

# A more powerful unconditional exact test of homogeneity for $2 \times c$ contingency table analysis

Louis Ehwerhemuepha, Heng Sok & Cyril Rakovski

To cite this article: Louis Ehwerhemuepha, Heng Sok & Cyril Rakovski (2019) A more powerful unconditional exact test of homogeneity for  $2 \times c$  contingency table analysis, Journal of Applied Statistics, 46:14, 2572-2582, DOI: [10.1080/02664763.2019.1601689](https://doi.org/10.1080/02664763.2019.1601689)

To link to this article: <https://doi.org/10.1080/02664763.2019.1601689>



© 2019 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group



Published online: 06 Apr 2019.



Submit your article to this journal [↗](#)



Article views: 1521



View related articles [↗](#)



View Crossmark data [↗](#)



Citing articles: 6 View citing articles [↗](#)

# A more powerful unconditional exact test of homogeneity for $2 \times c$ contingency table analysis

Louis Ehwerhemuepha<sup>a,b</sup>, Heng Sok<sup>a</sup> and Cyril Rakovski<sup>a</sup>

<sup>a</sup>School of Computational and Data Science, Chapman University, Orange, CA, USA; <sup>b</sup>Children's Hospital of Orange County, Orange, CA, USA

## ABSTRACT

The classical unconditional exact  $p$ -value test can be used to compare two multinomial distributions with small samples. This general hypothesis requires parameter estimation under the null which makes the test severely conservative. Similar property has been observed for Fisher's exact test with Barnard and Boschloo providing distinct adjustments that produce more powerful testing approaches. In this study, we develop a novel adjustment for the conservativeness of the unconditional multinomial exact  $p$ -value test that produces nominal type I error rate and increased power in comparison to all alternative approaches. We used a large simulation study to empirically estimate the 5th percentiles of the distributions of the  $p$ -values of the exact test over a range of scenarios and implemented a regression model to predict the values for two-sample multinomial settings. Our results show that the new test is uniformly more powerful than Fisher's, Barnard's, and Boschloo's tests with gains in power as large as several hundred percent in certain scenarios. Lastly, we provide a real-life data example where the unadjusted unconditional exact test wrongly fails to reject the null hypothesis and the corrected unconditional exact test rejects the null appropriately.

## ARTICLE HISTORY


Received 28 September 2018  
Accepted 26 March 2019

## KEYWORDS

Unconditional multinomial exact  $p$ -value test; empirical type I error; power; small samples

## 1. Introduction

Contingency tables are used to display sample data arising from given distributions with respect to either categories defined by characteristics inherent to the underlying distributions or by external factor variables. These tables facilitate subsequent analysis focused on the presence of relationships among the parameters of the distributions imposed on the data or among the external classification variables. In particular, a large sample comparison of  $r$  ( $r \geq 2$ ) multinomial distributions with  $c$  ( $c \geq 2$ ) categories is implemented via the classical chi-square testing procedure that contrasts the observed and expected cell counts for all samples under the null as shown by Agresti [2], Fisher [9], and Yates [19,20]. In the following study we will use the classical  $r \times c$  contingency table notation where  $n_{ij}$  denotes the element of the table on row  $i$  and column  $j$ , and corresponds to the observed number of

**CONTACT** Louis Ehwerhemuepha  [lehwerhemuepha@choc.org](mailto:lehwerhemuepha@choc.org)

$j$ th level observations in the  $i$ th sample and  $n_i$  denotes the  $i$ th sample size and  $n_j$  denotes the total number of the  $j$ th level observations  $i = 1, 2, \dots, r, j = 1, 2, \dots, c$ .

For a completely specified null hypothesis that the samples arise from the same specified multinomial distribution with category probabilities  $p_1, p_2, \dots, p_c$ , the test statistic  $\sum_{i=1}^r \sum_{j=1}^c (n_{ij} - e_{ij})^2 / e_{ij}$ , where  $e_{ij} = p_j n_i$  is the expected  $(i, j)$ th cell count under the null, follows a  $\chi^2$  distribution with  $r(c - 1)$  degrees of freedom. Similarly, for a general null hypothesis that the samples arise from the same but unspecified multinomial distribution, the test statistic  $\sum_{i=1}^r \sum_{j=1}^c (n_{ij} - \hat{e}_{ij})^2 / \hat{e}_{ij}$ , where  $\hat{e}_{ij} = \hat{p}_j n_i$  is the estimated expected cell count on row  $i$  and column  $j$  under the null, follows a  $\chi^2$  distribution with degrees of freedom (compared to the prior scenario) reduced by  $(c - 1)$ , and thus equal  $(r - 1)(c - 1)$ . Here, the reduction of the degrees of freedom is necessitated by the number of estimated parameters under the null  $\hat{p}_j = \sum_{i=1}^r n_{ij} / \sum_{i=1}^r \sum_{j=1}^c n_{ij}, j = 1, 2, \dots, c - 1$ . Via this adjustment, the unspecified null hypothesis is also naturally accommodated by the  $\chi^2$  distribution. These testing procedures are inappropriate for small sample or sparse data analyses, rather an implementation of exact testing procedures is required. According to Mehta et al. [15], small sample or sparse data is usually defined as settings where the expected contingency table cell counts do not exceed 5. The small sample testing alternatives include: the multinomial extension of the Fisher's exact tests that conditions on the margins of the contingency table by Fisher [8], and Mehta and Patel [14], permutation tests that conditions on the observed data by Efron and Tibshirani [7], and Hastie and Tibshirani, and lastly the classical unconditional exact  $p$ -value test (UEPT) that explicitly enumerates all possible contingency tables compatible with the observed data, evaluates their corresponding probabilities under the null and obtains an exact  $p$ -value by adding the probabilities of all data configurations as likely or less likely to occur than that of the observed data under the null [11–13].

