Interdiscip Sci Comput Life Sci (2016) 8:337–345
DOI 10.1007/s12539-015-0128-5

ORIGINAL RESEARCH ARTICLE

Comparing Performances of Multiple Comparison Methods in Commonly Used 2 × *C* Contingency Tables

Sengul Cangur[1] · **Handan Ankarali**[1] · **Ozge Pasin**[1]

Received: 27 January 2015 / Revised: 7 July 2015 / Accepted: 12 August 2015 / Published online: 10 October 2015

Abstract This study aims at mentioning briefly multiple comparison methods such as Bonferroni, Holm–Bonferroni, Hochberg, Hommel, Marascuilo, Tukey, Benjamini–Hochberg and Gavrilov–Benjamini–Sarkar for contingency tables, through the data obtained from a medical research and examining their performances by simulation study which was constructed as the total 36 scenarios to 2 × 4 contingency table. As results of simulation, it was observed that when the sample size is more than 100, the methods which can preserve the nominal alpha level are Gavrilov–Benjamini–Sarkar, Holm–Bonferroni and Bonferroni. Marascuilo method was found to be a more conservative than Bonferroni. It was found that Type I error rate for Hommel method is around 2 % in all scenarios. Moreover, when the proportions of the three populations are equal and the proportion value of the fourth population is far at a level of ±3 standard deviation from the other populations, the power value for Unadjusted All-Pairwise Comparison approach is at least a bit higher than the ones obtained by Gavrilov–Benjamini–Sarkar, Holm–Bonferroni and Bonferroni. Consequently, Gavrilov–Benjamini–Sarkar and Holm–Bonferroni methods have the best performance according to simulation. Hommel and Marascuilo methods are not recommended to be used because they have medium or lower performance. In addition, we have written a Minitab macro about multiple comparisons for use in scientific research.

Keywords Contingency table · Multiple comparison · Gavrilov–Benjamini–Sarkar · Holm–Bonferroni · Marascuilo · Hommel

1 Introduction

When the null hypothesis which is established in appropriate Chi-square test results is rejected in $R \times C$ contingency table, that is, if calculated P value smaller than initially determined Type I error rate (say 5 %), there is a significant relationship between the row and column variables. In this case, there is a need to determine rows or columns that cause significant relationship [1, 2]. This requirement is satisfied by multiple comparisons or post hoc tests in statistics literature. The post hoc tests can be categorized in terms of kept under control only Family-Wise Error Rate (FWER) and False Discovery Rate (FDR) which are different error rates. In multiple hypothesis tests, a number of researchers argue to control FWER, while another part FDR. Simulation studies in the literature are intended to determine the best post hoc test. But because the performances of these methods show differences in any case, still it is seen that studies are continued in this field [3–5].

In $R \times C$ tables to be significant relationship has same meaning and same way interpreted with to be significant of interaction in factorial analysis of variance models [6]. When the interaction is significant, the means of the levels

Electronic supplementary material The online version of this article (doi:10.1007/s12539-015-0128-5) contains supplementary material, which is available to authorized users.

✉ Sengul Cangur
sengulcangur@duzce.edu.tr

Handan Ankarali
handanankarali@gmail.com

Ozge Pasin
ozgepasin90@yahoo.com.tr

[1] Department of Biostatistics and Medical Informatics, Faculty of Medicine, Duzce University, Duzce, Turkey

of other factors are compared separately by post hoc test in levels of each factor, while the relationship is significant according to Chi-square test result, the comparions of columns (or rows) are made separately in each rows (or columns). But, the statistic compared here is proportion, not average. Based on these similarities, it is seen that there are various post hoc tests used in literature for contingency tables, and it has seen that there are limited simulations or numerical studies, which are intended to application of multiple comparison methods and to compare their performance [see: 2–12]. It is noticeable that only three of these studies are the simulation study [5, 11, 12].

Usually, researchers prefer traditional approach to determining differences between proportions when significant relationship is found as a result of Chi-square test. One of these approaches is to compare in pairs to determine rows or columns that are different in the contingency table [13]. However, in this case because comparisons are considered as independent, making the Type I error rate is increasing. Other classic approach is to apply Standardized and Adjusted Residuals Statistics (STAR) method [14]. In this method, the interpretation of the normal probability plots is getting quite harder with the increasing number of cells to be tested. Against these classical methods, post hoc tests described in this study are approaches that protect the predetermined Type I error rate and are easily applicable. Of the methods mentioned in this study, Benjamini–Hochberg and Gavrilov–Benjamini–Sarkar (GBS) methods control primary FDR, while Bonferroni, Holm–Bonferroni, Hochberg, Hommel, Marascuilo and Tukey methods control only FWER.

