A SIMULATION COMPARISON OF TWO METHODS FOR CONTROLLING THE EXPERIMENT-WISE TYPE I ERROR RATE OF CORRELATED TESTS FOR CONTRASTS IN ONE-WAY COMPLETELY RANDOMIZED DESIGNS

by

YUANFANG JIAO

B.S., Tianjin University, 2007
Ph.D., Kansas State University, 2012

A REPORT

submitted in partial fulfillment of the requirements for the degree

MASTER OF SCIENCE

Department of Statistics
College of Arts and Sciences

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2014

Approved by:

Major Professor
Dr. Paul I. Nelson

Abstract

A Bonferroni and an ordered P-value solution to the problem of controlling the experiment-wise Type I error rate are studied and compared in terms of actual size and power when carrying out correlated tests. Although both of these solutions can be used in a wide variety of settings, here they are only investigated in the context of multiple testing that specified pairwise comparisons of means, selected before data are collected, are all equal to zero in a completely randomized, balanced, one factor design where the data are independent random samples from normal distributions all having the same variance. Simulations indicate that both methods are very similar and effective in controlling experiment wise type error at a nominal rate of 0.05. Because the ordered P-value method has, almost uniformly, slightly greater power, it is my recommendation for use in the setting of this report.

Table of Contents

List of Figures

List of Tables

Acknowledgements

I would like to thank my advisor Dr. Paul Nelson for his guidance, instruction, and support during my graduate study. I learnt so much from him not only how to solve statistic problems but also how to think critically as a statistician. His rigorous study attitude and endless enthusiasm to science make a profound impression to me and motivate me to persist in future study.

I also thank all my committee members Dr. Weixin Yao and Dr. Christopher Vahl for their valuable advices and suggestions. I really appreciate for their time and efforts.

Finally, I am grateful for my family for all their love, support and guidance through my whole life.

Dedication

To my parents

Chapter 1 - Introduction

1.1 Research Motivation and Background

Hypothesis testing is a statistical method used to make inferences based on analyzing data collected from designed and observational studies. In testing a single hypothesis, we usually specify a maximum acceptable probability of rejecting the null hypothesis when it is true, which is defined as a Type I error. Comparing procedures for controlling the overall (Experiment-Wise) type I error rate when testing more than one hypothesis using the same data, the topic of this report, is an important and widely studied topic.

When multiple hypothesis testing is conducted, even though the Type I error rate of each test is controlled at some small desired level, the probability of making at least one Type I error can increase at a dramatic rate as the number of hypotheses being tested grows. For example, in genomics, scientists need to measure expression levels of tens of thousands of genes and genotypes for millions of genetic markers. An inflated experiment-wise Type I error rate may have very serious consequences for researchers by causing them to waste time and resources following false leads. This issue was raised a long time ago, and numerous statisticians have been working on it for decades since it occurs in a wide range of applied fields. Many methods have been proposed to deal with this problem. But so far, no general solution has been found to be best or even good for all situations.

In this study, we compare a Bonferroni and an ordered P-value method, described below, for their ability to approximately control the experiment wise Type I error rate associated with simultaneously testing multiple contrasts, which are linear combinations of two or more treatment means whose coefficients add up to zero, in a one way, completely randomized design (*CRD*). Specifically we will test that a series of pairwise comparisons of means, special contrasts selected before the data are examined, are all equal to zero. Using a one way *CRD*, instead of multifactorial study will make it easier to carry out the simulation. The criteria used here for comparing multiple testing methods are; (i) Closeness of actual experiment-wise Type I error rate to the traditional value of 0.05; (ii) Power when at least one contrast differs from zero. Note that I will not consider the validity of tests for the individual contrasts in comparing the two methods.

Although Tukey's *Honest Significant Difference* (*HSD*) provides an exact experiment-wise size α test in this to the problem of multiple testing in balanced designs, I do not investigate the *HSD* here since it has been widely studied and does not generalize beyond this limited setting.

1.2 Control of Experiment-Wise Type I Error Rate

The Experiment-Wise Type I error rate, also called the Family-Wise Type I error rate, is defined as the probability of at least one Type I error if all m null hypotheses $\{H_0(i);\ i=1,2,...,m\}$ being tested are true. Specifically, it is desired to control the Type I error rate of $H_0 = \bigcap_{i=1}^{m} H_0(i)$. Suppose the m hypotheses $\{H_0(i);\ i=1,2,...,m\}$ are tested individually at corresponding Type I error rates $\{\alpha_i\}$, and each decision is based on independent criteria. If all the hypotheses are true, the Experiment Wise Type I error (α_{EW}) can be expressed as

$$\alpha_{EW} = 1 - P(fail\ to\ reject\ H_i, i = 1,2,\ldots,m|H_0) = 1 - \prod_{i=1}^{m}(1-\alpha_i). \qquad (1.1)$$

When all the tests are carried out using the same type I error rate, denoted α_C , (1.1) becomes $\alpha_{EW} = 1-(1-\alpha_C)^m$. However, independence is not a common occurrence in multiple testing problems.

Without assuming independence, the well-known Boole's inequality implies that

$$\alpha_{EW} = P(\bigcup_{i=1}^{m}\{Reject\ H_0(i)\}|H_0) \leq \sum_{i=1}^{m} P(Reject\ H_0(i)|H_0) = \sum_{i=1}^{m}\alpha_i = m\alpha_C \quad (1.2)$$

1.3 Ordered P-values

For a single hypothesis test, given data, a *p-value* is the smallest type I error rate that would lead to rejection of the null hypothesis of interest. In many cases, a p-value may be equivalently defined as the probability under the null hypothesis of observing a test statistic as least as extreme as the value actually observed. In multiple testing of m hypothesis tests, the methods we explore for controlling the Type I error rate of $H_0 = \bigcap_{i=1}^{m} H_{0\,(i)}$ can be expressed as procedures for adjusting p-values computed for the individual hypotheses An ordered p-value is a p-value modified, usually by making it larger, to take multiple testing into account. Specifically, let $\{p_i\}$ be the original, unordered p-values for the null hypotheses $H_0(i)$, i = 1, 2,

…, m, that are uniformly distributed under their respective null hypotheses. Then, a decision to reject $H_0(i)$, a single inference, will be reached if $p_i \leq \alpha_i$, where $\{\alpha_i\}$ are the desired Type I error rates., i = 1,2,…,m. In a multiple testing setting, letting $\{\tilde{p}_i\}$ denote the corresponding ordered p-values defined by a particular method, a decision to reject H_0 is made if at least for one index i, $\tilde{p}_i \leq \alpha_i$, i = 1,2,…,m.

Another concept related to α_{EW} is the False Discovery rate (*FDR*) {SAS/STAT 9.2 User's Guide}, denoted here by Q and defined by Q = V/R, where Q is the proportion of false rejection, V equals the number of false rejections of true hypotheses and R equals the total number of rejections. The control of FDR is another important statistical method used in multiple-hypothesis testing {Benjamini, Y. and Yekutieli, D. (2001)}. From the definition above, one can tell that *FDR* aims to control the proportion of false discoveries (incorrectly rejected null hypotheses), which is less conservative and can result in more power compared to the Experiment Wise Type I error rate methods {Efron, B (2010)}.

1.4 Bonferroni Method

The upper bound in (1.2) can be much larger than α_{EW}. Consequently, using it to choose $\{\alpha_i\}$ in order to obtain a desired value of α_{EW} by equating the right sides of (1.1) can lead to procedures that have very small Type I error rates and correspondingly low power. For example, if it is desired that α_{EW} be at most α and that the individual rates $\{\alpha_i\}$ all equal a value, denoted α_C, the Bonferroni bound guarantees that taking

$$\alpha_C = \alpha/m \qquad (1.3)$$

provides an approximate solution to the problem of calibrating α_C. The corresponding ordered p-values for the Bonferroni method are then given by

$$\{\tilde{p}_i = mp_i\} \qquad (1.4)$$

For example, for α = 0.05 and m = 10 using (1.2) yields α_C = 0.005, a very small value that could result in low power for testing H_0. If the Bonferroni bound is approximately correct, as happens under independence, this method of calibrating α_C is close to the best that can be done without additional assumptions {Hommel, G. (1988)}. From (1.3), it then follows that under joint independence, unless some $\{\alpha_i\}$ are 'large', using the first order Taylor expansions$\{e^{-\alpha_i} \cong$

$1-\alpha_i\}$, so that using (1.1), $\alpha_{EW}=1-\prod_{i=1}^{m}(1-\alpha_i)\cong 1-e^{-\sum_{i=1}^{m}\alpha_i}\cong\sum_{i=1}^{m}\alpha_i$. Hence, under joint independence of the individual P-values$\{p_i\}$, rejecting H_0 if $p_i\le\alpha/m$ for at least one $i, i=1,2,\ldots,m,$ results in a test with approximate Type I error rate α.

1.5 Independent or Correlated Tests

In sum, using the Bonferroni upper bound to calibrate α_C as given in (1.3) is approximately correct under independence. However, in many practical applications, the statistics from which the individual ordered P-values $\{p_i\}$ are obtained are correlated and an assumption of independence among the ordered P-values $\{p_i\}$ is correspondingly questionable, even as a rough approximation.

A study of ordered *p-values* for multiple correlated tests is presented by Conneely and Boehnke in 2007 {Conneely, K. and Boehnke, M. (2007)}. In the application of genome-wide association (GWA) studies, scientists need to test association among hundreds of thousands of genetic variants; while hundreds of the tests are correlated with one another. Conneely and Boehnke proposed a new method to compute ordered P-values for correlated tests (P_{ACT}), which can attain the power of permutation tests but require much less computation time as well.

Efron {Efron, B. (2007)} did a lot of research on multiple testing and carried out large scale simulation studies. . He presented both computational and theoretical methods for large-scale testing which related to the correct choice of the null distribution and demonstrated those using massive data sets like breast cancer and HIV studies as well as other popular microarray analysis techniques.

1.6 An Ordered P-value Method

Simes in 1986 {Simes, R.J. (1986)} proposed an alternate to the Bonferroni solution to the calibration problem given in (1.2), given by: Reject H_0 at nominal experiment wise Type I Error Rate α if

$$p_{(i)}\le i*\alpha/m \text{ for any } i=1,2,\ldots,m \qquad (1.4)$$

where $\{p_{(i)}\}$ are the ordered p-values. The adjusted ordered p-values for this method are given by

$$\tilde{p}_{(i)} = mp_{(i)} / i \ .$$

Simes {Simes, R.J. (1986)} proved that (1.4) provides an exact size α test under independence and showed by simulation that it performs better than (1.3) in terms of yielding an actual α_{EW} closer to α and having higher power in some cases when $\{p_i\}$ are not independent. However, the scenarios he simulated do not appear to relate to any practical problems and may be only of academic interest.

Note that although there are many other methods for controlling experiment wise Type I error rate, my report only examines the two methods given above, Bonferroni and the adjusted ordered p-value of Simes {Simes, R.J. (1986)}. Also, note that $\tilde{p}_{(1)} \le \alpha$ if and only if {$\tilde{p}_i \le \alpha$; for at least one i}. Hence, if the Bonferroni method leads to rejecting H_0 at nominal experiment wise Type I error rate α , so will the Simes method. The converse is not true, raising the possibility that, in at least some cases the, Simes method is more powerful than the Bonferroni method while still maintaining a desired experiment wise type 1 rate. My report describes some of the other methods available in proc multitest of SAS, including those based on bootstrap and permutation procedures. Some other SAS procedures like GLM, MIXED can also adjust results for multiple tests as well.

