COMPARING DUNNETT'S TEST WITH THE FALSE DISCOVERY RATE METHOD: A SIMULATION STUDY

A Thesis
Submitted to the Graduate Faculty
of the
North Dakota State University
of Agriculture and Applied Science

By

Jamie Marie Kubat

In Partial Fulfillment of the Requirements
for the Degree of
MASTER OF SCIENCE

Major Department:
Statistics

June 2013

Fargo, North Dakota

North Dakota State University
Graduate School

Title

Comparing Dunnett's Test with the False Discovery Rate Method: A Simulation Study

By

Jamie Marie Kubat

The Supervisory Committee certifies that this ***disquisition*** complies with North Dakota State University's regulations and meets the accepted standards for the degree of

MASTER OF SCIENCE

SUPERVISORY COMMITTEE:

Curt Doetkott
Co-Chair

Rhonda Magel
Co-Chair

Megan Orr

Juan Osorno

Approved:

6/20/2013
Date

Rhonda Magel
Department Chair

ABSTRACT

Recently, the idea of multiple comparisons has been criticized because of its lack of power in datasets with a large number of treatments. Many family-wise error corrections are far too restrictive when large quantities of comparisons are being made. At the other extreme, a test like the least significant difference does not control the family-wise error rate, and therefore is not restrictive enough to identify true differences. A solution lies in multiple testing. The false discovery rate (FDR) uses a simple algorithm and can be applied to datasets with many treatments. The current research compares the FDR method to Dunnett's test using agronomic data from a study with 196 varieties of dry beans. Simulated data is used to assess type I error and power of the tests. In general, the FDR method provides a higher power than Dunnett's test while maintaining control of the type I error rate.

TABLE OF CONTENTS

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

LIST OF APPENDIX TABLES

LIST OF APPENDIX FIGURES

CHAPTER 1. INTRODUCTION

The issue of multiple testing comes up in statistical analysis for almost any discipline. From data collected in a survey to clinical trial data, there will most likely be more than one hypothesis of interest. Due to the fact that it is easier to run an additional hypothesis test rather than obtain an additional subject to complete a survey or to participate in a clinical trial, multiple testing is an important statistical concept (Westfall & Young, 1993). According to Storey (2003), with the growth in size of available datasets, it is common for the statistician to test "as many interesting features…as possible rather than test a very specific hypothesis on one item." In agricultural and biological sciences this is also the case. It would be easier to observe an additional trait or behavior on each subject of interest (i.e. plant, animal) than it would be to introduce an additional subject to the study. If we perform an analysis of variance (ANOVA) test to compare our treatments, we may reject the null hypothesis that all of the treatments perform equally. In this case, the alternative hypothesis does not tell us much other than the fact that at least one of the treatments is different from the others, but in practice this often means there are multiple differences among the treatments. Which are different? Which perform the same? These questions may be answered through multiple comparisons and multiple testing.

1.1. Multiple Comparisons

After finding treatment to be significant in the ANOVA model, post hoc testing is typically done with pairwise comparisons (t-tests) or some other contrasts. Testing each pairwise comparison, as in a Fisher least significant difference (LSD) test, will control the error rate at a level of $\alpha=0.05$ for each comparison, but does not control an overall α level (Montgomery, 2009). If we look at many such pairwise comparisons simultaneously, this will seriously inflate the overall type I error level (the probability of rejecting at least one null hypothesis when it is actually true). If we are evaluating the multiple tests overall, this inflation of the type I error rate will have severe consequences for interpretation (Shaffer, 1995). We want to control the type I error rate overall because we do not want to reject too many true null hypotheses.

In general, when we run an ANOVA test, we are testing m different hypotheses of interest, with m_0 of them being true (Benjamini & Hochberg, 1995). Table 1 shows the different outcomes of a multiple comparisons test. It is assumed that m is known, and R can be observed directly, while U, V, T, and S are not observable. A per comparison test of error rate like the LSD controls E(V/m). Controlling the overall type I error rate means controlling V.

Table 1: Outcomes of Testing *m* Hypotheses

	Conclusion		
	Null accepted	Null rejected	Total
True Null Hypothesis	U	V	m_0
Non-True Null Hypothesis	T	S	$m - m_0$
Total	$m - R$	R	m

1.1.1. Family-Wise Error Rate. Many different tests have been introduced to control for the type I error rate overall, including Bonferroni, Tukey, and Scheffé (see Shaffer, 1995, for a review of these and many other methods). These tests are known to control the family-wise error (FWE) rate, or "the probability of making one or more type I errors among all the hypotheses" (Storey, 2002). This can also be stated as P($V \geq 1$). Each FWE test is slightly different in what it controls for, be it all pairwise comparisons, all contrasts, or a combination of pairwise and contrast comparisons. Controlling the type I error rate results in more conservative conclusions, meaning it will be more difficult to reject the null under these constraints (Westfall & Young, 1993). While the three FWE controlling tests mentioned here look at all pairwise comparisons or multiple contrasts, another test exists to perform all pairwise comparisons to one control.

1.1.2. Dunnett's Test. Introduced by Charles Dunnett (1955), the FWE correction Dunnett's test compares all treatments to one specified control. Dunnett recognized limitations of the Tukey and Scheffé tests when one mean was being compared to each of the others; he introduced modified confidence intervals and tables for one-sided and two-sided t-tests comparing all treatments to one control. Howell (2002) also shows that Dunnett's test is more powerful in comparison to a Bonferroni test when comparing all of the data to one control treatment. In a follow-up to his 1955 paper, Dunnett (1964) gave exact values for making two-sided comparisons and extended the tables presented in 1955 to include more treatments.

In practice, when datasets have many treatments, instead of doing all pairwise tests, a control may be chosen and Dunnett's test performed. When the ANOVA shows significant differences exist, we can try to make a comparison to the highest mean or the lowest mean as the control, or any other control that is of interest to the researcher. We can look into how many other treatment means are significantly different from this control to try to draw conclusions from the results.

1.1.3. Limitations to Multiple Comparisons. After finding treatment to be a significant main effect in the ANOVA model, when there are many treatment levels, post hoc testing may show almost all of the treatments to be significantly different from each other or none of the treatments significantly different from each other. Different post hoc tests will yield different results on this spectrum, with LSD being the least conservative, and FWE rate corrections being more conservative. When we get into datasets with many different treatments, the FWE rate can become too restrictive, while an LSD test will yield many false rejections.

A simple example of this comes from Basford and Tukey (1999), which uses a plant breeding trial for demonstration. In this study, there are 58 different genotypes of soybeans.

Basford and Tukey point out that there are 1,653 (58 choose 2) pairwise comparisons that can be made. If we wanted to analyze these with post hoc testing, an LSD test would be expected to yield 83 (5% of 1,653) incorrect rejections among the pairwise comparisons. A Tukey test, controlling for the FWE rate, would only allow for an error rate of 0.00003 (0.05/1,653) on each comparison. In other words, the p-value for a specific comparison would have to be less than or equal to 0.00003 to be significant. As we can see, for the LSD it seems like there are far too many false rejections being seen as acceptable. On the other hand, the Tukey test seems far too restrictive to find significance for many of the pairwise tests. When getting into datasets with increasingly large numbers of treatments, FWE rate becomes even more restrictive than in this example, and LSD allows for even more false rejections.

1.2. False Discovery Rate

While FWE and LSD corrections seem to be two extremes, there is some middle ground found in the false discovery rate (FDR). This method, introduced by Benjamini and Hochberg (1995), calls for controlling the number of rejected hypotheses that were in error. We call V from Table 1 a false discovery and S a true discovery. Therefore, we want to control $Q=V/R$, or the proportion of false discoveries out of all discoveries (rejections). FDR is simply expressed as $E(Q)=E(V/R)$. When $R=0$, the FDR is defined to be 0. Benjamini and Yekutieli (2001) suggest that it is "preferable to control the proportion of errors rather than the probability of making even one error," especially as the number of rejected hypotheses rises.

The procedure for controlling the FDR presented by Benjamini and Hochberg (1995) involves looking at p-values from the individual hypothesis tests. In a situation similar to the plant data example from Basford and Tukey, we would have m hypothesis tests comparing different varieties of soybeans to each other. We can call these H_1, H_2,…, H_m, and the resulting p-values can be denoted $P_1, P_2, \ldots, P_m$. Let $P_{(1)} \leq P_{(2)} \leq \ldots \leq P_{(m)}$ represent the ordered p-values, then let k be the largest integer such that $P_{(k)} \leq \frac{k\alpha}{m}$ and reject all $H_{(i)}$ $i = 1, 2, \ldots, k$. If no such k exists, then no hypotheses are rejected.

This is clearly a less restrictive cutoff point than FWE procedures, as the α value here is being multiplied by some positive integer k. For illustration, Lazar (2012) describes a plot of ordered p-values, with 1 through m on the x-axis and 0 to 1 on the y-axis. The Bonferroni correction would draw a line at height α / m and deem any p-values below this line significant. The FDR, on the other hand, would draw a line with slope α / m and count the p-values below this line to be significant. A visual representation of this can be found in Figure 1 for m=20.

