# Examining the Probabilities of Type I Error for Unadjusted All Pairwise Comparisons and Bonferroni Adjustment Approaches in Hypothesis Testing for Proportions

Sengul Cangur* and Handan Ankaralı

*Department of Biostatistics, Faculty of Medicine, Duzce University, Turkey*

**Abstract:** The aim of this study is to examine the association among the probabilities of Type I error obtained by Unadjusted All Pairwise Comparisons (UAPC) and Bonferroni-adjustment approaches, the sample size and the frequency of occurrence of an event (prevalence, proportion) in hypothesis testing of difference among the proportions in studies. In the simulation experiment planned for this purpose, 4 groups were formed and the proportions in each group were chosen between 0.10 and 0.90 so that they will be equal at each experiment. Furthermore, the sample sizes were chosen from 20 to 1000. In accordance with these scenarios, the probabilities of Type I error were calculated by both of approaches. In each approach, a significant S-curve relationship was found between the probability of Type I error and sample size. However, a significant quadratic relationship was found between the probabilities of Type I error and the proportions in each group. Nonlinear functional relations were put forward in order to estimate the observed Type I error rates obtained by the two different approaches where sample size and the proportion in each group are known. Furthermore, it was founded that Bonferroni-adjustment approach cannot always protect Type I error level. It was observed that the probability of Type I error estimated by the functional relation on Type I error rate for UAPC approach is lower than the values calculated using the formula in the literature.

**Keywords:** Proportion comparison, type I error, bonferroni adjustment, unadjusted all pairwise comparisons.

## INTRODUCTION

Proportion comparison methods and their post-hoc approaches are performed frequently in medical studies which have diagnostic or therapeutic purposes. In cases where the relevant null hypothesis is rejected when more than two proportions are required to compare, classical approaches such as Unadjusted All Pairwise Comparisons (UAPC), Standardized and Adjusted Residuals Statistics (STAR) and multiple comparison procedures protecting the Type I error established at the beginning are used to determine the proportions leading to the difference [1,2].

As more than two proportions are compared as if they are two-by-two independents with UAPC approach, the possibility of making Type I error established at the beginning increases. And in STAR approach, the interpretation of the normal probability graph becomes quite hard as the number of proportions to be compared increases. Despite these approaches, Bonferroni-adjustment approach is a method that is often preferred in medical studies as it protects the familywise error rate (FWER) established at the beginning. However, this method is known to be a conservative test as the number of proportions to be compared increases [3].

*Address correspondence to this author at the Department of Biostatistics, Faculty of Medicine, Duzce University, Turkey; Tel: +90 5375956051; Fax: +90 380 542 13 02; E-mail: sengulcangur@duzce.edu.tr

When UAPC approach is performed, it is known that Type I error level increases subject also to the number of proportions to be compared. However, no such study is encountered that examines in detail how this relation varies subject to sample size and the frequency of occurrence of an event (prevalence, proportion). The literature generally includes such studies which use multiple comparison procedures in comparison of proportions or compare the performances of these procedures [see: 4-8]. While it is said that the observed Type I probability error decreases when the sample size is increases, the relation among them is not mentioned in these studies.

The aim of this study is to examine the effects of the change in the sample size and the proportions in each group on Type I error rates obtained by UAPC and Bonferroni-adjustment approaches in hypothesis testing of difference among proportions. Furthermore, under which conditions the probabilities of Type I error calculated according to both approaches yield appropriate solutions will be examined.

## METHODOLOGY

### Unadjusted All Pairwise Comparisons (UAPC) Approach

After detecting that more than two proportions are significantly different according to chi-square test statistics, $P_i$ values of the test statistics for each test are obtained as a result of comparing these proportions

E-ISSN: 1929-6029/14 

with *t* or *z* statistics. When $H_i$ is the null hypothesis constructed for *i*th comparison and $P_i$ is the unadjusted probability of error calculated about the test statistic in the comparison *i*th $i=1,2,...,k$, hypotheses of $H_i,...,H_k$ are constructed for each of a total of *k* comparisons and at the end of hypotheses testing, a total number of $P_1,...,P_k$ probabilities of error are calculated. These $P_i$ values are compared with the Type I error value ($\alpha$) determined at the beginning. If $P_i<\alpha$, $H_i\,(i=1,2,...,k)$ is rejected. This approach is known as UAPC.

