Likelihood Ratio Tests With Three-Way Tables

Philip E. CHENG, Michelle LIOU, and John A. D. ASTON

Likelihood ratio (LR) tests for association and for interaction are examined for three-way contingency tables, particularly the widely used $2 \times 2 \times K$ table. Mutual information identities are used to characterize the information decomposition and the logical relationship between the omnibus LR test for conditional independence across K strata and its two independent components, LR tests for interaction and for uniform association. The latter two tests are logically connected to formulating a natural two-step test for conditional independence. The proposed two-step test with reduced nominal levels is suggested instead of the Breslow–Day test and the Cochran–Mantel–Haenszel test. This yields efficient interval estimation for both the interaction parameter and the common odds ratio compared with using the Mantel–Haenszel estimate. This allows the development of power analysis for testing general hypotheses of varied interactions, using an invariant Pythagorean law of relative entropy.

KEY WORDS: Breslow–Day test; Cochran–Mantel–Haenszel test; Mutual information; Pearson test; Three-way interaction.

1. INTRODUCTION

The analysis of contingency tables with three-way classifications has been much studied in the literature, notably in the case of a $2 \times 2 \times K$ table. This analysis is often concerned with testing for association within and homogeneity across the strata. Overall, testing conditional independence of the table can be found by considering not only a direct test, but also a combination of the two previous tests. The aim of this article is to develop tests based on information identities that give independent tests for both steps, allowing them to be naturally combined into a two-step test for conditional independence. These tests are also contrasted with tests that are usually used to test each step individually, but do not necessarily have the required independence to allow them to be combined into a two-step test.

The literature concerning testing of all three quantities individually is well established. Bartlett (1935) initiated a test for no interaction across strata (i.e., equality or homogeneity of odds ratios) and derived an estimate of the common odds ratio (COR) with a pair of 2×2 tables. Norton (1945) extended the discussion to finite K tables, and Simpson (1951) supplied interpretations of various interactions. Woolf (1955) discussed estimation and testing for the COR, and Roy and Kastenbaum (1956) proved that Bartlett's estimate is a conditional maximum likelihood (ML) estimate given the margins of each stratum in an $I \times J \times K$ table. The classical approaches to testing homogeneity or, equivalently, to testing the converse (interaction) were based mainly on weighted chi-squared tests, and were further discussed by Plackett (1962) and Goodman (1964). On the other hand, a chi-squared test with 1 df for two-way independence across K strata was proposed by Cochran (1954) and Mantel and Haenszel (1959). This is the Cochran–Mantel–Haenszel (CMH) test for conditional independence or partial association (Birch 1964; Goodman 1969) that has been widely used in the literature.

These early studies led to further analyses of three-way tables for estimating the COR, testing association, and testing interaction across strata. For $2 \times 2 \times K$ tables, the Bartlett test for interaction involves inconvenient computations for the conditional ML estimate (given the margins of each 2×2 table) of the COR, as noted by Birch (1963). To relax the computational burden, Goodman (1964) discussed approximate chi-squared tests and tests using conditional or unconditional ML estimate (fixing only one margin of each 2×2 table) of the COR were discussed by Gart (1971), Zelen (1971), and Halperin et al. (1977). The score test for homogeneity (Breslow and Day 1980) gained popularity as the Breslow–Day (BD) test because of its ease of computation by using the Mantel–Haenszel (MH) COR estimate, although Paul and Donner (1992) found in a simulation study that score tests for interaction tend to be conservative when the odds ratios are unequal.

A related topic of special interest in biomedical research is the interval estimation of the COR between strata. Woolf (1955) introduced weighted logit COR estimators. The popular MH COR estimate is defined with intuitive appeal and is asymptotically efficient when the common odds ratio is equal to unity (see, e.g., Birch 1964; Nurminen 1981; McCullagh and Nelder 1989). For the finite strata case with moderate to large data sets, Tarone, Gart, and Hauck (1983) and Hauck (1984) showed through simulation studies that a conditional ML estimate is generally superior to the unconditional ML estimate or the MH estimate in terms of bias and precision.

Whereas testing homogeneity and estimating the COR have been much discussed in the literature, testing interaction and testing partial association seldom have been studied together. Goodman (1969, 1970) and Bishop, Fienberg, and Holland (1975) discussed these hypotheses using likelihood ratio (LR) tests for the parameters of log-linear models. In the literature of educational statistics and psychometrics, analysis of three-way tables using the log-linear models and the CMH test has been a popular approach to identifying test items exhibiting differential item functioning (DIF) between two social groups (e.g., Mellenberg 1982; Holland and Thayer 1988; Swaminathan and Rogers 1990; Wang and Yeh 2003). These studies generally first test for interaction using different methods, then test for partial

Philip E. Cheng is Research Fellow and Michelle Liou is Research Fellow, Institute of Statistical Science, Academia Sinica, Taipei 115, Taiwan. John A. D. Aston is Associate Professor, CRiSM, Department of Statistics, University of Warwick, Coventry CV4 7AL, U.K. (E-mail: *j.a.d.aston@warwick.ac.uk*). The authors thank the editor, associate editor, and two reviewers for comments that helped improve the article. They also thank Professor David Firth for his very useful feedback on the manuscript. This work was partially supported by the NSC (grant 2118-M-001-009, to P. C.) and the EPSRC (grant EP/H031936/1, to J. A.).

Journal of the American Statistical Association
June 2010, Vol. 105, No. 490, Theory and Methods
DOI: 10.1198/jasa.2010.tm09061

association using mainly the CMH test. However, they overlooked a fact, noted by Goodman (1969), that the log-likelihood of the conditional independence can be expressed as a sum of two independent terms. Specifically, the sum is the omnibus LR test for conditional independence. The first summand tests for interaction, the heterogeneity of odds ratios, termed the *nonuniform DIF* in psychometrics. The second summand is used to test for homogeneous or *uniform association* within strata, called the *uniform DIF*, or the partial association by Birch (1964) and numerous follow-up studies.

Unlike classical analysis of variance with continuous variables, these three LR tests are logically related; that is, a significant interaction tested by the first test implies the rejection of the other two hypotheses. Thus it is crucial to analyze how data information shared between the two independent summands affects interpretation of the three LR tests at their associated levels of significance. If the same nominal level is used for all three tests, then the omnibus test can be inconsistent with either one of the other two terms, such that it becomes less sensitive than the combination of the other two terms—that is, a naive two-step test that uses the second term to test for uniform association only when the first test for homogeneity is not rejected. The drawback can be corrected by using reduced nominal levels, against which the two independent terms are tested. In the literature, this logical relation between the two hypotheses has received little attention, and the two hypotheses are usually tested separately using the BD and the CMH tests. They often are assumed to be independent when used, which is not the case. A few examples will illustrate the difference between using the two popular classical tests and using the two-step LR tests for the two hypotheses.

The primary goal of this study was to examine certain likelihood information identities for the three LR tests in three-way tables, to develop a two-step test for comparison with the omnibus test for conditional independence. The secondary goal was to develop a power analysis for testing general hypotheses of unequal interactions in three-way tables, which extends the notion and application of a two-step test. For example, the probability of observing the data under arbitrary patterns of interactions can be evaluated, and efficient interval estimation of the interaction parameter and of the COR can be derived from the two-step test procedures.

The article is organized as follows. The background of testing hypotheses with a $2 \times 2 \times K$ table is briefly reviewed in Section 2. Information identities that divide the data log-likelihood into orthogonal components are discussed in Section 3. As an extension of an information identity of Cheng et al. (2008, lemma 2), power analysis at alternative varied interactions, or unequal ratios of odds ratios, is given in Theorem 1 of Section 3. The three LR tests are examined using the notion of a two-step test, which also yields efficient interval estimation of the interaction parameter and of the COR as a byproduct. Application of the two-step test is illustrated with examples in Section 4. Two data analyses of the BD and the CMH tests in the literature are examined against the proposed two-step LR tests, and the performances are compared in terms of p-values and interval estimates of the COR in Section 5. A power evaluation of unequal odds ratios as a consequence of the two-step test in the examples is illustrated. Overall, this study constructs a geometric frame of testing conditional independence, interaction, and uniform association and develops the two-step LR tests based on the mutual information identity.

2. TESTING HYPOTHESES FOR $2 \times 2 \times K$ TABLES

Statistical inference for association between two categorical variables is of considerable interest in many applications. Data from a case-control study with a dichotomous risk factor are often stratified into 2×2 tables by a third variable with K levels. Let (X, Y, Z) denote the three-way categorical vector, and let (X_k, Y_k) denote pairs of dichotomous variables, where Z is the K-level $(k = 1, \ldots, K)$ stratum variable. The observed data are frequency counts n_{ijk} of subjects with condition i [$i = 1$ (case), 2 (control)] and exposure j [$j = 1$ (exposed), 2 (nonexposed)], which fall in stratum k, $k = 1, \ldots, K$. Let $U = \{U_k = (n_{11k}, n_{12k}; n_{21k}, n_{22k}), k = 1, \ldots, K\}$ denote the observed K strata of 2×2 tables. We use a dot notation for summation over a subscript; that is, $n_{\cdots} = n$ denotes the total sample size, $n_{1\cdot k}$ is the number of cases in stratum k, and $n_{\cdot 2k}$ is the total number of nonexposed subjects in stratum k, and so on.