1.7 Research Questions of My Report

(1) How well do the Bonferroni and Simes methods perform in terms of controlling experiment wise Type I error rate?

(2) How much more powerful is the Simes method than the Bonferroni method?

(3) Are the gains in power attained by the Simes method big enough to compensate for possible inflation of the experiment wise Type I error rate?

1.8 Example from Kuehl (2000)

We present an example from Kuehl {Kuehl, R. (2000)} to illustrate the two multiple testing procedures being studied in this report for the multiple testing problem of the pairwise comparison of means in a one way, completely randomized design. In an animal physiology study about the pituitary function of hens, it was observed that hens will come back into

production after being forced to molt. The response associated with pituitary function is a measurement of compound (serum T3) concentration from hens. Physiologists measure compounds from hens at the premolt stage prior to the forced molt regimen and at the end of each different stages of the forced molt regimen. The five stages of the regimen were (1) premolt (control), (2) fasting for 8 days, (3) 60 grams of bran per day for 10 days, (4) 80 grams of the bran per day for 10 days, and (5) laying mash for 42 days, which were taken as five treatment of this study. There were 25 hens used for this study and 5 hens were used for each treatment.
Clearly, the design structure here is one-way CRD with 5 treatments; each treatment with 5 observations.

The ANOVA Table is shown as below in Table 1.1.

Table 1.1 ANOVA Table of Serum Example from Kuehl (2000)

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	48568.87638	12142.21909	78.08	<.0001
Error	20	3110.18912	155.50946		
Corrected Total	24	51679.06550			

R-Square	Coeff Var	Root MSE	serum Mean
0.939817	10.32231	12.47034	120.8096

If interest lies in making all m = 10 pairwise comparisons at type I error rates 0.05 each, and the unordered p-values given below were actually independent, the Experiment Wise Type I error $\alpha_{EW} = 1-(1-0.05)^{10} = 1- 0.5987 = 0.4013$, a large value. Table 1.2 presents these unordered p-values reported by SAS. Entries that are at most 0.05, indicating statistical significance, are highlighted in red.

Table 1.2 Pairwise Comparison of Unordered P-value

i/j	1	2	3	4	5
1		0.0031	<.0001	0.0059	0.8415
2	0.0031		<.0001	0.7872	0.0050
3	<.0001	<.0001		<.0001	<.0001
4	0.0059	0.7872	<.0001		0.0092
5	0.8415	0.0050	<.0001	0.0092	

Table 1.3 presents the SAS output of pairwise comparisons using the Bonferroni method ordered p-values given by $\tilde{p}_i = 10p_i$. With nominal α_{EW} = 0.05, the entries in red indicate statistical significance.

Table 1.3 Pairwise Comparison using Bonferroni Method Ordered P-value

i/j	1	2	3	4	5
1		0.0314	<.0001	0.0586	1.0000
2	0.0314		<.0001	1.0000	0.0499
3	<.0001	<.0001		<.0001	<.0001
4	0.0586	1.0000	<.0001		0.0924
5	1.0000	0.0499	<.0001	0.0924	

Actually, as described in Kuehl, the researchers are interested in four contrasts among the ten pairwise comparisons. These null hypotheses can be expressed as:

$H_0 = \{H_i, i = 1, 2, 3, 4\}$, while H_1: $\mu_1 = \mu_2$, H_2: $\mu_2 = \mu_3$, H_3: $\mu_3 = \mu_4$, H_4: $\mu_4 = \mu_5$.

The following are SAS output of the four contrasts.

Table 1.4 Four Contrasts of Serum Example from Kuehl (2000)

Contrast	DF	Contrast SS	Mean Square	F Value	Pr > F
premolt vs fasting	1	1752.71121	1752.71121	11.27	0.0031
fasting vs 60 grams	1	21764.09104	21764.09104	139.95	<.0001
60 grams vs 80 grams	1	22782.48361	22782.48361	146.50	<.0001
80 grams vs laying	1	1290.72321	1290.72321	8.30	0.0092

If using unordered p-values, with m = 4, we reject each H_i at α_i = 0.05, then, assuming independence, the α_{EW} = 1-(1-0.05)^4 = 1-0.8145 = 0.1855. Note that since p_1 = 0.0031< 0.05, p_2 < 0.0001 < 0.05, p_3 < 0.0001 < 0.05, p_4 = 0.00924 < 0.05, so that at least one comparison is statistically significant at type I error rate 0.05. But, using this procedure to reject H_0 provides no bound on the experiment wise error rate.

Using the Bonferroni ordered P-values, $\tilde{p}_1$ = 0.00314*4 < 0.05, $\tilde{p}_2$ = 0.0001*4 < 0.05, $\tilde{p}_3$ = 0.0001*4 < 0.05, $\tilde{p}_1$ = 0.00924*4 < 0.05, so that the same conclusion is reached but now with $\alpha_{EW} \leq 0.05$.

Using the ordered p-value method of Simes (1986), we have $p_3 < p_2 < p_1 < p_4$ and letting $p_3 = p_{(1)}$, $p_2 = p_{(2)}$, $p_1 = p_{(3)}$, $p_4 = p_{(4)}$, reject H_0 if $p_{(i)} \leq i*\alpha/m$ (i = 1,2,3,4; α = 0.05; m = 4). Hence, $p_3 = p_{(1)}$ < 0.0001 <1*0.05/4, $p_2 = p_{(2)}$ < 0.0001 < 2*0.05/4, $p_1 = p_{(3)}$ = 0.0031 < 3*0.05/4, $p_4 = p_{(4)}$ = 0.0092 < 4*0.05/4, also leading to the decision ‘ Reject H_0.'

Chapter 2 - Simulation Experiment

Let $\{y_{ij}; i=1,2,...,k;\ j=1,2,...,n\}$ be realizations of independent random variables $Y_{ij} \sim N(\mu_i, \sigma^2)$ and let $\{c_l = \mathbf{c}_l'\boldsymbol{\mu}, l = 1,2,...m\}$ be *m* contrasts of interests specified before the data are collected. The goal here is to design an experiment that specifies representative values *of k, n, m*, $\{\mu_i\}, \{\mathbf{c}_l\}$ and *t*, where *t* of the contrasts are zero and *m-t* are not. Without loss of generality take σ = 1.0 and set the target type I error rate α = .05. In this setting, the null hypotheses are given by $\{H_l : \mathbf{c}_l'\boldsymbol{\mu} = 0\}$. Each will be tested against a two sided alternative. Note that $\mu_1 = \mu_2 = \cdots = \mu_k$ if and only if $\{\mathbf{c}_l, l = 1,2,...,k-1\}$ are linearly independent and $H_0 = \bigcap H_0(i)$ holds. If the contrast coefficients $\{\mathbf{c}_l\}$ are orthogonal, and hence linearly independent, although the estimators $\{\hat{c}_l = \mathbf{c}_l'\bar{\mathbf{Y}} = \sum_{i=1}^{k} c_{li}\bar{Y}_i, l = 1,2,...m\}$ are independent, the p-values obtained from testing $\{H_0(i)\}$ are not independent. In this report I only consider pairwise comparisons (a class of contrasts) $\{\mu_r - \mu_s, r < s\}$, constructed as follows. For *k* = 4, 6, and 8, let *j* = *k*/2. Specify Δ > 0 and let $\mu_1 = \mu_2 = \cdots = \mu_j$ = 0 and $\mu_{j+1} = \mu_{j+2} = \cdots = \mu_k = \Delta$. Consider two cases.

<u>Case 1</u>

All m = j*(j-1)/2 +j*(j-1)/2 null hypotheses { $H_0(l)$ } consist of the true statements $\mu_r = \mu_s$ where (i) $1 \le r < s \le j$ and (ii) $j+1 \le r < s \le k$.

<u>Case 2</u>

t= j-1 of the hypotheses are the true statements of the form $\mu_r = \mu_{r+1}$, $1 \le r < j-1$ and the *m*-*t* = j*j false statements $\mu_r = \mu_s$, $1 \le r < j < s \le k$. Hence for Case 2, m = j-1 +j*j. Several values of Δ will be selected and representative values of sample sizes, *n*, used.

2.1 Simulation Procedure

For each selected scenario:

(i) Generate data from the assumed model.

(ii) Test $H_0 = \bigcap \{H_0(i) : \mathbf{c}_i'\boldsymbol{\mu} = 0, i = 1, 2, ..., m\}$ using (1.3) and (1.4) at the nominal value α_{EW} = 0.05.

(iii) Independently repeat (i) and (ii) N (= 5000) times.

(iv) Record the results by filling in a $N \times 2$ matrix score sheet as follows. Each row refers to one of the N replications. In column 1 of row *i* enter a '1' if (1.3) leads to the rejection of H_0 and a '0' otherwise. Do the same for column 2 of row i using (1.4), *i* = 1,2,…,N.

(v) Tally the proportion of ones using both methods as an estimate of the powers of the tests of H_0.

(vi) Carry out (i)-(iv) for all parameter settings, cases 1 and 2.

(vii) Display and summarize the results using graphs, plots, charts and statistical analyses. McNemar's test, described below, will be used to compare the Type I and Type II error rates of the two procedures.

(viii) Write a summary comparing the two results and make recommendations about using them.

Since each data set in my simulation experiment is a block to which both multiple testing procedures are applied, I used McNemar's test {McNemar, Q. (1947)} to test for the equality of the powers, denoted here simply by the proportions p_1 and p_2 , of the two multiple comparison tests under a particular parameter setting. McNemar's test is a normal approximation applied to count data in the form of the 2 by 2 contingency table given in Table 2, where a '+' sign indicates rejection of the null hypothesis that a contrast is zero. The symbols a, b, c, d denote the observed counts in N simulated data sets. Since the estimates of these proportions, $(a+b)/N$ and $(a+c)/N$, are correlated , this procedure is commonly referred to as a test for (more accurately called a test based on) correlated proportions.

Large values of The McNemar's Chi-square statistic $\chi^2 = \dfrac{(b-c)^2}{b+c}$ support rejection of the hypothesis of equal proportions. Its approximate critical value is obtained from a Chi-square distribution with degrees of freedom = 1.

Table 2.1 Example of McNemar's Test

	Test 2 (+)	Test 2 (-)	Row Total
Test 1 (+)	a	b	a + b
Test 1 (-)	c	d	c + d
Column Total	a + c	b + d	N

2.2 Simulation Results

The following simulation results were all carried out in R.