The aim of this study is to mention briefly the all-pairwise multiple comparison tests, which are recommended for use to determine significant differences between proportions in contingency table, and to examine the performance of these in terms of Type I error and power by a simulation study. Our simulation study that is constructed in the form of different scenarios unlike other studies will be compared performances of the methods which control both FWER and primarily FDR. Moreover, it has been aimed to write a macro of methods which have good performance in our simulation results. This macro will be used effectively among applied researches.

2 Methodology

In literature, there are many multiple comparison procedures that are classified differently. According to the objective of the researcher, multiple comparison procedures can be classified such as the all-pairwise multiple comparisons (MCA), multiple comparison with the best (MCB), multiple comparisons with a control (MCC) and multiple comparisons with the mean (MCM). If two or more groups are compared, MCA compares the pairwise differences of group parameters (means or proportions, etc.). MCB compares each group with the best of the other groups. Also MCC compares the differences between the mean (or proportion, etc.) of each groups and those of the control. MCM compares the differences between the mean of each groups and the overall mean of groups [15]. This study includes only some MCA procedures.

Apart from this general classification, multiple comparison procedures can be categorized according to their algorithm structure (single-step or stepwise), the distribution type of test statistic (marginal or joint), the feature of adaptivity (adaptive or non-adaptive) and the type of error which they primarily control (FWER, FDR). No matter what the type of classification is, the main concern of multiple testing is multiplicity problem. In other words, the problem is that probability of making at least one Type I error at the predetermined level increases quickly with the increase in the number of hypotheses that are tested. There are two main approaches in order to solve this problem. According to the first approach, FWER, which can be described as the probability of rejecting at least one hypothesis mistakenly in a certain set of hypothesis, is controlled when V is the number of false rejection [15]: $\text{FWER} = P(V > 0)$.

According to the second approach, FDR, which can be described as the expected rate of Type I errors among the rejected hypothesis, is controlled [16]. FDR is the expected value of the proportion of the number of falsely rejected hypothesis to the number of rejected hypothesis (R): $\text{FDR} = E\left(\frac{V}{R} R > 0\right) P(R > 0)$. If all the hypothesis tests are true null hypothesis, $k = m_0$. If all null hypothesis are true, then FDR = FWER. When $m_0 < k$, FDR is equal to or below FWER. When $R = 0$, $\text{FDR} = V/R = 0/0$ and it is a special description in the calculation of FDR.

2.1 Multiple Comparison Methods in $R \times C$ Tables

This study includes only the MCA approaches of the most commonly used in multiple comparison procedures literature. The methods of Holm–Bonferroni, Hochberg and Hommel are the modified versions of Bonferroni test. Holm–Bonferroni method is more powerful than simple Bonferroni correction, and it is easy to use. It is one of the marginal multiple comparison methods which use stepwise algorithm for making simultaneous inference [17]. Hochberg method is a closed form procedure developed from Simes' (1986) test. It is also known as the step-up version of Holm–Bonferroni method [18]. The Hommel approach is known as a closed test method, and it was developed from Simes' (1986) test [19]. Marascuilo method tests the

differences between the proportions simultaneously [8]. Tukey method was first proved by Ryan [7] that it can be used for proportion comparisons. Benjamini–Hochberg method which use step-up algorithm is also known as marginal multiple comparison procedure [16]. GBS method is a simpler procedure developed by means of revising the procedure of multiple-stage adaptive step-down proposed by Benjamini–Krieger–Yekutieli (2006) and based on the approach of Benjamini–Hochberg [3, 16]. The steps of algorithms for these methods were given broadly in Supplementary A.

3 Application

3.1 Numerical Data

In order to make the implementation steps of the multiple comparison methods and compare their results, we used the data obtained from the diagnostic scale of attention-deficit/hyperactivity disorders (ADH-D) which was applied to 261 students attending the 9th, 10th and 11th grades in science high school Table 1).

With Fisher–Freeman–Halton test, it was found out that there is a significant difference between the grades statistically in terms of ADH-D. According to this result, in order to determine between which grades the proportions belonging to each category of ADH-D case are different, Bonferroni, Holm–Bonferroni, Hochberg, Hommel, Tukey, Marascuilo, Benjamini–Hochberg and GBS approaches were performed, respectively.

