Some tests for detecting trends based on the modified Baumgartner-Weiß-Schindler statistics

Guogen Shan[1], Changxing Ma[1], Alan D. Hutson[1], and Gregory E. Wilding[1] *

[1] *Department of Biostatistics,University at Buffalo, 3435 Main Street, Buffalo, NY 14214*

July 7, 2011

Abstract

We propose a modified nonparametric Baumgartner-Weiß-Schindler test and investigate its use in testing for trends among K binomial populations. Exact conditional and unconditional approaches to p-value calculation are explored in conjunction with the statistic in addition to a similar test statistic proposed by Neuhäuser [24], the unconditional approaches considered including the maximization approach [6], the confidence interval approach [8], and the E+M approach [17]. The procedures are compared with regards to actual Type I error and power and examples are provided. The conditional approach and the E+M approach performed well, with the E+M approach having an actual level much closer to the nominal level. The E+M approach and the conditional approach are generally more powerful than the other p-value calculation approaches in the scenarios considered. The power difference between the conditional approach and the E+M approach is often small in the balance case. However, in the unbalanced case, the power comparison between those two approaches based on our proposed test statistic show that the E+M approach has higher power than the conditional approach.

Keywords: Baumgartner-Weiß-Schindler test, Exact Tests, E+M p-value, Test for trend, Unconditional test.

*Corresponding author. Department of Biostatistics,University at Buffalo, 3435 Main Street, Buffalo, NY, 14214, USA.

E-mail address:gwilding@buffalo.edu

1 Introduction

In categorical data analysis the problem of testing the equality of K binomial proportions against an ordered alternative has been studied for many years. The data structure corresponding to the problem can be presented as a $2 \times K$ contingency table where the binary outcome variable is represented by the rows and the column variable is ordinal in nature. Consider a dose response study where subjects are randomized to different doses of the same experimental compound and then observed for response. Let n_i be the number of subjects enrolled in the i-th group corresponding to dose d_i, $i = 1, 2, \cdots, K$, such that $d_1 < d_2 < \cdots < d_K$. It is reasonable in many scenarios to assume that if there does indeed exist an effect of the drug, the probability of response is a non-decreasing function of dosage. The null hypothesis of interest is then

$$H_0 : p_1 = p_2 = \cdots = p_K =: p,$$

which is tested against an ordered alternative of the form

$$H_a : p_1 \leq p_2 \leq \cdots \leq p_K \quad \text{and} \quad p_1 < p_K, \tag{1}$$

where p_i is the probability of the response at dose d_i, $i = 1, 2, \cdots, K$. Several testing procedures have been proposed for investigating binomial trends such as the Cochran-Armitage (CA) test (Cochran [10], Armitage [3]). The commonly used CA test is based upon the estimated regression coefficient from the weighted linear regression of the observed proportions of success on the fixed scores corresponding to the ordinal predictor. Several methods for p-value calculation in conjunction with the CA test statistic have been proposed; e.g. see Shan et al. [25] for a review of existing methods and an outline of a new unconditional approach.

Under the alternative hypothesis, we would expect the collection of ordinal variable levels (the d_i's) for the responders to be generally larger than the non-responders. Therefore the problem of testing for trend may be alternatively addressed through use of one of the many one-sided two-sample test statistics available. A linear rank test such as the Wilcoxon rank-sum test [27, 19, 28] can be applied to test H_0 against H_a by treating the observations as two-sample multinomial data where the multinomial variable takes on one of K levels corresponding to the d_i's [16]. In the case of an equal number of subjects corresponding to the ordinal groups and equally space d_i's , the Wilcoxon rank-sum statistic is mathematically identical to the CA statistic [14]. Other two-sample tests

may also be considered for testing for trends among binomial proportions.

A rank based test for comparing locations of two continuous populations was proposed by Baumgartern, Weiß, and Schindler (BWS) [7]. This nonparametric test is based on the squared value of the difference between the two empirical distribution functions weighted by the respective variance. This weighting places more emphasize on the tails of the distribution functions and was used in the popular Anderson Darling goodness-of-fit test [2]. Baumgartern et al. studied the asymptotic distribution of the test statistic under normality and showed that the BWS test is at least as powerful as well known nonparametric tests, such as the Wilcoxon rank sum test [13, 28], the Kolmogorov-Smirnov test [13], and the Cramer-von Mises test [1]. A permutation approach in conjunction with the BWS test statistic was suggested by Neuhäuser [22]. Neuhäuser showed that the exact version of the test is generally associated with a more accurate Type I error control and more power as compared to the exact Wilcoxon test for a variety of continuous distributions, and especially so when the underlying populations are exponentially distributed. In the case of normally distributed data, the difference in power between the BWS test and Wilcoxon test was negligible. Since the BWS test is not suitable for a one-sided alternative hypothesis, a modified BWS test was later proposed by Neuhäuser [23] for the two-sample problem and K-sample problem with ordered alternatives. In the context of detecting ordered alternative with K location-shift populations, the proposed test was compared with the commonly used Jonckheere-Terpstra test [15, 26]. The numerical studies revealed that the exact Jonckheere-Terpstra test is both more conservative and often times less powerful than the modified BWS test. In addition, the power of the modified BWS test was seen to be more consistent with respect to the actual trend pattern.

More recently, the use of the modified BWS test statistic with categorical outcomes has been explored. With ordinal populations the exact permutation versions of the modified BWS and Wilcoxon tests were compared with regards to the Type I error rate and simulated power for one-sided alternatives [22]; The exact modified BWS test was shown to be superior to the Wilcoxon test. Later on, Neuhäuser [24] applied the modified BWS test for detecting trends among binomial proportions. The procedure, based on conditioning upon the observed marginals for computing significance, was compared with the asymptotic CA test and the exact conditional CA test via Monte Carlo simulations. Both exact conditional tests guaranteed the nominal level, while the asymptotic CA

test suffered consistently from an inflated Type I error rate. Between the two exact tests, the test based on the BWS test statistic was found to be less conservative and associated with higher power than that based on the CA test statistic.

Murakami [20] made an additional modification to the BWS test, referred to as the BWS_M test, for use with continuous populations. This modification involved the use of a test statistic, which is a function of the exact mean and variance of ordered rank statistics. The properties of the BWS_M test were compared with the t-test, the Wilcoxon rank sum test, the Kolmogorov-Smirnov test, the Cramer-von Mises test, the Anderson-Darling test, and the original BWS test. The BWS test showed similar power as the BWS_M test in most cases, but the BWS_M test demonstrated a much higher power than that of the BWS test in some settings, especially in cases with unequal sample sizes. Analogues of the two-sample tests for the K-sample problem were compared based on simulations. Similar conclusions were found. Thus far, the BWS_M test has only been studied under the assumption of continuous underlying distributions.

Exact tests are generally preferred in small sample studies due to their guaranteed Type I error rates, but receive criticism due to the required computational resources. More recently the computational issue is of less concern given improvements in computing performance. The exact conditional modified BWS test for trends for binary outcomes was recommended by Neuhäuser [24], but the approach can be highly conservative when the sample sizes are small or the true proportions are at the $0-1$ boundaries. As an alternative to the tests described above, exact unconditional approaches may be explored, as they are usually less discrete and thus providing Type I error rates closer to the desired level. The commonly used unconditional approach is maximization, where the p-value is maximized over the range of the nuisance parameters. When spikes appear in the plot of the p-value versus the nuisance parameter, the maximization approach may be more conservative than the conditional approach. For this reason a partial maximization approach was developed by Berger and Boos [8] in order to reduce the conservatism of the unconditional test. Their method is based on maximization over a $100(1-\gamma)\%$ confidence region for the nuisance parameter instead of the full range. The estimated p-value is obtained by replacing the unknown nuisance parameter of the null distribution with its maximum likelihood estimate (MLE). The E+M p-value proposed by Lloyd [17] is then obtained by maximizing the p-value quantity using the estimated

p-value as the test statistic. All of these unconditional methods for p-value calculation have yet to be explored when used in conjunction with the test statistic proposed by Neuhäuser [24] for the purpose of detecting trends.