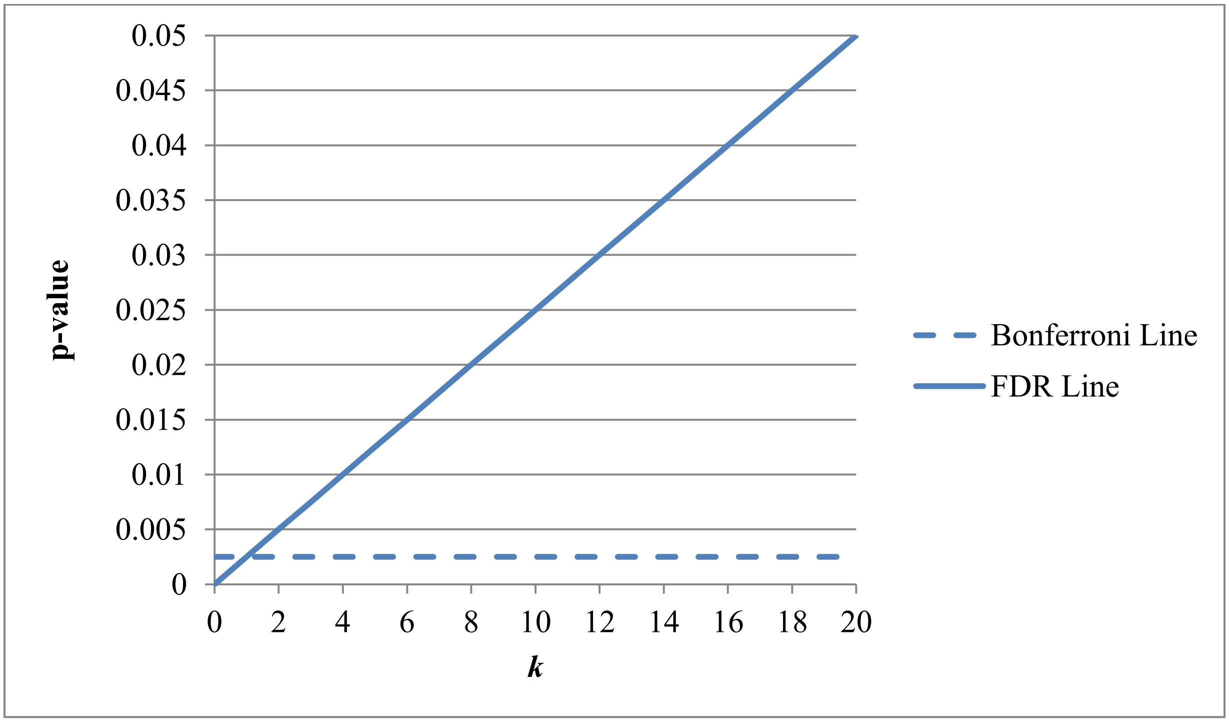

Figure 1: Comparison of Bonferroni and FDR Cutoff Points for *m*=20

The FDR has a weak control of the FWE rate. This means that if $m_0 = m$, or if all of the null hypotheses are true, the FDR and FWE rate are the same (Benjamini & Hochberg, 1995). If $m_0 < m$ then FDR ≤ FWE rate. Thus it can be said that all FWE tests control the FDR, but an

FDR test does not always control the FWE rate. If we look at an FDR test, we can allow for more errors when a lot of hypotheses are rejected (Benjamini & Yekutieli, 2001), which is not the case in FWE testing procedures that only have one cutoff point. Thus, FDR provides flexibility within different datasets.

While initially intended for use in situations similar to multiple comparisons, the FDR test has been shown to work well with extensive datasets such as those found in genetics (Lazar, 2012). While traditional multiple comparison approaches become more and more conservative with increasingly large numbers of treatments to compare, the FDR is able to handle data with many treatments. This is promising with the massive increases in computing power and availability of large datasets in recent years. The fact that power is maintained under the FDR method is also encouraging, because, as Benjamini and Yekutieli (2001) suggest, the loss in power when running multiple comparisons had led many in the field away from doing any sort of multiplicity control at all. Benjamini and Hochberg (1995) also show that a gain in power is likely with the FDR, especially when many null hypotheses are not true.

We explore this gain in power along with assessing type I error in a simulation study outlined in Chapter 3 after presenting materials and methods similar to the data from Basford and Tukey in Chapter 2. Chapter 4 presents the results and discussion of both the real and simulated data. We hope to find out which post hoc test will be more powerful through the simulation studies and we hypothesize that the FDR method will be favored. In Chapter 5 we explore some further analyses after finding some additional questions in the results that we wanted to address. These results should confirm which test is better under realistic conditions. Conclusions are presented in Chapter 6.

CHAPTER 2. MATERIALS AND METHODS

Interest in the multiplicity problem was encountered when working on a large dataset from plant sciences, similar to the dataset used in Basford and Tukey (1999). Seven different agronomic traits were recorded on over 300 different genotypes of edible dry beans. ANOVAs were run looking for differences in performance of genotypes for each agronomic trait, but no multiplicity control was examined. We wanted to use this data to compare Dunnett's test to the FDR method, and then run simulations based off of the real data to ensure the dataset was not anomalous. Information on the simulation study can be found in Chapter 3.

2.1. BeanCAP Research

The Common Bean Coordinated Agricultural Project (BeanCAP) is current research funded by the United States Department of Agriculture, National Institute of Food and Agriculture (USDA-NIFA) in partnership with many Universities across the United States (www.beancap.org). The aim of the BeanCAP research is to "[focus] on the genetics and genomics aspects of nutrition in this important food crop" and provide education and extension services (McClean, et al., n.d.). Genotypic and phenotypic data were collected and later used for Genome-Wide Association Studies (GWAS). This will provide bean breeders information on how different genes affect different traits including drought tolerance, micronutrient levels, plant height, or even seed yield.

The genotypes were planted in 2011 in four different states: Colorado, Michigan, North Dakota, and Nebraska. These four states currently produce about 90 percent of all dry beans grown in the U.S. (McClean, et al., n.d.). Beans are classified into Andean and Middle American gene pools, in which each pool is subdivided in three races: Durango, Jalisco and Mesoamerican for the Middle American gene pool and Nueva Granada, Peru, and Chile for the Andean gene pool (Singh et al., 1991). A total of three separate trials were grown in 2011 across locations based on their genetic background and race: 196 Durango genotypes, 108 Mesoamerican, and 49 Andean.

Mesoamerican beans comprise the market classes of navy and black beans (McClean, et al., n.d.). There were 108 different varieties of Mesoamerican beans represented at each of the four locations. Each variety was represented by two plots of approximately two hundred plants each. Four Durango bean varieties were also planted among the Mesoamerican crops for cross validation.

Durango beans, which include pinto and great northern market classes (McClean, et al., n.d.), were also planted at all four locations. Two replicated plots of 196 different varieties of Durango beans were sown, along with four Mesoamerican varieties for cross checks.

There were 49 different varieties of Andean beans, all of the Nueva Granada race, planted in North Dakota and Nebraska. In Michigan, only 35 of these 49 varieties were planted. Analysis was done with all 49 varieties at two locations and with 35 varieties across three locations. Each variety had two plots planted in each location.

For each bean variety plot, different agronomic traits were recorded at each location; a description of all of these traits can be found in McClean, et al. (n.d.). We outline seven of these agronomic traits here which are directly observable from the field and the harvested beans. The first trait is days to flower (DF). This is the number of days from planting until about half of the plants in each plot have at least one flower that is open. Days to maturity (DM) is the number of days from planting until about half of the plants in the plot have at least one dry seed pod. Plant height (PH) is taken in centimeters from the soil to the top node bearing at least one dry seed pod. The plant height of each variety is measured in the center of the plot after flowering occurred since the plants do not grow much taller after this point in time. Growth habit (GH) and lodging (LG) are both measures of how the plant is growing. These measures are taken two to three weeks before harvest. GH has three levels, where type 1 means the plant is standing erect, type 2 is indeterminate erect, and type 3 is indeterminate prostrate. LG score further defines the plant development for each plot with a scale of five values with 1 meaning 100% of the plant is standing erect and 5 meaning 100% of the plant is lying flat on the ground. Hundred seed weight (SW) is the weight in grams of 100 randomly selected undamaged beans from each plot. The final agronomic trait is seed yield (SY) which is measured in grams per plot and then extrapolated to kilograms per hectare.

For the purposes of this study, we elected to use the Durango dataset containing 196 varieties of dry beans. Due to variability across the four locations, we assessed each location separately, and for simplicity, we chose to use just one location as the basis of our simulation study. The North Dakota data was selected for use. We also decided to look at SY as our response variable because it was continuous and is the agronomic trait that may be most useful to a bean breeder or a farmer. It should be noted that all of the cross-check varieties were eliminated from the dataset to ensure we were only looking at Durango beans. It should also be mentioned that North Dakota had three different genotypes that did not have a value for SY in either replicate due to poor adaptation and photoperiod sensitivity, so we are dealing with a dataset of 193 different varieties instead of 196.

