Statistica Sinica **11**(2001), 807-826

AN OPTIMALITY THEORY FOR MID p–VALUES IN 2×2 CONTINGENCY TABLES

J. T. Gene Hwang and Ming-Chung Yang

Cornell University and National Central University

Abstract: The contingency table arises in nearly every application of statistics. However, even the basic problem of testing independence is not totally resolved. More than thirty–five years ago, Lancaster (1961) proposed using the mid p–value for testing independence in a contingency table. The mid p–value is defined as half the conditional probability of the observed statistic plus the conditional probability of more extreme values, given the marginal totals. Recently there seems to be recognition that the mid p–value is quite an attractive procedure. It tends to be less conservative than the p–value derived from Fisher's exact test. However, the procedure is considered to be somewhat ad–hoc.

In this paper we provide theory to justify mid p–values. We apply the Neyman–Pearson fundamental lemma and the *estimated truth approach*, to derive optimal procedures, named *expected p–values*. The estimated truth approach views p–values as estimators of the truth function which is one or zero depending on whether the null hypothesis holds or not. A decision theory approach is taken to compare the p–values using risk functions. In the one–sided case, the expected p–value is exactly the mid p–value. For the two–sided case, the expected p–value is a new procedure that can be constructed numerically. In a contingency table of two independent binomial samplings with balanced sample sizes, the expected p–value reduces to a two–sided mid p–value. Further, numerical evidence shows that the expected p–values lead to tests which have type one error very close to the nomial level. Our theory provides strong support for mid p–values.

Key words and phrases: Estimated truth approach, Fisher's exact test, expected p–value.

1. Introduction

Perhaps one of the simplest problems in statistics, yet one which remains controversial, is testing independence in a 2×2 contingency table. There are many procedures proposed in the literature and not much conclusive study as to their worth. In this paper, we exhibit one theory that leads decisively to an optimal procedure.

For further discussion, let y_{ij} and p_{ij} be as layed out as follows:

			Row total		
	y_{11}	y_{12}	n_1	p_{11}	p_{12}
	y_{21}	y_{22}	n_2	p_{21}	p_{22}
Column total	c	d			

We deal with three sampling schemes. In the first scheme, y_{11} and y_{21} are assumed to be two independent binomial observations with sizes n_1 and n_2 and success probabilities p_{11} and p_{21}, respectively. Also, $p_{12} = 1 - p_{11}$, $p_{22} = 1 - p_{21}$, $y_{12} = n_1 - y_{11}$, and $y_{22} = n_2 - y_{21}$. The second sampling scheme involves a multinomial observation $y = (y_{11}, y_{12}, y_{21}, y_{22})$ with total fixed number of cases $N = n_1 + n_2$, and with probability parameters $(p_{11}, p_{12}, p_{21}, p_{22})$. The third sampling scheme assumes that the y_{ij}'s are independent Poisson variables with means p_{ij} not necessarily bounded by 1. In any case, let

$$\theta = (p_{11}/p_{12})/(p_{21}/p_{22}).$$

Consider the one–sided and two–sided test settings:

$$H_0 : \theta \le \theta_0 \quad \text{vs} \quad H_1 : \theta > \theta_0. \tag{1.1}$$

$$H_0 : \theta = \theta_0 \quad \text{vs} \quad H_1 : \theta \ne \theta_0. \tag{1.2}$$

The most interesting and important case is $\theta_0 = 1$.

For the one–sided hypothesis, the most popular p–value is *the normal p–value*

$$P(Z \ge t), \tag{1.3}$$

where Z is a standard normal random variable and t is the realization of

$$T = (\widehat{p}_{11} - \widehat{p}_{21}) \Big/ \sqrt{\widehat{p}(1-\widehat{p})\Big(\frac{1}{n_1} + \frac{1}{n_2}\Big)}, \tag{1.4}$$

where $\widehat{p}_{11} = y_{11}/n_1$, $\widehat{p}_{21} = y_{21}/n_2$ and $\widehat{p} = \frac{y_{11}+y_{21}}{n_1+n_2}$.

Here and below, the α–level test corresponding to a p–value $\gamma(y)$ rejects H_0 if and only if $\gamma(y) \le \alpha$. The advantage of (1.3) is that it applies to general $r \times c$ contingency tables. The disadvantage of (1.3) is that it is not exactly *valid* (see Definition 3.1).

In the case of a 2×2 table, the type one error of the normal p–value converges asymptotically to the nominal level. However, its exact type one error can be twice as large as the nominal level when n_1 and n_2 are moderate and one is much larger than the other. See Hirji, Tan and Elashoff (1991).

Fisher (1934) derived an alternative p–value by conditioning on the marginal totals:

$$P_{\theta_0}(Y_{11} \ge y_{11} \mid \text{marginal totals}) = P_{\theta_0}(HY \ge y_{11}). \tag{1.5}$$

Here Y_{11} is the random variable with the realized value y_{11}. Further, HY represents the random variable having the hypergeometric distribution

$$f_{\theta_0}(t) = \binom{n_1}{t}\binom{n_2}{c-t}\theta_0^{y_{11}} \Big/ \sum \binom{n_1}{y_1}\binom{n_2}{c-y_{11}}\theta_0^{y_{11}}, \tag{1.6}$$

when $\max(0, c - n_2) \le t \le \min(n_1, c)$. In particular,

$$f_1(t) = \binom{n_1}{t}\binom{n_2}{c-t} \Big/ \binom{n_1+n_2}{c}.$$

Note that we need only focus on y_{11} in (1.5), because the data depends only on y_{11} after conditioning on the marginal totals.

The p–value (1.5) amounts to Fisher's exact test. The conditional distribution (1.6) is exact and Fisher's exact test is valid. It is often very conservative, having small type one error. See, for example, Upton (1982) and Hirji, Tan and Elashoff (1991).

There are dozens of alternative tests proposed in the literature, including the *mid p–value* that plays an important role in this paper. For the one–sided test, the mid p–value is defined as

$$P(HY > y_{11}) + \frac{1}{2}P(HY = y_{11}). \tag{1.7}$$

The mid p–value was proposed first by Lancaster (1961) and was endorsed by Plackett (in discussing Yates (1984)), Barnard (1989, 1990), Hirji, Tan and Elashoff (1991), Upton (1992) and Agresti (1992)). Although the mid p–value has nice properties in terms of type I error and power, it has been considered ad–hoc, having little theory attached to it (with the exception of Barnard (1990)).

The two–sided version of the normal p–value (1.3) is the chi–squared p-value proposed by Pearson (1900):

$$P(Z^2 > t^2), \tag{1.8}$$

where Z and t are defined below (1.3). Obviously Z^2 has a chi–squared distribution with one degree of freedom.

There are several two–sided versions of Fisher's p–value. See Section 5.2. According to our study, the following choice outperforms others based on Fisher's conditional distribution:

$$\sum_{\{t: f_{\theta_0}(t) \le f_{\theta_0}(y_{11})\}} f_{\theta_0}(t). \tag{1.9}$$

This is called Fisher's two–sided p–value.