In this study it is assumed that data can arise from a wide range of experiments, including multinomial, independent multinomial, Poisson, or hypergeometric sampling across strata. The inference in this study is essentially independent of these sampling schemes, because the issue of testing homogeneity and conditional independence between the two main variables, the case and the risk factor, is discussed via conditional likelihood inference across strata. Let the odds ratios of the 2×2 tables be defined by $\psi_k = p_{11k}p_{22k}/p_{12k}p_{21k}$, $k = 1, \ldots, K$, where $p_{ijk} = P(X = i, Y = j, Z = k)$, $i, j = 1$ or 2, are the cell proportions. The null hypothesis of independence between the two variables in each stratum is the hypothesis of conditional independence, denoted by

$$H_0 : \psi_k = 1 \quad \text{for } k \in \{1, \ldots, K\}. \tag{2.1}$$

The traditional test for H_0 is the Pearson chi-squared test using the statistic

$$\chi^2_{PE} = \sum_{k=1}^{K} \sum_{i,j=1}^{2} \frac{(n_{ijk} - n_{i\cdot k}n_{\cdot jk}/n_{\cdot\cdot k})^2}{n_{i\cdot k}n_{\cdot jk}/n_{\cdot\cdot k}}. \tag{2.2}$$

It approximates the chi-squared distribution with K df, denoted by χ^2_K, which admits a continuity correction for each stratum of small sample size (Yates 1934).

A widely discussed issue is the hypothesis of equal odds ratio across K strata. Testing equal odds ratios between strata was initially studied as the testing of no second-order interaction by Bartlett (1935). This is the hypothesis of COR,

$$H_1 : \psi_k = \psi \quad \text{for } k \in \{1, \ldots, K\} \tag{2.3}$$

for a positive constant ψ. By definition, (2.1) is a special case of (2.3), that is, $H_0 \subset H_1$. Thus rejection of H_1 logically implies rejection of H_0, and the implication is valid against a common nominal level of test, say $\alpha = 0.05$. The individual tests for H_0 and H_1 can yield varying significant results against this logical implication. Assuming COR, denoted by ψ under H_1, the MH estimate of ψ is defined as

$$\psi_{MH} = \frac{\sum_{k=1}^{K}(n_{11k}n_{22k}/n_{\cdot\cdot k})}{\sum_{k=1}^{K}(n_{12k}n_{21k}/n_{\cdot\cdot k})}. \tag{2.4}$$

A popular test that can be easily computed for H_1 uses the estimate ψ_{MH}. The BD score test is defined as

$$\chi^2_{BD} = \sum_k \frac{e_k^2}{\text{var}(n_{11k}|\psi_{MH})}. \tag{2.5}$$

Here the adjusted cell estimates, e_k, and the denominator variance can be easily found (e.g., Agresti 2002, p. 232). Under H_1, the test statistic (2.5) approximates the chi-squared distribution with $K-1$ df.

Also designed for testing H_0 is the CMH test, defined by the statistic

$$\chi^2_{CMH} = \frac{(\sum_{k=1}^K n_{11k} - \sum_{k=1}^K n_{1\cdot k}n_{\cdot 1k}/n_{\cdot\cdot k})^2}{\sum_{k=1}^K \{n_{1\cdot k}n_{2\cdot k}n_{\cdot 1k}n_{\cdot 2k}/n^2_{\cdot\cdot k}(n_{\cdot\cdot k}-1)\}}. \tag{2.6}$$

The test statistic (2.6), derived from an estimating equation, assumes equality of the K odds ratios, that is, $\psi_k = \psi$, the COR. In essence, it is used to test the hypothesis $H_2 = (H_0|H_1)$; that is, given a COR ψ,

$$H_2 : \psi = 1, \tag{2.7}$$

which defines the same hypothesis H_0 implicitly conditioned on that H_1 is not rejected. It follows that rejection of H_1 logically implies rejection of H_2. This logical relation is the basis of the two-step tests discussed in this study. The test statistic (2.6) has two equivalent versions that approximate the chi-squared distribution with 1 df. Cochran's test (1954) is defined with the independent binomial distributions model, while the MH version is derived from a Fisher's exact test hypergeometric distribution using the Yates continuity correction. Extension of (2.6) to a $I \times J \times K$ table was discussed by Landis, Heyman, and Koch (1978) and Somes (1986).

Although LR tests defined in Section 3.2, as well as score tests for H_1 using either conditional or unconditional ML estimates of the COR, have been widely studied, the notion of testing H_2 has received little attention. The relationships among testing the three hypotheses, H_0, H_1, and H_2 have not been discussed in the literature, with the exception of Goodman (1969). We report a systematic study of the LR tests after defining some terminology necessary to introduce likelihood information. In the literature, hypotheses H_0 and H_2 are usually termed conditional independence and partial association, respectively, but the two hypotheses are often mixed in use. To avoid ambiguity in this study, the hypothesis H_0 tested for conditional independence (across strata), H_1 tested for homogeneity (of odds ratios) or interaction (across strata), and H_2 tested for uniform association (within strata).

3. LIKELIHOOD RATIO TESTS FOR THREE–WAY TABLES

One main objective of this study is to clarify that the sample version of the Pythagorean law (3.4) yields simple LR tests for H_0, H_1, and H_2 and provides a unified inference framework compared with the BD and the CMH tests described in Section 2. Proposition 1 and Theorem 1 extend (3.4) to testing general hypotheses of conditional independence and nonunity interactions, respectively. The goal of comparing the inference of the derived statistics is illustrated using the notion of two-step test in Section 4.

3.1 An Information Identity

Let (X, Y, Z) be the variables of a three-way $I \times J \times K$ contingency table. Let $f(i,j,k) = P(X=i, Y=j, Z=k)$, $f(i)$, $g(j)$, $h(k)$; $i = 1,\ldots,I$, $j = 1,\ldots,J$, $k = 1,\ldots,K$, denote the joint and marginal probability density functions (pdf's). The well-known Shannon entropy defines a basic information identity

$$H(X) + H(Y) + H(Z) = I(X,Y,Z) + H(X,Y,Z), \tag{3.1}$$

where $H(X,Y,Z) = -\sum_{(i,j,k)} f(i,j,k) \cdot \log f(i,j,k)$ is the joint entropy, and marginal entropies such as $H(X)$ are defined likewise. The term $I(X,Y,Z) = \sum_{(i,j,k)} f(i,j,k) \cdot \log\{f(i,j,k)/f(i)g(j)h(k)\}$ denotes the mutual information between the three variables (see Gray 1990; Cover and Thomas 1991). One main contribution of this study is the formulation of a framework for testing associations based on the mutual information $I(X,Y,Z)$ of (3.1). The mutual information characterizes the minimum divergence from the joint pdf to a hyperplane of products of the marginal pdf, that is, the projection from the data to the parameter space of the independence hypothesis of the variables [Cheng et al. 2007, (2.9)]. Furthermore, $I(X,Y,Z)$ admits three equivalent expressions in terms of the obvious likelihood decomposition, for example,

$$\begin{aligned}\log\left\{\frac{f(i,j,k)}{f(i)g(j)h(k)}\right\} &= \log\left\{\frac{f(i,k)}{f(i)h(k)}\right\} + \log\left\{\frac{f(i,j,k)}{f(i,k)g(j)}\right\} \\ &= \log\left\{\frac{f(i,k)}{f(i)h(k)}\right\} + \log\left\{\frac{f(j,k)}{g(j)h(k)}\right\} \\ &\quad + \log\left\{\frac{f(i,j,k)/h(k)}{f(i|k)f(j|k)}\right\},\end{aligned} \tag{3.2}$$

where convenient notations $f(i,j)$ and $f(i|j)$ are used to denote the joint pdf and conditional pdf, respectively. By taking expectations of the sampling versions of both sides of (3.2), an orthogonal decomposition of the mutual information using Z as the (common) conditioning variable (CV) is expressed as

$$I(X,Y,Z) = I(X,Z) + I(Y,Z) + I(X,Y|Z). \tag{3.3}$$

Three information-equivalent forms of (3.2) and (3.3) are obtained using each of the three variables as the CV. The conditional mutual information $I(X,Y|Z)$ on the right side of (3.3) measures the association between X and Y at each level of Z, against which the hypothesis H_0 is defined and tested. A further decomposition leads to a key identity used in this study [Cheng et al. 2007, (2.12)],

$$I(X,Y|Z) = \text{Int}(X,Y,Z) + I(X,Y \parallel Z). \tag{3.4}$$

The first summand, $\text{Int}(X,Y,Z)$, on the right side of (3.4) defines the three-way interaction between X and Y across Z, which is unique because of the symmetry in the three variables. Its sample version can be computed by the iterative proportional fitting (IPF) scheme of Deming and Stephan (1940) (see, e.g., Agresti 2002, p. 343). The second summand, $I(X,Y \parallel Z)$, obtained from subtracting $\text{Int}(X,Y,Z)$ from $I(X,Y|Z)$, quantifies the uniform association between X and Y, given Z. Its sample version is an analog of the CMH/MH test statistic (2.6). Equations (3.3) and (3.4) are also valid with sample frequencies, and the sample analogs on the right side of (3.3) and (3.4) are statistically independent. The sample versions of the three terms of (3.4) approximate chi-squared distributions with $(I-1)(J-1)K$, $(I-1)(J-1)(K-1)$, and $(I-1)(J-1)$ df, respectively (Cheng et al. 2007). These are discussed in the next section.

