Exact tests for two-way symmetric contingency tables

JOHN W. McDONALD[1], **DAVID C. De ROURE**[2] and **DANIUS T. MICHAELIDES**[1]

[1]*Department of Social Statistics* and [2]*Department of Electronics and Computer Science, University of Southampton, Southampton SO17 1BJ, UK*

A two-way contingency table in which both variables have the same categories is termed a symmetric table. In many applications, because of the social processes involved, most of the observations lie on the main diagonal and the off-diagonal counts are small. For these tables, the model of independence is implausible and interest is then focussed on the off-diagonal cells and the models of quasi-independence and quasi-symmetry. For ordinal variables, a linear-by-linear association model can be used to model the interaction structure. For sparse tables, large-sample goodness-of-fit tests are often unreliable and one should use an exact test. In this paper, we review exact tests and the computing problems involved. We propose new recursive algorithms for exact goodness-of-fit tests of quasi-independence, quasi-symmetry, linear-by-linear association and some related models. We propose that all computations be carried out using symbolic computation and rational arithmetic in order to calculate the exact p-values accurately and describe how we implemented our proposals. Two examples are presented.

Keywords: Exact conditional tests, inter-observer agreement, linear-by-linear association, quasi-independence, quasi-symmetry, square tables, uniform association

1 Introduction

A two-way contingency table in which both variables have the same categories is termed a symmetric table. Such square tables occur in a variety of applications, including studies of: change over time, where each individual is classified by the same categorical variable at two points in time; pairs of individuals, where each member of the pair is classified by the same categorical variable; respondents answering two similar questions in a survey, each question having the same answers; confusion matrices, where subjective categorical data are compared with the truth; and inter-observer agreement of subjective ratings. For other examples, see Bishop *et al.* (1975, Chapters 7–8) and Agresti (1990, Chapters 10–11).

In many applications, because of the social processes involved, most of the observations lie on the main diagonal and the off-diagonal counts are small. For these tables, the model of independence (I) is implausible and interest is then focussed on the off-diagonal cells and the models of quasi-independence (QI) and quasi-symmetry (QS). For ordinal variables, a linear-by-linear association (LLA) model can be used to model the interaction structure. For sparse tables, large-sample goodness-of-fit tests are often unreliable and one should use an exact test, i.e., a test which uses the exact distribution of the test statistic, rather than a large-sample approximation such as the χ^2 distribution.

Exact tests for log-linear models are generalizations of Fisher's exact test of independence in a 2×2 table to more complicated models and larger tables. Recently much progress has been made in developing efficient computational algorithms for exact tests, e.g., the network algorithm, which has been developed in a series of papers by Mehta, Patel and co-workers. See Mehta (1994) for references and a description of available software for exact inference. Agresti (1992) surveyed the current theoretical and computational developments of exact methods for the analysis of contingency tables. In the section on future research, he stated 'By the turn of the century, we should see advances in applicability of exact methodology for contingency tables at least comparable to those of the past decade. One does not need a crystal ball to predict that computer speed will continue to increase and algorithms will be further improved, so that tables not now feasible for analysis soon will be. In addition, it is reasonable to handle new types of categorical data, in particular, more complex relationships for larger tables in higher dimensions'. However, Agresti's review made no mention of exact tests of QI, QS and LLA and these tests are not available in existing software packages for exact inference.

An exact test of independence requires the enumeration of a population of tables, namely, all tables with row and column marginal totals equal to those of the observed table. For large tables or sample sizes, this enumeration is usually infeasible, especially with near uniform marginals. With very skew marginal distributions, the enumeration is much easier. It has not been appreciated that for tests of QI, QS and LLA the required population of tables is much smaller because these models impose additional constraints on the population of tables and the enumeration is much easier. In this paper, we first review exact tests and the computational problems that need to be solved. We then propose new recursive algorithms for exact tests of QI, QS and LLA and some related models. We propose that all computations be carried out using symbolic computation and rational arithmetic in order to calculate the exact p-values accurately and describe how we implemented our algorithms. Two examples are presented.

2 Exact tests for two-way symmetric tables

One approach to making inferences about parameters of interest, when there are nuisance parameters, is the conditional approach. If sufficient statistics exist for both sets of parameters, then the distribution of the sufficient statistics for the parameters of interest, conditional on the sufficient statistics for the nuisance parameters, does not depend on the nuisance parameters. This conditional distribution, termed the reference distribution, may be used for inference about the parameters of interest and the resulting test is called an exact test. Exact inference for contingency tables is reviewed by Agresti (1992). We consider exact goodness-of-fit tests for a class of log-linear models used by Agresti (1988) to model agreement between ratings in an ordinal symmetric table. Note that for testing the goodness of fit of a given log-linear model, all the model parameters are nuisance parameters and the parameters not in the model are our parameters of interest (set to zero under the null hypothesis).

