IMS Collections
Borrowing Strength: Theory Powering Applications –
A Festschrift for Lawrence D. Brown
Vol. 6 (2010) 184–198

DOI: 10.1214/10-IMSCOLL613

Optimality results for mid p–values

Martin T. Wells[*,1]

Cornell University

Abstract: This article examines some classical optimality properties of various mid p–values. It is shown that the usual mid p–value arises naturally via a Rao–Blackwellization argument of a certain type of mid p–value. In $2 \times 2 \times K$ contingency tables and $I \times J$ tables with ordered alternatives it is shown that the type of alternative is crucial in the construction of an improved evidence assessment rule. It turns out that dimension of the sufficient statistic determines the amount of improvement gained by Rao–Blackwellization. In the $2 \times 2 \times K$ contingency tables and the uniform association model it is shown that the Cohen and Sackrowitz (1992) mid p–value can be improved. In each of these examples, the p–values are not based only on sufficient statistics. However, each can be shown to reduce to the usual mid p–value upon Rao–Blackwellization. Furthermore the mid p–value can, more generally, be derived as an optimal procedure for estimating the truth function using a UMP test. Hence the mid p–value serves a special role as a procedure suggested by both Neyman–Pearson and Rao–Blackwell theories as well as from a Bayesian perspective. Mid p–values for the matched pairs sample and logistic regression models are also studied.

1. Introduction

To lessen the effect of conservativeness of tests for discrete data researchers often use a mid p–value. A mid p–value equals half the probability of the observed data, plus the probability of observing data more extreme than the data at hand. Although a mid p–value is usually considered an *ad hoc* measure of evidence, Lancaster [23] and Plackett [29] (in his discussion of [34]) both recommend it as a good compromise between having a conservative test and using randomization on the boundary to eliminate problems of discreteness. Unlike exact tests with the ordinary p–value, a test using a mid p–value does not guarantee that the Type I error rate falls below the nominal level. However, the maximum of its true Type I error is approximately equal to the nominal level and is less conservative than many exact tests.

In this paper it is shown that the mid p–value arises quite naturally as a consequence of some fundamental statistical results, namely the Neyman–Pearson Lemma and the Rao–Blackwell Theorem. It is also demonstrated that a variety of reasonable proposals of evidence assessment can be improved upon by using the mid p–values. Estimators of the *truth indicator function*, the indicator function over

*This paper is dedicated to Larry Brown. Larry has contributed in deep and original ways to both theoretical and applied statistics. The statistics community has been immeasurably enriched by Larry's research contributions and intellectual leadership. This research was partially supported by NSF grant 06-12031 and NIH grant R01-GM083606-01.

[1]Department of Statistical Science, 1190 Comstock Hall, Ithaca, NY 14853 USA, e-mail: mtw1@cornell.edu

Keywords and phrases: hi–squared test, estimated truth functions, $I \times J$ tables with ordered alternatives, logistic regression, mid p–value, Rao-Blackwellization, $2 \times 2 \times K$ tables, uniform association.

AMS 2000 subject classifications: Primary 62C05, 62B05; secondary 62E15.

the null hypothesis space, are considered. For example, consider K 2×2 tables with common odds ratio θ. Suppose one wishes to test $H_0 : \theta \leq \theta_0$ vs $H_1 : \theta > \theta_0$, then the truth indicator function is $I(\theta \leq \theta_0)$. The decision problem is then formulated as finding a good estimator $\gamma(x)$ of the indicator function $I(\theta \leq \theta_0)$ with respect to the loss $L(I(\theta \leq \theta_0, \gamma(x))$. A possible candidate for the loss function is the squared-error loss, $(I(\theta \leq \theta_0) - \gamma(x))^2$. The squared-error loss can be justified on Bayesian grounds as it is a *proper loss.* That is, a proper loss is one such that the minimizer of the posterior expected loss is $\gamma(x) = P(\theta \leq \theta_0 | X = x)$, the natural Bayesian choice. A second justification for the squared-error loss is that under the loss $\gamma(x)$ will be an admissible estimator for the truth function indicator if and only if it is admissible for estimating the posterior probability of the null hypothesis (when the prior is unknown). This result can be established using arguments similar to [8]. Although Bayesian arguments are used to suggest good decision procedures we focus on frequentist properties of the decision rules.

The estimated truth approach to formulate hypothesis testing problems has been used before. In simple settings, without nuisance parameters, the approach has been applied to evaluate p–values in [16, 14, 17, 30, 33], and Blyth and Staudte [5]. In more complex discrete problems Hwang and Yang [18] and [5] apply these ideas. The estimated truth approach can be viewed more broadly as a way of constructing conditional inference procedures. In a series of papers [19, 20, 21] addressed the problem of developing conditional and estimated confidence theories to provide frequentist estimates of confidence. Berger [4] compared the Bayesian and frequentist approaches to this problem.

Interestingly, the estimated truth approach suggests a solution very much related to the Neyman–Pearson theory. Consider the case that one is interested in making inference for a parameter θ and the sampling model for $T|\theta$ has a monotone likelihood ratio in T. Then the uniformly most powerful (UMP) α-level test for $H_0 : \theta \leq \theta_0$ vs $H_0 : \theta > \theta_0$ exists by Theorem 2 of [25, p. 78], and its critical function has the form

$$\varphi(t) = 1 \quad t > \eta \\ = r \quad t = \eta \\ = 0 \quad \text{otherwise,} \tag{1.1}$$