3.2 Simulation Studies

3.2.1 Observed Type I Error Rate (When the Null Hypothesis Is True)

Because the number of groups compared is usually between 3 and 6 in medical studies, in our simulation study the number of groups compared was taken as four (2×4) in order to be compatible with numerical data set used in this study. We applied the proposed procedures in this study to the simulated 2×4 contingency tables, and we compared the empirical Type I error and the empirical power of six different multiple comparison procedures and the individual test under different scenarios. These procedures are Unadjusted All-Pairwise Comparisons (UAPC), Adjusted All-Pairwise Multiple Comparisons (AAPMC) such as Bonferroni, Holm–Bonferroni, Hommel, Marascuilo and GBS. Among these methods, Bonferroni, Holm–Bonferroni and Marascuilo control FWER. On the other hand, Hommel method is one that also controls FWER but is implemented on a step-by-step basis. And the method of GBS is one of the newest methods which controls primarily FDR and is implemented on a step-by-step basis. We have considered controlling actual Type I error at 0.05 in all procedures.

The proportions of the groups which are compared within the scenarios determined for Type I error were generated from multinomial distribution in a way to be small (0.10), medium (0.50) and big (0.80). Moreover, the sample sizes were determined to be 50, 100 and 500, respectively, in order to examine the effect of small, medium and large sample size on the results. Because each proportion in a given sample was drawn from the same underlying distribution, no significant differences were expected and any that were found were attributed to Type I error. Apart from that, the conditions when the number of observations is equal in each group are considered in the study because Type I error and power which are calculated under these conditions give the best result. When the balance in terms of the numbers of observations in each group is destroyed, these two tests start to give more poor results.

3.2.2 Power of Tests (When the Null Hypothesis Is False)

We planned that the proportion belonging to one of the four groups which are thought to be compared to calculate the power of the tests will be taken from a different population. The proportions of the population which is different are

Table 1 Distribution of attention-deficit/hyperactivity disorder according to the grades

Attention-deficit/hyperactivity disorder (ADH-D)	Grades						Total	
	9th		10th		11th			
	n	%	*n*	%	*n*	%	*n*	%
No evidence of ADH-D	1	1.02	5	5.62	1	1.35	7	2.68
Low-risk	5	5.10	5	5.62	4	5.41	14	5.36
Moderate-risk	79	80.61	64	71.91	40	54.05	183	70.11
High-risk	13	13.27	15	16.85	29	39.19	57	21.84
Total	98	100	89	100	74	100	261	100

calculated in a way to be far away from the proportion of population from which the other three groups have been taken at a rate of ±1, ±2 and ±3 standard deviation (SD), respectively. Moreover, the sample size was determined to be 50, 100 and 500, respectively.

As a result in this simulation study, 9 hypotheses (scenarios) and 27 hypotheses (scenarios) were considered for Type I error and power of tests, respectively. Each simulation was done with 5000 repetitions. Summary of simulation in this study is shown in Supplementary Table 1 and Table 2. We used a macro that we wrote in Minitab programme (ver. 16.) for simulation study. An example part of the codes used to simulation is shown in Supplementary B–C.

4 Results

4.1 Results of Numerical Data

When the data were evaluated, it was found that there is a significant relationship between ADH-D and the grades; in other words, the differences between the grades change in accordance with ADH-D (Fisher–Freeman–Halton test statistic $= 21.041$, $P < 0.001$). According to this result, the groups were compared for each ADH-D finding separately and the question of which grades cause the difference to come out was tried to be answered by means of multiple comparison methods. The related results are given in Table 2.

Among all the methods that were examined, it was found that there is no significant difference statistically between the grades when there is no finding of ADH-D and the students with a low level of risk are in question. When the ones which carry a risk of ADH-D at a medium level are evaluated, according to the results of the methods except for Marascuilo method, the frequency of ADH-D cases at the 11th grade (54.1 %) is significantly less than both the 10th (71.9 %) and 9th (80.6 %) grades. When the category which carries a high level of ADH-D risk is considered, the proportion of the 11th-grade students (39.2 %) was found out to be significantly higher than 9th- and 10th-grade students (13.3 %, 16.9 %). Besides, with regard to the significance of difference between the proportions compared under the same conditions, it was found that the methods of Tukey, Bonferroni and Hommel are similar and they determine the significance in a more difficult way. Under the same conditions, Holm–Bonferroni and Hochberg methods are rejected the null hypothesis more easily and their results are found similar. Among these methods, the ones that find the significant differences most easily are the methods of GBS and Benjamini–Hochberg. Moreover, in the majority of the conditions within medical example, Type I error rates calculated by GBS and Benjamini–Hochberg methods were found to be smaller than the value calculated according to the other methods. In addition to this, if the Type I error rates calculated by the methods of Bonferroni, Holm–Bonferroni, Hochberg, Hommel, Tukey, Benjamini–Hochberg and GBS are close to the critical alpha value and are found to be significant, a significant result cannot be obtained from the method of Marascuilo (Table 2).