A saturated log-linear model for a $r \times r$ symmetric table $\boldsymbol{Y}$ with either Poisson or (product) multinomial distributed cell counts Y_{ij}'s is

$$\log E(Y_{ij}) \; = \; \lambda \, + \, \alpha_i \, + \, \beta_j \, + \, \gamma_{ij} \qquad i, j = 1, \ldots, r. \tag{1}$$

The model of independence, I, corresponds to $\gamma_{ij} = 0$ for all i, j. Here the interaction parameters, γ_{ij}'s, are of interest while the constant λ and main effects, α_i's and β_j's, are nuisance parameters. The cell counts are sufficient statistics for the interaction parameters and the marginal totals are sufficient statistics for the nuisance parameters. The diagonal-parameter model, D, corresponds to $\gamma_{ij} = 0$, for $i \neq j$ and $\gamma_{ii} = \delta$ for all i. Here δ is an additional nuisance parameter with sufficient statistic $\sum_i y_{ii}$. The model of quasi-independence, QI, for the off-diagonal cells corresponds to $\gamma_{ij} = 0$, $i \neq j$. Here the γ_{ii}'s are additional nuisance parameters with sufficient statistics y_{ii}'s. The model of quasi-symmetry, QS, corresponds to $\gamma_{ij} = \gamma_{ji}$ for all i, j. Here the common interaction parameters are additional nuisance parameters with sufficient statistics the sum of symmetrically opposite cell counts, $y_{ij} + y_{ji}$ for $i \neq j$. The model of linear-by-linear association, LLA, corresponds to $\gamma_{ij} = \gamma u_i u_j$ for all i, j for known scores u_i. The model of uniform association, UA, corresponds to $\gamma_{ij} = \gamma u_i u_j$ for all i, j for known equal-interval scores u_i. For the model of LLA (and special case of UA), γ is an additional nuisance parameter with sufficient statistic $\sum_i \sum_j u_i u_j y_{ij}$. The models of LLA and UA may be considered for the cells off the main diagonal and are termed the quasi-linear-by-linear association (QLLA) and quasi-uniform association (QUA) models. Note than QLLA and QUA are special cases of the QS model, i.e., they are parsimonious QS models which have QI as the special case of $\gamma = 0$.

An exact goodness-of-fit test for I uses the conditional distribution of the cell counts given the margins. An exact goodness-of-fit test for D uses the conditional distribution of the cell counts given the margins and the sum of the diagonal counts. An exact goodness-of-fit test for QI uses the conditional distribution of the cell counts given the margins and the diagonal counts. An exact goodness-of-fit test for QS uses the conditional distribution of the cell counts given the margins, the diagonal counts and the sums of symmetrically opposite cell counts. An exact goodness-of-fit test for LLA uses the conditional distribution of the cell counts given the margins and $\sum_i \sum_j u_i u_j y_{ij}$, i.e., the 'covariance' between the scores. An exact goodness-of-fit test for QLLA uses the conditional distribution of the cell counts given the margins, the diagonal counts and $\sum_i \sum_j u_i u_j y_{ij}$.

An exact goodness-of-fit test is conceptually simple. Consider the set of possible tables with the same sufficient statistics for the nuisance parameters (the model parameters) as those of the observed table. This set is called the reference set and denoted by Γ, with subscript referring to a specific model if necessary. For a test statistic T and a value of T, T_{obs}, for the observed table, carry out the following steps: 1) enumerate all possible tables in Γ, 2) compute T for each table in Γ, and 3) calculate the exact p-value, $p_{obs} = P(T \geq T_{obs})$, by summing all null conditional probabilities of tables at least as extreme as the observed table.

3 Computational problems

Three computational problems need to be solved in order to calculate accurately the exact p-value: 1) how to enumerate Γ efficiently, 2) how to compare accurately the value of the test statistic for each table in Γ with the observed test statistic and 3) how to calculate accurately and sum the null probabilities of tables at least as extreme as the observed table. We consider each computational problem in turn and propose a new recursive algorithm for enumeration of tables with fixed margins as well as for enumeration of tables with fixed margins and fixed diagonal counts. We also propose the use of symbolic computation along with error-free rational arithmetic so as to calculate the exact p-values more accurately than existing algorithms.

3.1 Table enumeration

Verbeek and Kroonenberg (1985) survey algorithms for exact tests in $r \times c$ contingency tables with fixed margins. The algorithms for enumeration of Γ either involve straightforward filling in of the table counts, simulating a dynamic number of nested 'for-loops' or recursion. Only Boulton and Wallace (1973) proposed a truly recursive algorithm. Their ALGOL algorithm reduced the problem of enumerating all $r \times c$ tables with fixed margins to three subproblems: 1) enumerating all $(r-1) \times c$ tables with fixed margins (derived by collapsing the last two rows), 2) enumerating certain $2 \times c$ tables with fixed margins and 3) enumerating certain 2×2 tables with fixed margins. Unfortunately, computer languages most often used for statistical computation, such as FORTRAN 77 or earlier versions, did not support recursion, so this may explain why this algorithm has not been used by statisticians. Languages such as Lisp and C do support recursion. Also, versions of Lisp support distributed computing, which may be used to distribute subproblems to a network of computers or to a set of processors in a multiprocessor parallel computer. Distribution may be used for speed up of computations as well as for extending the range of feasible problems (De Roure and Michaelides, 1994; Michaelides, 1997).

For independence, we propose a new recursive algorithm for enumerating all tables with fixed margins. The algorithm of Boulton and Wallace (1973) does not generalize to the case of enumerating all tables with fixed margins and fixed diagonal counts, which is one of the computational problems for an exact test of QI. Our proposed algorithm decomposes the problem into generation of the top row and (recursive) generation of the subtable consisting of the remaining rows. After generating the top row of a table (or subtable), we subtract the counts in the top row from the corresponding column marginal totals. These adjusted column marginal totals form the new constraints for the generation of the subtable consisting of the remaining rows. The generation of the top row consists of filling in the feasible row counts from left to right, for each cell always starting with the highest possible value consistent with the row and column marginal totals. For cells with fixed counts, its value is subtracted from the corresponding row and column marginal totals and the cell is flagged as fixed. For tables with fixed cells, these fixed cells are skipped when generating feasible counts. This algorithm is described in further detail in the Appendix.

