Statistica Sinica **12**(2002), 491-504

COMPUTING THE NON-NULL ASYMPTOTIC VARIANCE OF n-WAY CONTINGENCY TABLES WITH FIXED ONE-WAY MARGINAL TOTALS, WITH APPLICATIONS TO GENE-ENVIRONMENT INTERACTION STUDIES

Craig B. Borkowf

National Cancer Institute

Abstract: Consider first the set of all possible n-way multinomial tables defined by certain mean cell proportions with a given number of total counts. Consider next the subset of these n-way tables that, in addition, satisfies certain one-way marginal totals obtained by summing the cell counts over all but one subscript. The subset of tables that satisfies these marginal constraints is said to have the multivariate extended hypergeometric (MXH) distribution. In this paper we develop a general algorithm for calculating the asymptotic variance of n-way MXH tables and present some explicit covariance formulas under independence and in other special cases. We also note that permutation tests defined by certain mean cell proportions and one-way marginal constraints essentially enumerate the entire set of MXH tables with those proportions and constraints. Thus, one can use the asymptotic MXH distribution to approximate the finite sample variances of statistics calculated under permutation tests for various null and alternative hypotheses. One can then use these results to construct confidence intervals for parameters of interest and to approximate the percentiles of test statistics under permutation tests, which is a significant advantage when these tests are computationally prohibitive. We illustrate the use of methods based on the asymptotic MXH distribution as complements and alternatives to permutation tests in the analysis of epidemiological studies of gene-environment interactions.

Key words and phrases: Contingency table, epidemiology, gene-environment interaction, multinomial, multivariate extended hypergeometric (MXH) distribution, permutation test.

1. Introduction

In categorical data analysis, one encounters the extended hypergeometric (XH) distribution in permutation tests for 2×2 contingency tables, such as Fisher's exact test (e.g., Agresti (1990), Plackett (1981)). One can consider the set of 2×2 XH tables to be a subset of all possible 2×2 multinomial (MULT) tables defined by certain mean cell proportions $\{\pi_{ij}\}$ with a given number of total counts t, with the additional requirement that these XH tables must satisfy certain prespecified marginal totals $(\{\pi_{i+}t\}, \{\pi_{+j}t\})$. As a result of these marginal

constraints, XH tables have only one free cell, unlike their MULT counterparts, which have three free cells. The term 'extended' (or 'noncentral') refers to the fact that the mean cell proportions need not be the products of the marginal proportions (i.e., $\pi_{ij} \neq \pi_{i+}\pi_{+j}$). Harkness (1965) first described the moments and asymptotic distribution of XH tables.

More generally, one can consider the set of n-way multivariate extended hypergeometric (MXH) tables to be a subset of all possible n-way MULT tables defined by certain mean cell proportions with a given number of total counts, again with the additional requirement that these MXH tables must satisfy certain one-way marginal totals obtained by summing the cell counts over all but one subscript. Thus, the probability of generating a particular MXH table is directly proportional to the probability of generating the corresponding MULT table, with the set of all MXH probabilities rescaled so that they sum to one. The subset of MXH tables becomes increasingly sparse relative to the set of MULT tables both as the number of free cells and the marginal totals increase. Hence, the calculation of the finite distribution of MXH tables can quickly become quite computationally intensive. (Also, it is generally not practical to simulate MXH tables directly, except under independence, for which Patefield (1981) gives an efficient algorithm.) We will show, however, that one can calculate the asymptotic means and variances of these tables more easily, and for moderate numbers of total counts these asymptotic approximations work reasonably well.

It is crucial to recognize that the one-way marginal constraints of the MXH distribution may differ significantly from the constraints assumed by other common tests. For example, Zelen's and Breslow and Day's tests for homogeneity of odds ratios in $2 \times 2 \times K$ tables both condition on the marginal totals within each of the K 2×2 subtables, as well as on the sum of the counts in the first cell over all K tables (Agresti (1990, p. 238)). These constraints reduce the degrees of freedom to $K - 1$ and thus make these tests computationally less intensive than tests based on the MXH distribution. It is important to remember that there are many possible marginal constraints that one could consider, and therefore one must decide which are reasonable and which are inappropriately restrictive for the problem at hand.

One major motivation for developing algorithms for calculating the asymptotic MXH distribution is their potential use as complements and alternatives to permutation tests. Permutation tests essentially enumerate the entire set of MXH tables generated by given mean cell proportions and marginal counts. Thus, these tests can rapidly become computationally prohibitive as the number of MXH tables to be enumerated increases. One may therefore employ the asymptotic MXH distribution to approximate the finite sample variances of statistics calculated under permutation tests for various null and alternative hypotheses. In turn, these

results can be used in order to construct confidence intervals for parameters of interest and to approximate percentiles and p-values for test statistics under permutation tests. One may also employ asymptotic MXH methods to determine the power of permutation tests under various alternative hypotheses quickly and to study the asymptotic covariance structure of tables under permutation tests, which can be particularly useful for testing multiple hypotheses.

A second motivation for developing these algorithms comes from the study of n-way contingency tables created by partitioning n-dimensional continuous data into categories defined by selected empirical quantiles (e.g., medians or quintiles) of the marginal data. These tables have the empirical multivariate quantile-partitioned (EMQP) distribution (Borkowf, Gail, Carroll and Gill (1997), Borkowf (2000)). Note that multivariate ranks and quantile-categories have the EMQP distribution, and hence statistics like Spearman's rank correlation and the multivariate intraclass correlation have distributions derived from the EMQP distribution. EMQP tables have fixed one-way marginal totals like MXH tables but a substantially more complicated asymptotic distribution. Nevertheless, one can show that EMQP tables have a similar asymptotic covariance structure to MXH tables and that the asymptotic variances of many common statistics calculated from MXH tables serve as lower bounds for those calculated from EMQP tables. Also, under independence, the EMQP and MXH distributions are identical by definition. The above-mentioned papers compare the asymptotic variances of certain statistics calculated from two- and three-way MULT, MXH, and EMQP tables.