When multiple comparison is made, it is known that Type I error increases quickly with the increase in the number of groups (in other words the number of proportions) to be compared. This relation is defined in literature as $1-(1-\alpha)^k$ subject to the nominal level and the number of groups to be compared [3].

### Bonferroni-Adjustment Approach

Bonferroni-adjustment method is designed to keep FWER under control. This method is a powerful test that is easy to implement. It makes simultaneous inference. In this approach, FWER which is the probability of rejecting at least one hypothesis incorrectly in a definite set of hypotheses is controlled [3,9].

$$FWER=E\left(\frac{V}{m_0}|m_0>0\right) \quad (1)$$

In the formula, $m_0$ is the number of true null hypotheses and *V* means the number of rejected true null hypothesis (the number of false rejection).

The process steps of this method may be summarized as the following. When $H_i$ is the null hypothesis constructed for *i*th comparison and $P_i$ is the unadjusted probability of error calculated about the test statistic in the comparison *i*th $i=1,2,...,k$, hypotheses of $H_i,...,H_k$ are constructed for each of a total of *k* comparisons and at the end of hypotheses testing, a total number of $P_1,...,P_k$ probabilities of error are calculated. Each $P_i$ value is compared with $\alpha/k$ and the acceptance or rejection of the hypothesis is decided. According to Bonferroni-adjustment approach, the adjusted values of $P_i$ is obtained as below.

$$\tilde{P}_i=\{kP_i,\quad for\, i=1,2,..,k \quad (2)$$

In this equality, *k* means the number of comparisons. All $\tilde{P}_i$ values are decided on by comparing with $\alpha$.

## MATERIAL AND METHODS

We applied the proposed procedures in this study to the simulated 2×4 contingency tables. For example this table contains gender (male, female) × groups (placebo, drug 1, drug 2 and drug 3). In this simulation experiment, the proportions for 4 groups were equal in each trial and derived from binominal distribution. These values are chosen between 0.10 and 0.90. Furthermore, sample sizes were chosen from 20 to 1000. Sixty scenarios were created taking into account the twelve sample sizes and the five different proportions in this simulation study. In each scenario, Type I error probabilities of unadjusted and after adjustment using Bonferroni-adjustment approach were calculated. It was considered controlling actual Type I error at 0.05 in two procedures.

We used a macro that we wrote in Minitab programme (ver. 16.) for simulation study. Each scenario was done with 10000 repetitions.

As a result of the simulation, the relations between the probabilities of Type I error obtained using the two approaches and the sample size and the proportions in each group were put forward using Levenberg-Marquardt technique, one of Nonlinear least squares model estimation techniques. The significance of the formulas found with both approaches was assessed using goodness-of-fit indices such as Sum of Squared Error (SSE), Root Mean Squared Error (RMSE) and $R^2$.

## RESULTS

The probabilities of Type I error which are unadjusted and are adjusted using Bonferroni-adjustment approach obtained according to the simulation study where 60 different scenarios were constructed are listed on Table **1**. The relations between the probabilities of Type I error obtained by two different approaches and both the sample size and the proportions in each group were examined separately. In model selection, the model with the lowest error value and standard error value of estimates and the biggest $R^2$ value was advised as the appropriate model. The relevant results are listed on Table **2**.

With both approaches, a significant S-curve relationship was found between the probability of Type I error and sample size. When compared to other relations researched, the error value and standard error value of estimates of the model showing this relation, in