What is the suitable two–sided version of a mid p–value? It seems appropriate to consider

$$\gamma_m(y_{11}) = \sum_{\{t: f_{\theta_0}(t) < f_{\theta_0}(y_{11})\}} f_{\theta_0}(t) + \frac{1}{2} \sum_{\{t: f_{\theta_0}(t) = f_{\theta_0}(y_{11})\}} f_{\theta_0}(t) \tag{1.10}$$

as the *two–sided mid p–value*.

Example 1.1. Assume two binomial experiments yield the following table

	Success	Failures	Row total
pop 1	3	1	4
pop 2	1	3	4
column total	4	4	

The one–sided test with $\theta_0 = 1$ is equivalent to

$$H_0 : p_{11} \le p_{21} \quad \text{vs} \quad H_1 : p_{11} > p_{21}. \tag{1.11}$$

Since the realization of T, in (1.4), is $(3/4-1/4)/(\frac{1}{2}\cdot\frac{1}{2}(\frac{1}{4}+\frac{1}{4}))^{\frac{1}{2}} = \sqrt{2}$, the normal p–value is $P(z \ge \sqrt{2}) = 0.079$. Fisher's p–value (1.5) is $P(HY \ge 3) = \left[\binom{4}{3}\binom{4}{1} + \binom{4}{4}\binom{4}{0}\right] \Big/ \binom{8}{4} = 0.243$ and the mid p–value is $\left[\frac{1}{2}\binom{4}{3}\binom{4}{1} + \binom{4}{4}\binom{4}{0}\right] \Big/ \binom{8}{4} = 0.1285$.

The two–sided hypotheses with $\theta_0 = 1$ reduces to

$$H_0 : p_{11} = p_{21} \quad \text{vs} \quad H_1 : p_{11} \ne p_{21}. \tag{1.12}$$

The chi–squared p–value (1.8) is $P(Z^2 > 2) = 2P(Z > \sqrt{2}) = 0.158$. The two–sided Fisher's p–value (1.9) is .486 whereas the two–sided mid p–value (1.10) is 0.257.

The data are from the well-known Fisher tea tasting experiment (1935). In the experiment, however, all marginal totals are fixed. Here, we assume the binomial model in order to relate to a normal p–value or chi–squared p–value, which make sense only for random marginal totals. The p–values are very different. (The discrepancy will be smaller for larger sample sizes.) The mid p–value falls between the conservative Fisher's p–value and the "radical" normal p–value or chi–squared p–value which, as demonstrated in this example, typically happens. It seems important to develop a systematic way to choose an "optimal" p–value.

In this paper, we take an approach called the *estimated truth approach*. See Hwang and Pemantle (1997) and Blyth and Staudte (1995, 1997). Section 2 gives an introduction to this approach. We apply the Rao–Blackwell Theorem to derive an optimal "p–value" which is called the *expected p–value*. It turns out, for the one–sided test, the expected p–value is the mid p–value (1.7), see Section 3. For the two–sided test, the expected p–value, in general, can only be evaluated numerically. However, it is exactly equal to (1.10) for two binominal populations with $n_1 = n_2$. See Section 4. Section 5 reports some numerical studies which show that the expected p–value is optimal.

2. Estimated Truth Approach

We briefly discuss the approach that will be used in this paper, namely the estimated truth approach. In it, one views p–values as estimators of the *truth*

indicator function which, by definition, is the *truth indicator function* over the null hypothesis space.

If the problem is to test the one–sided hypotheses (1.11) then the truth indicator function considered is $I(p_{11} \leq p_{21}) = 1$ or 0 depending on whether $p_{11} \leq p_{21}$ or not. Similarly for the two–sided hypotheses (1.12), the truth indicator function considered is $I(p_{11} = p_{21})$. It seems plausible that the p–value can be viewed as an estimator of the truth indicator, because a small p–value indicates that the null hypothesis is unlikely or the indicator function is nearly zero. Similarly a large p–value indicates that the null hypothesis is likely and hence the truth indicator function is one.

In the estimated truth approach, one uses a loss function, $L(I, \gamma(X))$, to evaluate an estimator $\gamma(X)$ of I, where X denotes the data. It seems natural to impose two basic requirements on L:

$$L(0, \gamma(X)) \text{ is increasing in } \gamma(X), \text{ and } L(1, \gamma(X)) \text{ is decreasing in } \gamma(X), \quad (2.1)$$

and these are assumed to hold throughout the paper. A special case of a loss function that satisfies (2.1) is the *squared error loss*

$$(I - \gamma(X))^2, \quad (2.2)$$

which can be justified as a proper loss function from a Bayesian point of view, see Hwang and Pemantle (1997). We shall, however, take a frequentist decision theory approach below. As in the usual decision theory, one then tries to find the estimator that minimizes, in some sense, the risk function

$$R(\theta, \gamma(X)) = E_\theta L(I, \gamma(X)). \quad (2.3)$$

In simple settings without nuisance parameters, the estimated truth approach has been applied to evaluate p–values in Hwang, Casella, Robert, Wells and Farrell (1992). Two parallel approaches are the estimated confidence approach and the estimated loss approach. The former approach, initiated in Berger (1985), addresses the problem of estimating $I(\theta \in C_X)$ for a given confidence set C_X. The latter approach, most recently studied by Lindsay and Li (1997), focusses on estimating the loss of a given estimator. Related papers are included in the references. A well written review of these three approaches is in Goutis and Casella (1995).

3. One–sided Test

3.1. General result

We begin with some definitions. The size of a test is the supremum of the type one error over the null hypothesis space. A test has α *level* if its size is bounded above by α.

Definition 3.1. An estimator $\gamma(X)$ is said to be *test–valid* if for every α, $0 < \alpha < 1$, the test that rejects if and only if $\gamma(X) \leq \alpha$ has level α. Furthermore, if the test has size α, then $\gamma(X)$ is said to be *test–exact*.

Below we focus on the case where $f_\theta(X)$ has a monotone likelihood ratio in X, where both X and θ are one–dimensional. In such a case the uniformly most powerful α level test for the one–sided test

$$H_0 : \theta \leq \theta_0 \qquad H_1 : \theta > \theta_0 \tag{3.1}$$

exists by Theorem 2 on page 78 of Lehmann (1997). Furthermore, it is given by the *critical function* (corresponding to a randomized test):

$$\begin{aligned} \varphi_M(X) &= 1 \quad X > c \\ &= \gamma \quad X = c \\ &= 0 \quad X < c \end{aligned} \tag{3.2}$$

where c and γ are chosen so that

$$E_{\theta_0}\varphi_M(X) = \alpha. \tag{3.3}$$

(A critical function of a test gives the probability of rejecting H_0.) Hence the power of a critical function $\varphi(X)$, i.e.,

$$\beta(\theta) = E_\theta\varphi(X) \tag{3.4}$$

is maximized by $\varphi(X) = \varphi_M(X)$ for $\theta > \theta_0$ among all tests of level α.

Below we focus on the case where X takes only integer values. Let U be a random variable uniformly distributed over $[0, 1]$ which is independent of X. The statistical problems based on observing X or $Z = X + U$ are equivalent. In particular, if we observe Z, then almost surely we have $X = [Z]$. Here and later, for any number a, $[a]$ denotes the largest integer less than a. We consider a test that rejects H_0 if

$$Z = X + U > k, \tag{3.5}$$

where k is chosen so that the test has size α. This corresponds to a test of the form (3.2), where $c = [k]$ and $\gamma = 1 - (k - [k])$, and hence is UMP.

