Information Identities and Testing Hypotheses: Power Analysis for Contingency Tables

Philip E. Cheng[1], Michelle Liou[1], John A. D. Aston[1,2] and Arthur C. Tsai[1]

[1] *Academia Sinica and* [2] *University of Warwick*

Statistica Sinica **18** (2008), 535-558

Abstract

An information theoretic approach to the evaluation of 2×2 contingency tables is proposed. By investigating the relationship between the Kullback-Leibler divergence and the maximum likelihood estimator, information identities are established for testing hypotheses, in particular, testing independence. These identities not only validate the calibration of p values, but also yield unified power analysis for the likelihood ratio test, Fisher's exact test and Pearson-Yates' chi-square test. It is shown that a widely discussed exact unconditional test for the equality of binomial parameters is ill-posed for testing independence, and that using this test to criticize Fisher's exact test as being conservative is logically flawed.

Key Words: Chi-square Test; Contingency Table; Exact Test; Kullback-Leibler Divergence; Likelihood Ratio Test; Mutual Information.

Correspondence: John Aston
Institute of Statistical Science
Academia Sinica
128 Academia Road, Sec 2
Taipei, 115
Taiwan ROC
Tel: +886-2-2783-5611 x 314
Email: jaston@stat.sinica.edu.tw

1. Introduction

Evaluation of association and independence between two categorical factors is a classic topic of interest in statistical inference. Pearson's celebrated goodness of fit test yielded the chi-square test in the analysis of a 2×2 contingency table (Pearson, 1900; 1904). Yule (1911) introduced a test for association through testing the equality of two independent binomial proportions of one dichotomous factor. Fisher (1934) characterized the combinatorial randomization of two-factor association using the extended hypergeometric distribution, which gave rise to his exact test.

By the 1930s the philosophy of hypothesis testing had been well established by Fisher (1925, 1935), and Neyman and Pearson (1928), among others. It also initiated the long debate concerning the two approaches: significance testing for Fisher and hypothesis testing for Neyman and Pearson. Testing independence for a 2×2 table was a notable example in these arguments. While the debate was focused on the notions of inductive inference, significance level, and decision theory for testing hypotheses, the importance of power evaluation was generally accepted (e.g., Fisher, 1946) with the adoption of the idea of identifying appropriate critical regions for constructing more sensitive tests. For example, in testing the equality of two binomial parameters by Yule's test, the *p* values and the power at alternatives can be computed from either the normal approximation or the exact distribution. However, unified power analysis has not been fully developed for Pearson's chi-square or Fisher's exact test for assessing independence in a 2×2 table. This will be investigated here.

Meanwhile, a controversial issue arises when using the exact test, due to its discrete nature. With the limited sample space defined by fixed row and column margins, it yields a conservative test when the sample size is not large. Barnard (1945, 1949) discussed this issue using the Convexity-Symmetry-Maximum (CSM) triple-condition test based on the sample space of the two independent binomials model. This led to studying the so-called unconditional test where only one margin of the 2×2 table is fixed. Another classic unconditional test proposed in the 1950s is essentially a mixture of the exact conditional tests (Bennet and Hsu, 1960). The test aims at finding a more powerful critical region subject to a nominal significance level. However, the advantage over Fisher's exact test can only be achieved by considering biased or raised levels for the conditional tests which are implemented in constructing the unconditional test (Boschloo, 1970).

The criticism of conservativeness of Fisher's exact test reached a climax when Berkson (1978) dispraised Fisher's exact test using arguments based on Yule's test for the equality of two independent binomial proportions. Since then, Yule's test has been the most widely discussed *exact unconditional test* in the literature. Yates (1984) gave supporting arguments for Fisher's exact test, noting that "*tests for independence in a* 2×2 *table must be conditioned on both margins*". Most discussants on Yates' paper agreed with his assertion. However, this remains a debated issue in the literature, primarily due to the lack of unified power analysis for both Pearson's chi-square test and Fisher's exact test. Indeed, a thorough comparison between conditional and unconditional tests has not been undertaken in the literature, and will be considered here.

The paper proceeds as follows. Tests for independence in a 2×2 contingency table will be defined in Section 2. This is followed by a calibration of the *p* values between the chi-square, the exact and the likelihood ratio tests over the common sample space of their null distributions. The calibration is derived together with a fundamental likelihood identity, defined using "mutual information", which yields proper representations of the *p* values based on the conditional distributions. In Section 3, the likelihood identity is generalized to yield an invariance property of information decompositions, which is used to develop the power analysis at alternative hypotheses where the odds ratios differ from one. This leads to the identification of the logical flaw in comparing Yule's test with Fisher's test for independence in a 2×2 table. Applications of the information identity to two-way tables for testing model-data fit for general association models are illustrated in Section 4. In conclusion, we note that Fisher's "most relevant set" (Fisher, 1935; Bartlett, 1984) is characterized, where a unified power analysis of Pearson's chi-square test and Fisher's exact test is validated.

2. Testing Independence in a 2×2 Contingency Table

In the analysis of categorical data, a fundamental problem is to decide whether an attribute *A* (or not *A*) is randomly allocated to two mutually exclusive subpopulations defined by another dichotomous factor. The statistical question is to test whether independence, or no association, holds between the two dichotomous factors. In certain designs of experiments, a random sample is often selected from the entire population to assess the odds of having the attribute A in the two subpopulations (e.g., Lehmann, 1986, Section 4.7). The observed data (with sample size *N*) are frequency counts, which are expressed as a 2×2 contingency table:

$$\begin{array}{ccccc} & A & \bar{A} & \text{Total} \\ \text{Group 1} & x_{11} & x_{12} & x_{1\cdot} \\ \text{Group 2} & x_{21} & x_{22} & x_{2\cdot} \\ \text{Total} & x_{\cdot 1} & x_{\cdot 2} & N \end{array} \tag{2.1}$$

A general probability structure of the 2×2 table of (2.1) is the multinomial model, which defines the distribution of the four mutually exclusive categories in the population. With a fixed sample total *N*, the data is illustrated by the probability model:

$$\begin{aligned} & P\{X=(X_{11}=x_{11},\ X_{12}=x_{12};\ X_{21}=x_{21},\ X_{22}=x_{22})\} \\ &= \frac{N!}{x_{11}!x_{12}!x_{21}!x_{22}!}\, p_{11}^{x_{11}} p_{12}^{x_{12}} p_{21}^{x_{21}} p_{22}^{x_{22}}\ , \end{aligned} \tag{2.2}$$

where

$$\mathrm{P}=\left\{(p_{11},\ p_{12};\ p_{21},\ p_{22}):\ \sum_{i=1}^{2}\sum_{j=1}^{2} p_{ij}=1\right\}$$

is the parameter space with three degrees of freedom (d.o.f.). The units of the two rows may be randomly selected from the two mutually exclusive subpopulations separately, and the units having factor *A* are counted. This defines two independent binomial samples with the row margins fixed, which form Groups 1 and 2 of table (2.1) (e.g., Yule, 1911; Barnard, 1947; Pearson, 1947). In this case, the total count $x_{\cdot 1}$ $(=x_{11}+x_{21})$ of factor *A* is a random variable, and conditioned on the row margins, formula (2.2) yields

$$\begin{aligned} & P\left\{X_{11}=x_{11},\ X_{21}=x_{21}\mid X_{11}+X_{12}=x_{1\cdot}\right\} \\ &= \binom{x_{1\cdot}}{x_{11}}\binom{x_{2\cdot}}{x_{21}} p_1^{x_{11}} q_1^{x_{12}} p_2^{x_{21}} q_2^{x_{22}} \\ &= \binom{x_{1\cdot}}{x_{11}}\binom{x_{2\cdot}}{x_{21}} q_1^{x_{1\cdot}} q_2^{x_{2\cdot}} \exp[x_{11}\log\psi + (x_{11}+x_{21})\log(p_2/q_2)]. \end{aligned} \tag{2.3}$$

Here $p_i = p_{i1}/(p_{i1}+p_{i2}),\ i=1,2,\ \ q_i=1-p_i$, form a parameter space with two degrees of freedom. The functional parameter $\psi = p_{11}p_{22}/p_{12}p_{21} = p_1q_2/p_2q_1$ is called the odds ratio or, the cross-product ratio. Clearly, knowing the p_{ij}'s implies knowing $p_i,\ i=1,2$, and thus, knowing ψ; but not conversely, except that $p_1=p_2$

holds when $\psi = 1$.

