TAMKANG JOURNAL OF MATHEMATICS
Volume 40, Number 4, 415-427, Winter 2009

NONINFORMATIVE BAYESIAN P-VALUES FOR TESTING MARGINAL HOMOGENEITY IN 2×2 CONTINGENCY TABLES

LEE-SHEN CHEN AND MING-CHUNG YANG

Abstract. This article considers the problem of testing marginal homogeneity in 2×2 contingency tables under the multinomial sampling scheme. From the frequentist perspective, McNemar's exact p-value ($p_{_{ME}}$) is the most commonly used p-value in practice, but it can be conservative for small to moderate sample sizes. On the other hand, from the Bayesian perspective, one can construct Bayesian p-values by using the proper prior and posterior distributions, which are called the prior predictive p-value (p_{prior}) and the posterior predictive p-value (p_{post}), respectively. Another Bayesian p-value is called the partial posterior predictive p-value (p_{ppost}), first proposed by [2], which can avoid the double use of the data that occurs in p_{post}. For the preceding problem, we derive p_{prior}, p_{post}, and p_{ppost} based on the noninformative uniform prior. Under the criterion of uniformity in the frequentist sense, comparisons among p_{prior}, $p_{_{ME}}$, p_{post} and p_{ppost} are given. Numerical results show that p_{ppost} has the best performance for small to moderately large sample sizes.

1. Introduction

Testing the null hypothesis or model having the equality of marginal proportions in a 2×2 contingency table is frequently conducted in clinical studies and many practical applications. The most commonly used measure of compatibility of the null model with the observed data is the p-value defined by the appropriate test statistic. However, the p-value has been criticized for a long time in the statistical literature. A brief summary of the controversy about the p-value can be found in [12]. As noted in [12], the major criticisms about the p-value have come from Bayesian viewpoints. For instance, the calculation of the p-value involves averaging over sample values which have not occurred, that is, a clear violation of the likelihood principle; see, for example, [14], [5], and [4]. In recent advances, [1], [2], and [20] gave thoughtful discussions about common misinterpretation of p-values from both Bayesian and frequentist perspectives. In [20], they also

Corresponding author: Ming-Chung Yang.

2000 *Mathematics Subject Classification.* Primary 62F15; Secondary 62F03.

Key words and phrases. Bayesian p-value, McNemar's exact p-value, contingency table, marginal homogeneity.

developed two procedures to calibrate p-values in testing precise null hypotheses, which can be interpreted in either a Bayesian or a frequentist way. However, several leading Bayesians, for example [9], [10], [7], and [19], have argued that the p-value, calculating a tail-area probability of a statistic can be a useful tool, even for Bayesian analysts in monitoring the adequacy of a model. This has led to several formulations of Bayesian p-values over past decades. [7] popularized the use of the prior distribution to construct the *prior predictive p-value* (p_{prior}). [19] used the posterior predictive distribution of a statistic to calculate the tail-area probability corresponding to the observed value of the statistic. This tail-area probability is called the *posterior predictive p-value* (p_{post}). However, the drawback of p_{prior} is that it can not be derived by using improper noninformative priors, which are often considered by the objective analyst from the beginning. Although the posterior (predictive) distribution is typically proper by using the improper noninformative prior, the main weakness of p_{post} is that its calculation involves double use of the data. [2] first proposed the *partial posterior predictive p-value* (p_{ppost}) to modify p_{post}. This improved Bayesian p-value can be derived from proper or improper priors and can avoid the double use of data.

The rest of this article is organized as follows. In Section 2, we first introduce McNemar's exact p-value, which is the most commonly used p-value in practice. Then, we consider the noninformative priors, from the objective Bayesian perspective, and take the uniform prior on the null hypothesis to derive p_{prior}, p_{post}, and p_{ppost}. Note that the uniform prior is proper due to the null parameter space being bounded. In Section 3, under the criterion of uniformity, comparisons among $p_{_{ME}}$, p_{prior}, p_{post}, and p_{ppost} are given. From the sense of frequentist, an appealing property for a random p-value is that it has the $U(0,1)$ distribution under the null model for all values of parameters. [2] also argued that if a p-value has uniformity under the null model in the frequentist sense, then it has the strong Bayesian property, which is the marginal uniformity under any proper prior. Hence, the uniformity can be adopted to evaluate Bayesian p-values in the numerical study. Our numerical results show that p_{ppost} has the best performance for all cases. The concluding remarks are given in Section 4.

