Exact Tests of Significance
in Higher Dimensional Tables

Daniel Zelterman\textsuperscript{1,2}
Ivan Siu-Fung Chan\textsuperscript{1}
Paul W. Mielke\textsuperscript{3}

February 7, 1995

\textsuperscript{1}Division of Biostatistics
University of Minnesota
Minneapolis, MN 55455

\textsuperscript{2}Division of Biostatistics
Yale University
New Haven, CT 06520

\textsuperscript{3}Department of Statistics
Colorado State University
Fort Collins, CO 80523

This research was supported by grant numbers U01-AG10328, R01-AI34110, N01-AI05073, and R01-HS07772 awarded by the U.S. National Institutes of Health, and a grant from the Minnesota Supercomputer Institute.
Abstract

We describe methods used to provide an exact test of significance of the hypothesis that all factors are mutually independent of each other in $2^3$ and $2^4$ contingency tables. Several numerical examples demonstrate the advantages of exact tests over approximate significance levels. We give bounds on the number of tables needed to perform this exact significance test. In 4 or more dimensions, the number of tables in this enumeration becomes astronomical with even modest sample sizes. Inverting the characteristic function of the exact distribution has proved useful in these situations.
1 Introduction

The Fisher exact test is well known in $2 \times 2$ tables. The central idea is to enumerate all possible outcomes consistent with a given set of marginal totals and add up the probabilities of those tables more extreme than the one observed. Conditional on the margins, a $2 \times 2$ table is a one-dimensional random variable so there are few conceptual difficulties. In larger tables, there are problems in the complete enumeration of all tables given a set of specified marginal totals. In tables larger than $2 \times 2$, three questions come to mind:

1. What is the probability of any one table in the enumeration?

2. How do we define the "tail" of the distribution?

and

3. How do we efficiently enumerate all possible tables?

The first question is the easiest to answer and the probability of $2^3$ and $2^4$ tables are given in the following section. The model or hypothesis we are testing is that of mutual independence of all factors. Other hypotheses involving interactions could be tested using similar methods given by Agresti (1992, Section 4.4).

The definition of the "tail" depends on the criteria being used. For an exact test based on the likelihood, any table with a smaller probability than the observed table is in the tail of that table. On the other hand, a table with a larger value of Pearson's chi-squared than the observed table is in the
tail. These two definitions do not necessarily coincide (see also Fisher 1950; and Radlow and Alf, 1975). We need not restrict our attention to Pearson's chi-squared either: any of the continuum of statistics of Cressie and Read (1984, 1989) might be used as well as the statistic suggested by Zelterman (1987) for use in sparse data situations.

The problem of enumeration is more difficult to answer and has received considerable attention in recent years. For $I \times J$ tables, Mehta and Patel (1983) developed algorithms and programmed them in the commercially available software package StatXact (1991). Agresti (1992) is a recent review for two and three dimensional exact methods. The popularity of $2 \times 2$ tables in epidemiologic studies has produced a literature on exact methods for estimating and testing the common odds ratio. For examples, see: Thomas (1975); Pagano and Tritchler (1983); Mehta, Patel, and Gray (1985); Morgan and Blumenstein (1991); and Vollset, Hirji, and Elashoff (1991). Some of the important issues in exact inference for high dimensional tables are described by Kreiner (1992). Dwass (1978) showed that likelihood methods generally produce most powerful tests in higher dimensional tables. Kreiner (1987) worked out several exact tests for a $2^7$ table. Stumpf and Steyn (1986) gave the likelihood function and exact moments of random variables in a 3-way table. These were generalized to $r$-way tables by Mielke and Berry (1988). For high dimensional tables, the characteristic function inversion methods of Baglivo, Olivier, and Pagano (1992) could be used.

To motivate exact methods consider the two $2^3$ artificial examples given in Tables 1 and 2. In both cases we want to test the hypothesis of mutual
independence of all three factors. In these tables, the counts are small enough to suggest that an exact enumeration is not only possible, but perhaps the preferred analysis.

