Comment: Exact Inference in Multidimensional Tables

Svend Kreiner

1. INTRODUCTION

Professor Agresti's authoritative and stimulating paper is filling a lacuna in the literature on exact conditional inference, and I find that he has done statistics, in general, and this reader, in particular, a great service in presenting this comprehensive survey on exact methods. I would just like to bring up a few additional points on the topic of exact inference in multidimensional tables.

It is correct that many problems remain to be solved before we have real exact inference in multidimensional tables. The situation is not completely hopeless, however. Limited exact inference is a practical possibility today, and it may have an important role to play in serious loglinear model building. To that end, however, it is necessary that we change our ideas about how strategies for loglinear modeling should work.

We have to distinguish between two different problems: (1) tests for higher order interactions and (2) tests for conditional independence. There seems to be no practical solution to the first problem in the predictable future. For the second problem, however, we have some solutions through Monte Carlo sampling. In connection with collapsibility, exact conditional tests may first of all be used for a larger number of situations than one perhaps would think possible at first glance. Second, they generalize without special problems to exact goodness of fits for decomposable loglinear models.

2. TESTING CONDITIONAL INDEPENDENCE IN MULTIDIMENSIONAL TABLES

Consider first a five-dimensional table, n_{ABCDE}, and the problem of testing conditional independence of two variables given the rest: $A \perp E \mid BCD$.

If nothing is assumed on a model for the five-way table, this is equivalent to fitting the loglinear ($ABCD$, $BCDE$) model against a saturated alternative. Under this assumption we may, without loss of information, combine the CDE variables into one stacked variable ($Z = B * C * D$) and treat the problem as if it were a problem of inference in a three-way table. The hypothesis we consider is the third of the hypotheses discussed by Agresti for $I \times J \times K$ tables. Exact conditional tests by Monte Carlo sampling of tables from the product multiple hypergeometric distribution is no problem with today's computing power. It works even for very large tables on IBM-PC compatible microcomputers.

We assumed nothing on a model for the five-way table. We may, therefore, think of the test of conditional independence as a non-parametric test. One is well-advised to be concerned with the power of chi-squared statistics in this case and use of ordinal statistics (e.g., the partial Goodman–Kruskall coefficient is strongly recommended whenever ordinal statistics are appropriate). We notice also, however, that there is no practical obstacle for tests of (AZ, ZE) against parametric alternatives for ordinal variables or models assuming constant AB-association (AE, AZ, ZE). These tests will, of course, be considerably more time-consuming than tests against the saturated alternative, because calculation of tests statistics requires iteration for each sampled table. That, however, is a problem where we can count on computer science for a solution, and it should not concern us here.

If a test against a nonsaturated model is required, we have to rely on collapsibility to guide us toward exact conditional tests. Assume for instance that we want to test conditional independence of A and E against a loglinear model (AC, CD, BD, ABE). The conditional distribution of the complete table given the AB, AC, CD, BD and BE marginals may seem inaccessible at first sight. Collapsibility properties implies that the test of (AB, AC, CD, BD, BE) against (AC, CD, BD, ABE) is equivalent to a test of (AB, BE) against the saturated alternative in the ABE marginal, because both models are collapsible in the sense discussed by Asmussen and Edwards (1983). Not only are the interaction parameters of interest, the same in the complete and marginal model, but estimates and test statistics will be exactly the same whether or not we calculate them in the complete or the marginal table. And, finally, the conditional probability of a test statistic defined on the ABE marginal, $T(n_{ABE})$, given the sufficient marginals,

Svend Kreiner is a Statistician and Senior Researcher, The Danish Institute for Educational Research, 28 Hermodsgade, DK-2200 Copenhagen North, Denmark.

are easily seen to be the same for the two models:

$$P(T(n_{ABE}) \mid n_{AB}, n_{AC}, n_{CD}, n_{BD}, n_{BE}) = P(T(n_{ABE}) \mid n_{AB}, n_{BE}).$$

This test is, therefore, also a test of conditional independence in a three-way table, where exact conditional procedures are available.

The weaker type of collapsibility discussed by Bishop, Fienberg and Holland (1975, page 47) implies that parameters of interest are recovered in the marginal model. Estimates and test statistics for these parameters may be different, however. Consider, for instance, a six-way table, n_{ABCDEF}, and the problem of testing conditional independence of A and B (AB, BC, CD, DE, AEF). We have collapsibility onto the $ABCDE$ marginal in the sense described by Asmussen and Edwards (1983), but the test of conditional independence in the five-dimensional table is not equivalent to a test of conditional independence in a three-way table. We notice, however, that (AB, BC, CD, DE, AEF) is parametric collapsible onto both the ABE marginal and the ABD marginal. Conditional independence in the six-way table, therefore, implies that A and B are conditional independent in both three-way marginals. If we have reason to doubt the validity of the asymptotic approximations for the five-way table, we would, of course, be well-advised to use an exact conditional test in one or both of the three-way tables. The likelihood ratio statistic will not be optimal compared with the likelihood ratio for the complete table; but, if ordinal statistics are appropriate, we may even find that we have greater power in the three-way tests than given by the likelihood ratio statistic for the complete table.

Situations like these may turn up many times during loglinear inference in multidimensional tables, thereby permitting exact conditional inference, at least to a certain degress. It requires, of course, that one takes the trouble to look for collapsibility or, even better, that the statistical software one uses does that automatically.

3. EXACT INFERENCE IN DECOMPOSABLE MODELS

One of the reasons that exact conditional tests are practical for tests of conditional independence in three-way tables is that both the saturated alternative and the model, assuming that two variables are conditionally independent, are decomposable models with direct estimates, test statistics and explicit formulas for conditional distributions. The computational problems are restricted because iterations are never needed. There is more to the decomposable models than that however.

