PERFORMANCE OF FDR TEST: A SIMULATION STUDY*

Aybüke Koca[1], İbrahim Kılıç[2], İsmet Doğan[3], Sinan Saraçlı[4]

[1]Afyon Kocatepe Üniversitesi, Fen Bilimleri Enstitüsü, İstatistik ABD, Afyonkarahisar, Türkiye
[2]Afyon Kocatepe Üniversitesi, Veteriner Fakültesi, Biyoistatistik ABD, Afyonkarahisar, Türkiye
[3]Afyon Kocatepe Üniversitesi, Tıp Fakültesi, Biyoistatistik ABD, Afyonkarahisar, Türkiye
[4]Afyon Kocatepe Üniversitesi, Fen-Edebiyat Fakültesi, İstatistik Bölümü, Afyonkarahisar, Türkiye
e-posta ssaracli@aku.edu.tr

ABSTRACT

Multiple comparison and multiple confidence tests are used to determine which group or groups are significantly different than others in analysis of variance. According to some assumptions, there are different kinds of multiple comparison and multiple confidence tests in the literature. In this study performance of the FDR (False Discovery Rate) test, which is one of the alternative test for the multiple comparison tests are examined. The significance value according to FDR is compared with the significance value for the t test. 3, 5 and 10 groups for the sample sizes 50, 100 and 200 are compared. MATLAB software is used to obtain the related data sets and to conclude the analysis. All of the results show that FDR is nonsensitive to the number of compared groups. The increase of FDR values is constant for numbers of compared groups even if the sample size increases. Other results are given in related tables.

Keywords: FDR, Multiple Comparison, Analysis of Variance

1. Introduction

As it is known, one of the statistical method to compare the differences between more than two groups is ANOVA; Analysis of Variance. But to apply Analysis of variance there are some assumptions to be met (Winer, 1971). These assumptions are some parametric factors like homogeneity, normality and additive (Ferguson, 1981).

ANOVA which is a parametric test has a quadratic form by an additive $\left(\sum_{i=1}^{nj}(X_{ij}-\bar{X})^2\right).\left(\sum_{j=1}^{k}\sum_{i=1}^{nj}(X_{ij}-\bar{X})^2\right)$ point (Ferguson, 1981). This test finds out whether there is a significant difference or not but it doesn't find out which group or groups cause this difference (Kayri, 2009).

If there is significant difference between groups, the statistics which determines this difference is called as Post-hoc (Köklü et all, 2006; Roscoe, 1975). There are kinds of post-hoc tests and some assumptions to use the most suitable one in a research (Kayri, 2009).

The term "Multiple Comparisons" refers to making several tests for statistical significance of differences between means (or proportions or variances, etc.) within a group. Statistical procedures that are designed to take into account and properly control for the multiplicity effect through some combined or joint measure of erroneous inferences are called multiple comparison procedures (MCPs). It is a fundamental problem of practical importance. They can be conducted in different ways. The following four types of multiple comparison procedures are seen in the literature based on the objective of the researcher: (Rao and Swarupchand, 2009).

(i) MCA (all-pairwise multiple comparisons) considers $\mu_i - \mu_j$ for all $i \neq j$ to be of primary interest.
(ii) MCB (multiple comparisons with the best) considers $\mu_i - \max \mu_j$ for i =1,..., k to be of primary interest.
(iii) MCC (multiple comparisons with a control) considers $\mu_i - \mu_k$ for i = 1,..., k-1 to be of primary interest.

[*] *This study is a part of Aybüke Koca's MS thesis supervised by İbrahim Kılıç at Afyon Kocatepe University Institute of Science, 2013.*

(iv) MCM (multiple comparisons with the mean) considers $\mu_i - \bar{\mu} \; or \; \mu_i - \bar{\bar{\mu}}$ for all i = 1,..., k to be of primary interest, where $\bar{\mu}$ and $\bar{\bar{\mu}}$ are the unweighted and the weighted means of the $\mu_i's$.

Except the MCA all other three types (MCB, MCC, and MCM) of multiple comparisons comes under the category many-to-one comparisons. The foundation of the subject of multiple comparisons was laid in the late 1940s and early 1950s, principally by David Duncan, S.N.Roy, R. C. Bose, Henry Scheffe and John W.Tukey, although some of the ideas appeared much earlier in the works of Fisher, Student, and others (Rao and Swarupchand, 2009).

