Likelihood Ratio Tests with Three-Way Tables

Philip E. Cheng, Michelle Liou, and John A. D. Aston

Institute of Statistical Science, Academia Sinica

CRiSM, Department of Statistics, The University of Warwick

January 23, 2009

Abstract

Likelihood ratio (LR) tests for association and for interaction are examined for three-way contingency tables, in particular, 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, the LR tests for homogeneity and for no association. The latter two tests are logically connected to formulating a natural two-step test for conditional independence. It is suggested to use the proposed two-step test with reduced nominal levels in contrast to 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 to 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 paper is to develop tests based on information identities which give independent tests for both the steps allowing them to be naturally combined into a two-step test for conditional independence. These tests will also be contrasted to the tests that are usually used to test each step individually, but which do not necessarily have the required independence to allow their combination into a two-step test.

The literature concerning testing of all three quantities individually is well established. Bartlett (1935) initiated a test for interaction (inhomogeneity of odds ratio) and derived an estimate of the common odds ratio (COR) with a pair of $2 \times 2$ tables. Norton (1945) extended the discussion to finite $K$ tables, and Simpson (1951) supplied interpretations of varied 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 estimate (MLE) given the margins of each stratum in an $I \times J \times K$ table. The classical approaches to testing inhomogeneity were mainly based on weighted chi-square tests, and were further discussed by Plackett (1962) and Goodman (1964). On the other hand, a chi-square test with one degree of freedom (d.f.) for two-way independence across $K$ strata was proposed by Cochran (1954), and by Mantel and Haenszel (1959). This is the celebrated CMH test for conditional independence or partial association (Birch (1964) and Goodman (1969)) that has been widely used in the literature.

These early studies led to further analyses of the three-way tables on 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 MLE (given the margins of each $2 \times 2$ table) of the COR, as noted by Birch (1963). To relax the computation burden, Goodman (1964) discussed approximate chi-square tests, and tests using conditional or unconditional MLE (fixing only one margin of each $2 \times 2$ table) of the COR were discussed by Gart (1971), Zelen (1971) and Halperin, et al. (1977). The Breslow-Day score test for inhomogeneity (1980) gained popularity due to its computational ease using the Mantel-Haenszel COR estimate, although Paul and Donner (1992) found through
a simulation study that score tests for inhomogeneity tend to be conservative when the odds ratios are either small or large.

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 Mantel-Haenszel (1959) COR estimate is defined with intuitive appeal, and is asymptotically efficient when the common odds ratio is equal to one, for example, Birch (1964), Nurminen (1981), and McCullagh and Nelder (1989). For the finite strata case with moderate to large data, Tarone, et al. (1983) and Hauck (1984) showed by simulation studies that a conditional MLE is generally superior to the unconditional MLE or the MH estimate in terms of bias and precision.

While testing interaction and estimating the COR have been much discussed in the literature, testing interaction and testing partial association were seldom studied together. Goodman (1969, 1970) and Bishop, et al. (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 testform items exhibiting differential item functioning (DIF) between two social groups, see for example, Mellenberg (1982), Holland and Thayer (1988), Swaminathan and Rogers (1990), Wang and Yeh (2003). These studies overlooked a fact noted by Goodman (1969) that the loglikelihood 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 characterizes the interaction or the inhomogeneity across strata which is called the non-uniform DIF in psychometrics, and the second summand is used to test homogeneous association within strata which is called the uniform DIF, or the partial association by Birch (1964) and many follow-up studies.

Unlike classical analysis of variance with continuous variables, these three LR tests are logically related, that is, rejection of homogeneity by the first summand implies the rejection of the other two tests. It is thus crucial to analyze how data information shared between the two independent summands affect the interpretation of these LR tests. If the same nominal level is used for all three tests, the omnibus test can be inconsistent with either one of the
other two terms, and becomes less sensitive than the combination of the other two, a naive two-step test, which tests the second term for homogeneous association only if the first test for homogeneity is not rejected. The drawback is simply corrected by using reduced nominal levels, against which the two independent terms are tested. A similar logical relation may exist between the BD and the CMH tests, but it is unclear as to how the results of the BD test can be used to precede the CMH test so as to form a genuine two-step test. A primary goal of this study is to examine the relations between the three LR tests and thus develop a two step test from the pair of two independent components with similar properties to the omnibus conditional independence test. Another goal of this study is to develop power analysis for testing general hypotheses of unequal interactions within the three-way tables. 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 are straightforward byproducts of the tests developed.

The rest of the paper is organised as follows. The background of testing hypotheses with a $2 \times 2 \times K$ table is briefly reviewed in Section 2. Equivalent mutual information identities that divide the data loglikelihood into orthogonal components are discussed in Section 3. Using an information identity, power analysis at hypotheses of unequal interactions is developed from an invariant Pythagorean law of relative entropy (Cheng, et al. (2008)). As a byproduct, efficient interval estimation of the interaction parameter (the ratio between two odds ratios) and of the COR can be easily derived. The relation between the three LR tests is studied in terms of the two-step tests, and illustrated by examples of 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 at alternative tables of unequal odds ratios is also exemplified. Thus, a three-way mutual information identity sets the foundation for both the entire frame of testing hypotheses and the notion of two-step LR tests.
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 $2 \times 2$ tables by a third variable with $K$ levels. Let $(X, Y, Z)$ denote the categorical data vector, and let $(X_k, Y_k)$ denote pairs of dichotomous variables, where $Z$ is the $K$-level ($k = 1, \ldots, K$) stratum variable. In practice, frequency counts are the observed data. Let $n_{ijk}$ represent the number of subjects having condition $i$, ($i = 1$ (case), 2 (control)), exposure $j$ ($j = 1$ (exposed), 2 (non-exposed)), and falling in stratum $k$, $k = 1, \ldots, K$. For each stratum, the $2 \times 2$ table can also be conveniently expressed as a fourfold vector $(n_{11k}, n_{12k}; n_{21k}, n_{22k})$. The dot notation will be used for summation over a subscript, say, $n_{1\cdot k}$ is the number of cases in stratum $k$, and $n_{\cdot 2k}$ is the total number of unexposed subjects in stratum $k$, and so on.