Plackett ((1981), Section 6.1) describes a somewhat complicated method for calculating the non-null asymptotic variance of two-way MXH tables. We present some notation and assumptions, and then develop a simplified algorithm that extends this method for n-way MXH tables (Sections 2.1-2.2). We also present an explicit covariance formula under independence (Section 2.3). Finally, we illustrate the use of asymptotic MXH methods as complements and alternatives to computationally intensive permutation tests in the analysis of epidemiological studies of gene-environment interactions (Section 3).

2. Algorithms and Formulas for the Asymptotic MXH Distribution

2.1. Preliminary notation and assumptions

For an n-way contingency table, let d_i denote the number of categories in the ith dimension of this table ($i = 1, \ldots, n$). Let $\underline{a} = (a_1, \ldots, a_n)' = \{a_i\}_{i=1}^n$ denote a column vector of indices ($1 \leq a_i \leq d_i$), and let A_0 denote the set of all index vectors. Note that A_0 has $N_0 = \prod_{i=1}^n d_i$ elements. Next, let $A_1 \subset A_0$ denote the nonsingular subset of index vectors such that at least two of the indices are less than their maxima, i.e., $\sum_{i=1}^n I\{a_i < d_i\} \geq 2$, where $I\{\cdot\}$ denotes the indicator

function. In turn, let $A_c = A_0 \backslash A_1$, and note that A_c has $N_c = \sum_{i=1}^{n} d_i - (n-1)$ elements, the number of unique one-way marginal constraints. Hence, A_1 has $N_1 = N_0 - N_c$ elements, the number of free cells in an n-way MXH table without structural zeros.

Now, let $\pi(\underline{a})$denote the mean cell proportion for the $\underline{a}$ th cell of a given table, $\underline{a} \in A_0$. In turn, let $\pi_i(j)$ denote the prespecified one-way marginal proportion for the jth category of the ith dimension $(i = 1, \ldots, n; j = 1, \ldots, d_i)$. That is, $\pi_i(j) = \sum_{\underline{b} \in B_{ij}} \pi(\underline{b})$, where B_{ij} denotes the set of all index vectors $\underline{b} = \{b_k\}_{k=1}^{n}$ such that $b_i = j$. Furthermore, let t denote the total counts in the MXH table. Obviously, $\pi_i(j)t$ must be an integer in order for the one-way marginal constraints to be satisfied in MXH tables with finite total counts.

Next, let X_{ki} denote the category in which the ith measurement of the kth subject occurs $(k = 1, \ldots, t)$. Then, the observed proportion of counts in the $\underline{a}$th cell can be written as

$$p(\underline{a}) = t^{-1} \sum_{k=1}^{t} \Big[\prod_{i=1}^{n} I\{X_{ki} = a_i\} \Big]. \tag{2.1}$$

Also, let $m(\underline{a}) = p(\underline{a})t$ denote the observed counts in the $\underline{a}$th cell. In turn, let $p_i(j)$ and $m_i(j)$ denote the observed one-way marginal proportion and counts, respectively, for the jth category of the ith dimension $(i = 1, \ldots, n; j = 1, \ldots, d_i)$. The N_c unique one-way marginal constraints yield $m_i(j)/t \equiv p_i(j) \equiv \pi_i(j)$.

Using this notation, one can define the MULT and MXH probabilities of contingency tables. Let $s = \{m(\underline{a}) | \underline{a} \in A_0\}$ denote an observed table of counts, let S_0 denote the set of all possible tables with t total counts, and let $S_1 \subset S_0$ denote the subset of those tables that satisfy the one-way marginal constraints $\{p_i(j)\} = \{\pi_i(j)\}$. Then, the probability of obtaining a particular table $s \in S_0$ under the MULT distribution generated by the mean cell proportions $\{\pi(\underline{a})\}$ and total counts t is

$$P_{MULT}(s) = P_{MULT}(\{m(\underline{a}) | \underline{a} \in A_0\}) = t! \prod_{\underline{a} \in A_0} [\pi(\underline{a})^{m(\underline{a})} / m(\underline{a})!]. \tag{2.2}$$

In turn, the probability of obtaining a particular table $s^* \in S_1$ under the MXH distribution with the same $\{\pi(\underline{a})\}$ and t and with the marginal constraints $\{\pi_i(j)\}$ is

$$P_{MXH}(s^*) = P_{MULT}(s^*) / \sum_{s \in S_1} P_{MULT}(s). \tag{2.3}$$

Finally, for random variables X and Y that converge in probability to their asymptotic means as functions of t, let $\sigma^2(X) = \lim_{t \to \infty} Var(t^{1/2}X), \sigma(X, Y) = \lim_{t \to \infty} Var(t^{1/2}X, t^{1/2}Y)$, and $\gamma(X, Y) = \lim_{t \to \infty} Corr(X, Y)$. Also, let $\delta_{xy} = \delta(x, y) = 1$ if $x = y$ and 0 otherwise.

2.2. An algorithm for computing the asymptotic variance of n-way MXH tables.

Using the above notation, note first that the MULT distribution has the simple finite sample covariance formula $\mathrm{Var}_{MULT}(p(\underline{a}), p(\underline{b})) = [\pi(\underline{a})\delta(\underline{a},\underline{b}) - \pi(\underline{a})\pi(\underline{b})]/t$.

By contrast, the MXH distribution has a much more complex covariance matrix. We simplify and generalize Plackett's results for two-way tables to obtain an expression for the non-null asymptotic variance of n-way MXH tables.

