Snežana Konjikušić[1]
Slađana Starčević [2]
Saša Mašić[3]

JEL: C10, C12, O30''
DOI: 10.5937/industrija47-22081
UDC:005.336.6:640.412(497.11)
005.311.11:303.62
Original Scientific Paper

Post Hoc Analysis of Serbian hotel ratings[4]

Article history:
Received: 10 July 2019
Sent for revision: 2 August 2019
Received in revised form: 2 September 2019
Accepted: 3 September 2019
Available online: 15 October 2019

Abstract: *The purpose of this paper is to present basic characteristics and highlight the differences between post hoc tests, as well as to show their application on concrete data of the research conducted. The said tests are applied on data obtained in the research which found evidence of 240 Serbian hotel ratings, given by their 71,700 guests. Each guest rated: cleanliness, comfort, location, facilities, staff, value for money, and free Wi-Fi in the hotel. A difference in ratings in relation to hotel category was observed and explained using several post hoc tests. The use of those tests is made much easier with the development of numerous statistical software packages.* Therefore, clearly differentiating each of the tests allows one to select the most appropriate test in the research process, according to the type of data and research objectives.*The paper presents the tests used when one-way analysis of variance, which is a method frequently used in statistical processing of experimental data, finds evidence of the existence of statistically significant differences in values of arithmetic mean in groups of data observed. The task of post hoc tests is to determine which group of data leads to the difference observed. Tests thus presented here are: the Fisher LSD, the Tukey HSD, the Bonferroni , the Newman-Keuls, the Dunnett and the Scheffé test.*

Keywords*: post hoc tests, hotel rating, sample arithmetic mean, statistical tests.*

[1] Metropolitan University – FEFA, Faculty of Economics, Finance and Administration, zansak@gmail.com
[2] Metropolitan University – FEFA, Department of Economics
[3]The College of Hotel Management, Belgrade
[4]This research paper was part of the research project No. 47028, financed by the Serbian Ministry of Science and Technological Development.

Post hoc analiza ocena hotela u Srbiji

Apstrakt: *Svrha ovog rada je da predstavi osnovne karakteristike i istakne specifičnosti različitih post hoc testova, kao i da se prikaže njihova primena nad konkretnim podacima sprovedenog istraživanja. Pomenuti testovi su primenjeni nad podacima prikupljenim u istraživanju kojim su evidentirane ocene za 240 domaćih hotela, od strane njihovih 71.700 gostiju. Svaki gost je davao ocenu za: ćistoću, komfor, lokaciju, servise, osoblje, dobijenu vrednost usluge za dati novac i besplatni WiFi internet hotela. Utvrđena je razlika u ocenama hotela u odnosu na kategoriju i ona se objašnjava primenom različitih post hoc testova. Upotreba tih testova je umnogome olakšana sa razvojem mnogobrojnih statističkih softvera. Iz tog razloga je značajno praviti jasnu razliku između ovih pojedinačnih testova da bi u procesu istraživanja bio izabran najadekvatniji, u skladu sa tipom podataka, ali i ciljevima istraživanja. Ovde će biti predstavljeni testovi koji se koriste kada jednofaktorska analiza varijansi, često korišćena u statističkoj obradi eksperimentalnih podataka, evidentira postojanje statistički značajne razlike u vrednostima aritmetičkih sredina kod posmatranih grupa podataka. Zadatak pos hoc testa je da ustanovi koja grupa podataka dovodi do opaženih razlika. Testovi višestrukog poređenja, odnosno post hoc testovi, predstavljeni u ovom radu su: Fisherov LSD, Tukey HSD, Bonferoni, Newman-Keuls, Dunnett i Scheffe test.*

Ključne reči*: post hoc testovi, ocene hotela, uzoračke aritmetičke sredine, statistički test.*

1. Introduction

The common problem of a research process is to determine whether there are differences between arithmetic means of groups; moreover, if differences are found, which of them are significant, in relation to the dependent variable. The one-way analysis of variance is used when the dependent variable is continuous with normal distribution and with homogenous variances within data divided into three or more groups, in relation to the independent variable. This analysis belongs to the group of parametric methods so, when a difference is found, the null hypothesis is rejected, and the next step in the process is the application of post hoc tests or multiple comparison tests (Westfall,P. et al. 2011). A disadvantage of this parametric method is that assumptions on the normal distribution and the homogeneity of variances must be met. When these conditions are not complied with, non-parametric methods are used, as well as their post hoc tests, which are not the subject of this paper.