Note that $\Gamma_{LLA} \subseteq \Gamma_I$ and $\Gamma_{QLLA} \subseteq \Gamma_{QS} \subseteq \Gamma_{QI} \subseteq \Gamma_D \subseteq \Gamma_I$. For enumerating Γ_{LLA}, we propose an enumerate–and–reject algorithm, namely, enumerate Γ_I, then reject any table in Γ_I whose covariance between the scores does not equal those of the observed table. A similar algorithm may be used for Γ_D. For enumerating Γ_{QS}, we propose an enumerate–and–reject algorithm, namely, enumerate Γ_{QI}, then reject any table in Γ_{QI} whose sums of symmetrically opposite counts does not equal those of the observed table, so as to obtain Γ_{QS}. A similar algorithm may be used for Γ_{QLLA}. The efficiency of any enumerate–and–reject algorithm depends on the rejection rate and the efficiency in which the larger set of tables can be enumerated and each table tested for rejection. When algorithms for the direct generation of the desired Γ do not exist, this may be the only possible way to enumerate the desired Γ. A similar simulate–and–reject idea was used by McDonald and Smith (1995) and Smith and McDonald (1995) to carry out Monte Carlo exact tests for QI.

3.2 Comparing test statistics

Fisher's exact test of independence orders the tables in Γ_I by the inverse of their hypergeometric probabilities. This idea was extended to $r \times c$ tables by Freeman and Halton (1951) and this test is referred to as the Freeman-Halton test. This test uses as test statistic, $-2log(\gamma pr(\mathbf{y}))$, where $pr(\mathbf{y})$ is the null table probability and γ is a known normalisation factor. The Freeman-Halton test statistic has an asymptotic χ^2 distribution with $(r-1)(c-1)$ degrees of freedom.

Three common ways of measuring departure from the null hypothesis are the Pearson X^2 statistic, the log-likelihood-ratio L^2 statistic and the null table probability (Freeman-Halton test). As different test statistics may order tables in the Γ differently, they may yield different p-values. Kim and Agresti (1995) suggest using two statistics so that the actual size of the test is closer to the desired nominal size than would be possible using a single statistic. This reduces the potential conservativeness of exact tests resulting from the discreteness of the exact distribution. They calculate a modified p-value by first partitioning the sample space using primary test statistic T, then within fixed values of T, further partitioning the sample space using a secondary test statistic T'. Let T_{obs} and T'_{obs} denote the observed values of the primary and secondary test statistic. Kim and Agresti defined their modified p-value, p^*, as

$$p^* = P(T > T_{obs}) \; + \; P(T = T_{obs}\, ,\, T' \geq T'_{obs}). \tag{2}$$

Note that if the primary statistic, e.g. L^2, depends only on the sufficient statistics under the alternative, then any other statistic, e.g. X^2, which also depends only on the sufficient statistics under the alternative cannot be used as a secondary statistic, as two tables that have the same value of the primary statistic will have the same value of the secondary statistic. Thus, we must base the secondary statistic on a more general alternative, in order to have a secondary partitioning of the sample space, e.g., we can use L^2 as the primary test statistic and the Freeman-Halton statistic as secondary test statistic.

How can we compare accurately the value of the test statistic T for each table in Γ with the observed test statistic T_{obs}? The tests 'if $T = T_{obs}$, then' or 'if $T \geq T_{obs}$, then' are problematic when T and T_{obs} are floating point numbers. In practice, a tolerance is set which declares two floating point numbers to be equal if the difference is less than the tolerance. Unfortunately, the results of a floating point test may depend on the compiler, the machine type, the size of the floating point mantissa, the order of computations and the CPU rounding process (Verbeek and Kroonenberg, 1985). Some of these problems may be avoided by using floating point arithmetic conforming to IEEE standards (Thisted, 1988, Section 2.4).

The most accurate form of a test statistic should be used for computation. We may write the X^2 statistic and the likelihood–ratio LR statistic (rather than $L^2 = 2 \log LR$) as

$$X^2 = \sum_{ij} \left(\frac{y_{ij}^2}{e_{ij}}\right) - y_{++} \tag{3}$$

$$LR = \prod_{ij} \left(\frac{y_{ij}}{e_{ij}}\right)^{y_{ij}} \tag{4}$$

where e_{ij} denotes the expected value in the ij cell. For independence, expected values and multivariate hypergeometric probabilities are rational numbers so that comparisons of the three commonly used test statistics may be made using rational arithmetic. Many versions of Lisp support BigNum (Serpette *et al.*, 1989), a C library of routines for arbitrary precision arithmetic, which supports rational number types. The calculation and comparison of rational test statistics can made without error by using BigNum. Hence, for an exact test of independence, the problematic comparison of floating point numbers may be avoided entirely.