The first two procedures reduce the computational complexity by condition on the margins of the table or on the observed data. However, in certain scenarios as indicated in Agresti [1], Fisher's Exact test has been known to be conservative and several approaches that correct for the level of conservativeness have been proposed by Barnard [4,5], Boschloo [6], Lin and Yang [10], Röhmel and Mansmann [17], and Routledge [18]. In a recent work, Oliveira et al. [16] showed that the exact likelihood ratio approach possesses performance advantages over Barnard's test and extended its use to handle the hypothesis of independence as well.

In this work we investigate the effect of parameter estimation on the type I error rates of the classical UEPT when applied to a general (unspecified) comparison of several multinomial distributions in small samples. We also provide the necessary correction for the conservativeness that adjusts the type I error rate to nominal levels with substantial increase in power. As mentioned above, in the case of a not-completely specified null hypothesis for equality of several multinomial distributions,  $c - 1$  parameters need to be estimated from the data in order to obtain the necessary parameters of the common distribution under the null. There is a natural test adjustment for the estimation of these parameters in large sample analysis – the decrease in degrees of freedom of the  $\chi^2$  distribution. However, the classical exact  $p$ -value test is distribution free and therefore does not allow such a straightforward mode of adjustment. Consequently, we derive regression-based estimations of the level of conservativeness and the 5th percentile of the exact  $p$ -values under the null including sample sizes, number of multinomial categories and their interactions. We provide a

real-world application of the results of our study by applying our corrections to ‘uncorrected’  $p$ -values from a multinomial test of homogeneity obtained in a study on the effect of exercise on the biobehavioral outcomes of fatigue during cancer treatment by Al-Majid et al. [3].

## 2. Methods

Formally, both the completely specified and the general null hypotheses that several independent samples (represented in rows of the corresponding contingency table) are drawn from the same multinomial distributions (the latter known as hypothesis of homogeneity) can respectively be written as

$$H_0 : p_{1i} = p_{2i} = \dots = p_{ri} = p_{0i}, \quad i = 1, 2, \dots, c, \tag{1}$$

when the  $p_{0i}$ ’s are given and,

$$H_0 : p_{1i} = p_{2i} = \dots = p_{ri}, \quad i = 1, 2, \dots, c. \tag{2}$$

In large samples, the corresponding test statistics are,

$$\sum_{j=1}^r \sum_{i=1}^c \frac{(n_{ij} - e_{ij})^2}{e_{ij}} \sim \chi^2(r(c - 1)) \tag{3}$$

and

$$\sum_{j=1}^r \sum_{i=1}^c \frac{(n_{ij} - \hat{e}_{ij})^2}{\hat{e}_{ij}} \sim \chi^2((r - 1)(c - 1)). \tag{4}$$

As discussed earlier, the asymptotic parametric case naturally adjusts the transition between hypotheses (1) and (2) by subtracting  $(c - 1)$  degrees of freedom from the  $\chi^2$  distribution due to the estimation of the  $(c - 1)$  common category probabilities of the underlying multinomial distribution,

$$\hat{p}_j = \frac{\sum_{i=1}^r n_{ij}}{\sum_{i=1}^r \sum_{j=1}^c n_{ij}}, \quad j = 1, 2, \dots, c - 1. \tag{5}$$

Moreover, in the latter case of a general unspecified null hypothesis, if one fails to adjust the degrees of freedom by subtracting the number of estimated parameters under the null, the resulting  $\alpha = 0.05$  type I error rate will be conservative as the integral below has the correct limits but reflects the area under the wrong curve,

$$T_I = \int_0^{\chi_{0.05}^2((r-1)(c-1))} f_{r(c-1)}(x) dx, \tag{6}$$

where  $f_{r(c-1)}(x) = [2^{r(c-1)/2} \Gamma(r(c - 1)/2)]^{-1} x^{r(c-1)/2-1} e^{-x/2}$ ,  $x > 0$  is the pdf of  $\chi^2$  distribution with  $r(c - 1)$  degrees of freedom.