Following Plackett, treat the counts $\{m(\underline{a})\} = \{p(\underline{a})t\}$ in an n-way MXH table as independent Poisson counts with means and variances $\{\pi(\underline{a})t\}$. Let $t \to \infty$ and $m(\underline{a}) \to \infty$ such that $m(\underline{a})/t = p(\underline{a}) \to \pi(\underline{a})$ for all $\underline{a} \in A_0$. Then the asymptotic density of $\{m(\underline{a})\} = \{p(\underline{a})t\}$ is proportional to the singular multivariate normal density

$$q(\{\pi(\underline{a})t\}) \exp\Big[-\frac{1}{2} t \sum_{a \in A_0} (p(\underline{a}) - \pi(\underline{a}))^2/\pi(\underline{a})\Big]. \tag{2.4}$$

In turn, let $V_0 = \lim_{t\to\infty} \mathrm{Var}\,(\{t^{1/2} p(\underline{a}) \,|\, \underline{a} \in A_0\})$ and $V_1 = \lim_{t\to\infty} \mathrm{Var}\,(\{t^{1/2} p(\underline{a}) \mid \underline{a} \in A_1\})$. Note that it is easier to calculate $U_1 = V_1^{-1}$ than V_1 directly. The elements of U_1 can be used to rewrite the quadratic term in (2.4), which involves all N_0 proportions indexed by A_0 , as a quadratic term that involves just the N_1 proportions indexed by A_1. That is,

$$\sum_{a \in A_0} (p(\underline{a}) - \pi(\underline{a}))^2/\pi(\underline{a}) = \sum_{\underline{a} \in A_1} \sum_{\underline{b} \in A_1} [p(\underline{a}) - \pi(\underline{a})] u_1(\underline{a},\underline{b}) [p(\underline{b}) - \pi(\underline{b})], \tag{2.5}$$

where $u_1(\underline{a},\underline{b})$ is the cell of U_1 corresponding to the pair $(p(\underline{a}), p(\underline{b}))$. For $n = 2$ (without structural zeros), the cell of U_1 corresponding to the pair $(p_{ij}, p_{kl}) = (p(\underline{a}), p(\underline{b}))$, where $\underline{a} = (i,j)$ and $\underline{b} = (k,l)$, is

$$u_1(\underline{a},\underline{b}) = \pi_{ij}^{-1}\delta_{ik}\delta_{jl} + \pi_{id_2}^{-1}\delta_{ik} + \pi_{d_1 j}^{-1}\delta_{jl} + \pi_{d_1 d_2}^{-1}. \tag{2.6}$$

Furthermore, for $n \geq 2$ (without structural zeros), one can generalize the formula for the cell of U_1 corresponding to the pair $(p(\underline{a}),\ p(\underline{b}))$ as

$$\begin{aligned} u_1(\underline{a},\underline{b}) = [\pi(\underline{a})]^{-1}\delta(\underline{a},\underline{b}) + \sum_{i=1}^{n} [\pi(\underline{c}_i(a_i))]^{-1}\delta(a_i,b_i) I\{a_i < d_i\} \\ + [\pi(\underline{d})]^{-1}\Big[\sum_{i=1}^{n} I\{a_i < d_i\} - 1\Big]\Big[\sum_{i=1}^{n} I\{b_i < d_i\} - 1\Big], \end{aligned} \tag{2.7}$$

where $\underline{d} = \{d_i\}_{i=1}^n$ and $\underline{c}_i(k) = \{c_{ij}\}_{j=1}^n$ with $c_{ii} = k$ and $c_{ij} = d_j (i,j = 1,\ldots,n; i \neq j)$. One can then invert U_1 to obtain V_1.

Now, note that $\underline{d}$ and $\underline{c}_i$ are elements of A_c. One can write the N_c proportions indexed by A_c in terms of the N_1 proportions indexed by A_1 using the one-way marginal constraints $\{\pi_i(j)\}$, namely

$$\pi(\underline{d}) = (1-n) + \sum_{i=1}^{n} \pi_i(d_i) + \sum_{\underline{a} \in A_1} \left[\sum_{i=1}^{n} I\{a_i < d_i\} - 1 \right] \pi(\underline{a}), \tag{2.8}$$

$$\pi(\underline{c}_i(k)) = \pi_i(k) - \sum_{\underline{a} \in A_1} \pi(\underline{a}) I\{a_i = k\}. \tag{2.9}$$

In turn, one can use equations (2.8) and (2.9) to expand V_1 to the singular covariance matrix V_0.

In general, one must compute the asymptotic MXH covariance matrix numerically, but in special cases, like independence (Section 2.3), one can obtain explicit covariance formulas. For example, for 2×2 MXH tables, which have only one free cell, the elements of the singular MXH covariance matrix V_0 are given by the formula

$$\sigma(p_{ij}, p_{kl}) = (-1)^{\delta_{ik} + \delta_{jl}} (\pi_{11}^{-1} + \pi_{12}^{-1} + \pi_{21}^{-1} + \pi_{22}^{-1})^{-1}. \tag{2.10}$$

In addition, McCullagh and Nelder (1989, p. 262) give explicit covariance formulas for $2 \times d$ MXH tables, and Borkowf (2000) gives explicit covariance formulas related to 2^n MXH tables.