The rest of this article is organized as follows. In Section 2, we review the BWS test and its modifications in further detail and propose an a new modified statistic for two types of statistical problems. We briefly review the proposed procedures for detecting binomial trends in Section 3. In Section 4 examples are used to illustrate the tests and in Section 5 we compare the performance of the competing tests, studying the exact test size and power results under a wide range of conditions. Section 6 is given to conclusions.

2 BWS statistics family

In this section we review previous work relevant to the problem of interest and develop the newly proposed test statistic.

2.1 Two-sided two-sample tests

A nonparametric two-sample test for determining whether the two samples are from the same population was proposed by Baumgartern, Weiß, and Schindler (BWS) [7]. Let the sample corresponding to one population be denoted by $\underset{\sim}{Y} = (Y_1, Y_2, \cdots, Y_{m_1})'$ and let $\underset{\sim}{Z} = (Z_1, Z_2, \cdots, Z_{m_2})'$ denote the sample for a second population. The hypotheses are

$$H_0 : F = G,$$

$$H_a : F \neq G,$$

where F and G represent the cumulative distribution functions corresponding to the random variables Y and Z, respectively. The metric $(\hat{F}(w) - \hat{G}(w))^2$ weighted inversely by $w(1-w)$ was used to construct the testing statistic given as

$$T(\underset{\sim}{Y}, \underset{\sim}{Z}) = \frac{m_1 m_2}{m_1 + m_2} \int_0^1 \frac{1}{w(1-w)} (\hat{F}(w) - \hat{G}(w))^2 dw, \tag{2}$$

where $\hat{F}(.)$ and $\hat{G}(.)$ are the empirical distribution functions. The statistic $T(\underset{\sim}{Y}, \underset{\sim}{Z})$ may be approximated by

$$B(\underset{\sim}{Y}, \underset{\sim}{Z}) = \frac{1}{2}(B_{\underset{\sim}{Y}} + B_{\underset{\sim}{Z}}), \tag{3}$$

where

$$B_{\underset{\sim}{Y}} = \frac{1}{m_1}\sum_{j=1}^{m_1}\frac{(R_j - \frac{m_1+m_2}{m_1}j)^2}{\frac{j}{m_1+1}(1-\frac{j}{m_1+1})\frac{m_2(m_1+m_2)}{m_1}},$$

$$B_{\underset{\sim}{Z}} = \frac{1}{m_2}\sum_{l=1}^{m_2}\frac{(H_l - \frac{m_1+m_2}{m_2}l)^2}{\frac{l}{m_2+1}(1-\frac{l}{m_2+1})\frac{m_1(m_1+m_2)}{m_2}},$$

and $R_j, j = 1, 2, \cdots, m_1$, and $H_l, l = 1, 2, \cdots, m_2$, are the ranks of the samples from the first and second populations, respectively, from the combined samples. Large values of $B(\underset{\sim}{Y}, \underset{\sim}{Z})$ would support the alternative hypothesis. Baumgartern et al. showed that their test is easy to apply and is at least as powerful as the Wilcoxon rank-sum test. Although the null distribution of $B(\underset{\sim}{Y}, \underset{\sim}{Z})$ converges to the limiting case extremely fast in the continuous setting as shown via simulation, the use of critical values from the limiting distribution results in an inflated Type I error rate for small sample sizes [22].

It is well known that the expected value and variance of R_j and H_l [4] are

$$E(R_j) = \frac{m_1+m_2+1}{m_1+1}j,$$

$$Var(R_j) = \frac{j}{m_1+1}(1-\frac{j}{m_1+1})\frac{m_2(m_1+m_2+1)}{m_1+2},$$

$$E(H_l) = \frac{m_1+m_2+1}{m_2+1}l,$$

and

$$Var(H_l) = \frac{l}{m_2+1}(1-\frac{l}{m_2+1})\frac{m_1(m_1+m_2+1)}{m_2+2},$$

respectively. Note that Murakami [20] proposed a modification of the BWS test statistic incorporating these quantities, specifically,

$$B^*(\underset{\sim}{Y}, \underset{\sim}{Z}) = \frac{1}{2}(B^*_{\underset{\sim}{Y}} + B^*_{\underset{\sim}{Z}}), \tag{4}$$

where

$$B^*_{\underset{\sim}{Y}} = \frac{1}{m_1}\sum_{j=1}^{m_1}\frac{(R_j - \frac{m_1+m_2+1}{m_1+1}j)^2}{\frac{j}{m_1+1}(1-\frac{j}{m_1+1})\frac{m_2(m_1+m_2+1)}{m_1+2}},$$

and

$$B^*_{\underset{\sim}{Z}} = \frac{1}{m_2}\sum_{l=1}^{m_2}\frac{(H_l - \frac{m_1+m_2+1}{m_2+1}l)^2}{\frac{l}{m_2+1}(1-\frac{l}{m_2+1})\frac{m_1(m_1+m_2+1)}{m_2+2}}.$$

Murakami [20] demonstrated that the test based on the statistic $B^*(\underset{\sim}{Y}, \underset{\sim}{Z})$ has similar power to that of the BWS test for most location-shift alternatives. However, the modified test is more powerful than the BWS test in the

case of unequal sample sizes in the two-sample setting, as well as for the K-sample problem for the general alternative when using the test statistic, $\frac{1}{K}\sum_{q=1}^{K}\frac{1}{m_q}\sum_{v=1}^{m_q}B^*_{qv}$, where B^*_{qv} denotes the quantity of the form (4) for the q-th and v-th populations.

Similar to the test based on $B(\underset{\sim}{X},\underset{\sim}{Z})$, the test based on $B^*(\underset{\sim}{X},\underset{\sim}{Z})$ is associated with an inflated Type I error in the finite sample setting when critical values from the asymptotic null distribution are used. It can be easily seen that the asymptotic distribution of the $B^*(\underset{\sim}{X},\underset{\sim}{Z})$ is the same as the $B(\underset{\sim}{X},\underset{\sim}{Z})$ statistic.

2.2 One-sided tests

In the two-sample problem of detecting if one population is stochastically larger than the other, neither the test based on $B(\underset{\sim}{X},\underset{\sim}{Z})$ nor the test based on $B^*(\underset{\sim}{X},\underset{\sim}{Z})$ is appropriate. Neuhäuser [23] modified the original BWS test for the one-sided hypothesis test of the form

$$H_0 : F(u) = G(u), u \in R,$$

$$H_a : F(u) \geq G(u), u \in R,$$

with strict inequality for some point u. He presented the test statistic

$$B^{\alpha}(\underset{\sim}{X},\underset{\sim}{Z}) = \frac{1}{2}(B^{\alpha}_{\underset{\sim}{Z}} - B^{\alpha}_{\underset{\sim}{X}}), \tag{5}$$

where

$$B^{\alpha}_{\underset{\sim}{X}} = \frac{1}{m_1}\sum_{j=1}^{m_1}\frac{(R_j - \frac{m_1+m_2}{m_1}j)|R_j - \frac{m_1+m_2}{m_1}j|}{\frac{j}{m_1+1}(1-\frac{j}{m_1+1})\frac{m_2(m_1+m_2)}{m_1}},$$

and

$$B^{\alpha}_{\underset{\sim}{Z}} = \frac{1}{m_2}\sum_{l=1}^{m_2}\frac{(H_l - \frac{m_1+m_2}{m_2}l)|H_l - \frac{m_1+m_2}{m_2}l|}{\frac{l}{m_2+1}(1-\frac{l}{m_2+1})\frac{m_1(m_1+m_2)}{m_2}}.$$

Since $B^{\alpha}_{\underset{\sim}{Z}}$ is an increasing function of the location difference between the two groups, and $B^{\alpha}_{\underset{\sim}{X}}$ is a decreasing function under location-shift alternatives, large values of the test statistic support the alternative.

