Using Monte-Carlo-algorithms for the estimation of ß-error: alternative inference strategies for multidimensional contingency tables

Kelle, Udo; Prein, Gerald

Veröffentlichungsversion / Published Version
Sammelwerksbeitrag / collection article

F. Faulbaum (Ed.) (1994). SoftStat '93
Advances in Statistical Software 4. Stuttgart - Jena - New York: Gustav Fischer, 559-566

Using Monte-Carlo-Algorithms for the Estimation of ß-Error Alternative Inference Strategies for Multidimensional Contingency Tables

G. Prein and U. Kelle

SUMMARY

Small samples and sparse cell frequencies cause major problems for statistical modelling with categorical data: Sampling zeros or small expected frequencies can lead to situations where asymptotic approximations of test statistics will be inadequate. In such cases one resorts to the use of exact tests or Monte-Carlo-simulations. But also in this case, inference can yield problematic results, as the power of tests is often extremely low and will therefore lead to the rejection of theoretically plausible hypotheses on the base of poor empirical material.
In this paper an alternative modelling strategy for small samples using Monte-Carlo-algorithms is presented. This strategy is extending the asymptotic power approximations presented by Cohen (1977) or Agresti (1990).

1. SMALL SAMPLES AND CATEGORICAL DATA ANALYSIS

The following considerations result from our work in the Special Collaborative Centre 186 *Status Passages and Risks in Life Course* at the University of Bremen which is sponsored by the "Deutsche Forschungsgemeinschaft". In our department of Methodology and Statistics we are confronted with practical problems in the statistical analysis of categorical data that arise in some of our eleven empirical projects.

With categorical data modelling is often restricted to more or less sophisticated contingency table analysis. Compared with parametric test procedures, the power of non-parametrics is relatively low and, proportionally, the probability to commit a type-II-error is relatively high. This problem increases with small sample sizes or complex statistical models, leading sometimes to extremely small expected frequencies. In such cases, "reality" may be far more complex than a model that can be accepted in a goodness-of-fit-test. When too many expected cell frequencies fall below certain limits ("Cochran's conditions"), asymptotic approximations for test statistics such as χ^2 can no longer yield valid results. Statistical software packages such as SPSS usually warn the user if χ^2-tests are used for the analysis of two-dimensional tables, but

seem to ignore this problem in the case of multivariate modelling procedures. Many problems relating to small sample sizes such as sampling zeros or small expected frequencies can be solved by statistical algorithms using Monte-Carlo- or exact inference methods - such as StatXact or LogXact (cf. Mehta & Patel 1983 and Hirji, Mehta & Patel, 1987). Although their power seems to be greater in case of small samples than the one of asymptotics, the general problem of type-II-error remains unsolved with these programms as well: interaction effects in contingency tables which are highly plausible for theoretical reasons often can hardly survive significance tests if sample sizes are small or models are very complex. However we do not want to argue against significance testing as a rational strategy for model selection but would opt for strategies that take into acount that the power of a test can be very low if the statistical analysis is based on small samples[1].

In order to estimate the type-II-error (β) or the power of a test ($1-\beta$) it is necessary to know (1) the sample size, (2) the critical α-level and (3) the distribution of the relevant statistical parameters given the alternative hypothesis. The third condition poses serious problems for usual inference strategies: the null hypothesis is generally tested against an infinite set of possible alternatives and not against a specific one. This overemphazises the importance of the α-error.

Using loglinear models for the analysis of multidimensional contingency tables, one can specify a data model under the conditions of the alternative hypothesis. If a specified model has to be rejected although it seems to be plausible from theoretical grounds, the probability of a type-II-error should be estimated and be presented to the reader. Thereby, the plausibility of the tested model (i.e. the alternative hypothesis) can be judged: it might differ from the null hypothesis although some effects could be too small to justify the rejection of the null hypothesis. This could lead to the replication of an empirical study - or to other attempts to find new empirical evidence for the initial hypothesis. In any case, this strategy will prevent us from rejecting theoretically plausible hypotheses on the basis of poor empirical material.