In small sample and sparse data scenarios, both asymptotic approaches (1) and (2) are inapplicable. The properties of the multinomial extension of the Fisher’s exact test

and permutation tests have previously been studied and are well-known. Here, we are particularly interested in studying the classical UEPT type I error rates encountered under the second type of hypotheses. In the subsequent presentation, we assume that  $X_i \sim Mult(n_i, p_{i1}, p_{i2}, \dots, p_{ic}), i = 1, \dots, r$  and equate every observed or possible  $k$  sample multinomial data with  $c$  categories with the corresponding contingency table  $T$  (representing its cross-classification representation). Given hypotheses (1) and (2), the classical unconditional exact test calculates a  $p$ -value associated with the observed data by adding the probabilities of all contingency tables that occur with probabilities not exceeding that of the observed data,

$$P_{\text{exact}} = \sum_i P(T_i)I\{P(T_i) \leq P(T_{\text{obs}})\}, \tag{7}$$

where  $T_i$  enumerates all possible contingency tables. Furthermore, all probabilities are calculated either under the completely specified null hypothesis (1) or under the general null hypothesis that the samples come from the same unspecified multinomial distribution (2). The bijective correspondence between multinomial samples and contingency tables allows us to immediately see that the number of possible values for  $X_i, i = 1, 2, \dots, r$  equals,

$$N_i = \binom{n_i + c - 1}{c} \tag{8}$$

and therefore the number of possible contingency tables associated with  $r$  multinomial samples with  $c$  categories and sample sizes  $n_1, n_2, \dots, n_r$  is given by

$$N = \prod_{i=1}^r \binom{n_i + c - 1}{c}. \tag{9}$$

The probability of each of these  $N$  tables can be directly calculated under hypothesis (1),

$$P(T) = \prod_{i=1}^r \frac{n_i!}{n_{i1}!n_{i2}! \dots n_{ic}!} p_{10}^{n_{i1}} p_{20}^{n_{i1}} \dots p_{c0}^{n_{ic}} \tag{10}$$

and estimated under hypothesis (2) after having performed calculations (5),

$$\hat{P}(T) = \prod_{i=1}^r \frac{n_i!}{n_{i1}!n_{i2}! \dots n_{ic}!} \hat{p}_1^{n_{i1}} \hat{p}_2^{n_{i1}} \dots \hat{p}_c^{n_{ic}}. \tag{11}$$

The absence of parametric distribution and its natural degrees of freedom adjustment between hypotheses (1) and (2) presents an interesting problem when the exact  $p$ -values test (7) is used for hypothesis (1) via probabilities (10) and for (2) via probabilities (11). The exact test possesses nominal type I error rates under the completely specified null hypotheses but the effect of parameter estimation (necessary under the general null) on the type I error rates is unknown. We implemented an extensive simulation study in an effort to empirically estimate the type I error rates and the correct 5th percentile of the exact  $p$ -value distribution under the general null hypothesis under a range of sample sizes, number of multinomial categories and various underlying null hypotheses. Further, we

**Table 1.** Simulation configuration.

$c$	$(n_1, n_2)$	$(p_1, p_2, \dots, p_c)$	Configurations
2	(3, 3), (3, 7), (3, 10)		45
2	(3, 15), (3, 25), (7, 7)		
2	(7, 10), (7, 15), (7, 25)	(0.1, 0.9), (0.3, 0.7), (0.5, 0.5)	
2	(10, 10), (10, 15), (10, 25)		
2	(15, 15), (15, 25), (25, 25)		20
3	(3, 3), (3, 7), (3, 10)	(0.1, 0.1, 0.8), (0.1, 0.3, 0.6)	
3	(3, 15), (3, 25)	(0.25, 0.25, 0.50), (0.33, 0.33, 0.34)	
3	(7, 7), (7, 10), (7, 15)	(0.1, 0.1, 0.8), (0.1, 0.3, 0.6)	
3	(7, 25), (10, 10), (10, 15)	(0.33, 0.33, 0.34)	27
3	(10, 25), (15, 15), (15, 25)		
3	(25, 25)	(0.1, 0.1, 0.8), (0.33, 0.33, 0.34)	2
4	(3, 3), (3, 7), (3, 10)	(0.1, 0.1, 0.1, 0.7), (0.1, 0.1, 0.3, 0.5)	
4	(3, 15), (3, 25), (7, 7)	(0.1, 0.3, 0.3, 0.3), (0.25, 0.25, 0.25, 0.25)	32
4	(7, 10), (7, 15)		

averaged the simulated data over the unobservable (under (2) null hypothesis) probabilities and used linear regression modeling to obtain the best predictive models for the type I error rates and the 5th percentile of the exact  $p$ -value distribution in an effort to derive an exact test alternative to the asymptotic formula (6). Lastly, the computational complexity of the exact  $p$ -value method combined with the simulation on tens of thousand instances makes this study possible only for  $r = 2$ , and  $c = 2, 3, 4$ . However, these happen to be the most common scenarios that occur in data analysis.