In the context of categorical data analysis, Neuhäuser [21] investigated the use of the midrank when ties are present and examined the Type I error control of the original BWS test when used in conjunction with ordinal data. He demonstrated that the more ties that are present, the more the Type I error rate is inflated. Neuhäuser applied the $B^{\alpha}(\underset{\sim}{X},\underset{\sim}{Z})$ test statistic to ordinal data [21], as well as to binomial data [24]. The one-sided test statistic

was shown to be less conservative, and often more powerful than the one-sided Wilcoxon rank-sum test for both continuous distributions and categorical distributions [21], partially due to the more discrete null distribution of the Wilcoxon exact test statistic.

Following Murakami [20], we propose a new modification of the BWS test statistic. In the context of the one-sided two-sample problem, the test statistic is of the form

$$B^{\beta}(\underset{\sim}{Y},\underset{\sim}{Z}) = \frac{1}{2}(B^{\beta}_{\underset{\sim}{Z}} - B^{\beta}_{\underset{\sim}{Y}}), \tag{6}$$

where

$$B^{\beta}_{\underset{\sim}{Y}} = \frac{1}{m_1}\sum_{j=1}^{m_1}\frac{(R_j - \frac{m_1+m_2+1}{m_1+1}j)|R_j - \frac{m_1+m_2+1}{m_1+1}j|}{\frac{j}{m_1+1}(1-\frac{j}{m_1+1})\frac{m_2(m_1+m_2+1)}{m_1+2}}$$

and

$$B^{\beta}_{\underset{\sim}{Z}} = \frac{1}{m_2}\sum_{l=1}^{m_2}\frac{(H_l - \frac{m_1+m_2+1}{m_2+1}l)|H_l - \frac{m_1+m_2+1}{m_2+1}l|}{\frac{l}{m_2+1}(1-\frac{l}{m_2+1})\frac{m_1(m_1+m_2+1)}{m_2+2}}.$$

The new test statistic $B^{\beta}(\underset{\sim}{Y},\underset{\sim}{Z})$ can be used for data with or without ties; in the case of ties the $R_j^{'}$s and $H_l^{'}$s are defined to be the midranks. The asymptotic distribution of the test statistic $B^{\beta}(\underset{\sim}{Y},\underset{\sim}{Z})$ is not easily derived. However, an exact permutation test can readily be performed in order to calculate the p-value conditional on a given data set. The study of the properties of the above test statistic when comparing two continuous populations will be the subject of a future manuscript and we restrict the rest of this note to the examination of the statistic when in use for detecting trends among binomial populations.

2.3 Test for binomial trends

As noted earlier, when testing for trends among K binomial proportions, as an alternative to the commonly used CA test, one-sided two-sample procedures may be employed. By redefining $\underset{\sim}{Y}$ and $\underset{\sim}{Z}$ to be samples representing the ordinal group levels of the non-responders and responders, respectively, the test statistics $B^{\alpha}(\underset{\sim}{Y},\underset{\sim}{Z})$ and $B^{\beta}(\underset{\sim}{Y},\underset{\sim}{Z})$ may be used. Let the test statistics for use in this context be denoted as $B^{\alpha}(\tilde{x})$ and $B^{\beta}(\tilde{x})$, where $\tilde{x} = (x_1, x_2, \cdots, x_K)$ is the vector of the numbers of responses for the K binomial samples, x_i being the number of responses within the $i-$th ordinal group. For the newly proposed test statistic, in terms of the notation of the problem of interest, the statistic is of the form

$$B^{\beta}(\tilde{x}) = \frac{1}{2}(B^{\beta}_1 - B^{\beta}_0), \tag{7}$$

where

$$B_0^{\beta} = \frac{1}{a_0} \sum_{j=1}^{a_0} \frac{(R_j - \frac{a_0+a_1+1}{a_0+1} j)|R_j - \frac{a_0+a_1+1}{a_0+1} j|}{\frac{j}{a_0+1}(1 - \frac{j}{a_0+1}) \frac{a_1(a_0+a_1+1)}{a_0+2}}$$

and

$$B_1^{\beta} = \frac{1}{a_1} \sum_{l=1}^{a_1} \frac{(H_l - \frac{a_0+a_1+1}{a_1+1} l)|H_l - \frac{a_0+a_1+1}{a_1+1} l|}{\frac{l}{a_1+1}(1 - \frac{l}{a_1+1}) \frac{a_0(a_0+a_1+1)}{a_1+2}}.$$

where $a_1 = \sum_{i=1}^{K} x_i, a_0 = \sum_{i=1}^{K} (n_i - x_i)$ are the total number of responses and non-responses, respectively, and R_j and H_l are now redefined to be the midranks corresponding to the ordinal group levels. In the balanced scenario, the midrank for the i-th ordinal group is $(n+1)/2 + (i-1)n$, $i = 1, 2, \cdots, K$, where n is the common sample size corresponding to each ordinal group. Similar formula can be defined for the test statistic $B^{\alpha}(\tilde{x})$.

The exact conditional approach [11], fixing both marginal totals, and the asymptotic approach were used to compare the performance of the $B^{\alpha}(\tilde{x})$ and the CA test statistics in a previous work [24]. The exact tests were seen to be preferable due to the guaranteed Type I error rate, and furthermore, the exact conditional $B^{\alpha}(\tilde{x})$ test was seen to be more powerful than exact conditional CA test in most cases. Given these results we further study exact approaches based on the test statistics $B^{\alpha}(\tilde{x})$ and $B^{\beta}(\tilde{x})$.

3 P-value calculation for $B^{\alpha}(\tilde{x})$ and $B^{\beta}(\tilde{x})$

Due to the intractable asymptotic distribution for the test statistics $B^{\alpha}(\tilde{x})$ and $B^{\beta}(\tilde{x})$, we investigate alternative exact approaches for obtaining significance in this article. Four methods are considered here and described in more detail below: (1) the conditional approach; (2) the maximization approach; (3) the confidence interval approach; and (4) the estimation and maximization approach. The conditional and unconditional methodologies based on the test statistic $T(\tilde{x})$, taken to be either $B^{\alpha}(\tilde{x})$ or $B^{\beta}(\tilde{x})$, will be evaluated for when detecting the ordered alternative in the $2 \times K$ contingency table. The test statistic $T(\tilde{x})$ will support the alternative for large values. Let $\tilde{x_0}$ be the observed vector of responses for a given data set, $S(\tilde{x_0})$ be the total number of responders, and the value of the test statistic T for this table be denoted as $T(\tilde{x_0})$.

One commonly used approach for p-value calculation in the analysis of contingency tables is to condition on the observed marginal totals. The reference distribution is obtained by calculating the value and corresponding probability of the test statistic T for all the possible tables with the same marginal values based on hypergeometric

probabilities. The p-value is then defined as the sum of the probabilities of tables with observations $\tilde{x}$, such that $T(\tilde{x}) \geq T(\tilde{x_0})$. The exact conditional p-value for testing for trend is

$$P_C(t) = Pr(T(\tilde{x}) \geq T(\tilde{x_0})|S(\tilde{x_0}) = s, H_0) = \sum_{\tilde{x} \in \Omega_C(\tilde{x_0})} \frac{\prod_{i=1}^{K} \binom{n_i}{x_i}}{\binom{n}{s}},$$

where $\Omega_C(\tilde{x_0}) = \{\tilde{x} : T_C(\tilde{x}) \geq T_C(\tilde{x_0}) \text{ and } S(\tilde{x}) = s\}$ is the rejection region of the test.

For the general categorical problem, the Type I error rate of the conditional approach has been shown to be far below the nominal level, for example, see Shan et al. [25]. An approach for reducing the conservatism of the test in this context is within the unconditional framework, where only one set of marginal totals is fixed. The appropriateness of such an approach is rooted in the study design, for example, contingency tables where the column totals represent fixed sample sizes. Given we consider a fixed number of subjects are to be allocated to each ordinal group, an unconditional approach may be seen as a reasonable alternative. A commonly used unconditional approach is conducted through use of maximization (the M approach) [5, 6]. In this case, the p-value (referred to as the M p-value) is defined to be the supremum of the p-value function over the whole range of the parameter space,

$$P_M(\tilde{x_0}) = \max_{0 \leq p \leq 1} \{ \sum_{\tilde{x} \in \Omega_M(\tilde{x_0})} \prod_{i=1}^{K} \binom{n_i}{x_i} p^{x_i} (1-p)^{n_i - x_i} \}.$$

The M approach is sometimes challenging due to the computational intensity of the maximization step. However, a grid search algorithm is usually a reasonable method to find the maximum for a given data set.