**Table 1: Observed Overall Type I Error Probabilities of Unadjusted All Pairwise Comparisons and Bonferroni-Adjustment Approaches**

| Scenarios for Type I error | | Observed Overall Type I Error (%) | |
|---|---|---|---|
| **Sample size** | **Proportions in each group ($P_i$ for $i$=1, 2, 3, 4)** | **Unadjusted All Pairwise Comparisons** | **Bonferroni-Adjustment** |
| 20 | 0.10 | 16.93 | 1.74 |
| | 0.30 | 21.63 | 4.88 |
| | 0.50 | 17.43 | 3.77 |
| | 0.70 | 21.12 | 4.44 |
| | 0.90 | 16.93 | 1.62 |
| 30 | 0.10 | 22.37 | 1.66 |
| | 0.30 | 19.41 | 4.28 |
| | 0.50 | 21.55 | 3.97 |
| | 0.70 | 20.02 | 4.26 |
| | 0.90 | 21.78 | 1.44 |
| 50 | 0.10 | 23.07 | 4.01 |
| | 0.30 | 20.24 | 3.97 |
| | 0.50 | 22.78 | 3.27 |
| | 0.70 | 20.04 | 4.07 |
| | 0.90 | 22.02 | 3.57 |
| 60 | 0.10 | 19.69 | 4.37 |
| | 0.30 | 20.47 | 3.95 |
| | 0.50 | 21.64 | 3.75 |
| | 0.70 | 20.96 | 3.79 |
| | 0.90 | 20.32 | 3.88 |
| 80 | 0.10 | 20.70 | 3.98 |
| | 0.30 | 20.28 | 4.40 |
| | 0.50 | 19.16 | 4.23 |
| | 0.70 | 20.90 | 4.07 |
| | 0.90 | 19.20 | 3.79 |
| 100 | 0.10 | 20.38 | 3.83 |
| | 0.30 | 21.82 | 4.39 |
| | 0.50 | 22.46 | 4.22 |
| | 0.70 | 21.15 | 4.44 |
| | 0.90 | 20.75 | 3.65 |
| 150 | 0.10 | 19.87 | 3.64 |
| | 0.30 | 20.47 | 4.31 |
| | 0.50 | 22.41 | 4.54 |
| | 0.70 | 20.47 | 4.17 |
| | 0.90 | 19.84 | 3.71 |
| 200 | 0.10 | 21.12 | 4.48 |
| | 0.30 | 21.25 | 4.18 |
| | 0.50 | 20.73 | 4.09 |
| | 0.70 | 20.91 | 4.28 |
| | 0.90 | 19.66 | 4.07 |

**(Table 1). Continued.**

| Scenarios for Type I error | | Observed Overall Type I Error (%) | |
|---|---|---|---|
| **Sample size** | **Proportions in each group ( $P_i$ for $i$=1, 2, 3, 4)** | **Unadjusted All Pairwise Comparisons** | **Bonferroni-Adjustment** |
| 300 | 0.10 | 20.56 | 4.12 |
| | 0.30 | 21.13 | 4.20 |
| | 0.50 | 21.94 | 3.98 |
| | 0.70 | 21.58 | 4.41 |
| | 0.90 | 20.28 | 3.90 |
| 500 | 0.10 | 19.75 | 4.13 |
| | 0.30 | 20.55 | 4.21 |
| | 0.50 | 22.25 | 4.21 |
| | 0.70 | 20.76 | 4.51 |
| | 0.90 | 20.30 | 4.09 |
| 750 | 0.10 | 20.34 | 3.93 |
| | 0.30 | 21.10 | 4.30 |
| | 0.50 | 21.04 | 4.41 |
| | 0.70 | 21.53 | 4.53 |
| | 0.90 | 20.20 | 4.07 |
| 1000 | 0.10 | 20.45 | 3.94 |
| | 0.30 | 20.77 | 4.29 |
| | 0.50 | 21.50 | 4.46 |
| | 0.70 | 20.97 | 4.45 |
| | 0.90 | 21.40 | 4.24 |