Multiple comparison tests are often used and, although present in the statistical literature, it is still impossible to confidently answer the question often raised: what is the best test for routine usage. When using post hoc tests, it is crucially important to take into account the probability of Type I errors. These tests control Type I error in different manners, most often due to subjective and technical issues, but rather come to the basic question of the maximum acceptable error to a researcher (Frane, 2015). A short overview of these tests enables a researcher to make a more rational decision suited to his own research.

The Fisher LSD test, or the Least Significant Difference test, was presented by Fisher in 1953, and is considered the weakest one. It is based on common t-tests between all pairs, where the significance level for each t-test is α, and thus more likely to reject more than required.

The most often used tests for multiple comparison of pairs are the Newman-Keuls test and the Tukey test. Newman-Keuls test was introduced by Newman in 1939 and further developed by Keuls in 1952. It is used when the sample sizes differ significantly. Its popularity somewhat decreased after the 50's and 60's, due to introduction and growing consideration of FWER (familywise error rate) approach. This test uses different critical values for different comparison pairs (Seaman et al. 1991). Hence, there is a higher likelihood of discovering significant differences between arithmetic means of samples as well as to make a Type I error, i.e. to reject the null hypothesis even when correct. The Newman-Keuls test is more powerful yet less conservative than the similar Tukey test, presented in 1952. The purpose of Tukey's test is to figure out which groups in the sample differ. It uses the "Honest Significant Difference," a number that represents the distance between groups, to compare every mean with every other mean. Tukey test is most commonly used for same size samples. When this is not the case, there is the Tukey-Kremer test.

The Bonferroni test uses α/k significance level for each comparison (where *k* is the number of samples). The Bonferroni does suffer from a loss of power. This is due to several reasons, including the fact that Type II error rates are high for each test. In other words, it overcorrects for Type I errors (Hochberg, 1988). The ordinary Bonferroni method is sometimes viewed as too conservative. Holm's sequential Bonferroni post hoc test is a less strict correction for multiple comparisons.

Another frequently used post hoc test is the Dunnett's test, developed in 1955 by a Canadian statistician - Charles Dunnett. The original test was revised in 1964 by introducing critical values. Unlike the aforementioned tests, this test is often used when there is a control group in research, i.e. when this control group is to be compared to each sample individually (Noble, 2009). With this

test, error is never greater than the presumed level of significance, and is often used in medicine and agronomy.

In statistics, Scheffé's test, named after the American statistician, Henry Scheffé, is most often used when post hoc comparisons of combinations of arithmetic means are conducted. It enables the estimation of differences of all potential contrasts, and not only the paired differences, as is the case with the aforementioned tests. In statistics, particularly in variance analysis, a contrast is a linear combination of variables (parameters or statistics) whose coefficients add up to zero. Scheffé's test is more conservative than the Newman-Keuls test, revealing less significant differences, when compared to other tests. On the other hand, the advantage of this test is its lower sensitivity both to deviance of data distribution from the normal distribution, and to the distortion of the variance homogeneity assumption.

The process of choosing the appropriate statistical test might be a challenging task, hence good knowledge and understanding of relevant statistical terms could help in better decision-making. It is especially important to know which data type is analyzed, how this data is organized, how many samples are observed, whether data is dependent or independent, and if it is necessary for the data to have normal distribution. When data is normally distributed with equal variances and if the measurements are independent, then it is needed to know if parametric tests are applicable. It is crucially important for the variance homogeneity assumption to be correct, as multiple comparative testing should not be performed if heteroscedasticity has been declared. Generally speaking, parametric methods require more assumptions than non-parametric methods. However, if all assumptions are correct, more accurate and precise results will be obtained. That is why they are considered statistically more reliable. It is desirable for multiple comparison tests to have all samples of the same size, for obtaining the most reliable and robust test, but it is often not the case in real life.

2. Research methodology

The research involved 240 hotels in Serbia, with 86 hotels in Belgrade, 23 in Novi Sad, 18 in Niš, 11 at Kopaonik, 10 in Kragujevac, 9 in Subotica and at Zlatibor, 6 at Vrnjačka Banja, 5 in Novi Pazar, 4 in Čačak, Kraljevo and Pirot, 3 in Jagodina, Kanjiža and Leskovac, 2 in Šabac, Zaječar, Bor, Stara Pazova, Ruma, Vršac, Bajina Bašta, Paraćin, Sremski Karlovci and Kruševac, and 1 in Smederevo, Vinča, Požarevac, Veliko Gradište, Ždrelo, Banja Koviljača, Vranje, Aleksinac, Sokobanja, Dimitrovgrad, Negotin, Knjaževac, Bačka Palanka, Vrbas, Vrdnik, Zrenjanin, Novi Bečej, Sombor, Požega, Perućac, Aranđelovac and Sjenica.