An alternative technique to make multiple comparisons according to results of ANOVA test can be considered as FDR (False Discovery Rate), to compare the groups, whether there is a statistically significant difference or not. This technique is recently used in many studies and in this study the performance of this technique is tried to put forward via a simulation study.

2. Method: FDR (False Discovery Rate)

Consider the problem of testing simultaneously m (null) hypotheses, of which m_0 are true. R is the number of hypotheses rejected. Table 1 summarizes the situation of a traditional form. The specific m hypotheses are assumed to be known in advance. R is an observable random variable; U, V, S and T are unobservable random variables. If each individual null hypotheses are tested separately at level α then R=R(α) is increasing in α. Lower case letters are used for their realized values (Benjamini and Hochberg, 1995).

Table 1. Number of errors committed when testing *m* null hypotheses.

	Declared Non-significant	Declared significant	*Total*
True null hypotheses	U	V	m_0
Non-true null hypotheses	T	S	m-m_0
	m-R	R	m

In terms of these random variables, the PCER is $E(V/m)$ and the FWER is $P(V \geq 1)$. Testing individually each hypothesis at level α/m guarantees that $P(V \geq 1) \leq \alpha$.

The proportion of errors committed by falsely rejecting null hypotheses can be viewed through the random variable $Q = V/(V+S)$ –the proportion of rejected null hypotheses which are erroneously rejected. Naturally, we define Q=0 when V+S=0, as no error of false rejection can be committed. Q is an unobserved (unknown) random variable, as we do not know v or s, and thus $q = v/(v+s)$, even after experimentation and data analysis. FDR Q_e is defined to be the expectation of Q,

$$Q_e = E(Q) = E\{V/(V+S)\} = E(V/R)$$

Two properties of this error rate are easily shown, yet are very important (Benjamini and Hochberg, 1995).

a. If all null hypotheses are true, the FDR is equivalent to the FWER: in this case s=0 and v =0 then Q=0, and if $v > 0$ then q=1 leading to $P(V \geq 1) = E(Q) = Q_e$. Therefore control of the FDR implies control of the FWER in the weak sense.
b. When $m_0 < m$ the FDR is smaller than or equal to the FWER: in this case, if $v > 0$ then $v/r \leq 1$, leading to $X_{(V\geq 1)} \geq Q$. Taking expectations on both sides we obtain $P(V \geq 1) \geq Q_e$ and the two can be quite different. As a result, any procedure that controls the FWER also controls the FDR. However, if a procedure controls the FDR only, it can be less stringent, and a gain in power may be expected. In particular, the larger the number of the non-true null hypotheses is, the larger S tends to be, and so is the difference between the error rates. As a result, the potential for increase in power is larger when more of the hypotheses are non-true.

In large-scale hypotheses generating studies such as microarray experiments, the FDR seems more relevant than the Family Wise Error Rate (FWER) defined by the probability of committing at least one false discovery (Hochberg and Tamhane, 1987).

FWER controlling procedures are often too conservative unless we increase the desired probability of making at least one false positive decision. However, when we search for induced genes among thousands we can allow false positives if this leads to a higher number of true positives and a better understanding of biological processes. The proportion of false positives among all positives is called False Discovery Rate (FDR). A Bonferroni-like method to control this rate was first introduced by Benjamini and Hochberg (1995) (Scheid and Spang, 2003).

3. Application

The performance of FDR, which is one of the alternatives for multiple comparison tests, is examined via a simulation study in this section. The significance values of tests statistics of FDR and t test are compared under the restrictions of three groups and the sample sizes of 50, 100 and 200. MATLAB software is used to obtain the related data sets and complete the comparison study and α level is set to 0,05 (Koca, 2013).

Table 2 shows the results of three group comparison for sample size 50. In this table groups are defined as A, B and C. For these groups the related data sets which are set to normal distribution and having different means are obtained by MATLAB software. Similarly the comparison results of FDR and t tests for sample sizes 100 and 200 are given in Table 3 and Table 4 respectively.

When there are five groups to compare, it's known that the number of all pairwise comparisons is ten. If we consider the groups as A, B, C, D and E, the results of FDR and p values for all of the possible pairwise comparisons for these groups with sample sizes 50, 100 and 200 are given in Table 5, Table 6 and Table 7 respectively.

Finally when there are ten groups called as A,B,C,D,E,F,G,H,I and J, results of FDR and p values for forty-five possible pairwise comparisons for the sample sizes 50, 100 and 200 are given in Table 8, Table 9 and Table 10 respectively. The p values of accepted hypotheses are given bold in all tables.