4.2 Results of Simulation Studies

The results of simulation study gathered in relation with Type I error in accordance with the scenarios in Supplementary Table 1 are given in Table 3. The graph related to the simulation results is shown in Supplementary Fig. 1. In all the scenarios, Type I error rates gathered through UAPC approach were found to be around 20 % (20.98–22.84 %). In addition to this, when the sample size is 100 under all conditions and the proportion value of each group is 0.50, Type I error rates belonging to the tests except for GBS are below 5 %. Moreover, when the proportion is around 0.50 in all groups, Type I error rate has been calculated to be closest to the expected value. In other words, when the group proportion values get closer to zero or one, Type I error rate belonging to the tests is found to be lower than expected. When the sample size is 100 and over, it was observed that the methods which can preserve the level of predetermined Type I error are GBS, Holm–Bonferroni and Bonferroni. On the other hand, Marascuilo method was found to be a more conservative test than Bonferroni method. In other words, it rejects fewer hypotheses than expected. In all the scenarios we have examined, it was found that Type I error rate obtained by Hommel method is around 2 % (1.14–2.30 %) and that the level of predetermined error cannot be preserved. Hence, it can be said that the three tests we have examined except for GBS and Holm–Bonferroni methods are more conservative tests. Moreover, Type I error rates were found to be less than the expected value in small sample sizes.

The power values of multiple comparison procedures obtained by different scenarios are given in Table 4 and Supplementary Fig. 2. In all sample sizes, no matter what the proportion of three populations is, when the proportion of the fourth population is further away from these populations at a level of ±2 SD and more, it was found out that the power values obtained by all the methods increase. It was observed that when the proportion of the three populations is 0.10 and the proportion of the fourth population is further at a level of ±3 SD, and when the sample size is more than 100, the power value increases. When the proportion of the three populations is 0.80 and the proportion of the fourth population is further away at a level of ±2 SD

Table 2 Multiple comparisons between grades in terms of attention-deficit/hyperactivity disorder