2. Asymptotic Estimations of Power for χ^2-Tests

Modelling strategies for categorical data often use χ^2 or G^2 as test statistics. Agresti offers an approach of power analysis based upon noncentral χ^2-distributions (Agresti 1990, 241): to approximate the power of a χ^2-test for a given model M with ν degrees of freedom and a significance level α, he develops a four-step analysis (ibid.):

[1] Such strategies have been proposed by Witte (1980, 86 ff). The general question of test power analysis is discussed by Cohen (1977).

(1) Choose a hypothetical set of **true** cell probabilities $\{\pi_i\}$;

(2) calculate the cell probabilities **given model M** $\{\pi_i (M)\}$;

(3) calculate the non-centrality parameter λ (cf. formula 7.11 for Pearson and 7.12 for likelihood statistic);

(4) determine the probability to observe a χ^2-value lying above the critical α-level, i. e. $p[X^2_{\nu,\lambda} > \chi^2_\nu(\alpha)]$. In case one can expect the test statistic to be asymptotically chi-squared distributed, tables for this are easily available (cf. Haynam, Govindarajulu & Leone 1970).

Modelling with multi-dimensional contingency tables often requires a slightly different strategy of test power analysis as it is often not necessary to compute power statistics for the entire model. Most of the time one wants to test conditional independence: the probability that single model parameters are unequal to zero. In **model based testing** one would test the goodness-of-fit of a given model M_2 under the conditions of another model M_1. To approximate the power of tests for partial associations, the strategy described above has to be modified: In this case probabilities given the more complex model M_1 $\{\pi_i(M_1)\}$ would correspond to $\{\pi_i\}$ and the probabilities given the restricted model M_2 $\{\pi_i(M_2)\}$ correspond to $\{\pi_i(M)\}$.

Let us take as an example a hierarchical loglinear model M_1 containing three random variables A, B, and C and implying the interaction effects {AB} and {BC}. In order to test the partial association between A and B, one can compare this model with a model M_2 not including this effect; i.e. the model {A} and {BC}.

Let the empirical distribution, represented in a contingency table, be

	C1		C2		C3	
	A1	A2	A1	A2	A1	A2
B1	2	3	3	2	0	0
B2	1	10	13	28	3	6

The goodness-of-fit test statistic (G^2) of the restricted model M_2 calculated by SPSS gives a value of 4.66 and a corresponding p-value of 0.199. The corresponding statistic of the more complex model M_1 gives a G^2-value of 2.76 and a p-value of 0.251. The partial-χ^2 for the

{AB}-interaction[2] is 3.31. With 2 degrees of freedom its asymptotic p-value is 0.069. This value could seem to be too small and the goodness-of-fit-statistics of the two models too similar to reject the null hypothesis that this interaction effect is null, and therefore could lead to the rejection of the more complex model M_1. This type of decision is always connected with the risk of committing a type-II-error: this would be the case, if the more complex model M_1 fits the true population parameters better than the restricted model M_2 (the null hypothesis) while, due to the sampling error, the observed sample does not support the rejection of the null hypothesis.

Following Agresti's strategy for power approximations one would have to calculate λ using the relative frequencies expected under M_1 as "true" probabilities and the relative frequencies expected under M_2 as "probabilities under M". One then would look up the corresponding test-power using tables for the noncentral χ^2-distribution. In the example we presented here, λ was 1.84, α was 0.05 and the corresponding power value 0.18. So the risk of a type-II-error seems too high in this case.

3. Calculating test power with Monte-Carlo-algorithms

The strategy Agresti proposes for the estimation of the power is suited only for relatively large samples as it is using asymptotic χ^2-approximations. As for significance tests, this could become a problem for sample sizes if expected cell frequencies are getting small. Different authors propose different limitations for the use of χ^2-distributed test statistics for the analysis of contingency tables. The most common recommandation ("Cochran's conditions") is to calculate probabilities for a specific χ^2-distributed test statistic only if not more than 20 % of the cells contain expected frequencies lower than five. If sample sizes are small or the tested models are very complex this is often not the case - like in the example we have given above.