Another commonly discussed experiment is the two comparative binomial trials (e.g., Barnard, 1947; Plackett, 1977; Kempthorne, 1978; Yates, 1984; Little, 1989; Greenland, 1991). The model assumes that $x_{1\cdot}$ out of N individuals are randomly assigned to one of two treatments, yielding Group 1, and the remaining $x_{2\cdot}$ to another, forming Group 2 of table (2.1). Under model (2.3), it is often assumed that the individual status of carrying attribute *A* is unchanged, and the column margins of table (2.1) are also considered fixed. Thus, randomization of the units, with or without the attribute *A*, characterizes the extended hypergeometric distribution (Fisher, 1934; Johnson and Kotz, 1969):

$$P\left\{X_{11} = x_{11} \mid X_{11} + X_{12} = x_{1\cdot},\ X_{11} + X_{21} = x_{\cdot 1}\right\} = \binom{x_{1\cdot}}{x_{11}}\binom{x_{2\cdot}}{x_{21}}\frac{\psi^{x_{11}}}{C_t(\psi)}, \qquad (2.4)$$

where

$$C_t(\psi) = \sum_{z=\max(0,\ x_{\cdot 1}-x_{2\cdot})}^{\min(x_{1\cdot},\ x_{\cdot 1})}\binom{x_{1\cdot}}{z}\binom{x_{2\cdot}}{x_{\cdot 1}-z}\psi^{z}\,.$$

It is well known that conditional on both margins $x_{1\cdot}$ and $x_{\cdot 1}$, any entry, say, x_{11}, is sufficient for ψ; and model (2.4) defines a case of one-parameter inference that can be fully illustrated by the likelihood principle, for example, Birnbaum (1962).

2.1 Classical Tests for Independence

Here three tests of independence will be considered. The notion of independence between the two factors is defined in the likelihood (probability) equation as $\psi = 1$, the odds ratio equals to 1. The null hypothesis of independence specifies a composite hypothesis with two d.o.f. (Kendall and Stuart, 1979, p.578):

$$H_0 = \left\{(p_{11}, p_{12};\ p_{21}, p_{22});\ \psi = p_{11}p_{22} / p_{12}p_{21} = 1\right\}. \qquad (2.5)$$

Pearson (1904) developed a chi-square test for H_0 based on his goodness-of-fit test (Pearson, 1900) under the multinomial model (2.2). The test is defined with both margins $x_{1\cdot}$ and $x_{\cdot 1}$ held fixed, without assuming independence between rows, and so, like (2.4), it is termed a conditional test (Yates, 1984). A simplified version is

$$\chi^2 = \frac{N(x_{11}x_{22} - x_{12}x_{21})^2}{x_{\cdot 1}x_{\cdot 2}x_{1\cdot}x_{2\cdot}} \cong \frac{N(|\, x_{11}x_{22} - x_{12}x_{21}\,| - N/2)^2}{x_{\cdot 1}x_{\cdot 2}x_{1\cdot}x_{2\cdot}} = \chi_c^2, \qquad (2.6)$$

where the second fraction, defined as χ_c^2 (Yates, 1934), includes the continuity correction for a more accurate χ^2 approximation to its distribution. The obtained χ^2 and χ_c^2 values can be compared to the table of chi-square distribution with one d.o.f. (Fisher, 1922).

Conditions for or against the use of continuity correction (Plackett, 1964; Grizzle, 1967), and the median probability alternative suggested by Lancaster (1949) have been much discussed, as reviewed by Upton (1982) and Yates (1984), among others. While care must be exercised with multiple common and small χ_c^2 values, when the table margins are small and symmetric, χ_c^2 generally performs well as evidenced by the calibration study of Section 2.4.

Yule (1911) initiated under model (2.3) a statistic for testing H_0, which actually discussed testing the equality H_0^p: $p_1 = p_2$ of the two binomial parameters:

$$Z_Y = \frac{x_{11}/x_{1\cdot} - x_{21}/x_{2\cdot}}{\sqrt{\frac{x_{\cdot 1}}{N}\left(1 - \frac{x_{\cdot 1}}{N}\right)\left(\frac{1}{x_{1\cdot}} + \frac{1}{x_{2\cdot}}\right)}}. \qquad (2.7)$$

The margin $x_{\cdot 1}$ is a sufficient statistic for the common value $p_1 = p_2$ under H_0^p, but not ancillary for ψ (cf. Plackett, 1977; Little, 1989) under H_0. The former, H_0^p has one d.o.f. under (2.3), while H_0 has two d.o.f. under (2.2). Since only the row margin $x_{1\cdot}$ is held fixed, Z_Y is an unconditional test for H_0^p. It follows from (2.6) and (2.7) that the equality $\chi^2 = Z_Y^2$ holds. However, whether the two tests yield equivalent effects for testing H_0, or H_0^p, has not been rigorously examined in the literature, but will be investigated in this study.

The third classical test is the widely discussed exact test (Fisher, 1934), whose test statistic is denoted by T_E. The test statistic can be represented by any entry of the

table (2.1), say, $X_{11}\,(=x_{11})$. Since the two margins $x_{1\cdot}$ and $x_{\cdot 1}$ are fixed, it is an exact conditional test. The null distribution of T_E is the conditional distribution of (2.4) with $\psi = 1$, namely, the hypergeometric distribution:

$$P\left\{X_{11} = x_{11} \mid X_{11} + X_{12} = x_{1\cdot},\ X_{11} + X_{21} = x_{\cdot 1}\right\} = \frac{\binom{x_{1\cdot}}{x_{11}}\binom{x_{2\cdot}}{x_{\cdot 1} - x_{11}}}{\binom{N}{x_{\cdot 1}}}. \qquad (2.8)$$

The finite (discrete lattice) sample space that supports the distribution (2.8) consists of all 2 x 2 tables having the same margins $x_{1\cdot}$ and $x_{\cdot 1}$, denoted by

$$\mathfrak{X} = \left\{(x_{11},\ x_{1\cdot} - x_{11};\ x_{\cdot 1} - x_{11},\ x_{2\cdot} - x_{\cdot 1} + x_{11}) :\ \max(0,\ x_{\cdot 1} - x_{2\cdot}) \le x_{11} \le \min(x_{\cdot 1},\ x_{1\cdot})\right\} \qquad (2.9)$$

For observed data (2.1), the *p* value of T_E under H_0 is the extremity probability, defined to be the sum of the probabilities, given by (2.8), for the members in $\mathfrak{X}$, whose probabilities are not greater than those of the observed data. The number of elements in $\mathfrak{X}$ is equal to "the minimum of the four margins plus 1," which is less than $(x_{1\cdot}+1)(x_{2\cdot}+1)$, the number of elements in the sample space of the independent binomial model (2.3). When the sample size N is small, hence $\mathfrak{X}$ is small, a *p* value of the exact test can be substantially less than a nominal significance level. While increasing the *p* value by randomization is often unacceptable, the exact test has been criticized as being rather conservative (Berkson, 1978). This is most remarkable when comparing T_E with Z_Y among others, for testing H_0 under model (2.3).

A common trait of the three tests χ^2, Z_Y and T_E is that they all measure the deviation from independence using both margins of data (2.1). While T_E is conservative in terms of p value defined by the hypergeometric distribution (2.8), it does enjoy a large sample approximation to normality under model (2.4). By Stirling's approximation, a standardized version of the test statistic T_E, or $X_{11}\,(=x_{11})$ of (2.1), is asymptotically standard normal under H_0 (Pearson, 1947; Feller, 1968; Lancaster, 1969; Cox and Snell, 1989, p. 48):

$$Z_E = \frac{x_{11} - \frac{x_{\cdot 1} x_{1\cdot}}{N}}{\sqrt{\frac{x_{\cdot 1} x_{\cdot 2} x_{1\cdot} x_{2\cdot}}{N^2 (N-1)}}}. \qquad (2.10)$$

In general, asymptotic normality of X_{11} holds under model (2.4) if, and only if, $x_{1\cdot} x_{2\cdot} x_{\cdot 1} x_{\cdot 2} / N^3$ tends to infinity as N does (Kou and Ying, 1996). It is seen that $Z_E^2 = \chi^2$ on $\mathfrak{X}$, and $Z_E = Z_Y$ if $(N-1)$ in (2.10) is replaced with N. Moreover, using fixed $x_{1\cdot}$ and $x_{\cdot 1}$, the test statistics χ^2 and Z_E are invariant with respect to "the sample odds ratios", and Z_Y is invariant with respect to "the difference between the two binomial sample rates" of data (2.1). Despite these common properties, asymptotic power evaluations under a simple alternative to H_0 have not been established for χ^2 and Z_E, hence T_E. However, Z_Y has the exact independent binomial power analysis. On this issue, it is noteworthy that a classical unconditional power analysis for testing $p_1 = p_2$ is based on selecting a critical region among the mixtures of critical regions defined by the hypergeometric distributions having the same $x_{1\cdot}$ and N, but different $x_{\cdot 1}$ of (2.8) (e.g., Bennet and Hsu (1960); Boschloo (1970); Gail and Gart (1973); and Mehta and Patel (1980)). The main concern in these studies is on finding a wider critical region, for which the conditional levels of significance can be raised above a nominal level in order to reach the unconditional nominal level. Since the derived power calculation is also a sum of independent binomial probabilities, comparable to those of Z_Y, similar discussions are omitted.