Attention-deficit/hyperactivity disorder (ADH-D)	Pairwise comparisons of grades								
	9th versus 10th			9th versus 11th			10th versus 11th		
	TS	P	$\tilde{P}$	TS	P	$\tilde{P}$	TS	P	$\tilde{P}$
Bonferroni									
No evidence of ADH	1.782	0.075[a]	0.225	0.211	0.833[a]	1.000	1.440	0.150[a]	0.450
Low-risk	0.157	0.876[a]	1.000	0.093	0.926[a]	1.000	0.060	0.952[a]	1.000
Moderate-risk	1.401	0.161[a]	0.483	3.928	**<0.001**[a]	**<0.001**	2.406	**0.016**[a]	**0.048**
High-risk	0.687	0.492[a]	1.000	4.121	**<0.001**[a]	**<0.001**	3.250	**0.001**[a]	**0.003**
Holm–Bonferroni									
No evidence of ADH	1.782	0.075[a]	0.224	0.211	0.833[c]	0.833	1.440	0.150[b]	0.300
Low-risk	0.157	0.876[a]	1.000	0.093	0.926[b]	1.000	0.060	0.952[c]	1.000
Moderate-risk	1.401	0.161[c]	0.161	3.928	**<0.001**[a]	**<0.001**	2.406	**0.016**[b]	**0.032**
High-risk	0.687	0.492[c]	0.492	4.121	**<0.001**[a]	**<0.001**	3.250	**0.001**[b]	**0.002**
Hochberg									
No evidence of ADH	1.782	0.075[a]	0.224	0.211	0.833[c]	0.833	1.440	0.150[b]	0.300
Low-risk	0.157	0.876[a]	0.952	0.093	0.926[b]	0.952	0.060	0.952[c]	0.952
Moderate-risk	1.401	0.161[c]	0.161	3.928	**<0.001**[a]	**<0.001**	2.406	**0.016**[b]	**0.032**
High-risk	0.687	0.492[c]	0.492	4.121	**<0.001**[a]	**<0.001**	3.250	**0.001**[b]	**0.002**
Hommel									
No evidence of ADH	1.782	0.075[a]	0.411	0.211	0.833[c]	1.000	1.440	0.150[b]	0.412
Low-risk	0.157	0.876[a]	1.000	0.093	0.926[b]	1.000	0.060	0.952[c]	1.000
Moderate-risk	1.401	0.161[c]	0.296	3.928	**<0.001**[a]	**<0.001**	2.406	**0.016**[b]	**0.044**
High-risk	0.687	0.492[c]	0.902	4.121	**<0.001**[a]	**<0.001**	3.250	**0.001**[b]	**0.003**
Benjamini–Hochberg									
No evidence of ADH	1.782	0.075[a]	0.224	0.211	0.833[c]	0.833	1.440	0.150[d]	0.225
Low-risk	0.157	0.876[a]	0.952	0.093	0.926[d]	0.952	0.060	0.952[c]	0.952
Moderate-risk	1.401	0.161[c]	0.161	3.928	**<0.001**[a]	**<0.001**	2.406	**0.016**[d]	**0.024**
High-risk	0.687	0.492[c]	0.492	4.121	**<0.001**[a]	**<0.001**	3.250	**0.001**[d]	**0.002**
Gavrilov–Benjamini–Sarkar									
No evidence of ADH	1.782	0.075[e]	0.243	0.211	0.833[g]	1.000	1.440	0.150[f]	0.176
Low-risk	0.157	0.876[e]	1.000	0.093	0.926[f]	1.000	0.060	0.952[g]	1.000
Moderate-risk	1.401	0.161[g]	0.064	3.928	**<0.001**[e]	**<0.001**	2.406	**0.016**[f]	**0.016**
High-risk	0.687	0.492[g]	0.323	4.121	**<0.001**[e]	**<0.001**	3.250	**0.001**[f]	**0.001**
Tukey		**CV**			**CV**			**CV**	
No evidence of ADH	2.477	3.31	0.187	0.336	3.31	0.969	1.978	3.31	0.342
Low-risk	0.229	3.31	0.986	0.173	3.31	0.992	0.044	3.31	1.000
Moderate-risk	1.965	3.31	0.347	**5.260**[h]	3.31	**<0.001**	**3.320**[h]	3.31	**0.0497**
High-risk	0.963	3.31	0.775	**5.521**[h]	3.31	**<0.001**	**4.509**[h]	3.31	**0.004**
Marascuilo									
No evidence of ADH	0.046	0.065	0.220	0.003	0.041	0.981	0.043	0.068	0.309
Low-risk	0.005	0.081	0.988	0.003	0.084	0.996	0.002	0.088	0.998
Moderate-risk	0.087	0.152	0.375	**0.266**[h]	0.172	**<0.001**	0.179	0.184	0.059
High-risk	0.036	0.128	0.791	**0.259**[h]	0.162	**<0.001**	**0.223**[h]	0.169	**0.006**

Bold values indicate $\alpha = 0.05$

TS test statistic, *CV* critical value, P unadjusted P value, $\tilde{P}$ adjusted P value, $\alpha = 0.05$

[a] $\alpha' = 0.017$, [b] $\alpha' = 0.025$, [c] $\alpha' = 0.05$, [d] $\alpha' = 0.033$, [e] $\alpha' = 0.016$, [f] $\alpha' = 0.048$, [g] $\alpha' = 0.130$, [h] If TS > CV, it is significant

Table 3 Observed overall Type I error probabilities of multiple comparison tests used in simulation study

Sample size	Proportions in each group (P_i for $i = 1, 2, 3, 4$)	Observed overall Type I error (%)					
		Unadjusted all-pairwise comparisons	Bonferroni	Holm–Bonferroni	Hommel	Marascuilo	Gavrilov–Benjamini–Sarkar
50	0.10	21.24	3.08	3.34	1.14	2.26	3.50
	0.50	22.84	3.72	3.85	2.14	2.80	4.11
	0.80	21.92	3.64	3.88	1.68	2.60	3.90
100	0.10	21.52	4.26	4.30	1.86	2.52	4.32
	0.50	22.34	4.76	4.86	2.20	2.86	5.04
	0.80	20.98	4.08	4.14	1.88	2.92	4.16
500	0.10	21.20	4.30	4.40	1.88	2.96	4.45
	0.50	21.72	4.80	4.88	2.30	2.90	4.92
	0.80	21.00	4.26	4.28	1.92	2.86	4.54