Hotel guests covered by this research provided their ratings from 1 to 10 on the Booking website for: cleanliness, comfort, location, facilities, staff, value for money, and free Wi-Fi. The cleanliness rating, for example, was formed by finding an average value of all ratings from hotel guests who participated in the hotel rating process on the Booking website. In the same vein, the ratings of other hotel categories were formed. For all those ratings, an average rating value was found for every hotel, thus giving one rating to the hotel in question, which was formed as an average value of all ratings of the hotel's characteristics. Total of 71,700 hotel guests participated in the rating of 240 hotels. The smallest number of guests who gave their hotel rating is 30 (Dunav Hotel in Sremska Kamenica; Nacional Hotel in Belgrade; Raj Hotel in Novi Pazar; Omorika Hotel in Bajina Bašta; Ozon Hotel in Brzeće; Patria Hotel in Subotica), and the biggest number of guests who participated in the research is 2,354 guests (Moskva Hotel in Belgrade).

The hotels observed belonged to the different categories, precisely, the hotels in a category of one to five stars were analyzed. By applying the one-way analysis of variance it is possible to determine whether there is a significant statistical difference in hotel rating in relation to hotel category (number of stars). A series of data, which represent hotel ratings, is not normal distribution, because of Skewness/SE=8.1 and Kurtosis/SE=7.4. Therefore, it is necessary to transform data of the observed series, with the goal of having a normal dispersion of the transformed data so as to apply the one-way analysis. A graphical preview of the series of data ratings indicated that data should be transformed into a new series using the f(x)=1/(11-x) function. For the data transformed by using the said function: Skewness/SE=1.6 and Kurtosis/SE=1.1. The formerly given measures show that the transformed data have a normal distribution.

The statistical package IBM SPSS 20.0 was used for data processing. The significance level of 0.05 was used throughout all tests.

3.Results and discussion

Since the one-way analysis of variance (F(4,71700)=4.826; p-value=0.001) determined the existence of significant differences between the ratings of hotels of different categories, it was necessary to use the post hoc tests, so as to determine what categories lead to these differences, i.e. what hotel categories differentiate per guest rating. There is a moderate difference between the arithmetic means of the ratings of hotels of different categories. The size of that difference, expressed as the eta-square meter, is 0.076. Averages of transformed ratings are rising, specifically for hotels: one-star, average=0.3658, SD=0.13888; with two-stars, average=0.4021, SD=0.14056;

with three-stars, average=0.4200, SD=0.11341; with four-stars, average=0.4731, SD=0.10554; with five-stars, average=0.4927, SD=0.8947.

Six post hoc tests will been shown as follows: theoretical first, and then the results of the presented test for hotel ratings. Finally, and certainly being the crucial for this paper, test results will been compared, with the aim of perceiving the importance of theoretical knowledge of post hoc tests, both in their application process and in the result interpretation.

3.1. LSD test

The Fisher LSD test (Least significant difference) is the simplest, yet the least reliable test used to compare the differences between arithmetic means of observed samples. Its application implies simultaneous multiple comparison, where each comparison is performed with the same α significance level. Hence, the error risk in comparing all means is most certainly higher than α. It can also be noted that this test is quite liberal, not controlled by the overall α. The formula for calculating the least significant difference is the following one:

$$\mathrm{LSD} = \mathrm{t}_{\nu,\frac{\alpha}{2}} \cdot \mathrm{s} \cdot \sqrt{\frac{1}{\mathrm{n_i}} + \frac{1}{\mathrm{n_j}}} \qquad (1)$$

where $s = \sqrt{\mathrm{MSE}}$, MSE – residual variance, $\mathrm{t}_{\nu,\frac{\alpha}{2}}$ critical value of two-tailed *t*-distribution test, $\nu = \mathrm{k} \cdot (\mathrm{n_i} - 1)$, k- number of samples, n_i – number of data in the *i-th* sample, *i* =1,2,...,*k*. It is not necessary for the sample sizes of n_i and n_j to be the same. For each pair of samples whose absolute difference of arithmetic means is higher or equal to the LSD value, it is concluded that there is a significant statistical difference between arithmetic means of observed samples (Davis, 2002). In all other cases, there is no difference.