2.2.1 Assessing the Correlation between P-values

Since this report studies the effect of correlation on the performance of two multiple comparison procedures, we begin the discussion of my simulation experiment with a brief assessment of the correlation among the P-values it generated. I explored two simple types of data $\{y_{ij}; i = 1, 2, ..., k; j = 1, 2, ..., n\}$; k = 3 and 4 and in each case n = 5, 10, 20, 50 respectively and N = 5000 (the number of simulated data sets). Each y_{ij} follows a normal distribution N (Δ, 1) , where Δ can be zero or non-zero, and is randomly generated by a computer algorithm. First, pairwise comparisons $\{c_l = \mathbf{c}_l'\boldsymbol{\mu}, l = 1, 2, ...m\}$ were specified for each case with m = 2. For each case, we are interested here in the P-values obtained from two estimated contrasts: $\hat{c}_1 = \bar{y}_{1.} - \bar{y}_{2.}$ and $\hat{c}_2 = \bar{y}_{2.} - \bar{y}_{3.}$ when k = 3; $\hat{c}_1 = \bar{y}_{1.} - \bar{y}_{2.}$ and $\hat{c}_2 = \bar{y}_{3.} - \bar{y}_{4.}$ when k = 4. For case one, $y_{ij} \sim N(0, 1)$ with i = 1, 2, ..., k; for case two, $y_{ij} \sim N(1, 1)$ with i = 1 and $y_{ij} \sim N(0, 1)$ with i = 2, 3, …, k. Therefore, in case one, both hypotheses (c_1 = 0 and c_2 = 0) are true; in case two only one hypothesis (c_2 = 0) is true and the other hypotheses (c_1 = 0) is false. Note that although $\hat{c}_1$ and $\hat{c}_2$ are correlated for k = 3 but independent for k = 4, the P-values are correlated in both settings since both test statistics use the same estimate of experimental error.

For each case, we calculated the correlation coefficient between estimated P-value of the two contrasts based on the N = 5000 (See Appendix B for more details). The results are shown in the following tables.

Table 2.2 Correlation Coefficient of P-values from Two Contrasts for Case One

Case 1	n = 5	n = 10	n = 20	n = 50
k = 3	0.2161717	0.1729556	0.178407	0.1753097
k = 4	0.03983468	0.02929927	-0.004706192	0.0001726341

Table 2.3 Correlation Coefficient of P-values from Two Contrasts for Case Two

Case 2	n = 5	n = 10	n = 20	n = 50
k = 3	-0.01253577	-0.0648454	-0.1240184	-0.06299838
k = 4	0.03695898	0.03583926	-0.01424717	-0.006119949

As we can observe from the results above, estimated correlation coefficients of P-values obtained from these contrasts are only moderately large for k = 3 in case 1, small otherwise. Generally, correlation coefficients of P-values when k = 3 fluctuate in a small range as n increases, which means the number of replications in a given normal distribution does not have much influence on the linear association between these particular P-values, for both case one and case two.

On the other hand, for both case one and case two, the absolute values of the correlation coefficients decrease when k = 4, as the common sample size n increases. As expected, since the contrast estimates are independent for k = 4 but not for k = 3 in both cases, the correlations for k = 4 are less in absolute value than the corresponding correlations for k = 3 in case 1. Surprisingly, this is only true for n > 5 in Case 2. In addition, scatter plots of estimated p-values in Appendix B for Case 1 are consistent with the relatively low illustrate seen in Table 2.2.

In order to further investigate the extent to which p-values in my simulation experiment are correlated, we tried another set of scenarios for case one $\{y_{ij}; i = 1, 2, ..., k; j = 1, 2, ..., n\}$ with k = 4, 6 and 8; n = 5, 10, 20, 50; and N = 5000, where $y_{ij} \sim N$ (0, 1). In each scenario, all possible pairwise comparisons were tested. Specifically, the pairwise contrasts $\{c_l = \mathbf{c}_l' \boldsymbol{\mu}, l = 1, 2, ... m\}$ of my design were used for each case with m = 6, 15, 20 as k = 4, 6, 8. Estimated P-values from every contrast were obtained and the correlation coefficients between P-value of each pair of contrasts were entered into an $m \times m$ correlation matrix. Although the matrix itself is too big to show, the determinant of the matrix is one summary value that can still provide some valuable information about the overall magnitude of the correlation coefficients taken as a group. For example, for a 2 by 2 correlation matrix with correlation ‘r’, the determinant is $1-r^2$, which goes to zero as r increases. In general, the smaller the determinant of a correlation matrix is, the greater the dependence will be among the variables on which it is based. From Table 2.4 we see that our simulation results indicate that: (i) the determinants of the correlation coefficient matrices are moderately large for k = 4; (ii) decrease as k increases, indicating increasing over all linear dependence among the P-values; (iii) are small for k = 8; (iv) are fairly constant for each k as common sample size n increases. Keep in mind that unlike the settings presented in Tables 2.2 and 2.3, there are many contrasts here for each k, some of whose estimators are independent and some dependent.

Table 2.4 Determinant of Correlation Coefficient Matrix at N = 5000

Case 1	n = 5	n = 10	n = 20	n = 50
k = 4	0.6641073	0.6880515	0.7278202	0.7362667
k = 6	0.2034491	0.2159213	0.2313283	0.2476259
k = 8	0.02393239	0.02370257	0.0275811	0.0296115

2.2.2 Comparison of Type I Error and Power of the Bonferroni Method and Ordered P-value Method

For simulation of this part, we chose k = 4, 6, 8; n= 5, 10, 20, 50; Δ = 0.5, 1; N = 5000. For pairwise comparisons (a class of contrasts) $\{\mu_r - \mu_s, r < s\}$, we set $\mu_1 = \mu_2 = \cdots = \mu_j$ = 0 and $\mu_{j+1} = \mu_{j+2} = \cdots = \mu_k = \Delta$ with j = k/2. Now recall our two cases as defined before, repeated for the convenience of the reader.

<u>Case 1</u>

All m = j*(j-1)/2 +j*(j-1)/2 null hypotheses { $H_0(l)$ } consist of the true statements $\mu_r = \mu_s$ where (i) $1 \le r < s \le j$ and (ii) $j+1 \le r < s \le k$.

Case 2

t= j-1 of the hypotheses are the true statements of the form $\mu_r = \mu_{r+1}$, $1 \le r < j-1$and the *m-t* = j*j false statements $\mu_r = \mu_s$, $1 \le r < j < s \le k$. Hence for Case 2, m = j-1 +j*j.

The following tables show, at each value of k, the number of true and false null hypotheses in each case and summarize the simulation results for N=5000, where both multiple comparison procedures were carried out at a nominal 0.05 experiment wise Type I error rate..

Table 2.5 Number of True and False Null Hypotheses for Each Case

	Case 1		Case 2	
	True Hypotheses	False Hypotheses	True Hypotheses	False Hypotheses
k = 4	2	0	1	4
k = 6	6	0	2	9
k = 8	12	0	3	16

From Tables 2.6-2.9 we see that for Case 1, estimated Type I Errors from the Bonferroni Method are a little smaller than those of the Ordered P-value Method for fixed n and k and that both are satisfactorily close to their nominal values of 0.05. Estimated standard errors of these estimated type I error rates are no greater than 0.0034. Individual power comparisons that are statistically significantly different at the 0.05 level using McNemar’s Test are marked with an asterisk, *. Since we make multiple power comparisons, individual McNemar P-values should only be interpreted as rough guide lines for indicating statistically significant differences. Also we noticed that as n and k increase, the results of McNemar’s test change from non-significant to significant (even though the trend is not very strong. See a more complete summary of these comparisons in Appendix C. Overall; we conclude that Bonferroni Method is slightly more conservative than Ordered P-value Method in terms of Type I Error.

Table 2.6 Comparison of Type I Error Rates of Bonferroni Method and Ordered P-value Method for Case 1 with n = 5 and N = 5000

Type I Error (Case 1)	Bonferroni Method	Ordered P-value Method
k = 4	0.0500	0.0504
k = 6 *	0.0366	0.0396
k = 8	0.0472	0.0476

Table 2.7 Comparison of Type I Error Rates of Bonferroni Method and Ordered P-value Method for Case 1 with n = 10 and N = 5000

Type I Error (Case 1)	Bonferroni Method	Ordered P-value Method
k = 4	0.0466	0.0472
k = 6 *	0.0442	0.0460
k = 8	0.0466	0.0532

Table 2.8 Comparison of Type I Error Rates of Bonferroni Method and Ordered P-value Method for Case 1 with n = 20 and N = 5000

Type I Error (Case 1)	Bonferroni Method	Ordered P-value Method
k = 4 *	0.0468	0.0476
k = 6 *	0.0482	0.0494
k = 8	0.0478	0.0522

Table 2.9 Comparison of Type I Error Rates of Bonferroni Method and Ordered P-value Method for Case 1 with n = 50 and N = 5000

Type I Error (Case 1)	Bonferroni Method	Ordered P-value Method
k = 4	0.0550	0.0556
k = 6 *	0.0442	0.0458
k = 8	0.0410	0.0432

From Tables 2.10- 2.13, for Case 2 with $\Delta = 1$, the estimated powers of the Ordered P-value Method are always a bit higher than the corresponding powers of the Bonferroni Method given n, k. Δ. Individual comparisons that are statistically significant at the 0.05 level using McNemar's Test are with *. Also we noticed that as n and k increase, the results of McNemar's test change from significant to non-significant (only when n = 50). We can draw conclusion that Ordered P-value Method is able to keep more power than Bonferroni Method. Therefore, the results are similar as we expected and we recommend Ordered P-value Method.

Table 2.10 Comparison of Power of Bonferroni Method and Ordered P-value Method for Case 2 with n = 5, Δ = 1 and N = 5000

Power (Case 2)	Bonferroni Method	Ordered P-value Method
k = 4 *	0.0958	0.1058
k = 6 *	0.3400	0.3832
k = 8 *	0.4058	0.4574

Table 2.11 Comparison of Power of Bonferroni Method and Ordered P-value Method for Case 2 with n = 10, Δ = 1 and N = 5000

Power (Case 2)	Bonferroni Method	Ordered P-value Method
k = 4 *	0.1868	0.2008
k = 6 *	0.7222	0.7720
k = 8 *	0.7754	0.8324

Table 2.12 Comparison of Power of Bonferroni Method and Ordered P-value Method for Case 2 with n = 20, Δ = 1 and N = 5000

Power (Case 2)	Bonferroni Method	Ordered P-value Method
k = 4 *	0.3724	0.3936
k = 6 *	0.9748	0.9860
k = 8 *	0.9884	0.9950

Table 2.13 Comparison of Power of Bonferroni Method and Ordered P-value Method for Case 2 with n = 50, Δ = 1 and N = 5000

Power (Case 2)	Bonferroni Method	Ordered P-value Method
k = 4 *	0.7960	0.8190
k = 6	1.000	1.000
k = 8	1.000	1.000

From the Tables above, we conclude that Bonferroni Method is slightly more conservative than Ordered P-value Method in terms of Type I Error; while Ordered P-value Method is able to keep more power than Bonferroni Method. In sum, the results are similar as we expected that Ordered P-value Method can perform better than Bonferroni Method.

The next step is to compare power from Bonferroni Method and Ordered P-value Method when Δ is given different values (e.g. 0 - 3). What we want to compare is at each Δ value how much difference of power between these two methods. In order to get a smooth curve, we calculate the power with Δ ranging from 0 to 3 sampled at every 0.01.

The following figures present and compare smoothed estimated powers between Bonferroni Method (line in black) and Ordered P-value Method (line in red) for Case 2 with k = 4, 6, 8; n = 5, 10, 20, 50; Δ = 0 ~ 3 and N = 5000. The y-axis represents power (from 0 ~ 1), while the x-axis is marked with the indices (1 ~ 300, with 300 points in total) where Δ = index/100. Therefore they are equivalent to plots of power (0 ~ 1) verses Δ (0 ~ 3 with points increasing by 0.01).