2.2. General Methods: Real Data Analysis

Using the data from North Dakota, we wanted to see if variety had a significant effect on SY. In other words, we wanted to see whether or not the different genotypes of beans were performing the same in North Dakota. Tests were completed to check for normality of the SY

North Dakota data, and then an ANOVA test was performed to find out if differences between varieties existed. See Table A1 in Appendix A for an example of SAS ANOVA output. Dunnett's test was run using two different lines as controls. The first time, genotype UI-126 was used as the control, and the second test was run with INTA Precoz as the control. These two lines had the highest seed yield and the lowest seed yield respectively. We wanted to see how many other varieties had different seed yields from these two lines.

Following Dunnett's test, the FDR method analysis was performed on this data. In order to compare Dunnett's test to FDR directly using SAS, we had to run all pairwise comparison tests and then extract only the p-values of comparisons in relation to our control. We again used UI-126 and INTA Precoz as control varieties. We ran all pairwise comparisons (18,528), and then found the 192 comparisons in relation to each control. After obtaining these 192 p-values, we ran the FDR analysis on the data. See Appendix A for the SAS code used for the real data analysis.

2.3. General Methods: Simulation Study

After performing these tests on the real data, we set up a simulation to 1) ensure our dataset was viable and 2) confirm our results from the real data analysis. All data analysis was run using SAS version 9.3 (SAS Institute Inc., 2011). In the first stage of our simulation study, we used the mean and standard deviation values of SY from the North Dakota dataset to simulate a new dataset with ten varieties. We then wanted to assess type I error overall to make sure our dataset was reasonable, and we also wanted to look at power. In stage two of the simulation study, we created a dataset with 30 different varieties in order to see the impact of a larger number of levels. Chapter 3 outlines the simulation study in detail.

CHAPTER 3. SIMULATION STUDY

3.1. Stage One: Type I Error, Power, and False Discovery Rate Analysis

The main goal of the simulation study is to compare Dunnett's test to the FDR method for power and type I error analysis. Based on the research presented in Chapter 1, we expect the FDR method to have a higher power than Dunnett's test, but we are uncertain about how type I error will be affected. We expect Dunnett's test to control the type I error rate at a level of 0.05, but are not sure if the increase in power of the FDR method is a result of an increase in type I error.

We first looked at ten simulated varieties for simplicity. The mean SY value from the North Dakota data was found to be 1,763.18, and the standard deviation for SY was 438.96. These two values were used to simulate SY data for ten varieties of beans. For the tests of type I error and power, we simulated five replicates and then ten replicates for each variety, because with only ten varieties, we wanted more than two replications for each (as in the original dataset). SAS code for stage one can be found in Appendix B.

3.1.1. Type I Error. To assess the type I error of ten varieties, we simulated all of them coming from the same population (mean 1763.18, s.d. 438.96). We ran an ANOVA and flagged samples when the overall F value was significant ($p \leq 0.05$) to find the overall type I error. We expect the ANOVA to reject the null hypothesis of no variety effect about five percent of the time. The simulation had 10,000 samples, so we expect about 500 of the overall ANOVAs to yield rejections of the null hypothesis.

In order to assess the type I error of Dunnett's test, we ran the post hoc test on the ANOVA and looked within each sample of the simulation to see if any of the nine pairwise comparisons to the control were significant. If any pairwise test out of the nine had a significant p-value (signifying a rejection of the null hypothesis), we marked that sample as significant and moved to the next one. At the end we had 10,000 evaluated samples for Dunnett's test. Again, we expect Dunnett's to control type I error at a rate of 0.05, so we expect significance in about 500 of the samples. It should be noted that in this simulation, the control value for Dunnett's test was just the first variety for each sample.

Finally we used this dataset, simulating ten varieties from the same population, to look at the type I error of the FDR method. Like with Dunnett's test, the control was just the first simulated variety. Similar to what we did with the real data, all pairwise comparisons were taken for the ten varieties and then only the p-values for the tests between variety one and each other variety were taken out for the FDR test. Then, FDR analysis was run and significant FDR p-values were flagged for each sample, as they were in Dunnett's test. As stated before, we were uncertain of what the expectation of type I error might be with the FDR method, but we wanted to ensure it was being controlled and that a potential increase in power was not solely due to an increase in type I error rate.

3.1.2. Rejection Percentage. Next, we assessed the percentage of rejection for both Dunnett's test and the FDR method, as well as by using the overall ANOVA. To find this percentage, we simulated some varieties coming from the same population with the North Dakota mean and standard deviation, and other varieties coming from a population four standard deviations away from the North Dakota mean. The standard deviation was maintained through this study, so we are only looking into how well the two post hoc tests detect a difference in means. One group had a mean of 1,763.18 and the other group had a mean of 3,519. With an effect size of four standard deviations, we expect to find at least some differences in the overall ANOVA and also in each of the post hoc tests. The percentage of overall ANOVA rejection was assessed by looking at the ANOVA table results and flagging any with significant ($p \leq 0.05$) overall F-values. Post hoc rejection percentages were assessed by looking at the number of rejections of the null hypothesis, or significant p-values, out of all 90,000 tests (10,000 samples each with nine comparisons). We tried all combinations of groups with the ten simulated varieties, performing each test with five replicates and then again with ten replicates of each variety.

One vs. Nine. The first analysis we did was looking at one control variety coming from the population with mean 3,519 and nine other varieties coming from the population with mean 1,763.18. We will call these the high mean group and North Dakota mean group respectively. In other words, we simulated one control treatment and nine other treatments each with a mean different from the control. We expect to detect this difference all of the time, since all of the varieties are from a different population than the control, and we are interested in how Dunnett's test and the FDR method perform with regards to rejection percentage.

Nine vs. One. The next way we wanted to look at the data was with nine varieties simulated from the high mean group and only one variety from the North Dakota mean group. Our control will always come from the high mean group, so in this case we are simulating one control and eight other varieties from the same population and only one variety from a different population. We expect the overall rejection rate for nine vs. one to be similar to the one vs. nine case; the ANOVA looks at the ten varieties and sees two groups: one group with nine varieties and the other group with one variety. Therefore, it does not matter which mean group the variety is coming from when assessing the overall ANOVA rejection. The rejection percentages of both Dunnett's test and the FDR method are likely to be low in this case since the control mean is coming from a population along with eight other varieties and we are trying to detect a difference between the control and any other variety.

Five vs. Five. The third combination of simulated varieties we chose to look at was an equal group case. We simulated five varieties coming from the high mean group – one control and four other varieties – and five coming from the North Dakota mean group. We again assessed the overall rejection percentage and the percentage for each post hoc test.

Based on the results obtained from these three scenarios, we decided to continue with all different combinations of simulating data from the high mean group and the North Dakota mean group. The first group always contains the control and any additional varieties being simulated from the high mean population and the second group contains varieties from the North Dakota mean group. Rejection percentages for the overall ANOVA, Dunnett's test, and the FDR method were assessed on each of the remaining combinations: two vs. eight; eight vs. two; three vs. seven; seven vs. three; four vs. six; and six vs. four.

3.1.3. False Discovery Rate and Power Analysis. With the rejection analysis set up to evaluate all different comparisons of the ten varieties, we also know which discoveries we are expecting to find. For instance, in the five vs. five case we know that four varieties are coming from the same population as the control and five others are coming from a different population. We expect to "discover", or flag as significant, the five varieties that are different, and we hope to not "discover" the varieties that come from the same population as the control. We can look into the rate of false discoveries (incorrect rejections of varieties the same as the control) and power (correct rejections of varieties different from the control) of these comparisons. To avoid confusion between this analysis and the FDR method used in post hoc testing, we will refer to this analysis as the “false discovery rate” and not use an abbreviation, whereas the post hoc test will be referred to as “the FDR method” or “the FDR test” for the rest of the paper.

Because each comparison has a different number of simulated varieties that are the same as the control, the calculations for false discovery rate and power will differ depending on what comparison we are making. In the one vs. nine case, we expect all varieties to show significance from the control, so we can assess the power to detect these nine varieties. Since we have no varieties coming from the same population as the control, we cannot compute the false discovery rate for this comparison.

For all of the other comparisons, we compute the false discovery rate by using frequency tables of the flagged values. If a variety comes from the same population as the control and is significant, this is a false discovery; we can count all of the times when this occurs and then divide by the number of varieties that are the same as the control multiplied by 10,000. For example, in the three vs. seven case, we have two varieties the same as the control, so for varieties 2 and 3 we would add up all of the significance flags and divide that number by 20,000. This applies to all of the comparisons with at least one variety coming from the same population as the control.

We can also assess the power in a similar way. We know the discoveries that we want to detect are from the simulated varieties coming from the North Dakota mean group. We expect high counts on the significance flags for these varieties. To determine the power, we add up all of the rejections and divide by the number of varieties simulated from the North Dakota mean group multiplied by 10,000. Again for demonstration we use the three vs. seven scenario. We add up all of the significance flags for varieties 4-10 then divide by 70,000. We can find this value for all of the comparisons, including the one vs. nine case where we expect all of the discoveries to be true discoveries and therefore significant.