where η and r are chosen such that $E_{\theta_0}\varphi(T) = \alpha$. The focus here is on the case where T is discrete and only takes on integer values. Let U be a $U(0, 1)$ random variable independent of T. Then the two statistical problems based on observing T or $Z = T + U$ are equivalent. If we observe Z, then $T = [Z]$, where $[\cdot]$ is the Gauss function, that is $[Z]$ denotes the largest integer less than Z. The test that rejects H_0 if $Z \geq c$ is a size α UMP test, where $\eta = [c]$ and $r = 1 - (c - [c])$. The p–value for this test is $p(z) = P_{\theta_0}(Z \geq z)$ where $z = t + u, t$, and u represent realizations of Z, T and U. In the one-sided testing problem, $p(t, u) = P_{\theta_0}(T \geq t) + (1 - u)P_{\theta_0}(T = t)$. However unlike the Neyman–Pearson theory, the estimated truth approach does not stop and recommend the randomized p–value, which is unacceptable from a scientific point of view. Note that since $p(t, u)$ is not a function of the sufficient statistic only, the estimated truth approach automatically suggests to Rao–Blackwellize it and end up with an improved p–value, having uniformly smaller risks. Taking the expectation with respect to U leads to the mid p–value $P_{\theta_0}(T \geq t) + \frac{1}{2}P_{\theta_0}(T = t)$. In the case of 2×2 contingency tables Hwang and Yang [18] show that the mid p–value has some decision-theoretic optimality properties and often behaves better than the ordinary exact p–value. In the two-sided problem one can also use a UMP randomized test to derive its p–value as the form $p(z) = P_{\theta_0}(Z \notin (\eta_1(z), \eta_2(z))$, where $\eta_1(z)$ and $\eta_2(z)$ are defined through the test. In this case, by taking an expectation with respect to U we obtain a new "mid p–value" $p_E(t) = \int_0^1 p(t+u)du$, which is the Rao–Blackwellization of $p(z|\ u)$. It is worth noting that constructing an accurate saddlepoint approximation to a mid p–value is feasible since Z and U are continuous and T and U are independent (see [32]).

The Rao–Blackwellization of various mid p–values is examined in Section 2. Note that it follows from the Rao–Blackwell theorem, that the Rao–Blackwellized p–value has a smaller risk. In particular, the $2\times2\times K$ contingency table problem is studied in Section 2.1. The p–values proposed in [10] are applied by Kim and Agresti [22] to try to avoid conservativeness by eliminating possible tables, a form of smoothing. It is shown that Rao–Blackwellizing a certain type of p–value used by Kim and Agresti [22] gives a mid p–value. $I \times J$ tables are considered in Section 2.2. In the $I \times J$ tables with ordered alternatives it is shown that the type of alternative is crucial in the construction of the improved p–value. It turns out that dimension of the sufficient statistic determines the amount of improvement gained by Rao–Blackwellization. In the uniform association model (see [2]) it is shown that the [10] mid p–value can be improved via Rao–Blackwellization. In each of these particular examples, these p–values are not based on sufficient statistics only. However, upon Rao–Blackwellization, each can be shown to reduce to the classical mid p–value. Hence the mid p–value serves a special role as a procedure suggested by both Neyman–Pearson and Rao–Blackwell theories. In Section 3 it is shown that the good properties of the classical mid p–value in Section 2.1 are a consequence of some general theory of estimated truth functions for UMP tests. Other models, these include the matched pairs sample model and logistic regression are discussed in Section 4. The final section gives some concluding remarks.

2. Rao–Blackwellization of p–values

Two particular models are studied in this section, $2\times2\times K$ and $I\times J$ contingency tables with ordered alternatives. It is shown that the Rao–Blackwellization of some type of mid p–values lead to the classical mid p–value.

2.1. $2 \times 2 \times K$ contingency tables

As in Kim and Agresti [22], the observations and probabilities in the k–th table, as illustrated below, are denoted by Y_{ijk} and p_{ijk} respectively, where i and j are either 1 or 2.

$$\begin{array}{lll} Y_{11k} & Y_{12k} & Y_{1+k} \quad p_{11k}\ p_{12k} \\ Y_{21k} & Y_{22k} & Y_{2+k} \quad p_{21k}\ p_{22k} \\ Y_{+1k} & Y_{+2k} & \end{array}$$

The plus notation in Y_{i+k} and Y_{+jk} denotes the sum over the index it replaces. Hence Y_{i+k} and Y_{+jk} are the marginal totals in the k–th table.

In each table, assume two independent binominal distributions with the common odds ratio

$$\theta = \frac{p_{11k}/(1-p_{11k})}{p_{12k}/(1-p_{12k})} \tag{2.1}$$

is independent of k. Under the assumption that the tables are independent, then joint sufficient statistics are the marginal totals Y_{+1k}, and

$$T(Y) = \sum_{k=1}^{K} Y_{11k}. \tag{2.2}$$

Let t be the observed value of T. The usual mid p–value (as defined in [23]), for testing $H_0 : \theta \le \theta_0$ vs $H_1 : \theta > \theta_0$ is

$$p_{mid} = P_{\theta_0}(T > t \mid Y_{+1k}, k = 1, \ldots, K) + \frac{1}{2} P_{\theta_0}(T = t \mid Y_{+1k}, k = 1, \ldots, K),$$

where $P_{\theta_0}(\cdot \mid Y_{+1k}, 1 \le k \le K)$ denotes the conditional probability distribution calculated when $\theta = \theta_0$. Note the conditional distribution of Y_{11k} given Y_{+1k} is a hypergeometric distribution with the hypergeometric discrete probability function $h_k(y_{11k})$

$$\begin{aligned} h_k(y_{11k}) &= P(Y_{11k} = y_{11k} \mid Y_{+1k} = y_{+1k}) \\ &= \binom{y_{1+k}}{y_{11k}} \binom{y_{2+k}}{y_{+1k} - y_{11k}} \Big/ \binom{y_{++k}}{y_{1+k}} \end{aligned}$$

for y_{11k} satisfying $max(0, y_{+1k} - y_{+2k}) \le y_{11k} \le min(y_{1+k}, y_{+1k})$. Otherwise $h_k(y_{11k}) = 0$. Let $h(y) = \prod_k h_k(y_{11k})$, where y is the vector consisting of y_{11k}, $1 \le k \le K$. Let $Z = (Z_1, \ldots, Z_K)$ be the discrete random vector with the probability function $h(z)$. Then the mid p–value can be expressed as

$$p_{mid}(t) = P(T(Z) > t) + \frac{1}{2} P(T(Z) = t), \tag{2.3}$$

where P denotes the probability corresponding to Z and $T(Z) = Z_1 + \cdots + Z_K$. Although (2.3) also depends on the marginal totals, we have suppressed them in writing $p_{mid}(t)$. By adopting the test of Cohen and Sackrowitz [10] originally proposed for $I \times J$ tables to this $2 \times 2 \times K$ model, Kim and Agresti [22] proposed