Results of this test for hotel ratings in relation to hotel categories are given in the following table:

Table 1. Results of the LCD test

(I)	(J)	Mean Difference (I-J)	Std. Error	Sig.	95% C.I. Lower Bound	95% C.I. Upper Bound
1	2	-.03636	.05067	.474	-.1362	.0635
	3	-.05425	.04816	.261	-.1491	.0406
	4	-.10736	.04794	.026	-.2018	-.0129
	5	-.12698	.06166	.041	-.2485	-.0055
2	3	-.01790	.02327	.443	-.0637	.0280
	4	-.07100	.02281	.002	-.1159	-.0261
	5	-.09062	.04500	.045	-.1793	-.0020
3	4	-.05310	.01649	.001	-.0856	-.0206
	5	-.07273	.04214	.086	-.1558	.0103
4	5	-.01962	.04189	.640	-.1022	.0629

Source: Authors` calculations

Based on the test's results, the difference in hotel rating in relation to hotel category was noticed for five pairs of categories. Namely, these ratings differ for pairs: one and four stars; one and five stars; two and four stars, two and five stars; three and four stars.

3.2. Tukey HSD (Honestly Significant Difference) test

The Tukey test is often used in the process of multiple comparison of arithmetic means of pairs of samples. If the analysis comprises *k* samples, the Tukey test analyses total of $\binom{k}{2}$ differences between arithmetic means. This test analyses the null hypothesis $H_0: \mu_i = \mu_j$ (i≠j) i,j=1,2,...,k compared to the alternative hypothesis, $H_1: \mu_i \neq \mu_j$, where μ_i and μ_j, arithmetic means. When the *q*-value is as follows:

$$q = \frac{|\mu_i - \mu_j|}{SE} \quad (2)$$

$$SE = \sqrt{\frac{MSE}{n_i}} \quad (3)$$

i.e. higher or equal to the critical table value of the Studentized range $q_{\alpha,\nu,k}$, the aforementioned null hypothesis is rejected. When samples whose arithmetic means are compared have an uneven number of data, the standard SE error is calculated as follows:

$$SE = \sqrt{\frac{MSE}{2}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)} \quad (4)$$

When this test is being applied, the arithmetic means of samples are sorted in descending order. First, the maximum value of arithmetic means is compared to the minimal one, then to the next minimal one, repeating the process until the maximum value of arithmetic mean is compared to the second highest arithmetic means value. The same process is carried out for the next highest value, et cetera.

It may happen that ANOVA indicates the existence of differences, while the Tukey test cannot determine which samples showed differences. In such cases, it is necessary to increase the number of observed units in a sample. This test offers optimal balance of the Type I and Type II error ratio.

The following table contains the results of this hotel rating test in relation to the categories observed:

Table 2. Results of the Tukey HSD test

(I)	(J)	Mean Difference (I-J)	Std. Error	Sig.	95% C.I. Lower Bound	95% C.I. Upper Bound
1	2	-.03636	.05067	.952	-.1757	.1030
	3	-.05425	.04816	.792	-.1866	.0781
	4	-.10736	.04794	.169	-.2391	.0244
	5	-.12698	.06166	.242	-.2965	.0425
2	3	-.01790	.02327	.939	-.0819	.0461
	4	-.07100	.02281	.018	-.1337	-.0083
	5	-.09062	.04500	.263	-.2143	.0331
3	4	-.05310	.01649	.013	-.0984	-.0078
	5	-.07273	.04214	.420	-.1886	.0431
4	5	-.01962	.04189	.990	-.1348	.0955

Source: Authors` calculations

Unlike the previous one, this test found differences in ratings for only two pairs of categories. Namely, this test reveals the difference in the ratings of two and four star hotels, as well as for hotels with three and four stars.

3.3. Newman-Keuls test (Student-Newman Keuls test)

The Newman-Keuls test compares the arithmetic means of all pairs of samples, similar to the Tukey test. The comparison is based on testing the null hypothesis, $H_0: \mu_i = \mu_j$ (i≠j) i,j=1,2,...,k as well as the alternative hypothesis, $H_1: \mu_i \neq \mu_j$, where μ_i and μ_j are the arithmetic means. The arithmetic means of samples are ranked, the differences between the arithmetic means pairs are calculated, then the standard error is calculated by applying the formula (3) or (4). The *q*-value in this test is the same as the one in the Tukey test, whereas critical values of these two tests differ. The critical value of Newman-Keuls test is $q_{\alpha,v,p}$, with α and *v* values explained above, while *p* is the number of arithmetic means in the range of analyzed arithmetic means. The next example illustrates the process of calculating the *p* value.