In Table 1 Pearson’s $\chi^2 = 7.34$ with 4 d.f. and an asymptotic significance level of $p = .119$. An exact analysis enumerates 53 tables that are consistent with these 3 one-way marginal totals and shows the exact significance of the chi-squared and the probability of observing a table this or less likely is .0424. It is no surprise that the exact distribution of the chi-squared statistic is very different from its asymptotic approximation with such a table of small counts.

For another example, consider Table 2. In this table, $\chi^2 = 6.519$ (4 d.f.) with an asymptotic significance level of $p = .164$. The exact enumeration of 259 tables shows that the exact significance level of $\chi^2$ is .178, or in fairly close agreement. On the other hand, the exact probability of observing this table or any table less likely is .0485. The lesson we learn from these two examples is that the behavior of test statistics, such as the Pearson $\chi^2$, may or may not agree with their asymptotic approximations. The only certain method for accurate analysis of tables with small counts is to perform exact methods based on the likelihood function.

This article describes the problem of testing mutual independence in 3 and 4 dimensions. Other hypotheses, such as testing specific interactions in log-linear models could be examined, as well as the analysis of tables in more than 4 dimensions. We could have also described tables with more than two categories in each dimension. In the following section we describe the
Algorithm in detail and sketch out the $2^4$ Algorithm. We give an upper bound on the number of $2^3$ tables consistent with a set of marginal totals in Section 3.

2 $2^3$ and $2^4$ Tables

For a three-way table, let $N$ denote the overall sample size and let $x_{ijk}$ ($i, j, k = 1, 2$) denote the non-negative integer counts in each of the eight ($= 2^3$) cells. A plus sign in the subscript denotes summation over that index so, for example, $N = x_{+++}$. Denote the one-way marginal totals by

$$n_1 = x_{1++} = \sum_j \sum_k x_{1jk}, \quad n_2 = x_{+1+}, \quad \text{and} \quad n_3 = x_{++1}.$$  

The full set of all one-way marginal totals is then $(n_s, N - n_s)$ for $s = 1, 2, 3$.

The exact likelihood of $\{x_{ijk}\}$ given the marginal totals is

$$P[x_{ijk}|n_s, N] = \frac{\prod_{s=1}^3 n_s!(N - n_s)!}{(N!)^2 \prod_{i,j,k} x_{ijk}!}$$  

(Stumpf and Steyn, 1986; Mielke and Berry, 1988).

The exact probability of observing Table 2, given the one-way marginal totals, is .00308. The $2^3$ Algorithm given in the following section shows that there are 259 tables whose marginal totals agree with those of Table 2. Of these, 188 have likelihoods smaller than that of Table 2 and the sum of these probabilities is .0485. Similarly, 206 tables in the complete enumeration have values of chi-squared greater than or equal to the observed value of 6.519. The probability of observing a value of chi-squared this large or greater is .178, without reference to the chi-square asymptotic distribution. Compared to
the 4 d.f. chi-squared asymptotic distribution, the approximate significance level is .164. As we mentioned in the introduction, there is no reason why the exact chi-squared tail area should agree with that of the exact likelihood. Tables that are more extreme than the observed table by one criteria are not necessarily more extreme by a different criteria.

There is no reason to restrict the analysis to tables with small counts. Table 3 gives another example of a $2^3$ table but the counts are quite large. There are a total of $N = 1663$ observations in this table. Respondents were categorized by their attitudes towards equal job opportunity for minorities, and which of two regions of the country they lived in. The third binary valued factor was the year of the survey. We will not describe the analysis in more detail except to say that $\chi^2 = 284.3$ (4 d.f.) so the test of mutual independence is highly significant. The exact probability of observing the data in Table 3 conditional on the one-way margins is $1.86 \times 10^{-73}$ under the hypothesis of mutual independence.