To address the issue of association, testing independence between the case and the risk factor, being conditioned on strata, will be discussed. Let the odds ratios of the $2 \times 2$ tables be defined by $\psi_k = n_{11k}n_{22k} / n_{12k}n_{21k}$, $k = 1, \ldots, K$. The null hypothesis of independence between disease occurrence and risk exposure across strata is conditional independence, denoted by

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

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

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

It approximates the chi-square distribution with $K$ d.f., denoted $\chi^2_K$, which admits a continuity correction for each stratum with small sample size (Yates, 1934).

A widely discussed issue is the hypothesis of equal odds ratio across $K$ strata. Testing inhomogeneity between strata was initially studied as testing the second-order interaction by Bartlett (1935), which is the hypothesis of COR

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

for a positive constant $\psi$. 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 $H_0$. However, this may fail to hold
statistically, especially when a common test size, say $\alpha$, is in use. Assuming COR, denoted by $\psi$ under $H_1$, the Mantel-Haenszel (1959) estimate of $\psi$ is defined to be

$$\psi_{MH} = \frac{\sum_{k=1}^{K} (n_{11k}n_{22k} / n_{-k})}{\sum_{k=1}^{K} (n_{12k}n_{21k} / n_{-k})},$$

(2.4)

A popular test for $H_1$ with simple computation using the estimate $\psi_{MH}$ is the Breslow-Day (BD, 1980) score test defined as

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

(2.5)

Here, the estimates $e_k$ are the solutions to the COR equations, for $k = 1, \ldots, K$,

$$(n_{11k} + e_k)(n_{22k} + e_k) = \psi_{MH}[(n_{12k} - e_k)(n_{21k} - e_k)],$$

and, the asymptotic variance estimate in the denominator of (2.5) is

$$\text{var}(n_{11k}|\psi_{MH})^{-1} = [(n_{11k} + e_k)^{-1} + (n_{22k} + e_k)^{-1} + (n_{12k} - e_k)^{-1} + (n_{21k} - e_k)^{-1}].$$

Under $H_1$, the BD test statistic approximates the chi-square distribution with $K - 1$ d.f.

Also designed for testing $H_0$ is the celebrated Cochran-Mantel-Haenszel test defined by the statistic

$$\chi^2_{CMH} = \frac{\left(\sum_{k=1}^{K} n_{11k} - \sum_{k=1}^{K} n_{1k}n_{-1k} / n_{-k}\right)^2}{\sum_{k=1}^{K} \{n_{1k}n_{2k}n_{1k}n_{2k} / n_{-k}^2(n_{-1k} - 1)\}}.$$

(2.6)

The test statistic (2.6), derived from an estimating equation, essentially assumes equality of the $K$ odds ratios, that is, $\psi_k = \psi$, the COR. In essence, it is used to test the hypothesis $H_2 = (H_0|H_1)$, which defines the same hypothesis $H_0$ implicitly conditioned upon that $H_1$ is not rejected. It follows that rejection of $H_1$ logically implies that of $H_2$. This logical relation is the basis of the two-step tests to be discussed in this study. The test statistic (2.6) has two equivalent versions that approximate the chi-square distribution with 1 d.f. Cochran’s test (1954) is defined with the independent binomial distributions model, while the Mantel-Haenszel (1959) version is derived from Fisher’s exact test hypergeometric distribution using the Yates continuity correction. Extension of (2.6) to $I \times J \times K$ table was discussed by Landis, et al. (1978) and Somes (1986).

Like weighted chi-square tests, LR tests and score tests for $H_1$ using either conditional or unconditional MLEs of the COR have been much studied in the literature, but little
emphasis has been placed on the discussion of testing $H_2$. In particular, the complete relationship between testing the three hypotheses $H_0$, $H_1$ and $H_2$ has not been discussed in the literature, with the exception of Goodman (1969). In the sequel, a systematic study of the LR tests will be carried out, but firstly the nomenclature will be clarified. In the literature, the hypotheses $H_0$ and $H_2$ are usually termed conditional association and partial association, respectively, but the two are often mixed in use. To avoid ambiguity, $H_0$ will be called the hypothesis of conditional independence (across strata); $H_1$, the hypothesis of homogeneity, or no interaction across strata; and $H_2$, the hypothesis of no (homogeneous) association within strata in the rest of this study.

3 Likelihood Ratio Tests for Three-Way Tables

3.1 An Information Identity

Let $(X, Y, Z)$ be the variables of a three-way $I \times J \times K$ contingency table. Let $f(i, j, k)$, $i = 1, \ldots, I$, $j = 1, \ldots, J$, $k = 1, \ldots, K$, denote the joint probability density function (j.p.d.f.) of $(X = i, Y = j, Z = k)$. The well-known Shannon entropy defines a basic information identity

$$H(X) + H(Y) + H(Z) = I(X; Y; Z) + H(X, Y, Z),$$

(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 likewise defined. Here, $I(X; Y; Z)$ denotes the mutual information between the three variables, see Cover and Thomas (1991), and Gray (1990). The mutual information characterizes the minimum divergence from the j.p.d.f. to a hyperplane of products of marginal p.d.f., that is, the projection from the data to the parameter space of the independence hypothesis of the variables (Cheng, et al. (2008), Lemma 1). Furthermore, $I(X; Y; Z)$ admits three equivalent expressions in terms of the obvious likelihood
decompositions, for example,

\[
\log \left\{ \frac{f(i, j, k)}{f(i)g(j)h(k)} \right\} = \log \left\{ \frac{f(i, k)}{f(i)h(k)} \right\} + \log \left\{ \frac{f(i, j)}{f(i)g(j)} \right\} \\
= \log \left\{ \frac{f(i, k)}{f(i)h(k)} \right\} + \log \left\{ \frac{f(j, k)}{g(j)h(k)} \right\} \\
+ \log \left\{ \frac{f(i, j, k)}{f(i)k}f(j)k \right\} ,
\]