Among the classical unconditional tests, it is well known that Barnard disqualified his CSM test (1945, 1949). Many studies with unconditional tests have been using Z_Y to define the p values and critical regions, e.g., Berkson (1978), Suissa and Shuster (1985), and Haber (1986). These authors discussed under model (2.3) the acquired level and power of the test Z_Y with the aim of selecting a more powerful (actually wider) critical region, see, for example, Santner and Duffy (1989), Agresti (1990), Berger and Boos (1994), and Berger (1996). The dispraise of the exact test by Berkson (1978) that strongly advocates Z_Y as a substitute for T_E, has gradually received less consensus, since Yates (1984, p. 433) argued that "testing for H_0 must be conditioned on both margins, whether data (2.1) is obtained from any one of the three experiments (2.2), (2.3) and (2.4)." Readers may refer to Berkson (1978) for the

introduction of the criticism, and to Barnard (1979), Upton (1982), and Yates (1984) for the details of the debate. The logic behind the comparisons between the conditional and unconditional tests were notably discussed by Little (1989) and Greenland (1991), who also deemed the use of the unconditional inference for testing H_0 suspect. Nevertheless, the notion of a conservative T_E, as compared to Z_Y, has continued to be acknowledged among many statisticians, including, Kempthorne (1978), Upton (1992) and Agresti (2002, p. 96).

The above literature on the classical tests for H_0 signals two important issues. First, if a unified power analysis holds for χ^2, Z_E and T_E altogether, then, it would likely justify that testing independence (H_0) should be conditioned on the sample space $\mathfrak{X}$. Naturally, the second issue is whether the unconditional test for H_0^p using Z_Y should be legitimately compared against the exact test T_E for testing H_0. These two issues will be addressed in this study using information identities developed through the likelihood ratio test.

2.2 Likelihood Ratio Test and Conditionality

It seems useful to examine the relationship between the chi-square test and the exact test, based on the likelihood ratio test (LRT) statistic. Additional notations are defined for ease of exposition. Let $P(X) = P(X = (x_{11}, x_{12}; x_{21}, x_{22})) = (x_{11}, x_{12}; x_{21}, x_{22})/N$ denote the observed sample proportion, that is, the empirical multinomial distribution. By (2.2), for $P = (p_{11}, p_{12}; p_{21}, p_{22}) \in \mathrm{P}$, let $P_{i\cdot}$ and $P_{\cdot j}$ be likewise defined as the row and column margin probabilities; moreover, let $(X; P)$ denote that "P is the true distribution of X", and let $f(X; P)$ be the corresponding likelihood function. Given data $X = x$ of (2.1), the LRT statistic $\lambda = \max_{Q \in H_0} f(X;Q)/f(X;P(X))$ for testing $P \in H_0$ satisfies the equation:

$$-2\log\lambda = 2\sum_{i,j=1}^{2} x_{ij}\log(Nx_{ij}/x_{i\cdot}x_{\cdot j}) = \chi^2\left(1+O_p(N^{-1/2})\right). \qquad (2.11)$$

Here, the first equation follows from maximum likelihood estimation and the second is asymptotically valid for large N (Kendall et al., 1979, p.579; Wilks, 1935). The

second term of (2.11), divided by the sample size, defines the Kullback-Leibler (1951) divergence

$$D(P(x) \,||\, \hat{P}(x)) = \sum_{i,j=1}^{2} (x_{ij} / N) \log(N x_{ij} / x_{i\cdot} x_{\cdot j}) = \sum p_{ij}(x) \log(p_{ij}(x) / p_{i\cdot}(x) p_{\cdot j}(x))$$

and characterizes the LRT statistic for the observed data x as

$$\max_{Q \in H_0} \frac{f(x;\, Q)}{f(x;\, P(x))} = \frac{f(x;\, \hat{P}(x))}{f(x;\, P(x))} = e^{-ND(P(x)\|\hat{P})}, \tag{2.12}$$

where $\hat{P}(x) = p_{i\cdot}(x) p_{\cdot j}(x) = (x_{1\cdot} x_{\cdot 1},\ x_{1\cdot} x_{\cdot 2};\ x_{2\cdot} x_{\cdot 1},\ x_{2\cdot} x_{\cdot 2}) / N^2$ is the unique MLE of $P(x)$ under H_0. Clearly, (2.12) is also valid for any table x in $\mathfrak{X}$, and $\hat{x} = N\hat{P}(x) = N\hat{P}$ defines the same $\hat{P}$ for all x. Although $\hat{x}$ need not be a member of $\mathfrak{X}$, it has the same margins as x, and lies in the continuum extension $\mathfrak{X}_C$ (defined in (2.15)) of the finite discrete lattice $\mathfrak{X}$. For the observed table X, values of (2.12) over the sample space $\mathfrak{X}$ may be normalized to form a discrete conditional distribution. Equivalently, let *CR* represent a one-sided critical region, that is, a one-sided boundary subset (see (2.16) for detailed formulation) of $\mathfrak{X}$, then the conditional distribution of the LRT (2.12) evaluates

$$P\left((X_i \in CR);\, \hat{P}(x)\right) = \sum_{x_j \in CR} \frac{e^{-ND(P(x_j)\|\hat{P})}}{S(\hat{P})}, \tag{2.13}$$

to yield the acquired p value, where $S(\hat{P}(x)) = S(\hat{P}) = \sum_{x_j \in \mathfrak{X}} e^{-ND(P(x_j)\|\hat{P})}$ is the normalizing constant. An analogue of both (2.12) and (2.13) can also be obtained for the hypergeometric distribution of the exact test T_E as

$$\frac{f(x;\, \hat{P}(x))}{f(x;\, P(x))} \cong e^{-(N+\frac{1}{2})D(P(x)\|\hat{P})}$$

and

$$P\left((X_i \in CR);\, \hat{P}(x)\right) = \sum_{X_j \in CR} \frac{e^{-(N+\frac{1}{2})D(P(X_j)\|\hat{P})}}{S_E(\hat{P})}, \tag{2.14}$$

where $S_E(\hat{P})$ is the same normalizing constant having the exponent N replaced by

$N+1/2$. Formula (2.14), derived from Stirling's formula, closely approximates the exact distribution (2.8).

Suppose that data $X = x$ has odds ratio $\psi_x = x_{11}x_{22}/x_{12}x_{21}$, and that x is situated on one side of $\hat{x} = N\hat{P}(x)$ (say, $\psi_x > \psi_{\hat{P}(x)} = 1$) on $\mathfrak{X}$. An enlarged ideal sample space can be defined as $\mathfrak{X}_C$ = the continuum (lattice extension) of $\mathfrak{X}$, which consists of all tables with non-negative entries (but not necessarily integers) and the same margins as any members in (the lattice) $\mathfrak{X}$. Specifically,

$$\mathfrak{X}_C = \left\{ x(\varepsilon) = (x_{11}+\varepsilon, x_{12}-\varepsilon, x_{21}-\varepsilon, x_{22}+\varepsilon) : \; x_{ij} \pm \varepsilon \geq 0, \text{ real } \varepsilon \right\}. \quad (2.15)$$

Figure 1 exhibits (non-convex) lattice hyperplanes of relative entropy, and $\mathfrak{X}_C$ can be visualised as the vertical (lattice) line segment that connects the data x to the MLE $\hat{x} = N\hat{P}(x)$, where it is perpendicular to the null hyperplane H_0.