Let $z = x + u$ where z, x and u are realizations of Z, X and U. As in Lehmann (1997, p.70), we define the p–value corresponding to a sequence of tests as the smallest type one error under which the corresponding test rejects H_0. The p–value corresponding to the randomized test (3.5) is

$$\gamma_R(z) = P_{\theta_0}(Z > z). \tag{3.6}$$

Putting these together with Theorem 2 on p. 78 of Lehmann (1997), we have the following theorem. The result seems to be known, but we have not found it in the literature.

Theorem 3.2. *Assume that L satisfies* (2.1). *The estimator $\gamma_R(Z)$ has the minimum risk function for every $\theta > \theta_0$ among all the test–valid estimators. That is, for every $\theta > \theta_0$,*

$$E_\theta L(I(\theta \le \theta_0), \gamma_R(Z)) \le E_\theta L(I(\theta \le \theta_0), \gamma(Z)), \tag{3.7}$$

as long as $\gamma(Z)$ is test–valid. Furthermore, if $\gamma(Z)$ is test–exact and its maximum type one error occurs at $\theta = \theta_0$, i.e.,

$$\max_{\theta \le \theta_0} P(\gamma(Z) \le \alpha) = P_{\theta_0}(\gamma(Z) \le \alpha) = \alpha, \tag{3.8}$$

then (3.7) *holds for every θ.*

Proof. Assume that $\gamma(Z)$ is test–valid. Compare the two tests with rejection regions $\{\gamma_R(Z) \le \alpha\}$ and $\{\gamma(Z) \le \alpha\}$. The former is a UMP α level test and the latter has level α as well. Hence $P_\theta(\gamma_R(Z) \le \alpha) \ge P_\theta(\gamma(Z) \le \alpha)$ for $\theta > \theta_0$. This implies that $\gamma_R(Z)$ is stochastically smaller or equal to $\gamma(Z)$ which, in turn, implies that $L(0, \gamma_R(Z))$ is stochastically smaller or equal to $L(0, \gamma(Z))$ by (2.1). Consequently (3.7) holds.

To establish the second assertion, all we need to do is to establish (3.7) for $\theta \le \theta_0$. By Theorem 2 (iv) on p.79 of Lehmann (1997), $P_\theta(\gamma(Z) \le \alpha)$ is minimized by the UMP test $\{\gamma_R(Z) \le \alpha\}$ as long as the second equation in (3.8) holds. Hence $P_\theta(\gamma_R(Z) \le \alpha) \le P_\theta(\gamma(Z) \le \alpha)$, for $\theta \le \theta_0$, and $\gamma_R(Z)$ is stochastically larger or equal to $\gamma(Z)$. Since $L(1, \cdot)$ is decreasing by (2.1), this implies that $E_\theta L(1, \gamma_R(Z)) \le E_\theta L(1, \gamma(Z))$ for $\theta \le \theta_0$.

3.2. Expected p–value

Thus far we have been discussing results mostly relating to randomized rules. Even though their corresponding p–value has the optimality described in Theorem 3.2, a randomized p–value is subject to criticisms. The main criticism is that the experimenter cannot base the decision on the observation only but is, in some situations, forced to toss a die or use computer randomization to reach a final decision. This does not seem reasonable. Therefore we propose a non–randomized p–value below.

From (3.6) we have $\gamma_R(z) = P_{\theta_0}(Z > z) = P_{\theta_0}(X > x) + P_{\theta_0}(X = x$ and $U > u)$. By independence of X and U, we conclude $\gamma_R(z) = P_{\theta_0}(X > x) + (1-u)P_{\theta_0}(X = x)$. Taking the conditional expectation of $\gamma_R(z)$ with respect to U while x is fixed leads to $\gamma_E(x) = P_{\theta_0}(X > x) + \frac{1}{2}P_{\theta_0}(X = x)$, which is called the expected p–value. We have the following theorem.

Theorem 3.3. *In addition to* (2.1)*, assume that* $L(I,\cdot)$ *is convex. Then*

$$E_\theta L(I(\theta \le \theta_0), \gamma_E(X)) \le E_\theta L(I(\theta \le \theta_0), \gamma(Z)) \tag{3.9}$$

for all $\theta > \theta_0$ *and any test–valid estimator* $\gamma(Z)$*. If* $\gamma(Z)$ *is test–exact and satisfies* (3.8)*, then* (3.9) *holds for every* θ*. Strict inequality holds in* (3.9) *if* $L(I,\cdot)$ *is strictly convex. In particular, for squared error loss and* $\theta = \theta_0$*,*

$$E_{\theta_0}(I(\theta \le \theta_0), \gamma_E(X))^2 < \frac{1}{3} = E_{\theta_0}(I(\theta \le \theta_0) - \gamma(Z))^2. \tag{3.10}$$

Proof. The theorem follows if we show that $\gamma_E(X)$ dominates $\gamma_R(Z)$. This follows easily from Jensen's inequality

$$\begin{aligned} E_\theta L(I(\theta \le \theta_0, \gamma_R(Z)) &= E_\theta E[L(I(\theta \le \theta_0), \gamma_R(Z))|X] \\ &\ge E_\theta L[I(\theta \le \theta_0), E(\gamma_R(Z)|X)] \\ &= E_\theta L[I(\theta \le \theta_0), \gamma_E(X)]. \end{aligned}$$

Also, by Jensen's inequality, the risk of $\gamma_E(X)$ is strictly smaller than $\gamma_R(Z)$ if $L(I,\cdot)$ is strictly convex. For (3.10), note that the squared error loss is strictly convex and hence (3.9) holds with strict inequality when L is replaced by squared error loss. Furthermore, $\gamma(Z)$ is uniformly distributed (see e.g., Lehmann (1997, p.170)) for $\theta = \theta_0$. Hence its risk with respect to the squared error loss is $\frac{1}{3}$, establishing (3.10).

In the above derivation, we use a uniform random variable U. It turns out that any continuous random variable supported on $[0,1]$ as its support leads to the same estimator γ_E.

Is $\gamma_E(X)$ test–valid? The answer is, unfortunately, no. Otherwise, for a strictly convex loss function, Theorem 3.3 leads to a strict inequality in (3.9) and contradicts (3.7). It is, however, very close to being test–valid. In particular a test–valid estimator has expectation $\frac{1}{2}$, which is the expectation of $\gamma_E(X)$.

To argue further for the assertion that $\gamma_E(X)$ is nearly test–valid, we note that $\gamma(X)$ is test–valid if and only if

$$E_\theta h(\gamma(X)) \le Eh(U),\ \theta \le \theta_0, \tag{3.11}$$

for every nonincreasing function h. Here U represents, as before, a uniform random variable over $[0,1]$. (Test–validity follows by taking $h(t)$ to be the indicator function, i.e., $h(t)$ equals one or zero depending on whether t is smaller than α or not.) Although γ_E does not satisfy (3.11) for every nonincreasing function h, it does satisfy the inequality for every nonincreasing convex function h.