(3.2)

where convenient notations \( f(i, j) \) and \( f(i|j) \) are used to denote j.p.d.f. and conditional p.d.f., respectively. Taking expectation of both sides of (3.2), an orthogonal decomposition of the mutual information, using \( Z \) as the conditioning variable, is

\[
I(X; Y; Z) = I(X; Z) + I(Y; Z) + I(X; Y | Z).
\]

(3.3)

Three equivalent forms of (3.2), or (3.3) can be given by choosing each of the variables to be the conditioning variable. Like (3.1), the equivalence between equations (3.2) and (3.3) presents a key important fact that the same identities hold in terms of either the sample frequencies or their expected p.d.f., and this is valid only in the form of likelihood ratios, or relative entropies, but not in the form of plain likelihoods as formulated in models such as hierarchical log-linear models. This is the key fact in modeling categorical data loglikelihood, where parameter estimates such as sample odds ratios observed from data (subject to marginal likelihood constraints on the identities) can be directly tested as model parameters of hypotheses. Thus, it suffices to use a few typical examples of contingency data tables, see for example, Section 4, to test a hypothesis or support a theory, without the need for simulation studies as commonly used in modeling and testing with continuous variables.

The conditional mutual information \( I(X; Y | Z) \) on the right-hand side of (3.3) measures the conditional association between \( X \) and \( Y \), given \( Z \), against which hypothesis \( H_0 \) is defined and tested. A further decomposition leads to a key identity:

\[
I(X; Y | Z) = Int(X; Y; Z) + I(X; Y \parallel Z),
\]

(3.4)

which is also satisfied by its sample version. The sample analogs of the terms of (3.4) approximate chi-square distributions with \((I - 1)(J - 1)K\), \((I - 1)(J - 1)(K - 1)\), and \((I - 1)(J - 1)\) d.f., respectively. The first summand \( Int(X; Y; Z) \) of (3.4) defines the
three-way interaction (the inhomogeneous association) between X and Y, across Z, which is unique by symmetry, irrespective of which conditioning variable is in use. The second term, \( I(X; Y \parallel Z) \), quantifies the homogeneous association between X and Y, within Z, which is an analogous term to the within-block variance of ANOVA with continuous variables.

A main purpose 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 unified inference compared to the BD and the CMH tests of Section 2. Furthermore, it will be shown that an extension of (3.4) will establish testing general hypotheses of inhomogeneous odds ratios, providing a complete power analysis for the testing frame of three way tables.

### 3.2 LR Tests for \( 2 \times 2 \times K \) Tables

To discuss \( K \) strata of \( 2 \times 2 \) tables in the variables (X, Y, Z), some notation will be introduced for ease of exposition. Set \( U_k = (n_{11k}, n_{12k}; n_{21k}, n_{22k}), k = 1, \ldots, K, \) to be the observed stratified \( 2 \times 2 \) tables. Let the conditional MLE under \( H_0 \) be denoted by \( W_k = (n_{11k}^*, n_{12k}^*; n_{21k}^*, n_{22k}^*), k = 1, \ldots, K, \) where \( n_{ijk}^* = n_{i,k}n_{-jk}/n_{-k}, \) for all \( (i, j, k), \) are the MLE of the cell proportions under \( H_0. \) Define the conditional MLE under \( H_1 \) by \( V_k = (\hat{n}_{11k}, \hat{n}_{12k}; \hat{n}_{21k}, \hat{n}_{22k}), k = 1, \ldots, K, \) which can be computed by the method of iterative proportional fitting (IPF, Deming and Stephan (1940)). The IPF finds the combined \( 2 \times 2 \) tables corresponding to the conditional MLE of the COR, denoted \( \hat{\psi}, \) when a unique solution exists with the observed data that is not ill-conditioned, ie not having zero and infinity sample odds ratios.

According to equations (3.3) and (3.4), \( I(X; Y \mid Z) \) characterizes the information attributed to the conditional independence between X and Y given Z, which defines the hypothesis \( H_0. \) It follows that the sample analog of \( I(X; Y \mid Z), \) as a sample relative entropy, yields a basic LR test. Specifically, the sample KL divergence defines the omnibus test for \( H_0, \) denoted by

\[
D_0 = D(U \parallel W) \equiv 2n\hat{I}(X; Y \mid Z) = 2 \sum_{k=1}^{K} \sum_{i=1}^{2} \sum_{j=1}^{2} n_{ijk} \log(n_{ijk} / n_{ijk}^*) \approx \chi^2_K(H_0). \tag{3.5}
\]

The last term explains its approximate distribution \( \chi^2_K \) under \( H_0, \) comparable to the Pearson chi-square test \( \chi^2_{PE} \) of (2.2). It follows from (3.4) and (3.5) that \( D_0 \) equals the sum of two
independent components. The first summand of (3.4) defines the LR test for $H_1$ as the sample KL divergence

$$D_1 = D(U \parallel V) = 2 \sum_{k=1}^K \sum_j \sum_i n_{ijk} \log(n_{ijk} / \hat{n}_{ijk}) \approx \chi^2_{K-1}(H_1).$$

(3.6)

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

$$D_2 = D(V \parallel W) = 2 \sum_{k=1}^K \sum_j \sum_i \hat{n}_{ijk} \log(\hat{n}_{ijk} / n_{ijk}^*) \approx \chi^2_{1}(H_0 | H_1),$$

(3.7)

which is comparable to the CMH test (2.6) and approximates $\chi^2_{1}$ under $H_2$. By (3.7), $D_2$ tests for $H_2 = (H_0 | H_1)$ conditioned upon that $H_1$ is not rejected, otherwise, it is ineffective because logically the rejection of $H_1$ implies that of $H_2$. This fact has been essentially overlooked in using the CMH test as mentioned in the Introduction. 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 $\chi^2_{K}(H_0) \approx \chi^2_{K-1}(H_1) + \chi^2_{1}(H_2)$ as stated below.