Table 2. Comparison results of FDR and p values for three groups and sample size n=50.

Groups Compared	p	FDR
A-B	0,024055	0,016667
A-C	**0,051984**	0,033333
B-C	**0,463199**	0,05

Table 3. Comparison results of FDR and p values for three groups and sample size n=100.

Groups Compared	p	FDR
A-B	0,002569	0,016667
A-C	0,010028	0,033333
B-C	**0,364527**	0,05

Table 4. Comparison results of FDR and p values for three groups and sample size n=200.

Groups Compared	p	FDR
A-B	0,0000114206098	0,016667
A-C	0,000250609	0,033333
B-C	**0,26414843**	0,05

If we examine Table 2, Table 3 and Table 4 given above, it can be seen from Table 2 that where the null hypotheses for one of the pairwise comparisons are rejected, two are accepted both for FDR and p values. According to Table 3 and Table 4 only one is accepted and the p values are decreasing for the increasing sample sizes.

Table 5. Comparison results of FDR and p values for five groups and sample size n=50.

Groups Compared	p	FDR
A-B	0,000034079	0,005
A-C	0,000309056	0,01
A-D	0,002947086	0,015
A-E	0,009709377	0,02
B-C	0,026519613	0,025
B-D	**0,053547710**	0,03
B-E	**0,129254107**	0,035
C-D	**0,232721538**	0,04
C-E	**0,433698404**	0,045
D-E	**0,714021009**	0,05

Table 6. Comparison results of FDR and p values for five groups and sample size n=100.

Groups Compared	p	FDR
A-B	0,0000000033	0,005
A-C	0,0000007187	0,01
A-D	0,0001276159	0,015
A-E	0,0006027161	0,02
B-C	0,0023604704	0,025
B-D	0,0086800038	0,03
B-E	0,0311203124	0,035
C-D	**0,0868255801**	0,04
C-E	**0,2639783271**	0,045
D-E	**0,5737905454**	0,05

Table 7. Comparison results of FDR and p values for five groups and sample size n=200.

Groups Compared	p	FDR
A-B	0,000000000000000022	0,005
A-C	0,000000000001040140	0,01
A-D	0,000000096212694750	0,015
A-E	0,000002088307544477	0,02
B-C	0,000027914260987232	0,025
B-D	0,000306740919718883	0,03
B-E	0,002911819373859250	0,035
C-D	0,017417007841139800	0,04
C-E	**0,089546481681295500**	0,045
D-E	**0,317129146898809000**	0,05

According to the results of ten possible pairwise comparisons for five groups, it can be seen that if the sample size is 50, where the fife of p values are lower than the α value (which means that the null hypotheses are rejected) the other fife are higher. By the increasing sample sizes, the rejected null hypothesis again increasing and now it's more clear than three group comparisons that for sample size

100 seven of ten and for sample size 200 eight of ten null hypothesis are rejected according to p values. It can also easily be seen that FDR values are not affected by the increasing sample sizes.

Table 8. Comparison results of FDR and p values for ten groups and sample size n=50.

Groups Compared	p	FDR
A-B	0,000000000000000000	0,001111
A-C	0,000000000000001240	0,002222
A-D	0,000000000000376480	0,003333
A-E	0,000000000044204308	0,004444
A-F	0,000000000344664272	0,005556
A-G	0,000000001815543751	0,006667
A-H	0,000000004833152341	0,007778
A-I	0,000000016825851582	0,008889
A-J	0,000000040869917662	0,01
B-C	0,000000103043573415	0,011111
B-D	0,000000270348236957	0,012222
B-E	0,000000778189080659	0,013333
B-F	0,000001762146871507	0,014444
B-G	0,000003708703263271	0,015556
B-H	0,000008013586227067	0,016667
B-I	0,000017752445402116	0,017778
B-J	0,000036269669303618	0,018889
C-D	0,000075086434022479	0,02
C-E	0,000140593684139160	0,021111
C-F	0,000262173352136893	0,022222
C-G	0,000444552272420111	0,023333
C-H	0,000793525133591124	0,024444
C-I	0,001363556428131020	0,025556
C-J	0,002245103594305900	0,026667
D-E	0,003633551149332790	0,027778
D-F	0,005842341723940550	0,028889
D-G	0,009171276566376540	0,03
D-H	0,013966350011112400	0,031111
D-I	0,020133176143261600	0,032222
D-J	0,028964275642826400	0,033333
E-F	0,040024797833452000	0,034444
E-G	**0,054198045658671900**	0,035556
E-H	**0,072537047198545100**	0,036667
E-I	**0,094539555660992700**	0,037778
E-J	**0,121978738596565000**	0,038889
F-G	**0,156239821648749000**	0,04
F-H	**0,195166446608266000**	0,041111
F-I	**0,244947756432680000**	0,042222
F-J	**0,305193495512790000**	0,043333
G-H	**0,370418992588437000**	0,044444
G-I	**0,450649244469801000**	0,045556
G-J	**0,542946009703164000**	0,046667
H-I	**0,647736264875323000**	0,047778
H-J	**0,759587797346495000**	0,048889
I-J	**0,883219859196012000**	0,05