When structural zeros occur in an n-way MXH table (i.e., $\pi(\underline{a}) = 0$), one must set $(p(\underline{a}) - \pi(\underline{a}))^2 / \pi(\underline{a}) \equiv 0$ in (2.4) and then derive an alternative expression for this quadratic in terms of the nonzero proportions indexed by A_1. For example, for 3×3 MXH tables with three structural zeros on the diagonal (i.e., $\pi_{ii} = 0; i = 1, 2, 3$), which also have just one free cell, the elements of the singular MXH covariance matrix V_0 are given by the formula ($i \neq j, k \neq l$)

$$\sigma(p_{ij}, p_{kl}) = (-1)^{\delta_{ik} + \delta_{jl}} (\pi_{12}^{-1} + \pi_{13}^{-1} + \pi_{21}^{-1} + \pi_{23}^{-1} + \pi_{31}^{-1} + \pi_{32}^{-1})^{-1}. \tag{2.11}$$

Furthermore, for any parameter vector $\underline{\lambda} = f(\{\pi(\underline{a})\})$, one can approximate the asymptotic covariance of the statistic $\hat{\underline{\lambda}} = f(\{p(\underline{a})\})$ in terms of $\{\pi(\underline{a})\}$ and $V_0 = \sigma^2(\{p(\underline{a})\})$ using the multivariate delta method (Bishop, Fienberg, and Holland (1975, pp. 492-497)). In addition, one can approximate the finite sample variance $\mathrm{Var}\,(t^{1/2}\{p(\underline{a})\})$ by the well-known approximation (e.g., McCullagh and Nelder (1989, p. 259))

$$\mathrm{Var}\,(t^{1/2}\{p(\underline{a})\}) = t(t-1)^{-1} \sigma^2(\{p(\underline{a})\}) + o(t^{-1}). \tag{2.12}$$

In turn, one can approximate the finite sample variance $\mathrm{Var}\,(t^{1/2}\hat{\underline{\lambda}})$ by

$$\mathrm{Var}\,(t^{1/2}\hat{\underline{\lambda}}) = t(t-1)^{-1} \sigma^2(\hat{\underline{\lambda}}) + o(t^{-1}). \tag{2.13}$$

Key concept: The covariance of any two cells in an n-way MXH table does not depend on the order of the categories within each dimension, so one can rearrange (relabel) the categories within each dimension for convenience. Furthermore, when no structural zeros occur, one can combine multiple categories within a dimension without changing the asymptotic covariances of the unmodified cells. Thus, the problem of computing a particular asymptotic covariance, $\sigma(p(\underline{a}), p(\underline{b}))$, for an arbitrarily large n-way MXH table without structural zeros reduces to that of computing the asymptotic covariance for an appropriately constructed 3^n MXH table (or smaller if $d_i = 2$ for some i).

2.3. Independence

Under n-way independence, $\pi(\underline{a}) = \prod_{i=1}^{n} \pi_i(a_i) \equiv \omega(\underline{a})$. One can compute the finite sample means and covariances of the cell proportions $\{p(\underline{a})\}$ by writing each proportion as the sum of n products of indicator variables, as in (2.1). Then

$$E[p(\underline{a})] = t^{-1} \sum_{k=1}^{t} \Big[\prod_{i=1}^{n} E[I\{X_{ki} = a_i\}] \Big] = \prod_{i=1}^{n} E[I\{X_{1i} = a_i\}] = \prod_{i=1}^{n} \pi_i(a_i) \equiv \omega(\underline{a}), \quad (2.14)$$

$$\begin{aligned}
&E[p(\underline{a})p(\underline{b})] \\
&= t^{-1} \prod_{i=1}^{n} E[I\{X_{1i} = a_i\} I\{X_{1i} = b_i\}] + t^{-1}(t-1) \prod_{i=1}^{n} E[I\{X_{1i} = a_i\} I\{X_{2i} = b_i\}] \\
&= t^{-1} \omega(\underline{a}) \delta(\underline{a}, \underline{b}) + t^{-1}(t-1) \omega(\underline{a}) \prod_{i=1}^{n} \{[\pi_i(b_i) t - \delta(a_i, b_i)]/(t-1)\} \\
&= t^{-1} \omega(\underline{a}) \delta(\underline{a}, \underline{b}) + t^{-1}(t-1) \omega(\underline{a}) \omega(\underline{b}) \Big[1 + (n-1) t^{-1} \\
&\quad - t^{-1} \sum_{i=1}^{n} \Big[\pi_i(a_i)^{-1} \delta(a_i, b_i) \Big] + o(t^{-2}). \qquad (2.15)
\end{aligned}$$

From (2.14) and (2.15), one obtains the finite covariance formula

$$\begin{aligned}
Cov[t^{1/2} p(\underline{a}), t^{1/2} p(\underline{b})] &= \omega(\underline{a}) \delta(\underline{a}, \underline{b}) \\
&\quad + \omega(\underline{a}) \omega(\underline{b}) \Big[(n-1) - \sum_{i=1}^{n} [\pi_i(a_i)]^{-1} \delta(a_i, b_i) \Big] + o(t^{-1}). \qquad (2.16)
\end{aligned}$$

Thus, as $t \to \infty$, one obtains the asymptotic variance

$$\sigma(p(\underline{a}), p(\underline{b})) = \pi(\underline{a}) \delta(\underline{a}, \underline{b}) + \pi(\underline{a}) \pi(\underline{b}) \Big[(n-1) - \sum_{i=1}^{n} [\pi_i(a_i)]^{-1} \delta(a_i, b_i) \Big]. \quad (2.17)$$

Finally, note that for two-way MXH tables, (2.16) reduces to the familiar covariance formula (e.g., Plackett (1981, p. 65))

$$Cov(t^{1/2}p_{ij}, t^{1/2}p_{kl}) = t(t-1)^{-1}\pi_{i+}\pi_{+j}(\delta_{ik} - \pi_{k+})(\delta_{jl} - \pi_{+l}). \qquad (2.18)$$

3. Applications of Asymptotic MXH Methods to Gene-Environment Interaction Studies

In gene-environment interaction studies, some marginal totals are fixed by design, while others are fixed by hypothesis. One might wish to perform permutation tests which condition on the one-way marginal totals to examine certain hypotheses concerning a given data set, since these marginal totals convey no information about the association between the variables of interest. We compare the use of permutation tests and asymptotic MXH methods with respect to numerical results and computational complexity. All computations for this paper were performed by programs written by the author in the GAUSS 3.2 programming language (Aptech Systems, Inc. (1997)).