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

$$D_0 = D_1 + D_2.$$  

(3.8)

Thus an alternative hypothesis of inhomogeneous odds ratios has been proposed to that of the BD and the CMH tests, with a clear independence relationship between the two LR tests. As a main aim of this study, power analysis for the three LR tests will be developed under general hypotheses including the null hypotheses $H_0$, $H_1$ and $H_2$ using this independence structure. This will be achieved by proving an invariant Pythagorean law of relative entropy which extends testing two-way independence (Cheng, et al. (2008)) to three-way tables.
3.3 Power Analysis of the LR Tests

From testing $H_0$ or $H_1$, it is useful to examine a class of alternative hypotheses $H'$ having unequal odds ratios between the $2 \times 2$ tables. For example, this can be used to estimate the sample size needed to meet a specificity and sensitivity criteria for certain patterns of unequal odds ratios of interest. For a strata of $K$ $2 \times 2$ 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 $K = 2$ suffices to clarify the theory. Additional notations will be used. 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 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 $V' = (V_1 = (a\hat{\cdot}, b\hat{\cdot}; c\hat{\cdot}, d\hat{\cdot}); V_2 = (e\hat{\cdot}, f\hat{\cdot}; g\hat{\cdot}, h\hat{\cdot}))$ would have the same margins as those of $U = (U_1; U_2)$, and the same ratio $\gamma$ as the interaction between $V_1$ and $V_2$ ($V_i$ need not have $\psi_i$ as its odds ratio). An analog of (3.8) holds. This is summarized in Theorem 1 and proved in Appendix A. Figure 1 presents the geometry to illustrate the hypotheses testing for $H_0$ and $H'$ using the Pythagorean law (3.8), and the extended Pythagorean law as stated below.

**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 $U$, odds ratios $(\psi_1, \ldots, \psi_K)$, and consecutive three-way 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

$$D(U \parallel W') = D(U \parallel V') + D(V' \parallel W').$$  \hspace{1cm} (3.9)

It is easily checked that the proof of Theorem 1 in Appendix A also concludes the proof for (3.8) as a special case. The sample analog of identity (3.9) generalizes those of (3.8), from testing null hypotheses $H_0$ and $H_1$, $\gamma_i = 1$, to testing $H'$ and $H'_1$, $\gamma_i \neq 1$; and, the same
Figure 1: Null Hypotheses: $D(U||W) = 0 = D(U||V) + D(V||W)$, $\gamma_i = 1$;
Alternative Hypotheses: $D(U||W') = 0 = D(U||V') + D(V'||W')$, $\gamma_i \neq 1$. 


asymptotic chi-square distributions for the sample analogs of (3.8) are also valid for those of (3.9). Thus, both \( p \) values at \( H_0 \) and \( H_1 \), and power values at alternative hypotheses \( H' \) and \( H'_1 \) can be so calculated. Figure 1 shows this similarity between equations (3.8) and (3.9) by parallel hyperplanes associated with the hypotheses \( H_1 \) and \( H'_1 \). An application of this power evaluation will be given in Example 4 of Section 5. Such power evaluations at \( 2 \times 2 \times K \) tables with unequal interactions have not been discussed in the literature.

For the case \( K = 2 \), it can be shown that (3.9) provides interval estimation of the interaction parameter \( \gamma (= \gamma_1) \) between the two-way tables. It is seen that \( D_1 \) of (3.8) tests for homogeneity \( \gamma = 1 \), and \( D(U \parallel V') \) of (3.9) tests for any specific value of \( \gamma \neq 1 \). Indeed, \( D(U \parallel V') \) can be used to provide an interval estimation of the parameter \( \gamma \) of a pair of \( 2 \times 2 \) tables, extending the same for the odds ratio \( \psi \) of a single \( 2 \times 2 \) table. For economy of exposition, the proof of Corollary 2 below is omitted but follows in a straightforward manner.

**Corollary 2** For \( K = 2 \), \( D(U \parallel V') \) tests for specific values of the interaction parameter \( \gamma (\neq 1) \), and provides an interval estimation of the parameter \( \gamma \) given the observed data \( U \).

## 4 Two-Step Tests for \( H_1 \) and \( H_2 \)

### 4.1 Interval estimation of the common odds ratio

Confidence interval (C.I.) estimation of the COR \( \psi \) was initially studied by Woolf (1955), Mantel and Haenszel (1959), and Gart (1962). Thereafter, the asymptotic variance of the MH estimate received most discussions, for example, Thomas (1975), McKinlay (1978), Breslow and Liang (1982), Tarone et al. (1983), and Robins et al. (1986). On the other hand, the conditional MLE \( \hat{\psi} \) (fixing both margins of each \( 2 \times 2 \) table) of the COR differs slightly in magnitude from either the MH estimate (2.4) or the conditional MLE using extended hypergeometric distributions (Gart, 1970); and, it also differs from the unconditional MLE (fixing one margin of each \( 2 \times 2 \) table). It was noted that \( \hat{\psi} \) may be preferred to the unconditional ones in terms of both bias and efficiency (Hauck, 1984). In contrast to these studies, a C.I. estimation of \( \psi \) is proposed here using the information identity (3.8).