Table 9. Comparison results of FDR and p values for ten groups and sample size n=100.

Groups Compared	p	FDR
A-B	0,0000000000000000000000000	0,001111
A-C	0,0000000000000000000000000	0,002222
A-D	0,0000000000000000000000000	0,003333
A-E	0,0000000000000000000000000	0,004444
A-F	0,0000000000000000009992007	0,005556
A-G	0,0000000000000000193178806	0,006667
A-H	0,0000000000000003652633752	0,007778
A-I	0,0000000000000022924995236	0,008889
A-J	0,0000000000000188135063098	0,01
B-C	0,0000000000001535711557939	0,011111
B-D	0,0000000000014621097665923	0,012222
B-E	0,0000000000077998787428868	0,013333
B-F	0,0000000000237520274559699	0,014444
B-G	0,0000000001043824651159040	0,015556
B-H	0,0000000003921834973441560	0,016667
B-I	0,0000000018187403394431800	0,017778
B-J	0,0000000109608980246589000	0,018889
C-D	0,0000000265060956192720000	0,02
C-E	0,0000000930611820035843000	0,021111
C-F	0,0000002857406433793840000	0,022222
C-G	0,0000009414849929322290000	0,023333
C-H	0,0000032865506035822900000	0,024444
C-I	0,0000109784351513302000000	0,025556
C-J	0,0000303153649629655000000	0,026667
D-E	0,0000696833015214867000000	0,027778
D-F	0,0001510747976395890000000	0,028889
D-G	0,0003479508671506970000000	0,03
D-H	0,0007527678385681180000000	0,031111
D-I	0,0014565421847827400000000	0,032222
D-J	0,0026145052203776200000000	0,033333
E-F	0,0045691768392574000000000	0,034444
E-G	0,0077106077204198800000000	0,035556
E-H	0,0123206492233679000000000	0,036667
E-I	0,0187592571301449000000000	0,037778
E-J	0,0289516092430585000000000	0,038889
F-G	0,0432595198034065000000000	0,04
F-H	**0,0628709259068479000000000**	0,041111
F-I	**0,0908828572543007000000000**	0,042222
F-J	**0,1278447198176540000000000**	0,043333
G-H	**0,1771375295143160000000000**	0,044444
G-I	**0,2525097363417940000000000**	0,045556
G-J	**0,3507193189482940000000000**	0,046667
H-I	**0,4724441968684370000000000**	0,047778
H-J	**0,6333766754534350000000000**	0,048889
I-J	**0,8143350242746520000000000**	0,05

Table 10. Comparison results of FDR and p values for ten groups and sample size n=200.