For instance, if arithmetic means of observed samples is the following $\bar{x}_5 = 27.3$, $\bar{x}_4 = 22.7$, $\bar{x}_2 = 19.8$, $\bar{x}_3 = 15.9$ and $\bar{x}_1 = 12.3$ in descending order, the *p* value of each pair of samples compared is given in the following table:

Table 3. Rank arithmetic means

Pairs	p	Pairs	p
5-1	5	4-3	3
5-3	4	4-2	2
5-2	3	2-1	3
5-3	4	2-3	2
4-1	4	3-1	2

Source: Authors` calculations

To compare the ranked arithmetic means, such as $\bar{x}_5$ and $\bar{x}_3$ (5-3) from the presented example, four arithmetic means should be considered, hence *p*=4. Since this test uses different critical values for different ranking of arithmetic means, it is often referred to as the Multiple range test. The results of this test are statistically more important than the results of the Tukey test, i.e. the results of this test are more reliable. Regardless of the aforementioned, these tests could, but do not have to, bring about the same conclusions. The Tukey test is, on the one hand, considered conservative, because according to Miller (1981), it records too little statistically significant differences, whereas, on the other hand, (Ramsey, 2009) emphasizes that this test could falsely state significant differences with likelihood higher than α.

Having been criticised, Tukey suggested using the arithmetic means of critical values of both Tukey and Newman-Keuls tests in the process of comparing differences of arithmetic means of critical values. This modified test is referred to as the Honestly significant difference test.

The results of this test for the observed ratings are given in the following table:

Table 4. Results of the S-N-K test

Categories	Subset for alpha = 0.05	
	1	2
1	.3658	
2	.4021	.4021
3	.4200	.4200
4	.4731	.4731
5		.4927
Sig.	.057	.143

Source: Authors` calculations

Unlike the previous two tests, this one yields just two groups of ratings that do not differ much. The first group is thus made of the ratings of the hotels with one, two, three and four stars, while the second group comprises those with two, three, four and five stars.

3.4. Bonferroni test

To maintain the Type I error risk at the α level, by applying the Bonferroni test, in the process of forming the reliability interval for the arithmetic means difference of two samples (of *k* samples), the significance level is not α but $\frac{\alpha}{C_2^k}$, where C_2^k – is the number of combinations of *k* elements taken two at a time, equaling the number of pairs whose arithmetic means are compared (Noble, 2009). Consequently, that makes this test a conservative one, so occasionally significant differences might not be discovered.

For the α significance level, $100 \cdot (1 - \alpha')\%$ confidence interval for μi-μj, i≠j, i,j=1,2,...,k is as follows:

$$(\bar{x}_i - \bar{x}_j) \pm t_{\alpha \cdot v} \cdot s \cdot \sqrt{\frac{1}{n_i} + \frac{1}{n_j}} \qquad (5)$$

where $s = \sqrt{MSE}$, $\alpha' = \frac{\alpha}{C_2^k}$, $t_{\alpha',v}$- the table value of *t*-distribution, and $v = k \cdot (n_i - 1)$.

By using this test, for each confidence interval formed this way, if containing zero, it is possible to conclude that there are no differences between the observed arithmetic means. In other cases, the differences do exist.

The following table contains the results of this test applied to hotel ratings. This test found the difference in the ratings of the two and four star hotels, as well as in the evaluation of those with three and four stars.

Table 5. Results of the Bonferroni test

(I)	(J)	Mean Difference (I-J)	Std. Error	Sig.	95% C.I. Lower Bound	95% C.I. Upper Bound
1	2	-.03636	.05067	1.000	-.1800	.1072
	3	-.05425	.04816	1.000	-.1907	.0822
	4	-.10736	.04794	.261	-.2432	.0285
	5	-.12698	.06166	.406	-.3017	.0478
2	3	-.01790	.02327	1.000	-.0838	.0480
	4	-.07100	.02281	.021	-.1356	-.0064
	5	-.09062	.04500	.452	-.2181	.0369
3	4	-.05310	.01649	.015	-.0998	-.0064
	5	-.07273	.04214	.857	-.1922	.0467
4	5	-.01962	.04189	1.000	-.1383	.0991