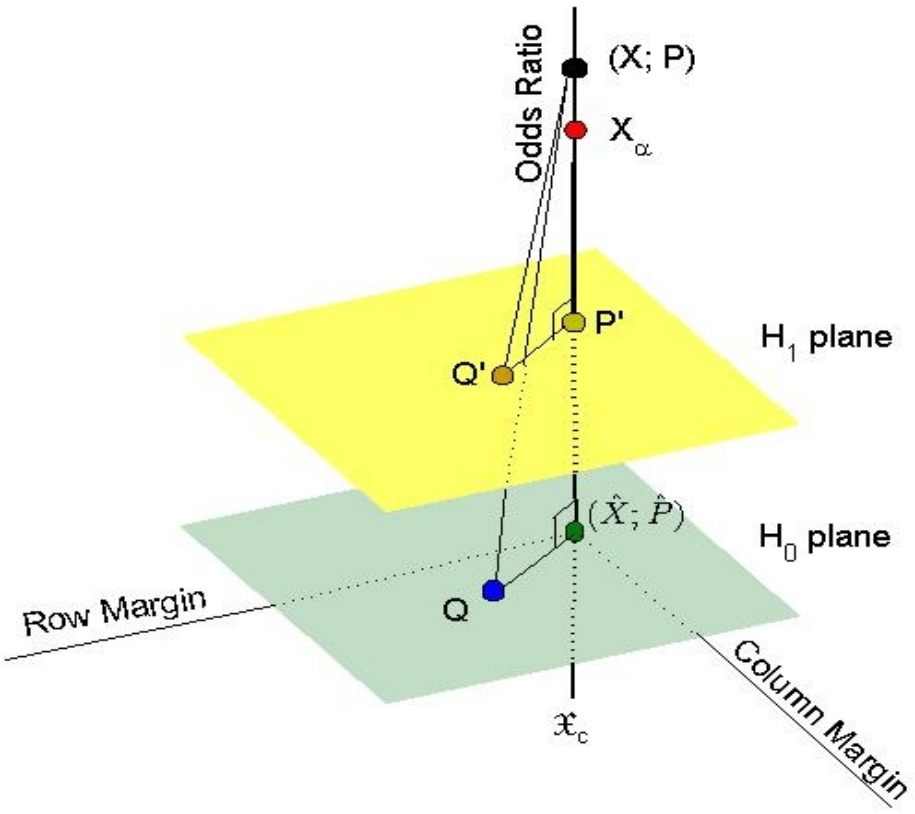

Figure 1. Central pillar $\mathfrak{X}_C$ is the continuum extension of the sample space $\mathfrak{X}$; H_0 is the null hyperplane with odds ratio 1; H_1 has a unique odds ratio.

Now, consider a one-sided critical-region subset of $\mathfrak{X}_C$. Let $X = x$ be observed with $\psi_x > 1$, and define the one-sided (boundary set) critical region by

$$CR_x = \left\{ x(\varepsilon) \in \mathfrak{X}_C : \ \psi_{x(\varepsilon)} = \frac{(x_{11}+\varepsilon)(x_{22}+\varepsilon)}{(x_{12}-\varepsilon)(x_{21}-\varepsilon)} \geq \psi_x \right\}. \qquad (2.16)$$

The approximation to the chi-square distribution (2.6), together with equation (2.13), establishes the standard weak convergence, that for large sample size N

$$\sum_{x_j \in CR_x} \frac{e^{-ND(P(x_j)\|\hat{P})}}{S(\hat{P})} \cong \frac{\int_{CR_x} e^{-ND(P(z)\|\hat{P})} d\psi_Z}{\int_{z \in \mathfrak{X}_C} e^{-ND(P(z)\|\hat{P})} d\psi_Z},$$

and

$$P\left\{(Z \in CR_x);\ \hat{P}(x)\right\} = \frac{\int_{z \in CR_x} e^{-ND(P(z)\|\hat{P})} d\psi_Z}{\int_{z \in \mathfrak{X}_C} e^{-ND(P(z)\|\hat{P})} d\psi_Z} \cong \frac{1}{2} P\left\{\chi^2 > 2ND(P(x) \,||\, \hat{P})\right\}, \quad (2.17)$$

where the random table Z is realized as a member z in $\mathfrak{X}_C$. The last term of (2.17) is replaced by $1 - P\{\chi^2 > 2ND(P(x) \,||\, \hat{P})\}/2$ when $\psi_x \leq 1$, which rarely occurs as a practical choice of a *CR*. Thus, (2.17) can be used to estimate the *p* value of any observed 2×2 table in $\mathfrak{X}_C$ with nonnegative entries and an arbitrary odds ratio.

The above analysis shows that the conditional distribution of the chi-square test and the LRT are closely comparable to that of the exact test. It is of interest to examine whether the same characterization from (2.11) to (2.17) holds over the entire parameter space of testing independence.

2.3 Likelihood Ratio Test and Mutual Information

The first step is to examine whether the calibration (2.13) would be valid not only for the MLE $\hat{P}$ but also for any member Q of the null hypotheses H_0. At the outset, this seems to be a redundant issue, since the LRT (2.12) is maximized over all members of H_0. However, the logical notion is: "Suppose any individual parameter Q of H_0 were a hypothetical alternative to $\hat{P}$, would it possibly affect the validity

of (2.13)?" This is answered below by Lemma 1, using the definition of mutual information (Gray, 1990). The observation also provides a fundamental characterization of the MLE, but differs from the additivity of the minimum discrimination information, discussed for asymptotically optimal hypothesis testing procedures (e.g., Gokhale and Kullback (1978)). The proof of lemma 1 is elementary and omitted.

Lemma 1. (The Pythagorean Law of Relative Entropy) For given data X, $P = P(X)$ and for any $Q \in H_0$, the mutual information yields the MLE $\hat{P}$ via the identity:

$$D(P \,||\, Q) = D(P \,||\, \hat{P}) + D(\hat{P} \,||\, Q). \tag{2.18}$$

The term "Pythagorean law" is coined by the fact that (2.18) partitions the approximate chi-square distribution with three d.o.f. for a 2×2 table (cf. Kendall et al. 1979, (33.117); Rao, 1973, (6d.2.6)), as shown by Figures 1 and 2. By (2.18), the term mutual information between a pair of random variables (X, Y) with joint probability density $f(x, y) \cong P$ can be equivalently defined as

$$I(X;Y) = D(P \,\|\, \hat{P}) = \min_{g(x,y) \in H_0} D(f \,\|\, g). \tag{2.19}$$

As a consequence of Lemma 1, (2.13) can be generalized over H_0. The following theorem follows by incorporating (2.18) into an analogue of (2.13), and cancelling the common factor $D(\hat{P} \,||\, Q)$, thus the proof is omitted.

Theorem 1. For data $(X, P(X))$ of (2.1), any $Q \in H_0$, and for each one-sided boundary subset CR of $\mathfrak{X}$, the following equation holds for testing $X \cong \hat{P}(X) = \hat{P}$ against $X \cong Q$ in distribution,

$$P\big((X_i \in CR);\ Q\big) = \sum_{x_j \in CR} \frac{e^{-ND(P(x_j)\|\hat{P})}}{S(\hat{P})}, \tag{2.20}$$

where $\hat{P}$ is the projection of the KL-divergence from $P(X)$ onto H_0.

The right-hand side of (2.20) is the same as that of (2.13), as expected. Theorem 1 establishes that, by the conditionality principle, testing the composite H_0 is equivalent to testing the single null parameter $P = \hat{P}$, and that the unconditional MLE under model (2.2) reduces to the same $\hat{P}$ as the conditional MLE under model (2.4); moreover, the reduction passes through model (2.3). It also characterizes the MLE of the LRT as the projection root of the KL-divergence, which is the mutual information under general hypothesis testing for independence. In the literature, the asymptotic chi-square distribution for the parametric LRT was also examined by Chernoff (1954); and, given a uniform (improper) prior supported on H_0, the posterior mode is the projection of the KL-divergence (Lindley, 1956).

2.4 Calibration of Conditional Tests

Distributions of the conditional test statistics were generated for a comparison study. Two 2×2 tables with different sample sizes were evaluated using the conditioned sample sets $\mathfrak{X}$ with fixed margins. The data table $X = (5, 6; 2, 5)$ yielded 8 members in $\mathfrak{X}$, and $X = (16, 8; 9, 15)$ yielded 24 tables. A large sample size table, say, $(135, 190; 75, 100)$ would show closer approximation between the test statistics, but for brevity, it is not reported. Tables 1 and 2 list the *p* values obtained by the four tests. These are one-sided *p* values associated with one-sided (upper) critical regions, consisting of tables in $\mathfrak{X}$ whose odds ratios are increasingly greater than those of the given data X. For example, Table 1 lists, for each test, eight ascending *p* values with corresponding odds ratio values ψ in the left-end column. The *p* values increase toward 1 as the values of ψ decrease toward 0, and a boxed *p* value corresponds to a one-sided critical region consisting of tables that are greater in ψ and more extreme (in probability) than the observed data X. In Tables 1 and 2, the two chi-square statistics of (2.6) plus their *p* values, the *p* values for the exact test (2.8), and those for the LRT (2.13) were all calibrated together on the same scale, by matching the same one-sided critical region with each individual member of the finite sample space $\mathfrak{X}$.