For the models of QI, QS, QLLA and LLA, the expected values may be irrational. However, rational fitted values may be obtained for any log-linear model by implementing the iterative proportional fitting (IPF) algorithm using rational arithmetic as follows. IPF maximizes the likelihood by searching along a series of fixed directions defined by the column vectors of the model matrix (Fienberg and Meyer, 1983). If we are maximizing the likelihood along direction $\mathbf{x}_j$, defined by the jth column of the model matrix, then the new updated vector of fitted values, $\mathbf{e}_{new}$, is obtained by proportional adjustment of the old vector of fitted values, $\mathbf{e}_{old}$, by

$$\mathbf{e}_{new} \;=\; \mathbf{e}_{old} \times \theta_j \, \mathbf{x}_j, \tag{5}$$

where the multiplication is coordinatewise. When $\mathbf{x}_j$ is a vector of zeros and ones, θ_j is the ratio of the lengths of $\mathbf{y}$ and $\mathbf{e}_{old}$ projected onto the $\mathbf{x}_j$ direction, i.e.,

$$\theta_j \;=\; (\frac{\mathbf{y}'\mathbf{x}_j}{\mathbf{e}'_{old}\mathbf{x}_j}). \tag{6}$$

When $\mathbf{x}_j$ is arbitrary, there is no direct estimate of θ_j and we are left with a one-dimensional maximization problem.

For QI and QS, the column vectors of the model matrix are vectors of zeros and ones. For these models, start with initial estimates of 1 for off-diagonal cells and 0 for the diagonal cells. Since each iterative step of IPF only involves successive proportional adjustment of fitted values at the previous step by a ratio of rationals to yield updated fitted values, the successive updated fitted values may be calculated exactly using rational arithmetic. A stopping rule may be used that ensures any desired degree of accuracy in the rational expected values. Comparison of the usual test statistics is based on error-free rational arithmetic using BigNum library routines.

For QLLA and LLA, the column vector corresponding to γ is not a vector of zeros and ones, so an one-dimensional maximization problem needs solving. By restricting the proportional adjustment scale factor θ to be rational, once again the successive updated fitted values may be calculated exactly using rational arithmetic. A stopping rule may be used that ensures any desired degree of accuracy in the rational fitted values. Comparison of the usual test statistics is based on error-free rational arithmetic using BigNum library routines. Note that alternative IPF methods, not involving searching, exist for fitting the UA model (Lawal and Upton, 1995), which can be implemented using only rational arithmetic.

3.3 Calculating the p-value

For a saturated log-linear model and null hypothesis that all the interest parameters equal zero, Forster *et al.* (1996) derived, up to a constant of proportionality, the conditional distribution of the sufficient statistics for the interest parameters, given the sufficient statistics for the nuisance parameters. For testing the goodness of fit of a log-linear model, all the model parameters are nuisance parameters and the conditional distribution for testing goodness of fit may be written explicitly in terms of the cell counts, up to a constant of proportionality, as

$$f(\boldsymbol{y}|X^T\boldsymbol{Y} = X^T\boldsymbol{y}_{obs}) \propto \frac{1}{\prod_{ij} y_{ij}!}, \tag{7}$$

where $\boldsymbol{y}$ denotes the random vector of counts, X^T is the transpose of the model matrix X, $\boldsymbol{y}_{obs}$ denotes the observed vector of counts, $X^T\boldsymbol{y}_{obs}$ are the observed sufficient statistics for the nuisance parameters and the right hand side is subject to the conditioning constraints. For independence, as is well known, this distribution is the multivariate hypergeometric. Unfortunately, in general, these null conditional probabilities cannot be represented in closed form. Except possibly for the case of independence, the normalization constant cannot be represented in closed form, but must be left as a summation over all tables whose sufficient statistics for the nuisance parameters are equal to their observed values. Note that for QI and QS, this distribution only involves the distribution of counts in the off-diagonal cells, so that the product in (**??**) may be taken as over the off-diagonal cells, see also Smith and McDonald (1995).

How to calculate accurately and sum the probabilities of tables at least as extreme as the observed table is problematic. The probabilities in (**??**) may be extremely small and usually the normalization constant must be calculated by summing (**??**) over all the tables in Γ. Most algorithms use log-factorials or the log-gamma function in order to prevent overflow or underflow in the calculations (see Verbeek and Kroonenberg (1985) for details). The most accurate approach uses symbolic calculation based on number theory. Any positive integer, say $n!$, can be written as a product of prime factors, so that $n! = p_1^{r_1} p_2^{r_2} p_3^{r_3} \cdots p_k^{r_k}$ for some positive integers k and $r_1, r_2, r_3, \ldots, r_k$, where the p_i's are prime numbers. Each factorial term can be represented by its prime factorization. Two terms in the summation of the normalization constant are added together only after factorial terms have been factorized into primes and the calculation simplified symbolically. This reduces the computational complexity to a minimum and achieves maximum accuracy, see also Wu (1993). The result is a rational number. A C routine was written to symbolically reduce each required calculation to its simplest irreducible rational form. Then the addition or division of rational terms was implemented using BIGNUM library routines yielding rational results without loss of any accuracy. While symbolic computation maintains accuracy, it incurs significant time penalties (about a factor of 3 or 4).

4 Couple's rating of sexual fun

Hout *et al.* (1987) studied the association of husbands' and wives' reports of sexual fun. Husbands and wives answered the question 'Sex is fun for me and my partner (a) never, (b) occasionally, (c) fairly often, (d) very often, (e) almost always'. The rare (two wives and one husband) 'never' responses were combined with the 'occasionally' responses because the data was sparse. After deleting a few cases for various reasons, they analysed the resulting 4×4 table based on 91 responses. These data are presented in Table **??** with asymptotic and exact p-values for the models of I, UA, QI, QUA and QS reported in Table **??**. For the model of I, the asymptotic p-value of 0.078 differs somewhat from the exact p-value of 0.114, calculated by enumerating the 947 766 430 tables in Γ_I. For the model of QI, the asymptotic p-value of 0.402 differs substantially from the exact p-value of 0.502, calculated by enumerating the 15 708 tables in Γ_{QI}. For the model of QS, the asymptotic p-value of 0.947 differs somewhat from the exact p-value of 1.000, calculated by enumerating the 161 tables in Γ_{QS}. While the models of QI and QS fit the data, they do not take into account that the categories are ordered.