2. McNemar's p-value and Bayesian p-values

For matched-pairs binary data, one shall consider the 2×2 contingency table with random cell counts X_{ij}, $i = 1, 2$, $j = 1, 2$, satisfying the multinomial distribution having total sample size n and cell proportions θ_{ij}, $i = 1, 2$, $j = 1, 2$, as layed out as follows:

			Row total		
	X_{11}	X_{12}	X_{1+}	θ_{11}	θ_{12}
	X_{21}	X_{22}	X_{2+}	θ_{21}	θ_{22}
Column total	X_{+1}	X_{2+}			

Let $\boldsymbol{X} = (X_{11}, X_{12}, X_{21}, X_{22})$, $\boldsymbol{\theta} = (\theta_{11}, \theta_{12}, \theta_{21}, \theta_{22})$, and $\sum_{i=1}^{2}\sum_{j=1}^{2} X_{ij} = n$. The joint *probability mass function* or joint *pmf* of $\boldsymbol{X}$ is

$$f(\boldsymbol{x};\boldsymbol{\theta}) = \frac{n!}{x_{11}!x_{12}!x_{21}!x_{22}!}\theta_{11}^{x_{11}}\theta_{12}^{x_{12}}\theta_{21}^{x_{21}}\theta_{22}^{x_{22}} \tag{2.1}$$

where the vector of cell counts $\boldsymbol{x} = (x_{11}, x_{12}, x_{21}, x_{22})$ satisfying $\sum_{i=1}^{2}\sum_{j=1}^{2} x_{ij} = n$ and $\sum_{i=1}^{2}\sum_{j=1}^{2} \theta_{ij} = 1$. Testing marginal homogeneity, which is $\theta_{11}+\theta_{12} = \theta_{11}+\theta_{21}$, is equivalent to testing $\theta_{12} = \theta_{21}$, and hence the interesting hypotheses to be tested is one sided setting as follows:

$$H_0 \ : \ \theta_{12} \ = \ \theta_{21} \quad \text{vs} \ H_1 \ : \ \theta_{12} \ > \ \theta_{21}\,. \tag{2.2}$$

Since the null hypothesis in (2.2) contains one parameter $\theta = \theta_{12} = \theta_{12}$, the data can be condensed as three cell counts $(X_{12}, X_{21}, n - X_{12} - X_{21})$ where $n - X_{12} - X_{21} = X_{11} + X_{22}$ denotes the number concordant pairs, which is fixed if both X_{12} and X_{12} are given. For simplicity, the notation $\boldsymbol{X} = (X_{11}, X_{12}, n - X_{21} - X_{22})$ is replaced by $\boldsymbol{X} = (X_1, X_2, n - X_1 - X_2)$. Now, the preceding pmf in (2.1) under the null hypothesis in (2.2) can be simplified as

$$f(\boldsymbol{x};\theta) = \frac{n!}{x_1!x_2!(n-x_1-x_2)!}\theta^{x_1+x_2}(1-2\theta)^{n-x_1-x_2} \tag{2.3}$$

where $0 \le x_1,\ x_2 \le n$, $x_1 + x_2 \le n$, and $0 \le \theta \le 1/2$. Obviously, the joint pmf in (2.3) is degenerate when $\theta = 0$.

2.1. McNemar's p-value

The most commonly used approach, from the frequentist perspective, to eliminate the nuisance parameters is to condition on their sufficient statistics. In our case, the sufficient statistic for θ in the null hypothesis (2.2) is $S = X_1 + X_2$. Conditioning on $S = X_1 + X_2$, the conditional distribution for $\boldsymbol{X}$ under H_0 in (2.2) is

$$\begin{aligned} f(\boldsymbol{x}|S=s) &= \frac{f(\boldsymbol{x};\theta)}{f(s;\theta)} \\ &= \binom{s}{x_1}\left(\frac{1}{2}\right)^{x_1}\left(\frac{1}{2}\right)^{s-x_1} \end{aligned} \tag{2.4}$$

where $s = x_1 + x_2$. Hence, given $S = s$, the conditional distribution of $\boldsymbol{X}$ does not contain the nuisance parameter. One can choose $T(\boldsymbol{X}) = X_1$ as a test statistic since the large value of X_1 indicates less compatibility of the null model. In fact, the one-sided testing problem in (2.2) was first treated by [17], who proposed an conditional asymptotic method, which is based on only S, the number of discordant pairs. Given the discordant pairs $S = s$, the distribution X_1 follows the binomial distribution. Then McNemar's test statistic is defined as

$$W \;=\; (X_1 - X_2)/\sqrt{X_1 + X_2} \;=\; (2\,X_1 - s)/\sqrt{s} \tag{2.5}$$

and the large value of W indicates the null hypothesis H_0 is unlikely. For one-sided hypotheses (2.2), the p-value is defined as

$$P(W \geq w) \tag{2.6}$$

where $w = (x_1 - x_2)/\sqrt{s} = (2x_1 - s)/\sqrt{s}$ denotes the realization of W. Since the null conditional distribution of W converges to the standard normal if the sample size is moderately large, one can use the asymptotic null distribution, the standard normal, to compute McNemar's asymptotic p-value. However, this p-value in (2.6) can also be derived from the sense of uniformly most powerful (UMP) unbiased test because McNemar's test statistic W in (2.5) is obtained as the test statistic in the UMP unbiased test; see [15] on page 169. From (2.3), the nonrandomized p-value corresponding to the UMP unbiased test for the hypotheses (2.2) can be expressed as

$$p_{_{ME}}(x_1) \;=\; P\left(X_1 \;\geq\; x_1 \,\middle|\, X_1 + X_2 = s\right) \;=\; P(B \;\geq\; x_1) \tag{2.7}$$

where x_1 is the realization of X_1 and B represents the random variable having the symmetric binomial probability mass function

$$g(c) = \binom{s}{c}\left(\frac{1}{2}\right)^{s}, \quad 0 \leq c \leq s\ . \tag{2.8}$$

The p-value in (2.7) is referred to as the *exact* conditional p-value, and is called *McNemar's exact p-value*. Here the term "*exact*" means the use of an exact conditional null distribution to calculate the p-value.