3.2 LR Tests for $2 \times 2 \times K$ Tables

Recall $U = \{U_k,\ k = 1, \ldots, K\}$, the observed data from K strata of 2×2 tables in the variables (X, Y, Z), as defined before (2.1). Denote the conditional ML estimate under H_0 by $W_k = (n^*_{11k}, n^*_{12k}; n^*_{21k}, n^*_{22k})$, $k = 1, \ldots, K$, where $n^*_{ijk} = n_{i\cdot k} n_{\cdot jk}/n_{\cdot\cdot k}$ are the conditional ML estimates of the cell proportions given the margins, which are the sufficient statistics, of each 2×2 table. The LR test for independence of (X, Y) in a 2×2 table, say $\{n_{ij},\ i, j = 1, 2\}$, can be expressed by a sample Kullback–Leibler (KL) divergence (Kullback and Leibler 1951). That is, maximizing the multinomial LR statistic $\Pi_{i,j}[\{f(i,j)^{n_{ij}};\ f(i,j) \in H_0\}/\{p(i,j)^{n_{ij}};\ p(i,j) = n_{ij}/n\}]$ for testing $p(i,j) \in H_0$ is equivalent to minimizing a scaled KL divergence $D(f = p : f = f^* \in H_0) = \sum_{i,j=1}^{2} n_{ij} \log(n \times n_{ij}/n_{i\cdot} \times n_{\cdot j})$ (see, e.g., Cheng et al. 2008, lemma 1). To extend testing H_0 of a 2×2 table to that of K independent 2×2 tables, the analogous LR test can be expressed as a sum of K KL divergence statistics. This is the sample analog of $I(X, Y|Z)$, often termed a deviance statistic when discussing generalized linear models (McCullagh and Nelder 1989). This LR test can be denoted by

$$D_0 = 2D(U : W) = 2 \sum_{k=1}^{K} \sum_{i=1}^{2} \sum_{j=1}^{2} n_{ijk} \log(n_{ijk}/n^*_{ijk}) \cong \chi^2_K(H_0). \tag{3.5}$$

Thus D_0 defines the LR test statistic, giving the omnibus test under H_0. The last term of (3.5) explains its approximate distribution, χ^2_K, under H_0, comparable to the Pearson chi-squared test, χ^2_{PE}, of (2.2). The first term on the right side of (3.4) characterizes the conditional ML estimate under H_1 by $V = \{V_k = (\hat{n}_{11k}, \hat{n}_{12k}; \hat{n}_{21k}, \hat{n}_{22k}),\ k = 1, \ldots, K\}$, which can be computed by the IPF scheme. The IPF finds V, the combined 2×2 tables given the conditional ML estimate of the COR, denoted by $\hat{\psi}$ (e.g., Bishop, Fienberg, and Holland 1975). It was noted after (3.4) that D_0 equals the sum of two independent components. The first summand of (3.4) defines the LR test for H_1 through the sample KL-divergence statistic

$$D_1 = 2D(U : V) = 2 \sum_{k=1}^{K} \sum_{j} \sum_{i} n_{ijk} \log(n_{ijk}/\hat{n}_{ijk}) \cong \chi^2_{K-1}(H_1). \tag{3.6}$$

As the sample version of $\mathrm{Int}(X, Y, Z)$, D_1 approximates a χ^2_{K-1} in distribution under H_1, directly comparable to the BD test of (2.5). The second component of (3.4) defines the sample analog of $I(X, Y \parallel Z)$ to be

$$D_2 = 2D(V : W) = 2 \sum_{k=1}^{K} \sum_{j} \sum_{i} \hat{n}_{ijk} \log(\hat{n}_{ijk}/n^*_{ijk}) \cong \chi^2_1(H_0|H_1), \tag{3.7}$$

which is comparable to the CMH test (2.6) and approximates a χ^2_1 under H_2. By (3.7), D_2 tests for $H_2 = (H_0|H_1)$ conditional on H_1 not being rejected; otherwise it is not necessary, because logically the rejection of H_1 implies the rejection of H_2. As mentioned in Section 1, this fact has essentially been overlooked when using the CMH test. By (3.4), the independent LR tests D_1 and D_2 are favorable competitors to the BD and the CMH tests, respectively. In short, the three LR test statistics (3.5), (3.6), and (3.7) satisfy the approximate equation in distribution $\chi^2_K(H_0) \cong \chi^2_{K-1}(H_1) + \chi^2_1(H_2)$.

Proposition 1. Let the data be the $2 \times 2 \times K$ table, $U = \{(n_{11k}, n_{12k}; n_{21k}, n_{22k}),\ k = 1, \ldots, K\}$. Let W be the ML estimate $(n^*_{11k}, n^*_{12k}; n^*_{21k}, n^*_{22k})$, $k = 1, \ldots, K$, under H_0, and let V be the ML estimate $(\hat{n}_{11k}, \hat{n}_{12k}; \hat{n}_{21k}, \hat{n}_{22k})$, $k = 1, \ldots, K$, under H_1. It then follows by (3.4) that the LR test statistics satisfy the same identity,

$$D_0 = D_1 + D_2. \tag{3.8}$$

3.3 Power Analysis of the LR Tests

As mentioned in Section 1, power evaluation at a class H' of patterns of unequal odds ratios across strata, against the null hypothesis H_0 or H_1, can be developed. This power analysis can be used to estimate the sample size needed to meet a criterion of specificity and sensitivity for testing H'. For a strata of K 2×2 tables, there are $K - 1$ ratios between consecutive pairs of the odds ratios or the consecutive three-way interactions. Without loss of generality, using the case where $K = 2$ suffices to clarify the theory. Additional notation is needed. Let the observed $2 \times 2 \times 2$ table be $U = (U_1; U_2)$, where $U_1 = (a, b; c, d)$ and $U_2 = (e, f; g, h)$. As an extension of Proposition 1, let $W' = (W_1; W_2)$ be a member of H', which is a class of alternatives to H'_1. Here it is convenient to assume that $W_1 = (a^*, b^*; c^*, d^*)$ and $W_2 = (e^*, f^*; g^*, h^*)$ have the same total counts as U_1 and U_2, respectively and unequal sample odds ratios, $\psi_1 = a^*d^*/b^*c^*$ and $\psi_2 = e^*h^*/f^*g^*$, such that $1 \neq \gamma = \psi_1/\psi_2 > 0$. Given $W' \in H'$, there exists a unique member V' in H'_1 such that $V' = (V_1 = (\hat{a}, \hat{b}; \hat{c}, \hat{d}); V_2 = (\hat{e}, \hat{f}; \hat{g}, \hat{h}))$ would have the same margins as those of $U = (U_1; U_2)$, and the same ratio γ as the interaction between V_1 and V_2 (V_i need not have ψ_i as its odds ratio). An analog of (3.8) holds. This is summarized in Theorem 1, which is proved in Appendix A, which also includes the proof of (3.8) as a special case.