If the nuisance parameter is unbounded maximization is further complicated. Berger and Boos [8] proposed a new approach, maximizing the tail probability over the confidence interval for the nuisance parameter instead of the whole range. For a given penalty value γ, this confidence interval p-value (referred to as the CI p-value) is defined as the supremum of the p-value function over $C(\tilde{x_0})$ plus the penalty value, where $C(\tilde{x_0})$ is the $100(1-\gamma)\%$ confidence interval of the parameter estimated using $\tilde{x_0}$. The CI p-value is defined by

$$P_{CI}(\tilde{x_0}) = \sup_{p \in C(\tilde{x_0})} \{ \sum_{\tilde{x} \in \Omega_{CI}(\tilde{x_0})} \prod_{i=1}^{K} \binom{n_i}{x_i} p^{x_i} (1-p)^{n_i - x_i} \} + \gamma,$$

where $\Omega_{CI}(\tilde{x_0}) = \Omega_M(\tilde{x_0})$, and $C(\tilde{x_0})$ is the Clopper-Pearson interval [9]. Small values of γ are desirable, and in this note we take $\gamma = 0.001$. This methodology has been applied to test for trend in a $2 \times K$ table using the CA test statistic, see Freidlin and Gastwirth [12].

The estimation approach is an older and is a much easier computational approach. The idea is to replace the nuisance parameter in the null likelihood with the maximum likelihood estimate (MLE) of the parameter under H_0. For a given data set $\tilde{x_0}$, the rejection region for this approach is

$$R_E(\tilde{x_0}) = \{\tilde{x} : \{T(\tilde{x}) \geq T(\tilde{x_0}))\}.$$

The simply calculated p-value, referred to as the E p-value, is

$$P_E(\tilde{x_0}) = \sum_{\tilde{x} \in R_E(\tilde{x_0})} \prod_{i=1}^{K} \binom{n_i}{x_i} \hat{p}^{x_i} (1-\hat{p})^{n_i - x_i},$$

where $\hat{p}$ is the value of maximum likelihood estimate of p under the null hypothesis, $\hat{p} = \sum_{i=1}^{K} x_i / \sum_{i=1}^{K} n_i$. While the E p-value is not valid and not exact, an exact p-value can be obtained after a maximization step, which is the basis of the method proposed by Lloyd [17]. In this case the E p-value is taken to be the test statistic. The corresponding rejection region of the so called E+M p-value for the test is

$$\Omega_{E+M}(\tilde{x_0}) = \{\tilde{x} : P_E(\tilde{x}) \leq P_E(\tilde{x_0})\},$$

and the E+M p-value is given by

$$P_{E+M}(\tilde{x_0}) = \sup_{p \in [0,1]} \{ \sum_{\tilde{x} \in \Omega_{E+M}(\tilde{x_0})} \prod_{i=1}^{K} \binom{n_i}{x_i} p^{x_i} (1-p)^{n_i - x_i} \}.$$

The E+M p-value $P_{E+M}(\tilde{x})$ is exact, while the CI p-value $P_{CI}(\tilde{x})$ and the E p-value $P_E(\tilde{x})$ are not [17].

The conditional approach based on $B^{\alpha}(\tilde{x})$ for testing for trends among ordered binomial proportions was investigated by Neuhäuser [24], and was shown to be less conservative and more powerful than that based on the CA test statistic. The unconditional approaches mentioned above (M, CI, E+M approaches) based on the modified one-sided BWS test statistics, and the conditional approach based on the test statistic $B^{\beta}(\tilde{x})$ have not been studied in this context. Before investigating the Type I error and power properties of these procedures, we illustrate the approaches through use of some examples.

4 Examples

We consider an inhalation study by Malley et al. [18]. The aim of the inhalation study was to determine the potential toxicity and/or potential neurotoxicity of cyclohexane. Groups of rats were exposed to 0, 500, 2000,

or 7000 ppm of the compound. Here, the control group was exposed to air alone. In the 90-day toxicity study, rats of 12 per group were studied. Both males and females had significantly increased incidence of stained chin when testing for the alternative of the form (1) using the CA test. The same trend was observed in the 90-day neurotoxicity study which was based on 20, 10, 10, and 20 rats in group 1 through 4, respectively. The data from both studies, which consisted of multiple binary endpoints each, is given in Table 1.

For the sixteen data examples the p-value profile as a function of the nuisance parameter was calculated using the four different approaches outlined in Section 3. Figure 1 shows the observed null likelihood based on the new test statistic $B^{\beta}(\tilde{x})$ as a function of the nuisance parameter under the null hypothesis. In 12 out of these 16 cases, the E+M approach has a smaller p-value than the M approach's; see Table 2. In the majority of cases, the E+M approach results in a much smaller p-value than that obtained via the conditional approach. The corresponding plots for the test statistic $B^{\alpha}(\tilde{x})$ are shown in the Figure 2, and as can be seen, for the majority of cases the plots are similar between the two test statistics. We focus specific attention to the comparison of the tests based on the first data set in Table 1 (Figures 1 a and 2 a). At the 0.05 significance level, the conclusions are not consistent among the tests. There is a big spike in the null likelihood graph based on $B^{\alpha}(\tilde{x})$ for the M approach, which results in a p-value greater than 0.1. However, all other p-values based on either test statistic are less than 0.05. The CI p-value based on $B^{\beta}(\tilde{x})$ is closer to the M p-value, but the CI p-value based on $B^{\alpha}(\tilde{x})$ is smaller than the M p-value in the plot. The E+M p-value is the smallest among these p-values based on either test statistic.

5 Method comparison

To evaluate the performance of the exact conditional and unconditional tests, we compare their actual significance levels as well as exact powers at the 0.05 nominal level. The computation is based on complete enumeration and no simulation is involved.

5.1 Type I error rate

Figure 3 shows the actual Type I error rates of the tests as a function of the value of the common binomial parameter under the null hypothesis. We studied the balanced case with sample sizes 10, 20, 30 and 50 for each group (1st, 2nd, 3rd, and 4th row, respectively, in Figure 3). Plots in the first column (plots (a), (c), (e), (g))

show the actual Type I error rates of the four exact tests based on test statistic $B^{\alpha}(\tilde{x})$ with $K = 3$. The plots in the second column are based on the test statistic $B^{\beta}(\tilde{x})$. There is a slight difference between the plots based on different test statistics. Large spikes are observed in the case of small sample sizes in the plots for the CI approach, whose maximization over the whole range is greater than the nominal level. The other three approaches preserve the nominal level of the test for all the cases. As shown in the Figure 3, the actual Type I error rate of the E+M approach is much closer to the nominal level as compared to the other approaches. This advantage is most apparent when the sample sizes are small and at two ends of the parameter range (close to 0 or 1). Although not presented here, similar results are obtained for larger K.

The Type I error rate comparisons with unequal sample sizes are shown on the Figure 4. The plots in the left side of the figure are based on the test statistic $B^{\alpha}(\tilde{x})$ and the right side has results for $B^{\beta}(\tilde{x})$. Two different sets of sample sizes (n_1, n_2, n_3) were considered for the $K = 3$ scenario: (8,9,12) which appears in plots (a) and (b), respectively, and (8,12,10) which correspond to the plots (c) and (d), respectively. As we can seen from the figure, there are very slight differences in the Type I error rates between the two test statistics. The interesting result is that the M approach is extremely conservative in these unbalanced cases, as is the CI approach. The conditional approach and the E+M approach performed well, with the E+M approach having an actual level much closer to the nominal level. Similar results are seen for larger values of K which are not provided in this manuscript.