There are a total of $3.68 \times 10^9$ tables that are consistent with the one-way margins of Table 3. Of these tables, $2.76 \times 10^9$ have probabilities smaller than that of Table 3 and the sum of these probabilities is $1.68 \times 10^{-66}$. The point of this third example is not to test the hypothesis so much as to show what can be done with computing resources. The complete enumeration of these tables took approximately 3 hours on a mainframe computer, and several days on a personal computer.

The example in Table 4 motivates exact tests for $2^4$ tables. This data summarizes an experiment on the sensitivity of nerve fiber endings in rats'
tongues to four different taste stimuli (Bishop, Fienberg, and Holland, 1975, pp. 220–1). Dead fibers are indistinguishable from those that do not respond to any of the four stimuli and this cell was treated as a structural zero (i.e., no observation is possible here) but we will treat this data as a complete $2^4$ table.

The likelihood for four-dimensional tables $\{x_{ijkl}\}$ given the four one-way marginal totals $\{n_s: s = 1, \ldots, 4\}$ and sample size $N$ is

$$P[x_{ijkl}|n_s, N] = \frac{\prod_{s=1}^{4} n_s!(N-n_s)!}{(N!)^3 \prod_{ijkl} x_{ijkl}!}$$

(Mielke and Berry, 1988). The notation is by analogy to the three-dimensional tables, namely, $n_1 = x_{1+++} = \sum_j \sum_k \sum_l x_{ijkl}$, $n_2 = x_{++1+}$, $n_3 = x_{++11}$, $n_4 = x_{++++}$, and $N = x_{++++}$.

The test of the null hypothesis that the rat taste buds act independently on the four different stimuli gives the chi-squared value $\chi^2 = 12.780$ (11 d.f.) and an approximate significance level of $p = .3079$ using the asymptotic chi-squared reference distribution.

The $2^4$ enumeration shows that there are 3,299,027 four dimensional tables whose one-way marginal totals agree with those of Table 4. Of these tables, 3,055,894 have values of chi-squared greater than or equal to the observed value of 12.780 in Table 4. The exact probability of observing a value of chi-squared at least as great as the observed value is .3163 using the $2^4$ likelihood function. This probability is fairly close to the approximate significance level of .3079 using the asymptotic chi-squared distribution.

The exact probability of observing Table 4 conditional on all four of the one-way margins is $9.043 \times 10^{-7}$. There are 3,061,883 tables with likelihoods
smaller than or equal to this value and the sum of the probabilities of those tables is .3047. In summary, the exact likelihood test, exact and approximate chi-squared tests are all in close agreement about testing the null hypothesis of independence of all four factors in Table 4.

3 Algorithms and Bounds for $2^3$ and $2^4$ Tables

In this section we describe an algorithm for generating all possible $2^3$ contingency tables that are consistent with a given set of one-way marginal totals. This enumeration is needed to perform an exact test of mutual independence of all 3 factors. In general, for $k$ dimensions, the hypothesis of mutual independence of all $k$ factors has $2^k - k - 1$ degrees of freedom. The generation of all $2^3$ tables requires $4 (= 2^3 - 3 - 1)$ nested loops each specifying one of the cells in the table. The remaining four cells are determined by the values in these cells. These methods represent the practical limits using enumeration. Higher dimensional tables could be analyzed using Fourier inversion such as proposed by Baglivo, et al. (1992).

When $N$ is large, efficiency of the algorithm becomes an important issue. The algorithm given here is efficient in the sense that every iteration of the innermost loop generates a proper table that is consistent with the marginal totals $\{n_x\}$. This algorithm was programmed in FORTRAN and, like the $2^4$ program, is available via the Internet from the first-named author. The algorithms of Pagano and Tritchler (1983) and Mehta and Patel (1983) are efficient in the sense that they only generate tables that are more extreme
than the one observed.