Equation (3.8) specifies a two-step test. It first tests \( H_1 \) by \( D_1 \), and, if \( H_1 \) is rejected, so
is $H_2$, stop and conclude the test; otherwise, proceed with testing $H_2$ using $D_2$ as the second step. A byproduct of the two LR tests is a simple interval estimation for $\psi$. If $H_1$ is rejected when $D_1$ is significant, then no C.I. for $\psi$ is expected. Otherwise, $H_1$ is not rejected, and a standard C.I. can be obtained by inverting the sampling distribution of the second-step test $D_2$ which approximates the $\chi^2$ distribution. To illustrate this case, Proposition 1 asserts that the first-step test $D_1$ computes the MLE $V$, which is the $2 \times 2 \times K$ table having the COR $\hat{\psi}$, a base value within a C.I. for $\psi$. Next, apply the second-step test $D_2$ to yield the two-sided C.I. $(W_l, W_u)$ centered at $V$ using the approximate distribution $\chi^2_1$; and, compute the pair of COR’s of the pair $(W_l, W_u)$ of $2 \times 2 \times K$ tables to yield $(\psi_l, \psi_u)$, which is the desired two-sided C.I. that includes $\hat{\psi}$. By standard LR test theory, this approach yields asymptotically efficient C.I. estimation for $\psi$, without computing an asymptotic variance estimate for $\hat{\psi}$, although it can be computed if desired. Such a proposed C.I. will be computed in Example 4 of Section 5.

In theory, the CMH test is formulated as a score test without using the COR $\psi_{MH}$, and it can not yield a C.I. by inverting the approximate chi-square distribution. A popular C.I. for $\psi$ derives from a logarithmic transformation using $\psi_{MH}$ and an asymptotic variance estimate (e.g., Robins et al., 1986). This COR estimation is shown to be asymptotically efficient under $H_0$, without using a first-step test for $H_1$. It is crucial that these two estimation methods, that often yield comparable intervals for the COR, are conceptually different in inference, and thus, the LR tests need to be further discussed.

### 4.2 A Naive Two-Step Test

The above illustration together with the right triangle (Figure 1) shows that the hypotenuse defines the omnibus test $D_0$ for $H_0$, and the two sides of the right triangle form a naive 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. Since $H_0 \subset H_1$, and $H_2 = (H_0|H_1)$, both $D_1$ and $D_2$ are regarded effective for testing $H_0$. As in analysis of variance with a continuous variable, the test for $H_0$ is concluded insignificant once the omnibus test $D_0$ is insignificantly small. The tests $D_1$ and $D_2$ need to be examined only when $D_0$ is significant. If, in the latter case, the same test size $\alpha$ is used for both $D_1$ and $D_2$ as usual, then the effective size for testing $H_0$ would be about $2\alpha$, twice
that of \(D_0\). Such an increase in test size particularly occurs when testing odds ratios with categorical variables, but not when testing equality of means in the analysis of variance with continuous variables. For a preliminary discussion of a two-step test, the same nominal level \(\alpha\) is tentatively used for each of the three LR tests.

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

\[
C = C_K = \{D_0 > q_{K,\alpha}\} \quad (4.1)
\]

be a level-\(\alpha\) critical region under \(H_0\), where \(q_{K,\alpha}\) denotes the 100\((1 - \alpha)\) percentile of the \(\chi^2_K\) 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 \(\chi^2_{K-1}\) and \(\chi^2_1\) distributions, respectively. Then, for fixed \(K\), the critical region of the naive two-step test is defined to be

\[
E = E_K = A \cup (A^c \cap B), \quad (4.2)
\]

where the superscript \(c\) denotes the complement of an event. Under \(H_1\) and \(H_2\), event \(E\) is governed by the chance \(P_{H_1}(A) = \alpha\) of rejecting \(H_1\) (hence \(H_2\)) by \(D_1\), such that \(P_{H_1}(A^c) = 1 - \alpha\) is the chance 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 \(2\alpha\). For \(K \geq 2\) and \(0 < \alpha \leq 0.10\), the following inequality holds from the properties of the chi-square distribution,

\[
q_{K,\alpha} < q_{K-1,\alpha} + q_{1,\alpha}. \quad (4.3)
\]

Thus, by (4.1) and (4.2), define for fixed \(K\), \(F = F_K = C \cap A^c \cap B^c\) to be 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

\[
P(F) \leq P(q_{K,\alpha} < D_0 \leq q_{K-1,\alpha} + q_{1,\alpha})
\leq P(q_{K,\alpha} < D_0) - P(q_{K-1,\alpha} \leq D_1 \text{ and } q_{1,\alpha} \leq D_2)
= \alpha - \alpha^2. \quad (4.4)
\]
On the other hand,

\[ P(G) = P\{(A \cap C^c) \cup (A^c \cap B \cap C^c)\} \]
\[ = P(C^c) - P(C^c \cap A^c \cap B^c) \]
\[ \geq P(C^c) - P(A^c)P(B^c) \]
\[ = \alpha - \alpha^2. \quad (4.5) \]

Thus, (4.4) and (4.5) conclude that

\[ P(F) \leq P(G). \quad (4.6) \]

It is noted in Section 4.1 that the COR interval estimation is measured within the event \( A^c \cap B \) of (4.2), given that testing for \( H_1 \) by \( D_1 \) is insignificant at level \( \alpha \). However, the test levels of \( D_1 \) and \( D_2 \) need not be equal to \( \alpha \), and this case will be examined in the sequel.