Example 1. A study of the effects of maternal smoking during pregnancy and a genetic polymorphism on the risk of cleft palate birth defects.

For this study, Hwang, et al. (1995) selected 68 cases of cleft palate birth defects and 281 controls (with non-cleft palate birth defects), and then classified these individuals based on whether their mothers smoked during pregnancy and on their transforming growth factor alpha (TGFα) TaqI genotype. The two levels of maternal smoking are (1) non-smoker and (2) smoker, while the two levels of the TGFα TaqI genotype are (1) wild-type and (2) variant. Table 1 shows that these data form a $2 \times 2 \times 2$ contingency table with 349 counts of environmental exposure by genotype by disease status ($i, j, k = 1, 2$).

Comment. In gene-environment interaction studies, one often assumes that the environmental exposure and genotype classifications are independent. Instead, one focuses on the interaction between the effects of environmental exposure and genotype on disease status. For example, in this study, mothers were assumed to be equally likely to smoke or not smoke during pregnancy regardless of their genotype, and interest mainly focuses on the joint effects of maternal smoking and genotype on birth defect status. (The independence assumption would not hold in a study of alcohol consumption and certain alcohol dehydrogenase polymorphisms on liver damage, because of the negative correlation between the environmental exposure and genotype.) While the independence assumption is essential for case-only studies (Khoury and Flanders (1996)), it also helps shape the hypotheses of interest in case-control studies.

Table 1. $2 \times 2 \times 2$ table of counts for 349 individuals classified by maternal smoking during pregnancy, TGFα TaqI genotype, and cleft palate birth defects.*

environmental exposure: $(i = 1, 2)$	genotype: $(j = 1, 2)$	cleft palate birth defect status: $(k = 1, 2)$	
maternal smoking during pregnancy	TGFα TaqI polymorphism	controls	cases
non-smoker	wild-type	167	36
non-smoker	variant	34	7
smoker	wild-type	69	12
smoker	variant	11	13
total		281	68

*Data are adapted from Hwang, et al. (1995), by permission. The two levels of maternal smoking are (1) non-smoker and (2) smoker; the two levels of the TGFα TaqI genotype are (1) wild-type (homozygous for the TaqI C1 alleles) and (2) variant (having one or two TaqI C2 alleles); and the two levels of disease status are (1) control (non-cleft palate birth defect) and (2) case (cleft palate birth defect). Thus, these data represent a $2 \times 2 \times 2$ table of counts (environment $\times$ genotype $\times$ disease status) $(i, j, k = 1, 2)$.

Khoury and Flanders (1996) analyzed these data to determine whether a synergistic effect exists between the environmental exposure and the genotype on disease status. Let $\theta_G = \pi_{111}\pi_{122}/\pi_{112}\pi_{121}$ denote the odds ratio (OR) for the genetic effect on disease status among non-smokers ($i = 1$), let $\theta_E = \pi_{111}\pi_{212}/\pi_{112}\pi_{211}$ denote the OR for the environmental exposure effect on disease status among individuals with the wild-type genotype ($j = 1$), and let $\theta_B = \pi_{111}\pi_{222}/\pi_{112}\pi_{221}$ denote the OR for the effect of both the environmental exposure and the variant genotype on disease status compared to neither exposure nor variant genotype. In turn, let $\theta_S = \theta_B/(\theta_G \times \theta_E)$ denote the synergistic effect, which indicates how much greater multiplicatively the effect of having both the environmental exposure and variant genotype is than the product of having either risk factor singly. Finally, let $\psi_X = \ln(\theta_X)$, where X denotes the appropriate subscripts.

Table 2 shows the point estimates of the adjusted ORs $\hat{\theta}^*_X$ and the log-ORs $\hat{\psi}^*_X$, calculated from the adjusted sample proportions, $p^*_{ijk} = p_{ijk} + (2t)^{-1}$. The adjustment of $(2t)^{-1}$ avoids the problem of division by zero when $p_{ijk} = 0$ and reduces the bias and mean squared error of the estimators, especially for tables with small counts (e.g., Agresti (1990, p. 54)). Table 2 also gives (1) the means and standard deviations of selected statistics and (2) the percentiles of the sample estimates $\hat{\psi}^*_X$, calculated by asymptotic MXH methods (using $\{\pi_{ijk}\}$)

and by permutation methods (using $\{p^*_{ijk}\}$) under two models for the mean cell proportions $\{\pi_{ijk}\}$. These models are (a) $\pi_{ijk} = p_{ijk}$ (the saturated model) and (b) $\pi_{ijk} = \pi_{i+k}\pi_{+jk}/\pi_{++k}$ (independence between exposure and genotype given disease status.) Note that the permutation methods enumerated $1,812,434$ MXH tables (with $N_1 = 4$ free cells) and took 2 hours on a Pentium III with a 600 MHz processor, while the asymptotic MXH methods took about a second.

One may wish to test the null hypothesis that the synergistic effect does not exist, namely $H_0 : \theta_S \leq 1$, versus the alternative hypothesis that it does exist, $H_A : \theta_S > 1$. The null hypothesis corresponds to model $(b)\pi_{ijk} = \pi_{i+k}\pi_{+jk}/\pi_{++k}$, under which $\theta_S = 1, \psi_S = 0$, and $\sigma(\hat{\psi}_S) = 12.316$. The estimated synergistic effect is $\hat{\theta}^*_S = 6.543$ ($\hat{\psi}^*_S = 1.878$).