Figure 2.1 Comparison of Power between Bonferroni Method and Ordered P-value Method for Case 2 with k = 4, n = 5 and N = 5000

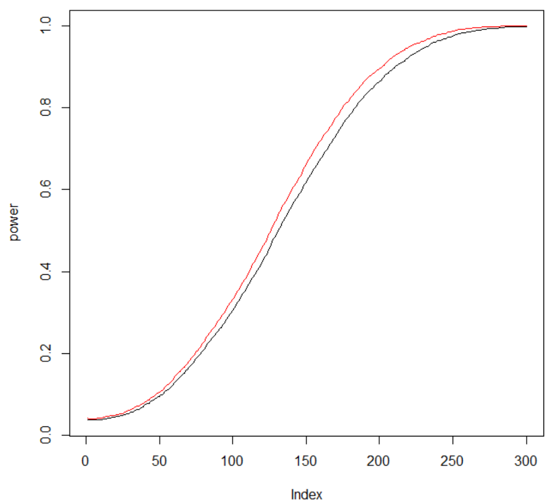


Figure 2.2 Comparison of Power between Bonferroni Method and Ordered P-value Method for Case 2 with k = 6, n = 5 and N = 5000

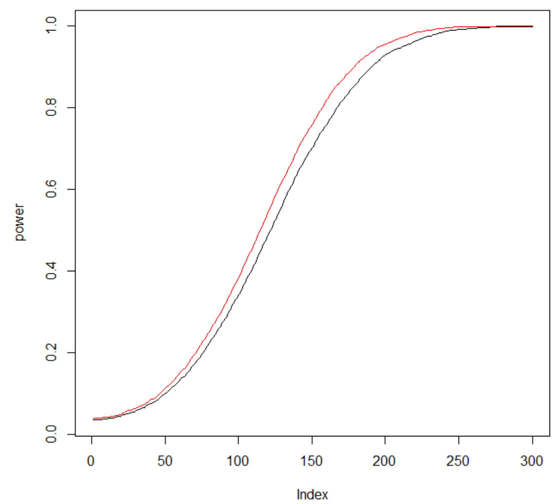


Figure 2.3 Comparison of Power between Bonferroni Method and Ordered P-value Method for Case 2 with k = 8, n = 5 and N = 5000

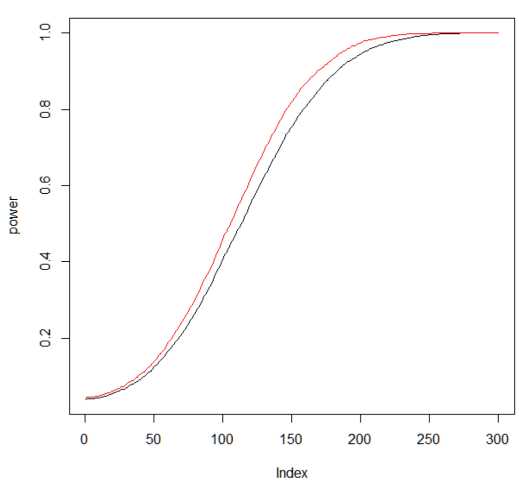


Figure 2.4 Comparison of Power between Bonferroni Method and Ordered P-value Method for Case 2 with k = 4, n = 10 and N = 5000

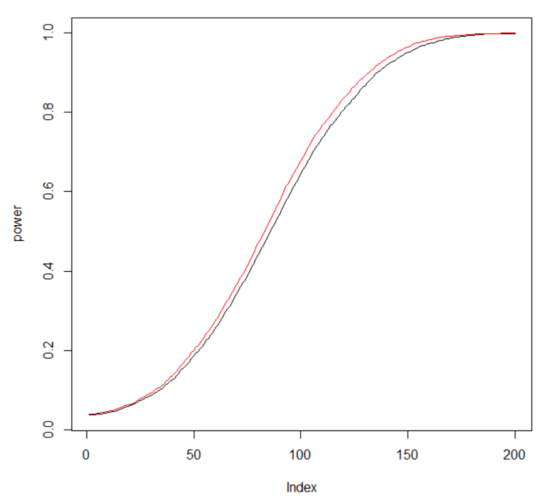


Figure 2.5 Comparison of Power between Bonferroni Method and Ordered P-value Method for Case 2 with k = 6, n = 10 and N = 5000

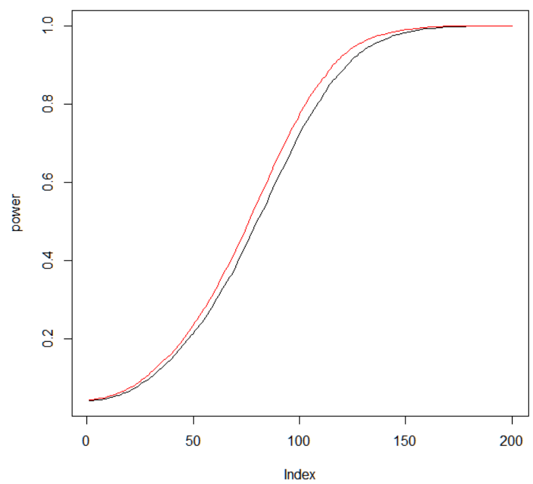


Figure 2.6 Comparison of Power between Bonferroni Method and Ordered P-value Method for Case 2 with k = 8, n = 10 and N = 5000

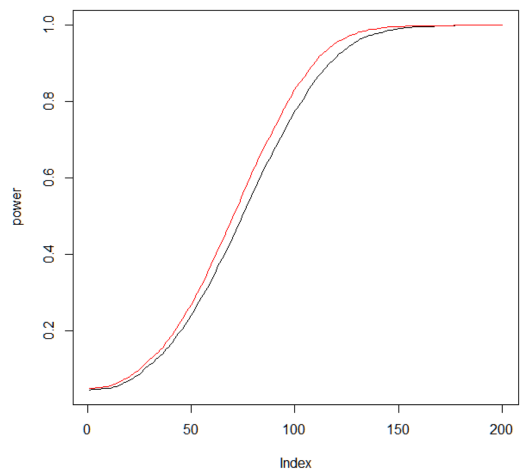


Figure 2.7 Comparison of Power between Bonferroni Method and Ordered P-value Method for Case 2 with k = 4, n = 20 and N = 5000

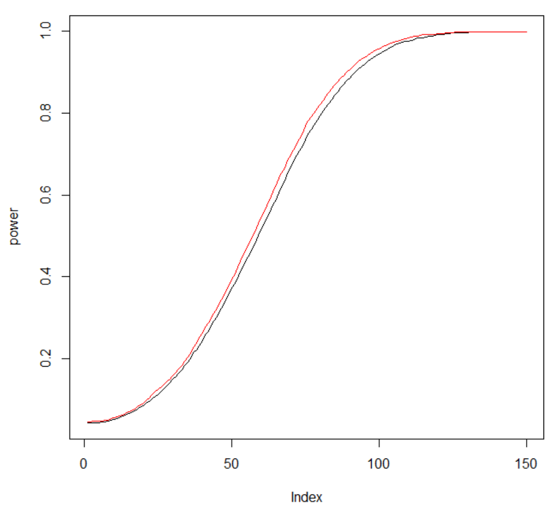


Figure 2.8 Comparison of Power between Bonferroni Method and Ordered P-value Method for Case 2 with k = 6, n = 20 and N = 5000

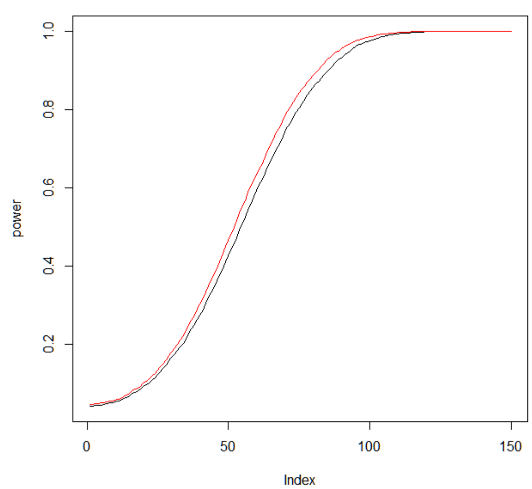


Figure 2.9 Comparison of Power between Bonferroni Method and Ordered P-value Method for Case 2 with k = 8, n = 20 and N = 5000

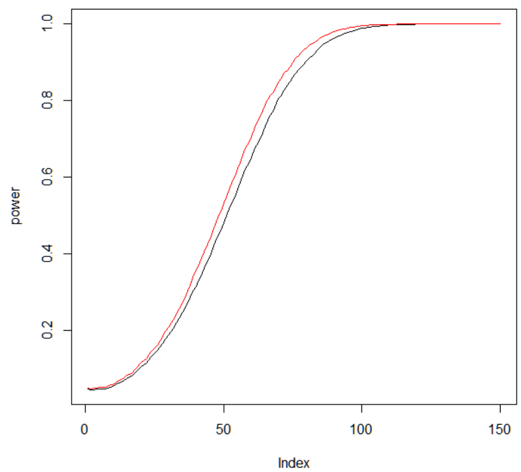


Figure 2.10 Comparison of Power between Bonferroni Method and Ordered P-value Method for Case 2 with k = 4, n = 50 and N = 5000

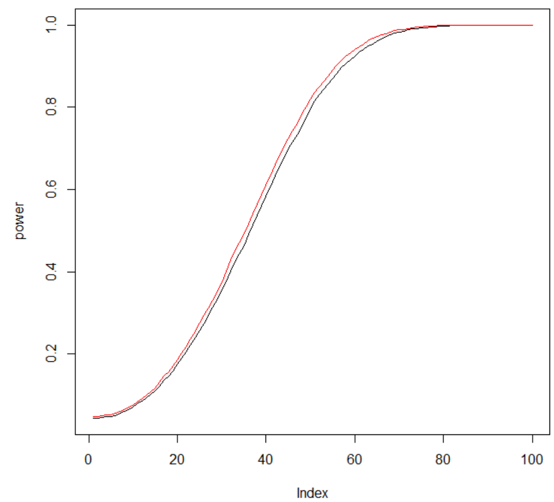


Figure 2.11 Comparison of Power between Bonferroni Method and Ordered P-value Method for Case 2 with k = 6, n = 50 and N = 5000

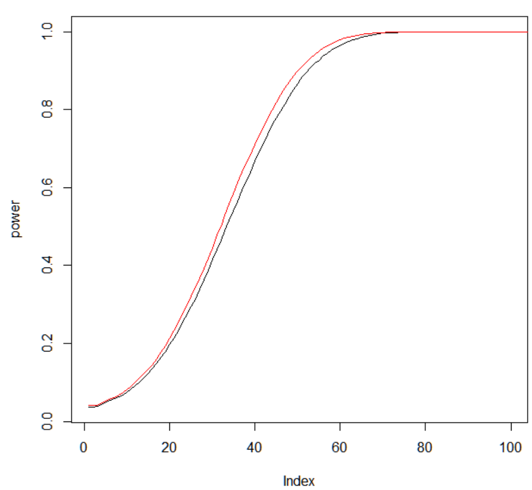
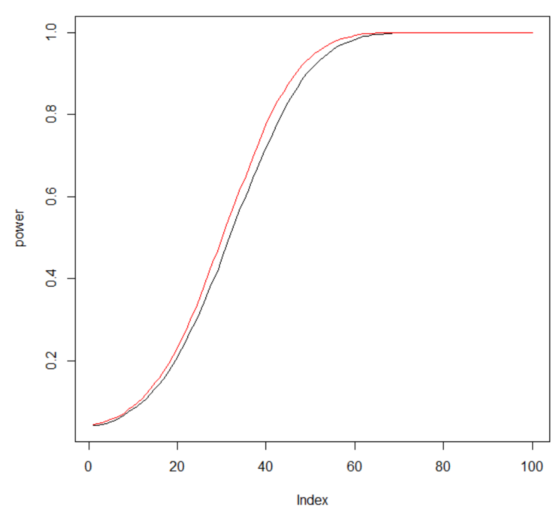


Figure 2.12 Comparison of Power between Bonferroni Method and Ordered P-value Method for Case 2 with k = 8, n = 50 and N = 5000



From the figures above, we can observe that red line is always on top of black curve, except at the starting and ending points, where both are close to 0.05 and 1.0, respectively, which means estimated power from the Ordered P-value Method is never lower than Bonferroni Method. Another interesting phenomenon is that power approaches 1.0 faster as the values of *k* and *n* increase, which indicates that, with more replication and/or more levels of factors, power grows more quickly.