We now consider the models of UA and QUA (scored never or occasionally $= 1, \ldots,$ almost always $= 4$). For the model of UA, the asymptotic p-value of 0.757 differs somewhat from the exact p-value of 0.795, calculated by enumerating the 8 137 492 tables in Γ_{UA}. For the model of QUA, the asymptotic p-value of 0.979 is close to the exact p-value of 1.000, calculated by enumerating the 251 tables in Γ_{QUA}.

The tests just discussed are tests of goodness of fit, i.e., they compare a given model to the saturated model. Exact tests against non-saturated alternatives are straightforward. An exact test of model I against UA (or QI against QUA) is based on the sufficient statistic for γ, i.e., the covariance between the scores. This linear-by-linear association test orders the tables in Γ_{UA} (and Γ_{QUA}) by the covariance between the scores. The exact test of model I against UA yields an exact p-value of 0.00079, calculated by enumerating the 947 766 430 tables in Γ_I. The exact test of model QI against QUA yields an exact p-value of 0.02113, calculated by enumerating the 15 708 tables in Γ_{QI}. While the models of I and QI fit the data, most analysts would reject these models in favour of either the model of UA or QUA on the basis of these exact tests against non-saturated alternatives. As the models of UA and QUA both fit the data well, most analysts would choose the model of uniform association as it is the most parsimonious.

5 Variability in classification of cancer by two pathologists

Agresti (1988) used the log-linear models described in Section 2 to study the agreement between two pathologists evaluating possible cervical cancer, using data given in Holmquist *et al.* (1967). These data are presented in Table **??**. The pathologists classified 118 specimens using the ordered categories: 1) negative, 2) atypical squamous hyperplasia, 3) carcinoma in situ, 4) squamous carcinoma with early stromal invasion and 5) invasive carcinoma.

Agresti noted that Table **??** was sparse with 12 off-diagonal zeros and that the distribution of L^2 is not well approximated by a χ^2 distribution. L^2, degrees of freedom, asymptotic and exact p-values for the test of goodness of fit of various models fitted to the data in Table **??** are reported in Table **??**. The I model is implausible and is not considered further. The asymptotic p-value of 0.009 for the D model differs slightly from the exact p-value of 0.001, calculated by enumerating the 845 489 tables in Γ_D. The asymptotic p-value of 0.3679 for the UA model (scored negative $= 1, \ldots,$ invasive carcinoma $= 5$) differs substantially from the exact p-value of 0.036, calculated by enumerating the 16 623 tables in Γ_{UA}. The asymptotic p-value of 0.8668 for the D + UA model differs substantially from the exact p-value of 0.459, calculated by enumerating the 3 350 tables in Γ_{D+UA}. The asymptotic p-value of 0.259 for the QI model differs substantially from the exact p-value of 0.023, calculated by enumerating the 435 tables in Γ_{QI}. The asymptotic p-value of 1.000 for the QUA model equals, to three decimal points, the exact p-value of 1.000, calculated by enumerating the 3 tables in Γ_{QUA}. The asymptotic p-value of 0.986 for the QS model is very close to the exact p-value of 1.000, calculated by enumerating the 3 tables in Γ_{QS}.

One advantage of exact tests is that the calculation of degrees of freedom is unnecessary, which for complicated log-linear models fitted to sparse tables is often difficult. For the model of QI, the diagonal counts are fitted exactly so that the last two columns of Table **??** do not contribute to L^2. We claim that degrees of freedom for the model of QI should be based on the 5×3 table obtained by deleting these last two columns. Hence, a χ^2 distribution with 5 df, rather than 11 df, should be used for the calculation of asymptotic p-values. Based on 5 df, QI has an asymptotic p-value of 0.019 which is much closer to the exact p-value of 0.023. See Smith and McDonald (1995) for further discussion. Similarly, for QUA and QS, degrees of freedom should be based on the 3×3 table obtained by deleting the last two columns and the last two rows of Table **??**. This conclusion is supported by the enumeration of Γ_{QUA} and Γ_{QS}. These reference sets contain the same three tables. These tables have identical counts in the last two columns and the last two rows, namely, those of the observed table. Note that the models of QUA and QS based on the 3×3 table obtained by deleting the last two columns and the last two rows of Table **??** have 1 df. Hence, a χ^2 distribution with 1 df, rather than 10 and 6 df, should be used for the calculation of asymptotic p-values for QUA and QS respectively. Based on 1 df, QUA has an asymptotic p-value of 0.263 which differs substantially from the exact p-value of 1.000. Based on 1 df, QS has an asymptotic p-value of 0.323 which differs substantially from the exact p-value of 1.000.

On the basis of exact goodness-of-fit tests we conclude, as does Agresti, that the diagonal parameter plus uniform-association model fits the data well. Therefore, there is agreement in excess of that occurring simply by chance (what would occur under independence of ratings) plus extra agreement due to a positive association between the ratings. For further discussion and interpretation in terms of parameter estimates and local odds-ratios see Agresti (1988).

6 Discussion

Exact tests can be used when the computations are feasible. Feasibility depends on the size of Γ and how fast we can enumerate Γ. Estimates of the size of Γ_I, $|\Gamma_I|$, have been reviewed by Agresti *et al.* (1979). Unfortunately, estimates of the size of Γ for more complicated models are unavailable and providing estimates an area for further research.