2.2. Bayesian p-values

It is well-known that Bayesian school has a natural way, which is to integrate the nuisance parameters out, to eliminate the nuisance parameters. In this subsection, we use a noninformative but proper prior distribution for θ and choose the McNemar's test statistics in (2.5) or (2.7), $T(\boldsymbol{X}) = X_1$, as the departure statistic to derive several Bayesian p-values. We consider the noninformative uniform prior distribution for θ

$$\pi(\theta) = \begin{cases} 2\ , & 0 \leq \theta \leq \frac{1}{2} \\ 0\ , & \text{otherwise} \end{cases}\ . \tag{2.9}$$

Then, the marginal distribution for $\mathbf{X}$ is as follows:

$$\begin{aligned} m(\boldsymbol{x}) &= \int_0^{1/2} f(\boldsymbol{x};\theta)\pi(\theta)d\theta \\ &= \frac{1}{n+1}\binom{x_1 + x_2}{x_1} 2^{-(x_1+x_2)}, \quad 0 \leq x_1, x_2 \leq n,\ 0 \leq x_1 + x_2 \leq n. \end{aligned} \tag{2.10}$$

The prior predictive p-value (p_{prior}) based on the prior in (2.9) and the observed data $\boldsymbol{x}_{obs} = (x_1^o, x_2^o)$ and $m(\boldsymbol{x})$ is then defined as

$$\begin{aligned} p_{prior}(\boldsymbol{x}_{obs}) &= P_r^{m(\cdot)}(X_1 \geq x_1^o) \\ &= \frac{1}{n+1} \sum_{x_1=x_1^o}^{n} \sum_{x_2=0}^{n-x_1} \binom{x_1+x_2}{x_1} 2^{-(x_1+x_2)}, \quad x_1^o = 0, 1, 2, \ldots, n. \end{aligned} \tag{2.11}$$

The derivation of $m(\boldsymbol{x})$ in (2.10) is provided in A.1 of the Appendix. The weakness of p_{prior} in (2.11) is that its performance heavily depends on the prior $\pi(\theta)$. [11] and [19] proposed to use the marginal distribution of $\boldsymbol{X}$, $m_1(\boldsymbol{x}|\boldsymbol{x}_{obs})$, which can be obtained by integrating $f(\boldsymbol{x};\theta)$ with respect to the posterior distribution $\pi(\theta|\boldsymbol{x}_{obs})$, instead of the $\pi(\theta)$ in (2.9). Then one can define the posterior predictive p-value (p_{post}) based on $m_1(\boldsymbol{x}|\boldsymbol{x}_{obs})$ as

$$p_{post}(\boldsymbol{x}_{obs}) = P_r^{m_1(\cdot|\boldsymbol{x}_{obs})}(X_1 \geq x_1^o), \;\; x_1^o = 0, \ldots, n \tag{2.12}$$

where

$$\begin{aligned} m_1(\boldsymbol{x}|\boldsymbol{x}_{obs}) &= \int_0^{1/2} f(\boldsymbol{x};\theta)\pi(\theta|\boldsymbol{x}_{obs})d\theta \\ &= \frac{(n+1)!}{x_1!x_2!(n-x_1-x_2)!} \binom{n}{x_1^o+x_2^o} (1/2)^{x_1+x_2} \\ &\qquad \times B(x_1^o+x_2^o+x_1+x_2+1, 2n-(x_1^o+x_2^o+x_1+x_2)+1). \end{aligned}$$

Here $B(\alpha,\beta)$ denotes the beta function with parameters α and β for which $B(\alpha,\beta) = \int_0^1 \theta^{\alpha-1}(1-\theta)^{\beta-1}d\theta$. The derivation of $m(\boldsymbol{x}|\boldsymbol{x}_{obs})$ is also given in A.2 of the Appendix. Generally, the p_{post} is much more heavily influenced by the model than by the prior. Another drawback of p_{post} is that it involves "double use" of the data, which we first use the data to obtain the posterior distribution $\pi(\theta|\boldsymbol{x}_{obs})$ for deriving $m(\boldsymbol{x}|\boldsymbol{x}_{obs})$ and use the data again to compute the p-value in (2.7). The double use of the data can induce unnatural behavior for p_{post}; see, for example, [2]. To avoid the double use of the data, [2] first proposed the partial posterior predictive p-value to improve p_{post}. The partial posterior predictive p-value (p_{ppost}) is defined as