In sum, based upon the results from Tables and Figures both reach the same conclusion: the Ordered P-value Method has comparable type I error rate and slightly more power than the Bonferroni Method. We therefore would like to recommend Ordered P-value Method.

Chapter 3 - Conclusion and Recommendation

3.1 Conclusion

Using simulation, we compared the Type I Error rates and Powers for testing pairwise contrasts in one way, completely randomized, balanced designs using both the Bonferroni Method and Ordered P-value Methods of multiple comparisons. We observed that the Bonferroni Method is a little more conservative than the Ordered P-value Method and that both are effective in controlling experiment wise Type I error rates ; while in terms of power, the Ordered P-value Method was a little better than the Bonferroni Method. Somewhat to our surprise, there was not a great advantage in using the Ordered P-value Method. Even though a clear difference exists, it is not as large as we had anticipated. In other words, the Bonferroni Method is not nearly as bad in the cases we studied as is often claimed. It is very competitive and easy to apply method.

In conclusion, however, we believe that the Ordered P-value Method should be more widely used to make multiple comparisons among a series of contrasts. Overall, the Ordered P-value Method is just as good as the Bonferroni Method in controlling the Experiment Wise Type I Error Rates and has greater power for correlated tests of contrasts in the one way completely randomized designs studied here.

3.2 Recommendations

For future work, we need to take the following in to consideration.

(i) test contrasts other than just pairwise comparisons

(ii) test the robustness of both procedures if variance is not constant across treatments

(iii) test data with distributions other than normal (e.g. with heavy tail, outliers)

(iv) investigate the effects of unequal sample sizes on the performance of both procedures.

(v) investigate settings where there are larger correlations among P-values than the values attained here.

References

Benjamini, Y. (2010). Simultaneous and Selective Inference: Current Successes and Future Challenges. *Biometrical Journal*, 52, 708-721.

Benjamini, Y. and Yekutieli, D. (2001). The Control of False Discovery Rate in Multiple Testing under dependency. *Annals of Statistics, 29*, 1165-1188.

Hommel, G. (1988). A Stagewise Rejective Multiple Test Procedure Based on a Modified Bonferroni Procedure. *Biometrika*, 75, 383-386.

Simes, R.J. (1986). An Improved Bonferroni Procedue for Multiple Tests of Significance. *Biometrika*, 73, 751-754.

SAS/STAT 9.2 User's Guide: The MULTTEST Procedure (Book Excerpt)

Kuehl, R. (2000). Design of Experiments: Statistical Principles of Research Design and Analysis (2nd Edition). *Duxbury Press*.

Westfall, P. and Young, S. (1993). Resampling-Based Multiple Testing: Examples and Methods for P-value Adjustment. *John Wiley & Sons, Inc.*

Efron, B (2010). Large-Scale inference: Empirical Bayes Methods for Estimation, Testing, and Prediction. *Cambridge University Press.*

Conneely, K. and Boehnke, M. (2007). So Many Correlated Tests, So Little Time! Rapid Adjustment of P Value for Multiple Correlated Tests. *The American Journal of Human Genetics*, *81*, 1158-1168.

Efron, B. (2007). Correlation and Large-Scale Simultaneous Significance Testing. *Journal of the American Statistical Association*, *102*, 93-103.

McNemar, Q. (1947). Note on the Sampling Error of the Difference between Correlated Proportions or Percentages. *Psychometrika, 12,* 153-157.

Appendix A - Example of Kuehl (2000)

The following are supplemental materials for example of Kuehl (2000).

A.1 Raw Data

The following table shows the raw data collected in the example of Kuehl (2000).

Table A.1 Data from Example of Kuehl (2000)

Treatment	Serum T3, (ng/dl) × 10^{-1}				
Premolt	94.09	90.45	99.38	73.56	74.39
Fasting	98.81	103.55	115.23	129.06	117.61
60g bran	197.18	207.31	177.50	226.05	222.74
80g bran	102.93	117.51	119.92	112.01	101.10
Laying mash	83.14	89.59	87.76	96.43	82.94

A.2 SAS Code and Output

```
data egg;
input trt $ serum @@;
datalines;
1 94.09 1 90.45 1 99.38 1 73.56 1 74.39
2 98.81 2 103.55 2 115.23 2 129.06 2 117.61
3 197.18 3 207.31 3 177.50 3 226.05 3 222.74
4 102.93 4 117.51 4 119.92 4 112.01 4 101.10
5 83.14 5 89.59 5 87.76 5 96.43 5 82.94
;
run;
proc glm data=egg;
class trt;
model serum=trt/solution;
lsmeans trt/stderr cl pdiff;
run;
```

The GLM Procedure

Class Level Information

Class	Levels	Values
trt	5	1 2 3 4 5

Number of Observations Read	25
Number of Observations Used	25

Dependent Variable: serum

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	48568.87638	12142.21909	78.08	<.0001
Error	20	3110.18912	155.50946		
Corrected Total	24	51679.06550			

R-Square	Coeff Var	Root MSE	serum Mean
0.939817	10.32231	12.47034	120.8096

Least Squares Means for effect trt

Pr > |t| for H0: LSMean(i)=LSMean(j)

Dependent Variable: serum

i/j	1	2	3	4	5
1		0.0031	<.0001	0.0059	0.8415
2	0.0031		<.0001	0.7872	0.0050
3	<.0001	<.0001		<.0001	<.0001
4	0.0059	0.7872	<.0001		0.0092
5	0.8415	0.0050	<.0001	0.0092	

trt	serum LSMEAN	95% Confidence Limits	
1	86.374000	74.740776	98.007224
2	112.852000	101.218776	124.485224
3	206.156000	194.522776	217.789224
4	110.694000	99.060776	122.327224
5	87.972000	76.338776	99.605224

```
proc glm data=egg;
class trt;
model serum=trt/solution;
lsmeans trt/stderr cl pdiff adjust=bon;
run;
```

Least Squares Means

Adjustment for Multiple Comparisons: Bonferroni

Least Squares Means for effect trt

Pr > |t| for H0: LSMean(i)=LSMean(j)

Dependent Variable: serum

i/j	1	2	3	4	5
1		0.0314	<.0001	0.0586	1.0000
2	0.0314		<.0001	1.0000	0.0499
3	<.0001	<.0001		<.0001	<.0001
4	0.0586	1.0000	<.0001		0.0924
5	1.0000	0.0499	<.0001	0.0924	

trt	serum LSMEAN	95% Confidence Limits	
1	86.374000	74.740776	98.007224
2	112.852000	101.218776	124.485224
3	206.156000	194.522776	217.789224
4	110.694000	99.060776	122.327224
5	87.972000	76.338776	99.605224

Least Squares Means for Effect trt

i	j	Difference Between Means	Simultaneous 95% Confidence Limits for LSMean(i)-LSMean(j)	
1	2	-26.478000	-51.348673	-1.607327
1	3	-119.782000	-144.652673	-94.911327
1	4	-24.320000	-49.190673	0.550673

1	5	-1.598000	-26.468673	23.272673
2	3	-93.304000	-118.174673	-68.433327
2	4	2.158000	-22.712673	27.028673
2	5	24.880000	0.009327	49.750673
3	4	95.462000	70.591327	120.332673
3	5	118.184000	93.313327	143.054673
4	5	22.722000	-2.148673	47.592673

```
estimate 'premolt vs fasting' trt 1 -1 0 0 0;
estimate 'fasting vs 60 grams' trt 0 1 -1 0 0;
estimate '60 grams vs 80 grams' trt 0 0 1 -1 0;
estimate '80 grams vs laying' trt 0 0 0 1 -1;
contrast 'premolt vs fasting' trt 1 -1 0 0 0;
contrast 'fasting vs 60 grams' trt 0 1 -1 0 0;
contrast '60 grams vs 80 grams' trt 0 0 1 -1 0;
contrast '80 grams vs laying' trt 0 0 0 1 -1;
```

Parameter	Estimate	Standard Error	t Value	Pr > |t|
premolt vs fasting	-26.4780000	7.88693745	-3.36	0.0031
fasting vs 60 grams	-93.3040000	7.88693745	-11.83	<.0001
60 grams vs 80 grams	95.4620000	7.88693745	12.10	<.0001
80 grams vs laying	22.7220000	7.88693745	2.88	0.0092

Contrast	DF	Contrast SS	Mean Square	F Value	Pr > F
premolt vs fasting	1	1752.71121	1752.71121	11.27	0.0031
fasting vs 60 grams	1	21764.09104	21764.09104	139.95	<.0001
60 grams vs 80 grams	1	22782.48361	22782.48361	146.50	<.0001
80 grams vs laying	1	1290.72321	1290.72321	8.30	0.0092

A.3 R Code and Output

```
a=c(94.09,90.45,99.38,73.56,74.39)
b=c(98.81,103.55,115.23,129.06,117.61)
c=c(197.18,207.31,177.50,226.05,222.74)
d=c(102.93,117.51,119.92,112.01,101.10)
e=c(83.14,89.59,87.76,96.43,82.94)

k=5
n=5
dfe=(n-1)*k
a1=mean(a)
b1=mean(b)
c1=mean(c)
d1=mean(d)
e1=mean(e)

ssa=0
ssb=0
ssc=0
ssd=0
sse=0

aa=c(rep(0,n))
bb=c(rep(0,n))
cc=c(rep(0,n))
dd=c(rep(0,n))
ee=c(rep(0,n))

for(i in 1:n){

aa[i]=(a[i]-a1)*(a[i]-a1)
```

```
bb[i]=(b[i]-b1)*(b[i]-b1)
cc[i]=(c[i]-c1)*(c[i]-c1)
dd[i]=(d[i]-d1)*(d[i]-d1)
ee[i]=(e[i]-e1)*(e[i]-e1)
ssa=ssa+aa[i]
ssb=ssb+bb[i]
ssc=ssc+cc[i]
ssd=ssd+dd[i]
sse=sse+ee[i]
}

mse=(ssa+ssb+ssc+ssd+sse)/dfe

tab=(a1-b1)/sqrt(mse*2/n)
pab=2*(1-pt(abs(tab),dfe))
tbc=(b1-c1)/sqrt(mse*2/n)
pbc=2*(1-pt(abs(tbc),dfe))
tcd=(c1-d1)/sqrt(mse*2/n)
pcd=2*(1-pt(abs(tcd),dfe))
tde=(d1-e1)/sqrt(mse*2/n)
pde=2*(1-pt(abs(tde),dfe))
```

	P-value	t-value	F-value
ab	0.003135828	-3.3572	11.27077
bc	1.74818E-10	-11.8302	139.9534
cd	1.16742E-10	12.10381	146.5022
de	0.009237985	2.880966	8.299965

Note: “a” represents “premolt”; “b” represents “fasting”; “c” represents “60 grams”; “d” represents “80 grams”; “e” represents “laying”.