**Table 2: Nonlinear Functional Relations Obtained by Two Approaches**

| Approaches | Factors | Functional relations | *p* | *SSE* (*RMSSE*) | $R^2$ |
|---|---|---|---|---|---|
| Unadjusted All Pairwise Comparisons | Sample size | $_U Y_{Unadj} = Exp\left(3.049 - 1.504\left(1/X_{ssize}\right)\right)$ | 0.005 | 0.003 (0.434) | 0.129 |
| | | $_S Y_{Unadj} = Exp\left(-0.359\left(1/X_{ssize}\right)\right)$ | | | |
| | Proportions | $_U Y_{Unadj} = 19.938 + 4.825X_{prop} - 4.984X_{prop}^2$ | 0.095 | 1.386 (8.890) | 0.079 |
| | | $_S Y_{Unadj} = 1.141X_{prop} - 1.211X_{prop}^2$ | | | |
| Bonferroni Adjusted | Sample size | $_U Y_{Bonf} = Exp\left(1.471 - 8.782\left(1/X_{ssize}\right)\right)$ | <0.001 | 0.046 (1.628) | 0.263 |
| | | $_S Y_{Bonf} = Exp\left(-0.513\left(1/X_{ssize}\right)\right)$ | | | |
| | Proportions | $_U Y_{Bonf} = 3.303 + 4.147X_{prop} - 4.295X_{prop}^2$ | 0.004 | 0.409 (4.827) | 0.179 |
| | | $_S Y_{Bonf} = 1.706X_{prop} - 1.816X_{prop}^2$ | | | |

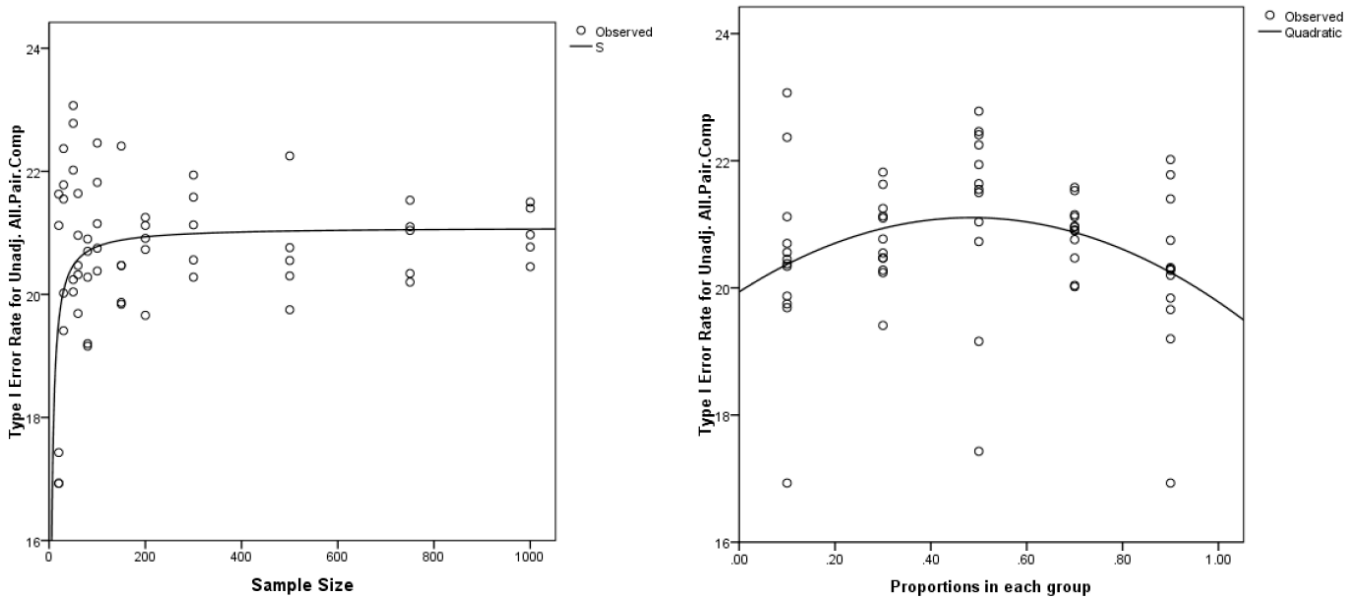
$Y_{Unad\,j}$ :Type I error rate values for Unadjusted All Pairwise Comparisons.

$Y_{Bonf}$ :Type I error rate values for Bonferroni Adjusted.

$X_{ssize}$ :Sample sizes, $X_{prop}$ : Proportion in each group.