The algorithm for $2^4$ tables is very similar to the $2^3$ algorithm. The most noticeable difference is that it consists of 11 nested loops and because of its length it will not be given here. For a given value of the sample size, $N$, there are a great many more $2^4$ tables than $2^3$ tables when all marginal totals $n_s$ are close to $N/2$. This observation alone discouraged us from examining higher dimensional tables. A program for $2^5$ tables, for example, would require 26 nested loops. While this program is well within our capabilities to write, it would run very slowly except on the largest supercomputers.

A reasonable question to ask is: How many tables are there in the complete enumeration? Before we start a program it is useful to have a rough estimate of how long it will take to complete. We want to give some ideas about the effort involved in performing the exact test in 3 and 4 dimensions.

The total number of tables in the complete enumeration depends on the overall sample size $N$, of course, but it also depends on the marginal totals in the observed table. If all of the marginal totals $n_s$ and $N - n_s$ are very close to 1 or $N - 1$, for example, the number of tables to enumerate won't be very large relative to $N$. On the other hand, if the marginal totals are all close to $N/2$ then the enumeration effort will be greatest for a given sample size $N$. We make this last assertion based on intuition and will not try to prove it. Below, we provide a bound on the number of $2^3$ tables that is only a function of the two smallest marginal totals $n_1$ and $n_2$.

Figure 1 plots the maximum number of tables needed to enumerate all possible $2^3$ and $2^4$ tables as a function of the sample size $N$. We assume
that all marginal tables $n_3$ are equal to the integer closest to $N/2$. The vertical scale is in log (base 10) units.

The most noticeable feature of Figure 1 is that the $2^4$ enumeration requires a great many more tables than the $2^3$ enumeration for any given sample size, $N$. When $N = 36$, for example, there are almost 106 million possible $2^4$ tables. This enumeration is just about the most one would attempt on a desktop computer. The sheer magnitude of this enumeration effort discouraged us from looking at $2^5$ tables or larger.

For $2^3$ tables, we derived the following bound on the number of tables in the complete enumeration. If $E_N = E_N(n_1, n_2, n_3)$ is the number of tables analyzed in the $2^3$ Algorithm with marginal totals satisfying $1 \leq n_1 \leq n_2 \leq n_3 \leq N/2$ then

$$E_N \leq (n_1 + 1)(n_1 + 2)(n_1 + 3)(2n_2 - n_1 + 2)/12.$$ 

The proof of this assertion is given in the Appendix. In rough terms, the number of $2^3$ tables is about $N^4/192$. This figure is the leading term of our bound for $n_1 = n_2 = N/2$.

This bound is not entirely useful for the small sample sizes in Figure 1. The bound is always about twice as large as the actual number of tables $E_N$ for samples smaller than $N = 256$. Empirically, however, we found that the relative error of this bound improves as the sample size grows. For an example, the data in Table 2 has $n_1 = 384$ (corresponding to South); $n_2 = 623$ (corresponding to the year 1963); and $N = 1663$. The bound for this data suggests that there are $4.14 \times 10^9$ tables in the enumeration. This figure is only 12% greater than the actual number of $3.68 \times 10^9$. A bound
similar to this one for $2^4$ tables can also be derived from the $2^4$ Algorithm but the algebraic effort is prohibitive.

Finally, we want to look at how much varying the marginal totals changes the enumeration effort. Figure 2 plots the total number of $2^3$ tables needed to test mutual independence of all three factors when the sample size $N$ is fixed at 48. The vertical scale is in log(base 10) units. The maximum number of tables (19,097) occurs when $n_1 = n_2 = n_3 = 24 (= N/2)$. The three solid curves in Figure 2 are (from top to bottom):

a) Set $n_2 = n_3 = 24$ and vary only $n_1$

b) Set $n_3 = 24$. Equate and vary $n_1 = n_2$

c) Equate and vary $n_1 = n_2 = n_3$

Figure 2 shows the marginal totals can be as important as the overall sample size in determining the amount of computing effort needed to enumerate all possible contingency tables in order to perform an exact test of significance in higher dimensions. The bound on the maximum number of tables is also plotted (dotted line) in Figure 2. The bound is best when $n_2 = n_3 = 24$ and $n_1$ is at its extremes.