One disadvantage of exact tests is that the discrete reference distribution may have small support or even be degenerate. Calculating the size of the reference set serves as a useful diagnostic tool. When the reference distribution has a small number of support points one should be cautious. Asymptotic methods can be unreliable when one is approximating a discrete distribution with a small number of support points by a continuous approximating distribution. In this case, exact methods can be conservative because of the high degree of discreteness of the test statistic and one should consider using a mid p-value or a modified p-value, which reduces conservativeness, but maintains exactness (Kim and Agresti, 1995).

Enumerating the reference set may be infeasible or so time consuming so as to inhibit data analysis, e.g., the calculations may take days. In this case, Monte Carlo methods may be used to provide point and interval estimates of the exact p-value by simulating from the required reference distribution. For each simulated table, the test statistic is calculated and the exact p-value is estimated by the proportion of simulated tables which are at least as discrepant from the null as the observed table. For independence, this has been described in detail by Agresti *et al.* (1979). For log-linear models, Forster *et al.* (1996) provide Markov chain Monte Carlo methods which may be used to estimate the exact p-value and its precision. Smith *et al.* (1996a, 1996b) use these methods for the models of I, QI, QS and QUA.

One advantage of estimating, rather than calculating, the exact p-value is that the required computational effort is much less dependent on the sample size, table size and distributions of the sufficient statistics for the nuisance parameters. One disadvantage is that the estimate depends on the starting seed used for random sampling. Senchaudhuri *et al.* (1995) estimate the required sample sizes so that the estimated p-value cannot differ in the first three decimal places, regardless of the starting seed, 99% of the time. These sample sizes vary monotonically from 2 651 244 for a p-value of 0.001 to 424 623 720 for a p-value of 0.200. Hence, much computational effort is needed to estimate exact p-values with very small Monte Carlo error.

One may want to calculate, rather than estimate, the true exact p-value for many reasons. The true value serves as the 'gold standard' for validating the Monte Carlo methods used to estimate the exact p-value. Monte Carlo error in the estimated p-value may be unacceptable, e.g., to regulatory agencies dealing with the licensing of drugs. The computational cost of calculating the exact p-value may be negligible when the calculation can run in the background and not interfere with other computer tasks. A Monte Carlo estimate may suffice for data analysis and the first submission of a paper to a journal. Many months may pass from paper submission until the final proof stage when exact p-values may be substituted for estimated p-values. While estimated p-values are usually sufficient for data analytic purposes, for many reasons, the goal should always be to calculate the true exact p-values.

This paper proposes recursive algorithms for exact tests of quasi-independence, quasi-uniform association and quasi-symmetry using symbolic computation and rational arithmetic in order to calculate the exact p-values accurately. One area for future research is whether the network algorithm can be modified to carry out exact tests of these models and whether it can be adapted to use rational arithmetic.

Appendix

Here we discuss the recursive enumeration algorithm in more detail as well as some of the implementation issues using Lisp and C.

The recursive step breaks down the table by removing the first row. At each level of recursion, i.e., each row in the table, all the possible rows given the constraints have to be generated. Generating a table is a simple extension of being able to generate rows given a set of constraints. The top row constraints are extracted from the general constraints for the table being generated. For each of the rows that is generated, the column totals are adjusted, and a recursive call is made passing in the new constraints for the subtable. The recursion terminates if the table to be generated consists of just one row. In this case the column margins are the elements of the final row. When generating under QI this may not be the case, since there may be cells in the final row that are fixed. Hence, if there are fixed cells in the final row, then the respective column margin must be zero, otherwise an invalid table has been generated.

Row generation can be easily achieved in a number of ways. Firstly, the constraints on the cell counts in the row have to be considered. The major constraint is that the row has to add up to a certain total, given by the row margin for this row. Additionally, each count in the row must be less than or equal to its column margin.

The algorithm must be able to generate rows of arbitrary length. One method would be to perform the operation recursively, and this approach was used in our Lisp implementation. This approach may return a long list of all the possible rows and becomes unworkable because of potential high storage requirements. In Lisp we solved this problem by working with delayed evaluation of 'streams' (not to be confused with I/O streams). In Lisp, such streams consist of head and tail elements of the list (Abelson and Sussman, 1985). The head stores the first element, whilst the tail is a function that will return the rest of the stream. To follow the tail of the stream, the function is evaluated, returning a new head and tail pair. Whilst this representation requires a minor performance penalty, it does allow very large lists.

In C, we used a slightly different method, which consisted of firstly setting all the row elements to zero, and starting out with the row total. Then starting at the first element, the maximum amount that the first element can hold is placed in it, and subtracted from the running row total. This is repeated for all the elements along the row, even if the running total is zero. Once the end of the row has been reached, the result should be a valid row, if the running total is zero. If not, then we have already generated all the rows possible. Once we have generated a row, another row is generated by resetting the running total to the value of the last element, then reversing along the row, looking for a nonzero element. If a nonzero element cannot be found, then we have again generated all the rows and can terminate. Once a nonzero element has been found, the value is decremented, the running total incremented and the procedure repeated starting at the next element. For row generation under QI, the only modification to the algorithm is that it should skip any fixed cells.

The major difficulty with this algorithm is expressing it in such a way that tables are generated one at a time. Using Lisp streams solved this problem cleanly, since the state of the computation is effectively stored in the tail of the stream. In C, the solution was to call a function whenever a table had been generated. This inside out approach does lead to some minor problems with the scope of variables, and breaking out of generating tables. For example, when generating under QS we discard tables that do not satisfy the constraints. With the C version we had to rewrite the function that is called from the final step of the recursion.