$$p^*(y) = P(T(Z) > t) + P(h(Z) \le h(y), T(Z) = t),$$

where y denotes the observed vector consisting of y_{11k}, $1 \le k \le K$. Obviously, p^* is less than the Fisher's type procedure $p_F = P(T(Z) \ge t)$ and hence is less conservative. It is interesting to note that p^* and p_F when used for testing are valid. For example, the test that rejects H_0 if and only if $p^*(y)$ is less than or equal to α has type I error less than or equal to α. A mid p–value type of procedure corresponding to p^* as proposed by Kim and Agresti [22], is

$$\begin{aligned} p^*_{mid}(y) = P(T(Z) > t) &+ P(h(Z) < h(y),\ T(Z) = t) \\ &+ \frac{1}{2} P(h(Z) = h(y),\ T(Z) = t). \end{aligned} \tag{2.4}$$

How does p^*_{mid} compare to p_{mid} in terms of evidence evaluation? One interesting feature of p^*_{mid} is that it depends on data more than the sufficient statistic consisting of t, the column sums, and the row sums, and it also depends on the individual y_{11k}, $1 \le k \le K$, as well. One may wonder what is the Rao–Blackwellized version of p^*_{mid}. The following theorem answers the question.

Theorem 2.1. $E(p^*_{mid} \mid T = t, y_{+1k}, 1 \le k \le K) = p_{mid}(t)$.

Proof. See the Appendix. □

Hence Theorem 2.1 implies that the regular mid p–value p_{mid} is preferred to p^*_{mid}. This follows from the direct application of the Rao–Blackwell theorem which is summarized below. We shall consider the loss function $L(I(\theta \in H_0), p(X))$ and its risk function $R(\theta, p) = EL(I(\theta \le \theta_0), p(X))$.

Theorem 2.2. *For each θ, assume that $L(I(\theta \le \theta_0), p(X))$ is a convex function in $p(X)$. Then $R(\theta, p_{mid}) \le R(\theta, p^*_{mid})$ for every θ. If $L(I(\theta \le \theta_0), \cdot)$ is strictly convex, then the last inequality is strict. In particular,*

$$E(p_{mid} - I(\theta \le \theta_0))^2 < E(p^*_{mid} - I(\theta \le \theta_0))^2 \quad \textit{for every } \theta.$$

Theorems 2.1 and 2.2 assume the K independent binomial models. Similar results can be established for K independent multinomial tables as well as for K independent Poisson tables when the "odds ratio" θ in (2.1) are assumed to be common in each table. Here odds ratios are in quotes, since in the Poisson case the p_{ij} in (2.1) are not probabilities but are means. To establish this, only a small change in the technical argument is needed. We condition on all the marginal totals including row totals and column totals. The p–values in question are the same as in (2.3) and (2.4), since in all these tables the conditional distribution of Y_{11k} are independent hypergeometric regardless of which model is assumed.

One may also Rao–Blackwellize $p^*(y)$ and obtain an improved p–value, which leads to

$$p_{mid}(t) + E(h(Z)|T(Z) = t).$$

This follows from the proof of Theorem 2.1. Similar results can be derived for settings relating to Theorem 2.3.

2.2. I × J contingency table with ordered alternatives

For an $I \times J$ contingency table the data is denoted as Y_{ij}, $1 \le i \le I$ and $1 \le j \le J$ and also the cell probabilities are denoted as p_{ij}. It is assumed that Y_{ij} has a multinomial model with parameter $n = \sum_{i,j} Y_{ij}$ the total number of IJ cells and cell probabilities p_{ij}. The testing problem is

$$H_0 : p_{ij} = p_{i+}p_{+j} \text{ vs } H_1 : \theta_{ij} = \log(p_{ij}p_{(i+1)(j+1)}/p_{i(j+1)}p_{(i+1)j}) \ge 0 \tag{2.5}$$

with strict inequality for at least one pair of (i, j). This problem has been considered in [9, 10]. They also commented that the alternative H_1 makes sense if the categories of the contingency table are ordered. The same problem has also been studied in [3, 15, 24, 28]. [10] suggested that the usual Fisher's p–value is

$$P_{H_0}(T(Y) \ge t \mid \text{marginal column and row totals}), \tag{2.6}$$

where

$$S_{ij} = \sum_{1 \le k \le j} \sum_{1 \le \ell \le j} Y_{k,\ell} \text{ and } T(Y) = \sum_{j=1}^{J-1} \sum_{i=1}^{I-1} S_{ij}(Y), \tag{2.7}$$

where $Y = (Y_{11}, \ldots, Y_{IJ})$ and t is the realization of $T(Y)$. Hence $t = T(y)$ and $y = (y_{11}, \ldots, y_{IJ})$ is the realization of Y. Note that under H_0, the conditional probability of $Y = (Y_{11}, \ldots, Y_{IJ})$ given the marginal totals $Y_{+j} = y_{+j}$ and $Y_{i+} = y_{i+}$ has the multivariate hypergeometric probability function

$$h(y) = \Big(\prod_{i=1}^{I} y_{i+} \prod_{j=1}^{J} y_{+j}\Big) \Big/ \Big[n! \prod_{i=1}^{I} \prod_{j=1}^{J} (y_{ij}!)\Big]. \tag{2.8}$$

Let $Z = (Z_{11}, \ldots, Z_{IJ})$ be a random vector with the probability function $h(Z)$ where $Z_{i+} = y_{i+}$ and $Z_{+j} = y_{+j}$. By using this notation, (2.6) can be written as $P(T(Z) \geq t)$.