Source: Authors` calculations

3.5. Dunnett test

The Dunnett test is used when there is a need to compare arithmetic means of one sample, most commonly referred to as the control sample, with arithmetic means of the remaining k-1 samples. Using this test, if the absolute difference (of two-tailed test) of arithmetic means of the *i-th* (*i*=1, 2,...,*k*-1) and the control sample is higher or equals

$$D = d_{\alpha,v,k} \cdot SE \qquad (6)$$

then the obtained difference between the observed arithmetic means of samples is regarded statistically significant. From the given formula, $d_{\alpha,v,k}$

$d_{\alpha,v,k}$ is the table critical value for the α significance level, $v=k(n_i-1)$ is degree of freedom , while *k* is the number of observed samples. If each sample has the same data size (*n*), then:

$$\mathrm{SE} = \sqrt{\frac{2\mathrm{MSE}}{\mathrm{n}}} \qquad (7)$$

When the analyzed samples contain a different number of observed units, the formula is as follows:

$$SE = \sqrt{MSE \cdot (\frac{1}{n_i} + \frac{1}{n_{control}})} \ . \qquad (8)$$

The control sample in this test should contain more data than any other sample. The optimal data size in the control sample should comply with the following formula:

$$n_{control} = \sqrt{(k-1)} \cdot n_i \qquad (9)$$

The results of this test are contained in the following table. The ratings of five star hotels are taken as control values. In this research, any group could be selected as a control group. It can be seen from the table that there is no difference between any group in relation to the control one, which is clearly opposite to the one-way analysis of variance.

Table 6. Results of the Dunnett test

(I)	(J)	Mean Difference (I-J)	Std.Error	Sig.	95% C.I.	
					Lower Bound	Upper Bound
1	5	-.12698	.06166	.097	-.2715	.0176
2	5	-.09062	.04500	.106	-.1961	.0149
3	5	-.07273	.04214	.190	-.1715	.0261
4	5	-.01962	.04189	.928	-.1178	.0786

Source: Authors` calculations

3.6. Scheffé test

The aforementioned tests perform multiple comparisons of the arithmetic means of the individual samples. Scheffé test is used when it is necessary to test the arithmetic means of several linked samples with arithmetic mean of one or more samples. It is often referred to as the S test. The differences analyzed in such cases are called *contrasts* (Rumsey, 2009). These contrasts could be linear, square, cubic (polynomial), orthogonal, etc. and are usually not predetermined. Significance level of each comparison in this test is α at most. The contrast is defined in the following way:

$$C = \sum_{i=1}^{k} c_i (\sum_{j=1}^{n_i} x_j) \quad (10)$$

where c_i, i=1,...,k are the contrast weights whose values comply with the equation $\sum_{i=1}^{k} c_i = 0$. For instance, contrast is used to determine whether the arithmetic means of $\bar{x}_1$, $\bar{x}_3$, andi $\bar{x}_4$ differ from the arithmetic mean of the fifth sample.

The null hypothesis of the Scheffé test could be stated as $H_0: \mu_A - \mu_B = 0$. Thus, the null hypothesis of the previous example is $H_0: \frac{\mu_1+\mu_3+\mu_4}{3} - \mu_5 = 0$, $where\ c_1 = c_3 = c_4 = \frac{1}{3},\ c_5 = -1$.
The test statistics of this test is as follows:

$$S = \frac{|\bar{x}_B - \bar{x}_A|}{SE} \quad (12)$$

$$\text{where } SE = \sqrt{MSE(\frac{1}{n_A} + \frac{1}{n_B})}, \quad (13)$$

MSE is the average square error of ANOVA, and the critical value of this test is as follows:

$$S_\alpha = \sqrt{(k-1)F_{\alpha,k-1,N-n}} \quad (14)$$

where *k* is the number of observed samples, *N* the overall size of data, and $F_{\alpha,k-1,N-n}$ the table value of Fisher distribution. When $S > S_\alpha$, the null hypothesis is rejected. Therefore, the Scheffé test is mostly used when post hoc (unplanned) comparisons of arithmetic means combinations are performed.

It's interesting that the results of this test, shown in the table below, coincide with the results obtained by the Tukey and Bonferroni test.