5.2 Power study

Having examined the true Type I error rate of each test, we now compare the procedures with respect to power. The power of testing procedures based on the test statistic $B^{\alpha}(\tilde{x})$ and our proposed test statistic $B^{\beta}(\tilde{x})$ were investigated for different true alternatives and results appear in Figures 5, 6, and 7. In Figures 5 and 6, the study was based on $K = 4$ with balanced sample sizes of 10 and 20 per group, respectively. The two alternatives studied were those considered by Neuhäuser [24]: (1) $p_1 = 0.1, p_2 = p_3 = 0.1 + x/2, p_4 = 0.1 + x$, and (2) $p_1 = 0.1, p_2 = p_3 = p_4 = 0.1 + x$, where $0 < x < 0.9$. Unbalanced case results are given in Figure 7 with $K = 3$ and sample sizes (10,12,10). The alternatives considered here were (1) $p_1 = 0.1, p_2 = 0.1 + x/4, p_3 = 0.1 + x$, and (2) $p_1 = 0.1, p_2 = 0.1 + x/2, p_3 = 0.1 + x$, where $0 < x < 0.9$. From the graphs of the balanced case, the conditional approach and the E+M approach are almost uniformly more powerful than the other p-value

calculation approaches. The power difference between the conditional approach and the E+M approach is often small in the balance case, as is the difference in power between the M and CI approach. In the unbalanced case, there is no clear winner between the conditional approach and the E+M approach based on the test statistic $B^{\alpha}(\tilde{x})$. However, the power comparison between those two approaches based on $B^{\beta}(\tilde{x})$ show that the E+M approach has higher power than the conditional approach.

6 Conclusions

In this article we proposed a modified test statistic based on the BWS test denoted $B^{\beta}(\tilde{x})$. An exact conditional approach based on $B^{\alpha}(\tilde{x})$ was studied by Neuhäuser [24], which was shown to have superior properties to the well known CA test, and we also considered such an approach for p-value calculation in this note. Furthermore, we applied the two commonly used unconditional approaches, the M and CI approach, as well as the recently proposed E+M approach due to Lloyd [17]. Maximizing the p-value based on using the estimated p-value as the test statistic resulted in an exact test, guaranteeing the nominal level of the test. Although all considered procedures were satisfactory, we recommend use of the E+M approach in conjunction with $B^{\alpha}(\tilde{x})$ or $B^{\beta}(\tilde{x})$ for use in practice.

The focus of this manuscript has been on the problem of comparing K binomial populations versus an ordered alternative. The exploration of the use of the $B^{\beta}(\underset{\sim}{Y}, \underset{\sim}{Z})$ test statistic for comparing two continuous or ordinal populations with an one-sided alternative is currently underway. In addition, one may consider a test of trend among K continuous or ordinal populations using the test statistic (following the approach of Jonckheere [15] and Terpstra [26])

$$T_{B^{\beta}} = \sum_{q=1}^{K-1} \sum_{v=q+1}^{K} B_{qv}^{\beta}$$

where B_{qv}^{β} is the proposed two-sample test statistic of the form (6) for comparing q-th and v-th populations.

References

[1] T. W. Anderson. On the Distribution of the Two-Sample Cramer-von Mises Criterion. *The Annals of Mathematical Statistics*, 33(3):1148–1159, September 1962.

[2] T. W. Anderson and D. A. Darling. Asymptotic Theory of Certain "Goodness of Fit" Criteria Based on Stochastic Processes. *The Annals of Mathematical Statistics*, 23(2):193–212, June 1952.

[3] P. Armitage. Tests for Linear Trends in Proportions and Frequencies. *Biometrics*, 11(3):375–386, 1955.

[4] B. C. Arnold, N. Balakrishnan, and H. N. Nagaraja. *A First Course in Order Statistics (Wiley Series in Probability and Statistics)*. Wiley-Interscience.

[5] G. A. Barnard. Significance Tests for 22 Tables. *Biometrika*, 34(1/2), 1947.

[6] D. Basu. On the Elimination of Nuisance Parameters. *Journal of the American Statistical Association*, 72(358):355–366, 1977.

[7] W. Baumgartner, P. Weiß, and H. Schindler. A Nonparametric Test for the General Two-Sample Problem. *Biometrics*, 54(3):1129–1135, 1998.

[8] R. L. Berger and D. D. Boos. P Values Maximized Over a Confidence Set for the Nuisance Parameter. *Journal of the American Statistical Association*, 89(427):1012–1016, 1994.

[9] C. J. Clopper and E. S. Pearson. The Use of Confidence or Fiducial Limits Illustrated in the Case of the Binomial. *Biometrika*, 26(4):404–413, 1934.

[10] W. G. Cochran. Some methods for strengthening the common χ^2 tests. *Biometrics*, 10(4):417–451, 1954.

[11] R. A. Fisher. *The Design of Experiments*. Macmillan Pub Co, 9 edition, 1935.

[12] B. Freidlin and J. L. Gastwirth. Unconditional Versions of Several Tests Commonly Used in the Analysis of Contingency Tables. *Biometrics*, 55(1):264–267, 1999.

[13] J. D. Gibbons and S. Chakraborti. *Nonparametric Statistical Inference (Statistics: A Series of Textbooks and Monographs)*. CRC Press, 3rd rev/ex edition, May 1992.

[14] J. F. Hilton. The appropriateness of the wilcoxon test in ordinal data. *Statist. Med.*, 15(6):631–645, 1996.

[15] A. R. Jonckheere. A Distribution-Free k-Sample Test Against Ordered Alternatives. *Biometrika*, 41(1/2):133–145, 1954.

[16] S.-H. Jung and S.-H. Kang. Tests for 2K contingency tables with clustered ordered categorical data. *Statist. Med.*, 20(5):785–794, 2001.

[17] C. J. Lloyd. Exact p-values for discrete models obtained by estimation and maximization. *Australian and New Zealand Journal of Statistics*, 50(4):329–345, 2008.

[18] L. A. Malley, J. R. Bamberger, J. C. Stadler, G. S. Elliott, J. F. Hansen, T. Chiu, J. S. Grabowski, and K. L. Pavkov. Subchronic toxicity of cyclohexane in rats and mice by inhalation exposure. *Drug and chemical toxicology*, 23(4):513–537, November 2000.

[19] H. B. Mann and D. R. Whitney. On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics*, 1947.

[20] H. Murakami. A k-sample rank test based on modified Baumgartner statistic and its power comparison. *Journal of the Japanese Society of Computational Statistics*, 19(1):1–13, December 2006.

[21] M. Neuhäuser. Exact tests based on the Baumgartner-Weiß-Schindler statistic A survey. *Statistical Papers*, 46(1):1–29, 2005.

[22] M. Neuhäuser. An exact two-sample test based on the baumgartner-weiß-schindler statistic and a modification of lepage's test. *Communications in Statistics - Theory and Methods*, 29(1):67–78, 2000.

[23] M. Neuhäuser. One-sided two-sample and trend tests based on a modified baumgartner-weiß-schindler statistic. *Journal of Nonparametric Statistics*, 13(5):729–739, 2001.

[24] M. Neuhäuser. An exact test for trend among binomial proportions based on a modified Baumgartner-Weiß-Schindler statistic. *Journal of Applied Statistics*, 33(1):79–88, 2006.

[25] G. Shan, C. Ma, A. Hutson, and G. Wilding. An efficient and exact approach for detecting trends with binary endpoints. Technical report, 2011.

[26] T. J. Terpstra. The asymptotic normality and consistency of Kendall's test against trend, when ties are present in one ranking. *Indigationes Mathematicae*, 14:327–333, 1952.

[27] F. Wilcoxon. Individual Comparisons by Ranking Methods. *Biometrics Bulletin*, 1(6):80–83, 1945.

[28] G. E. Wilding, G. Shan, and A. D. Hutson. Exact Two-Stage Designs for Phase II Clinical Trials with Rank-Based Endpoints. Technical report, Department of Biostatistics, University at Buffalo, 2010.

Table 1: Sixteen data sets from Malley et al.