and more, the power value increases in all the methods that have been examined. When the proportion value of the fourth population is further away at a level of ±1 SD no matter what the proportion of the three populations is in all sample sizes, the power value obtained by UAPC approach is about 34 % (29.50–39.52 %). However, the power value has been found to be around 9 % (5.40–14.28 %) in adjusted procedures (except for Hommel). Moreover, it was found that when the proportion of the fourth population is far away at a level of ±3 SD, the power obtained by the approach of UAPC is at least slightly higher than the one obtained from adjusted procedures (GBS, Holm–Bonferroni and Bonferroni). Only under this circumstance, the methods which are the most powerful can be listed as UAPC, GBS, Holm–Bonferroni and Bonferroni.

Consequently, it can be said that it would be true to choose the adjusted procedures such as GBS and Holm–Bonferroni in terms of preserving the predetermined Type I error rate and its power.

5 Discussion

In this study, UAPC and AAPMC procedures which can be used to compare more than two proportions and which consider different types of error have been implemented on numerical data set, and their performances have been compared in terms of Type I error and power.

When literature is examined, there are limited number of studies about implementing multiple comparison procedures on $R \times C$ contingency tables, and especially comparing them by simulation studies. The first studies about this topic were carried out by Ryan [7], Holland and Copenhaver [9] and Westfall and Young [10]. The fact that the studies with simulations are few in number is a great disadvantage in terms of evaluating the performances of the methods. Apart from this, in literature, it has been impossible to come across a study about implementing GBS, which is one of the newest multiple comparison methods to control primarily FDR, on contingency tables and comparing it with other methods. Moreover, any simulation study cannot be found about Marascuilo method, which is suggested in comparing more than two proportions.

At the end of this study, it was found that Type I error rates obtained by GBS and Holm–Bonferroni methods used for comparing more than two proportions are closer to 5 % than those of other methods. Moreover, Marascuilo method Hommel method can be said to be conservative tests. Although there are studies about practicability of Marascuilo procedure in comparing the proportion in literature [see: 6], a simulation study which includes this method cannot be found. Horne and Plaehn [12] achieved that the methods of Hommel and Bonferroni give similar results in simulation study. In our study, it was obtained that Hommel method is a more conservative test than Bonferroni method. When Oden et al. [5] examined the performances of Hochberg, Bonferroni, Hommel, Shaffer (S1, S2) and Pairwise Closed procedures in terms of Type I error (proportional, minimal) in the analysis of 2×2, 2×3 tables, they stated that all the methods could control FWER, but when compared to other methods, Type I error rate for Bonferroni method is further from the value of 5 %. They found that the method which is closest to the expected value in terms of proportional FWER is Pairwise Closed. When we compare our findings with the results of the study carried out by Oden et al. [5], it was observed that the examined common methods have given different results in terms of Type I error. Kim et al. [11] found that Bonferroni, Benjamini–Hochberg, Storey and Efron methods are more powerful when the number of significant

Table 4 Observed power values of multiple comparison methods for 27 simulation scenarios

Multiple comparison methods	Proportion values of three population ($P_{1,\ 2,\ 3}$)	P_4 away from $P_{1,\ 2,\ 3}$			Sample size
		1 SD	2 SD	3 SD	
		Power (%)			
Unadjusted all-pairwise comparisons	0.10	33.48	51.28	71.00	50
		29.50	50.04	70.22	100
		39.52	58.08	88.14	500
	0.50	35.16	62.44	86.52	50
		33.36	56.80	82.82	100
		38.70	62.72	85.58	500
	0.80	32.96	71.46	97.08	50
		31.30	61.52	96.74	100
		32.46	64.26	90.40	500
Bonferroni	0.10	8.80	19.30	36.46	50
		7.00	19.64	38.46	100
		14.28	26.98	65.76	500
	0.50	8.54	26.30	59.48	50
		9.00	23.96	55.00	100
		12.98	30.66	58.06	500
	0.80	8.10	35.16	81.12	50
		8.22	29.38	69.06	100
		9.88	33.66	70.18	500
Holm–Bonferroni	0.10	8.52	19.78	36.48	50
		7.30	18.50	39.90	100
		13.62	27.04	65.20	500
	0.50	9.08	27.54	59.44	50
		8.60	24.60	54.26	100
		13.04	31.86	58.24	500
	0.80	8.90	34.88	82.44	50
		9.52	29.00	68.20	100
		9.50	33.36	71.54	500
Hommel	0.10	3.60	10.36	25.80	50
		3.72	10.98	27.94	100
		7.50	18.16	53.52	500
	0.50	5.58	18.38	46.92	50
		4.52	17.00	43.02	100
		7.36	21.44	46.28	500
	0.80	4.22	23.44	66.78	50
		4.36	19.78	56.40	100
		4.84	22.38	60.06	500
Marascuilo	0.10	5.80	15.86	32.52	50
		5.40	15.52	34.38	100
		11.20	21.40	59.40	500
	0.50	9.02	25.28	55.86	50
		6.12	19.88	47.80	100
		9.64	24.40	52.76	500
	0.80	8.44	33.02	75.00	50
		6.08	23.84	63.62	100
		7.24	25.82	66.32	500

