A COMPARISON OF THE FALSE DISCOVERY RATE WITH DUNNETT'S TEST FOR A LARGE NUMBER OF TREATMENTS

Kayeromi D. Gomez[1], Curt Doetkott[2], and Rhonda C. Magel[3*]

[1,2,3]Department of Statistics North Dakota State University Fargo, ND 58108

*Corresponding Author: -

Abstract:

Probabilities of two types of errors are estimated for the False Discovery Rate Procedure (FDR), Dunnett's test, and conducting individual t-tests when treatments are being compared against a control. Probabilities for two types of powers are also estimated. The number of treatments considered is 50, 100, 150, and 300 with the first treatment being the control. Differing numbers of treatments equal to the control are considered with equal sample sizes of 3 and 5.

Keywords: *ANOVA, multiple comparisons, family wise error rate, false discovery rate*

I.

1. INTRODUCTION

In recent years, there has been a revolution of massive amounts of data in many areas including genomics, astrophysics, plant sciences, social sciences, and others. Researchers may wish to compare several treatments at a time in order to determine which treatments differ. In this case, several may be a few hundred treatments. There are two distinct scenarios that researchers may use to test if there are differences among several treatments; the multiple comparison scenario which is a two-stage process, and the multiple testing scenario which conducts several simultaneously individual tests. In the multiple comparison scenario, we first conduct an overall test to see if there are any treatment differences. If the overall test is rejected, a post-hoc multiple comparison procedure is used to determine where these differences lie. Some of the common post-hoc procedures include Sheffe's, Fisher's Least Significant Difference, and Tukey's [1]. Dunnett's test is a common multiple comparison testing procedure used when comparing all the treatments with a single control [1]. In this case, the total number of comparisons is reduced since not all treatments are being compared against each other. As an example, suppose that 6 treatments are being compared. In this case, 15 pairwise comparisons would need to be made, but if there were 5 treatments being compared against 1 control, only 5 comparisons would need to be made. Multiple comparison procedures were designed to control for the family-wise error rate (FWER) which is defined as the probability of at least one Type I error when testing for differences among the entire family of treatments [1].

When doing multiple testing of simultaneous hypotheses, one could also control the FWER. Suppose we are testing 50 null hypotheses using a Bonferroni type of control of the FWER [1]. The probability of finding at least one significant result due to chance = 1-P (no significant results)^{50} = 1 - (1-0.05/50)^{50} = 0.0488. Meaning there is a 4.88 % chance of rejecting at least one null hypothesis. The 0.05 here is a per-comparison kind of type I error rate. The more tests we perform, the more difficult it becomes to detect significance in at least one test as the adjusted significance level of alpha gets extremely small ($\alpha *= \alpha /m$, where m is the total number of null hypotheses being tested). On the other hand if we are testing 50 hypotheses simultaneously at a significance level of $\alpha = 0.05$ where no multiple comparison post-hoc FWER control is specified, the probability of observing significance in at least one test due to chance will be P (at least one significant result) = 1- P (no significant results)^{50} = 1- (1-0.05)^{50} = 0.9231. Hence in 50 tests, we have 92.31% chance of finding at least one significant result even if none of the tests is actually significant. With a large number of tests, this probability approaches 1.

The problem with controlling the FWER is that if the number of hypotheses to be tested becomes large (or as the number of treatments to be compared becomes large), the probability of rejecting a null hypothesis will become very small even when the null hypothesis is incorrect. In other words, the power is low and one may rarely determine a difference in two treatments even when there is one. As an alternative to controlling the FWER, Benjamini and Hochberg [2] introduced the False Discovery Rate Controlling Procedure (FDR). This is defined to be "the expected ratio of erroneous rejections to the number of rejected hypotheses" [2]. One of the areas in which the FDR method has been used extensively is in genomics research where thousands of genes are often simultaneously being compared and their relationship to particular diseases are investigated. Scientists confronted by this problem of large number of comparisons often find the FDR to be an appealing quantification of error [3,4].

Since the introduction of the False Discovery Rate Controlling Procedure (FDR), various researchers have suggested possible changes to it. This includes the introduction of the Benjamini Hochberg procedure [5], and work done by [6,7,8,9,10,11]. In particular, Efron [10] conducted research on correlation and the effect that it had on the FDR. Gordon, Glazko Qiu, and Yakovlev [12] did some comparison of the Benjamini Hochberg procedure associated with the FDR with the FWER using Bonferroni. Additional work needs to be conducted to more fully understand the error rate being controlled for in the FDR [3]. Research conducted in this paper was an attempt to do this in a limited setting.