Notice first that decomposable models are defined in terms of conditional independence assumptions. They belong to the class of loglinear graphical models (Whittaker, 1990), where they are characterized as graphical models whose independent graphs are triangulated. Edwards (1984) discusses model search for this class of models. It is an immediate consequence of the strong collapsibility properties that removal of an edge from a decomposable model is equivalent to a test of conditional independence in a marginal table against a saturated marginal model *if and only if the resulting model is decomposable*. One may, therefore, map out an exact backward elimination procedure for model search among decomposable models. The procedure starts with the saturated model. It then removes edges corresponding to interactions where tests for conditional independence are equivalent to tests for conditional independence against saturated alternatives in marginal models.

A second attractive property of decomposable models, as seen from the point of view of exact conditional inference, is that exact goodness of fit may be calculated using Monte Carlo sampling analogous to the exact tests for conditional independence.

Agresti presents two examples for three-way tables: tests of (X, Y, Z) and tests of (X, YZ) against the saturated model that are special cases of this property.

To construct exact conditional tests for decomposable models, we need a general algorithm for Monte Carlo sampling in this case. One such algorithm is based on partitioning of the simultaneous distribution of the complete set of variables into a sequence of conditional distribution. This will permit us to generate the complete table in several steps, thus generating larger and larger marginal tables.

As an example consider a five-dimensional table and the decomposable model (ABD, BCD, CDE). This model is defined by conditional independencies, $A \perp C \mid BDE$, $A \perp E \mid BCD$ and $B \perp E \mid ACD$.

In accordance with Whittaker (1990, Chapter 12), we may factorize and reduce the conditional probability for this table given the sufficient marginals in the following way:

$$\begin{aligned} P(n_{ABCDE} \mid n_{ABD}, n_{BCD}, n_{CDE}) &= P(n_{ABCDE} \mid n_{ABCD}, n_{ABD}, n_{BCD}, n_{CDE}) \\ &\quad \cdot P(n_{ABCD} \mid n_{ABD}, n_{BCD}, n_{CDE}) \\ &= P(n_{ABCDE} \mid n_{ABCD}, n_{CDE}) \\ &\quad \cdot P(n_{ABCD} \mid n_{ABD}, n_{BCD}). \end{aligned}$$

Define three stacked variables, $X = A^*B$, $Y = C^*D$ and $Z = B^*D$, and rewrite the probability as

$$P(n_{ABCDE} \mid n_{ABD}, n_{BCD}, n_{CDE}) = P(n_{XYE} \mid n_{XY}, n_{YE})\, P(n_{AZC} \mid n_{AZ}, n_{ZC}).$$

Both probabilities at the right-hand side of this equation are product multiple hypergeometric distributions of the same kind as considered in connection with exact inference in three-way tables. We may, therefore, generate a complete random five-dimensional table in a two-step procedure. The first step generates a four-dimensional *ABCD* marginal, whereas the second step uses this marginal to generate the complete table.

4. MODEL SEARCH STRATEGIES

Given the limited possibilities for exact conditional tests in multi- and high-dimensional tables, how should the statistical analysis be planned in order to capitalize on these possibilities?

Collapsibility is the key to exact conditional tests. Collapsibility will, in many cases, lead to either very simple tests of conditional independence against saturated alternatives in marginal tables, or—at the very least—to marginal tables, where exact conditional tests will be almost as good in terms of power as the optimal model-based tests. It, therefore, seems natural to design strategies that capitalize on collapsibility.

Standard strategies for loglinear model building prescribe reduction on the order of interaction before anything else. There are, of course, good and valid reasons for this approach, but is has the unfortunate side-effect that it destroys collapsibility right from the beginning of the analysis. Therefore, it is not to be recommended, at least for analysis of high-dimensional large, space tables.

To utilize collapsibility to the fullest, we suggest a strategy consisting of three different steps.

Step 1. Restricted Model Search for a Decomposable Model

This step searches for a parsimonious decomposable model that fits the data. All statistics calculated during this step address conditional independence hypotheses against (marginal) saturated models. Exact conditional tests should be used here.

Step 2. Restricted Model Search for a Graphical Model

This step takes the decomposable model as a starting point and searches for a more parsimonious graphical model. We are still testing conditional independence, but the situation will not always be as simple as in the first step. Parametric collapsibility will guide us to the smallest marginals, where exact conditional tests make sense, but they may imply a certain loss in power. Compared with the problems with asymptotics in large, sparse tables, the loss of power may nevertheless be the smallest of the problems we have to face.

Step 3. Search for a Hierarchical Loglinear Model

Finally, we search for a parsimonious loglinear model, starting from and using, collapsibility properties implied by the graphical model found above. This final analysis, in most cases, will not require much work. It is a question of determining the order of interactions in (hopefully) small marginal tables, where we can rely on asymptotic tests and estimates.

5. SOFTWARE

At least two programs have been developed offering exact conditional tests for multi- and high-dimensional contingency tables. Both are based on the class of graphical models, and both have strategies for model search along the lines described previously. DIGRAM (Kreiner, 1989) includes ordinal statistics and a comprehensive analysis of collapsibility properties. COCO (Badsberg, 1991) has the complete range of exact conditional tests for decomposable models, but only the standard likelihood ratio and chi-squared statistics. Both are available for IBM-PCs and compatibles running under DOS. COCO is also available in a version for UNIX workstations.

6. CONCLUSIONS

It has been the purpose of this comment to point out that exact conditional inference of multidimensional tables is a practical possibility today, despite the many shortcomings and remaining problems. I cannot but agree that we have so far seen nothing but the top of the iceberg—that is indeed more true for the multidimensional tables than for all the other problems covered by Professor Agresti's paper—but the view from the top of the iceberg does not look so bad.