3.2. Stage Two: Extension of Stage One to Include More Varieties

Given that our real dataset contains 196 different varieties, we wanted to implement a second stage of simulations with more than ten varieties being simulated. We were interested if there would be a change in results with 30 varieties in comparison to the ten in stage one. For this part of the simulation study, we decided to look at three different comparisons: one vs. twenty-nine; twenty-nine vs. one; and fifteen vs. fifteen. We kept the means for the different groups the same as before (an effect size of four standard deviations) and simulated a dataset with two replicates of each of the 30 varieties, like in our real data. We then assessed type I error and rejection percentages for each post hoc test in the same manner as with ten varieties. False discovery rate and power analysis was also performed as described in the ten variety case above. The SAS code for stage two can be found in Appendix C. Results and discussion of the real data analysis and simulation study can be found in Chapter 4.

CHAPTER 4. RESULTS AND DISCUSSION

4.1. Real Data Analysis

As stated in Chapter 2, North Dakota data was used to test the effects of variety on SY. A test for normality of the data was run to make sure this subset of data did not violate the normality assumption. The Kolmogorov-Smirnov test for normality had a p-value > 0.15. This means we do not reject the null hypothesis that the data is normal. See Figure A1 in Appendix A for a histogram of the North Dakota SY data. An ANOVA test was then run on the North Dakota SY data to test for an effect of variety; see Table A1 in Appendix A for this ANOVA table. The p-value for this test was highly significant at p < 0.0001 (F=3.68, 192 df). This merited the post hoc testing on the dataset.

After finding the ANOVA significant, we found the mean seed yield for each variety. Variety UI-126 had the highest mean SY and variety INTA Precoz had the lowest mean SY. These two varieties were used as the control varieties for Dunnett's test and the FDR method. Using Dunnett's test, 109 varieties were found to be different from UI-126 and 128 varieties were different from INTA Precoz. The FDR test found 171 varieties to be different from UI-126 and 185 varieties different from INTA Precoz. 192 total comparisons were made for each control. We can see that the FDR method is finding more significant differences than Dunnett's test. This is expected since the FDR method is known to have higher power than other FWE rate tests. Based off of these results, we now will look at the simulation study to see if the tests perform as expected while controlling effect size and number of replicates.

4.2. Simulation Stage One

4.2.1. Type I Error. Before starting the comparison simulations, an overall test of type I error was found for both the dataset with five replicates and the dataset with ten replicates. We expect the overall ANOVA to be rejected about five percent of the time when the varieties are all coming from the same population. Ten varieties were simulated from the North Dakota mean population and an ANOVA was run on each of the 10,000 samples. When the varieties had five replicates, 478 of the ANOVA tests were marked as significant. With ten replicates, 511 ANOVAs had significant overall F-values. Both of these numbers are right around 500, which is the expected number of rejections if we are controlling type I error at the 0.05 level.

We also expect the type I error of Dunnett's test to be about five percent. With five replicates, we found 480 of the 10,000 samples to have at least one of the nine pairwise comparisons being flagged as significant. With ten replicates we found 481 of the samples to have at least one rejection of the null hypothesis. These are both right around 500 which is what we would expect for Dunnett's test.

The FDR method was also found to control the type I error rate at a level less than five percent. Out of the 10,000 samples, 401 had significance flags on at least one of the nine comparisons when five replicates were present. When ten replicates were used, the type I error rate for the FDR method was 4.15%. These results mean that an increase in power is not due to any increase in the type I error rate of the FDR test.

4.2.2. Rejection Percentage. The rejection percentage analysis was run for Dunnett's test, the FDR test, and the overall ANOVA with five and ten replicates and an effect size of four standard deviations. ANOVA rejection was high in both cases, which makes sense because of the large difference in means of the two groups. Results from the simulation with five replicates can be found in Table 2, and Table 3 displays the results from the trial done with ten replicates.

As we can see on both Tables 2 and 3, the FDR method has a higher percentage of rejection in comparison to Dunnett's test for all nine comparisons being made. The difference between these percentages is not extreme, with the largest difference being 1.32 percent with five replicates and 1.29 percent with ten replicates, but the FDR method is consistent in having a higher rejection rate than Dunnett's test.

Table 2: Rejection Analysis Results for Five Replicates

Combination	Dunnett's Test†	FDR Method†	ANOVA††
1 vs. 9	99.95	99.99	100
2 vs. 8	88.94	89.41	100
3 vs. 7	77.91	78.80	100
4 vs. 6	66.91	68.03	100
5 vs. 5	55.88	57.07	100
6 vs. 4	44.87	46.19	100
7 vs. 3	33.85	35.10	100
8 vs. 2	22.81	23.78	100
9 vs. 1	11.75	12.46	99.997

†Note: Values are expressed as percentages taken out of 90,000.
††Note: Values are expressed as percentages taken out of 10,000.

Table 3: Rejection Analysis Results for Ten Replicates

Combination	Dunnett's Test†	FDR Method†	ANOVA††
1 vs. 9	100	100	100
2 vs. 8	88.97	89.45	100
3 vs. 7	77.95	78.77	100
4 vs. 6	66.94	68.04	100
5 vs. 5	55.89	57.09	100
6 vs. 4	44.85	46.12	100
7 vs. 3	33.87	35.16	100
8 vs. 2	22.76	23.72	100
9 vs. 1	11.76	12.41	100

†Note: Values are expressed as percentages taken out of 90,000.
††Note: Values are expressed as percentages taken out of 10,000.

These rejection values also seem to make sense in relation to which comparison we are making. For instance, in the one versus nine case, we only have the control coming from the high mean group, so we expect to have a very high percentage of rejection because the comparison is between the control and any other variety. In the two versus eight case, the rejection percentage goes down to around 88% (about 8/9). This makes sense because we have the control variety and one other coming from the high mean group and the other eight coming from the North Dakota mean group, so we expect to find differences between the control and the varieties that are truly different from the control, and not find differences between the ones that are the same. However, when thinking about how we expect the rejection rate to behave with these comparisons, it appears that the FDR method is performing higher than we would expect for each of the tests. We further examine the source of this increase in rejection percentage in the following section.

4.2.3. False Discovery Rate and Power Analysis. We found the false discovery rate and power for each of the post hoc tests as described in Chapter 3. To calculate the false discovery rate, the number of rejections found in the varieties simulated from the same population as the control is divided by *i* x 10,000, where *i* is the number of varieties from the same population as the control. Power is found by summing the rejections of the varieties coming from the North Dakota mean group, and dividing by *j* x 10,000, where *j* is the number of varieties coming from the North Dakota mean group. See the description of Tables B3 and B4 in Appendix B for an example. Table 4 displays the results for each test with five replicates.

Dunnett's test appears to be very conservative in the control of the false discovery rate. Although the FDR method has higher chance of making a false discovery than Dunnett's test in general, it is still controlling for the false discoveries at a rate of five percent. This is typically the level at which we want to control the false discovery rate anyway. It also appears that as the number of varieties coming from the same population as the control group increases, the percentage of false discoveries when using the FDR method decreases. The FDR method's power seems to be slightly better than the Dunnett test power when five replicates are present. However, they are both fairly close to each other. This may be due to the fact that the effect size is large, so both tests are able to detect this difference of four standard deviations.

Table 4: False Discovery Rate and Power for Five Replicates Expressed as Percentages

	False Discovery Rate			Power		
Comparison	Dunnett's Test	FDR Method	*i*	Dunnett's Test	FDR Method	*j*
1 vs. 9	--	--	--	99.95	99.997	9
2 vs. 8	0.760	4.72	1	99.96	99.999	8
3 vs. 7	0.770	4.60	2	99.95	99.996	7
4 vs. 6	0.827	4.09	3	99.95	99.997	6
5 vs. 5	0.775	3.40	4	99.96	100	5
6 vs. 4	0.810	3.16	5	99.95	99.993	4
7 vs. 3	0.780	2.66	6	99.98	99.993	3
8 vs. 2	0.770	2.01	7	99.97	99.980	2
9 vs. 1	0.731	1.52	8	99.94	99.920	1

The false discovery rate for the FDR method is much higher than for Dunnett's test, which is almost certainly a factor when looking at the increase in rejection percentage from Table 2. In Section 4.2.2 we mentioned that the FDR method had a rejection rate that was perhaps higher than we would expect. This is most likely due to the fact that Table 2 is taking into account any significant discoveries (false and true) out of the 90,000 tests. Table 4 shows that the FDR method has a higher number of false discoveries, so this is contributing to the increased rejection rate for the FDR method.

Table 5 shows the false discovery rate and power of the two post hoc tests for each of the nine comparisons with ten replicates. Again, Dunnett's test appears to be very conservative in controlling for false discoveries. In the ten replicate case, the power for both tests is always one hundred percent. While it is a good thing for power to be high, we do not expect it to be this high in most cases. Here, the power of predicting the test is most likely affected by the higher number of replicates and the large effect size. These results are not very helpful in determining which test is better.