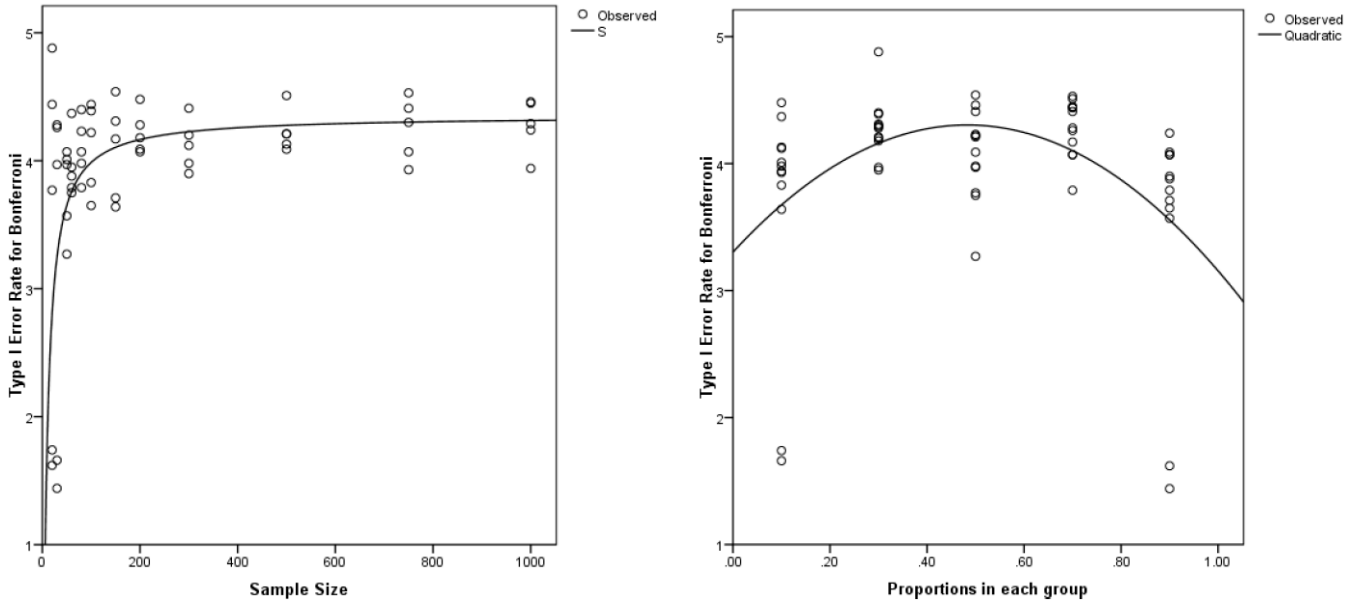
$_S Y$ : Nonlinear Standardized Regression Equation, $_U Y$ : Nonlinear Unstandardized Regression Equation.

SSE : Sum of Squared Errors, RMSSE :Root Mean Sum of Squared Errors.




(**a**) (**b**)

**Figure 1:** (**a**) S-curve relationship between sample size and Type I error rate for Unadjusted All Pairwise Comparisons (UAPC) approach (**b**) Quadratic relationship between Type I error rate for UAPC approach and the proportions in each group.




(**a**) (**b**)

**Figure 2:** (**a**) S-curve relationship between Type I error rate for Bonferroni-adjustment approach and sample size (**b**) Quadratic relationship between Type I error rate for Bonferroni-adjustment approach and the proportions in each group.

other words, the advised models, are the lowest and its $R^2$ is the biggest (Table **2**). The curves of the model are given in Figure **1a** and **2a** respectively. That extreme Type I error values emerge especially when the sample size is very small can be seen in these curves as well. A quadratic relationship was found between Type I error rates obtained by both approaches and the proportions in each group. Again, these relationships are shown in Figure **1b** and **2b** respectively.

That extreme Type I error values for Bonferroni-adjustment method emerge especially in the cases when the proportion in each group is close to 0 and 1 can be seen in the relevant curve (Figure **2b**).