### 4.3 The Two-Step Test

As illustrated above, a natural competitor to the omnibus test \( D_0 \) is a two-step test that combines together 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 that of \( H_0 \), and also of \( H_2 \). Suppose two distinct test levels \( \alpha_1 \) and \( \alpha_2 \), are separately used for \( D_1 \) and \( D_2 \), 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 keep the same test level \( \alpha \) as in (4.1), the following equation

\[ P(E) = P_{H_1}(A) + \{1 - P_{H_1}(A)\}P_{H_2}(B) \]
\[ = \alpha_1 + (1 - \alpha_1)\alpha_2 = \alpha \quad (4.7) \]

must be satisfied, where \( P_{H_0|H_1}(A^c \cap B) = P_{H_2}(B) \) since \( D_1 \) 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. In case a common level \( \alpha_1 = \alpha' = \alpha_2 \) is used, 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 \( \alpha = 0.050 \). Since each test is more conservative with \( \alpha_i < \alpha \), or \( \alpha' \simeq \alpha/2 \), it is unlikely that they would yield as high a sensitivity as the omnibus test.
Indeed, the converse of inequality (4.6) holds, stated as (4.8) in the next Proposition and proved in Appendix B.

Proposition 2 Define the same events $C$, $E$, $F$ and $G$ according to (4.1) to (4.5), except that the test sizes of $D_1$ and $D_2$ are now replaced by $\alpha_1$ and $\alpha_2$, respectively. Then, the converse of (4.6) holds for the two step test,

$$P(F) \geq P(G). \quad (4.8)$$

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 as an empty set, in view of inequality (4.3). This fact and (4.3) continue to be true when the levels $\alpha_1$ and $\alpha_2$, smaller than $\alpha$, are used in a two-step test. To illustrate the seven possibly nonempty events, it can be checked that the data given in Cheng, et al (2006, Table 1, plus evaluations of (3.2) and (3.3)) give a case of the event $A \cap B \cap C$ that has all three tests significant. It is also easy to exemplify a case having all three tests insignificantly small in magnitude, that is, $A^c \cap B^c \cap C^c$ occurs when each table’s cell counts are in essence homogeneous. Examples 1 to 5 below will present empirical data that characterizes the remaining five cases, and illustrates the performance of the two-step LR tests compared to those of the BD and the CMH tests in these settings.

Example 1 Consider a $2 \times 2 \times 3$ table, where the odds ratios of the three $2 \times 2$ tables are spread about unity as follows. Table (1) is denoted by the fourfold vector $(9, 19; 18, 18)$, $\psi_1 = 0.47$; table (2) is $(6, 15; 10, 9)$, $\psi_2 = 0.36$; and table (3) is $(22, 8; 12, 11)$, $\psi_3 = 2.52$.

Set $\alpha = 0.05$. 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. However, $\psi_{MH} = 0.77$, and $\chi^2_{BD} = 6.41$, $p = 0.04$, which suggests rejecting $H_1$, hence $H_0$, since $H_0 \subset H_1$. On the other hand, the LR test computes the hypotenuse to be $D_0 = 7.04$, $p = 0.07$; $D_1 = 6.39$, $p = 0.04$; and $D_2 = 0.66$, $p = 0.42$. Thus the event $A \cap B^c \cap C^c$ occurs. Like Pearson’s chi-square test, the LR test for $H_0$ is insignificant at the nominal level $\alpha$. Although the test
$D_1$ for $H_1$ is significant at level $\alpha$, it is not at a lower level $\approx \alpha/2$ chosen to accord to an equal level ($\alpha_1 = \alpha_2$) two-step test, from which the test $D_2$ for no uniform association is also insignificant. The results of the LR tests are therefore consistent, while the classical tests at the level $\alpha$ yield inconsistent results in this case.

**Example 2** Consider another $2 \times 2 \times 3$ table, where table (1) equals to (1, 3; 4, 39), $\psi_1 = 3.25$; table (2) is (80, 9; 8, 3), $\psi_2 = 3.33$; and, table (3) is (3, 9; 8, 71), $\psi_3 = 2.96$.

Again set $\alpha = 0.05$. 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 $\chi^2_{CMH} = 5.74$, $p = 0.017$; and, 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 usual CMH test (without continuity correction) shows a significant result but the LR test yields an insignificant two-step test at level $\alpha/2$, even though $D_2$ is significant at level $\alpha$. The MH test for $H_2$ in this case is comparable to the second-step test $D_2$ but again would be rejected at the level $\alpha$.

Examples 1 and 2 have illustrated that excessive sensitivity of either the BD or the CMH test, or that of either one of the two independent LR tests, may be relieved using the adjusted test sizes of a two-step test. Meanwhile, a special case may arise from the geometry of (3.8); it is likely to have a significant $D_0$ at level $\alpha$, but insignificant $D_1$ and $D_2$.

**Example 3** Consider a $2 \times 2 \times 2$ table, where table (1) is the four-tuple $(5, 1; 1, 6)$, $\psi_1 = 30$, and, table (2) is $(3, 4; 2, 3)$, $\psi_2 = 1.125$.

For this data, $D_0 = 6.81$, $p = 0.033$, and $D_1 = 3.14$ with $p = 0.076$, 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 to 3 above have shown a subset of the comparison conditions between the omnibus test $D_0$ and the two-step test, $D_1$ followed by $D_2$. To complement the illustration
of the remaining conditions, two real data examples from the literature are discussed in this section. In addition to illustrating the two-step LR tests against the BD and the CMH tests, a basic use of the power analysis of Theorem 1 will be shown in Example 4. For the next two examples, analyses of the previous authors will be reported and compared against those obtained from the LR tests. The level $\alpha = 0.05$ will be used unless otherwise stated.