Pseudo-code for the heart of the recursive routine follows:

```
define recursive(row-margins, column-margins)
 if final-row(row-margins)
     maketable(column-margins);
 else
    foreach x in permute(head(row-margins), column-margins) do
      appendfront(x, recursive(tail(row-margins),
                              sublist(column-margins, x)));
```

The `permute` function takes a total and a list of constraints, and generates a list of rows that have the given total and each element is less than or equal to the respective constraint.

Acknowledgements

McDonald and Michaelides were supported by the Economic and Social Research Council's Analysis of Large and Complex Datasets Programme (award H519255005 and Ph.D. studentship respectively). This work was also supported by the IBM Shared University Research programme which provided an IBM SP2 multiprocessor parallel computer for computations.

References

Abelson, H. and Sussman, G. (1985) *Structure and Interpretation of Computer Programs*, MIT Press, Cambridge, MA.

Agresti, A. (1988) A model for agreement between ratings on an ordinal scale. *Biometrics*, **44**, 539–548.

Agresti, A. (1990) *Categorical Data Analysis*, Wiley, New York.

Agresti, A. (1992) A survey of exact inference for contingency tables (with discussion). *Statistical Science*, **7**, 131–177.

Agresti, A., Wackerly, D. and Boyett, J. M. (1979) Exact conditional tests for cross-classifications: approximation of attained significance levels. *Psychometrika*, **44**, 75–83.

Bishop, Y. M. M., Fienberg, S. E. and Holland, P. W. (1975) *Discrete Multivariate Analysis: Theory and Practice*, MIT Press, Cambridge, MA.

Boulton, D. M. and Wallace, C.S. (1973) Occupancy of a rectangular array. *Computer Journal*, **16**, 57–63.

De Roure, D. C. and Michaelides, D. (1994) A distributed LISP–STAT environment, in *Proceedings of COMPSTAT 1994*, R. Dutter and W. Grossman (eds), Physica–Verlag, Heidelberg, pp. 371–376.

Forster, J. J., McDonald, J. W. and Smith, P. W. F. (1996) Monte Carlo exact conditional tests for log-linear and logistic models. *Journal of the Royal Statistical Society, Series B*, **58**, 445–453.

Freeman, G. H. and Halton, J. H. (1951) Note on an exact treatment of contingency, goodness of fit and other problems of significance. *Biometrika*, **38**, 141–149.

Holmquist, N. S., McMahon, C. A. and Williams, O. D. (1967) Variability in classification of carcinoma in situ of the uterine cervix. *Archives of Pathology*, **84**, 334–345.

Hout, M., Duncan, O. D. and Sobel, M. E. (1987) Association and heterogeneity: structural models of similarities and differences, in *Sociological Methodology 1987*, C. C. Clogg (ed), American Sociological Association, Washington DC, pp. 145–184.

Kim, D. and Agresti, A. (1995) Improved exact inference about conditional association in three-way contingency tables. *Journal of the American Statistical Association*, **90**, 632–639.

Fienberg, S. E. and Meyer, M. M. (1983) Iterative proportional fitting, in *Encyclopedia of Statistical Sciences*, Vol. 4, S. Kotz (ed), Wiley, New York, pp. 275–279.

Lawal, H. B. and Upton, G. J. G. (1995) An algorithm for fitting models to $N \times N$ contingency tables having ordered categories. *Communications in Statistics - Simulation*, **24**, 793–805.

McDonald, J. W. and Smith, P. W. F. (1995) Exact conditional tests of quasi-independence for triangular contingency tables: estimating attained significance levels. *Applied Statistics*, **44**, 143–151.

Mehta, C. R. (1994) The exact analysis of contingency tables in medical research. *Statistical Methods in Medical Research*, **3**, 135–156.

Michaelides, D. T. (1997) Exact Tests via Complete Enumeration: A Distributed Computing Approach. Ph. D. thesis under submission. Department of Social Statistics, University of Southampton.

Senchaudhuri, P., Mehta, C. R. and Patel N. R. (1995) Estimating exact p-values by the method of control variates, or Monte Carlo Rescue. *Journal of the American Statistical Association*, **90**, 640–648.

Serpette, B., Vuillemin, J. and Hervé, J. (1989) BigNum: a portable and efficient package for arbitrary-precision arithmetic. Research Report 2, Digital Equipment Corporation Paris Research Laboratory. Available at http://pam.devinci.fr/documentation.html.

Smith, P. W. F. and McDonald, J. W. (1995) Exact conditional tests for incomplete contingency tables: estimating attained significance levels. *Statistics and Computing*, **5**, 253–256.

Smith, P. W. F., Forster, J. J., and McDonald, J. W. (1996a) Monte Carlo exact tests for square contingency tables. *Journal of the Royal Statistical Society, Series A*, **159**, 309–321.

Smith, P. W. F., McDonald, J. W., Forster, J. J., and Berrington, A. M. (1996b) Monte Carlo exact methods used for analysing interethnic unions in Great Britain. *Applied Statistics*, **45**, 191–202.

Thisted, R. A. (1988) *Elements of Statistical Computing*, Chapman and Hall, New York.

Verbeek, A. and Kroonenberg, P. M. (1985) A survey of algorithms for exact distributions of test statistics in $r \times c$ contingency tables with fixed margins. *Computational Statistics & Data Analysis*, **3**, 159–185.