### 3. Simulation design

We carried out a large-scale simulation study by varying sample sizes, number of multinomial categories and null hypotheses probabilities in an effort to empirically estimate the type I error rates for the UEPT used for the general multinomial null hypothesis (2) and the 5th percentile of the distribution. In particular, the following combinations of number of categories, samples sizes, and a set of corresponding probabilities were implemented in our study design as shown in Table 1. We varied the actual values that define the grid and showed that the performance of the method is robust to such changes.

For each particular combination of parameters,  $10^4$  datasets were simulated, the exact  $p$ -value calculated, the corresponding empirical type I error rate at  $\alpha$ -level of 0.05, and the 5th percentile of the distribution of  $p$ -values were obtained. Consequently, the study was based on a total of  $1.26 \times 10^6$  simulated contingency tables. Below is an enumeration of the steps we took for clarity to the reader:

- (1) Set the number of multinomial categories,  $c$
- (2) Set the sample sizes,  $n_1$  and  $n_2$
- (3) Set the multinomial category probabilities  $p_1^{(1)} = p_1^{(2)}, p_2^{(1)} = p_2^{(2)}, \dots, p_c^{(1)} = p_c^{(2)}$
- (4) Generate 10,000 multinomially distributed  $2 \times c$  tables using values from steps 1 to 3.
- (5) Calculate the unconditional multinomial exact  $p$ -value under the null (since  $p_i^{(1)} = p_i^{(2)}$  for  $i = 1, 2, \dots, c$ )
- (6) Estimate the empirical 5th percentiles of the distributions of the  $p$ -values,  $P_{0.05}$
- (7) Repeat steps one to six for all combination of parameters shown in Table 1

**Table 2.** Estimation of the 5th percentile of the fully unconditional multinomial exact test as an  $\alpha$ -level correction for conservativeness of the test.

Coefficients	$\beta$	Standard error	$t$ -Value	$p$ -Value
$n_1$	0.0095	0.00450	2.10	.04
$n_2$	-0.0039	0.00230	-1.66	.10
$c$	0.0950	0.00410	23.01	< .01
$n_1:c$	-0.0054	0.00170	-3.25	< .01
$n_2:c$	0.0015	0.00080	1.89	.06

- (8) Use the empirical 5th percentiles of the distributions of the  $p$ -values,  $P_{0.05}$ , as the dependent variable in a linear regression model with the number of multinomial categories  $c$ , sample sizes  $n_1$  and  $n_2$ , and all two way statistical interactions. Pick the best model from all possible models.

## 4. Results

### 4.1. Regression-based adjustment for the 5th percentiles UEPT

Our results indicate that in all scenarios defined by number of samples, number of multinomial categories, and particular choices of the multinomial category probabilities under the null, the type I error rates of the unadjusted unconditional multinomial exact test were severely conservative and, consequently, the 5th percentiles of the distributions of  $p$ -values of the exact test were severely inflated in comparison to the nominal 0.05 level. The probabilities of the multinomial categories are unknown and could not appear in the regression as predictors. We simulated data over a range of multinomial category probabilities but integrated over these unknown parameters by averaging the distributions of the  $p$ -values over the scenarios with common sample sizes and number of multinomial categories. In essence, we used the empirically estimated 5th percentiles of the distribution of  $p$ -values of the exact test as the outcome variable of interest in a best subset linear regression model building step with the sample sizes and number of categories and all their possible interactions as candidate predictors. The resulting model is the basis of our proposed adjustment for the conservativeness of the unconditional multinomial exact  $p$ -value test, and it depends on the sample sizes and number of multinomial categories. The regression formula of the best model is shown in Equation (12), where  $n_1 \leq n_2 \leq 30$  and  $2 \leq c \leq 5$ . In other words, whenever  $n_1 \neq n_2$ ,  $n_1$  is chosen to be the smaller of the two. Additional details of the regression fit such as standard errors and  $p$ -values are shown in Table 2.

$$P_{0.05} = 0.0095n_1 - 0.0039n_2 + 0.095c - 0.0054n_1c + 0.0015n_2c. \tag{12}$$

We used the predicted 5th percentile of the exact  $p$ -value tests (using Equation (12)) as the value for  $\alpha$  and estimated the type I error rates from data simulated under the null. The average type I error rate was 0.06 indicating that our proposed adjustment corrects the conservativeness of the multinomial exact  $p$ -value test. We proceeded to simulate data under the alternative (by setting  $p_i^{(1)} \neq p_i^{(2)}$ , for any  $i = 1, 2, \dots, c$ ) in order to estimate the resulting gain in statistical power.