Numerical studies in Figure 2 of Hirji, Tan and Elashoff (1991, p.1148) indicate that $\gamma_E(X)$ is nearly test–valid when applied to a 2×2 contingency table.

In summary, $\gamma_E(X)$ should be preferred to $\gamma_R(Z)$ for two reasons. First, the expected p–value involves no randomness and seems to make more sense in practice. Second, it has smaller risk than $\gamma_R(Z)$ for squared error loss and all other convex nondecreasing losses.

3.3. Application to a 2×2 contingency table

As shown below the estimated truth approach, although unconditional, leads to a p–value based on the distribution of $Y = (y_{11}, y_{12}, y_{21}, y_{22})$ conditional on the marginal totals. To describe the conditional distribution of Y, it suffices to consider y_{11}. Regardless of the sampling scheme, the conditional distribution of y_{11} given the marginal totals of c, d, n_1, and n_2 has the probability function

$$f_\theta(y_{11}) = \binom{n_1}{y_{11}}\binom{n_2}{c-y_{11}}\theta^{y_{11}} \Big/ \sum_{y_{11}} \binom{n_1}{y_{11}}\binom{n_2}{c-y_{11}}\theta^{y_{11}}. \tag{3.12}$$

To describe the optimal solution, we begin with the randomized uniform most powerful unbiased test which is also based on Fisher's conditional distribution (see Tocher (1950) or Lehmann (1997)). Consider $Z = y_{11} + U$, where U is an independent uniform random variable over $[0, 1]$. Similar to the earlier development that led to (3.6), the p–value (based on the randomized test) is

$$\begin{aligned} \gamma_R(z) &= P_{\theta_0}(Z > z \mid \text{marginal totals}) \\ &= \sum_{t > y_{11}} f_{\theta_0}(t) + (1-u) f_{\theta_0}(y_{11}) \end{aligned} \tag{3.13}$$

where $y_{11} = [z]$ and $u = z - y_{11}$. Taking the conditional expectation of $\gamma_R(z)$ with respect to U while fixing y_{11}, gives the expected p–value

$$\gamma_E(y_{11}) = \sum_{t > y_{11}} f_{\theta_0}(t) + \frac{1}{2} f_{\theta_0}(y_{11}). \tag{3.14}$$

This is identical to (1.7) and is exactly the mid p–value of Lancaster (1961).

We call a test *α–level unbiased* if its probability of rejection is at most α under the null hypothesis and its probability of rejection is at least α under the alternative hypothesis. Obviously any α–level unbiased test has size α under a continuity assumption. An estimator $\gamma(Z)$ is said to be *test–unbiased* if for every α the rejection region $\{\gamma(X) \le \alpha\}$ is α–level unbiased.

Theorem 3.4. *Assume the loss function $L(I, \gamma(Z))$ satisfies* (2.1) *and also $L(I, \cdot)$ is convex. Let $\gamma_E(y_{11})$ be as in* (3.14). *Then $EL(I(\theta \le \theta_0), \gamma_E(y_{11})) \le$*

$EL(I(\theta \le \theta_0), \gamma(Z))$, $\forall$ p_{ij}, where the expectation is taken with respect to p_{ij}'s and $\gamma(Z)$ is any test–unbiased estimator. Strict inequality holds in the last inequality if $L(I, \cdot)$ is strictly convex. In particular, $E[(I(\theta \le \theta_0) - \gamma_E(y_{11}))^2] < \frac{1}{3} = E[I(\theta \le \theta_0) - \gamma(Z)]^2$, for all p_{ij}'s such that $\theta = \theta_0$.

Note that Theorem 3.4 holds with respect to binominal sampling, multinominal sampling and the Poisson sampling. This is due to the fact that $\gamma_E(X)$ dominates $\gamma(Z)$ even if the criterion is based on comparing the conditional risks $E[L(I(\theta \le \theta_0), \gamma(Z)) \mid \text{marginal totals}]$.

4. Two–sided p–value

4.1. General result

Assume the distribution of X belongs to a discrete exponential family, i.e.,

$$P(X = x) = f_\theta(x) = \begin{cases} \pi(\theta)h(x)e^{\theta x} & \text{if } x \text{ is an integer} \\ 0 & \text{otherwise.} \end{cases}$$

As in Section 3, we first construct an exact randomized test based on $Z = X + U$, where U has a uniform distribution over $[0, 1]$.

The goal here is to test $H_0 : \theta = \theta_0$ vs. $H_1 : \theta \neq \theta_0$. We first derive the p–value corresponding to the class of UMP unbiased tests. By the argument below (3.5) and by Lehmann (1997, p.135), the α–level UMP unbiased test uses a critical function that can be written as

$$\varphi(Z) = \begin{cases} 1 & \text{if } Z \notin (c_1, c_2) \\ 0 & \text{otherwise} \end{cases} \tag{4.1}$$

where c_1 and c_2 satisfy

$$E_{\theta_0}\varphi(Z) = \alpha \tag{4.2}$$

and $E_{\theta_0}\{[Z]\varphi(Z)\} = E_{\theta_0}([Z])E_{\theta_0}\varphi(Z)$. The latter equation is equivalent to $E_{\theta_0}([Z] - m)\varphi(Z) = 0$ where $m = E_{\theta_0}([Z])$. The equation is, in turn, equivalent to

$$E_{\theta_0}([Z] - m)I_{(c_1,c_2)}(Z) = 0, \tag{4.3}$$

where $I_{(c_1,c_2)}(Z)$ is the indicator function of (c_1, c_2). Note that if $c_1 \neq c_2$ satisfy (4.3) then

$$[c_1] \le m \le [c_2]. \tag{4.4}$$

Otherwise, suppose that one of the inequalities fails, say $m < [c_1]$. This implies that, for $Z > c_1$, $[Z] \ge [c_1] > m$ and hence $[Z] - m > 0$, contradicting (4.3).

For a fixed c_2 satisfying (4.4), let $B_1(c_2)$ be the smallest c_1 such that (4.3) is satisfied. Similarly, for a fixed c_1 satisfying (4.4), let $B_2(c_1)$ be the largest c_2

satisfying (4.3). Now we are ready to state Theorem 4.1, which is proven in the Appendix.

Theorem 4.1. *The p–value $\gamma_R(z)$ corresponding to a sequence of UMP unbiased randomized tests is*

$$\gamma_R(z) = P_{\theta_0}(Z \notin (c_1, c_2)), \tag{4.5}$$

where c_1 and c_2 depend on z as follows:

$$[z] > m, \text{ then } c_1 = B_1(z) \text{ and } c_2 = z;$$
$$[z] < m, \text{ then } c_1 = z \text{ and } c_2 = B_2(z); \tag{4.6}$$
$$[z] = m, \text{ then } c_1 = m \text{ and } c_2 = z. \tag{4.7}$$

Similar to Theorem 3.2, the next theorem can be established and hence the proof is omitted.