Theorem 1. Let U be a $2 \times 2 \times K$ table. Let $W' \in H'$ be another $2 \times 2 \times K$ table having the same table totals as those of U, sample odds ratios $(\psi_1, \ldots, \psi_K)$, and consecutive three-way sample interactions $1 \neq \gamma_i = \psi_i/\psi_{i+1} > 0$, $i = 1, \ldots, K - 1$. Then there is a unique $2 \times 2 \times K$ table V', $V' \in H'_1$, having the same table margins as those of U, such that an extension of (3.8) holds,

$$D(U : W') = D(U : V') + D(V' : W'). \tag{3.9}$$

Equation (3.9) is a generalization of the invariant Pythagorean law of relative entropy in testing two-way independence (e.g., Cheng et al. 2008, lemma 2). It is seen that eq. (3.9) generalizes (3.8) from testing H_0 and H_1 with interaction parameter $\gamma_i = 1$ to testing H' and H'_1 with $\gamma_i \neq 1$. The same asymptotic chi-squared distributions for (3.8) are also valid for (3.9) and yield the same calculated p-values for H_0 and H_1 as for H' and H'_1. Figure 1 presents the geometry illustrating the similarity between eq. (3.8) (lower hyperplane with $\gamma = 1$) and eq. (3.9) (upper hyperplanes with $\gamma \neq 1$), associated with hypotheses H_1 and H'_1, respectively. For the case where $K = 2$ with a pair of 2×2 tables, D_1 of (3.8) tests for the

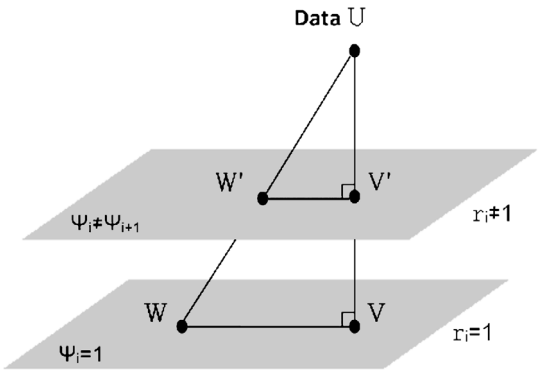

Figure 1. Null hypotheses: $D(U:W) = 0 = D(U:V) + D(V:W)$, $\gamma_i = 1$; alternative hypotheses: $D(U:W') = 0 = D(U:V') + D(V':W')$, $\gamma_i \neq 1$.

hypothesis H_1 of no interaction, that is, $\gamma = 1$. It also leads to an interval estimation for the interaction parameter γ through using $D(U:V')$ of (3.9).

In Section 4 we show that eq. (3.8) leads to the development of a two-step test for conditional independence across strata. Moreover, eq. (3.9) is key to developing a two-step test for unequal interaction parameters or for unequal odds ratios. An application of the power evaluation of (3.9) to a simple case, a pair of unequal odds ratios, is given in Example 4 of Section 5. Indeed, Theorem 1 can yield a particular inference framework; $D(U:V')$ is useful for testing a general hypothesis of consecutive interaction parameters $\gamma_i \neq 1$ between several 2×2 tables, for example, to test a trend of odds ratios. Overall, the inference derived from Theorem 1 directly provides a complete framework for testing hypotheses, of which the basic frame is similar to the logit modeling for testing conditional independence with a nonzero interaction parameter (e.g., Swaminathan and Rogers 1990; Agresti 2002, secs. 5.4 and 6.3). In the case of sparse tables with small sample sizes, LR tests might not yield desirable inference, particularly with a large number of sparse tables. In these cases, exact conditional inference is usually recommended (cf. Mehta and Patel 1983; Agresti 2002, sec. 6.7), for which our proposed two-step test procedure can be analogously defined. These facts can be summarized in the following corollary. For economy of exposition, a straightforward proof is omitted.

Corollary 1. For $K = 2$, the statistic $D(U:V')$ tests for a specific value of the interaction parameter γ ($\neq 1$) and provides an interval estimation for the parameter γ of the observed data U.

3.4 Relation to Log-Linear Models

At this point, it is of interest to note that the likelihood ratio tests for $2 \times 2 \times K$ tables as expressed by the foregoing information identities can be equivalently obtained by considering tests within a log-linear model setting. If the data are considered using a log-linear model, then D_1 can be expressed as an LR test statistic for the three-factor interaction term in the log-linear model. In addition, D_0 can be expressed as the LR test statistic for testing no association between X and Y and no three-way interaction. This also provides an easy way to obtain D_2, by finding D_0 and D_1 from the log-linear model and then subtracting. However, the relationship between the test statistics and any particular individual term in the log-linear model is not straightforward in general. For example, in Example 1 of Section 4.3, D_2 can be derived by subtraction, although this does not follow naturally from considering the terms in the hierarchical log-linear model, because D_0 is computed while ignoring the test significance for the interaction. We discuss the role of two-step tests in more detail in the next section. Even if the log-linear model were not required to be hierarchical and thus an interaction term was still included when testing the association between X and Y, this would not determine D_2. This is because the terms in the log-linear model are not generally orthogonal, whereas the information tests are orthogonal by construction. Thus, for higher-way tables, the terms cannot be straightforwardly matched on a term-by-term basis between the information and log-linear approaches, but possibly could be obtained by considering linear combinations of log-linear model LR test statistics that span orthogonal subspaces. In turn, this also could provide a way to examine the size and power relationships of tests when carrying out sequential tests of parameters in log-linear models, as the two-step tests for examining H_1 and H_2 discussed next.

4. TWO–STEP TESTS FOR H_1 AND H_2

4.1 Interval Estimation of the Common Odds Ratio

Confidence interval (CI) estimation of the COR, ψ, was first studied by Woolf (1955), Mantel and Haenszel (1959), and Gart (1962). Thereafter, the asymptotic variance of the MH estimate received most discussion (e.g., Thomas 1975; McKinlay 1978; Breslow and Liang 1982; Tarone, Gart, and Hauck 1983; Robins, Breslow, and Greenland 1986). On the other hand, the conditional ML estimate, $\hat{\psi}$ (fixing both margins of each 2×2 table), of the COR differs slightly in magnitude from either the MH estimate (2.4) or the conditional ML estimate using extended hypergeometric distributions (Gart 1970), as well as the unconditional ML estimate (fixing one margin of each 2×2 table). It has been noted that $\hat{\psi}$ might be preferred to the unconditional ML estimate in terms of both bias and efficiency (Hauck 1984). In contrast to those studies, here we propose CI estimation of ψ using the information identity (3.8).

The right side of eq. (3.8) can be used to define a two-step test for H_0, in contrast to the left-side test D_0. It first tests H_1 using D_1, and if H_1 is rejected, then so logically is H_2, and the test is concluded, noting that there is evidence of significant interaction; otherwise, testing H_2 proceeds, using D_2 as the second step. A byproduct of the two LR tests is a simple interval estimation procedure for the COR ψ. If H_1 is rejected, then the KL divergence statistic D_1 from the data to the line of COR (Figure 1, lower hyperplane) is significantly large. This implies that defining a COR is not statistically justified, even though a CI for ψ can be estimated using the foregoing equations. This would however not be valid as rejection of H_1 implies there is no COR. If H_1 is not rejected, then a standard CI of the COR ψ can be estimated by inverting the approximating χ^2_1 distribution of D_2. To illustrate this, Proposition 1 asserts that the first-step test D_1 computes the ML estimate V, which is the $2 \times 2 \times K$ table with the COR $\hat{\psi}$, a base value within a CI for ψ. Next, applying D_2 yields the two-sided CI (W_l, W_u) of $2 \times 2 \times K$ tables

centered at V using the approximate distribution χ_1^2; computing the CORs of the pair (W_l, W_u) yields (ψ_l, ψ_u), which is the desired two-sided CI that includes $\hat{\psi}$. By standard LR test theory, this approach yields asymptotically efficient CI estimation for ψ, without formulating the asymptotic variance for $\hat{\psi}$. An example of the proposed CI is given in Example 4 of Section 5.

In theory, the CMH test is formulated as a score test without using ψ_{MH}; thus a CI of the COR cannot be found by inverting a chi-squared distribution. A popular CI is derived from a logarithmic transformation of ψ_{MH} with an asymptotic variance estimate (e.g., Robins, Breslow, and Greenland 1986), which is known to be asymptotically efficient under H_0, and the resulting CI is usually comparable to that given by the two-step LR test. However, the difference between these two interval estimation procedures focuses on the logical implications of the two-step procedure. The proposed two-step LR test uses D_2 to estimate a CI only when D_1 is insignificant under the same joint likelihood testing framework using the statistical independence between the two tests. In the literature, a CI of the COR using ψ_{MH} has not been preceded by a test of H_1, because no independent test is available to precede with the CMH test.

4.2 A Naive Two-Step Test

Figure 1 illustrates that the hypotenuse defines the omnibus test D_0 for H_0, and that the two sides of the right triangle form a basic pair of LR tests. First, D_1 is used for testing H_1, then D_2 is used to test H_2 only if H_1 is not rejected. The two-step test arises from the identity (3.8), which is true for the LR tests but not for other tests, such as BD and CMH. This two-step procedure also holds for logit models, as mentioned in Section 3.3, because there are popular logit-normal random-effects models for testing and estimating H_0, H_1, and H_2 (e.g., Skene and Wakefield 1990; Liu and Pierce 1993; Agresti and Hartzel 2000). For the LR tests, because $H_0 \subset H_1$, and $H_2 = (H_0|H_1)$, both D_1 and D_2 are considered effective for testing H_0. But care must be taken when choosing the size of the tests. If the same test level α is used for both D_1 and D_2 as usual, then the effective test size for H_0 would be about 2α, twice that of D_0. To illustrate this further, the initial discussion of a two-step test that follows takes the same nominal level α for each of the three LR tests.