**Table 3.** Average power comparisons with Fisher’s, Barnard’s, and Boschloo’s test for  $2 \times 2$  tables.

$(n_1, n_2)$	$p_1^{(1)}$	$p_1^{(2)}$	Power				Power gain (%) over			
			Fisher	Barnard	Boschloo	Multinomial	$\alpha^a$	Fisher	Barnard	Boschloo
(3, 3)	0.10	0.20	0	0.005	0.005	0.095	0.183	–	1636	1636
(3, 3)	0.15	0.30	0	0.018	0.018	0.167	0.183	–	820	820
(3, 3)	0.25	0.50	0	0.056	0.056	0.056	0.183	–	0	0
(3, 3)	0.20	0.80	0	0.268	0.268	0.268	0.183	–	0	0
(3, 15)	0.10	0.20	0.001	0.011	0.011	0.015	0.173	1227	35	35
(3, 15)	0.15	0.30	0.003	0.012	0.012	0.039	0.173	1404	215	215
(3, 15)	0.25	0.50	0.025	0.073	0.073	0.137	0.173	452	88	88
(3, 15)	0.20	0.80	0.441	0.639	0.639	0.66	0.173	50	3	3
(7, 10)	0.10	0.20	0.016	0.019	0.018	0.084	0.172	414	346	358
(7, 10)	0.15	0.30	0.051	0.055	0.055	0.139	0.172	171	150	152
(7, 10)	0.25	0.50	0.137	0.144	0.156	0.158	0.172	15	10	2
(7, 10)	0.20	0.80	0.651	0.680	0.732	0.854	0.172	31	26	17
(10, 10)	0.10	0.20	0.011	0.043	0.043	0.177	0.168	1521	311	311
(10, 10)	0.15	0.30	0.035	0.086	0.086	0.135	0.168	290	57	57
(10, 10)	0.25	0.50	0.084	0.170	0.170	0.170	0.168	101	0	0
(10, 10)	0.20	0.80	0.634	0.802	0.802	0.91	0.168	44	13	13
(15, 15)	0.10	0.20	0.039	0.083	0.094	0.142	0.157	263	70	51
(15, 15)	0.15	0.30	0.076	0.132	0.141	0.156	0.157	104	19	10
(15, 15)	0.25	0.50	0.161	0.282	0.258	0.388	0.157	141	38	50
(15, 15)	0.20	0.80	0.871	0.946	0.944	0.976	0.157	12	3	3
(15, 25)	0.10	0.20	0.053	0.096	0.096	0.126	0.148	138	31	31
(15, 25)	0.15	0.30	0.118	0.164	0.164	0.173	0.148	47	6	6
(15, 25)	0.25	0.50	0.280	0.325	0.346	0.369	0.148	32	14	7
(15, 25)	0.20	0.80	0.968	0.979	0.976	0.988	0.148	2	1	1

<sup>a</sup>Adjusted  $\alpha$ -level, multinomial exact test.

**4.2. Average power gain by the corrected unadjusted exact p-value test (CUEPT)**

We estimated the average power of the CUEPT over a range of alternative hypothesis. As mentioned earlier, simulation under the alternative hypothesis was achieved by setting  $p_i^{(1)} \neq p_i^{(2)}$ , for any  $i = 1, 2, \dots, c$ . We show the power gain obtained in comparison to Fisher’s, Barnard’s, and Boschloo’s test for  $2 \times 2$ ,  $2 \times 3$ , and  $2 \times 4$  tables in Tables 3, 4, and 5 respectively.

In all cases considered, the average power of the CUEPT increased with increase in the values of  $n_1$  and  $n_2$ . The gain in average power was greatest for pair of the smallest sample size, (3, 3), compared to Fisher’s, Barnard’s, Boschloo’s and the multinomial UEPT. The adjustment to the multinomial exact test led to equal or substantial percent gain in power over the other tests. As sample size increased, the power of all tests/methods increased but the gain in power due to our adjustment decreased.

**5. Real data example**

In a study on the effects of exercise on behavioral outcomes of fatigue during cancer treatment, Al-Majid et al. [3] randomly assigned 14 women (who completed the study) into 2 groups consisting of subjects who were assigned to exercise, and those who receive ‘usual care’ as the control group. The study was a randomized prospective longitudinal study on cancer patients from two cancer centers in Virginia and Southern California. Inclusion criteria for the study included female patients 21 years or older who have been diagnosed with Stage I or II breast cancer and scheduled to receive chemotherapy. They described the