Table 2. Selected point estimates and parameters using the asymptotic MXH methods and permutation methods under two models for the data in Table 1.

	observed		asymptotic MXH methods				permutation methods			
parameter	$\hat{\theta}^*_X$	$\hat{\psi}^*_X$	θ_X	ψ_X	$\sigma(\hat{\psi}_X)$	$\%ile$	$E_p(\hat{\theta}^*_X)$	$E_p(\hat{\psi}^*_X)$	$\sigma_p(\hat{\psi}^*_X)$	$\%ile$
Model(a):$\pi_{ijk} = p_{ijk}$.										
θ_G	0.998	-0.002	0.955	-0.046	8.480	0.538	0.955	-0.046	8.607	0.513
θ_E	0.825	-0.192	0.807	-0.215	6.777	0.525	0.807	-0.215	6.824	0.509
θ_B	5.387	1.684	5.482	1.702	8.388	0.484	5.483	1.702	8.455	0.489
θ_S	6.543	1.878	7.115	1.962	12.831	0.451	7.122	1.963	12.986	0.462
Model(b):$\pi_{ijk} = \pi_{i+k}\pi_{+jk}/\pi_{++k}$.										
θ_G	0.998	-0.002	2.185	0.782	7.211	0.021	2.185	0.782	7.239	0.026
θ_E	0.825	-0.192	1.461	0.379	6.208	0.043	1.461	0.379	6.218	0.046
θ_B	5.387	1.684	3.192	1.161	9.395	0.851	3.191	1.160	9.524	0.854
θ_S	6.543	1.878	1.000	0.000	12.316	0.998	1.000	0.000	12.437	0.997

*The rows of this table show the ORs for genotype effect (θ_G), environmental exposure effect (θ_E), joint effect (θ_B), and the synergistic effect (θ_S). The columns of this table show the point estimates of the adjusted ORs $\hat{\theta}^*_X$ and log-ORs $\hat{\psi}^*_X$ calculated from the adjusted sample proportions, $p^*_{ijk} = p_{ijk} + (2t)^{-1}$; the asymptotic MXH means θ_X and ψ_X , standard deviations $\sigma(\hat{\psi}_X)$, and percentiles of $\hat{\psi}^*_X$; and the permutation means $E_p(\hat{\theta}^*_X)$ and $E_p(\hat{\psi}^*_X)$, standard deviations $\sigma_p(\hat{\psi}^*_X)$, and percentiles of $\hat{\psi}^*_X$ The asymptotic MXH methods and the permutation methods are performed under two models for the mean cell proportions, namely (a) $\pi_{ijk} = p_{ijk}$ (the saturated model) and (b) $\pi_{ijk} = \pi_{i+k}\pi_{+jk}/\pi_{++k}$ (independence of exposure and genotype given disease status).

To test the hypothesis of no synergistic effect using asymptotic MXH methods, one can use the test statistic $T = (t-1)^{1/2}(\hat{\psi}^*_S - 0)/\sigma(\hat{\psi}^*_S) = 2.845$, which yields a p-value of 0.0022 (on the standard normal distribution scale), as compared to the permutation p-value of 0.0026. We hence strongly reject the null hypothesis and conclude that there exists a highly significant synergistic effect between maternal smoking during pregnancy and the TGFα TaqI genotype on

the risk of cleft palate birth defects. In turn, we can construct an asymptotic 95% confidence interval for θ_S of the form $\exp(\hat{\psi}_S^* \pm 1.96\sigma(\hat{\psi}_S)(t-1)^{-1/2}) = (1.7, 25.1)$, using the standard deviation under the saturated model.

Example 2. A study of the effects of long-term smoking and a genetic polymorphism on the risk of lung cancer.

Asymptotic MXH methods become particularly useful as the dimensions of the table increase, and when one wishes to test multiple hypotheses. In a study, Nakachi, et al. (1991) selected 45 lung cancer cases and 135 healthy controls (all but two of whom were smokers), and then classified these individuals based on their lifetime cigarette consumption and on their P450IA1 MspI genotype. The three levels of smoking are (1) light, (2) moderate, and (3) heavy, while the two levels of the P450IA1 MspI genotype are (1) wild-type and (2) variant. Table 3 shows that these data form a $3 \times 2 \times 2$ contingency table with 180 counts of environmental exposure by genotype by disease status ($i = 1, 2, 3; j, k = 1, 2$).

Table 3. $3 \times 2 \times 2$ table of counts for 180 individuals classified by smoking status, P450IA1 genotype, and lung cancer.*

environmental exposure: ($i = 1, 2, 3$)	genotype: ($j = 1, 2$)	lung cancer: ($k = 1, 2$)	
smoking status	P450IA1 polymorphism	controls	cases
light	wild-type	79	6
light	variant	9	5
moderate	wild-type	22	11
moderate	variant	4	4
heavy	wild-type	18	16
heavy	variant	3	3
total		135	45

*Data are adapted from Nakachi, et al. (1991), by permission. The three levels of smoking are (1) light, (2) moderate, and (3) heavy (corresponding to lifetime cigarette consumption doses of < 3, $3 - 4$, and > 4 ($\times 10^5$), respectively); the two levels of the P450IA1 MspI genotype are (1) wild-type (homozygous or heterozygous for the more common alleles) and (2) variant (homozygous for the rare form of the alleles); and the two levels of disease status are (1) control (non-cancer) and (2) lung cancer. Thus, these data represent a $3 \times 2 \times 2$ table of counts (environment $\times$ genotype $\times$ disease status) ($i = 1, 2, 3$; j, $k = 1, 2$).