Concerning "ordinary" tests of significance most recent developments can help to overcome this problem: The use of complex algorithms for exact inference or the estimation of the parameters of the test distribution by using Monte-Carlo-methods would allow to calculate the needed probabilities if asymptotic tests would be not appropriate because of sparse cell frequencies or highly imbalanced contingency tables. But when we employ such tests as modelling strategies, it would be completely illogical to use asymptotics for the estimation of power. While such algorithms are now easily available for ordinary significance testing (cf. Mehta, Patel 1983), there

[2] calculated by SPSS by comparing model {AB}{AC}{BC} and model {AC}{BC}

are no programmes available using exact inference or Monte-Carlo-simulation for test power analysis.

Confronted with the necessity of constructing multi-variate models on the basis of small samples we decided to develop a method for the approximation of the type-II-error by combining Agresti's method described above and algorithms using Monte-Carlo-methods. The use of exact methods however was abandoned instead, because the investigated sample were still large enough to cause unacceptable long computation times.

In order to develop a transparent procedure we linked a standard spreadsheet (Microsoft Excel 4.0 under Microsoft Windows 3.1) with a Monte-Carlo-link-library written in C++. This was done to combine the advantages of a spreadsheet - an environment with an acceptable user interface and the possibility of macro-programming - with those of a fast and flexible programming language.[3] The algorithms we are actually using for the generation of random tables are a C++-clone of Boyett's FORTRAN-subroutine RCONT (Boyett 1979) and a second subroutine generating random tables with fixed margins and cell probabilities given by the alternative hypothesis (or the more complex model). C++ provides the use of dynamic memory allocation and, thereby, makes it possible to overcome certain limitations of RCONT - such as the restricted number of cases. The algorithms are published in a paper of Prein, Kluge and Kelle (1993). For the moment the use of these algorithms restricts us to the simulation of hierarchical loglinear models with direct estimates.[4] We are currently working on a modified version that permits simulation models having no direct estimates.

In order to test a multi-dimensional contingency table using this Monte-Carlo-algorithm, it has to be divided into subtables or nested tables. For each of them the minimal sufficient statistics have to be determined. A first C++-function is drawing a random subtable from a hypergeometric distribution. This has to be repeated for each subtable. The use of the spreadsheet then allows to put these subtables together and consequently produce a random table with fitted marginal distributions for the tested model. A second function then calculates G^2 or χ^2 for this table. This whole operation is repeated and so a test distribution equivalent to the non-central χ^2-distribution is generated. By comparing the resulting values for each random table to the

[3] The generation of random tables using the Excel macro-language is possible - but extremely slow: a Monte-Carlo simulation with 2000 tables would have taken several days on a personal computer with an Intel 486 processor.

[4] This is due to the initial conditions we were confronted with when the problem of low test power arose: There was a small, hierarchical loglinear model {AB}, {BC} containing three variables. Furthermore, this model had direct estimates, i.e. all expected frequencies could be calculated directly without using iterative algorithms.

critical value $\chi^2(\alpha)$, the probability to get a greater value than $\chi^2(\alpha)$ (i.e. $p[X^2_{\nu,\lambda}>\chi^2_\nu(\alpha)]$, the power of the test) can be calculated.

The application of this subroutine requires a minimum of programming knowledge, because it is used through an Excel-macro controlling the loop around these functions and passing over the model parameters. For the algorithm one can refer to the paper of Prein, Kluge and Kelle which is mentioned above. To apply it to a specific contingency table, you have to take the following steps:

(1) Put your empirical data into the spreadsheet.

(2) Calculate the expected frequencies according to the restricted model M_2

(3) Determine the critical χ^2-value for ν degrees of freedom.

(4) Start a Monte-Carlo-simulation for model M_1. If you have to divide the table into different subtables, define a Monte-Carlo-process for each subtable. Put the subtables together and calculate the noncentral $X^2_{\lambda,\nu}$ for each resulting table using the frequencies expected under the restricted model M_2 as expected frequencies.