Cohen and Sackrowitz [10] also came up with a less conservative p–value

$$P(T(Z) > t) + P(T(Z) = t, h(Z) \leq h(y)). \tag{2.9}$$

One may also consider a mid p–value modification of (2.9)

$$P(T(Z) > t) + P(T(Z) = t, h(Z) < h(y)) + \frac{1}{2} P(T(Z) = t, h(Z) = h(y)). \tag{2.10}$$

Can one Rao–Blackwellize (2.10)? To see this, note that the sufficient statistic for the problem under the full model (that is, under the union of H_0 and H_1) is (Y_{ij}, Y_{i+}, Y_{+j}), $1 \leq i \leq I - 1$ and $1 \leq j \leq J - 1$ (see equation (2.3) of [9]). This sufficient statistic, however, is a trivial sufficient statistic, since it is a one–one function of any of the $IJ - 1$ points of Y_{ij} (note that the dimension of Y_{ij} is $IJ - 1$, since $\sum Y_{ij}$ is fixed). Hence the Rao–Blackwellized (2.10) is equals itself and does not lead to anything interesting.

However, we may consider a slightly modified problem by assuming θ_{ij}, defined in (2.5), to be a constant θ. This model is known as the uniform association (see [2]). Consider testing the hypothesis

$$H_0 : \theta = 0 \;\; vs \;\; H_1 : \theta > 0. \tag{2.11}$$

Hence H_0 in (2.11) is equivalent to (2.5). Now from equation (2.3) of [9], we see that T, Y_{i+}, and Y_{+j} is a sufficient statistic under the model that all $\theta_{ij} = \theta$ which is now taken to be a full model. Consequently, one may Rao–Blackwellize (2.10). The theorem below summaries the result.

Theorem 2.3. *Assume the $I \times J$ uniform association model and hence $\theta_{ij} = \theta$. Then the Rao–Blackwellized estimator based on (2.10) is*

$$p_{mid}(t) = P(T(Z) > t) + \frac{1}{2} P(T(Z) = t), \tag{2.12}$$

where T is defined by (2.7). That is, $p_{mid}(t)$ equals the conditional expectation of (2.10) given the sufficient statistic consisting of T and the marginal totals.

As in Theorem 2.2, Theorem 2.3 and the Rao–Blackwell Theorem obviously implies that p_{mid} dominates (2.10) in estimating $I(\theta = \theta_0)$ with respect to a convex loss.

3. A general theory for the mid p–value in $2 \times 2 \times K$ and $I \times J$ tables

As in Section 2, we shall use a loss function $L(I(\theta \leq \theta_0), \gamma(X))$. Also it is assumed that

(3.1) $\quad L(1, \gamma(X))$ decreasing in $\gamma(X)$ and $L(0, \gamma(X))$ increasing in $\gamma(X)$

where $L(1,\cdot)$ or $L(0,\cdot)$ denote $L(I(\theta \le \theta_0), \gamma(X))$ when $I(\theta \le \theta_0) = 1$ or 0, respectively. Assumption (3.1) is reasonable when $\gamma(X)$ is viewed as an estimator of $I(\theta \le \theta_0)$.

We shall first consider the $2\times 2\times K$ table with common "odds ratio" in Section 2, under the three different sampling schemes: (i) Each 2×2 table has a multinomial distribution; (ii) Each table has two independent binomial random samplings; (iii) Each table has four independently Poisson distributed random observations. In all the cases, the conditional distributions given all the marginal totals are independent hypergeometric. In the case of a $2 \times 2 \times K$ table with common θ, the α–level UMP unbiased test for $H_0 : \theta \le \theta_0$ vs $H_1 : \theta > \theta_0$ is to reject with the probability

$$\varphi(T) = 1 \quad \text{if} \quad T > \eta \qquad (3.2)$$
$$= r \quad \text{if} \quad T = \eta$$
$$= 0 \quad \text{otherwise,}$$

where T is as defined in (2.2) and η is a positve integer. See, for example, [24, p. 163], for an outline of Neyman–Pearson's theory for this problem. Note r and η, functions of the marginal totals, are determined so that $E_{\theta_0}[\varphi(T) \mid \text{marginal totals}] = \alpha$.

We introduce a uniform random variable U independent of T and consider a test of this form

$$T + U > c, \qquad (3.3)$$

where c may depend on the marginal totals. This test reduces to (3.1) for $\eta = [c]$ and $r = 1 - (c - [c])$. Note $[\cdot]$ is the Gauss function previously defined just below (1.1). Hence (3.2) is a UMP unbiased test as well. A corresponding p–value is then

$$\begin{aligned} p_R(t,u) &= P_{\theta_0}(T + U > t + u \mid \text{marginal totals}) \\ &= P_{\theta_0}(T > t \mid \text{marginal totals}) + (1-u)P_{\theta_0}(T = t \mid \text{marginal totals}), \end{aligned}$$

where t and u are the realizations of T and U. If we use Rao–Blackwellization in the above expression, we end up with an improved estimator. Hence consider

$$\begin{aligned} &E_{\theta_0}[p_R(t,u) \mid T = t, \text{marginal totals}] \\ &\quad = P_{\theta_0}(T > t \mid \text{marginal totals}) + \frac{1}{2}P_{\theta_0}(T = t \mid \text{marginal totals}). \end{aligned}$$

This expression, of course, reduces to p_{mid} in (2.2). It turns out that the p_{mid} not only performs better than p_R but is also optimal as stated below.

In the following theorem, we shall say an estimator $p(Z)$ to be test–unbiased if for every α the following test procedure is α leveled unbiased H_0 if $p(Z) \le \alpha$. Similar to Theorem 2.3, we have the following theorem whose proof is omitted.

Theorem 3.1. *Assume that the loss function $L(I, p(Z))$ satisfies (3.1) and $L(I,\cdot)$ is convex. Then for any test unbiased estimator $p(Z)$,*

$$EL(I(\theta \le \theta_0), p_{mid}(Z)) \le EL(I(\theta \le \theta_0), p(Z)) \quad \textit{for every } \theta.$$

Furthermore strict inequality holds in the above inequality if $L(I,\cdot)$ is strictly convex. In particular,

$$E(I(\theta \le \theta_0) - p_{mid}(Z))^2 < E(I(\theta \le \theta_0) - p(Z))^2 \quad \textit{for every } \theta.$$

The above theorem is proved based on Neyman–Pearson and Rao–Blackwell theorems which led us to the optimal procedure, the mid p–value. The above theorem, however, does not imply Theorem 2.2, because p^*_{mid} is not test–unbiased. It thus is quite interesting that the Rao–Blackwellization of p^*_{mid} and the approach of the above theorem both point to the classical mid p–value p_{mid}.