**Table 4.** Statistical power comparisons with Fisher’s test for  $2 \times 3$  tables.

$(n_1, n_2)$	$(p_1^{(1)}, p_2^{(1)})$	$(p_1^{(2)}, p_2^{(2)})$	Power		$\alpha^a$	Power gain (%)
			Fisher	Multinomial		
(3, 3)	(0.10, 0.20)	(0.20, 0.30)	0	0.087	0.267	–
(3, 3)	(0.15, 0.50)	(0.30, 0.30)	0	0.060	0.267	–
(3, 3)	(0.20, 0.10)	(0.80, 0.10)	0	0.469	0.267	–
(3, 3)	(0.25, 0.40)	(0.50, 0.10)	0	0.088	0.267	–
(3, 15)	(0.10, 0.20)	(0.20, 0.30)	0.013	0.045	0.274	245
(3, 15)	(0.15, 0.50)	(0.30, 0.30)	0.047	0.060	0.274	30
(3, 15)	(0.20, 0.10)	(0.80, 0.10)	0.523	0.721	0.274	38
(3, 15)	(0.25, 0.40)	(0.50, 0.10)	0.185	0.292	0.274	58
(7, 10)	(0.10, 0.20)	(0.20, 0.30)	0.066	0.128	0.244	93
(7, 10)	(0.15, 0.50)	(0.30, 0.30)	0.082	0.136	0.244	66
(7, 10)	(0.20, 0.10)	(0.80, 0.10)	0.721	0.881	0.244	22
(7, 10)	(0.25, 0.40)	(0.50, 0.10)	0.226	0.333	0.244	47
(10, 10)	(0.10, 0.20)	(0.20, 0.30)	0.089	0.136	0.224	53
(10, 10)	(0.15, 0.50)	(0.30, 0.30)	0.105	0.150	0.224	43
(10, 10)	(0.20, 0.10)	(0.80, 0.10)	0.784	0.922	0.224	18
(10, 10)	(0.25, 0.40)	(0.50, 0.10)	0.234	0.355	0.224	52
(15, 15)	(0.10, 0.20)	(0.20, 0.30)	0.135	0.152	0.193	13
(15, 15)	(0.15, 0.50)	(0.30, 0.30)	0.159	0.186	0.193	17
(15, 15)	(0.20, 0.10)	(0.80, 0.10)	0.943	0.975	0.193	3
(15, 15)	(0.25, 0.40)	(0.50, 0.10)	0.382	0.480	0.193	26
(15, 25)	(0.10, 0.20)	(0.20, 0.30)	0.153	0.188	0.200	23
(15, 25)	(0.15, 0.50)	(0.30, 0.30)	0.196	0.220	0.200	12
(15, 25)	(0.20, 0.10)	(0.80, 0.10)	0.981	0.990	0.200	1
(15, 25)	(0.25, 0.40)	(0.50, 0.10)	0.535	0.607	0.200	13

<sup>a</sup>Adjusted  $\alpha$ -level, multinomial exact test.

**Table 5.** Average power comparisons with Fisher’s test for  $2 \times 4$  tables.

$(n_1, n_2)$	$(p_1^{(1)}, p_2^{(1)}, p_3^{(1)})$	$(p_1^{(2)}, p_2^{(2)}, p_3^{(2)})$	Power		$\alpha^a$	Power gain (%)
			Fisher	Multinomial		
(3, 3)	(0.10, 0.10, 0.10)	(0.20, 0.20, 0.10)	0	0.096	0.350	–
(3, 3)	(0.15, 0.50, 0.20)	(0.30, 0.30, 0.15)	0	0.067	0.350	–
(3, 3)	(0.25, 0.40, 0.20)	(0.50, 0.10, 0.20)	0	0.085	0.350	–
(3, 3)	(0.20, 0.10, 0.50)	(0.80, 0.10, 0.05)	0	0.377	0.350	–
(3, 15)	(0.10, 0.10, 0.10)	(0.20, 0.20, 0.10)	0.014	0.064	0.375	356
(3, 15)	(0.15, 0.50, 0.20)	(0.30, 0.30, 0.15)	0.050	0.062	0.375	24
(3, 15)	(0.25, 0.40, 0.20)	(0.50, 0.10, 0.20)	0.171	0.227	0.375	33
(3, 15)	(0.20, 0.10, 0.50)	(0.80, 0.10, 0.05)	0.564	0.704	0.375	25
(7, 10)	(0.10, 0.10, 0.10)	(0.20, 0.20, 0.10)	0.057	0.109	0.316	92
(7, 10)	(0.15, 0.50, 0.20)	(0.30, 0.30, 0.15)	0.099	0.132	0.316	33
(7, 10)	(0.25, 0.40, 0.20)	(0.50, 0.10, 0.20)	0.227	0.277	0.316	22
(7, 10)	(0.20, 0.10, 0.50)	(0.80, 0.10, 0.05)	0.738	0.822	0.316	11
(10, 10)	(0.10, 0.10, 0.10)	(0.20, 0.20, 0.10)	0.100	0.129	0.280	29
(10, 10)	(0.15, 0.50, 0.20)	(0.30, 0.30, 0.15)	0.124	0.138	0.280	12
(10, 10)	(0.25, 0.40, 0.20)	(0.50, 0.10, 0.20)	0.244	0.300	0.280	23
(10, 10)	(0.20, 0.10, 0.50)	(0.80, 0.10, 0.05)	0.780	0.873	0.280	12

<sup>a</sup>Adjusted  $\alpha$ -level, multinomial exact test.

baseline demographic characteristics of the participants by comparing the ‘exercise’ and the control group using exact two-sample multinomial test without correction for the conservativeness of the test under an unadjusted significance level of 0.05. The  $p$ -value of the (uncorrected) exact two-sample multinomial test indicated that there was no difference in demographics between the two groups of patients at the 0.05  $\alpha$ -level. We set out to assess whether the conclusion of no difference in demographics for both groups of patients holds

**Table 6.** Baseline demographic characteristics of study participants – effect of exercise on biobehavioral outcomes of fatigue during cancer treatment study.