Table 5: False Discovery Rate and Power for Ten Replicates Expressed as Percentages

	False Discovery Rate			Power		
Comparison	Dunnett's Test	FDR Method	*i*	Dunnett's Test	FDR Method	*j*
1 vs. 9	--	--	--	100	100	9
2 vs. 8	0.740	5.02	1	100	100	8
3 vs. 7	0.755	4.45	2	100	100	7
4 vs. 6	0.743	4.11	3	100	100	6
5 vs. 5	0.748	3.46	4	100	100	5
6 vs. 4	0.738	3.02	5	100	100	4
7 vs. 3	0.798	2.74	6	100	100	3
8 vs. 2	0.686	1.92	7	100	100	2
9 vs. 1	0.731	1.46	8	100	100	1

Based off of Tables 4 and 5, one could say that in the case with many replicates and a large effect size, Dunnett's test has the same power and a lower false discovery rate than the FDR method, so it is the better test. We also can see in Tables 2 and 3 that Dunnett's test performs how we expect it to, whereas the FDR method has a higher rejection rate than expected. This also points to using Dunnett's test with many replicates. However, in practice, a large number of replicates is not the norm. We will explore further analyses in stage two of the simulation, as well as in Chapter 5, in order to test more realistic numbers of replicates.

4.3. Simulation Stage Two

4.3.1. Type I Error. Thirty varieties, with two replicates, were simulated for stage two analysis. The type I error was assessed on each of the post hoc tests as well as overall using the

ANOVA results. The overall ANOVA error rate was 4.71%. Dunnett's test yielded a type I error rate of 5.08%. Again, we expect these two values to be right around five percent, and with some minor simulation error, these values meet our expectations. The FDR method produced a type I error rate of 3.5%, which again shows that the error is being controlled, albeit conservatively. Since the error is controlled at a rate less than five percent for the FDR method, we conclude that any increase in power is not simply due to an increase in type I error.

4.3.2. Rejection Percentage. The rejection rate was assessed on three different comparisons for stage two of the simulation study. The mean groups were the same as in the ten variety simulation, with an effect size of four standard deviations. Table 6 displays the results. The FDR method is consistent in outperforming Dunnett's test, especially for the one vs. twenty-nine comparison. Similar to the ten variety simulation in stage one, the rejection percentages for both post hoc tests drop as the number of varieties being simulated with the control rises.

Table 6: Rejection Analysis Results for Stage Two

Comparison	Dunnett's Test†	FDR Method†	ANOVA††
1 vs. 29	77.89	96.42	64.04
15 vs. 15	40.29	49.47	100
29 vs. 1	2.97	3.21	64.21

†Note: Values are expressed as percentages taken out of 290,000.
††Note: Values are expressed as percentages taken out of 10,000.

The ANOVA rejection rate shows what seems to be a symmetric curve in how well the test will perform relative to the comparison being made. We expect the ANOVA rejection percentages to be similar for opposite groups. In other words, we expect the ANOVA test to not see a large difference overall in comparing one to twenty-nine and comparing twenty-nine to one. The test just sees two distinct groups. This might also explain why the percentage is very high in the equal groups case. Since each group has fifteen varieties, there will be many varieties that are actually different from each other, and it is easier to detect these differences when the groups are closer in size. When we get to the extremes, it is harder to find significance because many of the varieties are actually the same.

Table 6 shows a significant gain in rejection rate for the FDR method over Dunnett's test in cases where the number of varieties coming from the same population as the control is low. As this number rises, the increase in the FDR method percentage seems to diminish. However, here the FDR method is performing close to what we would expect without going over as it did in stage one, while Dunnett's test is lower than we would expect, especially in the first two comparisons. This large difference in percentage between the two post hoc tests was not seen in stage one of the simulation study which raises some questions about where the difference between simulation stages comes from. Does increasing the number of varieties affect the rejection percentage increase? Is this rate higher due to the smaller number of replicates in stage two? Do the increased number of varieties and the high effect size combine to the benefit of the FDR method? We will attempt to address some of these questions in Chapter 5.

4.3.3. False Discovery Rate and Power Analysis. The number of false discoveries was again taken into account with the 30 simulated varieties. Table 7 displays these results. Dunnett's test is consistently conservative in making any false discoveries. This seems to make sense knowing that Dunnett's test is a fairly conservative FWE rate correction to begin with. The FDR method controls the false discoveries at a rate below five percent also. The number of true discoveries that were accurately found is consistent across all comparisons for Dunnett's test. The FDR method shows far greater power when the comparison has fewer simulated varieties that are the same as the control, but appears to decrease as this number increases. We did not run all 29 different comparisons for this dataset with 30 varieties, but it would be interesting to see where the power for Dunnett's test and the FDR method cross.

Table 7: False Discovery Rate and Power for Stage Two Expressed as Percentages

	False Discovery Rate			Power		
Comparison	Dunnett's Test	FDR Method	*i*	Dunnett's Test	FDR Method	*j*
1 vs. 29	--	--	--	77.89	96.42	29
15 vs. 15	0.303	2.61	14	77.61	93.22	15
29 vs. 1	0.325	0.793	28	76.97	70.78	1

Based off of the results found in Tables 6 and 7, it appears that the FDR method is a more beneficial test when there are a small number of replicates. There remain questions, however, because the number of treatments changed between stages one and two, and so we are not able to directly compare these results. Follow-up research is necessary to make any firm conclusions.

There are many different angles of follow-up research that stem from these results. We will address a few of them in Chapter 5. First, experience suggests that five and ten replicates are unrealistic to expect when it comes to real data collection, so we assess the rejection rate, power, and false discovery rate on a simulated set of ten varieties with two replicates each and an effect size of four standard deviations. We can then directly compare these results to the results found in Tables 2-5. The next piece we wanted to explore was the effect size. While initially we chose a very large effect size to make sure the difference would be detected, one this large is unrealistic. We will explore data with ten varieties, five replicates, and effect sizes of one, two, and three standard deviations.

CHAPTER 5. FURTHER ANALYSIS

While the results in simulation stage one found in Chapter 4 show that the FDR method controls the type I error rate and has slightly higher rejection rate than Dunnett's test, these results do not make a strong case for either method, but lean towards Dunnett's test being more effective with a large number of replicates. This is presumably because the rejection percentage is so high and the effect size is so large that there is really no room to show important differences. The stage two simulation, under a more realistic scenario of two replicates, was more useful in showing that the FDR method yielded a higher rejection rate and power than Dunnett's test, but there are still questions as to where the increase in power came from. There were also more gaps with the 30 varieties, since we did not have the time to run all 29 different comparisons. We wanted to explore some additional tests to see if the FDR method was consistent in outperforming Dunnett's test in the power analysis. We also looked into the false discovery analysis after changing the number of replicates and the effect size.

5.1. Analysis with Two Replicates

Similar to the stage one analysis, ten varieties were simulated from the same two mean groups, and all nine comparisons were analyzed. For this analysis, only two replicates for each variety were simulated. Two replicates were chosen to help address two main concerns that came up in both stages of the simulation: Are five and ten replicates actually reasonable in practice? And does the increase in power in stage two come from the larger number of varieties or the smaller number of replicates? As we stated in Chapter 4, obtaining ten replicates in practice is an unreasonable goal, and even five replicates is not typical, so we decided to use two so we could have a more realistic number. We also chose two replicates to emulate the North Dakota BeanCAP dataset we are working with.

5.1.1. Type I Error. We ran all of the same tests as in stage one of the simulation study. The type I error results were very similar to stage one, and again performed as expected. Overall type I error was 4.91% and Dunnett's error rate was 5.01%; both of these are very close to the five percent mark, the point at which we expect them to control the error rate. The FDR method type I error rate was 3.77% which is again somewhat conservative, but still less than five percent, so we know it is being controlled in some way.

5.1.2. Rejection Percentage. The rejection rate was also assessed for each of the post hoc tests and the overall ANOVA. The results from the rejection analysis are presented in Table 8. Figure 2 displays the rejection rate curves from Table 8. Similar to the results from stage two, the overall ANOVA rejection seems to have a symmetric curve, being highest when the two groups are equal. The FDR method rejection appears to outperform Dunnett's test in all of the comparisons except for when eight varieties are being simulated from the same population as the control. The difference in rejection rate in the nine vs. one case is marginal, but note that the FDR method percentage is lower than Dunnett's in this case.

Table 8: Rejection Analysis Results for Two Replicates

Comparison	Dunnett's Test†	FDR Method†	ANOVA††
1 vs. 9	75.23	92.18	74.68
2 vs. 8	67.48	82.02	95.65
3 vs. 7	58.94	71.03	98.84
4 vs. 6	50.56	60.59	99.63
5 vs. 5	42.57	49.80	99.63
6 vs. 4	33.86	38.88	99.56
7 vs. 3	25.65	28.46	99.06
8 vs. 2	17.40	18.27	95.75
9 vs. 1	9.11	8.82	75.13

†Note: Values are expressed as percentages taken out of 90,000.
††Note: Values are expressed as percentages taken out of 10,000.