Functional relations were found in order to examine how the probabilities of Type I error obtained using both approaches subject to sample size and the proportion in each group. The functional relations which

**Table 3: Nonlinear Regression Models Obtained by Two Approaches**

| Approaches | Functional relations | *p* | *SSE* (*RMSSE*) | $R^2$ |
|---|---|---|---|---|
| Unadjusted All Pairwise Comparisons | $_U Y_{Unadj} = 47.587 + 4.825X_{prop} - 4.984X^2_{prop} - 27.276Exp\left(1/X_{ssize}\right)$ | 0.008 | 1.241 (8.334) | 0.191 |
| | $_S Y_{Unadj} = 1.141X_{prop} - 1.211X^2_{prop} - 0.334Exp\left(1/X_{ssize}\right)$ | | | |
| Bonferroni Adjusted | $_U Y_{Bonf} = 26.769 + 4.147X_{prop} - 4.295X^2_{prop} - 23.149Exp\left(1/X_{ssize}\right)$ | <0.001 | 0.293 (4.051) | 0.421 |
| | $_S Y_{Bonf} = 1.706X_{prop} - 1.816X^2_{prop} - 0.493Exp\left(1/X_{ssize}\right)$ | | | |

$Y_{Unadj}$ :Type I error rate values for Unadjusted All Pairwise Comparisons.
$Y_{Bonf}$ :Type I error rate values for Bonferroni Adjusted.
$X_{ssize}$ :Sample sizes, $X_{prop}$ : Proportion in each group.
$_S Y$ : Nonlinear Standardized Regression Equation, $_U Y$ : Nonlinear Unstandardized Regression Equation.
SSE : Sum of Squared Errors, RMSSE :Root Mean Sum of Squared Errors.

were found significant or which reflect the association among the factors are given in Table **3**. The model error values and standard error values of estimates of the advised models for both approaches among the models examined are quite low and their $R^2$ values are the biggest (Table **3**).

The 3D presentation of the functional relations showing how the probabilities of Type I error unadjusted or adjusted using Bonferroni-adjustment approach change subject to sample size and the proportions in each group are given in Figure **3** and **4**.

While the proportions in each group is 0.1 and lower or 0.9 and higher, in cases when the sample size is lower than 30, both Bonferroni-adjusted and unadjusted Type I error rates were found quite low when compared to other conditions (Table **1**). When 3D graphs were examined, it was observed that diffraction fault is formed on both surfaces when the sample size is small and the proportion is low. It was detected that Type I errors with low extreme value exist in the two edges of the surface where the sample size is small and the proportion is close to the lower and upper limit