Data	Observation	Gender	Response			
90-day nuerotoxicity study with sample sizes (20,10,10,20)						
a	Stained face	M	0	0	1	3
b	Stained chin	M	0	0	1	10
c	Wet chin	M	1	1	0	8
d	Stained face	F	1	7	8	3
e	Stained chin	F	0	1	4	9
f	Wet chin	F	0	0	0	11
90-day rat study with sample sizes (12,12,12,12)						
g	Colored discharge mouth	M	0	0	1	18
h	Stained chin	M	0	0	0	15
i	Stained perineum	M	0	1	1	0
j	Wet chin	M	0	0	0	9
k	Wet perinuem	M	0	0	1	1
l	Colored discharge mouth	F	0	0	0	16
m	Stained chin	F	0	0	0	11
n	Stained perineum	F	1	2	1	4
o	Wet chin	F	0	0	0	6
p	Wet perinuem	F	1	2	5	14

Table 2: Comparison of p-values of tests using sixteen data sets based on test statistic $B^{\beta}(\tilde{x})$.

	E+M approach	M approach	CI approach	Conditional approach
E+M approach		12	15	15
M approach			14	10
CI approach				1

Numeric value is the number of data sets out of the sixteen total where the row test's p-value is less than the column test's p-value.

Figure 1: P-value profiles based on test statistic $B^{\beta}(\tilde{x})$.

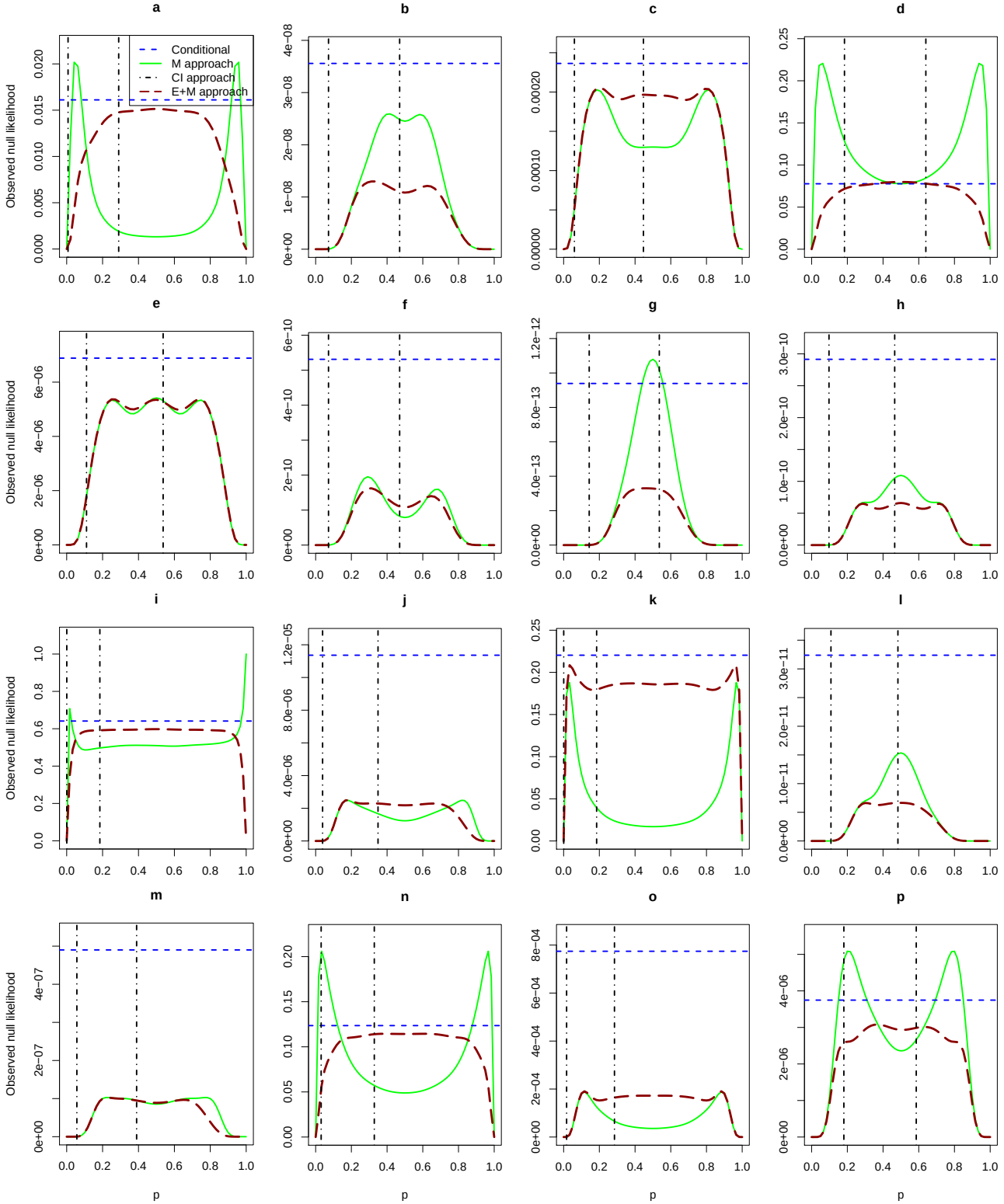

Figure 2: P-value profiles based on test statistic $B^{\alpha}(\tilde{x})$.