Nakachi, et al. (1991) analyzed these data to determine whether the effects of the variant genotype, which was thought to increase the risk of lung cancer,

became "washed-out" at higher levels of cigarette consumption. Let $\theta_{G|Ei} = \pi_{i11}\pi_{i22}/\pi_{i12}\pi_{i21}$ denote the OR for the genetic effect on disease status at the ith level of the environmental exposure ($i = 1, 2, 3$). In turn, let $\theta_{Ril} = \theta_{G|Ei}/\theta_{G|El}$ denote the ratio of the two ORs for exposure levels i and l. As before, let $\psi_X = \ln(\theta_X)$, where X denotes the appropriate subscripts. Note that $\theta_{R31} = \theta_{R21}\theta_{R32}$ and hence $\psi_{R31} = \psi_{R21} + \psi_{R32}$.

Table 4 shows the point estimates of the adjusted ORs $\hat{\theta}^*_X$ and the log-ORs $\hat{\psi}^*_X$, calculated from the adjusted sample proportions, $p^*_{ijk} = p_{ijk} + (2t)^{-1}$. Table 4 also gives (1) the means and standard deviations of selected statistics and (2) the percentiles of the sample estimates $\hat{\psi}^*_X$, calculated by asymptotic MXH methods (using $\{\pi_{ijk}\}$) and permutation methods (using $\{p^*_{ijk}\}$) under two models for the mean cell proportions $\{\pi_{ijk}\}$. These models are (a) $\pi_{ijk} = p_{ijk}$ (the saturated model) and (b) $\pi_{ijk} = \pi_{i++}\pi_{+jk}$ (independence between exposure and genotype $\times$ disease status). Note that the permutation methods enumerated 120,943,907 MXH tables (with $N_1 = 7$ free cells) and took 5 days on a Pentium III with a 600 MHz processor, while the asymptotic MXH methods again took about a second.

Next, one may wish to test the hypothesis that the wash-out effect does not exist, $H_0 : \theta_{G|E1} = \theta_{G|E2} = \theta_{G|E3}$ (homogeneity of ORs), against the alternative hypothesis that the wash-out effect does exist, $H_A : \theta_{G|E1} \geq \theta_{G|E2} \geq \theta_{G|E3}$, with at least one strict inequality. The null hypothesis corresponds to model $(b)\pi_{ijk} = \pi_{i++}\pi_{+jk}$, under which $\theta^*_{G|Ei} = 2.705$ at each level of exposure and hence $\theta_{Ril} = 1$ ($\psi_{Ril} = 0$). The sample estimates of the ORs are $\theta^*_{G|E1} = 7.081, \theta^*_{G|E2} = 1.957$, and $\theta^*_{G|E2} = 1.121$, and the estimated ratios are $\theta^*_{R21} = 0.276$, $\theta^*_{R31} = 0.158$, and $\theta^*_{R32} = 0.573$.

To test the hypothesis of no wash-out effect using asymptotic MXH methods, one can construct test statistics of the form $T_{il} = (t-1)^{1/2}(\hat{\psi}^*_{Ril} - 0)/\sigma(\hat{\psi}_{Ril})$. We obtain $T_{21} = -1.198$, $T_{31} = -1.702$ and $T_{32} = -0.433$, which yield percentiles of 0.115, 0.0444, and 0.332 (on the standard normal distribution scale), as compared to the permutation percentiles of 0.115, 0.0497, and 0.330. Thus, while we can reject $\theta_{G|E1} = \theta_{G|E3}$ in favor of $\theta_{G|E1} > \theta_{G|E3} (\psi_{R31} < 0)$ at the 5% significance level, we cannot further partition this result to reject either $\theta_{G|E1} = \theta_{G|E2}$ or $\theta_{G|E2} = \theta_{G|E3}$. We hence conclude that an overall wash-out effect appears to exist, but the local wash-out effects are not statistically significant.

The subject of multiple comparisons in contingency tables is an active area of research, and most methods are based on multivariate normal approximations (Hochberg and Tamhani (1987, pp. 278-281)). One may wish to use the correlations between statistics of interest to design tests that are less conservative than those based on Bonferroni methods, especially when these statistics are positively correlated. In this example, however, model (b) yields asymptotic MXH correlations $\gamma(\hat{\psi}_{R21}, \hat{\psi}_{R31}) = 0.290$, $\gamma(\hat{\psi}_{R21}, \hat{\psi}_{R32}) = -0.591$, and $\gamma(\hat{\psi}_{R31}, \hat{\psi}_{R32}) =$

0.600, as compared to the permutation correlations $\gamma_p(\hat{\psi}^*_{R21}, \hat{\psi}^*_{R31}) = 0.281$, $\gamma_p(\hat{\psi}^*_{R21}, \hat{\psi}^*_{R31}) = -0.595$, and $\gamma_p(\hat{\psi}^*_{R31}, \hat{\psi}^*_{R32}) = 0.604$. One can use the χ^{2+}_2 test (Follmann (1996)) to test for an overall wash-out effect. Let $\underline{x} = (\hat{\psi}_{R21}, \hat{\psi}_{R32})'$ and $V = (t-1)^{-1}\hat{\sigma}^2(\underline{x})$. Then $\underline{x}'V^{-1}\ \underline{x} = 3.439$, which corresponds to the 0.179 percentile of the χ^2_2 distribution. Since $\hat{\psi}_{R31} = \hat{\psi}_{R21} + \hat{\psi}_{R32} < 0$, one obtains a p-value of 0.090 using the χ^{2+}_2 test. We hence fail to establish a wash-out effect with this alternative test.