Table 4 continued

Multiple comparison methods	Proportion values of three population ($P_{1,\ 2,\ 3}$)	P_4 away from $P_{1,\ 2,\ 3}$			Sample size
		1 SD Power (%)	2 SD	3 SD	
Gavrilov–Benjamini–Sarkar	0.10	8.62	19.82	35.80	50
		7.04	18.58	38.18	100
		12.92	27.36	65.60	500
	0.50	8.14	25.88	60.14	50
		8.58	25.30	55.82	100
		12.82	30.16	57.50	500
	0.80	8.32	34.56	82.82	50
		8.78	28.32	69.38	100
		10.18	32.98	71.60	500

SD standard deviation

hypotheses increases. On the other hand, according to the results of our study, the distances given in the units of standard deviation decrease when the sample size increases since standard deviation is inversely proportional to the sample size. Under this circumstance, as the sample size increases, the proportion of the different sample gets closer to the proportions of the other sample. Because of this reason, the powers of the tests do not increase at an important level as the sample size increases. Westfall and Young [10] implemented multiple comparison procedures to the smallest three unadjusted *P* values which were chosen among the 24 Fisher' exact test results. When they compared tests in terms of adjusted *P* values, they decided that Bonferroni method is a more conservative method when compared to the methods of Sidak, Boostrap and Permutation style.

When simulation results were assessed, Type I error value was calculated to be around 20 % by UAPC approach, but this value was found to be 5 % lower than the result obtained through the formula of $(1 - (1 - \alpha)^{k(k-1)/2})$ included on the topic in literature. In terms of adjusted procedures' capability of preserving Type I error rate, they were found to be better than the approach of UAPC, like the findings of Cangur and Ankarali [20].

When observed power values obtained from the methods are evaluated in this study, power values of tests were found to be high when the proportion of three populations is equal in all sample sizes and the proportion of the fourth population is far away at a level of ±2 SD and more. However, under these conditions, power values for the methods of GBS, Holm–Bonferroni and Bonferroni except for UAPC were found to be higher than those of the other methods and calculated to be around 80 % on average. The expected power of the tests was stayed around 80 % because the power values were calculated under the condition when only one group shows a significant difference. On the other hand, it is an expected result that the power values obtained from UAPC approach were found to be around 95 %. This is because the fact that Type I error rate of this approach is found to be high is interpreted as finding a significant difference more than necessary in comparisons. According to these results, we think that it is important biologically, but in the small sample size we have difficulty in finding the difference, UAPC approach is recommended for comparisons. At the end of the simulation study, Oden et al. [5] found out that Pairwise Closed method is more powerful than the other methods they had examined. Moreover, they found that Hommel method is more powerful than Bonferroni method. On the other hand, they stated that Bonferroni procedure is less powerful than the other methods. Accordingly, it is confirmed that the some findings of Oden et al. [5] are compatible with the results of our study. Ryan [7] showed the application the methods of Adjusted Significance Levels and Tukey on a sample data set in terms of comparing proportions, and stated that Tukey method is obtained smaller interval values in terms of significance, and so it is a bit more powerful than the method of Adjusted Significance Levels. On the other hand, when the results of the sample data set are evaluated in this study, the results of the Tukey method have been found to be similar to the results of Bonferroni procedure. However, the generalizability of the results obtained from a single sample is weak. On the other hand, Horne and Plaehn [12] showed in their simulation study that Tukey method is at least slightly more powerful than Bonferroni procedure.