$$\begin{aligned} p_{ppost}(\boldsymbol{x}_{obs}) &= P_r^{m_2(\cdot|\boldsymbol{x}_{obs}\backslash t_{obs})}(T \geq t_{obs}) \\ &= 1 - c(x_1^o, x_2^o) \sum_{x_1=0}^{x_1^o-1} \binom{n}{x_1} \sum_{k=0}^{x_1^o-x_1} (1/2)^{x_1^o+x_2^o-k+1}(-1)^{x_1^o-x_1-k}\binom{x_1^o-x_1}{k} \\ &\qquad \times B(x_1^o+x_2^o-k+1, n-x_1^o-x_2^o+1) \end{aligned} \tag{2.13}$$

where $T = T(\boldsymbol{X}) = X_1$ and $t_{obs} = t(\boldsymbol{x}_{obs}) = x_1^o = 0, \ldots, n$. In order to obtain p_{ppost}, we compute

$$m_2(x_1|\boldsymbol{x}_{obs}\backslash t_{obs}) = \int_0^{1/2} f(x_1|\theta)\,\pi\,(\theta|\boldsymbol{x}_{obs}\backslash t_{obs})d\theta$$

$$= c(x_1^o, x_2^o)\binom{n}{x_1}\sum_{k=0}^{x_1^o - x_1}(1/2)^{x_1^o+x_2^o-k+1}(-1)^{x_1^o-x_1-k}\binom{x_1^o - x_1}{k}$$

$$\times B(x_1^o + x_2^o - k + 1, n - x_1^o - x_2^o + 1) \qquad (2.14)$$

where

$$\pi(\theta|\boldsymbol{x}_{obs}\backslash t_{obs}) \propto f(\boldsymbol{x}_{obs}|t_{obs};\theta)\,\pi\,(\theta) \propto \frac{f(\boldsymbol{x}_{obs};\theta)\pi(\theta)}{f(t_{obs};\theta)}. \qquad (2.15)$$

The derivations of $m_2(t|\boldsymbol{x}_{obs}\backslash t_{obs})$ and $p_{ppost}(\boldsymbol{x}_{obs})$ can be found in A.3 of the Appendix. Since the contribution of $t_{obs} = x_1^o$ to the posterior is "removed" in (2.15) before eliminating the θ by integration, obviously p_{ppost} avoids the double use of the data that occurs in p_{post}. To indicate this, the notation "$\boldsymbol{x}_{obs}\backslash t_{obs}$" is used for the marginal distribution of $T = X_1$ in (2.14) and the partial posterior in (2.15).

3. Numerical Study

From the frequentist viewpoint to evaluate p-values, one appealing property for a random p-value is that it has $U(0,1)$ distribution for all parameters in the null model. This uniformity property can be used to judge a proposed p-value whether to be conservative or anticonservative in a frequentist sense. For a Bayesian, the uniformity property is also adopted to evaluate Bayesian p-values as mentioned in Section 1; also see [18], [2], [3], and [16]. However, the distributions of p-values would not be $U(0,1)$ in our cases due to the discrete sample space. Therefore, we may compute the absolute distances (AD) to $U(0,1)$ of the p-values. Given a p-value $p(\boldsymbol{X})$, the absolute distance between $U(0,1)$ and $p(\boldsymbol{X})$ is defined as

$$\int_0^1 \Big|F_\theta(\alpha) - U(\alpha)\Big|d\alpha \qquad (3.1)$$

where F_θ is the distribution function of $p(\boldsymbol{X})$ with the parameter θ and U denotes the distribution function of $U(0,1)$. It is also interesting in measuring the "local" uniformity of p-values by specifying the range of nominal levels and computing the "local" absolute distance (LAD) between $U(0,1)$ and the p-value $p(\boldsymbol{X})$ as

$$\int_{\alpha_1}^{\alpha_2} \Big|F_\theta(\alpha) - U(\alpha)\Big|d\alpha$$

where F_θ and U are both defined in (3.1) and $0 \leq \alpha_1 < \alpha_2 \leq 1$. For instance, in a frequentist sense, the commonly used nominal level α varies from 0.01 to 0.1 for testing the null model.

In the following calculations, we consider various θ values and sample sizes n to calculate ADs in the table 1 and LADs in the table 2 for $p_{_{ME}}$, p_{prior}, p_{post}, and p_{ppost}.

In Table 1, we observe that for any selected θ value, the AD decreases when the sample size increases. For any selected sample size, Table 1 also shows that all the four p-values have the largest AD when $\theta = 0.05$ and $p_{_{ME}}$, p_{post}, and p_{ppost} have the smallest AD when $\theta = 0.45$; except the case of $n = 5$, p_{prior} has the smallest AD when $\theta = 0.25$. Moreover, from Table 1 we find that p_{ppost} has the smallest AD for all cases, and hence p_{ppost} has remarkable best performance. It indicates that the distribution of p_{ppost} is closest to $U(0, 1)$ than those of other three p-values, and p_{ppost} substantially improves p_{post} for all cases.