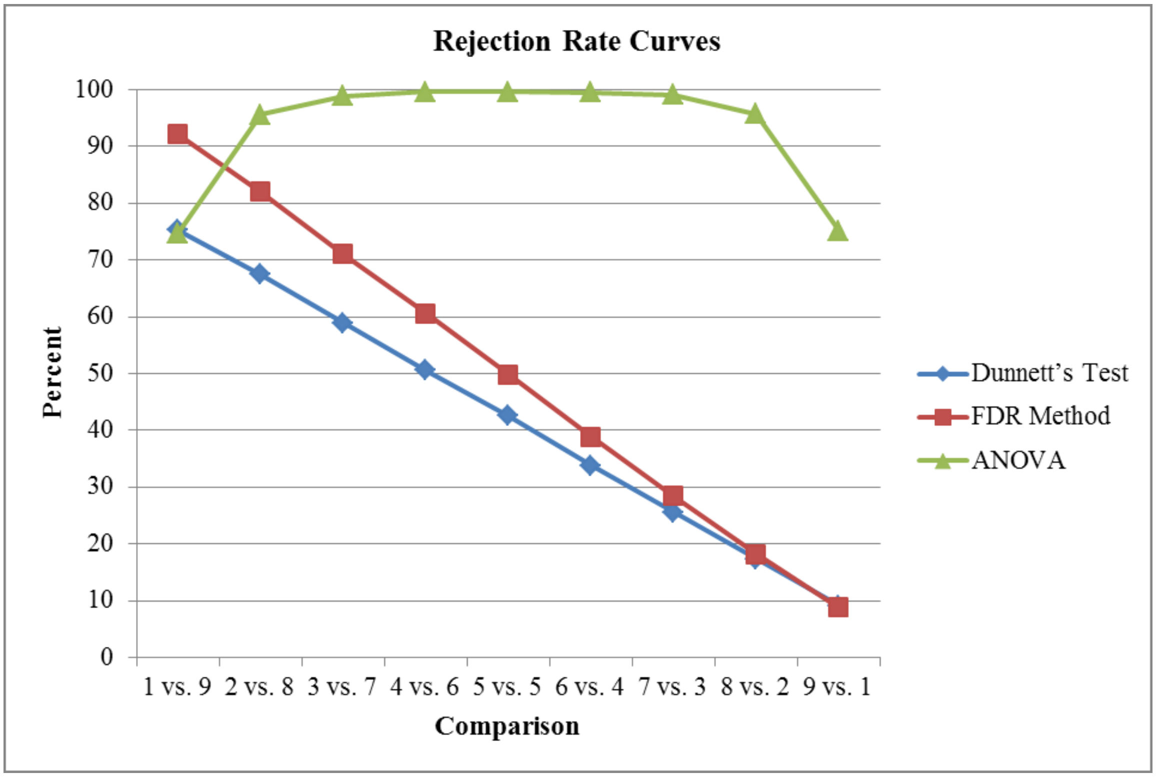

Figure 2: Rejection Rate Curves from Simulations with Two Replicates

5.1.3. False Discovery Rate and Power Analysis. The false discoveries were also flagged as in stage one. Table 9 displays the results of this analysis on the two replicates. The power to detect the true discoveries was very high for the FDR method until the final comparison, nine vs. one, where it dips below Dunnett's power level. Dunnett's test is very consistent in the power level. The rate of detecting false discoveries is below five percent for both post hoc testing methods, where Dunnett's test again is very conservative and fairly consistent, and the FDR method has a higher rate than Dunnett's test. It appears that the trade-off for the higher false discovery rate is an increase in power of over 15% compared to Dunnett's test in the first few comparisons. While this difference diminishes, it seems to suggest that the FDR method is favorable under this set of conditions. See Figure 3 for a visual representation of Table 9.

Table 9: False Discovery Rate and Power for Two Replicates Expressed as Percentages

	False Discovery Rate			Power		
Comparison	Dunnett's Test	FDR Method	*i*	Dunnett's Test	FDR Method	*j*
1 vs. 9	--	--	--	75.23	92.18	9
2 vs. 8	1.02	4.86	1	75.79	91.66	8
3 vs. 7	0.950	4.40	2	75.51	90.07	7
4 vs. 6	0.900	3.86	3	75.40	88.95	6
5 vs. 5	0.895	3.38	4	75.92	86.94	5
6 vs. 4	0.916	3.00	5	75.04	83.74	4
7 vs. 3	0.863	2.43	6	75.21	80.50	3
8 vs. 2	0.877	1.87	7	75.25	75.68	2
9 vs. 1	0.836	1.47	8	75.28	67.68	1

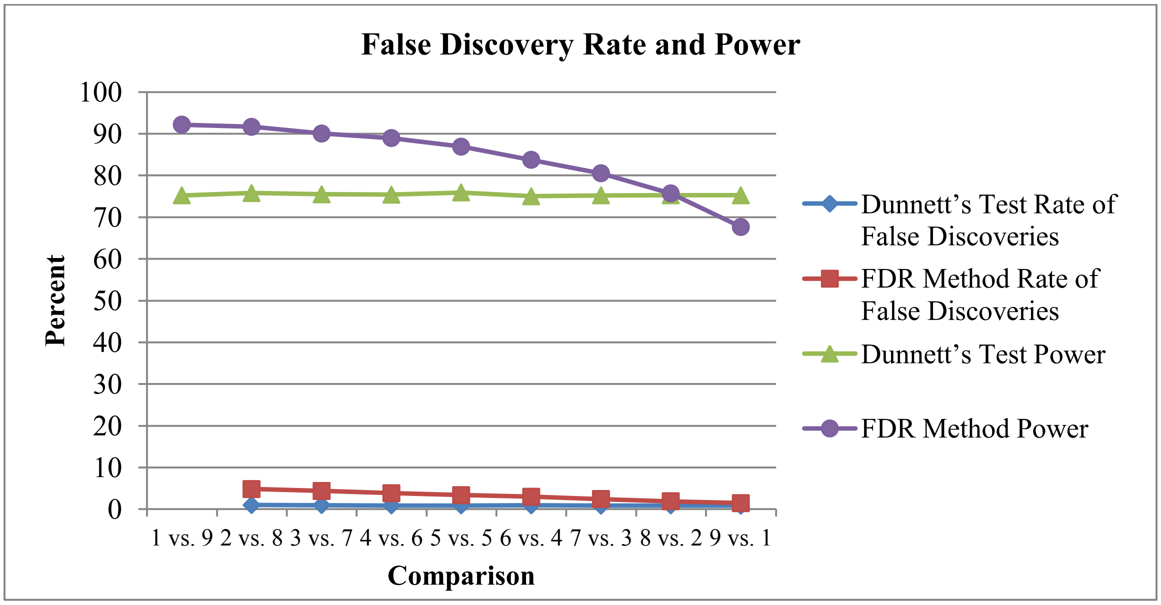

Figure 3: False Discovery Rate and Power Curves for Two Replicates

Overall, it appears that a smaller number of replicates may be a contributing factor to being able to see a difference between the FDR method and Dunnett's test. As also seen in simulation stage two, the FDR method has much higher power than Dunnett's test for most of the comparisons. If using real data with a small number of replicates, the FDR method appears to have an advantage over Dunnett's test when performing comparisons in relation to a control.

5.2. Differing Effect Sizes

The next concern we wanted to address was the large effect size of the simulation study. Initially, we wanted to make sure the effect size was large enough so a difference would be detected. Upon reviewing the results of stage one, it appears that this large difference was indeed detected, at times with 100% power. We realize that an effect size this large is not likely in practice, so we chose to investigate the rejection rate, power, and false discovery rate results on data with ten varieties and effect sizes of one, two, and three standard deviations.

5.2.1. Effect Size of Three Standard Deviations. With an effect size of three standard deviations, the high mean population now is being simulated with a mean of 3,080 and the same standard deviation of 438.96. Five replicates were used with this smaller effect size.

Rejection Percentage. The results from the rejection rate analysis can be found in Table 10 and are expressed visually in Figure 4. With the smaller effect size, the overall ANOVA rejection rate comes down for the two tests comparing one and nine varieties, but we can see that the overall ANOVA rejection is still 100% in all other cases. However, we can see that the overall ANOVA is starting to show the symmetric curve as was seen in the analysis on two replicates. The FDR method rejection rate is again higher than that of Dunnett's test on all of the comparisons, but only marginally so. We also see the FDR method seems to again have a slightly higher rejection rate than expected, as it did with an effect size of four standard deviations. This still suggests using Dunnett's test in cases where the effect size is large.

Table 10: Rejection Analysis Results for Five Replicates, Effect Size of Three SDs

Comparison	Dunnett's Test†	FDR Method†	ANOVA††
1 vs. 9	96.89	99.57	99.26
2 vs. 8	86.22	89.03	100
3 vs. 7	75.47	78.34	100
4 vs. 6	64.78	67.51	100
5 vs. 5	54.09	56.64	100
6 vs. 4	43.47	45.62	100
7 vs. 3	32.81	34.63	100
8 vs. 2	22.11	23.24	100
9 vs. 1	11.42	12.07	99.36

†Note: Values are expressed as percentages taken out of 90,000.
††Note: Values are expressed as percentages taken out of 10,000.