Theorem 4.2. *Assume that the loss function $L(I, \gamma(Z))$ satisfies* (2.1) *and that $L(I, \cdot)$ is convex. Then*

$$E_\theta L(I(\theta = \theta_0), \gamma_R(Z)) \leq E_\theta L(I(\theta = \theta_0), \gamma(Z)) \tag{4.8}$$

for all θ and for any test–unbiased estimator $\gamma(Z)$.

4.2. Expected p–value

The p–value defined in Theorem 4.1 corresponds to a randomized test. It is desirable to come up with a p–value corresponding to a nonrandomized test. We propose to use the expected p–value

$$\gamma_E(x) = \int_0^1 \gamma_R(x + u)du, \tag{4.9}$$

which can be numerically evaluated. Similar to Theorem 3.3, we have the following theorem whose proof is omitted.

Theorem 4.3. *Assume that $L(I, \gamma(Z))$ satisfies* (2.1) *and also that $L(I, \cdot)$ is convex. Then*

$$E_\theta L(I(\theta = \theta_0), \gamma_E(X)) \leq E_\theta L(I((\theta = \theta_0), \gamma(Z)) \tag{4.10}$$

for θ and for any test–unbiased estimator $\gamma(Z)$. Strict inequality holds in (4.10) *if $L(I, \cdot)$ is strictly convex. In particular, for squared error loss and $\theta = \theta_0$,*

$$E_{\theta_0}(I(\theta = \theta_0) - \gamma_E(X))^2 < \frac{1}{3} = E_{\theta_0}(I(\theta = \theta_0) - \gamma(Z))^2. \tag{4.11}$$

4.3. Application to a 2×2 contingency table

Similar to Theorem 4.1, the UMP unbiased tests (e.g., Lehmann (1997, p.155)) give the corresponding p–value

$$\gamma_R(z) = P(Z \notin (c_1, c_2)). \tag{4.12}$$

The Z in (4.1) and (4.12) is defined as $Z = Y_{11} + U$ where Y_{11} has probability function (3.12) and U is a uniform random variable independent of Y_{11}. Furthermore $z = y_{11} + u$ is the realization of $Z = Y_{11} + U$, c_1 and c_2 are defined in (4.6), and B_1 and B_2 are defined right before Theorem 4.1. In other words, all the expectations and probabilities are calculated with respect to the conditional distribution given the marginal totals. The Rao–Blackwellized p–value is

$$\gamma_E(y_{11}) = \int_0^1 \gamma_R(y_{11} + u)du. \tag{4.13}$$

As is demonstrated in the next section, γ_E works very well in risk. One reason is given in Theorem 4.4 (The proof is similar to that of Theorem 4.3 and is omitted).

Theorem 4.4. *The statements in Theorem* 3.4 *hold if* $I(\theta \leq \theta_0)$ *is replaced by* $I(\theta = \theta_0)$ *and* $\gamma_E(y_{11})$ *refers to* (4.13) *instead of* (3.14). *Further, the results apply to binominal sampling, multinominal sampling, and Poisson sampling.*

In general, we do not have an explicit expression for $\gamma_E(y_{11})$ and numerical integration is applied in practice. However, for independent binomial sampling, we may show that, for $\theta_0 = 1$ and $n_1 = n_2$,

$$\gamma_E(y_{11}) = \gamma_m(y_{11}), \tag{4.14}$$

where $\gamma_m(y_{11})$ is defined in (1.10) and is the two–sided mid p–value, also known as quasi–exact p–value (Hirji, Tan and Elashoff (1991)). The proof of (4.14) is not included here. It will be reported elsewhere, see Yang, Lee and Hwang (2000).

5. Numerical Comparisons

In this section, we present the risk functions of several p–values. We focus on squared error loss and 2×2 contingency tables sampled from two independent binominal distributions. All numerical results reported are obtained using exact calculations. Based on the risk functions, we conclude that the expected p–value performs better or as well as all the other p–values. We expect that similar conclusions hold for other sampling distributions, such as multinomial distributions and Poisson distributions, due to the fact that Theorems 3.4 and 4.4 hold for these distributions as well.

5.1. One–sided case

For the one–sided hypothesis (1.1) with $\theta_0 = 1$, we calculate the risk functions of the expected p–value (i.e., mid p–value, (3.14)), Fisher's p–value (1.5), the normal p–value (1.3), and the *exact normal p–value* to be described below.

According to the numerical studies, the normal p–value and the expected p–value perform the best among all the procedures we consider. Both are much better than Fisher's test. The normal p–value, however, has type I error substantially larger than the nominal level. See Hirji, Tan and Elashoff (1991).

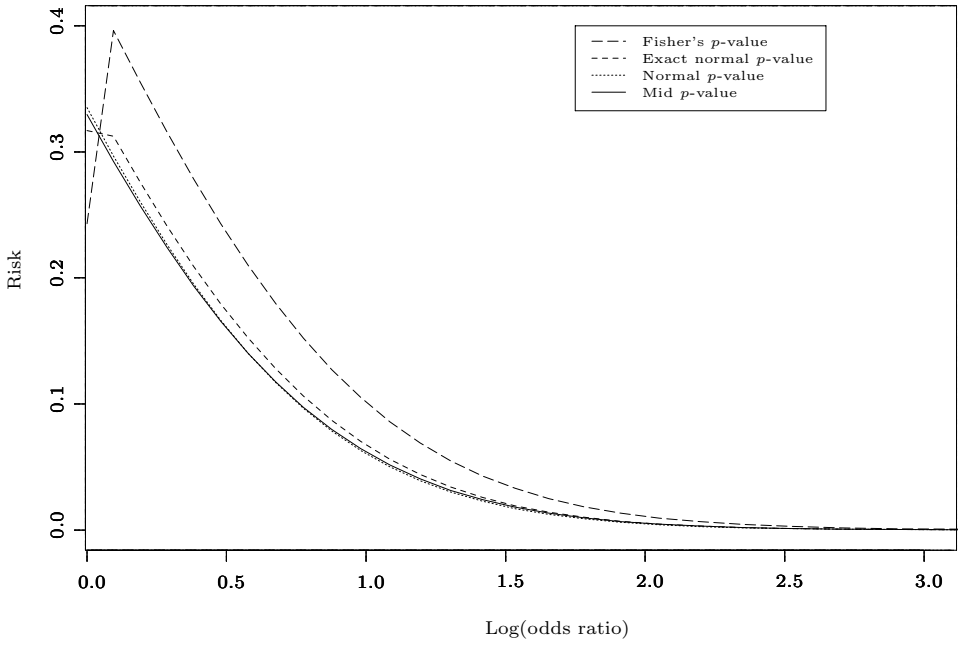

Figure 1. (One-sided test) The risk of the p-values plotted against the natural logarithm of odds ratio based on $p_{11} = 0.3 + 0.01t$ & $p_{21} = 0.3 - 0.01t$, $t = 0, \ldots, 20$, and $n_1 = n_2 = 20$.

We also consider the *exact normal p–value* which is valid, $P_{\theta_0}(T \geq t \mid$ marginal totals), where T and t are defined in and around (1.4). One might think that since this is a conditional procedure, it should do better than (1.3), perhaps because theorems such as Theorem 3.3 prove that optimal estimators are based on Fisher's conditional distribution. However, the numerical results in Figure 1 show the opposite. In fact the exact normal p–value does worse than the normal p–value. This apparently is due to the fact that the exact normal

p–value is based on a discrete distribution whereas its unconditional version, the normal p–value, is based on a continuous distribution.