From Above, the results from SAS is same as output from R, which illustrates that our R code is compiled correctly.

Appendix B - Correlation Coefficient

B.1 Calculation of Correlation coefficient

The following shows the R code to calculate correlation coefficient of P-values from two contrasts with different k values.

Table B.1 R Code to Calculate Correlation Coefficient of P-value from two contrasts with Different k Values

k = 3	k = 4

nrep=5000
k=3
n=5
dfe=(n-1)*k
tab=c(rep(0,nrep))
tbc=c(rep(0,nrep))
pab=c(rep(0,nrep))
pbc=c(rep(0,nrep))
 	
for (j in 1:nrep){
set.seed(567+j)
a=rnorm(n,m=0,sd=1)
# a=rnorm(n,m=1,sd=1) 	
b=rnorm(n,m=0,sd=1)
c=rnorm(n,m=0,sd=1)
 	
a1=mean(a)
b1=mean(b)
c1=mean(c)
	nrep=5000
	k=4
	n=5
	dfe=(n-1)*k
	tab=c(rep(0,nrep))
	tcd=c(rep(0,nrep))
	pab=c(rep(0,nrep))
	pcd=c(rep(0,nrep))
	
	for (j in 1:nrep){
	set.seed(567+j)
	a=rnorm(n,m=0,sd=1)
	# a=rnorm(n,m=1,sd=1)
	b=rnorm(n,m=0,sd=1)
	c=rnorm(n,m=0,sd=1)
	d=rnorm(n,m=0,sd=1)
	
	a1=mean(a)
	b1=mean(b)
	c1=mean(c)
	d1=mean(d)

<table>
<tr>
<td>ssa=0
ssb=0
ssc=0

aa=c(rep(0,n))
bb=c(rep(0,n))
cc=c(rep(0,n))

for(i in 1:n){
aa[i]=(a[i]-a1)*(a[i]-a1)
bb[i]=(b[i]-b1)*(b[i]-b1)
cc[i]=(c[i]-c1)*(c[i]-c1)

ssa=ssa+aa[i]
ssb=ssb+bb[i]
ssc=ssc+cc[i]
}

mse=(ssa+ssb+ssc)/dfe
tab[j]=(a1-b1)/sqrt(mse*2/n)
pab[j]=2*(1-pt(abs(tab[j]),dfe))
tbc[j]=(b1-c1)/sqrt(mse*2/n)
pbc[j]=2*(1-pt(abs(tbc[j]),dfe))
}

pab1=mean(pab)
pbc1=mean(pbc)</td>
<td>ssa=0
ssb=0
ssc=0
ssd=0

aa=c(rep(0,n))
bb=c(rep(0,n))
cc=c(rep(0,n))
dd=c(rep(0,n))

for(i in 1:n){
aa[i]=(a[i]-a1)*(a[i]-a1)
bb[i]=(b[i]-b1)*(b[i]-b1)
cc[i]=(c[i]-c1)*(c[i]-c1)
dd[i]=(d[i]-d1)*(d[i]-d1)

ssa=ssa+aa[i]
ssb=ssb+bb[i]
ssc=ssc+cc[i]
ssd=ssd+dd[i]
}

mse=(ssa+ssb+ssc+ssd)/dfe
tab[j]=(a1-b1)/sqrt(mse*2/n)
pab[j]=2*(1-pt(abs(tab[j]),dfe))
tcd[j]=(c1-d1)/sqrt(mse*2/n)
pcd[j]=2*(1-pt(abs(tcd[j]),dfe))
}

pab1=mean(pab)
pcd1=mean(pcd)</td>
</tr>
</table>

ab=c(rep(0,nrep))
bc=c(rep(0,nrep))
abc=c(rep(0,nrep))
sab=0
sbc=0
sabc=0
 	
for(i in 1:nrep){
ab[i]=(pab[i]-pab1)*(pab[i]-pab1)
sab=sab+ab[i]
bc[i]=(pbc[i]-pbc1)*(pbc[i]-pbc1)
sbc=sbc+bc[i]
abc[i]=(pab[i]-pab1)*(pbc[i]-pbc1)
sabc=sabc+abc[i]
}
cov=sabc/sqrt(sab)/sqrt(sbc)
	ab=c(rep(0,nrep))
	cd=c(rep(0,nrep))
	abcd=c(rep(0,nrep))
	sab=0
	scd=0
	sabcd=0
	
	for(i in 1:nrep){
	ab[i]=(pab[i]-pab1)*(pab[i]-pab1)
	sab=sab+ab[i]
	cd[i]=(pcd[i]-pcd1)*(pcd[i]-pcd1)
	scd=scd+cd[i]
	abcd[i]=(pab[i]-pab1)*(pcd[i]-pcd1)
	sabcd=sabcd+abcd[i]
	}
	cov=sabcd/sqrt(sab)/sqrt(scd)

Table B.2 Results of Correlation Coefficient of P-value from Two Contrasts with k = 3, n = 5, 10, 20, 50 and Δ = 1, 1.5, 2, 2.5

n		5	10	20	50
Δ	1	-0.03333	-0.06204	-0.0837	-0.10436
	1.5	-0.07714	-0.07932	-0.0321	-0.10436
	2	-0.07967	-0.07024	0.000562	-0.04672
	2.5	-0.05631	-0.05951	0.014385	-0.04442

Figure B.1 Scatter Plot of P-value of Contrasts (c_2 verse c_1) with k = 3, n = 5 and N = 5000 for Case One

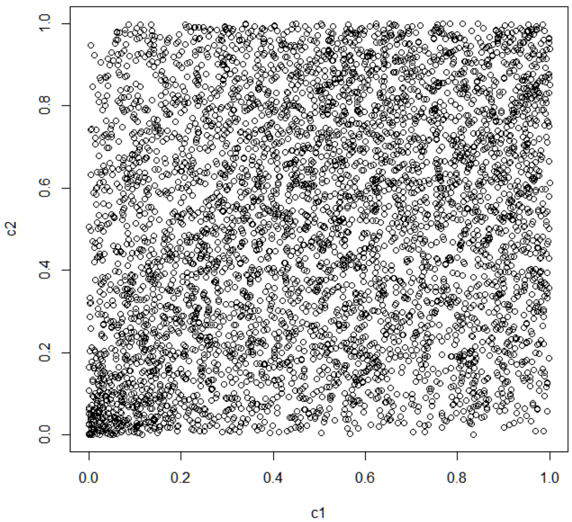


Figure B.2 Scatter Plot of P-value of Contrasts (c_2 verse c_1) with k = 4, n = 5 and N = 5000 for Case One

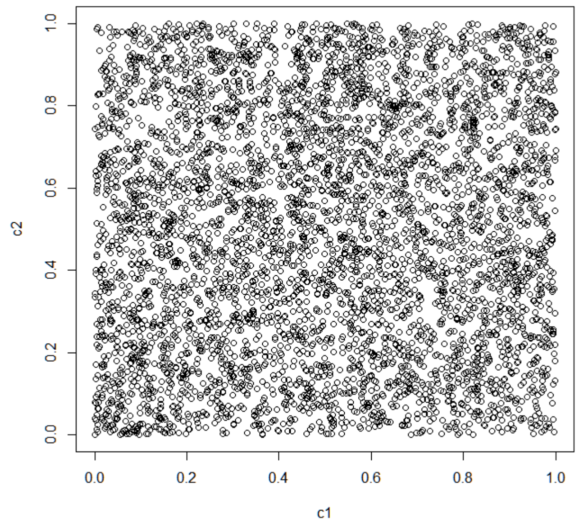


Figure B.3 Scatter Plot of P-value of Contrasts (c_2 verse c_1) with k = 3, n = 10 and N = 5000 for Case One

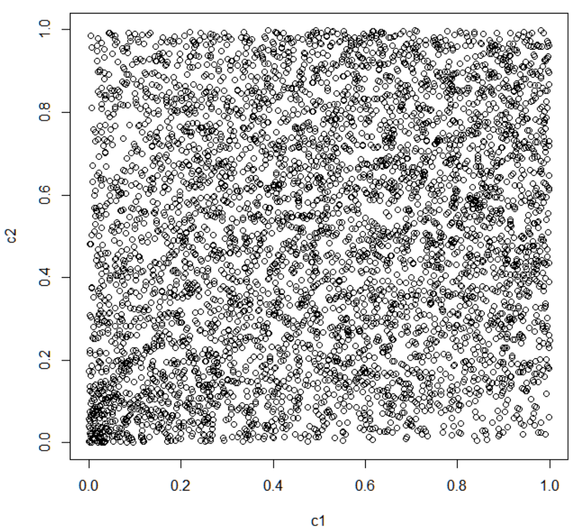


Figure B.4 Scatter Plot of P-value of Contrasts (c_2 verse c_1) with k = 4, n = 10 and N = 5000 for Case One

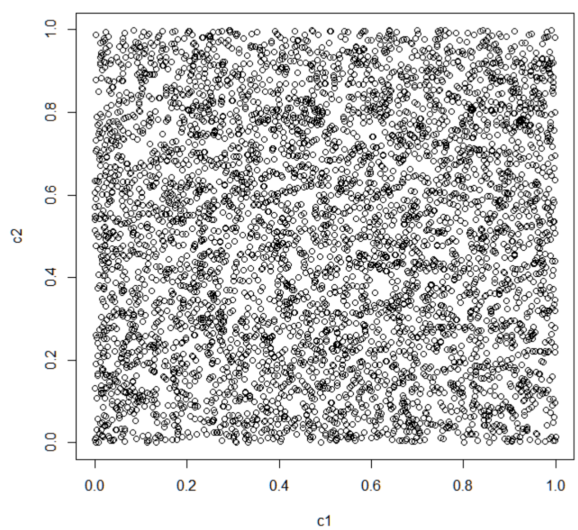


Figure B.5 Scatter Plot of P-value of Contrasts (c_2 verse c_1) with k = 3, n = 20 and N = 5000 for Case One