For fixed K and a usual nominal level range, say $0 < \alpha \leq 0.10$, let

$$C = C_K = \{D_0 > q_{K,\alpha}\} \tag{4.1}$$

be a level-α critical region under H_0, where $q_{K,\alpha}$ denotes the $100(1-\alpha)$ percentile of the χ_K^2 distribution. Similar critical regions are defined as the events $A = A_{K-1} = \{D_1 > q_{K-1,\alpha}\}$ and $B = \{D_2 > q_{1,\alpha}\}$, based on the quantiles of χ_{K-1}^2 and χ_1^2 distributions, respectively. Then, for fixed K, the critical region of the naive two-step test is defined to be

$$E = E_K = A \cup (A^c \cap B), \tag{4.2}$$

where the superscript c denotes the complement of an event. Under H_1 and H_2, event E is governed by the likelihood, $P_{H_1}(A) = \alpha$, of rejecting H_1 (hence H_2) by D_1, such that $P_{H_1}(A^c) = 1-\alpha$ is the likelihood of using D_2 when H_1 is not rejected. The size of this two-step test (for H_0) is $\alpha + (1-\alpha)\alpha = 2\alpha - \alpha^2$, slightly less than 2α. For $K \geq 2$ and $0 < \alpha \leq 0.10$, the following inequality holds from the properties of the chi-squared distribution:

$$q_{K,\alpha} < q_{K-1,\alpha} + q_{1,\alpha}. \tag{4.3}$$

Thus, by (4.1) and (4.2), define, for fixed K, $F = F_K = C \cap A^c \cap B^c$ as the subset of C not contained in E. In contrast, let the disjoint union $G = (A \cap C^c) \cup (A^c \cap B \cap C^c)$ denote the subset of E not included in C. It can be shown that the naive two-step test (for H_0) is more sensitive than the omnibus test D_0; that is, it follows from (4.3) that

$$\begin{aligned} P(F) &\leq P(q_{K,\alpha} < D_0 \leq q_{K-1,\alpha} + q_{1,\alpha}) \\ &\leq P(q_{K,\alpha} < D_0) - P(q_{K-1,\alpha} \leq D_1 \text{ and } q_{1,\alpha} \leq D_2) \\ &= \alpha - \alpha^2. \end{aligned} \tag{4.4}$$

In contrast,

$$\begin{aligned} P(G) &= P\{(A \cap C^c) \cup (A^c \cap B \cap C^c)\} \\ &= P(C^c) - P(C^c \cap A^c \cap B^c) \\ &\geq P(C^c) - P(A^c)P(B^c) \\ &= \alpha - \alpha^2. \end{aligned} \tag{4.5}$$

Thus (4.4) and (4.5) conclude that

$$P(F) \leq P(G). \tag{4.6}$$

As noted in Section 4.1, the COR interval estimate arises from the event $A^c \cap B$ of (4.2), given that testing for H_1 by D_1 is insignificant at level α. But the test levels of D_1 and D_2 need not be equal to α, as we discuss in the next section.

4.3 The Two-Step Test

As illustrated earlier, a natural competitor to the omnibus test D_0 is a two-step test that combines the first test D_1 for H_1 and the second test D_2 for H_2, because of the fact that rejection of H_1 implies rejection of H_0, and also of H_2. Suppose that two distinct test levels, α_1 and α_2, are separately used for D_1 and D_2; then the critical regions $A = \{D_1 > q_{K-1,\alpha_1}\}$ and $B = \{D_2 > q_{1,\alpha_2}\}$ can be defined by analogy with (4.2) against the same critical region (4.1) of the test D_0. To maintain the same test level α as in (4.1), the equation

$$\begin{aligned} P(E) &= P_{H_1}(A) + \big\{1 - P_{H_1}(A)\big\}P_{H_2}(B) \\ &= \alpha_1 + (1-\alpha_1)\alpha_2 = \alpha \end{aligned} \tag{4.7}$$

must be satisfied, where $P_{H_0|H_1}(A^c \cap B) = P_{H_2}(B)$, because D_1 and D_2 are independent. Equation (4.7) admits numerous solutions having $\alpha_1 + \alpha_2 \simeq \alpha$ within the range $0 < \alpha \leq 0.10$, when the product $\alpha_1\alpha_2$ is rather small. To choose the levels α_1 and α_2, given (3.8), the following can be used for a given level α:

$$\underset{\alpha_1,\alpha_2}{\operatorname{argmin}}\big\{q_{K-1,\alpha_1} + q_{1,\alpha_2} | \alpha_1 + (1-\alpha_1)\alpha_2 = \alpha\big\}. \tag{4.8}$$

Thus, when $K = 2$, a common level $\alpha_1 = \alpha' = \alpha_2$ is recommended, for which the solution is $\alpha' = 1 - \sqrt{1-\alpha}$, which is slightly greater than $\alpha/2$ for $0 < \alpha \leq 0.10$; for example, $\alpha' = 0.0253$ when $\alpha = 0.050$. For $K \geq 3$, the choices of the pair (α_1, α_2) will depend on the upper percentiles of the chi-squared distributions. Because $\alpha_i < \alpha$, it seems unlikely that a two-step test for H_0 will be as sensitive as the omnibus test. Indeed, the converse of inequality (4.6) holds, stated as (4.9) in the next proposition, which is proved in Appendix B.

Proposition 2. Define the same events C, E, F, and G as in (4.1)–(4.5), except that the test levels of D_1 and D_2 are now replaced by α_1 and α_2 according to (4.7). Then the converse of (4.6) holds for the two-step test, that is,

$$P(F) \geq P(G). \tag{4.9}$$

The critical regions A, B, and C of the three LR tests, as defined by (4.1) and (4.2), present eight possible combinations of significant and insignificant tests. This corresponds to dividing the sample space into distinct subsets, among which the event $A \cap B \cap C^c$ is impossible (i.e., is the empty set), in view of inequality (4.3). This fact and (4.3) continue to be true when the levels α_1 and α_2 (smaller than α) are used in a two-step test. To illustrate the seven possibly nonempty events, it can be checked that the data given by Cheng et al. [2006, table 1, (3.2) and (3.3)] give a case of the event $A \cap B \cap C$ that has all three tests significant. It also is easy to identify a case with all three tests insignificantly small in magnitude; that is, $A^c \cap B^c \cap C^c$ occurs when the cell counts in the tables are in essence homogeneous. Examples 1–5 present empirical data that characterizes the remaining five cases and illustrates the performance of the two-step LR tests compared with the BD and CMH tests in these settings, as well as the performance of the adjusted and unadjusted test sizes.

In the examples that follow, $\alpha = 0.05$, and the values of α_1 and α_2 are specified.

Example 1. Consider a $2 \times 2 \times 3$ table (Table 1) that has odds ratios $\psi_1 = 0.47$, $\psi_2 = 0.36$, and $\psi_3 = 2.52$, which vary around unity.

Computing the usual test statistics of Section 2, $\chi^2_{PE} = 7.00$, $p = 0.07$ and $\chi^2_{CMH} = 0.642$, $p = 0.42$, indicating that H_0 is not rejected. But $\psi_{MH} = 0.77$ and $\chi^2_{BD} = 6.41$, $p = 0.04$, which suggests rejecting H_1, and hence H_0, because $H_0 \subset H_1$. In contrast, the LR tests yield $D_0 = 7.04$, $p = 0.07$; $D_1 = 6.39$, $p = 0.04$; and $D_2 = 0.66$, $p = 0.42$. Like χ^2_{PE}, the LR test D_0 for H_0 is insignificant, and it seems that the event $A \cap B^c \cap C^c$ occurs for LR tests as well as for the classical tests. Although the test D_1 for H_1 is significant, it is not at a lower level (using $\alpha_1 = 0.027$ and $\alpha_2 = 0.024$ as $K = 3$) using the two-step test. The results of the two-step LR tests are consistent, whereas the BD and CMH tests are inconsistent, because the significant result for the BD test implies rejection of all three hypotheses.

Example 2. Consider a $2 \times 2 \times 3$ table (Table 2) that has similar odds ratios: $\psi_1 = 3.25$, $\psi_2 = 3.33$, and $\psi_3 = 2.96$.

The Pearson test yields $\chi^2_{PE} = 5.84$, $p = 0.14$; $\psi_{MH} = 3.15$, $\chi^2_{BD} = 0.013$, $p = 0.99$; and $\chi^2_{CMH} = 5.74$, $p = 0.017$. The MH test (using continuity correction) yields $\chi^2_{MH} = 4.33$, with $p = 0.037$. For the LR tests, the omnibus test is insignificant as $D_0 = 4.73$, $p = 0.19$; moreover, $D_1 = 0.013$, $p = 0.99$ and $D_2 = 4.72$, $p = 0.03$, and thus the event $A^c \cap B \cap C^c$ occurs. The CMH test is significant, but the LR test yields an insignificant two-step test at level $\alpha_2 = 0.024$, despite that D_2 is significant at level α. The MH test for H_2 was significant, but it could be comparable to the second-step test D_2, if the same adjusted level 0.024 were used. In this case, the MH test may be preferred to the CMH test, because there are a few cell counts below 5.