It is interesting to pay some attention to $\theta_0 = 1$ or $\ln \theta_0 = 0$. The risk function $E_{\theta_0}(1-\gamma(Z))^2$ at such a point for a test–exact estimator $\gamma(Z)$ should be $1/3$. For the mid p–value, the risk is .330, slightly conservative, as drawn in Figure 1, whereas the risk of the normal p–value is 0.335, slightly nonconservative. The other p–values have risk functions far away from $1/3$ and behave quite differently than the test–exact estimator.

Hence these numerical calculations are consistent with Theorem 3.4, which all point to the superiority of γ_E. We did similar calculations for the balanced case $n_1 = n_2 = n$, $n = 10, 40$. We also graphed the risks for the unbalanced cases with $(n_1, n_2) = (10, 20)$ and $(n_1, n_2) = (20, 40)$. The graphs are all similar to Figure 1 but are not reported here. Furthermore, for $\theta = 1$, the risks of the mid p–value are close to $1/3$ in all the cases we examined.

5.2. Two–sided case

As discussed in Agresti (1992, p.135, (a),(b),(c)) three p–values based on the Fisher's conditional argument are proposed: (a) *Fisher's double p–value*, (b) *Fisher's (two–sided) p–value*, and (c) *exact chi–squared p–value*. These p–values will be defined specifically below. *Fisher's double p–value* (a) is

$$\min\{1, 2P(\mathrm{HY} \geq y_{11})\} \tag{5.1}$$

(we make a modification to make sure that it is not greater than one). However, its risk do not fit in Figure 2, where smaller risks are plotted including those of Fisher's p–value (b) a valid test defined in (1.9). As shown in Figure 2, Fisher's p–value performs much worse than the chi–squared p–value (1.8). The expected p–value (4.13) is doing the best for two reasons: it is optimal in the sense of Theorem 4.3; it corrects the "discreteness" effect by taking the expectation with respect to U. It seems that discreteness has a more important effect than conditioning, as evidenced by the poor performance of Fisher's p–value.

One may wonder whether one can improve upon the chi–squared p–value by conditioning. This leads to the *exact chi–squared p–value* (c) defined as $P(|T| \geq |t| \mid \text{all the marginal totals})$, where the conditional probability is evaluated using the hypergeometric distribution. This p–value, however, performs worse than the chi–squared p–value. Apparently, the "discreteness" effect makes the exact chi–squared p–value perform worse even though it may have done the right thing by conditioning.

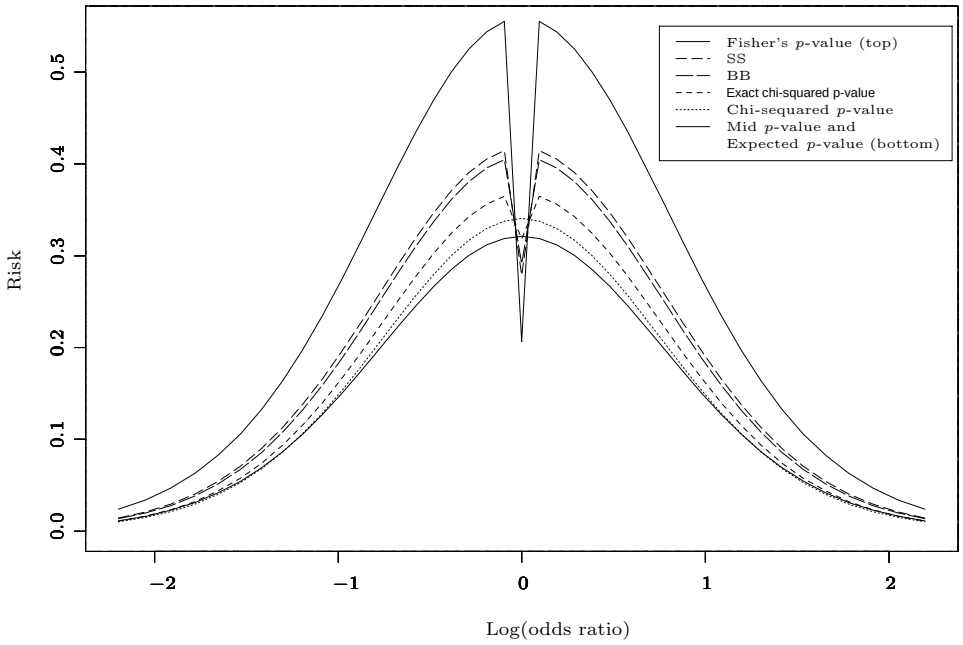

Figure 2. (Two-sided tests) The risk of p-values plotted against the natural logarithm of odds ratio based on $p_{11} = 0.3+$ (or $-$)$0.01t$ & $p_{21} = 0.3-$ (or $+$)$0.01t$, $t = 0, \ldots, 20$, and $n_1 = n_2 = 20$.

We also consider alternative p–values, denoted by SS and BB, which stand respectively for the p–values proposed by Suissa and Shuster (1985) and Berger and Boos (1994). The procedure SS is defined as $\sup_p P_p(|T| \geq |t|)$ where P_p denotes the exact probability evaluated under the two binomial populations with $p_{11} = p_{21} = p$ and T is defined as in (1.4) with the realization t. The p–value BB denotes $\min\left(1, \sup_{p \in C_\beta} P_R(|T| \geq |t|) + \beta\right)$ where $\beta = 0.001$ and C_β stands for a $1 - \beta = .999$ confidence interval for p. For our studies with $n_1 = n_2 = n$, ≤ 20, BB has a risk almost identical, but slightly smaller, than SS. For larger n, the difference may become larger. Fisher's p–value, SS, and BB have the advantage of being test–valid. The risks of these three p–values, however, are not good.

The p–value with the smallest risk function is the expected p–value. It is calculated by numerically averaging 1000 p–values $\gamma_R(y_{11} + u)$ where u is taken to be 1000 equal space points in $[0, 1]$. Note that, as demonstrated in Figure 2 for the balanced case and Figure 3 for the unbalanced case, it has the smallest risk functions for $\theta \neq 1$; also its risk at $\theta = 1$ is 0.321 in both Figures 2 and 3, close to the ideal value 1/3. These two cases are, however, not all what we have

considered. Although not reported in this paper, we also calculated numerically the risk for $(n_1, n_2) = (10, 10), (20, 40)$ and for the sequence of (p_{11}, p_{12}) given in the title of Figure 2. All the graphs are similar to Figures 2 or 3. They demonstrate that the risks of the expected p–value is the smallest among p-values discussed, or nearly so.