Variable	Exercise group <sup>a</sup>	Control group <sup>a</sup>	<i>p</i> <sup>b</sup>	Regression adjusted $\alpha$ -level
Age	47.9 ± 10.4	52.7 ± 10.7	–	–
Cancer stage	2.0 ± 0.5	1.6 ± 0.6	–	–
Ethnicity			1.00	0.17
Hispanic	2 (28.6)	2 (28.6)		
Non-hispanic	5 (71.4)	5 (71.4)		
Marital status			0.59	0.17
Married	5 (71.4)	3 (42.9)		
Unmarried	2 (28.6)	4 (57.1)		
Education			0.37	0.38
< High school	1 (14.3)	0 (0.0)		
High school	1 (14.3)	3 (42.9)		
Technical school	0 (0.0)	3 (42.9)		
College	2 (28.6)	0 (0.0)		
Postcollege	3 (42.9)	1 (14.3)		
Employment			0.91	0.17
Employed	4 (57.1)	5 (71.4)		
Unemployed	3 (42.9)	2 (28.6)		

<sup>a</sup>*n*(%) or *M* ± *SD*.<sup>b</sup>Unconditional exact multinomial *p*-value.

after correcting the multinomial test using the method we have described herein as well as in comparison to Boschloo's and Barnard's test.

In Table 6, we replicate the data from the study by Al-Majid et al. [3] including our model-based correction for the significance level. Based on our correction of the  $\alpha$ -level, there is a statistically significant difference between the exercise and control group by Education which would otherwise be unknown given the conservativeness of the unconditional multinomial exact *p*-value used to compare the treatment arms in the study.

## 6. Discussion

Parameter estimation adjustment in the degrees of freedom of the asymptotic  $\chi^2$  statistic when testing a general null hypothesis of equality of several multinomial distributions contrasted with the completely specified null that multinomial distributions are the same with given category probabilities is well known. In an effort to extend the theory to small sample analysis, we carried out a methodological study that explores the effect of parameter estimation on the classical exact *p*-value test where the absence of parametric distribution makes proper adjustment intricate. We implemented a large-scale simulation design in our study to empirically estimate the 5th percentiles of the exact *p*-values of the UEPT under extensive set of data generating parameters. In particular, we considered 126 parameter combinations including sample sizes, number of categories, and multinomial category probabilities and simulated  $10^4$  datasets for each scenario –  $10^4$  simulated datasets is the most optimal choice computationally and there is only a marginal effect on results if the number of simulated datasets is increased. We addressed the most common data analysis settings that are computationally feasible. The sample sizes encompass a grid of values that span small sample territory. The *p* was chosen based on a grid of multinomial categories that covers a wide range of values. We investigated both the dependence of the results on the values of the parameters that define each simulation data scenario and on the number of datasets simulated for each scenario. Further, we aggregated the data over all unobservable parameters (the multinomial category probabilities) and linearly regressed

the 5th percentiles of the distributions of the classical unconditional multinomial exact  $p$ -value against the sample sizes, the number of multinomial categories and their interactions. In all scenarios, parameter estimation of the unknown multinomial category probabilities, using data from both samples, leads to inflation of the 5th percentiles of the distribution of the  $p$ -values of UEPT compared to the nominal level. Consequently, estimating the empirical 5th percentile of the distribution of  $p$ -values of the UEPT can be used as adjustment required to correct the type I error rate to nominal levels while increasing the power of the test. Our subsequent regression analysis reveals the proper adjustment of the 5th percentile as a function of the sample sizes, number of multinomial categories and their interactions. This regression model achieved an  $R^2$  value of 0.95 and this correction establishes the correct hypothesis testing under the assumption of a multinomial distribution. Results of this study indicate that our adjustment results in a test more power than Fisher's, Barnard's, and Boschloo's tests – the most common alternatives to the unconditional multinomial exact  $p$ -value test.

Boschloo's test, Boschloo [6], is conceptually similar to our idea. It provides an adjustment for the Fisher's exact test by a scenario-specific significance level inflation for all  $2 \times 2$  contingency tables. The UEPT is conservative in the presence of parameter estimation and our adjustment via a scenario-specific significance level inflation for all  $c \times r$  contingency tables based on a regression model applied to a dense set of simulated data leads to improved power better than Barnard's and Boschloo's tests. Barnard's test, Barnard [4,5], similar to our approach, is fully unconditional as it considers all tables compatible with the observed sample sizes and the two categories since it applies to only for binomial data. However, this test does not estimate the common value of  $p$  under the null via maximum likelihood, but it finds the value of  $p$  that maximizes the exact  $p$ -value (defined as the sum of the probabilities of contingency tables as likely or less likely to occur than the observed table). Thus, Barnard's test treats the exact  $p$ -value expression as a 'new' likelihood for  $p$  that is maximized. Barnard treats the value of the maximum it attained as the exact  $p$ -value. Taking the maximum of the exact  $p$ -value expression as a function of  $p$  is an alternative way to adjust for the conservativeness of the UEPT. In contrast, we estimate  $p$  using maximum likelihood on the product of multinomial likelihoods (which does not maximize the exact  $p$ -value expression) and then adjust using scenario-specific significance level inflation for all  $2 \times c$  contingency table based on the aforementioned regression model.