Groups Compared	p	FDR
A-B	0,0000000000000000000000000	0,001111
A-C	0,0000000000000000000000000	0,002222
A-D	0,0000000000000000000000000	0,003333
A-E	0,0000000000000000000000000	0,004444
A-F	0,0000000000000000000000000	0,005556
A-G	0,0000000000000000000000000	0,006667
A-H	0,0000000000000000000000000	0,007778
A-I	0,0000000000000000000000000	0,008889
A-J	0,0000000000000000000000000	0,01
B-C	0,0000000000000000000000000	0,011111
B-D	0,0000000000000000000000000	0,012222
B-E	0,0000000000000000000000000	0,013333
B-F	0,0000000000000000000000000	0,014444
B-G	0,0000000000000000000000000	0,015556
B-H	0,0000000000000000002220446	0,016667
B-I	0,0000000000000000124344979	0,017778
B-J	0,0000000000000001794120408	0,018889
C-D	0,0000000000000500761654365	0,02
C-E	0,0000000000001165127994085	0,021111
C-F	0,0000000000039925138750618	0,022222
C-G	0,0000000000280903342808614	0,023333
C-H	0,0000000002352973988983820	0,024444
C-I	0,0000000019056126101535400	0,025556
C-J	0,0000000099129453947632800	0,026667
D-E	0,0000000619835252937406000	0,027778
D-F	0,0000002226417140480660000	0,028889
D-G	0,0000007486843286350630000	0,03
D-H	0,0000027270789359268300000	0,031111
D-I	0,0000097220594839388900000	0,032222
D-J	0,0000337464786794413000000	0,033333
E-F	0,0000877645436575134000000	0,034444
E-G	0,0002149456803242940000000	0,035556
E-H	0,0004796718525772590000000	0,036667
E-I	0,0010091998396619000000000	0,037778
E-J	0,0021863971447260500000000	0,038889
F-G	0,0047072595860724700000000	0,04
F-H	0,0084841536240105000000000	0,041111
F-I	0,0163538642442905000000000	0,042222
F-J	0,0289740982818153000000000	0,043333
G-H	0,0490194299411033000000000	0,044444
G-I	**0,0828179428948482000000000**	0,045556
G-J	**0,1445185201491050000000000**	0,046667
H-I	**0,2411097454516110000000000**	0,047778
H-J	**0,4199678196323640000000000**	0,048889
I-J	**0,6828134153085580000000000**	0,05

The results of forty-five possible pairwise comparisons for ten groups shows that all of the p values are lower according to early group comparisons and the percentage of the rejected null hypothesis increase according to p values related with the sample size even if the FDR values are same. For sample size 50 given in Table 8, there are fourteen, for sample size 100, given in Table 9 there are 9 and for sample size 200 given in Table 10, there are 5 hypotheses accepted whereas the others are rejected.

4. Results and Conclusion

Multiple comparison tests are used to determine the significant differences between groups after the null hypothesis is rejected according to results of ANOVA. One of the disadvantages of ANOVA is: it uses a constant α level to compare the mean differences for each group. To avoid from this situation there

are some studies in the literature. It’s very important to determine the common error ratio correctly to obtain reliable results. None of the other classical techniques consider about false discover ratio. By this study we wanted to emphasis on FDR and tried to show the performance of it via a simulation study.
According to the simulation tests the results can be summarized as below:

- Sample size is important on the numbers of rejected hypothesis according to p values.
- By increasing multiple comparison number, the number of accepted null hypothesis is decreasing.
- The p values of the rejected null hypothesis are decreasing by the number of compared groups. It is also decreasing related with the increasing sample sizes.
- If the number of compared groups is same, the p value is decreasing by the increased sample sizes.
- If both the number of compared group and sample size increase, whereas the FDR values are same, the p values are decreasing.

All of the results show that FDR is nonsensitive to the number of compared groups. The increase of FDR values is constant for numbers of compared groups even if the sample size increases. The performance of FDR test is always better than multiple comparison test because it considers the proportion of false positives among all positives.

References

Benjamini, Y., Hochberg, Y., 1995. Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing. Journal of Royal Statistical Society B Series, 57(1), 289-300.

Ferguson, G. A. (1981). Statistical analysis in psychology and education. New York: McGraw-Hill Book Company.

Hochberg,Y. and Tamhane,A. (1987) Multiple Comparison Procedures. Wiley.

Kayri M., (2009), “The Multiple Comparison (Post-Hoc) Techniques to Determine the Difference Between Groups in Researches” Fırat University Journal of Social Science Vol 19, No: 1, pp: 51-64.

Koca, A. (2013), Multiple Comparison Tests in Variance Analysis and a Simulation Application on FDR Test, Master of Science Thesis, Afyon Kocatepe University, Institute of Science, Afyonkarahisar.

Köklü, N., Büyüköztürk Ş. And Bökeoğlu, Ç.Ö. (2006). Sosyal bilimler için istatistik. Ankara: PegemA Yayıncılık.

Rao, C.V. and Swarupchand, U. (2009), Multiple Comparison Procedures - a Note and a Bibliography, Journal of Statistics, Volume 16, pp.66-109.

Roscoe, J. T. (1975). Fundemental research statistics for the behavioral sciences. New York: Holt, Rinehart and Winston, Inc.

Scheid S., Spang R.,(2003) “A False Discovery Rate Approach to Separate the Score Distributions of Induced and Non-induced Genes”, Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003) March 20–22, Vienna, Austria

Winer, B. J. (1971). Statistical principles in experimental design. New York: McGraw-Hill Book Company.