In Table 2, $\alpha_1 = 0.01$ and $\alpha_2 = 0.1$ are taken to calculate LADs. For any selected sample size, Table 2 shows that $p_{_{ME}}$, p_{post}, and p_{ppost} have the largest LAD when $\theta = 0.05$. On the other hand, $p_{_{ME}}$ and p_{post} have the smallest LAD when $\theta = 0.45$ while p_{prior} and p_{ppost} have the smallest LAD when $\theta = 0.35$. As expected, we also find that p_{ppost} has the smallest LAD for all cases in Table 2, and it is the best one among the four p-values. It is worth noting that for $n = 10$, the LAD of p_{post} is larger than that of $p_{_{ME}}$ for any selected θ value. This result demonstrates the impact of the double use of the data, and the performance of p_{post} is even worse than the classical (frequentist) p-value, $p_{_{ME}}$, which is conservative in small sample cases.

4. Concluding Remarks

Testing the equality of marginal proportions for matched-pairs binary data, $p_{_{ME}}$ is the most commonly used p-value in many practical applications, but it can be conservative for small to moderate sample sizes due to its discreteness of distribution; see, for example, [21], [6], and [8]. It is worth to note that using the mid p-value can reduce the high discreteness of distribution occurred in $p_{_{ME}}$, and the optimality of the mid p-value can also be found in [13]. Here, we do not attempt to study the mid p-value from the Bayesian viewpoint. In contrast to McNemar's exact p-value, $p_{_{ME}}$, from the frequentist perspective, we are interested in the performance of the Bayesian p-value from the objective Bayesian perspective. The main goal of this article is to investigate the behavior of Bayesian p-values using the noninformative uniform prior on the parameter space of null hypothesis. In Section 2, we derive three Bayesian p-values, including p_{prior}, p_{post}, and p_{ppost}.

In comparisons of Bayesian p-values and frequentist p-values, one can take the criterion of uniformity, which will be acceptable for (objective) Bayesians and (conditional) frequentists. Then the absolute distance (AD) and local absolute distance (LAD) of p-values can be calculated to make comparisons among the preceding four p-values. Our numerical calculations show that p_{ppost} with respect to the uniform prior has the smallest AD and LAD for any selected sample size and θ value. Furthermore, numerical results indicate that p_{ppost} is closest to $U(0, 1)$ than the other p-values for all cases. It is not

Table 1. The AD between the p-value and $U(0,1)$ with sample size n and parameter θ.

n	θ	p_{ME}	p_{Prior}	p_{post}	p_{ppost}
5	0.05	0.39282	0.41881°	0.38938	**0.36951**
	0.15	0.26610°	0.26246	0.23799	**0.20025**
	0.25	0.20929°	0.12451	0.14745	**0.10782**
	0.35	0.17297°	0.07244	0.09028	**0.05880**
	0.45	0.13566	0.14086°	0.05330	**0.04346**
10	0.05	0.32086	0.40913°	0.32056	**0.29418**
	0.15	0.18777	0.23057°	0.15904	**0.12320**
	0.25	0.13819°	0.08993	0.09433	**0.06055**
	0.35	0.11670°	0.10706	0.05722	**0.03167**
	0.45	0.09523	0.22936°	0.02937	**0.02252**
15	0.05	0.28962	0.40628°	0.27074	**0.24293**
	0.15	0.15664	0.22119°	0.12373	**0.09241**
	0.25	0.11751°	0.09821	0.07538	**0.04477**
	0.35	0.09403	0.13743°	0.04736	**0.02286**
	0.45	0.08643	0.27050°	0.02376	**0.01716**
20	0.05	0.24807	0.40477°	0.23397	**0.20675**
	0.15	0.13176	0.21804°	0.10451	**0.07642**
	0.25	0.09728	0.11104°	0.06798	**0.36690**
	0.35	0.08304	0.15605°	0.04394	**0.01845**
	0.45	0.07406	0.29467°	0.02144	**0.01419**
25	0.05	0.21822	0.40385°	0.20631	**0.18048**
	0.15	0.11546	0.21892°	0.09357	**0.06649**
	0.25	0.08653	0.12241°	0.06448	**0.03167**
	0.35	0.07387	0.16873°	0.04230	**0.01574**
	0.45	0.06051	0.31086°	0.02012	**0.01233**
30	0.05	0.19508	0.40323°	0.18511	**0.16088**
	0.15	0.10486	0.22103°	0.08768	**0.05960**
	0.25	0.08035	0.13120°	0.06242	**0.02821**
	0.35	0.06763	0.17808°	0.04133	**0.01388**
	0.45	0.05804	0.32259°	0.01926	**0.01101**
40	0.05	0.16348	0.40244°	0.15532	**0.13411**
	0.15	0.08942	0.22663°	0.08174	**0.05045**
	0.25	0.06873	0.14498°	0.06011	**0.02365**
	0.35	0.05717	0.19122°	0.04022	**0.01149**
	0.45	0.05090	0.33864°	0.01815	**0.00930**
50	0.05	0.14285	0.40196°	0.13573	**0.11692**
	0.15	0.07916	0.23082°	0.07856	**0.04450**
	0.25	0.06095	0.15508°	0.05882	**0.02073**
	0.35	0.05139	0.20037°	0.03958	**0.00998**
	0.45	0.04437	0.34923°	0.01749	**0.00818**