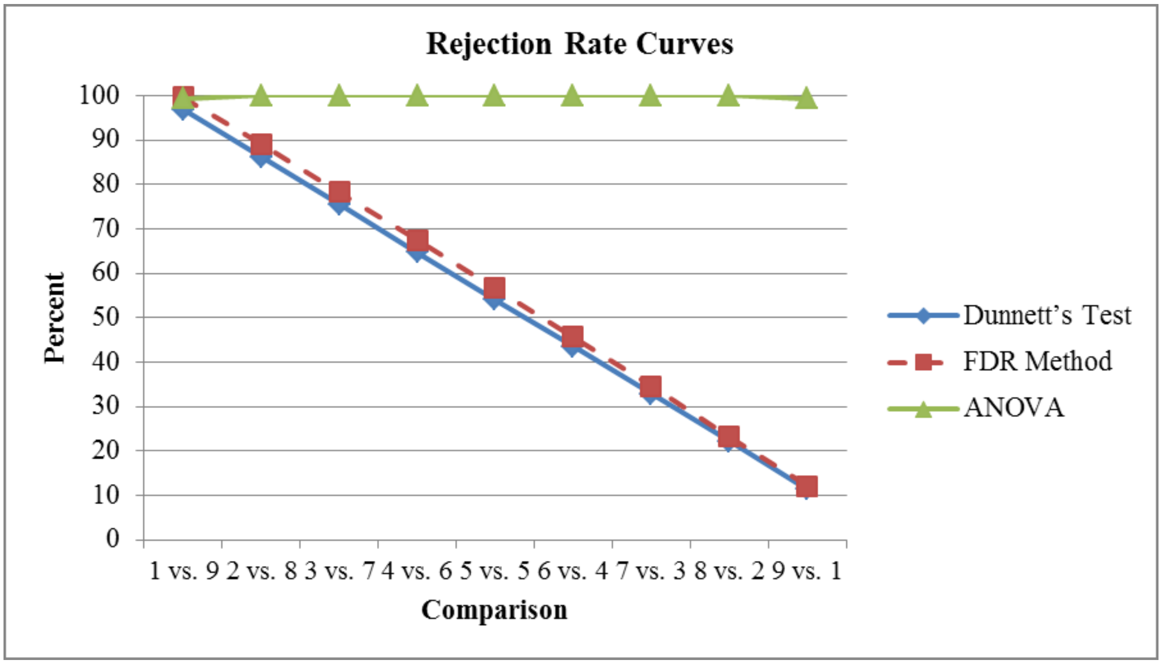

Figure 4: Rejection Rate Curves, Effect Size of Three SDs, Five Replicates

False Discovery Rate and Power Analysis. We also ran the false discovery and power analyses on these data comparisons. Results can be found in Table 11. Similar to the effect size of four standard deviations, we can see that the FDR method power is higher than Dunnett's test, until the last comparison, where it is slightly lower. We can also see that the false discovery rate is being controlled very conservatively for Dunnett's test; the FDR method false discovery rate never exceeds five percent. These results are consistent with the analyses in stages one and two of the simulation study.

Table 11: False Discovery Rate and Power for Five Replicates, Effect Size of Three SDs

	False Discovery Rate			Power		
Comparison	Dunnett's Test	FDR Method	*i*	Dunnett's Test	FDR Method	*j*
1 vs. 9	--	--	--	96.89	99.57	9
2 vs. 8	0.700	4.79	1	96.91	99.57	8
3 vs. 7	0.725	4.61	2	96.83	99.41	7
4 vs. 6	0.777	3.94	3	96.79	99.29	6
5 vs. 5	0.765	3.60	4	96.76	99.07	5
6 vs. 4	0.744	3.01	5	96.87	98.88	4
7 vs. 3	0.800	2.65	6	96.82	98.59	3
8 vs. 2	0.753	1.96	7	96.87	97.73	2
9 vs. 1	0.733	1.55	8	96.88	96.27	1

An effect size of three standard deviations is still quite high, or we expect to find many differences since the means are far apart. Although the power is slightly better for the FDR method with an effect size of three standard deviations, the results from Tables 10 and 11 still suggest that Dunnett's test is favorable overall. We now look at an effect size of two standard deviations to see if the preferred method changes.

5.2.2. Effect Size of Two Standard Deviations. With an effect size of two standard deviations, the high mean group now is being sampled from a population with mean 2,641 and standard deviation 438.96. We again used five replicates for each variety in the simulation.

Rejection Percentage. Results for the rejection rate analysis can be found in Table 12 and Figure 5. Again we see the symmetric curve forming for the overall ANOVA rejection rate, with the highest rejection percentage coming from the equal groups comparison (5 vs. 5). We can also see that the FDR method has a higher rejection rate than Dunnett's test in all of the comparisons. All of the rejection rates are lower than the expected value, but this is due to the effect size being only two standard deviations. The FDR method does a much better job of detecting a difference, especially when the number of varieties being simulated from the high mean group is low.

Table 12: Rejection Analysis Results for Five Replicates, Effect Size of Two SDs

Comparison	Dunnett's Test†	FDR Method†	ANOVA††
1 vs. 9	63.91	82.91	77.07
2 vs. 8	56.62	72.95	97.00
3 vs. 7	49.67	62.99	99.44
4 vs. 6	42.59	52.96	99.83
5 vs. 5	35.68	43.41	99.88
6 vs. 4	28.75	33.93	99.84
7 vs. 3	21.76	24.71	99.35
8 vs. 2	14.74	15.94	97.07
9 vs. 1	7.71	7.73	77.11

†Note: Values are expressed as percentages taken out of 90,000.
††Note: Values are expressed as percentages taken out of 10,000.

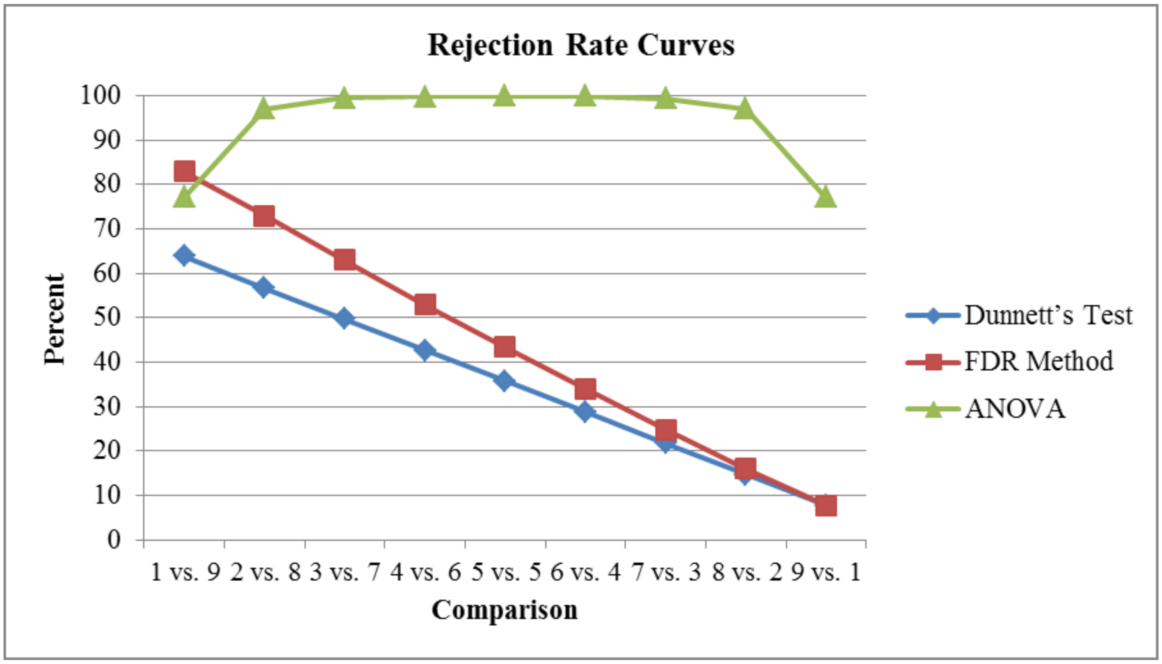

Figure 5: Rejection Rate Curves, Effect Size of Two SDs, Five Replicates

False Discovery Rate and Power Analysis. The false discovery rate and power results for an effect size of two standard deviations can be found in Table 13 and Figure 6. As with all of the previous results, Dunnett's test has a very conservative and consistent false discovery rate while maintaining a consistent power. The false discovery rate of the FDR method is again being controlled at a rate less than five percent while the power shows a significant increase over that of Dunnett's test. This increase diminishes as the number of varieties coming from the North Dakota mean group decreases, but the FDR method power is only lower than Dunnett's power in the final comparison. These results are consistent with the results from stage two in Chapter 4 and the simulation with two replicates from the beginning of Chapter 5.