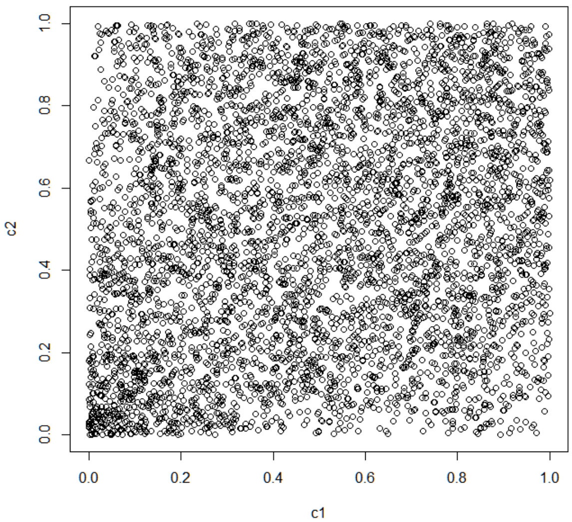


Figure B.6 Scatter Plot of P-value of Contrasts (c_2 verse c_1) with k = 4, n = 20 and N = 5000 for Case One

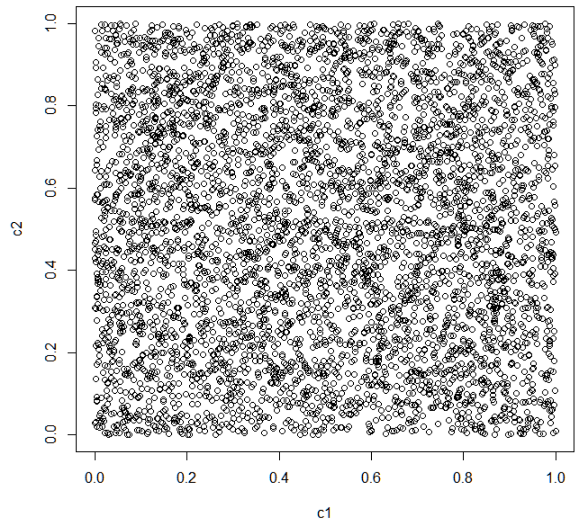


Figure B.7 Scatter Plot of P-value of Contrasts (c_2 verse c_1) with k = 3, n = 50 and N = 5000 for Case One

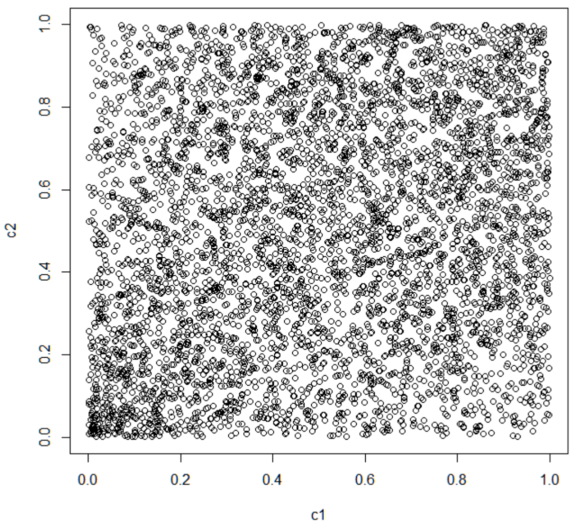
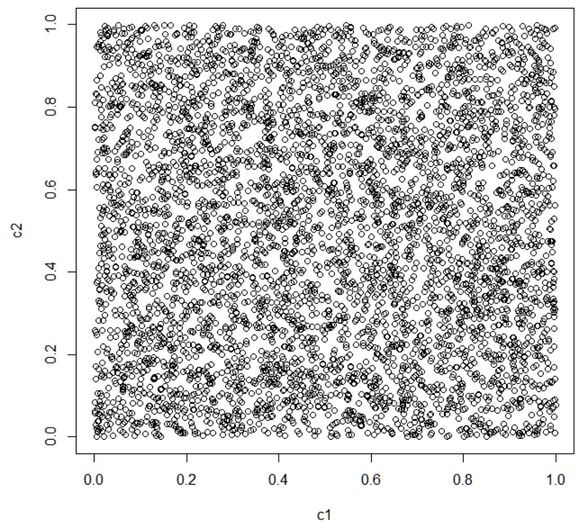


Figure B.8 Scatter Plot of P-value of Contrasts (c_2 verse c_1) with k = 4, n = 50 and N = 5000 for Case One



Appendix C - Comparison of Bonferroni Method and Ordered P-value Method

C.1 R Code

The following is R code using McNemar's Test to compare Bonferroni Method and Ordered P-value Method.

```
nrep=1000
k=4
n=20
dfe=(n-1)*k
tab=c(rep(0,nrep))
tcd=c(rep(0,nrep))
tac=c(rep(0,nrep))
tad=c(rep(0,nrep))
tbc=c(rep(0,nrep))
tbd=c(rep(0,nrep))
pab=c(rep(0,nrep))
pcd=c(rep(0,nrep))
pac=c(rep(0,nrep))
pad=c(rep(0,nrep))
pbc=c(rep(0,nrep))
pbd=c(rep(0,nrep))
bon1=c(rep(NA,nrep))
bon2=c(rep(NA,nrep))
adj1=c(rep(NA,nrep))
adj2=c(rep(NA,nrep))
n1_00=c(rep(0,nrep))
n1_10=c(rep(0,nrep))
n1_01=c(rep(0,nrep))
n1_11=c(rep(0,nrep))
```

```
n2_00=c(rep(0,nrep))
n2_10=c(rep(0,nrep))
n2_01=c(rep(0,nrep))
n2_11=c(rep(0,nrep))

for (j in 1:nrep){
set.seed(5677+j)
a=rnorm(n,m=0,sd=1)
b=rnorm(n,m=0,sd=1)
c=rnorm(n,m=1,sd=1) #delta=1
d=rnorm(n,m=1,sd=1)
a1=mean(a)
b1=mean(b)
c1=mean(c)
d1=mean(d)
ssa=0
ssb=0
ssc=0
ssd=0
aa=c(rep(0,n))
bb=c(rep(0,n))
cc=c(rep(0,n))
dd=c(rep(0,n))

for(i in 1:n){
aa[i]=(a[i]-a1)*(a[i]-a1)
bb[i]=(b[i]-b1)*(b[i]-b1)
cc[i]=(c[i]-c1)*(c[i]-c1)
dd[i]=(d[i]-d1)*(d[i]-d1)
ssa=ssa+aa[i]
ssb=ssb+bb[i]
```

```
ssc=ssc+cc[i]
ssd=ssd+dd[i]
}
sse=ssa+ssb+ssc+ssd
mse=sse/dfe
tab[j]=(a1-b1)/sqrt(mse*2/n) #H0
pab[j]=2*(1-pt(abs(tab[j]),dfe))
tcd[j]=(c1-d1)/sqrt(mse*2/n)
pcd[j]=2*(1-pt(abs(tcd[j]),dfe))
tac[j]=(a1-c1)/sqrt(mse*2/n) #Ha
pac[j]=2*(1-pt(abs(tac[j]),dfe))
tad[j]=(a1-d1)/sqrt(mse*2/n)
pad[j]=2*(1-pt(abs(tad[j]),dfe))
tbc[j]=(b1-c1)/sqrt(mse*2/n)
pbc[j]=2*(1-pt(abs(tbc[j]),dfe))
tbd[j]=(b1-d1)/sqrt(mse*2/n)
pbd[j]=2*(1-pt(abs(tbd[j]),dfe))

#bon
bon1[j]=1-(pab[j]>0.05/2)*(pcd[j]>0.05/2)    #case1 ab,cd m=2 with 1-reject 0-not reject
bon2[j]=1-(pab[j]>0.05/5)*(pac[j]>0.05/5)*(pad[j]>0.05/5)*(pbc[j]>0.05/5)*(pbd[j]> 0.05/5)
#case2 ab,ac,ad,bc,bd m=1+4 with 1-reject

#adj P
x1=c(pab[j],pcd[j])
y1=sort(x1)
adj1[j]=1-(y1[1]>0.05/2)*(y1[2]>0.05*2/2)        #case1 ab,cd m=2 with 1-reject

x2=c(pab[j],pac[j],pad[j],pbc[j],pbd[j])
y2=sort(x2)
```

```
adj2[j]=1-(y2[1]>0.05/5)*(y2[2]>0.05*2/5)*(y2[3]>0.05*3/5)*(y2[4]>0.05*4/5)*(y2[5]>
0.05*5/5)          #case2 ab,ac,ad,bc,bd m=1+4 with 1-reject

#McNemar's test
n1_00[j][bon1[j]==0 & adj1[j]==0]<-1
n1_11[j][bon1[j]==1 & adj1[j]==1]<-1
n1_01[j][bon1[j]==1 & adj1[j]==0]<-1
n1_10[j][bon1[j]==0 & adj1[j]==1]<-1

n2_00[j][bon2[j]==0 & adj2[j]==0]<-1
n2_11[j][bon2[j]==1 & adj2[j]==1]<-1
n2_01[j][bon2[j]==1 & adj2[j]==0]<-1
n2_10[j][bon2[j]==0 & adj2[j]==1]<-1

}
```

For case 1, both two contrasts are true, our type I error is 0.048 for Bonferroni Method and 0.049 for Ordered P-value Method, both are close to 0.05

```
> mean(bon1)
[1] 0.048
> mean(adj1)
[1] 0.049
> sum(n1_00)
[1] 951
> sum(n1_11)
[1] 48
> sum(n1_10)
[1] 1
> sum(n1_01)
[1] 0
```

```
> sum(bon1)
[1] 48
> sum(adj1)
[1] 49
```

Ho n_01=n_10 Ha n_01≠n_10

$\chi^2 = (n\_01 - n\_10)^2 /(n\_01 + n\_10)=(1-0)^2/(1+0)=1$ P-value = 0.3173

For case 2, we have five contrasts with one of them is true and the other four are false. The power is 0.945 for Bonferroni Method and 0.954 for Ordered P-value Method. Ordered P-value Method is better than Bonferroni Method in terms of power.

```
> mean(bon2)
[1] 0.945
> mean(adj2)
[1] 0.954
> sum(n2_00)
[1] 46
> sum(n2_11)
[1] 945
> sum(n2_10)
[1] 9
> sum(n2_01)
[1] 0
> sum(adj2)
[1] 954
> sum(bon2)
[1] 945
```

Ho n_01=n_10 Ha n_01≠n_10

$\chi2 = (n\_01 - n\_10)2 /(n\_01 + n\_10)=(9-0)2/(9+0)=9$ P-value = 0.0026

C.2 McNemar’s Test Result for Case 1 and 2

The following table shows the result of comparisons of Type I Error and Power between Bonferroni Method and Ordered P-value Method and each corresponding McNemar’s Test with n = 5, 10, 20, 50; k = 4, 6, 8; Δ = 0, 0.5, 1; N = 1000, 5000 raw data collected in the example of Kuehl (2000).