Note: The smallest AD among p-values is bolded, and the largest AD is marked by " ° ".

surprising that p_{post} performs worse than McNemar's exact p-value for some of small sample sizes. This is because p_{post} involves the double use of the data. The final remark here is that, from either the Bayesian or frequentist viewpoint, the p-values should be

Table 2. The LAD between the p-value and $U(0,1)$ with sample size n and parameter θ.

n	θ	p_{ME}	p_{Prior}	p_{post}	p_{ppost}
5	0.05	0.00495°	0.00495°	0.00495°	**0.00490**
	0.15	0.00488	0.00482	0.00489°	**0.00407**
	0.25	0.00452	0.00401	0.00458°	**0.00226**
	0.35	0.00374	0.00179	0.00384°	**0.00078**
	0.45	0.00291	0.00383°	0.00281	**0.00194**
10	0.05	0.00492	0.00495°	0.00495°	**0.00478**
	0.15	0.00420	0.00491	0.00492°	**0.00306**
	0.25	0.00339	0.00433	0.00460°	**0.00130**
	0.35	0.00313	0.00174	0.00366°	**0.00039**
	0.45	0.00248	0.00648°	0.00197	**0.00077**
15	0.05	0.00482	0.00495°	0.00495°	**0.00467**
	0.15	0.00350	0.00494°	0.00490	**0.00241**
	0.25	0.00279	0.00468°	0.00453	**0.00098**
	0.35	0.00232	0.00230	0.00355°	**0.00028**
	0.45	0.00204	0.00723°	0.00176	**0.00061**
20	0.05	0.00465	0.00495°	0.00495°	**0.00454**
	0.15	0.00308	0.00495°	0.00487	**0.00201**
	0.25	0.00240	0.00478°	0.00444	**0.00080**
	0.35	0.00205	0.00232	0.00346°	**0.00023**
	0.45	0.00182	0.00924°	0.00169	**0.00044**
25	0.05	0.00443	0.00495°	0.00495°	**0.00426**
	0.15	0.00274	0.00495°	0.00485	**0.00173**
	0.25	0.00216	0.00484°	0.00440	**0.00072**
	0.35	0.00187	0.00252	0.00343°	**0.00016**
	0.45	0.00170	0.01051°	0.00164	**0.00034**
30	0.05	0.00419	0.00495°	0.00495°	**0.00397**
	0.15	0.00252	0.00495°	0.00483	**0.00156**
	0.25	0.00200	0.00489°	0.00439	**0.00066**
	0.35	0.00174	0.00285	0.00343°	**0.00016**
	0.45	0.00155	0.01113°	0.00157	**0.00034**
40	0.05	0.00372	0.00495°	0.00495°	**0.00345**
	0.15	0.00221	0.00495°	0.00480	**0.00136**
	0.25	0.00178	0.00493°	0.00432	**0.00057**
	0.35	0.00155	0.00343°	0.00336	**0.00013**
	0.45	0.00138	0.01198°	0.00153	**0.00026**
50	0.05	0.00333	0.00495°	0.00495°	**0.00306**
	0.15	0.00201	0.00495°	0.00476	**0.00122**
	0.25	0.00162	0.00494°	0.00428	**0.00050**
	0.35	0.00140	0.00376°	0.00331	**0.00010**
	0.45	0.00125	0.01285°	0.00150	**0.00021**

Note: The smallest LAD among p-values is bolded, and the largest LAD is marked by ” ◦ ”.

interpreted carefully to avoid misuse.

Appendix

A.1 Derive $m(\boldsymbol{x})$ in (2.10) used to compute p_{prior} in (2.11).

$$\begin{aligned}
m(\boldsymbol{x}) &= \int_0^{1/2} f(\boldsymbol{x};\theta)\pi(\theta)d\theta \\
&= \frac{2(n!)}{x_1!x_2!(n-x_1-x_2)!}\int_0^{1/2}\theta^{x_1+x_2}(1-2\theta)^{n-x_1-x_2}d\theta \\
&= \frac{n!}{(2^{x_1+x_2})x_1!x_2!(n-x_1-x_2)!}\int_0^{1} y^{x_1+x_2}(1-y)^{n-x_1-x_2}dy, \\
&= \frac{n!}{(2^{x_1+x_2})x_1!x_2!(n-x_1-x_2)!}B(x_1+x_2+1, n-x_1-x_2+1) \\
&= \frac{1}{n+1}\binom{x_1+x_2}{x_1}2^{-(x_1+x_2)}, \quad 0 \le x_1, x_2 \le n, \quad 0 \le x_1+x_2 \le n.
\end{aligned}$$

Here $B(\alpha,\beta)$ denotes the beta function with parameters α and β for which $B(\alpha,\beta) = \int_0^1 \theta^{\alpha-1}(1-\theta)^{\beta-1}d\theta$. Now,

$$\begin{aligned}
p_{prior}(\boldsymbol{x}_{obs}) &= {P_r}^{m(\cdot)}(X_1 \ge x_1^o) \\
&= \frac{1}{n+1}\sum_{x_1=x_1^o}^{n}\sum_{x_2=0}^{n-x_1}\binom{x_1+x_2}{x_1}2^{-(x_1+x_2)}, x_1^o = 0,1,2,\ldots,n.
\end{aligned}$$

A.2 Derive $m_1(\boldsymbol{x}|\boldsymbol{x}_{obs})$ used to compute p_{post} in (2.12).