2. Proposed Research

The research conducted in this study was designed to compare results of using Dunnett's test with the results of using the FDR method. Dunnett's test [1, 13, 14] is a well-known multiple comparison test designed to control FWER. It is used after an ANOVA is performed and the null hypothesis that all treatments are the same is rejected. Dunnett's test then compares the control (1st treatment) with each of the other treatments to determine which of the other treatments differs significantly with the control [15]. In order to compare the FDR method on an equal basis with Dunnett's test, the FDR method was also used only to compare each treatment with the control (1st treatment).

A simulation study was designed with the number of treatments being 50,100, 150 and then 300. The variance of each treatment population was set to be the same throughout all of the treatments. Equal sample sizes of 3 and then 5 were used for each treatment. A variety of situations for each sample size and number of treatments were considered. For each situation, 10,000 samples were simulated. Tests were conducted on each of the samples and it was determined for each test whether or not the null hypothesis was rejected. SAS version 9.3 [16] was used in simulating the samples and conducting the tests. The FDR method was conducted by using PROC MULTTEST in SAS [17, 18, 19].

This simulation study used only two different means. The first mean was μ = 1763.18 (control mean) and the second mean was μ = 2641.1 (this was the mean of treatments if they were different from the control). The standard deviation, σ, was equal to 438.96. The underlying distributions were normal. These selected values for the means and standard deviation were used because they came from a study comparing the mean seed yield of different genotypes of edible dry beans conducted by Kubat [20].

When 50 treatments were considered with the first treatment being the control, the following settings were simulated: 1:49; 10:40; 25:25; 40:10; and 49:1. The first number given is the number of treatments with the low mean including the control and the second number given is the number of treatments with the high mean. In the case 1:49, only the control

has the low mean and the other 49 treatments are all different than the control with the high mean. In the case 10:40, 9 treatments have the same mean as the control and 40 treatments have a higher mean than the control. It is also noted that the means are set up so that the higher mean is 2 standard deviations above the first mean $\mu = 2641.1$ ($1763.18 + 2 \times 438.96$).

When 100 treatments were considered with the first treatment being the control, the following settings were simulated: 1:99; 25:75; 50:50; 75:25, and 99:1. For 150 treatments, the following settings were simulated: 1:149; 25:125; 50;100; 75:75; 100:50; 125:25; and 149:1. In the case of 300 treatments, the following settings were simulated: 1:299; 50:250; 100:200; 150:150; 200:100; 250:50; and 299:1.

In addition to having the underlying distributions be all normal distributions, we also conducted simulations in which the underlying distributions were all contaminated normal distributions. In this case, all of the distributions were 90% normal with a standard deviation of 438.96. The remaining 10% was normal with the mean the same mean as in the first population, but the standard deviation was twice the standard deviation of the original population. The same settings were simulated as before with the two means.

Four measures were calculated for each set of the 10,000 samples in order to allow us to compare the FDR method with that of Dunnett. In addition to the FDR method and Dunnett, we also calculated each of these measures with respect to conducting individual t-tests for each case (this was referred to as RAW in this study in the tables). The first measure calculated was the Relative Power percentage of all of the tests in each setting. In calculating the Relative Power percentage (RP), we counted the total number of rejections when comparing the treatment groups to the control group when the null hypothesis is false and divided this by the total number of comparisons between each treatment and the control in which the null hypothesis was false. This was multiplied by 100 to get the percentage. The formula for calculated RP is given in (1)

$$RP = \frac{total\ count\ of\ all\ rejections\ when\ Ho\ is\ false}{total\ number\ of\ comparions\ when\ Ho\ is\ false} \times 100 \tag{1}$$

If we consider the first scenario of 50 simulated treatments (1:49), we count the total number of rejections made divided by a total of 490,000 comparisons (49 comparisons multiplied by 10,000 simulations). This is done for each of the three methods.