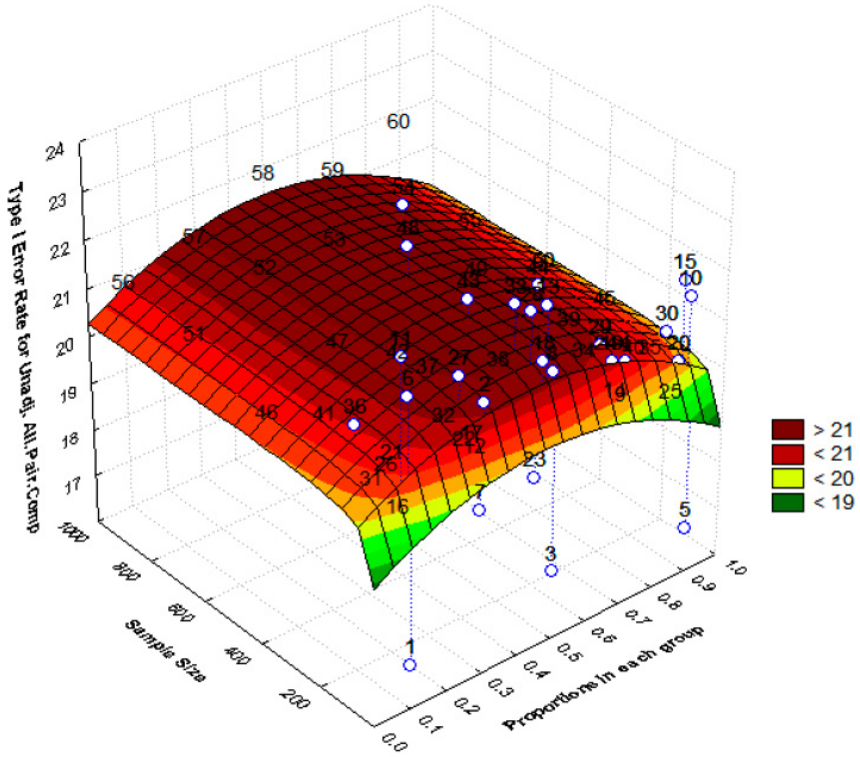



**Figure 3:** 3D function graph of Type I error rate for Unadjusted All Pairwise Comparisons approach.

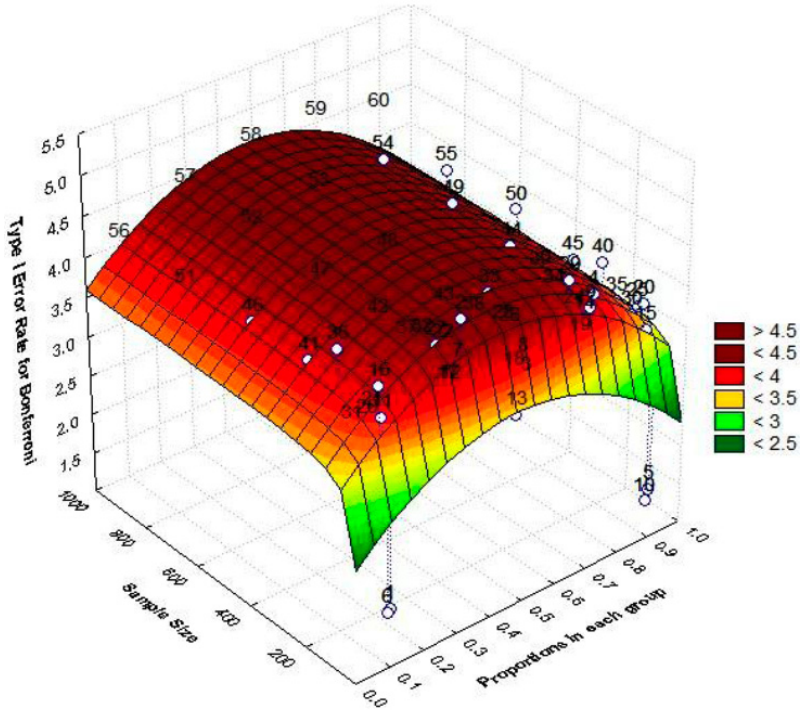



**Figure 4:** 3D function graph of Type I error rate for Bonferroni-adjustment approach.

values. This appeared more obviously on the relevant surface with Bonferroni-adjustment approach.

It was observed that Bonferroni-adjusted Type I error value is below 2% where the sample size is lower than 30 and the proportion in each group is 0.1 and 0.9. Furthermore, Bonferroni-adjusted Type I error value is around 4% (or closer to 5%) where the sample size is lower than 30 and the proportion in each group is between 0.3 and 0.7 (Table **1** and Figure **4**). By contrast, Type I error rate for UAPC approach is 16.93% on average and it was observed that these rates vary between 17%-23% (Table **1** and Figure **3**). For example, a researcher examined four different generic drugs which were applied to the total of 450 individuals in terms of drug-related side effect which the proportions of abdominal pain side-effect are expected to vary from 0.10 to 0.30, and achieved difference at significance level of 5%. It was supposed that researcher was used Bonferroni-adjustment and UAPC approaches as the post-hoc test in order to determine the drug/drugs that cause(s) to this significant difference. Researchers can easily predict the observed Type I error rates of two approaches described in this study, with the help of the functional relationships in Table **3**. When sample size is 450 $\left(X_{ssize}\right)$ and the average abdominal pain side-effect rate of four different generic drugs is 0.18 $\left(X_{prop}\right)$, the amount of observed Type I error will be 20.96% by UAPC approach $\left({}_{U}Y_{Unadj}\right)$. When pooled proportion is 8% instead of 18%, it is seen that the Type I error value for Bonferroni-adjustment approach $\left({}_{U}Y_{Bonf}\right)$ decreases (3.87%). Based on the results of this example, we can say that Type I error value of Bonferroni-adjustment approach changes in different conditions and generally less than 5% of value.

## CONCLUSION

In diagnostic or therapeutic purposes medical studies conducted on human or animal subjects, it is important that the sensitivity shown to the rules of ethics and experiment is also continued in the statistical process particularly in terms of obtaining of unbiased results. This can be provided with the protection of the Type I error which is thought to depend on study conditions (sample size, scale type etc.) and statistical methods.