For the $I \times J$ uniform association model, we may develop a theorem similar to Theorem 3.1. To see this, the probability function of Y_{ij}, as in [9], is

$$\beta(\theta, b, d) \exp\Big(\theta \sum_{ij}^{I-1,J-1} Y_{ij} + \sum_{i=1}^{I-1} Y_{i+} b_i + \sum_{j=1}^{J-1} Y_{+j} d_j\Big) g(y, m),$$

where $m = (Y_{1+}, \ldots, Y_{I+}, Y_{+1}, \ldots, Y_{+J})$ consists of the marginal totals,

$$\begin{aligned} b &= (b_1, \ldots, b_I), \\ b_i &= \ln(P_{iJ}/P_{IJ}), \\ d_j &= \ln(P_{Ij}/P_{IJ}). \end{aligned} \tag{3.4}$$

Note that when $\theta = 0$, y_{ij}, b_i and d_j, $1 \le i \le I$, $1 \le j \le J$ form a complete sufficient statistic. Hence by Theorem 3 of Lehmann (1986, p. 147), the following randomized test with the critical function φ is UMP

$$\varphi(t) = 1 \quad t > \eta \tag{3.5}$$

$$= r \quad t = \eta \tag{3.6}$$

$$= 0 \quad \text{otherwise}, \tag{3.7}$$

where r and η, depending on m, satisfy $E_0[(\varphi_1(T) \mid m] = \alpha$.

Following the proof of Theorem 3.1, we may conclude that the following theorem holds.

Theorem 3.2. *Assume the same loss function L as in Theorem 3.1. Then for any test unbiased estimator $p(Y)$,*

$$EL(I(\theta = \theta_0), p_{mid}(T)) \le EL(I(\theta = \theta_0), p(Y))$$

for any parameter θ. Furthermore strict inequality holds in the above inequality if $L(I, \cdot)$ is strictly convex. In particular

$$E(I(\theta = \theta_0) - p_{mid}(Y))^2 < E(I(\theta = \theta_0) - p(Y))^2 \quad \textit{for every } \theta.$$

The above theorem holds as well if $I(\theta = \theta_0)$ is replaced by $I(\theta \le \theta_0)$, which corresponds to testing $H_0 : \theta \le \theta_0$ vs $H_1 : \theta > \theta_0$.

One unfortunate fact, from the theoretical point of view, is that the optimal p–value in Theorems 3.1 and 3.2 are not test unbiased. It is interesting, however, that the criterion of estimated truth approach leads us to the discovery of some p–value that dominates the p–value corresponding to the UMP unbiased randomized tests. In contrast, the criterion of comparing power functions leads to the UMP unbiased randomized test as the optimal test even though this randomized test is recognized to be unsatisfactory from the practical and scientific point of view.

Two other advantages of the estimated truth approach are in order. First, since the optimal solution is an improvement over randomized UMP unbiased test, it is likely to produce p–values with type I error very close to the nominal level. This is supported by the many numerical studies including the one by Hwang and Yang

[18] which shows that in a 2×2 contingency table the mid p–value has level very close to the nominal level. The optimal p–value derived in [5] using an estimated truth approach slightly different from our approach has a good risk function but fails to have a good type I error.

4. Other models

4.1. 2×2 matched pair sample

The model is a 2×2 contingency table with observations Y_{ij} satisfying the multinomial distribution with probabilities p_{ij} as laid out in the following tables:

$$\begin{array}{ccccc} Y_{11} & Y_{12} & Y_{1+} & p_{11} & p_{12} \\ Y_{21} & Y_{22} & Y_{2+} & p_{21} & p_{22} \\ Y_{+1} & Y_{+2} & N, & & \end{array}$$

where the plus notation denotes the sum over the index it replaces. However, the hypotheses to be tested are

$$H_0 : p_{12} \le p_{21} \quad \text{vs} \quad H_1 : p_{12} > p_{21}. \tag{4.1}$$

For a recent survey of this model and some new solutions, see [31]. Earlier this problem was treated by McNemar [26] who proposed an asymptotic test. The UMP unbiased test, however, is based on the conditional distribution of Y_{12} given $Y_{12} + Y_{21} = m$, which is binomial (m, θ) where the probability

$$\theta = \frac{p_{12}}{p_{12} + p_{21}}.$$

Using this notation θ, the hypotheses in (4.1) can be rewritten as

$$H_0 : \theta \le \frac{1}{2} \quad \text{vs} \quad H_1 : \theta > \frac{1}{2}.$$

The UMP unbiased test has the randomized critical function

$$\begin{aligned} \varphi(Y_{12}) &= 1 \quad Y_{12} > \eta \\ &= r \quad Y_{12} = \eta \\ &= 0 \quad \text{otherwise}, \end{aligned} \tag{4.2}$$

where η and r, depending on m, are chosen so that $E_{\frac{1}{2}}(\varphi(Y_{12}) \mid Y_{12} + Y_{21} = m) = \alpha$. See [13] and [25], p. 169. The above expectation is taken with respect to the conditional distribution of Y_{12} at $\theta = \frac{1}{2}$, given $Y_{12} + Y_{21} = m$, namely the binomial $(m, \frac{1}{2})$ distribution.

Similar to Section 2, we may define a randomized p–value by using an independent uniform random variable U. Hence the test that rejects if and only if $Y_{12} + U > c$ where c depends on m is UMP since it corresponds to the critical function (4.2) with $\eta = [c]$ and $r = 1 - (c - [c])$.

An argument similar to that as what leads to Theorem 3.1 leads to the mid p–value

$$p_{mid} = P(B > y_{12}) + \frac{1}{2} P(B = y_{12}) \tag{4.3}$$

where B is a random variable with binomial distribution $(m, \frac{1}{2})$ and y_{12} is the realization of Y_{12}. Theorem 3.1 then holds for this situation, where p_{mid} refers to (4.3) and $I(\theta \le \theta_0)$ is replaced by $I(\theta \le \frac{1}{2})$.