To compute the posterior predictive p-value, we have to derive the posterior distribution of θ.

$$\begin{aligned}
\pi(\theta|\boldsymbol{x}_{obs}) &= f(\boldsymbol{x}_{obs}|\theta)\pi(\theta)/m(\boldsymbol{x}_{obs}) \\
&= \frac{n!}{x_1^o!x_2^o!(n-x_1^o-x_2^o)!}\theta^{x_1^o+x_2^o}(1-2\theta)^{n-x_1^o-x_2^o}\frac{2(n+1)2^{x_1^o+x_2^o}}{\binom{x_1^o+x_2^o}{x_1^o}} \\
&= 2(n+1)\binom{n}{x_1^o+x_2^o}(2\theta)^{x_1^o+x_2^o}(1-2\theta)^{n-x_1^o-x_2^o}.
\end{aligned}$$

$f(\boldsymbol{x}_{obs}|\theta)\pi(\theta)/m(\boldsymbol{x}_{obs})$

$$= \frac{(n+1)!}{x_1!x_2!(n-x_1-x_2)!}\binom{n}{x_1^o+x_2^o}(1/2)^{x_1+x_2-1}(2\theta)^{x_1^o+x_2^o+x_1+x_2}(1-2\theta)^{2n-x_1^o-x_2^o-x_1-x_2}.$$

Therefore,

$$m_1(\boldsymbol{x}|\boldsymbol{x}_{obs}) = \int_0^{1/2} f(\boldsymbol{x}|\theta)\pi(\theta|\boldsymbol{x}_{obs})d\theta$$

$$= \frac{(n+1)!}{x_1!x_2!(n-x_1-x_2)!}\binom{n}{x_1^o+x_2^o}(1/2)^{x_1+x_2} \times$$

$$B(x_1^o+x_2^o+x_1+x_2+1, 2n-(x_1^o+x_2^o+x_1+x_2)+1).$$

$p_{post}(\boldsymbol{x}_{obs}) = P_r{}^{m_1(\cdot|\boldsymbol{x}_{obs})}(X_1 \geq x_1^o),\ x_1^o = 0, \ldots, n.$

A.3 Derive $m_2(t|\boldsymbol{x}_{obs}\backslash t_{obs})$ in (2.14) used to compute p_{ppost} in (2.13).

The partial posterior $\pi(\theta|\boldsymbol{x}_{obs}\backslash t_{obs})$ in (2.11) under H_0 based on $\boldsymbol{x}_{obs} = (x_1^o, x_2^o)$ and $t_{obs} = t(\boldsymbol{x}_{obs}) = x_1^o$ is proportional to $f(\boldsymbol{x}_{obs};\theta)\pi(\theta)/f(t_{obs};\theta)$. Thus, one has

$$\pi(\theta|\boldsymbol{x}_{obs}\backslash t_{obs}) = c(x_1^o, x_2^o)\theta^{x_2^o}(1-2\theta)^{n-x_1^o-x_2^o}(1-\theta)^{x_1^o-n}$$

where $c(x_1^o, x_2^o) = (\int_0^{1/2} \theta^{x_2^o}(1-2\theta)^{n-x_1^o-x_2^o}(1-\theta)^{x_1^o-n}d\theta)^{-1}$. Then,

$$m_2(x_1|\boldsymbol{x}_{obs}\backslash t_{obs}) = \int_0^{1/2} f(x_1|\theta)\pi(\theta|\boldsymbol{x}\backslash t_{obs})d\theta$$

$$= c(x_1^o, x_2^o)\int_0^{1/2}\binom{n}{x_1}\theta^{x_1}(1-\theta)^{x_1^o-x_1}\theta^{x_2^o}(1-2\theta)^{n-x_1^o-x_2^o}d\theta.$$

Suppose $x_1^o > x_1$, and $(1-\theta)^{x_1^o-x_1}$ can be expressed as

$$\sum_{k=0}^{x_1^o-x_1}\binom{x_1^o-x_1}{k}(-1)^{x_1^o-x_1-k}\theta^{x_1^o-x_1-k}.$$