In this study, functional relations which may reveal simultaneously the effects of the proportions in each group and sample size in hypothesis testing of difference among the proportions on Type I error rates obtained using UAPC and Bonferroni-adjustment approaches were investigated.

While a significant S-curve relationship was found between Type I error rate and sample size with both approaches, a quadric relationship was found between the probability of Type I error and the proportion in each group. It was observed that the extreme Type I

error values emerge either when the sample size is quite small or the proportion is close to 0 or 1 as in Bonferroni-adjustment approach. When the results are assessed in terms of both conditions, it was concluded that the probabilities of Type I error are again much higher than the expected value and Bonferroni-adjusted Type I error rates are much lower than the expected value, in other words, it produces strict results.

In medical research compared the difference among the proportions, usually it is thought that Bonferroni-adjustment approach has best performance. Therefore many of the studies in health field are given the findings of this method in terms of significant differences. However according to the results of our study, Bonferroni-adjustment approach cannot always protect the level of error at the beginning and the test yields strict results when the sample size is below 30 and the proportion is 0.1 and 0.9. It may be advisable not to use this approach which finds difficult the significant differences in such cases.

In addition, many researchers express based on knowledge of the literature that the error rate increases when UAPC approach is used. But according to study conditions it is not known how much of the amount of the error. However in our study, it was observed that the probability of Type I error (17%-23%) estimated from the functional relation on Type I error for UAPC approach is lower than the value calculated using the formula $\left(1-(1-\alpha)^k\right)$ in the literature –which is only subject to the level of error at the beginning and the number of groups to be compared.

By means of the functional relations that we advised as a result of this study, the researchers may estimate the observed Type I error values obtained by two different approaches where the sample size and the proportion in each group are known. And this may be important at least for the selection of the appropriate multiple comparison procedure which will ensure that Type I error rate remains at nominal level at the end of the study in terms of continuing sensitivity shown during medical researches with human subject or laboratory animal.

## REFERENCES

[1] Lancaster MB. The derivation and partition of $\chi^2$ in certain discrete distributions. Biometrika 1949; 36: 117-29. http://biomet.oxfordjournals.org/content/36/1-2/117.full.pdf+html

[2] Haberman SJ. The analysis of residuals in cross-classified tables. Biometrics 1973; 29: 205-20. http://dx.doi.org/10.2307/2529686

[3] Zar JH. Biostatistical analysis. 4th ed. Upper Saddle River, NJ: Prentice-Hall 1999.

[4] Ryan TA. Significance tests for multiple comparison of proportions, variances and other statistics. Psychol Bull 1960; 57: 318-28. http://dx.doi.org/10.1037/h0044320

[5] Elliott AC, Reisch JS. Implementing a multiple comparison test for proportions in a 2xc crosstabulation in SAS®. Proceedings of the 31st Annual SAS® Users Group International Conference; 2006: March 26-29; San Francisco, CA. Cary, NC: SAS Institute Inc 2006: pp. 204-31.

[6] Kim SB, Tsui KL, Borodovsky M. Multiple hypothesis testing in large-scale contingency tables: inferring patterns of pair-wise amino acid association in β-sheets. IJBRA 2006; 2: 193-217. http://dx.doi.org/10.1504/IJBRA.2006.009768

[7] Horne J, Plaehn D. Multiple comparisons on 2xc proportions. SAS Conference Proceedings: Pacific Northwest SAS Users Group 2007 (PNWSUG). Seattle, WA: Pacific Northwest SAS Users Group; 2007. Available from: www.pnwsug.org/content/multiple-comparisons-2xc-proportions

[8] Teriokhin AT, de Meeûs T, Guégan JF. On the power of some binomial modifications of the Bonferroni multiple test. Zh Obshch Biol 2007; 68(5): 332-40.

[9] Toothaker LE. Multiple comparison procedures. Newbury Park, CA: Sage 1993.

Received on 09-08-2014 Accepted on 11-10-2014 Published on 06-11-2014

http://dx.doi.org/10.6000/1929-6029.2014.03.04.9