The calibration results of the tables in $\mathfrak{X}$, including the few examples presented here, can be summarized as follows. By formula (2.13), the computed *p* values of the exact test T_E of (2.8), the LRT (2.12), and the Yates χ_c^2 of (2.6) are very close to each other. As is well known, Yates' *p* values can be over-corrected when the table margins are small and symmetric, however, in other situations it closely approximates the *p* values of the exact test T_E. The *p* values of the LRT are more leptokurtic in the

center and lighter in the tails, reflecting the well-known most powerful (unbiased) property of the LRT. However, the *p* values of the Pearson χ^2 are generally much smaller, giving the most liberal results among the four tests. In general, the exact, the LRT and the Yates chi-square tests yield similar *p* values consistently, including small values near the commonly used nominal levels such as $\alpha = 0.1$, 0.05 and 0.01.

3. Power Analysis for Testing Independence

It is well known that the odds ratio plays an important role in the application of generalized linear models for studying biomedical, environmental, epidemiological and pharmaceutical experiments. The conditional distribution of X given the row and column margins depends on a single parameter, say, the odds ratio. Lemma 1 and Theorem 1 have shown that the (lattice) hyperplane of odds ratio 1 identifies the composite null hypothesis H_0 with two d.o.f. In contrast, each hyperplane of the composite alternative hypothesis H_1 consists of the probability vectors having the same odds ratio unequal to 1. It is meaningful to extend the scope of Lemma 1 from the null hypothesis to general alternatives, that is, to examine whether the conditioning argument (2.20) would be valid if H_0 is replaced with H_1.

3.1 Invariance of the Entropy Identity

To develop the power analysis, the notations used in Section 2 plus some others will be reorganized for ease of exposition. Let $(X = (x_{ij},\ i, j = 1,2);\ P(X))$ be the observed data. Let $(Y = (x_{ij}^*);\ Q'(Y))$ be any member of H_1, having odds ratio $\psi = x_{11}^* x_{22}^* / x_{12}^* x_{21}^* \neq 1$. It is straightforward to find the unique fourfold vector $(X' = (x_{ij}^{'});\ P(X') = P')$ on the continuum $\mathfrak{X}_C$, having the same odds ratio ψ (see Figure 1). An invariance property of the conditional distributions with respect to both H_0 and H_1 will hold as an extension of the information identity of Lemma 1. The proof of Lemma 2 is given in the Appendix. Subsequently, notations will be simplified, say, P and Q will be used, instead of $P(X)$ and $Q(Y)$.

Lemma 2. (Extended Pythagorean Law) Let $(X;\ P)$ be an observed 2×2 table

of (2.1). Let $(Y;\ Q') \in H_1$ and $(X';\ P')$ have the same odds ratio $(\psi \neq 1)$, where X and X' are members of $\mathfrak{X}_C$. The following additive law of information holds:

$$D(P \,||\, Q') = D(P \,||\, P') + D(P' \,||\, Q') . \tag{3.1}$$

It is noted that P' is the root of projection from P onto the hyperplane $H(\psi)$ of fourfold vectors having the common odds ratio ψ. In the null case of Lemma 1, $Y \in H_0 = H(\psi = 1)$, then $P' = \hat{P}$ and (3.1) reduces to (2.18). Lemma 2 thus extends Lemma 1 from the null hypothesis to the entire parameter space of non-negative odds ratios.

The main purpose is to characterize the power analysis at any alternative in H_1 based on the test for H_0. The next theorem, being a natural extension of Theorem 1, fulfils this goal. The proof directly follows by using Lemma 2 together with a similar argument to Theorem 1.

Theorem 2. Let $(X;\ P)$ be a 2×2 table. Let $Q \in H_1$ have odds ratio ψ. Then, for a CR subset of $\mathfrak{X}$, and a normalizing constant defined by (2.13), the following equation holds for testing H_0 against Q in distribution

$$P\big((X \in CR);\ Q\big) = \sum_{X_j \in CR} \frac{e^{-ND(P(X_j)\|P')}}{S(P')}, \tag{3.2}$$

where P' is the projection of the KL-divergence from P onto $H(\psi)$.

Like Lemma 2, if Q is a member of H_0 then (3.2) reduces to (2.20). Theorem 2 thus generalizes Theorem 1 and verifies that the conditional distributions of the LRT are invariant with respect to each common odds ratio hyperplane. Analogous to (2.17) for testing H_0, (3.2) leads to the power evaluations discussed below.

Let $(X;\ P)$ be an observed 2×2 table with $\psi_P = x_{11}x_{22} / x_{12}x_{21} > 1$, and let $(Y;\ Q') = (x_{ij}^*) \in H_1$ be any alternative with $\psi_{Q'} > 1$, where X and Y are situated on the same side of $\hat{P}$ $(\psi_{\hat{P}} = 1)$. Given a nominal level $\alpha < 1/2$, let $CR_{X,\alpha} \subset \mathfrak{X}_C$ be a one-sided boundary set, as defined by (2.16), satisfying an analogue of (2.17):

$$\alpha = P\left\{(Z \in CR_{X,\alpha});\ \hat{P}\right\} \cong \frac{1}{2} P\left\{\chi^2 > 2ND(P(X_\alpha) \,\|\, \hat{P})\right\}, \qquad (3.3)$$

where $D(P(X_\alpha) \,\|\, \hat{P}) = \min_{Z \in CR_{x,\alpha}} D(P(Z) \,\|\, \hat{P})$. Note that, for $\alpha = 1/2$, the obvious choice is $X_\alpha = \hat{P}$. It is straightforward to compute the power of the test (defined by $CR_{X,\alpha}$) at the alternative hypothesis $(Y;\ Q')$ according to (3.2),

$$P(Z \in CR_{X,\alpha} \mid Q') \cong \frac{1}{2} P\left\{\chi^2 > 2ND(P(X_\alpha) \,\|\, P')\right\}, \quad \text{if } \psi_{X_\alpha} > \psi_{P'} \geq 1; \text{ or,}$$

$$\cong 1 - \frac{1}{2} P\left\{\chi^2 > 2ND(P(X_\alpha) \,\|\, P')\right\}, \quad \text{if } \psi_{P'} \geq \psi_{X_\alpha} > 1. \qquad (3.4)$$

Fisher (1962) illustrated a confidence interval for the odds ratio parameter given a 2×2 table. The analysis was essentially an analogue of (3.4). Theorem 2 has conveyed two practical messages through (3.3) and (3.4). First, a critical region with unbiased level via (3.3), plus a desired sensitivity via (3.4), can be designed within the continuum sample space $\mathfrak{X}_C$. Thus, the information identity (3.2) establishes a Neyman-Pearson decision inference within this testing frame. Second, *p* values of (2.17) and power computations with (3.3) and (3.4) are validated not only for the LRT and the Pearson-Yates chi-square test, but also for the Fisher exact test in lieu of (2.14). The exact test obtains the power evaluation by the LRT approximation to the KL-divergence defined on $\mathfrak{X}_C$, the extension of $\mathfrak{X}$, and also of the support of the extended hypergeometric distribution. Altogether, the above discussion has addressed the first issue of section 2.1: *testing H_0 is essentially conditioned on the sample space $\mathfrak{X}$*.

3.2 Power Analyses in Practice

Data describing an experiment of vaccine inoculation, for the immunization of cattle from tuberculosis (Kendall and Stuart, 1979, Table 33.4, p. 616), is used for illustration. The 2×2 table is $X = (\text{nv-a} = 8,\ \text{nv-na} = 3;\ \text{v-a} = 6,\ \text{v-na} = 13)$, where the row letters "nv" and "v" stand for no-vaccine and vaccine-inoculated, with margins 11 and 19; the column letters "a" and "na" stand for tuberculosis-affected and unaffected, respectively, with margins 14 and 16. The odds ratio of data X is $\psi_X =$ 5.78. Under H_0, the MLE is $(\hat{X};\ \hat{P} = P(\hat{X}))$ with

$\hat{P} = (5.13, 5.87; 8.87, 10.13)/30$, the observed one-sided p value of the Pearson χ^2 is 0.015, and similar p values of the χ_c^2 and the exact T_E are close to 0.036. For power evaluations in accordance with the Neyman-Pearson theory, the nominal significance level $\alpha = 0.05$ is chosen for a detailed discussion below.