Table 1. Example $2 \times 2 \times 3$ data table with odds ratios varying around unity

9	19	6	15	22	8
18	18	10	9	12	11

Table 2. Example $2 \times 2 \times 3$ data table with very similar odds ratios ($\psi_1 = 3.25$, $\psi_2 = 3.33$, and $\psi_3 = 2.96$)

1	3	80	9	3	9
4	39	8	3	8	71

Examples 1 and 2 illustrate that the excessive sensitivity of the two independent LR tests may be relieved using the adjusted test sizes of a two-step test. However, given the nonindependence of the BD and CMH tests, it is difficult to formulate an appropriate adjustment for the test sizes of the combined BD and CMH tests.

Example 3. Consider a $2 \times 2 \times 2$ table (Table 3) with odds ratios $\psi_1 = 30$ and $\psi_2 = 1.125$.

For these data, $D_0 = 6.81$, $p = 0.033$; $D_1 = 3.14$, $p = 0.076$; and $D_2 = 3.67$, $p = 0.056$. This is a special case where the omnibus (hypotenuse) LR test is significant, yet the two independent components are insignificant, and thus the event $A^c \cap B^c \cap C$ occurs. In accordance with Proposition 2, a two-step test can be less sensitive than the omnibus test.

5. EMPIRICAL STUDY

Examples 1–3 present a subset of the comparison conditions between the omnibus test D_0 and the two-step test, D_1 followed by D_2, along with the corresponding BD and CMH tests. To complement the illustration of the remaining conditions, in this section we present two real data examples from the literature. In addition, a basic use of the power analysis of Theorem 1 is explained in Example 4. In the next two examples, analyses of the previous authors are reported and compared with those obtained from the LR tests. The level $\alpha = 0.05$ is used unless stated otherwise.

Example 4 refers to data extracted from part of an empirical study that examined the association between allele frequency of a type (or genotype), and a case-control diabetes type across population subdivision strata (Ardlie, Lunetta, and Seielstad 2002, table 2).

Example 4. Data of two 2×2 tables are genotypes and allele frequencies for certain polymorphisms in the Polish and U.S. samples. Odds ratios between the frequencies of allele type and of case-control, and p-values of association tests, including the MH test, were studied. The data from their table 2 are briefly illustrated in Table 4. The data are a pair of 2×2 tables denoted

Table 3. Example $2 \times 2 \times 2$ data table with very different odds ratios ($\psi_1 = 30$ and $\psi_2 = 1.125$)

5	1	3	4
1	6	2	3

Table 4. Genotypes and allele frequencies for the PPARg Pro12Ala polymorphism in the Polish and U.S. DM2 samples (from Ardlie, Lunetta, and Seielstad 2002, table 2)

	Poland		U.S.	
Allele freq.\Genotype	C	G	C	G
Case	62	419	48	447
Control	92	371	51	445

by $U = (U_1, U_2)$. The row factors were case and control, and the column factors were C and G allele types. U_1 is the table for Poland, with sample size $n_1 = 944$, and U_2 is the table for U.S. with $n_2 = 991$.

The authors computed the sample odds ratios, 0.597 and 0.937, for the two tables, respectively, and the COR estimate $\psi_{MH} = 0.719$, with a 95% CI $(0.60, 0.87)$ that excludes 0.937 and barely includes the other sample odds ratio, 0.597. The CMH test yields $\chi^2_{CMH} = 5.88$ with $p = 0.015$ (or $\chi^2_{MH} = 5.56$ with $p = 0.018$), which led to a significant conclusion that "the two odds ratios are different."

The omnibus test is $D_0 = 8.55$ with $p = 0.014$, and $K = 2$ df. Then $D_1 = 2.646$ with $p = 0.104$, and the conditional ML estimate is $\hat{\psi} = 0.718$; further, $\psi_{MH} = 0.719$ and $\chi^2_{BD} = 2.653$, $p = 0.103$. Thus, both tests for interaction are insignificant, justifying a common odds ratio for Poland and the U.S. because their distribution patterns are alike. Because the omnibus test D_0 is significant, and the test for H_1 is insignificant, H_2 is tested. Now the second-step test yields $D_2 = 5.905$ with $p = 0.015$, which is significant at level $\alpha_2 \approx \alpha/2 = 0.025$. This yields a significant two-step test at level 0.05, a similar result to the significant CMH test together with the insignificant BD test, and the event $A^c \cap B \cap C$ occurs. The conclusion follows that there is evidence that the odds ratios differ from 1, but no evidence that they differ from each other.

It follows from an insignificant first-step test for H_1 that a 95% CI, $(0.549, 0.938)$, of the COR ψ is legitimately computed by inverting the sampling distribution χ^2_1 of the statistic D_2, based on and centered at $\hat{\psi}$. The result is closely comparable to the popular CI $(0.551, 0.941)$ obtained from estimating the standard error of a transformed MH estimate, as illustrated at the end of Section 4.1, although the first CI does not require the asymptotic variance estimate that is needed for the second CI.

Suppose that there was doubt regarding possible undercounts or missing cases behind the observed data U. In accordance with Theorem 1, it may be assumed that an alternate table could have been observed, say $W' = (W'_1, W'_2)$ (see Table 4′). The odds ratio of the alternative table W'_1 for Poland is $\psi_1 = 0.693$, and that of W'_2 for the U.S. is $\psi_2 = 1.00$, such that the interaction parameter is $\gamma' = \psi_1/\psi_2 = 0.693$. The question is, if the table of unequal odds ratios $W' \in H'$ were a valid hypothesis, would it be supported by the observed data U? It is seen that $2D(U : W') = 5.450$, with $p = 0.066$ against the χ^2_2 distribution, where W' is normalized to have the same total as the data U. Thus the pair of odds ratios of W' are not far from those of U, and the omnibus test is insignificant. Meanwhile, by eq. (3.9) there exists a table $V' \in H'_1$ with the same margins of data U and the same interaction 0.693 of W'. It follows that the first-step test, $2D(U : V') = 0.092$, is insignificantly small, but the second-step test, $2D(V' : W') = 5.358$, with $p = 0.020$, is significant at level 0.025. Thus the event $A^c \cap B \cap C^c$ occurs, with the interpretation that testing W' with $(\gamma' = 0.693)$, which is closer to the sample ratio $\gamma = 0.597/0.937 = 0.637$ of the data U than any member of H_1 $(\gamma = 1)$, can be seemingly insignificant by the omnibus test, but not by the two-step LR test. Theorem 1 illustrates that testing a hypothetical table of unequal odds ratios (H'), and testing its ratio of odds ratios, the interaction parameter (H'_1), are two different LR tests, but they can be connected through a two-step LR test.

Table 4′. Possible alternative table of genotypes and allele frequencies

	Poland		U.S.	
Allele freq.\Genotype	C	G	C	G
Case	80	420	51	445
Control	110	400	51	445

The next example refers to data from a study of the effect of progestogens versus placebo on miscarriage, stillbirth, or neonatal death, which cites a previous study on "hormone administration for maintenance of pregnancy" by Reis, Hirji, and Afifi (1999, table III).

Example 5. Pregnancy data from a meta-analysis of clinical trials (Reis, Hirji, and Afifi 1999) with seven 2×2 tables (their table III) of two dichotomous factors: the case-control and the treatment conditions, progestogens and placebo.

For the data in their table III, Reis, Hirji, and Afifi (1999) computed several exact and asymptotic tests for H_1. The BD test yielded $\chi^2_{BD} = 11.26$, with $p = 0.081$ and $K - 1 = 6$ df. All other tests yielded similar results, with p-values in the range $[0.055, 0.09]$, except that $p = 0.032$ was given by the unconditional LR test, the only test that did not support H_1. It was suspected that "the unconditional LR test was too liberal in terms of size."

The omnibus test for H_0 yields $D_0 = 14.13$, $p = 0.049$, with 7 df, which is marginally significant. The LR test for H_1 yields $D_1 = 13.77$, $p = 0.032$, with 6 df, which agrees with the omnibus test and the unconditional LR test, but not with the BD test, at the same level, $\alpha = 0.05$. As mentioned in Section 1, this presents a case where the BD test may be more conservative than the LR test D_1, as noted by Hauck (1984) and Paul and Donner (1992). Meanwhile, $\chi^2_{CMH} = 0.353$, $p = 0.553$, which is comparable to the insignificant LR test $D_2 = 0.358$, $p = 0.550$. Thus the event $A \cap B^c \cap C$ occurs with the LR tests with all tests at the level α.

