the case. The test of the model implies assumptions about the magnitude of the error component which is a great advantage of the technique. It leads to a quicker rejection of an incorrect model.

REFERENCES

1. Andersen, E.B. (1977). Multiplicative Poisson models with unequal cell rates. Scand. J. Statist. 4.

2. Bishop, Y.M.M., Fienberg, S.E. & Holland, P.W. (1975). Discrete Multivariate Analysis; Theory and Practice. MIT-Press, London, 1975.

3. Darroch, J.N. (1974). Multiplicative and additive interaction in contingency tables. Biometrika, 1974, p. 207.

4. De Leeuw, J. & Oppe, S. (1976). The analysis of contingency tables; Log-linear Poisson models for weighted numbers. R-76-31. SWOV, Voorburg, 1976.

5. Foldvary, L.A. & Lane, J.C. (1974). The effectiveness of compulsory wearing of seat belts in casualty reduction. A.A.P., 6, p. 59-81.

6. Hamerslag, R. (1977). Het gebruik van het multi-proportionele schattingsmodel bij ongevallenanalyse. Ingenieursbureau Dwars, Heederik en Verhey, Amersfoort, The Netherlands, 1977.

7. Kruskal, J.B. (1965). Analysis of factorial experiments by estimating monotone transformations of the data. J.R. Statist. Soc., B, 27, 1965.

8. Nelder, J.A. & Wedderburn, R.W.M. (1972). Generalized Linear Models. J.R. Statist. Soc., A, 1972, p. 370.

9. Rasch, G. (1973). Two applications of the multiplicative Poisson models in road accidents statistics. In: Proc. of the 38th Session of the ISI, Wien, 1973.