The discrete sample space $\mathfrak{X}$, induced by the observed table X, contains 12 members. It follows by (3.3) that $X_\alpha = (7.3, 3.7; 6.7, 11.3) \in \mathfrak{X}_C$ defines the boundary of a one-sided (larger-odds-ratios) critical region at level $\alpha = 0.05$. To give an example of a case of power analysis using Thoerem 2, choose a member $(X' = (7, 4; 7, 12); P'=P(X'))$ in $\mathfrak{X}$, with odds ratio $\psi_{P'} = 3$. Let H_1 denote the lattice hyperplane of all 2×2 tables having the same odds ratio 3 and sample size $N = 30$. Thus, X' is located on H_1, indeed, $X' = \mathfrak{X} \cap H_1$. Given the level-$\alpha$, a one-sided critical region with boundary X_α, the power evaluation at $(X'; P')$ yields 0.438 as the computed χ_c^2 in (3.4).

In addition to the classical comparison in terms of p values, it is meaningful to compare the conditional tests with the unconditional test Z_Y based on the power evaluations carried out in the data example above. This will be examined using model (2.3) as a common ground for comparison. Thus, suppose the null hypothesis specifies that the proportions of the vaccine-inoculated units are the same across the column factor, affected and unaffected, denoted by H_0^p: $p_1 = p_2$. Under model (2.3), the sample space is characterized by the column lattice hyperplane H_c that consists of 255 members of 2×2 tables with the same column margins. Using test scores of Z_Y, the boundary table of a typical one-sided critical region is found to be $(Y_\alpha;\ P(Y_\alpha) = (13, 11; 1, 5)/30 = Q_\alpha)$ having p value 0.0498, and odds ratio 5.91. Meanwhile, consider the alternative table $(Y_c;\ P(Y_c) = (9, 6; 5, 10)/30 = Q_c)$ with odds ratio 3. It is located on the lattice line $H_1 \cap H_c$, which contains the table $(X';\ P' = (7, 4; 7, 12)/30) = H_1 \cap \mathfrak{X}_C$ (see Figure 2). For the pairs $(Q_\alpha;\ P')$ and $(Q_\alpha;\ Q_c)$, the test Z_Y evaluates the exact binomial probability as the power at P' and Q_c to be 0.428 and 0.436, respectively. The two power values are not equal, though not far from 0.438, the constant previously obtained for both pairs $(X_\alpha; P')$ and $(X_\alpha; Q_c)$ by (3.4), because P' and Q_c have the same odds ratio. Obviously, one can not expect the test Z_Y to be more powerful than the LRT, χ^2 and the exact

T_E.

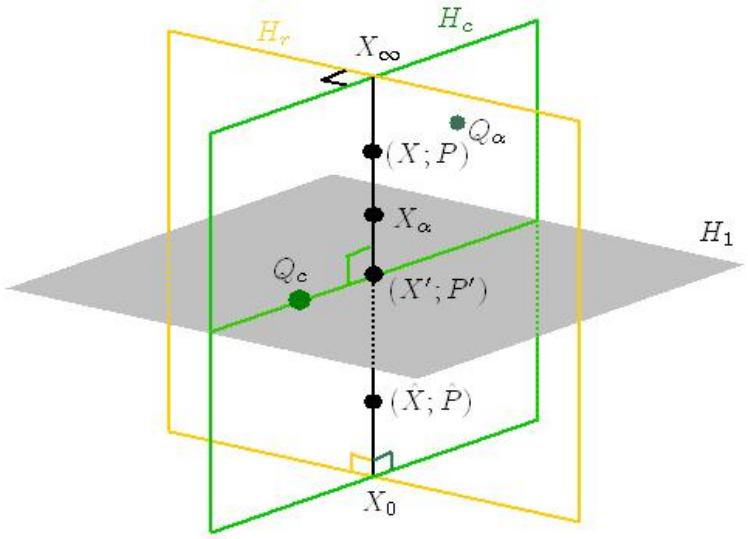

Figure 2. H_c, or H_r, is a binomial product sample space with fixed column, or row, margins, respectively; H_1 is the horizontal hyperplane of odds ratio 3 perpendicular to $\mathfrak{X}_C$; $X' = H_1 \cap \mathfrak{X}_C$ and $\mathfrak{X} = H_r \cap H_c$.

By symmetry, similar comparisons of power could also be obtained using the null hypothesis that the proportions of affected units are equal across the other factor, vaccine-inoculated or no-vaccine. It would, however, yield different critical region and power calculations on a row hyperplane H_r (Figure 2), from those derived with H_c, noting that H_r and H_c are perpendicular planes as remarked after (2.18).

3.3 The Logic of Testing Independence

It was noted in Section 2.1 that the test statistics χ^2, Z_Y and Z_E are essentially equal, based on the same margins of a 2×2 table. The statistic Z_Y has been widely used for testing H_0^p: $p_1 = p_2$ under model (2.3) with exact power computations at specified p_i's. In contrast, Theorem 2 addresses the first question of Section 2.1 by proving that both χ^2 and Z_E, hence T_E, evaluate asymptotic power in terms of usual approximations to chi-square distributions as used for testing H_0.

The second issue of Section 2.1 is thus addressed as it is not legitimate to compare the unconditional test Z_Y against the exact test T_E, or Z_E, for testing H_0 under

model (2.3). The hypothesis of independence H_0, $\psi = 1$, is universally defined and irrelevant to whichever model (2.2), (2.3) or (2.4) is assumed. From model (2.2) to model (2.3), the two d.o.f. H_0 is reduced to the one d.o.f. H_0^p ($p_1 = p_2$); and conversely, the alternative hypotheses parameter spaces are of one, and two, d.o.f., respectively, as shown by Figures 1 and 2. Recall the example of Section 3.2, where the same power values are obtained by the conditional tests at the alternatives P' and Q_c, having the same odds ratio on H_1. But, the unconditional test Z_Y must treat $P' = (p_1 = 0.5;\ p_2 = 0.25)$ and $Q_c = (p_1 = 0.644;\ p_2 = 0.375)$ differently, due to unequal ratios p_1 / p_2, which are 2 and 1.71, respectively. The interpretations of the two tests are different in meaning, or in purpose. Since Fisher's exact test was defined for testing independence, this illustrates that it is logically flawed to compare an unconditional test to a conditional test for testing independence under model (2.3).

4. Applications of the Information Identity

Beyond the 2×2 tables, multivariate data structures in the form of contingency tables have been widely studied in the literature. To illustrate the idea, the conditioning argument of Section 3 can be applied to testing basic association models in two-way contingency tables. Applications to general multi-way contingency tables will be presented in forthcoming studies.

4.1 The Basic $2 \times J$ Tables

It is well known that testing for uniform association or for independence within a $2 \times J$ table, $J \geq 3$, is equivalent to testing that the consecutive odds ratios between the two rows are all equal to 1. The LRT or the Pearson chi-square test is commonly used with d.o.f. $J - 1$, which is the number of consecutive odds ratios that can be estimated or tested within the table. As an extension of Lemma 2, it is often useful to test an alternative hypothesis such as a goodness-of-fit test, where the consecutive odds ratios may follow specified proportions or a trend. In Sections 4.1 and 4.2, the geometric viewpoint of Section 3 is used to illustrate the division of information in testing for these general forms of association.

Let X and Y be two $2 \times J$ tables with equal sample size. Assume that the $J - 1$ odds ratio parameters of Y are known or specified, or identically equal to a positive constant as a special case of uniform association. Treating Y as an alternative hypothesis, a test of goodness-of-fit or similarity between X and Y can be examined by an information identity like (3.1) of Lemma 2. Extensions of Lemma

3 to $I \times J$ tables will be discussed in Section 4.2.

Lemma 3. Assume X and Y are $2 \times J$ tables as defined above. Then, there exists a unique $2 \times J$ table Z, having the same row and column margins as X, and having the same $J-1$ consecutive odds ratios as those of Y, and an analogue of Lemma 2 holds:

$$D(X \,\|\, Y) = D(X \,\|\, Z) + D(Z \,\|\, Y). \tag{4.1}$$

A proof of Lemma 3 is given in the Appendix. The asymptotic distributions associated with the sample versions of (4.1) satisfy that $\chi^2_{J-1} = \chi^2_{J-2} + \chi^2_1$.

4.2 The $I \times J$ Tables

A general family of two-way association models for the $I \times J$ tables will be discussed. This subsection will discuss an application of Lemma 3 for testing certain uniform association models for two-way tables. The technical extension of (4.1) is based on the assumption that the odds ratios along the rows, columns, or rows-by-columns are in proportion. It is found that computing the MLE's of the odds ratio parameters by minimizing the relative entropy of (4.1) is an alternative method to standard log-linear modelling for fitting row and column effects of homogeneous association (cf. Goodman, 1984, Chapter 4, Table 4).