Table 4. Selected point estimates and parameters using the asymptotic MXH methods and permutation methods under two models for the data in Table 3.

	observed		asymptotic MXH methods				permutation methods			
parameter	$\hat{\theta}^*_X$	$\hat{\psi}^*_X$	θ_X	ψ_X	$\sigma(\hat{\psi}_X)$	$\%ile$	$E_p(\hat{\theta}^*_X)$	$E_p(\hat{\psi}^*_X)$	$\sigma_p(\hat{\psi}^*_X)$	$\%ile$
Model(a): $\pi_{ijk} = p_{ijk}$.										
$\theta_{G\|E1}$	7.081	1.957	7.315	1.990	9.396	0.482	7.314	1.990	9.666	0.488
$\theta_{G\|E2}$	1.957	0.671	2.000	0.693	10.703	0.489	2.001	0.693	11.088	0.493
$\theta_{G\|E3}$	1.121	0.114	1.125	0.118	11.885	0.498	1.125	0.118	12.329	0.494
θ_{R21}	0.276	1.286	0.273	-1.297	14.242	0.504	0.274	-1.296	14.709	0.504
θ_{R31}	0.158	-1.843	0.154	-1.872	15.150	0.510	0.154	-1.872	15.666	0.510
θ_{R32}	0.573	-0.557	0.563	-0.575	15.994	0.506	0.563	-0.575	16.581	0.506
Model(b): $\pi_{ijk} = \pi_{i++}\pi_{+jk}$.										
$\theta_{G\|E1}$	7.081	1.957	2.705	0.995	7.771	0.951	2.703	0.994	7.861	0.951
$\theta_{G\|E2}$	1.957	0.671	2.705	0.995	12.076	0.360	2.714	0.998	12.510	0.346
$\theta_{G\|E3}$	1.121	0.114	2.705	0.995	12.226	0.168	2.716	0.999	12.652	0.160
θ_{R21}	0.276	-1.286	1.000	-0.000	14.361	0.115	1.004	0.004	14.775	0.115
θ_{R31}	0.158	-1.843	1.000	0.000	14.487	0.044	1.005	0.005	14.895	0.050
θ_{R32}	0.573	-0.557	1.000	0.000	17.185	0.332	1.001	0.001	17.793	0.330

*The rows of this table show the ORs for genotype effect at each level of environmental exposure ($\Theta_{G|E_i}$), and the ratio of these odds ratios ($\theta_{Ril} = \Theta_{G|E_i}/\theta_{G|El}$). The columns of this table show the point estimates of the adjusted ORs $\hat{\theta}^*_X$ and log-ORs $\hat{\psi}^*_X$ calculated from the adjusted sample proportions, $p^*_{ijk} = p_{ijk} + (2t)^{-1}$; the asymptotic MXH means θ_X and ψ_X , standard deviations $\sigma(\hat{\psi}_X)$, and percentiles of $\hat{\psi}^*_X$; and the permutation means $E_p(\hat{\theta}^*_X)$ and $E_p(\hat{\psi}^*_X)$, standard deviations $\sigma_p(\hat{\psi}^*_X)$, and percentiles of $\hat{\psi}^*_X$. The asymptotic MXH methods and the permutation methods are performed under two models for the mean cell proportions, namely (a) $\pi_{ijk} = p_{ijk}$ (the saturated model) and (b) $\pi_{ijk} = \pi_{i++}\pi_{+jk}$ (independence of exposure and genotype × disease status).

Acknowledgement

This research was performed while the author held a CRTA Postdoctoral Fellowship at the National Cancer Institute's Cancer Prevention Studies Branch. He wishes to thank Mitchell H. Gail and Dean A. Follmann for helpful discussions, *American Journal of Epidemiology* and *Cancer Research* for permission to use

the data sets in the examples, and an associate editor and two referees for helpful comments on this paper.

References

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

Aptech Systems, Inc. (1997). *GAUSS Mathematical and Statistical System*. Aptech Systems, Inc., Maple Valley, WA.

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

Borkowf, C. B., Gail, M. H., Carroll, R. J. and Gill, R. D. (1997). Analyzing bivariate continuous data grouped into categories defined by empirical quantiles of the marginal distributions. *Biometrics* **53**, 1054-1069.

Borkowf, C. B. (2000). On multidimensional contingency tables with categories defined by the empirical quantiles of the marginal data. *J. Statist. Plann. Inference* **91**, 33-51.

Follmann, D. (1996). A simple multivariate test for one-sided alternatives. *J. Amer. Statist. Assoc.* **91**, 854-861.

Harkness, W. L. (1965). Properties of the extended hypergeometric distribution. *Ann. Math. Statist.* **36**, 938-945.

Hochberg, Y. and Tamhane, A. C. (1987). *Multiple Comparison Procedures.* John Wiley, New York.

Hwang, S.-J., Beaty, T. H., Panny, S. R., Street, N. A., Joseph, J. M., Gordon, S., McIntosh, I. and Francomano, C. A. (1995). Association study of transforming growth factor alpha (TGFα) TaqI Polymorphism and oral clefts: Indication of gene-environment interaction in a population-based sample of infants with birth defects. *Amer. J. Epidemiology* **141**, 629-636.

Khoury, M. J. and Flanders, W. D. (1996). Nontraditional epidemiologic approaches in the analysis of gene-environment interaction: Case-control studies with no controls! *Amer. J. Epidemiology* **144**, 207-213.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. 2nd edition. Chapman and Hall/CRC, Boca Raton.

Nakachi, K., Imai, K., Hayashi, S.-I., Watanabe, J. and Kawajiri, K. (1991). Genetic susceptibility to squamous cell carcinoma of the lung in relation to cigarette smoking dose. *Cancer Research* **51**, 5177-5180.

Patefield, W. M. (1981). An efficient method of generating random $r \times c$ tables with given row and column totals. *Appl. Statist.* **30**, 91-97.

Plackett, R. L. (1981). *The Analysis of Categorical Data*. 2nd edition. Macmillan, New York.

CRTA Postdoctoral Fellow, National Cancer Institute (NCI), Center for Cancer Research (CCR), Cancer Prevention Studies Branch (CPSB), 6116 Executive Blvd., Suite 705, MSC 8314, Bethesda, MD 20893-8314, U.S.A.

E-mail: borkowfc@mail.nih.gov

(Received September 2000; accepted September 2001)