In this work we proposed and implemented a novel approach for adjustment of the 5th percentiles of the exact  $p$ -value test when applied to comparisons of several multinomial distributions in small samples. We used best subset regression model applied to a dense set of simulated data to develop a scenario-specific adjustment to the significance level resulting in nominal type I error rates and increased power that equals several hundred percent over all other competing testing approach. This approach can be extended to more than two multinomial samples as well as including corrections for tests of independence.

## Disclosure statement

No potential conflict of interest was reported by the authors.

## References

- [1] A. Agresti, *Exact inference for categorical data: Recent advances and continuing controversies*, Stat. Med. 20 (2001), pp. 2709–2722.

- [2] A. Agresti, *Categorical Data Analysis*, Wiley, New York, 2002.
- [3] S. Al-Majid, L.D. Wilson, C. Rakovski, and J.W. Coburn, *Effects of exercise on biobehavioral outcomes of fatigue during cancer treatment results of a feasibility study*, *Biol. Res. Nurs.* 17 (2015), pp. 40–48.
- [4] G. Barnard, *A new test for  $2 \times 2$  tables*, *Nature* 156 (1945), pp. 177.
- [5] G. Barnard, *Significance tests for  $2 \times 2$  tables*, *Biometrika* 34 (1947), pp. 123–138.
- [6] R. Boschloo, *Raised conditional level of significance for the  $2 \times 2$ -table when testing the equality of two probabilities*, *Stat. Neerl.* 24 (1970), pp. 1–9.
- [7] B. Efron and R.J. Tibshirani, *Permutation tests*, in *An Introduction to the Bootstrap*, Springer, New York, 1993, pp. 202–219.
- [8] R.A. Fisher, *On the interpretation of  $\chi^2$  from contingency tables, and the calculation of P*, *J. R. Stat. Soc.* 85 (1922), pp. 87–94.
- [9] R.A. Fisher, *Confidence limits for a cross-product ratio*, *Aust. J. Stat.* 4 (1962), pp. 41–41.
- [10] C.-Y. Lin and M.-C. Yang, *Improved p-value tests for comparing two independent binomial proportions*, *Comm. Stat. Simulation Comput.* 38 (2008), pp. 78–91.
- [11] A.S. Mato and A.M. Andrés, *Simplifying the calculation of the p-value for Barnard's test and its derivatives*, *Stat. Comput.* 7 (1997), pp. 137–143.
- [12] D.V. Mehrotra, I.S. Chan, and R.L. Berger, *A cautionary note on exact unconditional inference for a difference between two independent binomial proportions*, *Biometrics* 59 (2003), pp. 441–450.
- [13] C.R. Mehta and J.F. Hilton, *Exact power of conditional and unconditional tests: going beyond the  $2 \times 2$  contingency table*, *Am. Stat.* 47 (1993), pp. 91–98.
- [14] C.R. Mehta and N.R. Patel, *Algorithm 643: Fexact: A fortran subroutine for Fisher's exact test on unordered  $r \times c$  contingency tables*, *ACM Trans. Math. Softw.* 12 (1986), pp. 154–161.
- [15] C.R. Mehta, N.R. Patel, and A.A. Tsiatis, *Exact significance testing to establish treatment equivalence with ordered categorical data*, *Biometrics* 40 (1984), pp. 819–825.
- [16] N.L. Oliveira, C.A.d.B. Pereira, M.A. Diniz, and A Polpo, *A discussion on significance indices for contingency tables under small sample sizes*, *PLoS One* 13 (2018), p. e0199102.
- [17] J. Röhmel and U. Mansmann, *Unconditional non-asymptotic one-sided tests for independent binomial proportions when the interest lies in showing non-inferiority and/or superiority*, *Biom. J.* 41 (1999), pp. 149–170.
- [18] R. Routledge, *Resolving the conflict over Fisher's exact test*, *Can. J. Stat.* 20 (1992), pp. 201–209.
- [19] F. Yates, *Contingency tables involving small numbers and the  $\chi^2$  test*, *Suppl. J. R. Stat. Soc.* 1 (1934), pp. 217–235.
- [20] F. Yates, *Tests of significance for  $2 \times 2$  contingency tables*, *J. R. Stat. Soc. Ser. A* 147 (1984), pp. 426–449.