Now, for $t_{obs} = x_1^o > x_1$,

$$m_2(x_1|\boldsymbol{x}_{obs}\backslash t_{obs})$$

$$= c(x_1^o, x_2^o)\binom{n}{x_1}\sum_{k=0}^{x_1^o-x_1}(-1)^{x_1^o-x_1-k}\binom{x_1^o-x_1}{k}\int_0^{1/2}\theta^{x_1^o+x_2^o-k}(1-2\theta)^{n-x_1^o-x_2^o}d\theta$$

$$= c(x_1^o, x_2^o)\binom{n}{x_1}\sum_{k=0}^{x_1^o-x_1}(1/2)^{x_1^o+x_2^o-k+1}(-1)^{x_1^o-x_1-k}\binom{x_1^o-x_1}{k}$$

$$\times B(x_1^o + x_2^o - k + 1, n - x_1^o - x_2^o + 1).$$

Then,

$$p_{ppost}(\boldsymbol{x}_{obs}) = P_r^{m_2(\cdot|\boldsymbol{x}_{obs}\backslash t_{obs})}(X_1 \geq x_1^o)$$

$$= 1 - P_r^{m_2(\cdot|\boldsymbol{x}_{obs}\backslash t_{obs})}(X_1 < x_1^o)$$

$$= 1 - c(x_1^o, x_2^o)\sum_{x_1=0}^{x_1^o-1}\binom{n}{x_1}\sum_{k=0}^{x_1^o-x_1}(1/2)^{x_1^o+x_2^o-k+1}(-1)^{x_1^o-x_1-k}\binom{x_1^o-x_1}{k}$$

$$\times B(x_1^o + x_2^o - k + 1, n - x_1^o - x_2^o + 1).$$

References

[1] M. J. Bayarri and J. O. Berger, *Quantifying surprise in the data and model verification*, In Bayesian Statistics 6, eds. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, London: Oxford University Press, 53–82, 1999.

[2] M. J. Bayarri and J. O. Berger, *P values for composite null models (with discussion)*, J. Amer. Statist. Assoc. **95** (2000), 1127–1170.

[3] M. J. Bayarri and J. O. Berger, *The interplay of Bayesian and frequentist analysis*, Statist. Sci. **19** (2004), 58–80.

[4] J. O. Berger and M. Delampady, *Testing precise hypotheses (with discussion)*, Statist. Sci. **2** (1987), 317–352.

[5] J. O. Berger and R. L. Wolpert, The Likelihood Principle, Institute of Mathematical Statistics, Hayward, CA, 1984.

[6] R. L. Berger and K. Sidik, *Exact unconditional tests for a* 2×2 *matched-pairs design*, Statistical Methods in Medical Research **12** (2003), 91-108.

[7] G. E. P. Box, *Sampling and Bayes inference in scientific modeling and robustness*, J. Roy. Statist. Soc. Ser. A **143** (1980), 383–430.

[8] L.-S. Chen and M.-C. Yang, *Improved p-values for testing marginal homogeneity in 2×2 contingency tables*, Commun. Statist.-Theory and Methods **38** (2009), 1649-1663.

[9] A. P. Dempster, *Model searching and estimation in the logic of inference (with discussion)*, In Foundations of Statistical Inference (V.P. Godambe and D.A. Sprott, eds.) 56–81. Holt, Rinehart and Winston, Toronto,1971.

[10] A. P. Dempster, *The direct use of likelihood for significance testing (with discussion)*, In Proceedings of Conference on Foundational Questions in Statistical Inference (O. Barndorff-Nielsen, P. Blaeslid and G. Schou, eds.) 335–354. Department of Theoretical Statistics, University of Aarhus, Denmark, 1973.

[11] I. Guttman, *The use of the concept of a future observation in goodness-of-fit problems*, J. Roy. Statist. Soc. Ser. B **29** (1967), 83–100.

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

[13] J. T. Hwang and M.-C. Yang, *An optimality theory for mid p-values in 2×2 contingency tables* , Statist. Sinica **11** (2001), 807–826.

[14] H. Jeffreys, Theory of Probability, 3rd edition. London: Oxford University Press, 1967.

[15] E. L. Lehmann, Testing Statistical Hypotheses, 2nd edition, Springer, New York, 1997.

[16] A. Lewin, S. Richardson, C. Marshall, A. Glazier and T. Aitman, *Bayesian modeling of differential gene expression*, Biometrics **62** (2006), 1–9.

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

[18] X. L. Meng, *Posterior predictive p-values*, Ann. Statist. **22** (1994), 1142–1160.

[19] D. B. Rubin, *Bayesianly justifiable and relevant frequency calculations for the applied statistician*, Ann. Statist. **12** (1984), 1151–1172.

[20] T. Sellke, M. J. Bayarri and J. O. Berger, *Calibration of p-values for testing precise null hypotheses*, Amer. Statist. **55** (2001), 62–71.

[21] S. Suissa and J. J. Shuster, *The* 2×2 *matched-pairs trial: Exact unconditional design and analysis*, Biometrics **47** (1991), 361–372.

Department of Applied Statistics and Information Science, Ming Chuan University, Taoyuan, Taiwan, R.O.C.

E-mail: lschen@mail.mcu.edu.tw

Graduate Institute of Statistics, National Central University, Chung-Li, Taiwan, R.O.C.

E-mail: yang@stat.ncu.edu.tw