4.2. Logistic regression

Assume that $Y = (Y_1, \dots, Y_p)'$ where Y_i's are independent binomial random observations with parameters (n_i, p_i). It is also assumed that p_i is related to a covariate X_i, which is a q–dimensional column vector, through the logistic model

$$p_i = \frac{e^{\theta' x_i}}{1 + e^{\theta' x_i}}.$$

Hence the probability function of Y is

$$P(Y = y) = \prod_i \binom{n_i}{y_i} e^{\theta' x' y} \Big/ \prod_i (1 + e^{\theta' x_i}), \tag{4.4}$$

where $X' = (x_1, \dots, x_p)$ and hence the size of X is $p \times q$. Obviously, the statistic $T = X'Y = (T_1, \dots, T_q)'$ is a sufficient statistic. Let $\theta = (\theta_1, \dots, \theta_q)'$ and consider the problem of testing $H_0 : \theta_1 \leq \theta_1^0$ vs $H_1 : \theta_1 > \theta_1^0$ for some fixed constant θ_1^0. The conditional test against H_0 is based on the conditional distribution of $T_1 = t_1$ given $(T_2, \dots, T_q) = (t_2, \dots, t_q)$, of which the probability function is

$$f_{\theta_1}(t_1 \mid t_2, \dots, t_q) = \frac{c(t_1, t_2) \exp(\theta_1 t_1)}{\sum_{\tau \in S} c(\tau, t_2) \exp(\theta_1 u_1)}, \tag{4.5}$$

where S consists of all possible values of T_1. See for example [11, 2] and [27].

It can be shown (as in Lehmann 1997, p. 178) that the following test with the critical function φ is α–level UMP unbiased test

$$\begin{aligned} \varphi(T_1) &= 1 \quad T_1 > \eta \\ &= r \quad T_1 = \eta \\ &= 0 \quad \text{otherwise}, \end{aligned} \tag{4.6}$$

where $\eta \in S$ is a possible value of t_1 and r is a number such that $0 \leq r \leq 1$. Both r and η depends on $(t_2, \dots, t_q)$ and are chosen such that $E_{\theta_1^0}[\varphi(T_1) \mid t_2, \dots t_q] = \alpha$. Note that the expectation is taken with respect to the probability function (4.5) with $\theta_1 = \theta_1^0$. The computational aspect is discussed in [27].

To derive a good p–value let $\tau_1 < \tau_2 < \cdots < \tau_k$ be all the possible values of T_1. Introduce two other artificial numbers $\tau_0 < \tau_1$ and $\tau_{k+1} > \tau_k$ for ease of notation. Let F be the "forward" function defined on S, i.e., $F(\tau_j) = \tau_{j+1}$ for $0 \leq j \leq k$. Now we may take an independent uniform random variable U on $[0, 1]$ and consider the test that rejects H_0 if and only if

$$T_1 + (F(T_1) - T_1)U > c, \tag{4.7}$$

where assume without loss of generality that $c \leq \tau_{k+1}$. Define a Gauss function $[\cdot]_S$ with respect to S as follows

$$\begin{aligned} [c]_S &= \text{largest } \tau_i'\text{s}, 1 \leq i \leq k+1, \text{ such that } \tau_i < c \text{ if } \tau_1 < c \\ &= \tau_0 \text{ if } c \leq \tau_1. \end{aligned} \tag{4.8}$$

Test (4.7) corresponds to a critical function in (4.6) if

$$\eta = [c]_S \quad \text{and} \quad r = 1 - \frac{c - \eta}{F(\eta) - \eta}. \tag{4.9}$$

The proof of this is given in the Appendix. Consequently, the randomized p–value corresponding to the UMP unbiased test can be written as

$$P_{\theta_0}(T_1 + (F(T_1) - T_1)U > t_1 + (F(T_1) - t_1)U \mid T_2, \ldots, T_q) \tag{4.10}$$

where t_1 and U are the realizations of T_1 and U respectively. Equation (4.10) can also be written as $p_R(t_1, U) = P(T_1(Z) + (F(T_1(Z)) - T_1(Z))U > t_1 + (F(t_1) - t_1)U)$ where Z is a random variable, independent of U, having probability function (4.5) with $\theta_1 = \theta_1^0$. This is a p–value that depends on U which is not a sufficient statistic. The conditional expectation of this p–value given the sufficient statistic equals

$$p_{mid} = P(T_1(Z) > t_1) + \frac{1}{2}P(T_1(Z) = t_1), \tag{4.11}$$

as established in the Appendix. One can deduce a result similar to Theorem 3.1 which states that p_{mid}, not only dominates p_R, but is also optimal.

5. Conclusion

In this paper, we derive classical mid p–values as the optimal p–values for various models. The theory developed is based on Neyman–Pearson and Rao–Blackwell theorems. The mid p–value has been considered to be a sound procedure based on numerical evidence and the theory of this paper supports the claim.

Although we have worked with only one–sided hypotheses, the similar theory can be developed for the two–sided case using Neyman–Pearson and Rao–Blackwell theorems. However the resultant optimal p–value, called the expected p–value, does not have an analytic closed form in general. In the situation of a 2×2 table with two independent binomial observations with equal sample sizes, the expected p–value reduces to a two–sided mid p–value [18]. In a superb pair of articles, Brown, Cai, and DasGupta [6, 7] show that the equal-tailed Jeffreys prior (beta with parameters 0.5 and 0.5) interval or the binomial parameter has some additional interesting connections to the mid p–value of Clopper-Pearson interval and does well from a frequentist perspective.

The mid p–value serves a special role as a procedure suggested by both Neyman–Pearson and Rao–Blackwell theories as well as from a Bayesian prospective. It seems interesting that the mid p–value, which was considered to be *ad hoc*, turns out to be theoretically well justified.

Appendix

Proof of Theorem 2.1. Comparing (2.3) and (2.4) and canceling out the first term on both sides, it suffices to establish that the second terms

$$E[D(T, Y) \mid T = t, Y_{+1k}, 1 \le k \le K] = \frac{1}{2}P(T = t \mid Y_{+1k}, 1 \le k \le K),$$

where

$$D(t, y) = P(h(Z) < h(y), T(Z) = t) + \frac{1}{2}P(h(Z) = h(y), T(Z) = t).$$

Let $z_1, \ldots, z_\ell$ be all the $2 \times 2 \times K$ tables z with marginal totals equal to that of y and such that $T(z) = t$.