Continuing with the tests but at an adjusted level, it is found that, using the scheme recommended in Section 4.3, the levels $(\alpha_1, \alpha_2) \simeq (0.030, 0.020)$ should be used. In this case, neither H_1 nor H_2 is rejected (although H_1 is fairly marginal). The COR estimate $\psi_{MH} = 0.853$, or $\hat{\psi} = 0.850$, is marginally meaningful, although H_2 implies that there is no evidence that $\psi \neq 1$. Thus this example demonstrates that when tests of level α are used, the event $A \cap B^c \cap C$ occurs, whereas the adjusted tests yield the event $A^c \cap B^c \cap C$, showing the decrease in sensitivity of the adjusted two-step test in accordance with Proposition 2.

Overall, the examples demonstrate that the use of LR tests developed through the information decomposition allows a description of the power and sensitivity of the omnibus and two-step tests due to the LR tests' independent nature, something that is not possible when using other tests.

6. CONCLUDING REMARKS

The two major achievements of this study—developing a power analysis for tables with unequal odds ratios or varied interactions, and decomposing the LR test for conditional independence as a two-step test—are based on a log-likelihood decomposition for three-way contingency tables. The key to these findings is the geometry of an invariant Pythagorean law of relative entropy, which is derived from the mutual information identities (3.3) and (3.4) for three-way tables. Extensions of mutual information identities to multiway tables are equivalently straightforward. It follows that testing interaction and testing uniform association are the two independent components of the omnibus test for conditional independence. Application to the special case of $2 \times 2 \times K$ tables establishes both Proposition 1 and Theorem 1, which leads to the development of the two-step tests. This allows for a comparison study with the widely used BD and CMH tests as presented in Section 4.

In practice, the omnibus LR test for conditional independence is first examined, followed by checking the first-step LR test for interaction and the second-step test for uniform association. The three LR tests of Figure 1 are examined together, where consistent or inconsistent significant results between the omnibus test and the two-step test can occur randomly, as indicated by Proposition 2. In contrast, the BD and the CMH tests are computed separately, and their individual results can differ from those of the two-step tests. In addition, because of the nonindependence of the BD and CMH tests, it is difficult to formulate a joint testing framework based on these two tests. On the other hand, each of the two-step LR tests is independent and thus can be combined easily. In addition, the two-step test corresponds to testing parameters in a logit model, as noted in Section 4.2. But when using the adjusted test levels, the two-step test likely will differ from the popular logit models, with or without random effects. In conclusion, we remark that mutual information identities are useful for testing the hypotheses of independence, uniform association, and arbitrary interaction through the definition of two-step LR tests.

APPENDIX A: PROOF OF THEOREM 1

Before presenting the proof, it is useful to remark on the difference between this proof and that (for a goodness-of-fit test in a $2 \times J$ table; $J \geq 3$) of lemma 3 of Cheng et al. (2008). The proof can be directly extended to K (≥ 3) 2×2 tables with different consecutive ratios of odds ratios say, $\gamma_1 \neq \gamma_2 \neq \cdots$, whereas the similar proof for the two-way table of lemma 3 requires that $\gamma_1 = \gamma_2 = \cdots = \psi$ to satisfy the goodness of fit between two rows or uniform association.

For a strata of K 2×2 tables, it takes $K - 1$ estimates of the interactions between the $K - 1$ consecutive pairs of tables. Without loss of generality, it suffices to confine the proof to the case where $K = 2$, and W' can be scaled to have the same size of each 2×2 table of the observed data U. Recall that $U = (U_1; U_2)$, where $U_1 = (a, b; c, d)$ and $U_2 = (e, f; g, h)$ are two 2×2 tables, and $W = (W_1; W_2)$, where the ratio of the two odds ratios of $W_1 = (a^*, b^*; c^*, d^*)$ and $W_2 = (e^*, f^*; g^*, h^*)$ is $\gamma > 0$. The task is to prove (3.9), that there is a unique $V' = (V_1 = (\hat{a}, \hat{b}; \hat{c}, \hat{d}); V_2 = (\hat{e}, \hat{f}; \hat{g}, \hat{h}))$ with the same margins as those of $U = (U_1; U_2)$ and the same ratio γ as the interaction between V_1 and V_2. It suffices to prove that

$$a\log(\hat{a}/a^*) + b\log(\hat{b}/b^*) + c\log(\hat{c}/c^*) + d\log(\hat{d}/d^*) = \hat{a}\log(\hat{a}/a^*) + \hat{b}\log(\hat{b}/b^*) + \hat{c}\log(\hat{c}/c^*) + \hat{d}\log(\hat{d}/d^*), \quad \text{(A.1)}$$

due to the basic equation $a\log(a/a^*) = a[\log(a/\hat{a}) + \log(\hat{a}/a^*)]$. By (A.1), it is easy to check that the proof reduces to verifying the equation

$$\begin{aligned}
&\hat{a}\log\left[\frac{(\hat{a}/a^*)/(\hat{b}/b^*)}{(\hat{c}/c^*)/(\hat{d}/d^*)}\right] + (\hat{a}+\hat{c})\log\left\{\frac{\hat{c}/c^*}{\hat{d}/d^*}\right\} \\
&\quad + (\hat{a}+\hat{b})\log(\hat{b}/b^*) + (\hat{c}+\hat{d})\log(\hat{d}/d^*) \\
&\quad + \hat{e}\log\left[\frac{(\hat{e}/e^*)/(\hat{f}/f^*)}{(\hat{g}/g^*)/(\hat{h}/h^*)}\right] + (\hat{e}+\hat{g})\log\left\{\frac{\hat{g}/g^*}{\hat{h}/h^*}\right\} \\
&\quad + (\hat{e}+\hat{f})\log(\hat{f}/f^*) + (\hat{g}+\hat{h})\log(\hat{h}/h^*) \\
&= a\log\left[\frac{(\hat{a}/a^*)/(\hat{b}/b^*)}{(\hat{c}/c^*)/(\hat{d}/d^*)}\right] + (a+c)\log\left\{\frac{\hat{c}/c^*}{\hat{d}/d^*}\right\} \\
&\quad + (a+b)\log(\hat{b}/b^*) + (c+d)\log(\hat{d}/d^*) \\
&\quad + e\log\left[\frac{(\hat{e}/e^*)/(\hat{f}/f^*)}{(\hat{g}/g^*)/(\hat{h}/h^*)}\right] + (e+g)\log\left\{\frac{\hat{g}/g^*}{\hat{h}/h^*}\right\} \\
&\quad + (e+f)\log(\hat{f}/f^*) + (g+h)\log(\hat{h}/h^*). \qquad \text{(A.2)}
\end{aligned}$$

By assumption, the four large bracketed logarithmic arguments are equal to the common ratio γ between two odds ratios, and these corresponding terms are equal, because $(a + e) = (\hat{a} + \hat{e})$ as U and V have equal margins. The terms associated with the four braces are also equal, because $(a + c) = (\hat{a} + \hat{c})$ and $(e + g) = (\hat{e} + \hat{g})$. Furthermore, the remaining terms are equal as well, because U and V' have the same margins. Equation (A.2) is valid, and the proof of Theorem 1 is complete.

APPENDIX B: PROOF OF PROPOSITION 2

By definition, the same notations of (4.1)–(4.5) apply to the tests D_0, D_1, and D_2, using the test sizes α, α_1, and α_2, respectively. Because $0 < \alpha_1, \alpha_2 < \alpha$, (4.3) is replaced by the inequality $q_{K,\alpha} < q_{K-1,\alpha_1} + q_{1,\alpha_2}$, which is valid as well. The next two equations are straightforward:

$$\begin{aligned}
P(F) &= P(C \cap A^c \cap B^c) \\
&= P(C) - P(C \cap A^c \cap B) \\
&\quad - P(C \cap A \cap B) - P(C \cap A \cap B^c) \qquad \text{(B.1)}
\end{aligned}$$

and

$$\begin{aligned}
P(G) &= P(C^c \cap A) + P(C^c \cap A^c \cap B) \\
&= P(C^c \cap A^c \cap B) + P(C^c \cap A \cap B^c), \qquad \text{(B.2)}
\end{aligned}$$

because the event $C^c \cap A \cap B$ is empty by (4.3). Taking the difference between (B.1) and (B.2), it is seen from (4.7) that

$$\begin{aligned}
P(G) - P(F) &= P(A^c \cap B) - P(C) + P(C \cap A \cap B) + P(A \cap B^c) \\
&\leq (1-\alpha_1)\alpha_2 - \alpha + \alpha_1\alpha_2 + \alpha_1(1-\alpha_2) \\
&= 0. \qquad \text{(B.3)}
\end{aligned}$$

This proves Proposition 2.

[Received January 2009. Revised February 2010.]