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

**Example 4** Data of two $2 \times 2$ 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 by the authors. The data from their Table 2 are briefly illustrated: the row factors were case and control, and the column factors were $C$ and $G$ allele types; the row-by-column frequency tables were $U = (U_1, U_2)$ where $U_1 = (n_{11} = 62, n_{12} = 419; n_{21} = 92, n_{22} = 371)$ with the total $n_\cdot = 944$ for Poland, and $U_2 = (48, 447; 51, 445)$ with $n_\cdot = 991$ for U.S., respectively.

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% confidence interval (0.60, 0.87) which excludes 0.937, and barely includes the other, 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 “the two odds ratios are different”.

By Figure 1, the hypotenuse test is $D_0 = 8.55$ with $p = 0.014$, $K = 2$ d.f. Then, $D_1 = 2.646$ with $p = 0.104$, and the conditional MLE $\hat{\psi} = 0.718$; 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 between Poland and U.S. because their distribution patterns are alike. Since the omnibus test $D_0$ is significant, and the test for $H_1$ is not rejected, $H_2$ 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 given the insignificant BD test, as the event $A^c \cap B \cap C$ occurs. The conclusion therefore follows that there is evidence that the odds ratios differ from one, but
It follows from Section 4.1 that a 95% C.I., (0.549, 0.938) of the COR $\psi$ is easily computed from inverting the sampling distribution of the statistic $D_2$, based on $\hat{\psi}$. This result is closely comparable to the popular C.I. (0.551, 0.941) obtained from estimating the standard error of a transformed MH estimate, as illustrated in Section 4.1.

Suppose there was a doubt of 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, $W' = (W'_1, W'_2)$, where $W'_1 = (80, 420; 110, 400)$ with odds ratio $\psi_1 = 0.693$, and $W'_2 = (51, 445; 51, 445)$, with $\psi_2 = 1.00$, such that $\gamma' = \psi_1/\psi_2 = 0.693$. That is, were the table of unequal odds ratios $W' \in H'$ a valid hypothesis, would it be supported by the observed data $U$? Such a question has not been addressed in the literature. It is seen that $D(U \parallel W') = 5.450$, with $p = 0.066$ against the $\chi^2_2$ distribution; the pair of odds ratios of $W'$ are not far from those of $U$ such that the hypotenuse $D(U \parallel W')$ is insignificant. Meanwhile, there exists by equation (3.9) of Theorem 1 a table $V' \in H'_{1}$ having the same margins as data $U$, and the same interaction 0.693 of $W'$, such that $D(U \parallel V') = 0.092$ which is insignificantly small. Moreover, by (3.9) of Theorem 1, $D(V' \parallel W') = 5.358$, with $p = 0.020$, which is significant at level 0.025 against $\chi^2_1$. This illustrates that testing $W'$ for a pair of unequal odds ratios can be seemingly insignificant by the omnibus test but not by the two-step test. When a non-unity interaction parameter $\gamma' = \psi_1/\psi_2 = 0.693$ of $W' \in H'_{1}$ is involved, the second-step test $D(V' \parallel W')$ for $W' \in H'$ may be significant at level 0.025, given that the first-step test $D(U \parallel V')$ is insignificant. This example yields an event $A^c \cap B \cap C^c$, and the pair of unequal odds ratios of $W'$ is not supported by the data $U$, according to Theorem 1 and a two-step LR test. Thus, testing unequal odds ratios together with the interaction parameter may not be a simple task.

The next example refers to a data from studying the effect of progestogens versus placebo on miscarriage, stillbirth or neonatal death, which cites a previous study about “hormone administration for maintenance of pregnancy” by Reis, et al. (1999, Table III).

**Example 5** Pregnancy data from a meta-analysis of clinical trials, cited by Reis et al. (1999), listed seven $2 \times 2$ tables of the two dichotomous factors: the case-control and the treatment conditions, progestogens and placebo. The tables are listed as fourfold vectors:
(n_{11} = 4, n_{12} = 74; n_{21} = 2, n_{22} = 74) with odds ratio (\psi = 1.99) from Study-1; (3, 77; 3, 85) with \psi = 1.10 from S-2; (8, 31; 3, 37) with \psi = 3.14 from S-3; (4, 11; 7, 7) with \psi = 0.36 from S-4; (0, 18; 7, 18) with \psi = 0.00 from S-5; (1, 96; 1, 97) with \psi = 1.01 from S-6; and (12, 48; 13, 40) with \psi = 0.77 from S-7.

For this data set, the authors computed several exact and asymptotic tests for \( H_1 \). The BD test yielded \( \chi^2_{BD} = 11.26 \), with \( p = 0.081 \), \( K - 1 = 6 \) d.f., and all other tests yielded similar results, having \( p \) values in the range \( [0.055, 0.09] \); except that \( p = 0.032 \) was given by the unconditional LR test, which was the only test that does 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 d.f.; which is marginally significant. The LR test for \( H_1 \) yields \( D_1 = 13.77, p = 0.032 \), with 6 d.f., which agrees with the omnibus test and the unconditional LR test, but not the BD test, at the same level \( \alpha = 0.05 \). As mentioned in the Introduction, this presents a case when the BD test may be more conservative than the LR test \( D_1 \), as noted by Hauck (1984), Paul and Donner (1992). Meanwhile, \( \chi^2_{CMH} = 0.353, p = 0.553 \), which is comparable to the insignificant LR test \( D_2 = 0.358, p = 0.550 \). This is a case that the event \( A \cap B^c \cap C \) occurs. The COR estimate \( \psi_{MH} = 0.853 \), or \( \hat{\psi} = 0.850 \), may not be meaningful when \( H_1 \) is rejected at a level, say 0.04. This happens when a two-step test of Section 4.3 is used with \( \alpha_1 (\ge 0.04) \). Thus, this example illustrates that the BD and the CMH tests may not form a useful pair of two-step tests as they may not provide as equally precise and sensitive inference (at a known level) as is the case in the two-step LR test.