According to the results of this study, GBS and Holm–Bonferroni methods are recommended to be used because they preserve the predetermined Type I error rate in

contingency tables and they display a better performance in terms of power. On the other hand, Hommel and Marascuilo methods are not recommended to be used as they have a performance that is at the level or even below average. Although Marascuilo method is preferred or suggested to be used in terms of comparing proportions in literature [6], this study has showed that this method has a poorer performance.

Moreover, the number of proportions which were compared in applied studies has been observed generally to range from 3 to 6. Because of this reason, we use the results of comparisons belonging to four proportions (2×4 table) that have been given in our simulation study to have it shed light on applied researches. But we cannot use other $2 \times c$ tables, especially for $c > 10$. In addition, the results of 2×4 table can be generalized to 2×3, 2×5, and 2×6 tables. Also, when the results are examined, it can be said that these results can be generalized for more proportion comparisons. In addition, our study is different from other studies and is of great importance in that one of newest methods controlling primarily FDR as well as FWER has been chosen and included in our simulation study. Moreover, we have not come across the modules of the procedures suggested in our study in common package programs. Because of this reason, preparing a macro related to the usage of multiple comparisons procedures of GBS and Holm–Bonferroni is an innovation for the field of implementation.

References

1. Zar JH (1999) Biostatistical analysis, 4th edn. Prentice-Hall, Upper Saddle River, NJ
2. Elliott AC, Reisch JS (2006) Implementing a multiple comparison test for proportions in a 2×c crosstabulation in SAS®. Proceedings of the 31st annual SAS users group international conference. SAS Institute Inc, Cary, NC
3. Gavrilov Y, Benjamini Y, Sarkar SK (2009) An adaptive step-down procedure with proven FDR control under independence. Ann Stat 37:619–629
4. Blakesley RE, Mazumdar S, Dew MA, Houck PR, Tang G, Reynolds CF 3rd et al (2009) Comparisons of methods for multiple hypothesis testing in neuropsychological research. Neuropsychology 23:255–264. doi:10.1037/a0012850
5. Oden N, VanVeldhuisen PC, Ingrid US, Ip MS, SCORE Study Investigator Group (2010) SCORE Study report 8: closed tests for all pair-wise comparisons of means. Drug Inf J 44:405–420
6. Michael GA (2007) A significance test of interaction in $2 \times K$ designs with proportions. Tutor Quant Methods Psychol 3:1–7
7. Ryan TA (1960) Significance tests for multiple comparison of proportions, variances and other statistics. Psychol Bull 57:318–328
8. Marascuilo L (1966) Large-sample multiple comparisons. Psychol Bull 65:280–290
9. Holland BS, Copenhaver MD (1988) Improved Bonferroni-type multiple testing procedures. Psychol Bull 104:145–149
10. Westfall PH, Young SS (1989) P value adjustments for multiple tests in multivariate binomial models. J Am Stat Assoc 84:780–786
11. Kim SB, Tsui KL, Borodovsky M (2006) Multiple hypothesis testing in large-scale contingency tables: inferring patterns of pair-wise amino acid association in β-sheets. Int J Bioinform Res Appl 2:193–217
12. Horne J, Plaehn D (2007) Multiple comparisons on 2×c proportions. SAS conference proceedings: Pacific Northwest SAS Users Group 2007. Pacific Northwest SAS Users Group, Seattle, WA. www.pnwsug.org/content/multiple-comparisons-2xc-proportions. Accessed March 2014
13. Lancaster MB (1949) The derivation and partition of χ^2 in certain discrete distributions. Biometrika 36:117–129
14. Haberman SJ (1973) The analysis of residuals in cross-classified tables. Biometrics 29:205–220
15. Hsu JC (1996) Multiple comparisons: theory and methods. CRC Press, Boca Raton
16. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc B 57:289–300
17. Holm S (1979) A simple sequentially rejective multiple test procedure. Scand J Stat 6:65–70
18. Hochberg Y (1988) A sharper Bonferroni procedure for multiple tests of significance. Biometrika 75:800–802
19. Hommel G (1988) A stagewise rejective multiple test procedure based on a modified Bonferroni test. Biometrika 75:383–386
20. Cangur S, Ankarali H (2014) Examining the probabilities of type i error for unadjusted all pairwise comparisons and Bonferroni adjustment approaches in hypothesis testing for proportions. Int J Stat Med Res 3(4):404–411. doi:10.6000/1929-6029.2014.03.04.9