Table 7. Results of the Scheffé test

(I)	(J)	Mean Difference (I-J)	Std. Error	Sig.	95% C.I. Lower Bound	95% C.I. Upper Bound
1	2	-.03636	.05067	.972	-.1937	.120
	3	-.05425	.04816	.866	-.2038	.0953
	4	-.10736	.04794	.289	-.2562	.0415
	5	-.12698	.06166	.377	-.3184	.0645
2	3	-.01790	.02327	.964	-.0902	.054
	4	-.07100	.02281	.049	-.1418	-.0002
	5	-.09062	.04500	.401	-.2303	.0491
3	4	-.05310	.01649	.037	-.1043	-.0019
	5	-.07273	.04214	.563	-.2036	.0581
4	5	-.01962	.04189	.994	-.1497	.1105

Source: Authors` calculations

Therefore, the results of the tests conducted on the same data are different, but are in accordance with their theoretical background. The goal of this paper is to definitely highlight the importance of selecting an adequate test which complies with the research goal. The LSD test found the biggest number of pairs, which differ in ratings based on hotel category. The result obtained is expected, because it is known that this test makes the biggest mistake. If the t-test of independent samples is used for rating arithmetic means for all pairs of different categories, the same results would be obtained as with the LSD test. This test placed one-, two-, and three-star hotels into one group, and four- and five-star hotels into another. Moreover, besides the LSD, Newman-Keuls test is considered to be in a more liberal category group of tests. This test classified the average hotel rating values into two groups. Although these two tests are the most liberal ones in the group of the presented tests, they gave the best results for the data analyzed. The difference between average hotel rating values comes from the great difference between one- and five-star hotel ratings. Additionally, it is evident that there is no difference between one-, two-, and three-star hotels, as well as between four- and five-star hotel ratings.

The sample analyzed here contains 240 hotels, out of which 6 hotels are one-star, 33 are two-star, 89 have three stars, 104 have four stars, and 8 hotels have five stars. Given the above, it is obvious that subsamples do not contain a similar number of hotels observed, and the reason behind this may come from getting an illogical result of the Tukey test. Obviously, the presented results of the Bonferroni test match the Tukey test results. Although the Bonferroni test is the most conservative one when compared to Tukey and Scheffé tests, and in such situations it might be that the application of the Scheffé test might lead to different results. However, this was not the case for this research data, even though Scheffé test is more appropriate than Tukey test when group (subsample) sizes are different. It is clear that it is not appropriate to use the Dunnett test in this type of research, so the results obtained make no sense.

4. Conclusion

The research which covered 240 hotels in the Republic of Serbia and 71,700 ratings from hotel guests showed that there are significant differences in average hotel rating values based on hotel category, even though differences were not expected, because it was assumed that service would be aligned with certain categorization. The obtained results may be explained by the fact that guests who use low-category services have either higher expectations or give a subjective rating, ignoring the hotel category, or by the fact that certain hotels may, in time, start deviating from the standard after they get hotel

categorization certificate. Again, some may gradually fulfill a higher category conditions but after a while keep doing business within the previously allotted category.

In this paper, the most commonly used post hoc tests of one-way analysis of variance are briefly presented, discussed and applied using actual research data. Their advantages and disadvantages have been clearly shown, with an intention to enable easier decision-making for researchers in the process of choosing the most adequate test in certain research conditions. Numerous modern statistical software packages offering these tests are available, which greatly facilitates certain research stages, yet their successful application still requires a good knowledge of the essential differences between them.

The usual concern with practical usage of post hoc analyses is to identify either any relevant pattern or relationship that exist between formed entities within sampled data. These might well be unnoticed if analyses were limited just to a priori tests and methods. The methods referred to as post hoc, or a posteriori tests, are powerful statistical tools that often gain better results and conclusions of a research. They often enable a researcher to either additionally verify or to disqualify relations which were assumed to exist between subgroups of sampled populations. Unfortunately, this way of result checking is not too common, so published papers often miss preventative a posteriori control of the type I error rate.

High effectiveness of post hoc methods in detection of false positives makes them especially valuable in testing of multivariate hypotheses. By reducing the chances of getting wrong conclusions, these tests ultimately enable researchers to more precisely formulate their a priori hypotheses, make conclusions with greater confidence and get better final results.

Finally, it is important to highlight that an adequate post hoc test should be chosen based on the sample and the research goal. There is no ready-made, best or most frequently used test for this. Thus, the choice should primarily depend on the data structure and the questions to be addressed.

References

Baka,V. (2016). The becoming of user-generated reviews: Looking at the past to understand the future of managing reputation in the travel sector. Tourism Management, Vol. 53, 148-162.