6 Concluding Remarks

The two major goals of this study of establishing a two-step test for conditional independence and of giving power analysis for the tests are accomplished using the loglikelihood decomposition of three-way contingency tables. Testing hypotheses of association and of interaction in a \( 2 \times 2 \times K \) table is fully discussed using the LR tests, together with the popular BD and the CMH tests. Theorem 1 develops power analysis of general LR tests for interaction, which directly evaluates the probability of the observed data under any specified odds ratios.
and interactions between the $2 \times 2$ tables of any three-way tables. This extends the invariant Pythagorean law of relative entropy from two-way tables to three-way tables, as illustrated by the data in Example 4.

By using mutual information identities, the notion of two-step LR tests specifically defined for $2 \times 2 \times K$ tables is introduced. A natural scheme for testing is to use the omnibus LR test for conditional independence, followed by checking the two-step LR test for homogeneity and for no association. Crucial findings are that LR test procedures present efficient and proper analyses in contrast to those obtained by the classical BD and CMH tests. Section 4 examines the logically connected two-step orthogonal component LR test. In particular, this natural scheme is in principle different from the classical scheme that first tests for interaction, which is common in methods such as in hierarchical log-linear modeling. Like the second-step LR test for no homogeneous association, the CMH or MH test could be used given that the first-step LR test for interaction is not rejected; otherwise, it is difficult to interpret its result regardless of its significance. However, care must be taken to ensure the correct interpretation of the latter two tests in light of their structure in relation to the initial test for interaction. This is in contrast to the second-step LR test which has been shown to be independent of the first step. To summarize, mutual information is useful both for evaluating the probability of arbitrary patterns of interaction, and for developing the notion of two-step LR tests with three-way tables.

References


**Appendix A: Proof of Theorem 1**

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 $K = 2$, and that $W'$ can be scaled to have the same size of each 2 × 2 table of the observed data $U$. Recall the notations: $U = (U_1; U_2)$, 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 between 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 $V' = (V_1 = (\hat{a}, \hat{b}; \hat{c}, \hat{d}); V_2 = (\hat{e}, \hat{f}; \hat{g}, \hat{h}))$, having the same margins as those of $U = (U_1; U_2)$, and the same $\gamma$ as the interaction between $V_1$ and $V_2$. It suffices to prove that

$$a \log(\hat{a}/a^*) + b \log(\hat{b}/b^*) + c \log(\hat{c}/c^*) + d \log(\hat{d}/d^*)$$

$$= a \log(\hat{a}/a^*) + \hat{b} \log(\hat{b}/b^*) + \hat{c} \log(\hat{c}/c^*) + \hat{d} \log(\hat{d}/d^*),$$  \hspace{1cm} (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 following equation

$$\hat{a} \log \left[ \frac{\hat{a}/a^*}{\hat{c}/c^*} \right] + (\hat{a} + \hat{c}) \log \left\{ \frac{\hat{c}/c^*}{\hat{d}/d^*} \right\} + (\hat{a} + \hat{b}) \log(\hat{b}/b^*) + (\hat{c} + \hat{d}) \log(\hat{d}/d^*)$$

$$+ \hat{e} \log \left[ \frac{\hat{e}/e^*}{\hat{f}/f^*} \right] + (\hat{e} + \hat{g}) \log \left\{ \frac{\hat{g}/g^*}{\hat{h}/h^*} \right\} + (\hat{e} + \hat{f}) \log(\hat{f}/f^*) + (\hat{g} + \hat{h}) \log(\hat{h}/h^*)$$

$$= a \log \left[ \frac{\hat{a}/a^*}{\hat{b}/b^*} \right] + (a + c) \log \left\{ \frac{\hat{c}/c^*}{\hat{d}/d^*} \right\} + (a + b) \log(\hat{b}/b^*) + (c + d) \log(\hat{d}/d^*)$$

$$+ e \log \left[ \frac{\hat{e}/e^*}{\hat{f}/f^*} \right] + (e + g) \log \left\{ \frac{\hat{g}/g^*}{\hat{h}/h^*} \right\} + (e + f) \log(\hat{f}/f^*) + (g + h) \log(\hat{h}/h^*).$$  \hspace{1cm} (A.2)

By assumption, the four large bracketed logarithmic arguments are equal to the common ratio $\gamma$ 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 since \((a + c) = (\hat{a} + \hat{c})\) and \((e + g) = (\hat{e} + \hat{g})\). Furthermore, the remaining terms are also equal due to that \(U\) 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) through (4.5) apply to the tests \(D_0, D_1\) and \(D_2\), using the test sizes \(\alpha, \alpha_1\) and \(\alpha_2\), respectively. Since \(0 < \alpha_1, \alpha_2 < \alpha\), (4.3) is replaced by the inequality \(q_{K,\alpha} < q_{K-1,\alpha_1} + q_{1,\alpha_2}\), which is also valid. The next two equations are straightforward:

\[
P(F) = P(C \cap A^c \cap B^c)
= P(C) - P(C \cap A^c \cap B) - P(C \cap A \cap B) - P(C \cap A \cap B^c),
\]

(B.1)

and

\[
P(G) = P(C^c \cap A) + P(C^c \cap A^c \cap B)
= P(C^c \cap A \cap B) + P(C^c \cap A \cap B^c),
\]

(B.2)

since 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

\[
P(G) - P(F) = P(A^c \cap B) - P(C) + P(C \cap A \cap B) + P(A \cap B^c)
\leq (1 - \alpha_1)\alpha_2 - \alpha + \alpha_1\alpha_2 + \alpha_1(1 - \alpha_2)
= 0.
\]

This proves Proposition 2.

Correspondence:

John A. D. Aston

Centre for Research in Statistical Methodology
Department of Statistics
The University of Warwick
Coventry, UK.
Email: j.a.d.aston@warwick.ac.uk