We may assume without loss of generality that $p_i = h(z_i)$ is nondecreasing in i. For the ease of argument, we shall assume for now that p_i is strictly increasing. (See the remark at the end for the case that some of the p_i's are equal.) Hence

$$D(t, z_i) = p_1 + \cdots + p_{i-1} + \frac{1}{2}p_i.$$

To calculate the conditional expectation of D given T and the marginal totals, we first consider $P(Y = z_i \mid \text{marginal totals}, T(Y) = t)$. At first glance, the probability depends on θ. But by sufficiency of T and marginal totals or a direct calculation, we can establish easily that it does not depend on θ. Hence we take $\theta = 1$ without loss of generality. This shows that the conditional probability equals $p_i / \sum_{j=1}^{\ell} p_j$. Hence,

$$\begin{aligned}
\text{(A.1)} \qquad & E[D(T,Y) \mid \text{marginal totals}, T(Y) = t] \\
&= \sum_{i=1}^{\ell} \Big(p_1 + \cdots + p_{i-1} + \frac{1}{2}p_i\Big)p_i \Big/ \Big(\sum_{j=1}^{\ell} p_j\Big) \\
&= \frac{1}{2}\Big(2\sum_{i=1}^{\ell}\Big(\sum_{j=1}^{i-1} p_j\Big)p_i + \sum_{i=1}^{\ell} p_i^2\Big) \Big/ \Big(\sum_{j=1}^{\ell} p_j\Big) \\
&= \frac{1}{2}\Big(\sum_{j=1}^{\ell} p_j\Big)^2 \Big/ \Big(\sum_{j=1}^{\ell} p_j\Big) \\
&= \frac{1}{2}\Big(\sum_{j=1}^{\ell} p_j\Big) \\
&= \frac{1}{2}P(T = t | \text{marginal totals}).
\end{aligned}$$

Note that in the above argument we have assumed that p_j is strictly increasing. If p_j is only nondecreasing, we may classify z_i's according to g groups such that the following two statements hold. For every z in G_k, $1 \le k \le g$, $h(z)$ remains the same. Also as K increases, the probability $h(z)$ strictly increases for z in G_k. Define

$$p^*(k) = \sum_{z \in G_k} h(z),$$

which equals $h(z)$ times the size of G_k, because $h(z)$ remains constant for $z \in G_k$. After some careful thinking, (A.1) holds true if all p_i is replaced by p_i^* and $\sum_{i=1}^{\ell}$ is replaced by $\sum_{i=1}^{g}$. Hence Theorem 2.1 is established for this more general case. □

Proof of Theorem 2.3. Let $z_1, \ldots, z_\ell$ be all the $I \times J$ tables such that $T(z_i) = t$ and such that z_i has the same marginal totals as y. Assume without loss of generality that $p_i = h(z_i)$ is increasing in i, where $h(z)$ is defined in (2.8). We, however, only prove the theorem assuming that p_i is strictly increasing. The case when p_i is increasing but not strictly increasing can be proved similar to the proof of Theorem 2.1. Now to show that the Rao–Blackwellized expression of (2.9) equals (2.10), it suffices to show that

$$E[D(T, n) \mid m, T = t] = \frac{1}{2}P(T = t \mid m).$$

Here m represents the marginal column and row totals (see the precise definition in the second paragraph after the statement of Theorem 3.1). Also,

$$D(t,y) = P(h(Z) < h(y), T(Z) = t) + \frac{1}{2}P(h(Z) = h(y), T(Z) = t).$$

Now $D(t, z_i) = p_1 + \cdots + p_{i-1} + \frac{1}{2}p_i$. Hence,

$$E[D(T,N) \mid m, T = t] = \sum_{i=1}^{\ell} \Big(p_1 + \cdots + p_{i-1} + \frac{1}{2}p_i\Big)p_i \Big/ \sum_{j=1}^{\ell} p_j = \frac{1}{2}P(T = t \mid m),$$

where the last equation follows the same argument as in the proof of Theorem 2.1 previously given in the Appendix. □

Proof of (4.9). We shall prove that (4.7) is equivalent to test (4.6) with η and γ given in (4.9) by considering three cases: (i) $\tau_1 < c \le \tau_{k+1}$; case (ii) $c \le \tau_1$ and (iii) $c > \tau_{k+1}$.

For case (i) $\tau_1 < c \le \tau_{k+1}$: The condition that $\tau_1 < c$ implies that $\eta = [c]_s = \tau_j$ for some j, $1 \le j \le k$. When $T_1 > \eta = \tau_j$, $T_1 \ge \tau_{j+1} \ge c$. Consequently (4.7) holds with probability one, agreeing with (4.6). When $T_1 = \eta$, (4.7) is equivalent to $(F(\eta) - \eta)U > c - \eta$, which holds with probability $1 - \frac{c-\eta}{F(\eta)-\eta}$. Hence H_0 is rejected with probability γ for $T_1 = \eta$, agreeing with (4.6). Finally if $T_1 < \eta$, then (4.7) is equivalent to $(F(T_1) - T_1)U > c - T_1$ or $U > (c - T_1)/(F(T_1) - T_1)$. However, $\frac{c-T_1}{F(T_1)-T_1} \ge \frac{\eta-T_1}{F(T_1)-T_1} \ge 1$. Hence given $T_1 < c$, the last displayed inequality holds with zero probability, which is what is stated in (4.6). The assertion holds for case (i).

For case (ii) $c \le \tau_1$: Note (4.7) always holds and H_0 is always rejected using (4.7). Also $\eta = [c]_s = \tau_0$. Hence $T_1 > \eta$ and test (4.6) always rejects as well.