Models of no association, homogeneous association, and specified odds ratios of Section 4.1 are the basic models. For $I \geq 3$, it follows from Goodman's terminology (1984, pp. 89-91) that models of interest are specified by the row effects, the column effects, and the row-by-column effects, with model parameters :

$$\psi_{i+} = \psi \eta_{i+} \tag{4.2}$$

$$\psi_{+j} = \psi \eta_{+j} \tag{4.3}$$

$$\psi_{ij} = \psi \eta_{i+} \eta_{+j}, \tag{4.4}$$

respectively, where

$$\prod_{i=1}^{I-1} \eta_{i+} = 1, \quad \text{and} \quad \prod_{j=1}^{J-1} \eta_{+j} = 1. \tag{4.5}$$

Under the first constraint of (4.5), the row effect model (4.2) will estimate $I-1$ row effect parameters under the null model, and enjoy $(I-1)(J-1)-(I-1)=(I-1)(J-2)$ d.o.f. Lemma 3 can be used to yield $I-1$ pairs of rows having the same $J-1$ odds ratio in each pair, while minimum KL divergence is achieved by keeping the row and column margins unchanged with each pair. To match the column margins when combining the pairs of rows, it suffices to rescale either the rows or the columns by appropriate positive constants to satisfy the solutions for the equations in J columns. Thus, it estimates the fitted table Z for testing model (4.2) such that asymptotically $D(X \| Z)$ enjoys the chi-square distribution with $(I-1)(J-2)$ d.o.f. The argument is likewise applicable to test the column effect model (4.3), and then the row and column effect model (4.4) and (4.5). This follows from Lemma 3, to yiled an alternative approach to the method of logl-inear models as illustrated by Goodman (1984, Chapter 4).

5. Conclusion

It is well known that factorization of the likelihood defines two important notions, independence and sufficiency, and together they constitute the likelihood approach to statistical inference. The LRT, mostly notable in the likelihood approach, has been widely used in testing hypotheses via Neyman-Pearson theory, in particular, testing independence with 2×2 contingency tables. The calibration of the p values of the conditional tests, a key idea due to Fisher (1935), relies upon the LRT given the margins, which can be derived from the mutual information identity. The invariance of information identity leads to the development of the (asymptotic) power analyses for both Pearson's chi-square test and Fisher's exact test. It also illustrates that the conditioned (and extended) sample space $\mathfrak{X}_{\mathrm{C}}$ offers an answer to Fisher's "most relevant set" (Barlett, 1984), where conditional distributions and unified power analysis of the LRT, the chi-square test and the exact test are validated. This last observation also resolves the long-term debate on the criticism of Fisher's exact test. That is, Berkson's dispraise against the exact test, in terms of conservative p values and improved power evaluations, was logically flawed due to the different models and hypotheses under evaluation.

Acknowledgements

The authors are grateful to Shaowei Cheng for careful reading of the manuscript plus suggestions; and, are also indebted to the Associate Editor and Referees for their many helpful comments.

References

Agresti, A. (1990) (2nd ed., 2002). *Categorical Data Analysis*. New York: Wiley.

Barnard, G. A. (1945). A new test for 2×2 tables. *Nature* **156**, 177.

Barnard, G. A. (1947). Significance tests for 2×2 tables. *Biometrika*, **34**, 123-138.

Barnard, G. A. (1949). Statistical inference. *J. Royal Statist. Soc., B*, **11**, 115-139.

Barnard, G. A. (1979). In contradiction to J. Berkson's dispraise: conditional tests can be more efficient. *J. Statist. Planning and Inference*, **3**, 181-187.

Bartlett, M.S. (1984). Discussion on tests of significance for 2x2 contingency tables (by F. Yates). *J. Royal Statist. Soc., A*, **147**, 453.

Bennett, B. M. and Hsu, P. (1960). On the power function of the exact test for the 2×2 contingency table. *Biometrika*, **47**, 393-398 (correction 48 [1961], 475).

Berger, R. L. and Boos, D. D. (1994). P-values maximized over a confidence set for the nuisance parameter. *J. Am. Statist. Assoc.*, **89**, 1012-1016.

Berger, R. L. (1996). More powerful tests from confidence interval p values. *Amer. Statist.*, **50**, 314-318.

Berkson, J. (1978). In dispraise of the exact test. *J. Statist. Planning and Inference*, **2**, 27-42.

Birnbaum, A. (1962). On the foundations of statistical inference (with discussion). *J. Am. Statist. Assoc.*, **57**, 269-326.

Boschloo, R. D. (1970). Raised conditional level of significance for the 2×2 table when testing the equality of probabilities. *Statistica Neerlandica*, **24**, 1-35.

Chernoff, H. (1954). On the distribution of the likelihood ratio. *Ann. Math. Statist.*, **25**, 573-578.

Cox, D. R. and Snell, E. J. (1989). *The Analysis of Binary Data. 2nd Edition*, London: Chapman and Hall.

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. *Ann. Math. Statist.*, **11**, 427-444.

Feller, W. (1968). *An Introduction to Probability Theory and its Applications. 3rd Edition*, New York: Wiley.

Fisher, R. A. (1922). On the interpretation of χ^2 from contingency tables, and the calculation of *P*. *J. Royal Statist. Soc.*, **85**, 87-94.

Fisher, R. A. (1925) (5th ed., 1934; 10th ed., 1946). *Statistical Methods for Research Workers*, Edinburgh: Oliver & Boyd.

Fisher, R. A. (1935). The logic of inductive inference. *J. Royal Statist. Soc., A*, **98**,

39-54.

Fisher, R. A. (1962). Confidence limits for a cross-product ratio. *Australian Journal of Statistics*, **4**, 41.

Gail, M. and Gart, J. J. (1973). The determination of sample sizes for use with the exact conditional test in 2×2 comparative trials. *Biometrics*, **29**, 441-448.

Gokhale, D. V. and Kullback, S. (1978). *The Information in Contingency Tables*. New York: Marcel Dekker.

Goodman, L. A. (1984). *The Analysis of Cross-Classified Data Having Ordered Categories*. Cambridge: Harvard University Press.

Gray, R. M. (1990). *Entropy and information theory*. New York: Springer-Verlag.

Greenland, S. (1991). On the logical justification of conditional tests for two-by-two contingency tables. *Amer. Statist*., **45**, 248-251.

Grizzle, J. E. (1967). Continuity correction in the χ^2 test for 2×2 tables. *Amer. Statist*., **21**, 28-32.

Haber M. (1986). An exact unconditional test for the 2×2 comparative trial. *Psychol. Bull*., **99**, 129-132.

Johnson, N. L. and Kotz, S. (1969). *Discrete Distributions*. New York: Wiley.

Kempthorne, O. (1978). Comments on J. Berkson's paper "In Dispraise of the Exact Test". *J. Statist. Planning and Inference*, **3**, 199-213.

Kendall, M. G. and Stuart, A. (1979). *The Advanced Theory of Statistics. 4th Edition*, Vol. 2, London: Charles Griffin.

Kou, S. G. and Ying, Z. (1996). Asymptotics for a 2×2 table with fixed margins. *Statistica Sinica*, **6**, 809-829.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *Ann. Math. Statist*., **22**, 79-86.

Lancaster, H. O. (1949). The combination of probabilities arising from data in discrete distributions. *Biometrika*, **36**, 370-382, Corrig. **37**, 452.

Lancaster, H. O. (1969). *The Chi-squared Distributions*. New York: Wiley.

Lehmann, E. L. (1986). *Testing Statistical Hypotheses. 2nd Edition*, New York: Wiley.

Lindley, D. V. (1956). On a measure of the information provided by an experiment. *Ann. Math. Statist*., **27**, 986-1005.

Little, R. J. A. (1989). Testing the equality of two independent binomial proportions. *Amer. Statist*., **43**, 283-288.

Mehta, C. R. and Patel, N. R. (1980). A network algorithm for the exact treatment of the $2 \times K$ contingency table. *Comm. Statist., Ser. B*9(6), 649-664.

Neyman, J. and Pearson, E. S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference. *Biometrika*, **20**A, 263-274.

Pearson, E. S. (1947). The choice of statistical tests illustrated on the interpretation of

data classed in a 2×2 table. *Biometrika*, **34**, 139-167.

Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Phil. Mag., Series* 5, **50**, 157-175.

Pearson, K. (1904). Mathematical contributions to the theory of evolution XIII: On the theory of contingency and its relation to association and normal correlation. *Draper's Co. Research Memoirs, Biometric Series*, no. 1. (Reprinted in *Karl Pearson's Early Papers*, ed. E. S. Pearson, Cambridge: Cambridge University Press, 1948.)

Plackett, R. L. (1964). The continuity correction in 2 x 2 tables. **51**, 327-337.