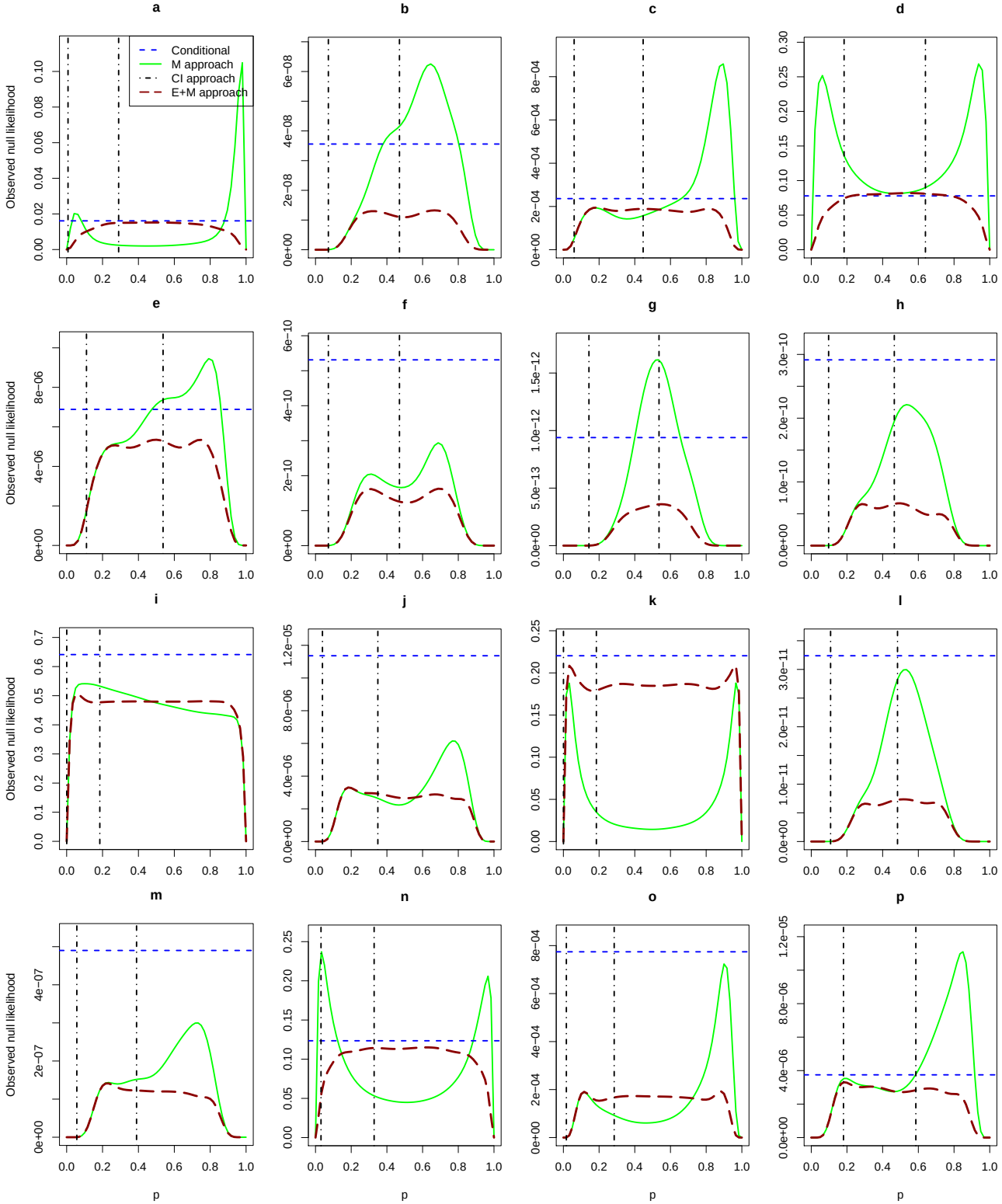

Figure 3: Actual Type I error of the tests under the null hypothesis with balanced sample sizes and $K = 3$ for test statistics $B^{\alpha}(\tilde{x})$ (left) and $B^{\beta}(\tilde{x})$ (right).

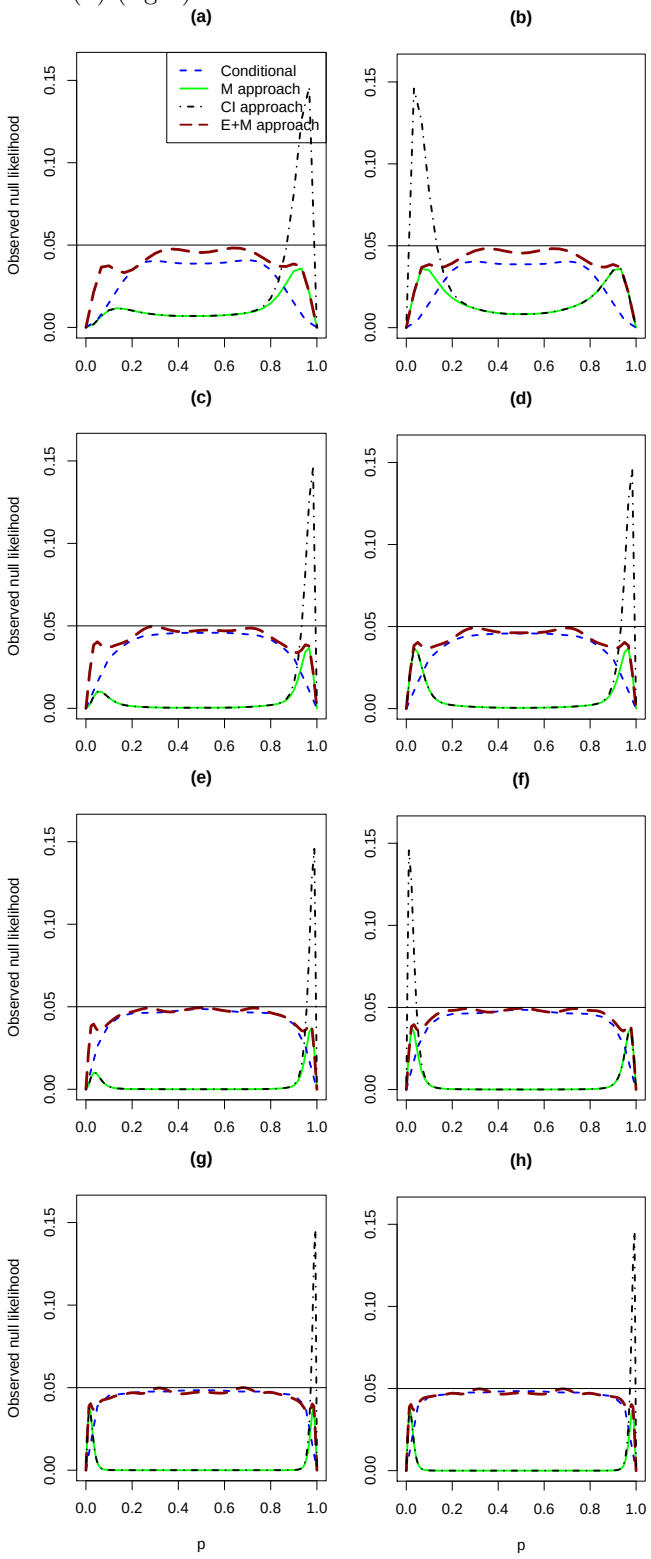

Figure 4: Actual Type I error of the tests under ther null hypothesis with unequal sample sizes and $K = 3$ for test statistics $B^{\alpha}(\tilde{x})$ (left) and $B^{\beta}(\tilde{x})$ (right).

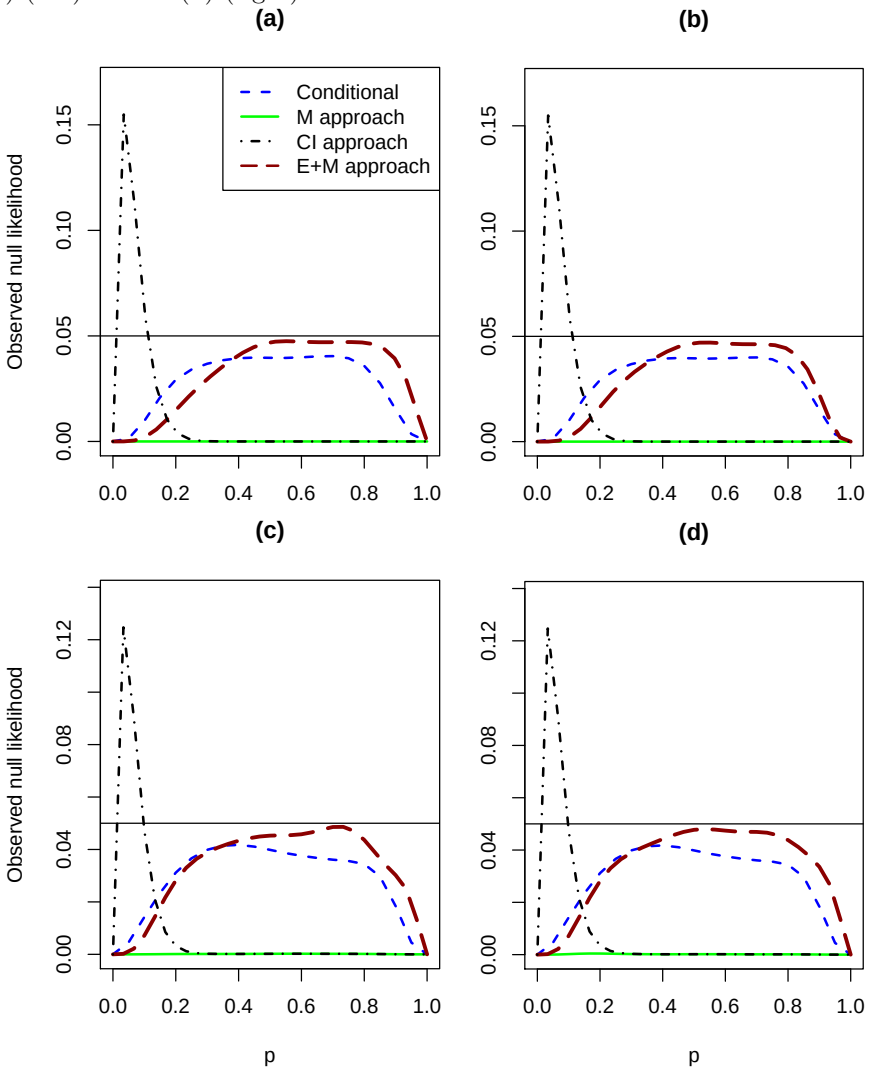

Figure 5: Power study with $K = 4$, $n = 10$ per group for test statistics $B^{\alpha}(\tilde{x})$ (left) and $B^{\beta}(\tilde{x})$ (right), under the alternative $p_1 = 0.1, p_2 = 0.1 + x/2, p_3 = 0.1 + x/2, p_4 = 0.1 + x$ (the first row) and $p_1 = 0.1, p_2 = 0.1 + x, p_3 = 0.1 + x, p_4 = 0.1 + x$ (the second row).