The next measure that was calculated for each method was the Actual Power percentage. In calculating the Actual Power percentage (AP), we counted the number of times that a test had at least one rejection in a sample when the null hypotheses was false divided by the total number of simulations and then multiplied by 100 to get the percentage. This would give us some idea of the ability of each of the tests to find at least one difference when there was at least one difference using the information in a sample. The AP is given by equation (2).

$$AP = \frac{total\ counts\ of\ at\ least\ one\ rejection\ when\ Ho\ is\ false\ per\ simulated\ sample}{total\ number\ of\ simulations\ (10,000)} \times 100 \tag{2}$$

The Relative Type I Error Rate (RE) was calculated for each method. In this case, we look at every sample that was simulated and every comparison made within a sample and counted the number of times the null hypothesis was rejected when it was true divided by the number of times the null hypothesis was true and multiplied by 100. For example in the setting 25:25, 24 samples are simulated each time that are from the same population as the control and this is done 10,000 times. Hence the number of true null hypothesis in this case would be 240,000 (24 X 10,000). The value in the denominator when calculating the RE (in this setting) would then be 240,000 in the equation given in (3)

$$RE = \frac{total\ counts\ of\ rejections\ when\ Ho\ is\ true}{total\ number\ of\ true\ null\ hypotheses} \times 100 \tag{3}$$

We calculated the Actual Type 1 error rate percentage for each of the methods in the study. This is also known as the Family Wise Error Rate (FWER). The Actual Type 1 error rate (AE) or FWER was estimated by counting the number of times of getting at least one false rejection in a simulated sample divided by the number of simulations and then multiplied by 100 to get the percentage. The AE (or FWER) is given by equation (4).

$$AE = \frac{total\ counts\ of\ at\ least\ one\ false\ rejection\ per\ simulated\ sample}{total\ number\ of\ simulations\ (10,000)} \times 100 \tag{4}$$

3. Results

The results for 50 treatments all having a sample size of 3 are given in Table 1 for each of the four measures. Recall that in the setting 1:49, there was one control with a mean of 1763.18 and the remaining 49 treatments had a mean of 2641.1. All of the treatment populations and the control population had the same standard deviation of 438.96. It is noted that the RE for the individual t-tests (Raw) stays at around 5% which is what we had set alpha equal to for the individual tests. The RE of Dunnett's test stays well below 5% since we are controlling for FWER and there are several treatments. The RE for the FDR methods lies between the RE of the other two methods. The AE or FWER for the individual t-tests does get large and is at 44.63% for the setting 25:25. The FWER stays approximately at 5% or below for the Dunnett's test since this is what is controlled for. The FWER does get higher than 5% for the FDR Method, but it does stay quite a bit lower than for the individual t-tests, being 14.69% compared to 44.63% in the setting 25:25. The RP is also higher for the FDR Method compared with Dunnett's test except for the 49:1 setting. It is noted that the RP stays the same for both the individual t-tests and Dunnett's test because all of the comparisons being done are at the same level of alpha with alpha being approximately 0.20% for Dunnett and 5% for the individual t-tests for all the settings. The alpha level changes with each setting for the FDR method ranging from approximately 0 .53% to 2.26%

since it is the false discovery rate being controlled for which changes with the differing number of

Table 1: Summary of Rejections Percentages in case of 50 treatments with sample size 3

SETTINGS	TESTS	Type I error		Power	
		Relative (RE)	Actual (AE/FWER)	Relative (RP)	Actual (AP)
1:49	Raw	—	—	68.15	99.58
	Dunnett's test			23.48	85.05
	FDR Method			56.13	83.93
10:40	Raw	5.05	26.93	68.15	99.44
	Dunnett's test	0.20	1.56	23.61	83.76
	FDR Method	2.26	12.22	52.68	82.11
25:25	Raw	4.98	44.63	67.93	98.92
	Dunnett's test	0.18	2.92	23.41	78.51
	FDR Method	1.54	14.69	44.49	75.82
40:10	Raw	5.02	55.79	67.63	96.67
	Dunnett's test	0.20	4.35	23.27	64.88
	FDR Method	0.96	10.02	32.46	59.82
49:1	Raw	4.98	59.35	67.56	67.56
	Dunnett's test	0.19	5.06	23.01	23.01
	FDR Method	0.53	4.37	18.56	18.56

Table 2: Summary of Rejections Percentages in case of 300 treatments with sample size 3