REFERENCES

Agresti, A. (2002), *Categorical Data Analysis*, Hoboken, NJ: Wiley. [742,744]

Agresti, A., and Hartzel, J. (2000), "Strategies for Comparing Treatments on a Binary Response With Multi-Centre Data," *Statistics in Medicine*, 19, 1115–1139. [745]

Ardlie, K. C., Lunetta, K. L., and Seielstad, M. (2002), "Testing for Population Subdivision and Association in Four Case-Control Studies," *The American Journal of Human Genetics*, 71, 304–311. [746,747]

Bartlett, M. S. (1935), "Contingency Table Interactions," *Journal of the Royal Statistical Society, Ser. B*, 2, 248–252. [740,741]

Birch, M. W. (1963), "Maximum Likelihood in Three-Way Contingency Tables," *Journal of the Royal Statistical Society, Ser. B*, 25, 220–233. [740]

——— (1964), "The Detection of Partial Association, I: The 2×2 Case," *Journal of the Royal Statistical Society, Ser. B*, 26, 313–324. [740,741]

Bishop, Y. M. M., Fienberg, S. E., and Holland, P. W. (1975), *Discrete Multivariate Analysis*, Cambridge, MA: MIT Press. [740,743]

Breslow, N. E., and Day, N. E. (1980), *Statistical Methods in Cancer Research. Volume* 1—*The Analysis of Case-Control Studies*, Lyon: IARC. [740]

Breslow, N. E., and Liang, K. Y. (1982), "The Variance of the Mantel–Haenszel Estimator," *Biometrics*, 38, 943–952. [744]

Cheng, P. E., Liou, J. W., Liou, M., and Aston, J. A. D. (2006), "Data Information in Contingency Tables: A Fallacy of Hierarchical Log-Linear Models," *Journal of Data Science*, 4, 387–398. [746]

——— (2007), "Linear Information Models: An Introduction," *Journal of Data Science*, 5, 297–313. [742]

Cheng, P. E., Liou, M., Aston, J. A. D., and Tsai, A. C. (2008), "Information Identities and Testing Hypotheses: Power Analysis for Contingency Tables," *Statistica Sinica*, 18, 535–558. [741,743,748]

Cochran, W. G. (1954), "Some Methods for Strengthening the Common Chi-Square Tests," *Biometrics*, 24, 315–327. [740,742]

Cover, T. M., and Thomas, J. A. (1991), *Elements of Information Theory*, Hoboken, NJ: Wiley. [742]

Deming, W. E., and Stephan, F. F. (1940), "On a Least Squares Adjustment of a Sampled Frequency Table When the Expected Marginal Totals Are Known," *The Annals of Mathematical Statistics*, 11, 427–444. [742]

Gart, J. J. (1962), "On the Combination of Relative Risks," *Biometrics*, 18, 601–610. [744]

——— (1970), "Point and Interval Estimation of the Common Odds Ratio in the Combination of 2×2 Tables With Fixed Marginals," *Biometrika*, 57, 471–475. [744]

——— (1971), "The Comparison of Proportions: A Review of Significance Tests, Confidence Intervals and Adjustments for Stratification," *International Statistical Review*, 39, 148–169. [740]

Goodman, L. A. (1964), "Simple Methods for Analyzing Three-Factor Interaction in Contingency Tables," *Journal of the American Statistical Association*, 59, 319–352. [740]

——— (1969), "On Partitioning χ^2 and Detecting Partial Association in Three-Way Contingency Tables," *Journal of the Royal Statistical Society, Ser. B*, 31, 486–498. [740-742]

——— (1970), "The Multivariate Analysis of Qualitative Data: Interactions Among Multiple Classifications," *Journal of the American Statistical Association*, 65, 226–256. [740]

Gray, R. M. (1990), *Entropy and Information Theory*, New York: Springer-Verlag. [742]

Halperin, M., Ware, J. H., Byar, D. P., Mantel, N., Brown, C. C., Koziol, J., Gail, M., and Green, S. B. (1977), "Testing for Interaction in an $I \times J \times K$ Contingency Table," *Biometrika*, 64, 271–275. [740]

Hauck, W. W. (1984), "A Comparitive Study of Conditional Maximum Likelihood Estimation of a Common Odds Ratio," *Biometrics*, 40, 1117–1123. [740,744,747]

Holland, P. W., and Thayer, D. T. (1988), "Differential Item Performance and the Mantel–Haenszel Procedure," in *Test Validity*, eds. H. Wainer and H. I. Braun, Hillsdale, NJ: Erlbaum, pp. 129–145. [740]

Kullback, S., and Leibler, R. A. (1951), "On Information and Sufficiency," *The Annals of Mathematical Statistics*, 22, 79–86. [743]

Landis, J. R., Heyman, E. R., and Koch, G. G. (1978), "Average Partial Association in Three-Way Contingency Tables: A Review and Discussion of Alternative Tests," *Review of International Statistical Institute*, 46, 237–254. [742]

Liu, Q., and Pierce, D. A. (1993), "Heterogeneity in Mantel–Haenszel-Type Models," *Biometrika*, 80, 543–556. [745]

Mantel, N., and Haenszel, W. (1959), "Statistical Aspects of the Analysis of Data From Retrospective Studies of Disease," *Journal of National Cancer Institute*, 22, 719–748. [740,744]

McCullagh, P., and Nelder, J. A. (1989), *Generalized Linear Models* (2nd ed.), London: Chapman & Hall. [740,743]

McKinlay, S. M. (1978), "The Effect of Nonzero Second-Order Interaction on Combined Estimators of the Odds Ratio," *Biometrika*, 65, 191–202. [744]

Mehta, C. R., and Patel, N. R. (1983), "A Network Algorithm for Performing Fisher's Exact Test in $r \times c$ Contingency Tables," *Journal of the American Statistical Association*, 78, 427–434. [744]

Mellenberg, G. J. (1982), "Contingency Table Models for Assessing Item Bias," *Journal of Educational Statistics*, 7, 105–118. [740]

Norton, H. W. (1945), "Calculation of Chi-Square for Complex Contingency Tables," *Journal of the American Statistical Association*, 40, 251–258. [740]

Nurminen, N. (1981), "Asymptotic Efficiency of General Noniterative Estimators of Common Relative Risk," *Biometrika*, 68, 525–530. [740]

Paul, S. R., and Donner, A. (1992), "Small Sample Performance of Tests of Homogeneity of Odds Ratios in K 2×2 Tables," *Statistics in Medicine*, 11, 159–165. [740,747]

Plackett, R. L. (1962), "A Note on Interactions in Contingency Tables," *Journal of the Royal Statistical Society, Ser. B*, 24, 162–166. [740]

Reis, I. M., Hirji, K. F., and Afifi, A. A. (1999), "Exact and Asymptotic Tests for Homogeneity in Several 2×2 Tables," *Statistics in Medicine*, 18, 893–906. [747]

Robins, J., Breslow, N., and Greenland, S. (1986), "Estimators of the Mantel–Haenszel Variance Consistent in Both Sparse Data and Large Strata Limiting Models," *Biometrics*, 42, 311–325. [744,745]

Roy, S. N., and Kastenbaum, M. A. (1956), "On the Hypothesis of no "Interaction" in a Multi-Way Contingency Table," *The Annals of Mathematical Statistics*, 27, 749–757. [740]

Simpson, E. H. (1951), "The Interpretation of Interaction in Contingency Tables," *Journal of the Royal Statistical Society, Ser. B*, 13, 238–241. [740]

Skene, A. M., and Wakefield, J. C. (1990), "Hierarchical Models for Multicentre Binary Response Studies," *Statistics in Medicine*, 9, 919–929. [745]

Somes, G. W. (1986), "The Generalized Mantel–Haenszel Statistic," *The American Statistician*, 40, 106–108. [742]

Swaminathan, H., and Rogers, H. J. (1990), "Detecting Differential Item Functioning Using Logistic Regression Procedures," *Journal of Educational Measurement*, 27, 361–370. [740,744]

Tarone, R. E., Gart, J. J., and Hauck, W. W. (1983), "On the Asymptotic Inefficiency of Certain Noniterative Estimators of a Common Relative Risk or Odds Ratio," *Biometrika*, 70, 519–522. [740,744]

Thomas, D. G. (1975), "Exact and Asymptotic Methods for the Combination of 2×2 Tables," *Computers in Biomedical Research*, 8, 423–446. [744]

Wang, W. C., and Yeh, Y. L. (2003), "Effects of Anchor Item Methods on Differential Item Functioning Detection With the Likelihood Ratio Test," *Applied Psychological Measurement*, 27, 479–498. [740]

Woolf, B. (1955), "On Estimating the Relation Between Blood Group and Disease," *Annals of Human Genetics*, 19, 251–253. [740,744]

Yates, F. (1934), "Contingency Tables Involving Small Numbers and the χ^2 Test," *Journal of the Royal Statistical Society, Suppl.*, 1, 217–235. [741]

Zelen, M. (1971), "The Analysis of Several 2×2 Contingency Tables," *Biometrika*, 58, 129–137. [740]