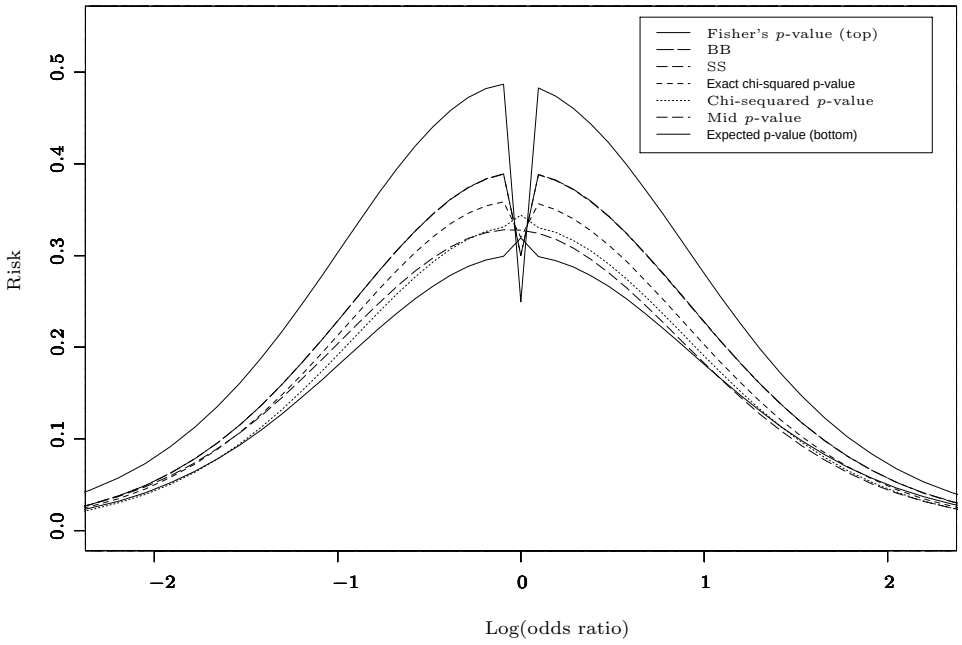

Figure 3. (Two-sided tests) The risk of p-values plotted against the natural logarithm of odds ratio based on $p_{11} = 0.3+$ (or $-$)$0.01t$ & $p_{21} = 0.3-$ (or $+$)$0.01t$, $t = 0, \ldots, 20$, and $n_1 = 10, n_2 = 20$.

Finally, it is desirable to develop a procedure which works well in risk and which is computationally less intensive than $\gamma_E(y_{11})$. The *two–sided mid p–value* $\gamma_m(y_{11})$, defined in (1.10), is such a procedure. For the balanced case in Figure 2, the risk functions of γ_E and γ_m are identical, agreeing with the analytically verified equation (4.14). For the unbalanced case γ_m has the second best risk function, only slightly larger than γ_E. Of course, the advantage of γ_m is its simplicity in form and in computation. Even the calculation for γ_E is still much faster than for SS and BB. Of course, the price we pay in using γ_E is that it is only approximately valid unlike SS and BB.

Actual significance levels. In Figures 4 and 5, we study the actual significance level corresponding to each of the various p–values, i.e., the test that rejects if and only if the p–value is less than α. For $\alpha = .05$, and unequal sample sizes, these pictures graph the *actual significance level* which is the supremum of the type I error over all p_{11} and p_{12} satisyfing H_0. The two figures demonstrate the well–known result that Fisher's p–value is conservative, whereas the chi–square p–value is too radical. See also Hirji, Tan and Elashoff (1991). Note that the expected p–value, BB and SS have actual significance levels closest to $\alpha = 0.05$, much closer than the chi–squared p–value. The mid p–value performs quite reasonably, albeit not as well as the expected p–value. For the balanced case similar results are obtained, although the failure of chi–squared test is not as drastic.

We also numerically calculated the power of these tests in many different cases. Our numerical results, not reported here, show that the tests corresponding to SS, BB, the mid p–value and the expected p–value are all quite similar.

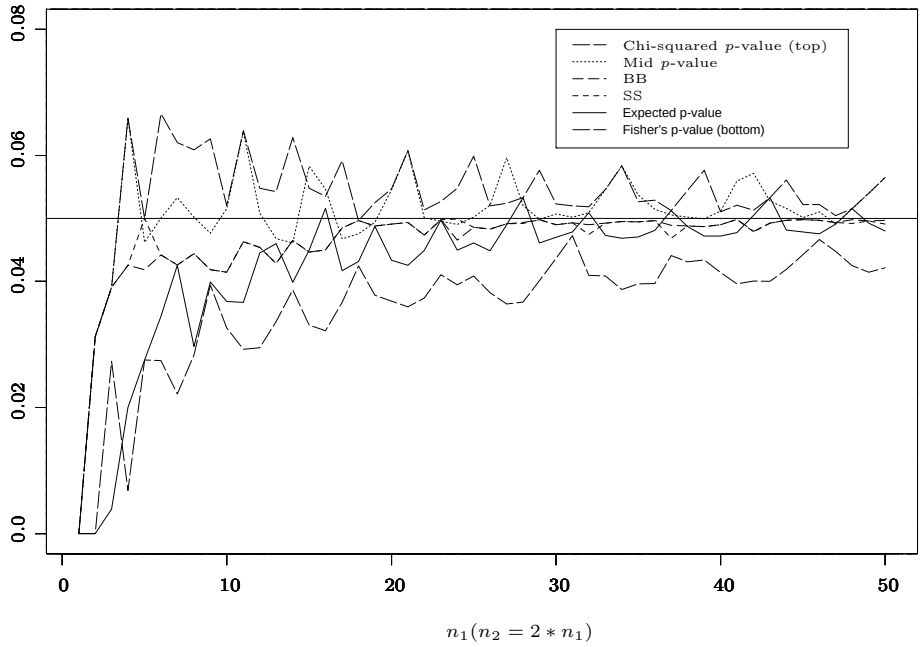

Figure 4. The actual significance levels of two-sided p-values under the nominal level= 0.05.

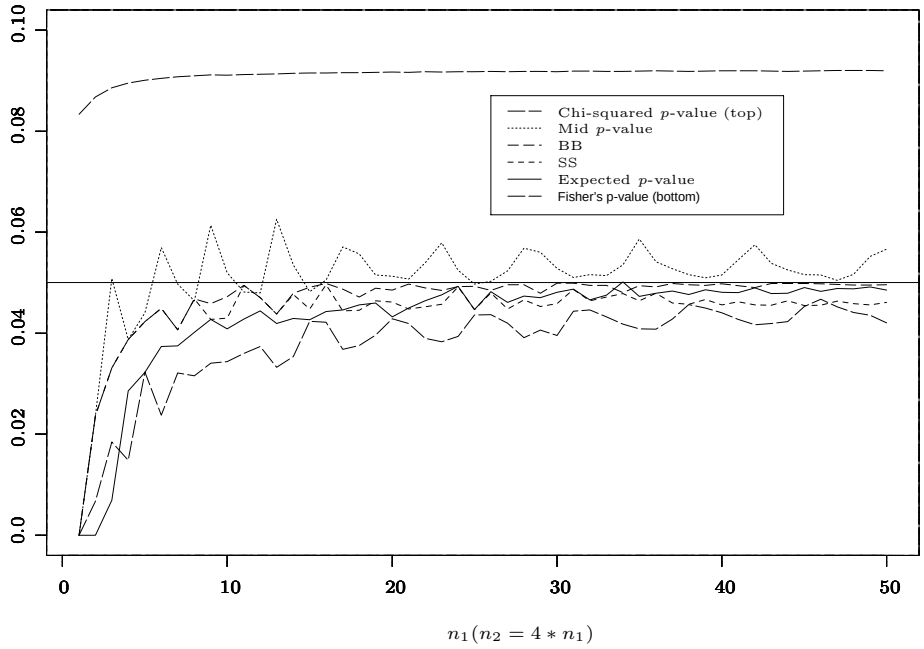

Figure 5. The actual significance levels of two-sided p-values under the nominal level= 0.05.