SETTINGS	TESTS	Type I error		Power	
		Relative (RE)	Actual (AE/FWER)	Relative (RP)	Actual (AP)
1:299	Raw			68.54	99.97
	Dunnett's test	—	—	13.55	88.76
	FDR Method			56.49	86.36
50:250	Raw	5.08	61.78	68.25	99.98
	Dunnett's test	0.05	1.49	13.76	87.14
	FDR Method	2.17	27.40	52.60	84.27
100:200	Raw	5.00	76.67	68.53	99.99
	Dunnett's test	0.04	2.41	13.74	86.71
	FDR Method	1.77	30.21	49.04	83.04
150:150	Raw	5.03	84.61	69.09	99.90
	Dunnett's test	0.05	2.90	14.03	84.95
	FDR Method	1.50	28.68	45.03	80.45
200:100	Raw	5.19	90.12	68.63	99.82
	Dunnett's test	0.05	4.39	14.00	80.64
	FDR Method	1.12	24.31	38.39	75.32
250:50	Raw	4.84	92.99	68.85	99.71
	Dunnett's test	0.04	4.24	13.54	73.94
	FDR Method	0.65	16.68	28.69	65.98
299:1	Raw	4.95	94.98	68.94	68.94
	Dunnett's test	0.04	5.00	13.33	13.33
	FDR Method	0.35	3.13	9.13	9.13

treatments equal to the control. The relative power of the FDR changes with higher relative power when there are a larger number treatments that are different from the control. The AP gives the percentage of times that each of the methods picked up at least one treatment different than the control for each sample. Dunnett and FDR method have about the same AP with Dunnett being just a little higher. In the setting 1:49 the RE and FWER are both zero for all of the methods since all of the treatment populations differ from the control. Table 2 given the values of the four measures for each of the methods based on equal sample sizes of 3 when there are 300 different treatment populations with one of these being the control. The same patterns are seen in this case.

Tables 3 and 4 give the RE, AE, RP, and AP for all of the methods in the case of 50 and then 300 treatments with equal sample sizes of 5. Again, similar patterns are seen. Table 5 and 6 give the results for the contaminated normal distribution with equal sample sizes of 3 and number of treatments being 50 and 300, respectively. Comparing the results in Table 1 for normal distributions to those in Table 5 with contaminated normal distributions, one can see that the error percentages are about the same in both cases, but the powers are lower in the contaminated normal case. Similar results can be seen when comparing the results in Table 3 (normal distributions) with the results in Table 6 (contaminated normal distributions).

Table 3: Summary of Rejections Percentages in case of 50 treatments with sample size 5

SETTINGS	TESTS	Type I error		Power	
		Relative (RE)	Actual (AE/FWER)	Relative (RP)	Actual (AP)
1:49	Raw	——	——	88.40	100.00
	Dunnett's test			50.16	98.04
	FDR Method			84.42	97.82
10:40	Raw	4.97	26.51	88.35	99.99
	Dunnett's test	0.19	1.29	49.93	97.31
	FDR Method	2.79	16.10	82.18	97.42
25:25	Raw	4.85	44.97	88.52	99.96
	Dunnett's test	0.16	2.52	50.41	95.86
	FDR Method	1.78	19.65	76.88	95.58
40:10	Raw	4.77	55.80	88.41	99.84
	Dunnett's test	0.15	3.65	50.32	90.78
	FDR Method	0.87	12.00	64.93	89.50
49:1	Raw	5.06	60.20	88.15	88.15
	Dunnett's test	0.18	4.73	50.84	50.84
	FDR Method	0.51	4.72	44.48	44.48

Table 4: Summary of Rejections Percentages in case of 300 treatments with sample size 5

SETTINGS	TESTS	Type I error		Power	
		Relative (RE)	Actual (AE/FWER)	Relative (RP)	Actual (AP)
1:299	Raw	——	——	88.68	100.00
	Dunnett's test			35.75	98.43
	FDR Method			84.83	98.00
50:250	Raw	4.80	61.45 1.40	88.80	100.00
	Dunnett's test	0.04	45.06	35.58	98.50
	FDR Method	2.60		82.97	98.09
100:200	Raw	4.80	78.06 2.31	88.62	100.00
	Dunnett's test	0.04	52.37	35.65	98.22
	FDR Method	2.16		80.13	97.73
150:150	Raw	5.03	85.36 3.21	88.39	100.00
	Dunnett's test	0.04	48.49	35.27	97.53
	FDR Method	1.59		76.20	96.98
200:100	Raw	5.01	90.21 1.06	88.59	100.00
	Dunnett's test	0.04	37.25	35.58	96.69
	FDR Method	1.13		71.25	95.82
250:50	Raw	4.95	93.08	88.63	99.99
	Dunnett's test	0.04	4.35	35.53	94.25
	FDR Method	0.68	22.86	61.93	92.23
299:1	Raw	4.93	94.92	88.79	88.79
	Dunnett's test	0.04	4.98	35.74	35.74
	FDR Method	0.29	3.46	27.45	27.45