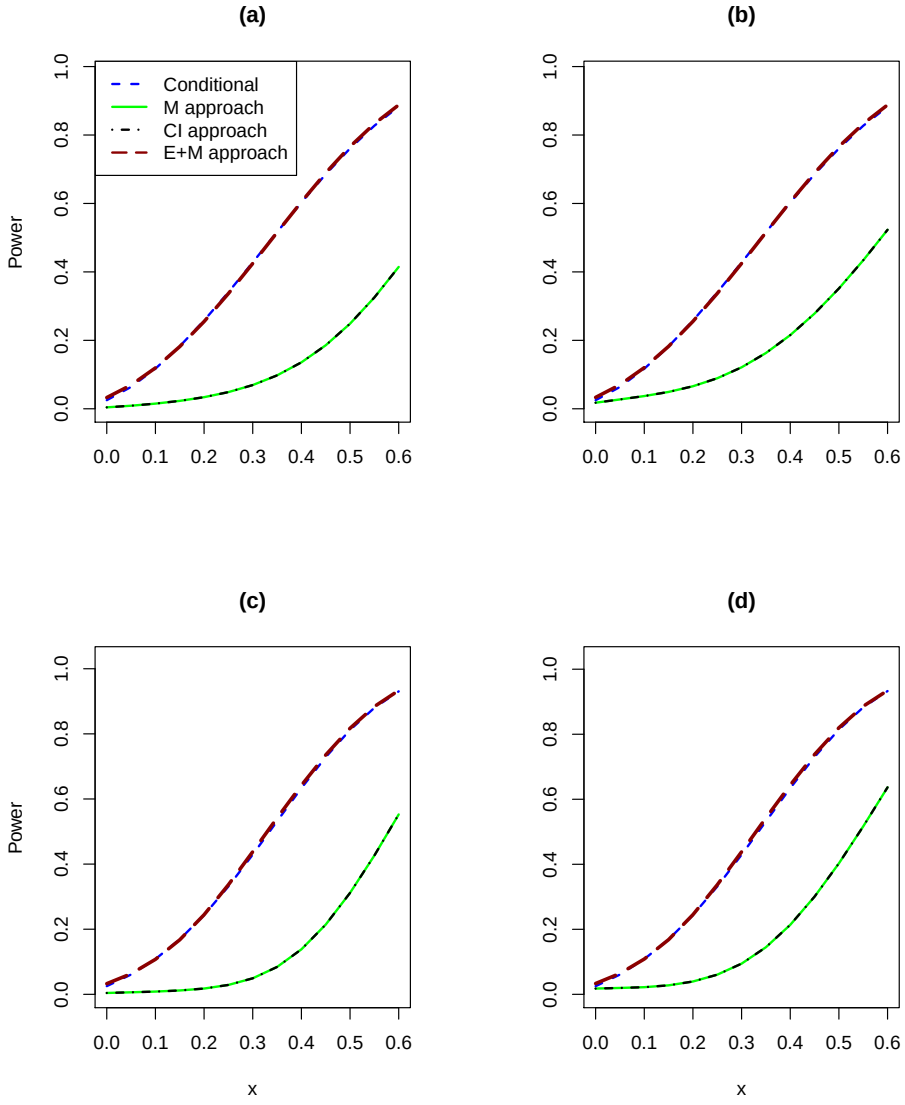

Figure 6: Power study with $K = 4$, $n = 20$ per group for test statistics $B^{\alpha}(\tilde{x})$ (left) and $B^{\beta}(\tilde{x})$ (right), under the alternative $p_1 = 0.1, p_2 = 0.1 + x/2, p_3 = 0.1 + x/2, p_4 = 0.1 + x$ (the first row) and $p_1 = 0.1, p_2 = 0.1 + x, p_3 = 0.1 + x, p_4 = 0.1 + x$ (the second row).

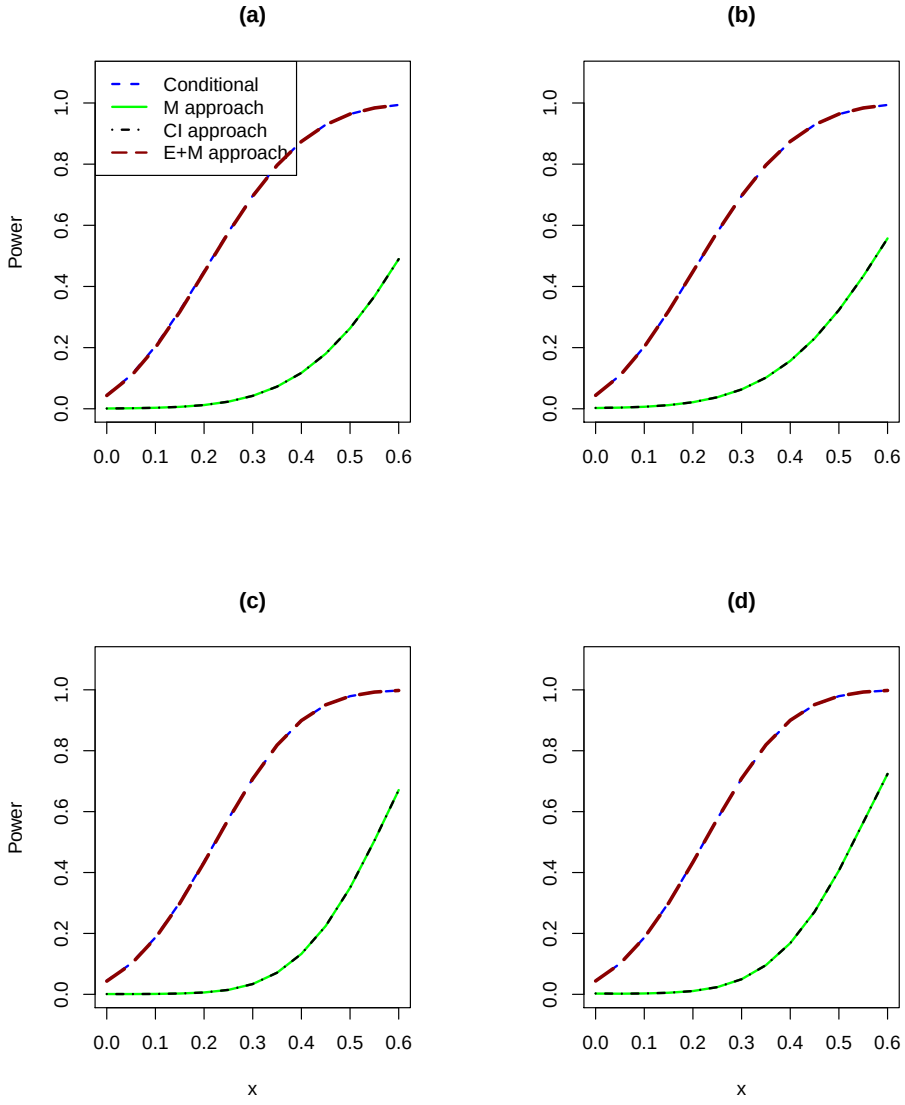

Figure 7: Power study with $K = 3$, $n = (10, 12, 10)$ for test statistics $B^{\alpha}(\tilde{x})$ (left) and $B^{\beta}(\tilde{x})$ (right), under the alternative $p_1 = 0.1, p_2 = 0.1 + x/2, p_3 = 0.1 + x$ (the first row) and $p_1 = 0.1, p_2 = 0.1 + x/4, p_3 = 0.1 + x$ (the second row).