Plackett, R. L. (1977). The marginal totals of a 2×2 table. *Biometrika*, **64**, 37-42.

Rao, C. R. (1973). *Linear Statistical Inference and Its Applications*. 2nd edition, New York: Wiley.

Santner, T. J. and Duffy, D. E. (1989). *The Statistical Analysis of Discrete Data*. New York: Springer-Verlag.

Suissa, S. and Shuster, J. (1985). Exact unconditional sample sizes for the 2×2 binomial trial. *J. Royal Statist. Soc., A*, **148**, 317-327.

Upton, G. J. G. (1982). A comparison of alternative tests for the 2×2 comparative trial. *J. Royal Statist. Soc., A*, **145**, 86-105.

Upton, G. J. G. (1992). Fisher's exact test. *J. Royal Statist. Soc., A*, **155, 395-402.**

Wilks, S. S. (1935). The likelihood test of independence in contingency tables. *Ann. Math. Statist.*, **6**, 190-196.

Yates, F. (1934). Contingency tables involving small numbers and the χ^2 test. *J. Royal Statist. Soc., Suppl.*, **1**, 217-235.

Yates, F. (1984). Tests of Significance for 2×2 contingency tables (with discussion). *J. Royal Statist. Soc., A*, **147**, 426-463.

Yule, G. U. (1911). *An Introduction to the Theory of Statistics*. London: Griffin.

Appendix

Proofs for Lemmas 2 and 3 will be carried out by a naive approach using assumptions under the conditioning argument. For ease of exposition, the following notation for the 2×2 table will be used:

	A	$\bar{A}$
Group 1	a	b
Group 2	c	d

Proof of Lemma 2. Suppose that the above table defines the fourfold vector $X = (a, b; c, d)$ with $X' = (a', b'; c', d')$ and $Y = (a^*, b^*; c^*, d^*)$ analogously defined. By the basic identity $a\log(a/a^*) = a[\log(a/a') + \log(a'/a^*)]$, (3.1) is valid if the next equation holds:

$$\begin{aligned} & a\log(a'/a^*) + b\log(b'/b^*) + c\log(c'/c^*) + d\log(d'/d^*) \\ = \; & a'\log(a'/a^*) + b'\log(b'/b^*) + c'\log(c'/c^*) + d'\log(d'/d^*). \end{aligned} \qquad \text{(A.1)}$$

Using the common odds ratio $a'd'/b'c' = \psi = a^*d^*/b^*c^*$, it is found that (A.1) is equivalent to the following equation:

$$\begin{aligned} & a\log\left(\frac{a'/a^*}{b'/b^*}\right) + (a+b)\log(b'/b^*) + c\log\left(\frac{c'/c^*}{d'/d^*}\right) + (c+d)\log(d'/d^*) \\ = \; & a'\log\left(\frac{a'/a^*}{b'/b^*}\right) + (a'+b')\log(b'/b^*) + c'\log\left(\frac{c'/c^*}{d'/d^*}\right) + (c'+d')\log(d'/d^*). \end{aligned} \qquad \text{(A.2)}$$

It is seen that the four log terms with large brackets are equal due to the common odds ratio ψ and $a + c = a' + c'$ since $(X; P)$ and $(X'; P')$ have the same margins. Likewise, as $a + b = a' + b'$ and $c + d = c' + d'$, (A.2) holds and the proof is complete.

Proof of Lemma 3 (A revision of Lemma 3, *Statistica Sinica*, 2008). It suffices to prove the case $J = 3$ without loss of generality. To fix notation, let X be a 2×3 table with the first row (a_{11}, a_{12}, a_{13}) and the second row (a_{21}, a_{22}, a_{23}), likewise, Y has its first row $(a_{11}^*, a_{12}^*, a_{13}^*)$ and second row $(a_{21}^*, a_{22}^*, a_{23}^*)$; and write in short $Z = ((a_{21}^{'}, a_{22}^{'}, a_{23}^{'}),\ (a_{21}^{'}, a_{22}^{'}, a_{23}^{'}))$. By definition, it suffices to prove (4.1) as an equation between three relative divergence terms, where each one is a sum of six log-likelihood ratios. By (A.1), it is equivalent to verifying the following equation between two similar terms which differ only by the entries a_{ij} and $a_{ij}^{'}$:

$$\sum_{i=1}^{2}\sum_{j=1}^{3} a_{ij} \log\left(\frac{a'_{ij}}{a^*_{ij}}\right) = \sum_{i=1}^{2}\sum_{j=1}^{3} a'_{ij} \log\left(\frac{a'_{ij}}{a^*_{ij}}\right). \qquad \text{(A.3)}$$

The assumption on the odds ratios of the tables Y and Z specifies that for $\gamma \neq 0$

$$\frac{(a'_{11}a'_{22} / a'_{12}a'_{21})}{(a^*_{11}a^*_{22} / a^*_{12}a^*_{21})} = \gamma = \frac{(a'_{12}a'_{23} / a'_{13}a'_{22})}{(a^*_{12}a^*_{23} / a^*_{13}a^*_{22})}. \qquad \text{(A.4)}$$

The left-hand-side (l.h.s.) of (A.3) satisfies the equation:

$$\begin{aligned}\sum_{i=1}^{2}\sum_{j=1}^{3} a_{ij} \log\left(\frac{a'_{ij}}{a^*_{ij}}\right) = {} & 2a_{11}\log\gamma + a_{12}\log\gamma + \left\{(a_{11}+a_{12}+a_{13})\log\left(\frac{a'_{13}}{a^*_{13}}\right)\right\} \\ & + \left\{(a_{11}+a_{21})\log\left(\frac{a'_{21}}{a^*_{21}}\right)\right\} + \left\{(a_{12}+a_{22})\log\left(\frac{a'_{22}}{a^*_{22}}\right)\right\} \\ & + \left\{[(a_{13}+a_{23})-(a_{11}+a_{12}+a_{13})]\log\left(\frac{a'_{23}}{a^*_{23}}\right)\right\}. \qquad \text{(A.5)}\end{aligned}$$

And, the r.h.s. of (A.3) also yields (A.5) with the entries a_{ij} replaced by a'_{ij}, because the row margins and column margins of Z are equal to those of data X. It follows that $(2a_{11}+a_{12})\log\gamma = (2a'_{11}+a'_{12})\log\gamma$, but $2a_{11}+a_{12}$ and $2a'_{11}+a'_{12}$ need not be equal, and therefore, $\gamma = 1$ as desired. This completes the proof of Lemma 3, which is an extension of Lemma 2 to $2\times J$ tables.

This proof also indicates that similar results for $I\times 2$ tables, and for $I\times J$ tables, can be useful for the discussion of Section 4.2.

Table 1

$X = (5, 6; 2, 5)$, $N = 18$, upper $CR_X = \{\text{Y: } \psi_Y \geq \psi_X = 2.08\}$

$\mathfrak{X} = \{X_a = (a,\ 11-a;\ 7-a,\ a),\ 0 \leq a \leq 7\}$

Odds ratio ψ	χ_c^2	Yates *p*-value	χ^2	Pearson *p*-value	LR *p*-value	Exact *p*-value
.00	14.04	1.000	18.00	1.000	1.000	1.000
.02	7.59	.997	10.57	.999	1.000	1.000
.09	3.11	.961	5.10	.988	.999	.998
.28	0.60	.780	1.61	.898	.970	.961
.76	0.05	.587	0.08	.609	.788	.780
2.08	0.05	.413	0.51	.237	.398	.417
7.20	1.47	.113	2.92	.044	.087	.112
∞	4.86	.014	7.29	.004	.003	.010

Table 2

$X = (16, 8; 9, 15)$, $N = 48$, upper $CR_X = \{\text{Y: } \psi_Y \geq \psi_X = 3.33\}$

$\mathfrak{X} = \{X_a = (a,\ 24-a;\ 25-a,\ a-1),\ 1 \leq a \leq 24\}$

$\psi > 1$	$\chi_{1,c}^2$	Yates *p*-value	χ_1^2	Pearson *p*-value	LR *p*-value	Exact *p*-value
1.18	0.00	.500	0.08	.386	.500	.500
1.66	0.33	.282	0.75	.193	.278	.282
2.33	1.34	.124	2.09	.074	.119	.124
3.33	3.01	.042	4.09	.022	.038	.041
4.86	5.34	.010	6.76	.005	.009	.010
7.29	8.35	.002	10.11	.001	.001	.002
11.40	12.02	.000	14.11	.000	.000	.000
19.00	16.36	.000	18.78	.000	.000	.000
:	:	:	:	:	:	:
∞	40.40	.000	44.16	.000	.000	.000