Table C.1 Comparison of Type I Error between Bonferroni Method and Ordered P-value Method for Case 1 with n = 5 and N = 1000

Type I Error (Case 1)	Bonferroni Method	Ordered P-value Method
k = 4	0.048	0.049
k = 6	0.042	0.045
k = 8	0.045	0.049

Table C.2 Result of McNemar’s Test Comparing between Bonferroni Method and Ordered P-value Method for Case 1 with n = 5 and N = 1000

McNemar's test	χ^2	P-value
k=4	1	0.3173
k=6	3	0.0832
k=8	1.333	0.2482

Table C.3 Comparison of Type I Error between Bonferroni Method and Ordered P-value Method for Case 1 with n = 10 and N = 1000

Type I Error (Case 1)	Bonferroni Method	Ordered P-value Method
k = 4	0.038	0.038
k = 6	0.040	0.040
k = 8	0.041	0.043

Table C.4 Result of McNemar’s Test Comparing between Bonferroni Method and Ordered P-value Method for Case 1 with n = 10 and N = 1000

McNemar's test	χ^2	P-value
k=4	0	1
k=6	0	1
k=8	0.667	0.4142

Table C.5 Comparison of Type I Error between Bonferroni Method and Ordered P-value Method for Case 1 with n = 20 and N = 1000

Type I Error (Case 1)	Bonferroni Method	Ordered P-value Method
k = 4	0.048	0.049
k = 6	0.053	0.054
k = 8	0.046	0.055

Table C.6 Result of McNemar’s Test Comparing between Bonferroni Method and Ordered P-value Method for Case 1 with n = 20 and N = 1000

McNemar's test	χ^2	P-value
k=4	1	0.3173
k=6	1	0.3173
k=8	7.3636	0.0066

Table C.7 Comparison of Type I Error between Bonferroni Method and Ordered P-value Method for Case 1 with n = 50 and N = 1000

Type I Error (Case 1)	Bonferroni Method	Ordered P-value Method
k = 4	0.059	0.060
k = 6	0.057	0.060
k = 8	0.047	0.054

Table C.8 Result of McNemar's Test Comparing between Bonferroni Method and Ordered P-value Method for Case 1 with n = 50 and N = 1000

McNemar's test	χ^2	P-value
k=4	1	0.3173
k=6	3	0.0832
k=8	3.7692	0.0522

Table C.9 Comparison of Power between Bonferroni Method and Ordered P-value Method for Case 2 with n = 5, Δ = 1 and N = 1000

Power (Case 2)	Bonferroni Method	Ordered P-value Method
k = 4	0.288	0.309
k = 6	0.332	0.370
k = 8	0.396	0.439

Table C.10 Result of McNemar's Test Comparing between Bonferroni Method and Ordered P-value Method for Case 2 with n = 5 , Δ = 1 and N = 1000

McNemar's test	χ^2	P-value
k=4	21	0.0000
k=6	38	0.000
k=8	43	0.000

Table C.11 Comparison of Power between Bonferroni Method and Ordered P-value Method for Case 2 with n = 10, Δ = 1 and N = 1000

Power (Case 2)	Bonferroni Method	Ordered P-value Method
k = 4	0.631	0.660
k = 6	0.708	0.761
k = 8	0.776	0.837

Table C.12 Result of McNemar's Test Comparing between Bonferroni Method and Ordered P-value Method for Case 2 with n = 10, Δ = 1 and N = 1000

McNemar's test	χ^2	P-value
k=4	29	7.2254e-8
k=6	53	0.0000
k=8	61	0.0000

Table C.13 Comparison of Power between Bonferroni Method and Ordered P-value Method for Case 2 with n = 20, Δ = 1 and N = 1000

Power (Case 2)	Bonferroni Method	Ordered P-value Method
k = 4	0.945	0.954
k = 6	0.970	0.987
k = 8	0.988	0.994

Table C.14 Result of McNemar's Test Comparing between Bonferroni Method and Ordered P-value Method for Case 2 with n = 20, Δ = 1 and N = 1000

McNemar's test	χ^2	P-value
k=4	9	0.0026
k=6	17	0.0000
k=8	6	0.0143

Table C.15 Comparison of Power between Bonferroni Method and Ordered P-value Method for Case 2 with n = 50, Δ = 1 and N = 1000

Power (Case 2)	Bonferroni Method	Ordered P-value Method
k = 4	1.000	1.000
k = 6	1.000	1.000
k = 8	1.000	1.000

Table C.16 Result of McNemar's Test Comparing between Bonferroni Method and Ordered P-value Method for Case 2 with n = 50, Δ = 1 and N = 1000

McNemar's test	χ^2	P-value
k=4	0	1
k=6	0	1
k=8	0	1

Table C.17 Comparison of Power between Bonferroni Method and Ordered P-value Method for Case 2 with n = 5, Δ = 0.5 and N = 1000

Power (Case 2)	Bonferroni Method	Ordered P-value Method
k = 4	0.100	0.114
k = 6	0.105	0.116
k = 8	0.132	0.147

Table C.18 Result of McNemar's Test Comparing between Bonferroni Method and Ordered P-value Method for Case 2 with n = 5, Δ = 0.5 and N = 1000

McNemar's test	χ^2	P-value
k=4	14	0.0001
k=6	11	0.0009
k=8	15	0.0001

Table C.19 Comparison of Power between Bonferroni Method and Ordered P-value Method for Case 2 with n = 10, Δ = 0.5 and N = 1000

Power (Case 2)	Bonferroni Method	Ordered P-value Method
k = 4	0.186	0.205
k = 6	0.213	0.230
k = 8	0.237	0.260

Table C.20 Result of McNemar's Test Comparing between Bonferroni Method and Ordered P-value Method for Case 2 with n = 10, Δ = 0.5 and N = 1000

McNemar's test	χ^2	P-value
k=4	19	0.0000
k=6	17	0.0000
k=8	23	0.0000

Table C.21 Comparison of Power between Bonferroni Method and Ordered P-value Method for Case 2 with n = 20, Δ = 0.5 and N = 1000

Power (Case 2)	Bonferroni Method	Ordered P-value Method
k = 4	0.361	0.385
k = 6	0.412	0.449
k = 8	0.490	0.538

Table C.22 Result of McNemar's Test Comparing between Bonferroni Method and Ordered P-value Method for Case 2 with n = 20, Δ = 0.5 and N = 1000

McNemar's test	χ^2	P-value
k=4	24	9.6355e-7
k=6	37	0.0000
k=8	48	0.0000

Table C.23 Comparison of Power between Bonferroni Method and Ordered P-value Method for Case 2 with n = 50, Δ = 0.5 and N = 1000

Power (Case 2)	Bonferroni Method	Ordered P-value Method
k = 4	0.801	0.827
k = 6	0.874	0.905
k = 8	0.906	0.937

Table C.24 Result of McNemar's Test Comparing between Bonferroni Method and Ordered P-value Method for Case 2 with n = 50, Δ = 0.5 and N = 1000

McNemar's test	χ^2	P-value
k=4	26	3.4157e-7
k=6	31	2.5894e-8
k=8	31	2.5894e-8

Table C.25 Comparison of Type I Error between Bonferroni Method and Ordered P-value Method for Case 1 with n = 5 and N = 5000

Type I Error (Case 1)	Bonferroni Method	Ordered P-value Method
k = 4	0.05	0.0504
k = 6	0.0366	0.0396
k = 8	0.0472	0.0476

Table C.26 Result of McNemar's Test Comparing between Bonferroni Method and Ordered P-value Method for Case 1 with n = 5 and N = 5000

McNemar's test	χ^2	P-value
k=4	2	0.1572
k=6	15	0.0001
k=8	0.04545	0.8311

Table C.27 Comparison of Type I Error between Bonferroni Method and Ordered P-value Method for Case 1 with n = 10 and N = 5000

Type I Error (Case 1)	Bonferroni Method	Ordered P-value Method
k = 4	0.0466	0.0472
k = 6	0.0442	0.046
k = 8	0.0466	0.0532

Table C.28 Result of McNemar's Test Comparing between Bonferroni Method and Ordered P-value Method for Case 1 with n = 10 and N = 5000

McNemar's test	χ^2	P-value
k=4	3	0.0832
k=6	9	0.0026
k=8	0.5789	0.4467

Table C.29 Comparison of Type I Error between Bonferroni Method and Ordered P-value Method for Case 1 with n = 20 and N = 5000

Type I Error (Case 1)	Bonferroni Method	Ordered P-value Method
k = 4	0.0468	0.0476
k = 6	0.0482	0.0494
k = 8	0.0478	0.0522

Table C.30 Result of McNemar's Test Comparing between Bonferroni Method and Ordered P-value Method for Case 1 with n = 20 and N = 5000

McNemar's test	χ^2	P-value
k=4	4	0.0455
k=6	6	0.0143
k=8	0.4783	0.4891

Table C.31 Comparison of Type I Error between Bonferroni Method and Ordered P-value Method for Case 1 with n = 50 and N = 5000

Type I Error (Case 1)	Bonferroni Method	Ordered P-value Method
k = 4	0.055	0.0556
k = 6	0.0442	0.0458
k = 8	0.041	0.0432

Table C.32 Result of McNemar's Test Comparing between Bonferroni Method and Ordered P-value Method for Case 1 with n = 50 and N = 5000

McNemar's test	χ^2	P-value
k=4	3	0.0832
k=6	8	0.0046
k=8	0.2683	0.6044

Table C.33 Comparison of Power between Bonferroni Method and Ordered P-value Method for Case 2 with n = 5, Δ = 1 and N = 5000

Power (Case 2)	Bonferroni Method	Ordered P-value Method
k = 4	0.0958	0.1058
k = 6	0.34	0.3832
k = 8	0.4058	0.4574

Table C.34 Result of McNemar's Test Comparing between Bonferroni Method and Ordered P-value Method for Case 2 with n = 5, Δ = 1 and N = 5000

McNemar's test	χ^2	P-value
k=4	50	0.0000
k=6	216	0.0000
k=8	258	0.0000

Table C.35 Comparison of Power between Bonferroni Method and Ordered P-value Method for Case 2 with n = 10, Δ = 1 and N = 5000

Power (Case 2)	Bonferroni Method	Ordered P-value Method
k = 4	0.1868	0.2008
k = 6	0.7222	0.772
k = 8	0.7754	0.8324

Table C.36 Result of McNemar's Test Comparing between Bonferroni Method and Ordered P-value Method for Case 2 with n = 10, Δ = 1 and N = 5000

McNemar's test	χ^2	P-value
k=4	70	0.0000
k=6	249	0.0000
k=8	285	0.0000

Table C.37 Comparison of Power between Bonferroni Method and Ordered P-value Method for Case 2 with n = 20, Δ = 1 and N = 5000

Power (Case 2)	Bonferroni Method	Ordered P-value Method
k = 4	0.3724	0.3936
k = 6	0.9748	0.986
k = 8	0.9884	0.995

Table C.38 Result of McNemar's Test Comparing between Bonferroni Method and Ordered P-value Method for Case 2 with n = 20, Δ = 1 and N = 5000

McNemar's test	χ^2	P-value
k=4	106	0.0000
k=6	56	0.0000
k=8	33	0.0000

Table C.39 Comparison of Power between Bonferroni Method and Ordered P-value Method for Case 2 with n = 50, Δ = 1 and N = 5000

Power (Case 2)	Bonferroni Method	Ordered P-value Method
k = 4	0.796	0.819
k = 6	1	1
k = 8	1	1

Table C.40 Result of McNemar's Test Comparing between Bonferroni Method and Ordered P-value Method for Case 2 with n = 50, Δ = 1 and N = 5000

McNemar's test	χ^2	P-value
k=4	115	0.0000
k=6	0	1
k=8	0	1