Given these bounds and the availability of software, the serious data analyst should consider using exact methods in higher dimensional tables of categorical data. Exact methods avoid relying on asymptotic approximations for chi-squared statistics and likelihood based inference will generally provide greater power.
Acknowledgements

This research was supported by grants U01-AG10328, R01-AI34110, N01-AI05073 and R01-HS07772 awarded by the U.S. National Institutes of Health, and by a grant from the Minnesota Supercomputer Institute. The FORTRAN source code for the $2^3$ and $2^4$ programs is available from the first author via the Internet at dan@muskie.biostat.umn.edu.

References


Table 1. An artificial example for testing mutual independence of 3 factors in a $2^3$ table. The exact chi-squared distribution agrees with the exact likelihood significance but is very different from the asymptotic chi-squared approximation.

<table>
<thead>
<tr>
<th>A/B</th>
<th>$\bar{A}/\bar{B}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>C/\bar{C}</td>
<td>\begin{pmatrix} 1 &amp; 0 \ 3 &amp; 9 \end{pmatrix}</td>
</tr>
</tbody>
</table>

Table 2. An artificial example for testing mutual independence of 3 factors. The exact chi-squared distribution agrees with its asymptotic approximation but is very different from the exact likelihood tail area.

<table>
<thead>
<tr>
<th>A/B</th>
<th>$\bar{A}/\bar{B}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>C/\bar{C}</td>
<td>\begin{pmatrix} 1 &amp; 0 \ 8 &amp; 17 \end{pmatrix}</td>
</tr>
</tbody>
</table>

Table 3. A $2^3$ table with large counts (Pomar, 1984). Respondents were asked if they believed that blacks should be given equal job opportunity (No, Yes). Respondents were categorized by region of the country in which they lived (North, South) and by the year of the survey (1946, 1963).

<table>
<thead>
<tr>
<th>Year</th>
<th>North No</th>
<th>Yes</th>
<th>South No</th>
<th>Yes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1963</td>
<td>410</td>
<td>56</td>
<td>126</td>
<td>31</td>
</tr>
<tr>
<td>1946</td>
<td>439</td>
<td>374</td>
<td>64</td>
<td>163</td>
</tr>
</tbody>
</table>
Table 4. Sensitivity of chorda tympani fibers in a rat’s tongue to $N =$ sodium chloride, $H =$ hydrogen chloride, $Q =$ quinine, and $S =$ sucrose. (Bishop, Fienberg, and Holland, 1975; pp. 220–1)

<table>
<thead>
<tr>
<th></th>
<th>yes</th>
<th>yes</th>
<th>no</th>
<th>no</th>
</tr>
</thead>
<tbody>
<tr>
<td>$N$</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>$S$</td>
<td>yes</td>
<td>yes</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>yes</td>
<td>yes</td>
<td>1</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>no</td>
<td>yes</td>
<td>3</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>no</td>
<td>no</td>
<td>3</td>
<td>1</td>
<td>4</td>
</tr>
</tbody>
</table>

*This cell is a structural zero (i.e., no observation is possible here) but we are treating it as an observed zero.*
The $2^3$ algorithm: Given sample size $N$ and one-way marginal totals $n_s$, generate all possible $2^3$ tables consistent with these margins. We assume that $1 \leq n_1 \leq n_2 \leq n_3 \leq N/2$.