Barjaktarović, L., Mašić, S. (2014).Investments and Investment Expectations in Serbian Hotel Industry. Possibilities and Perspectives for foreign Direct Investments in the Republic of Serbia, IIPE, Belgrade,Thematic Proceedings, 536-557.

Benjamini, Y., Hochberg, Y. (1995).Controlling the false discovery rate: a new and powerful approach to multiple testing. JRSS, series B, Methodological 57: 289–300.

Blair, R. C. (1981). A reaction to Consequences of failure to meet assumptions underlying the fixed effects analysis of variance and covariance. Review of Educational Research, 51.

Čačić, K., Mašić, S. (2013). Uticaj portala Tripadvisor na poslovanje hotela u Srbiji. Marketing, Vol 44, No. 3, 211-220.

Davis, C. S. (2002). Statistical Methods for the Analysis of Repeated Measurements. Springer, New York.

Frane, A. (2015). Are per-family Type I error rates relevant in social and behavioral science. Journal of Modern Applied Statistical Methods,14.

Field, A., Hole, G. (2002). How to design and report experiments. Sage.

Hochberg, Y. (1988). A Sharper Bonferroni Procedure for Multiple Tests of Significance. Biometrika. 75.

Keppel, G., & Wickens, T.D. (2004). Design and analysis: A researchers handbook (4rd Edition). Upper Saddle River, NJ: Pearson.

Klockars, A.J.,Hancock, G.R., & McAweeney, M.J. (1995). Power of unweighted and weighted versions of simultaneous and sequential multiple-comparison procedures. Psychological Bulletin, 118.

Kim, W.G., Limb, H., Brymer, R.A. (2015). The effectiveness of managing social media on hotel performance. International Journal of Hospitality Management, Vol. 44, 165–171.

Kirk, R. (1995). Experimental Design: Procedures For The Behavioral Sciences (3 ed.). Pacific Grove, CA.

Knežević, M., Barjaktarović, D., Obradović, P. (2014). Ocenjivanje kvaliteta hotelskih usluga putem Interneta. Međunarodna naučna konferencija SINTEZA 2014, Beograd, 767-771

Landan, S., Everitt, B. (2004). A handbook of statistical analysis using SPSS. Vol.1. FL:Chapman&Hall/CRC.

Mašić, S., Konjikušić, S. (2017). CONSUMERS' PERCEPTIONS OF SERBIA'S HOTEL PRODUCT QUALITY. *TISC - Tourism International Scientific Conference Vrnjačka Banja*, *2*(1), 59-76.

Ministry of Trade, Tourism and Telecommunications of the Republic of Serbia - Categorized accommodation objects, http://mtt.gov.rs/sektori/ sektor-za-turizam/korisne-informacije-turisticki-promet-srbija-kategorizacija/ (February, 2019)

Montgomery, Douglas C. (2001). Design and Analysis of Experiments (5th ed.). New York: Wiley.

Moore, David S.; McCabe, George P. (2003). Introduction to the Practice of Statistics (4th ed.). W H Freeman & Co.

Murphy, H.C., Chen, M., Cossutta M. (2016). An investigation of multiple devices and information sources used in the hotel booking process. Tourism Management, Vol. 52, 44-51.

Noble, S. W. (2009). How does multple testing correction work? Nature Biotechnology 27.

Radojević, T., Stanišić, N., Stanić, N. (2017). Inside the Rating Scores - A Multilevel Analysis of the Factors Influencing Customer Satisfaction in the Hotel Industry. Cornell Hospitality Quarterly (First published online January 9, 2017).

Rumsey, D. (2009). Statistics II for dummies. John Wiley&Sons.
Seaman, M.A., Levin, J.R., & Serlin, R.C. (1991). New developments in pairwise multiple comparisons: Some powerful and practicable procedures. Psychological Bulletin*, 110*.
Singer, D., J., Willett, B. J. (2003). Applied Longitudinal Data Analyses – Modeling Change and Event Occurrence. Oxford University Press.
Westfall,P.,Tobias,R, Wolfinger,R. (2011). Multiple comparisons and multiple testing using SAS. 2nd end, SAS Institute.
Živković, R., Njeguš, A., Gajić, J., Brdar, I., Mijajlović, I. (2015). Upravljanje onlajn zajednicama u hotelijerstvu. International Scientific Conference SITCON 2015, Belgrade, 133-139.