Table 5: Summary of Rejections Percentages in case of 50 treatments with sample size 3 in Contaminated Normal

SETTINGS	TESTS	Type I error		Power	
		Relative (RE)	Actual (AE/FWER)	Relative (RP)	Actual (AP)
1:49	Raw	___	___	56.82	98.74
	Dunnett's test			15.24	76.31
	FDR Method			41.56	73.85
10:40	Raw	5.19	27.82 1.97	57.01	98.52
	Dunnett's test	0.26	11.20	15.35	74.01
	FDR Method	2.15		38.39	71.39
25:25	Raw	5.26	48.16 4.19	57.45	97.61
	Dunnett's test	0.28	14.97	15.87	67.66
	FDR Method	1.68		32.06	64.24
40:10	Raw	5.30	59.89 6.30	57.14	93.52
	Dunnett's test	0.29	11.27	15.55	52.33
	FDR Method	1.08		21.81	47.20
49:1	Raw	5.08	64.57	58.23	58.23
	Dunnett's test	0.28	6.82	16.18	16.18
	FDR Method	0.68	5.81	12.53	12.53

Table 6: Summary of Rejections Percentages in case of 300 treatments with sample size 3 in Contaminated Normal

SETTINGS	TESTS	Type I error		Power	
		Relative (RE)	Actual (AE/FWER)	Relative (RP)	Actual (AP)
1:299	Raw			58.14 8.13	99.95
	Dunnett's test	___	___	42.40	86.30
	FDR Method				82.29
50:250	Raw	5.21	65.74 2.97	57.76 8.07	99.96
	Dunnett's test	0.11	26.08	38.65	84.03
	FDR Method	2.13			79.55
100:200	Raw	5.12	83.07 5.26	57.54 8.12	99.90
	Dunnett's test	0.09	31.05	35.04	81.50
	FDR Method	1.74			76.29
150:150	Raw	5.06	90.88 6.79	57.71 8.11	99.83
	Dunnett's test	0.11	30.50	30.41	78.89
	FDR Method	1.53			72.85
200:100	Raw	5.11	94.93	57.21 7.86	99.68
	Dunnett's test	0.10	8.74	24.52	72.78
	FDR Method	1.18	26.01		65.72
250:50	Raw	5.19	97.13 9.62	57.25 7.85	99.03
	Dunnett's test	0.09	19.42	17.37	62.14
	FDR Method	0.90			54.21
299:1	Raw	5.16	98.26	**57.77**	**57.77**
	Dunnett's test	0.10	11.41	**8.08**	**8.08**
	FDR Method	0.53	7.35	**5.79**	**5.79**

4. Conclusion

In the end, the FDR method is a compromise between using multiple t-tests and using Dunnett's test. The choice of an appropriate test will depend on the objectives at stake in an individual research project. When type I error is detrimental to the outcomes of the research, Dunnett's test offers a better alternative. Though the power of Dunnett's test decreases as the number of treatments increases, its Family Wise Error Rate is maintained at a constant level across all treatments. This option might be of value in a marketing situation in which a type I error could cause the loss of a considerable amount of money and the power of the test is not as important as severely limiting the probability of a type I error.

However, when the objective of the researcher is to detect as many of the differences as possible, and a type I error although not desired is not as detrimental, then the FDR method is preferred. The probability of a type I error increases as

the number of treatments gets higher using the FDR method, but not as much as using multiple t-tests. The strength in the FDR method is that it can find considerably more actual differences than Dunnett's test.

When there are a total of 300 treatments with 49 treatments that are the same as the control and 250 treatments different from the control and we have equal samples of size 3, Dunnett's test will generally not find any of the treatments that are the same as the control as being significantly different (49 X 0.0004 = 0.02 from Table 2). However, Dunnett's test will only find approximately 34 of the 250 treatments that are different from the control as being significantly different (250 X 0.1376 = 34.3). The FDR method will find approximately 1 of the 49 treatments the same as the control as being significantly different 49 X .0217 = 1.06). However, the FDR method will find approximately 132 of the 250 treatments different from the control as being significantly different (250 X 0.5260=131.5). Hence, by using the FDR method over Dunnett's test there will be 98 additional differences found where there are differences (over Dunnett's test) and only 1 extra difference marked as significantly different where no actual difference exists.