Wu, T. (1993) An accurate computation of the hypergeometric distribution function. *ACM Transactions on Mathematical Software*, **19**, 33–43.

Table 1: Rating of Sexual Fun: Husband's Response by Wife's Response

	Wife's Response			
Husband's Response	Never or Occasionally	Fairly Often	Very Often	Almost Always
Never or occasionally	7	7	2	3
Fairly often	2	8	3	7
Very often	1	5	4	9
Almost always	2	8	9	14

Table 2: Likelihood ratio goodness-of-fit test statistics and p-values for models for Table 1

Model	L^2	Degrees of freedom	Asymptotic p-value	Exact p-value
I	15.49	9	0.078	0.114
UA	5.00	8	0.757	0.795
QI	5.12	5	0.402	0.502
QUA	0.44	4	0.979	1.000
QS	0.37	3	0.947	1.000

Table 3: Cross-classification of pathologist ratings

	Pathologist B				
Pathologist A	1	2	3	4	5
1	22	2	2	0	0
2	5	7	14	0	0
3	0	2	36	0	0
4	0	1	14	7	0
5	0	0	3	0	3

Table 4: Likelihood ratio goodness-of-fit test statistics and p-values for models for Table 3

Model	L^2	Degrees of freedom	Asymptotic p-value	Exact p-value
D	30.90	15	0.009	0.001
UA	16.21	15	0.368	0.036
D+UA	8.41	14	0.867	0.459
QI	13.56	11	0.259	0.023
QUA	1.25	10	1.000	1.000
QS	0.98	6	0.986	1.000

296/96R: Exact tests for two-way symmetric contingency tables

McDONALD, De ROURE and**MICHAELIDES**

Changes made in response to comments of referee B

Referee B first paragraph: Referee B caught a minor error. We did not distinguish between the models of linear-by-linear association (LLA) FOR THE WHOLE TABLE and quasi linear-by-linear association (QLLA) FOR THE OFF-DIAGONAL CELLS. While LLA is not nested within QS, QLLA is nested within QS. We now use both terms in the paper.

The following text addressing these points is new.

The model of linear-by-linear association, LLA, corresponds to $\gamma_{ij} = \gamma u_i u_j$ for all i, j for known scores u_i. The model of uniform association, UA, corresponds to $\gamma_{ij} = \gamma u_i u_j$ for all i, j for known equal-interval scores u_i. For the model of LLA (and special case of UA), γ is an additional nuisance parameter with sufficient statistic $\sum_i \sum_j u_i u_j y_{ij}$. The models of LLA and UA may be considered for the cells off the main diagonal and are termed the quasi-linear-by-linear association (QLLA) and quasi-uniform association (QUA) models. Note than QLLA and QUA are special cases of the QS model, i.e., they are parsimonious QS models which have QI as the special case of $\gamma = 0$.

An exact goodness-of-fit test for LLA uses the conditional distribution of the cell counts given the margins and $\sum_i \sum_j u_i u_j y_{ij}$, i.e., the 'covariance' between the scores. An exact goodness-of-fit test for QLLA uses the conditional distribution of the cell counts given the margins, the diagonal counts and $\sum_i \sum_j u_i u_j y_{ij}$.

Referee B second paragraph: We did use the correct set of tables and algorithms. The referee preferred the term quasi-uniform association and notation QUA to our notation QI + UA. We agree and have made the necessary changes.

Referee B last paragraph: As the categories in Table 1 are ordered, we should have considered the model of LLA and its goodness of fit. We have done so with the UA and QUA models. The referee then carried out a Monte Carlo test of linear-by-linear association using StatXact. This is not a test of goodness of fit but can be viewed as a test of the model of I against a nonsaturated alternative. i.e., the model of UA. Using our algorithms we carried out two exact tests against nonsaturated alternatives, i.e., I against UA and QI against QUA. Note that our exact test of I against UA yielded an exact p-value of 0.0007949 versus the estimated p-value from StatXact of 0.0015 $\pm$ .0003 with 100,000 Monte Carlo samples. While our calculated p-value is outside the Monte Carlo interval estimate, we attribute the difference to Monte Carlo error.

The following text addressing these points is new.

While the models of QI and QS fit the data, they do not take into account that the categories are ordered.

We now consider the models of UA and QUA (scored never or occasionally $= 1, \ldots,$ almost always $= 4$). For the model of UA, the asymptotic p-value of 0.757 differs somewhat from the exact p-value of 0.795, calculated by enumerating the 8 137 492 tables in Γ_{UA}. For the model of QUA, the asymptotic p-value of 0.979 is close to the exact p-value of 1.000, calculated by enumerating the 251 tables in Γ_{QUA}.

The tests just discussed are tests of goodness of fit, i.e., they compare a given model to the saturated model. Exact tests against non-saturated alternatives are straightforward. An exact test of model I against UA (or QI against QUA) is based on the sufficient statistic for γ, i.e., the covariance between the scores. This linear-by-linear association test orders the tables in Γ_{UA} (and Γ_{QUA}) by the covariance between the scores. The exact test of model I against UA yields an exact p-value of 0.00079, calculated by enumerating the 947 766 430 tables in Γ_I. The exact test of model QI against QUA yields an exact p-value of 0.02113, calculated by enumerating the 15 708 tables in Γ_{QI}. While the models of I and QI fit the data, most analysts would reject these models in favour of either the model of UA or QUA on the basis of these exact tests against non-saturated alternatives. As the models of UA and QUA both fit the data well, most analysts would choose the model of uniform association as it is the most parsimonious.