for $x_{111} = 0$ to $n_1$ do
  for $x_{112} = 0$ to $(n_1 - x_{111})$ do
    for $x_{121} = 0$ to $(n_1 - x_{111} - x_{112})$ do
      $x_{122} = n_1 - x_{111} - x_{112} - x_{121}$
      $r_1 = \max(0, n_2 + n_3 - N - x_{111} + x_{122})$
      $r_2 = \min(n_2 - x_{111} - x_{112}, n_3 - x_{111} - x_{121})$
      for $x_{211} = r_1$ to $r_2$ do
        $x_{212} = n_2 - x_{111} - x_{112} - x_{211}$
        $x_{221} = n_3 - x_{111} - x_{121} - x_{211}$
        $x_{222} = N - n_1 - x_{211} - x_{212} - x_{221}$
        [Accumulate the $2^3$ table $\{x_{ijk}\}$ in the enumeration here]
      next $x_{211}$
    next $x_{121}$
  next $x_{112}$
next $x_{111}$
Figure 1. Maximum number of $2^3$ and $2^4$ tables needed to enumerate the exact test of significance of mutual independence of all factors. The maximum number of tables occurs when all one-way marginal totals are equal to the integer closest to $N/2$ where $N$ is the sample size. The number of tables needed is plotted on a log (base 10) scale. The dotted line is a bound on the number of $2^3$ tables. This bound is explained in Section 3 and derived in the Appendix.
Figure 2. Number of tables needed to enumerate a $2^3$ table with sample size $N = 48$ as a function of the marginal totals. The number of tables is plotted along the vertical axis on a log (base 10) scale. The three solid curves are obtained by varying one, two, or all three marginal totals, top to bottom. The dotted line represents the bound on the maximum number of tables in the enumeration.
Appendix: Bounds on the Number of $2^3$ Tables

Here we derive bounds on the number of tables $E_N$ needed to completely enumerate all $2^3$ tables with a given sample size $N$ and marginal totals $n_s$ ($s = 1, 2, 3$). We can permute the indices of the table so that the three one-way marginal totals satisfy:

$$1 \leq n_1 \leq n_2 \leq n_3 \leq N/2.$$  

The $2^3$ Algorithm shows that

$$E_N = E_N(n_1, n_2, n_3) = \sum_{i=0}^{n_1} \sum_{j=0}^{n_2-i} \sum_{k=0}^{n_3-i-j} \sum_{l=r_1}^{r_2} 1$$

where

$$r_1 = r_1(i, j, k, n_2, n_3, N) = \max(0, n_2 + n_3 - N - 2i - j - k)$$

and

$$r_2 = r_2(i, j, k, n_2, n_3) = \min(n_2 - i - j, n_3 - i - k).$$

An algebraic expression for $E_N$ is difficult to find but an upper bound can be derived replacing $r_1$ by 0 and replacing $r_2$ by either $n_2 - i - j$ or $n_3 - i - k$.

That is, define

$$T_1 = \sum_{i=0}^{n_1} \sum_{j=0}^{n_2-i} \sum_{k=0}^{n_3-i-j} \sum_{l=0}^{n_2-i-j} 1$$

and

$$T_2 = \sum_{i=0}^{n_1} \sum_{j=0}^{n_2-i} \sum_{k=0}^{n_3-i-j} \sum_{l=0}^{n_3-i-k} 1.$$
We then have $E_N \leq \min(T_1, T_2)$. Expressions for $T_1$ and $T_2$ can be found from formulas for sums of powers of consecutive integers. The algebra is rather lengthy and will be omitted. A simplified expression for $T_1$ is

$$T_1 = (n_1 + 1)(n_1 + 2)(n_1 + 3)(2n_2 - n_1 + 2)/12.$$ 

An analogous expression for $T_2$ is

$$T_2 = (n_1 + 1)(n_1 + 2)(n_1 + 3)(2n_3 - n_1 + 2)/12.$$ 

We assumed that $n_2 \leq n_3$ so $T_1 \leq T_2$ and $T_1$ is a sharper upper bound for $E_N$. 

20