When half of the treatments are different from the control, the FDR method finds approximately 68 of the 150 (150 X 0.4503=67.54) treatments that are different from the control as being significantly different from the control while finding approximately 2 treatments that are the same as the control to be significantly different from the control (149 X 0.015=2.24). In this situation, Dunnett's test is only able to identify 21 out of the 150 differences (150 X 0.1403=21.05) that are significantly different from the control as being significantly different from control; while finding at most 1 treatment (0.0005 X 149=0.075) that is the same as the control as significantly different from the control. Hence by using the FDR method, over Dunnett's test, there will be 47 additional differences found where there are differences and only 2 extra differences marked as significantly different where no actual difference exists.

5. References

[1].Montgomery, Douglas C. (1991). Design and Analysis of Experiments, Third Edition New York: John Wiley & Sons.

[2].Benjamini Y., and Hochberg Y. (1995). Controlling the false discovery rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological), 57*(1), 289-300.

[3].Benjamini, Yoav. (2010). Discovering the false discovery rate. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 72*(4), 405-416. doi: 10.1111/j.1467-9868.2010. 00746.x

[4].Osborne, Jason A. (2006). Estimating the False Discovery Rate using SAS. *SAS Users Group International Proceedings, 190*, 1-10

[5].Benjamini Y., and Yekutieli D. (2001). The Control of the False Discovery Rate in Multiple Testing under Dependency. *The Annals of Statistics, 29*(4), 1165-1188.

[6].Yekutieli, D. and Benjamini, Y. (1999). Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. *Journal of Statistical Planning and Inference, 82,* 171-196.

[7].Benjamini Y., and Hochberg Y. (2000). On the Adaptive Control of the False Discovery Rate in Multiple Testing with Independent Statistics. *Journal of Educational and Behavioral Statistics, 25*(1 (Spring, 2000)), 60-83.

[8].Storey, J.D. (2001). The positive False Discovery Rate: A Bayesian interpretation and the q-value. *Annals of Statistics*

[9].Storey, J.D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society B, 64,* 479-498.

[10]. Efron, Bradley. (2007). Size, Power and False Discovery Rates. *The Annals of Statistics, 35*(4), 1351-1377. doi: 10.1214/009053606000001460

[11]. Benjamini, Yoav (2010). Discovering the false discovery rate. *Journal of the Royal Statistical Society: Series B, 72*(4), 405-416.

[12]. Gordon, Alexander, Glazko, Galina, Qiu, Xing, and Yakovlev, Andrei (2007). Contol of the mean number of false discoveries, Bonferroni and stability of multiple testing. *Annals of Applied Statistics*, 1(1), 179190.

[13]. Dunnett, Charles W. (1955). A Multiple Comparison Procedure for Comparing Several Treatments with a Control. *Journal of the American Statistical Association, 50*(272), 1096-1121.

[14]. Dunnett, C. W. (1964). New Tables for Multiple Comparisons with a Control. *Biometrics, 20*(3), 482-491.

[15]. Hari Mukerjee, Tim Robertson and F. T. Wright. (1987). Comparison of Several Treatments with a Control Using Multiple Contrasts. *Journal of the American Statistical Association, 82*(399), 902-910.

[16]. SAS Institute Inc., (2011). SAS/STAT® 9.3 User's Guide. Cary, NC: SAS Institute Inc.

[17]. Westfall, P. H. and Soper, K. A. (1994). *Nonstandard Uses of PROC MULTTEST: Permutational Peto Tests; Permutational and Unconditional and Binomial Tests.* Paper presented at the Proceedings of the Nineteenth Annual SAS Users Group International Conference, Cary, NC.

[18]. Westfall, P. H. and Wolfinger, R. D. (2000). Closed Multiple Testing Procedures and PROC MULTTEST. *Observations*.

[19]. Westfall, P. H., Wolfinger, R. D. and Tobias, Randall D. (2011). *Multiple Comparisons and Multiple Tests Using SAS* (Second ed.): SAS Institute.

[20]. Kubat, Jamie Marie. (2013). *Comparing Dunnett's test with the false discovery rate method: A simulation study.* (Masters), North Dakota State University.