6. Conclusion

In this paper, we take the estimated truth approach and derive the corresponding optimal procedure which turns out to be the expected p–value. By evaluating the risk functions, we find that the expected p–value performs the best. It dominates both the p–values of the Fisher's exact test and the Pearson's chi–squared test among others. This appears to be the first type of result where one uses a criterion to conclude decisively that a nonrandomized p–value, namely the expected p–value, dominates several other p–values proposed in the literature.

For the one–sided hypothesis, the expected p–value is the mid p–value which has been recognized as a very good p–value in practice.

For the two–sided hypothesis, the expected p–value is new. However, for the balanced two binomial sampling 2×2 table, it reduces to the two–sided mid p–value. Also for the unbalanced case, the expected p–value perform similarly to the mid p–value.

In both cases, we also numerically show that the actual significance level of the expected p–value is very close to the nominal level α. Hence the expected p–value, although not exactly valid, is approximately valid. We feel that this is

due to the fact that the expected p–value, although nonrandomized, mimics the randomized p–value which has the actual significance levels exactly equal to α.

It is interesting that the unconditional estimated truth approach suggests a conditional solution. It also handles the discreteness by Rao–Blackwellization. Handling the discreteness properly can reduce the risk function a lot. In contrast, conditioning has a smaller effect.

In this paper we have provided a theory which strongly supported the mid p–values.

Acknowledgement

It is the authors' pleasure to thank Profesor Alan Agresti who pointed out several papers whose procedures were studied in this paper. The authors are grateful to Professors Martin Wells, Charles McCulloch, Roger Farrell and to Mark Jacobs for their encouraging comments during the preparation of this paper and for their careful reading of the previous drafts.

Appendix

Proof of Theorem 4.1. We first assume that $[z] > m$ (the proof for $[z] < m$ is similar and is omitted). For such a case, we prove that the p–value in (4.5) is the smallest p–value among all (c_1, c_2) satisfying (4.3) such that H_0 is rejected, i.e., $z \notin (c_1, c_2)$. This and the assumption that $[z] > m$ imply that

$$z \geq c_2. \tag{A.1}$$

(Otherwise $z \leq c_1$ and $[z] \leq [c_1] \leq m$ by (4.4), which contradicts $[z] > m$.)

It follows directly from (4.3) that $B_1(c_2)$ is decreasing in c_2. Hence $B_1(z) \leq B_1(c_2) \leq c_1$ where the last inequality follows directly from the definition of B_1. The inequality and (A.1) imply that $(B_1(z), z) \supseteq (c_1, c_2)$. Hence $P_{\theta_0}(Z \notin (B_1(z), z)) \leq P(Z \notin (c_1, c_2))$, establishing the assertion for $[z] > m$.

Now for $[z] = m$. Note that in this case, the UMP unbiased test is nonunique, since any pair of $(c_1, c_2) \subset (m, m+1)$ will satisfy (4.3) and will give the same power as long as $c_1 - c_2$ is fixed. In order to define a specific nested sequence of rejection regions, we focus on the rejection regions (c_1, c_2) when $c_1 = m$. (It leads to similar expected p–values for whatever choice of sequence.) Hence, we end up with (4.7).

References

Agresti, A. (1992). A survey of exact inference for contingency tables (with discussion). *Statist. Sci.* **7**, 131-177.

Barnard, G. A. (1989). On alleged gains in power from lower p–values. *Statist. Medicine* **8**, 1469-1477.

Barnard, G. A. (1990). Must clinical trials be large? The interpretation of P–values and the combination of test results. *Statist. Medicine* **9**, 601-614.

Berger, J. O. (1985). The frequentist viewpont and conditioning. In *Proc. Berkeley Conference in Honor of Jersy Neyman and Jack Kiefer* (Edited by L. M. LeCam and R. A. Olsen), vol.1, 15-44. Wadsworth, Monterey.

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

Blyth, C. R. and Staudte, R. G. (1995). Estimating statistical hypothesis. *Probab. Statist. Lett.* **23**, 45-52.

Blyth, C. R. and Staudte, R. G. (1997). Hypothesis estimates and acceptability profiles for 2×2 contingency tables. *J. Amer. Statist. Assoc.* **92**, 694-699.

Fisher, R. A. (1934). *Statistical Methods for Research Workers* (Originally published 1925, 14th edition 1970), Oliver and Boyd, Edinburgh.

Fisher, R. A. (1935). The logic of inductive inference. *J. Roy. Statist. Soc. Ser. A* **98**, 39-54.

Goutis, C. and Casella, G. (1995). Frequentist post–data inference. *Internat. Statist. Rev.* **63**, 325-344.

Hirji, K. F., Tan, S. J. and Elashoff, R. M. (1991). A quasi–exact test for comparing two binominal proportions. *Statist. Medicine* **10**, 1137-1153.

Hwang, J. T., Casella, G., Robert, C., Wells, M. and Farrell, R. (1992). Estimation of accuracy of testing. *Ann. Statist.* **20**, 490-509.

Hwang, J. T. and Pemantle, R. (1997). Estimating the truth indicator function of a statistical hypothesis under a class of proper loss functions. *Statist. Decisions* **15**, 103–128.

Lancaster, H. O. (1961). Significance tests in discrete distributions. *J. Amer. Statist. Assoc.* **56**, 223-234.

Lehmann, E. L. (1997). *Testing Statistical Hypotheses.* 2nd edition. Springer, New York.

Lindsay, B. and Li, B. (1997). On second–order optimality of the observed Fisher information. *Ann. Statist.* **25**, 2172-2199.

Pearson, K. (1900). On a 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. *Philos. Mag. Ser. 5* **50**, 157–175. (Reprinted 1948 in Karl Pearson's Early Statistical Papers, edited by E. S. Pearson, Cambridge University Press).

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

Tocher, K. D. (1950). Extension of the Neyman–Pearson theory of tests to discontinuous variates. *Biometrika* **37**, 130-144.

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

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

Yang, M. C., Lee, D. W. and Hwang, J. T. (2000). The equivalence of mid p–value and expected p–value for testing equality of two balanced binomial proportions. Cornell Statistics Center Technical Report.

Yates, F. (1984). Test of significance for 2×2 contingency tables. *J. Roy. Statist. Soc. Ser. A* **147**, 426-463.

Department of Mathematics, Malott Hall Cornell University, Ithaca, NY 14853-4201, U.S.A.

E-mail: hwang@math.cornell.edu

Graduate Institute of Statistics, National Central University, Chung-Li 32054, Taiwan.

E-mail: yangmc@cc.ncu.edu.tw

(Received October 1999; accepted January 2001)