For case (iii), $c > \tau_{k+1}$: Note that $\eta = [c]_s = \tau_{k+1}$, and consequently inequality (4.7) can never hold leading to acceptance of H_0 with probability one. Also obviously $T_1 < \eta = \tau_{k+1}$ and hence (4.6) always accept H_0 as well. Hence the assertion (4.9) is established. □

Proof of (4.11). Note that (4.10) equals

$$P(T_1(Z) > t_1) + P[T_1(Z) = t_1, (F(T_1(Z) - T_1(Z))U > (F(t_1) - t_1)U].$$

When $T_1(Z) = t_1$, $F(T_1(z) - T_1(z)) = F(t_1) - t_1$; hence the second term of the last displayed equation equals $P(T_1(Z) = t_1, U > u) = P(T_1(Z) = t_1)(1 - u)$. The conditional expectation of this expression given t_1 is $\frac{1}{2}P(T(Z) = t_1)$, establishing (4.11). □

Acknowledgements

It is the author's pleasure to thank Gene Hwang for the extensive collaboration on this project as well as Alan Agresti, Roger Berger, Joe Gastwirth, and Rajendran Narayanan for their helpful comments on the paper.

References

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

[2] Agresti, A. (2002). *Categorical Data Analysis*, 2nd ed. Wiley, New York.
[3] Agresti A., Mehta C. R. and Patel N. R. (1990). Exact inference for contingency tables with ordered categories. *J. Amer. Statist. Assoc.* **85** 453–458.
[4] Berger, J. O. (1985). The frequentist viewpoint and conditioning. In *Proceedings of Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer* (L. LeCam and R. Olshen eds.) 15–44. Wadsworth, Belmont.
[5] 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.
[6] Brown, L. D., Cai, T. and DasGupta, A. (2001). Interval estimation for a binomial proportion (with discussion). *Statist. Sci.* **16** 101–133.
[7] Brown, L. D., Cai, T. and DasGupta, A. (2002). Confidence intervals for a binomial proportion and asymptotic expansions. *Ann. Statist.* **30** 160–201.
[8] Brown, L. D. and Hwang, J. T. (1991). Admissibility of confidence estimators. In *Proceedings of the 1990 Taipei Symposium in Statistics, June 28–30* (M. T. Chao and P. E. Cheng, eds.) 1–10. Institute of Statistical Science, Academia Sinica, Taiwan.
[9] Cohen, A. and Sackrowitz, H. B. (1991). Tests for independence in contingency tables with ordered categories. *J. Multivariate Anal.* **36** 56–67.
[10] Cohen, A. and Sackrowitz, H. B. (1992). An evaluation of some tests of trend in contingency tables. *J. Amer. Statist. Assoc.* **87** 470–475.
[11] Cox, D. R. and Snell, E. J. (1989). *The Analysis of Binary Data*, 2nd ed. Chapman and Hall, London.
[12] Fisher, R. A. (1934). *Statistical Methods For Research Workers* (Originally published 1925, 14th ed. 1970). Oliver and Boyd, Edinburgh.
[13] Frisén, M. (1980). Consequences of the use of conditional inference in the analysis of a correlated contingency table. *Biometrika* **67** 23–30.
[14] Goutis, C., Casella, G. and Wells, M. T. (1996). Assessing evidence in multiple hypotheses. *J. Amer. Statist. Assoc.* **91** 1268–1277.
[15] Hirotsu, C. (1982). Use of cumulative efficient scores for testing ordered alternatives in discrete models. *Biometrika* **69** 567–577.
[16] Hwang, J. T., Casella, G., Robert, C., Wells, M. T. and Farrell, R. (1992). Estimation of accuracy of testing. *Ann. Statist.* **20** 490–509.
[17] Hwang, J. T. and Pemantle, R. (1997). Estimating the truth indicator function of a statistical hypothesis under a class of proper loss functions. *Statistics and Decisions* **15** 103–128.
[18] Hwang, J. T. and Yang, M. (2001). An optimality theory for mid p–values in 2×2 contingency tables. *Statist. Sinica* **11** 807–826.
[19] Kiefer, J. (1975). Conditional confidence approach in multi-decision problems. In *Multivariate Analysis IV* (P. R. Krishnaiah, ed.). Academic Press, New York.
[20] Kiefer, J. (1976). Admissibility of conditional confidence procedures. *Ann. Statist.* **4** 836–865.
[21] Kiefer, J. (1977). Conditional confidence statements and confidence estimators. *J. Amer. Statist. Assoc.* **72** 789–827.
[22] Kim, D. and Agresti, A. (1995). Improved exact inference about conditional association in three–way contingency tables. *J. Amer. Statist. Assoc.* **90** 632–639.
[23] Lancaster, H. O. (1961). Significance tests in discrete distributions. *J. Amer. Statist. Assoc.* **56** 223–234.

[24] Ledwina, T. (1984). A note on admissibility of some tests of independence against "positive dependence" in $R \times C$ contingency table. *Math. Operationsforsch. Statist. Ser. Statist.* **15** 565–570.

[25] Lehmann, E. L. (1997). *Testing Statistical Hypotheses*, 2nd ed. Springer, New York.

[26] McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* **12** 153–157.

[27] Mehta, C. R., Patel, N. R. and Senchaudhuri, P. (2000). Efficient Monte Carlo Methods for Conditional Logistic Regression. *J. Amer. Statist. Assoc.* **95** 99–108.

[28] Patefield, W. M. (1982). Exact tests for trends for ordered contingency tables. *Appl. Statist.* **31** 32–43.

[29] Plackett, R. L. (1984). Discussion of *Test of significance for* 2×2 *contingency tables* by Yates. *J. Roy. Statist. Soc. Ser. B* **147** 459.

[30] Schaafsma, W., Tolboom, J. and Van der Meulen, E. A. (1989). Discussing truth or falsity by computing a Q–value. *Statist. Data Anal. Inference* 85–100.

[31] Sidik, K. (1997). Exact unconditional tests for discrete data. Ph.D. thesis, Department of Statistics, North Carolina State Univ.

[32] Strawderman, R. L. and Wells, M. T. (1998). Approximately exact inference for contingency tables. *J. Amer. Statist. Assoc.* **93** 1294–1307.

[33] Van der Meulen, E. A. (1992). Assessing weights of evidence for discussing classical statistical hypotheses. Ph.D. thesis, Univ. Groningen, The Netherlands.

[34] Yates, F. (1984). Test of significance for 2×2 contingency tables. *J. Roy. Statist. Soc. Ser. B* **147** 426–463.