(5) Calculate the proportion of non-central χ^2-values greater than the critical value $\chi^2_\nu(\alpha)$ in order to determine the power of the test $(1-\beta)$. The probability of a type-II-error can simply be calculated as $1-(1-\beta)$.

For the example we have given in the previous chapter, the power estimation using the Monte-Carlo-algorithm was 0.32. Compared to the value 0.18 the test power seems to be greater than indicated by asymptotic tests. Nevertheless, the probability to commit a type-II-error remains very high with 0.68, so that a further study should be carried out.

4. CONSEQUENCES

A modelling strategy including the computation of type-II-error and test power can yield four different outcomes shown in the following table:

In cases one and two consequences are evident. If the probability of committing a type-II-error is low and therefore the power of the test we use is high, we can follow ordinary inference strategies to decide whether a given model must be rejected. Cases three and four refer to those problematic outcomes we could not detect by using "traditional" significance tests:

- Case three could be the result of the use of simplifying models with large data sets: the assumption of independence can be rejected at significance level α, but the test power remains low. Therefore, a different - perhaps more complex - model might be more appropriate.

- Case four could be the result if small samples, complex models or detrended and sparse tables are used as empirical material: the null hypothesis - complete independence or in case of multidimensional contingency tables a model containing few interaction effects - cannot be rejected on a given α-level. As the probability for committing a type-II-error is high, the alternative hypothesis should not be abandoned. In this case it would be more appropriate to replicate the study using a much larger sample. In case this cannot be done it should be mentioned at least that the more complex model M_1 might be theoretically more plausible, but has a lower goodness-of-fit-statistic.

	$p(\chi^2)$ > critical α-level	**$p(\chi^2) \leq$ critical α-level**
power high **type-II-error low**	(1) Reject H_0 at significance level α. Accept H_1.	(2) Do not reject H_0 at significance level α. Do not accept H_1.
power low **type-II-error high**	(3) Reject H_0 at significance level α. H_1 may not be appropriate.	(4) Do not reject H_0 at significance level α. H_1 might be plausible as well.

The strategy we propose is meant to identify those cases (3 and 4) where the "conservative" significance testing can have devastating impacts for modelling: with small samples it can lead to the rejection of theoretically probable alternative hypotheses. On the contrary, an alternative strategy of inference combining ordinary significance testing with test power analysis can help the researcher to detect those cases where the empirical material is not suited as a basis for a rational decision for or against the tested hypotheses or models.

5. REFERENCES:

Agresti, A. (1990): *Categorical Data Analysis.* New York; Chichester; Brisbane; Toronto; Singapore: John Wiley & Sons.

Boyett, J.M. (1979): Algorithm AS 144: Random R x C Tables with Given Row and Column Totals. In: *Applied Statistics. Vol. 28,* 1979. p. 329-332.

Cohen, J. (1977): *Statistical Power Analysis for the Behavioral Sciences.* New York; San Francisco; London: Academic Press.

Haynam, G.E.; Govindarajulu, Z.; Leone, F.C. (1970): Tables of the Cumulative Chi-square Distribution. In: *Selected Tables in Mathematical Statistics,* ed. by H. L. Harter and D. B. Owen. Chicago: Markham.

Hirji, K.F.; Mehta, C.R. & Patel, N.R. (1987): Computing Distribtions for Exact Logistic Regression. In: *JASA 82,* p. 1110-1117.

Mehta, C.R. & Patel, N.R. (1983): A Network Algorithm for Performing Fisher's Exact Test in r x c Contingency Tables. In: *JASA 78 (382),* p. 427-434.

Prein, G.; Kluge, S. & Kelle, U.: *Strategien zur Sicherung von Repräsentativität und Stichprobenvalidität bei kleinen Samples.* Bremen: Arbeitspapiere des Sfb 186, Nr. 18.

Witte, E.H. (1980): *Signifikanztest und statistische Inferenz.* Analysen, Probleme, Alternativen. Stuttgart: Enke.