Table 13: False Discovery Rate and Power for Five Replicates, Effect Size of Two SDs

	False Discovery Rate			Power		
Comparison	Dunnett's Test	FDR Method	*i*	Dunnett's Test	FDR Method	*j*
1 vs. 9	--	--	--	63.91	82.91	9
2 vs. 8	0.640	3.73	1	63.62	81.61	8
3 vs. 7	0.835	3.24	2	63.63	80.10	7
4 vs. 6	0.740	3.07	3	63.52	77.91	6
5 vs. 5	0.750	2.84	4	63.63	75.87	5
6 vs. 4	0.760	2.30	5	63.74	73.46	4
7 vs. 3	0.787	2.01	6	63.71	70.12	3
8 vs. 2	0.777	1.72	7	63.61	65.72	2
9 vs. 1	0.696	1.26	8	63.82	59.48	1

With an effect size of two standard deviations, we find the FDR method favorable because of the increase in power and the false discovery rate being maintained. The rejection rate is also closer to what we would expect using the FDR method. This implies that in cases where the effect size is smaller, the FDR method is preferred.

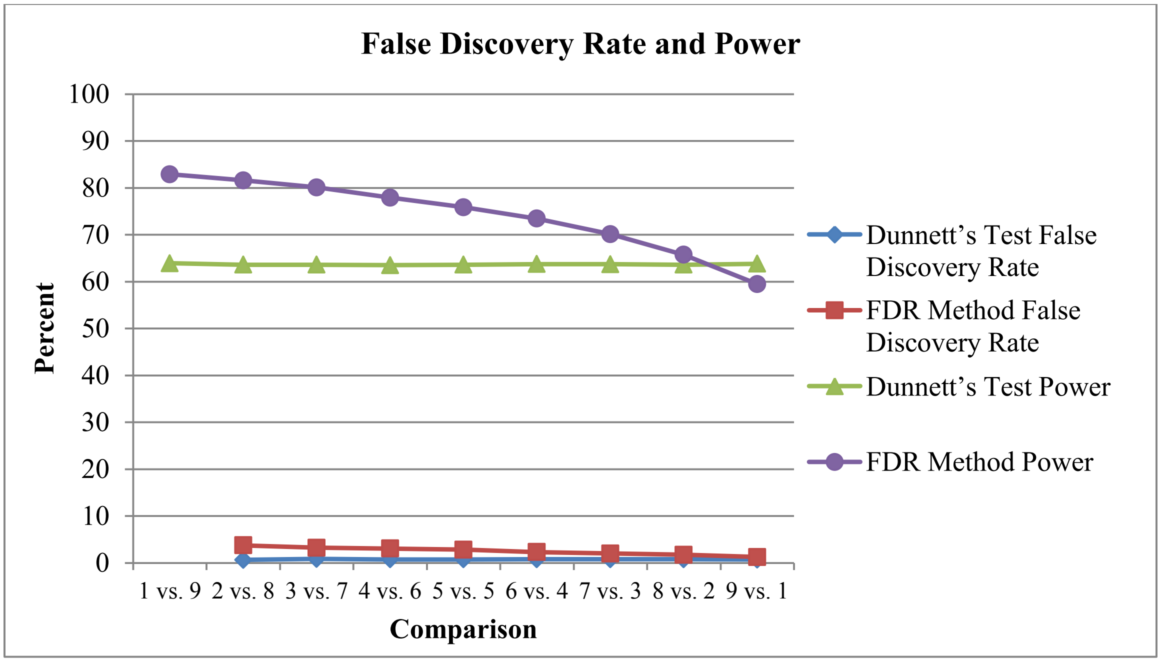

Figure 6: False Discovery Rate and Power Curves for an Effect Size of Two SDs

5.2.3. Effect Size of One Standard Deviation. The final effect size we wanted to assess was that of one standard deviation. This is perhaps the most likely scenario when it comes to real data, so we are interested in how the two post hoc tests compare. The high mean group is being sampled from a population with mean 2,202 and the same standard deviation of 438.96. Five replicates were assessed in accordance with the previous two tests.

Rejection Percentage. The rejection rate analysis results can be found in Table 14 and Figure 7. The overall ANOVA rejection rate again is highest in the five vs. five comparison and has a symmetric curve. While the rejection rates for both post hoc tests are very low, which is to be expected with the small effect size, we can see that the FDR method has a higher rejection rate than Dunnett's test. These results are consistent with all of the previous results and they seem to make sense within the context of the small effect size.

Table 14: Rejection Analysis Results for Five Replicates, Effect Size of One SD

Comparison	Dunnett's Test†	FDR Method†	ANOVA††
1 vs. 9	12.43	21.12	20.81
2 vs. 8	11.23	17.79	37.41
3 vs. 7	9.73	15.01	49.10
4 vs. 6	8.76	12.88	56.03
5 vs. 5	7.20	9.80	57.94
6 vs. 4	5.90	7.60	55.09
7 vs. 3	4.66	5.62	48.01
8 vs. 2	3.50	3.93	37.82
9 vs. 1	2.21	2.50	20.74

†Note: Values are expressed as percentages taken out of 90,000.
††Note: Values are expressed as percentages taken out of 10,000.

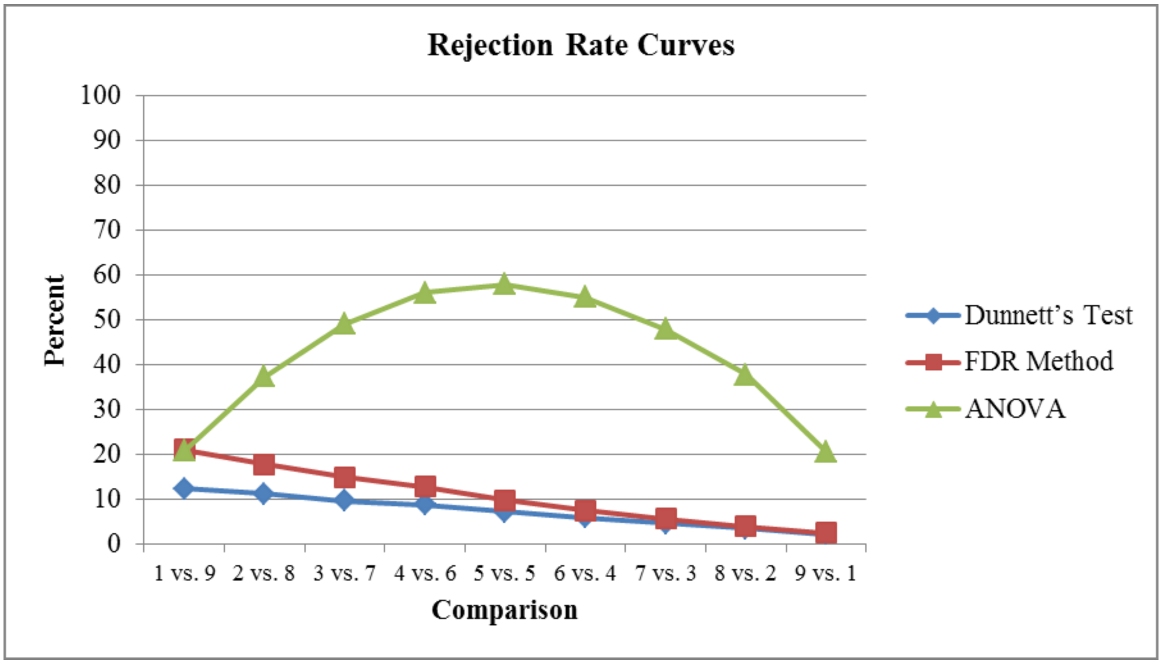

Figure 7: Rejection Rate Curves, Effect Size of One SD, Five Replicates

False Discovery Rate and Power Analysis. The false discovery rate and power analysis was also run for the simulation with an effect size of one standard deviation. These results can be found in Table 15 and Figure 8. We see that Dunnett's test is very conservative in making false discoveries and also very consistent in terms of power. The FDR method shows an increase in power of almost nine percent in the first comparison and then this increase diminishes along with the comparisons. Dunnett's test has higher power than the FDR method in the final two comparisons. These results are consistent with what we saw in the previous simulations and they make sense with the small effect size being tested.

Table 15: False Discovery Rate and Power for Five Replicates, Effect Size of One SD

	False Discovery Rate			Power		
Comparison	Dunnett's Test	FDR Method	*i*	Dunnett's Test	FDR Method	*j*
1 vs. 9	--	--	--	12.43	21.12	9
2 vs. 8	0.810	2.39	1	12.53	19.71	8
3 vs. 7	0.730	2.06	2	12.31	18.72	7
4 vs. 6	0.707	1.89	3	12.78	18.38	6
5 vs. 5	0.745	1.70	4	12.37	16.29	5
6 vs. 4	0.756	1.61	5	12.32	15.09	4
7 vs. 3	0.778	1.60	6	12.42	13.67	3
8 vs. 2	0.824	1.41	7	12.89	12.77	2
9 vs. 1	0.878	1.41	8	12.91	11.26	1

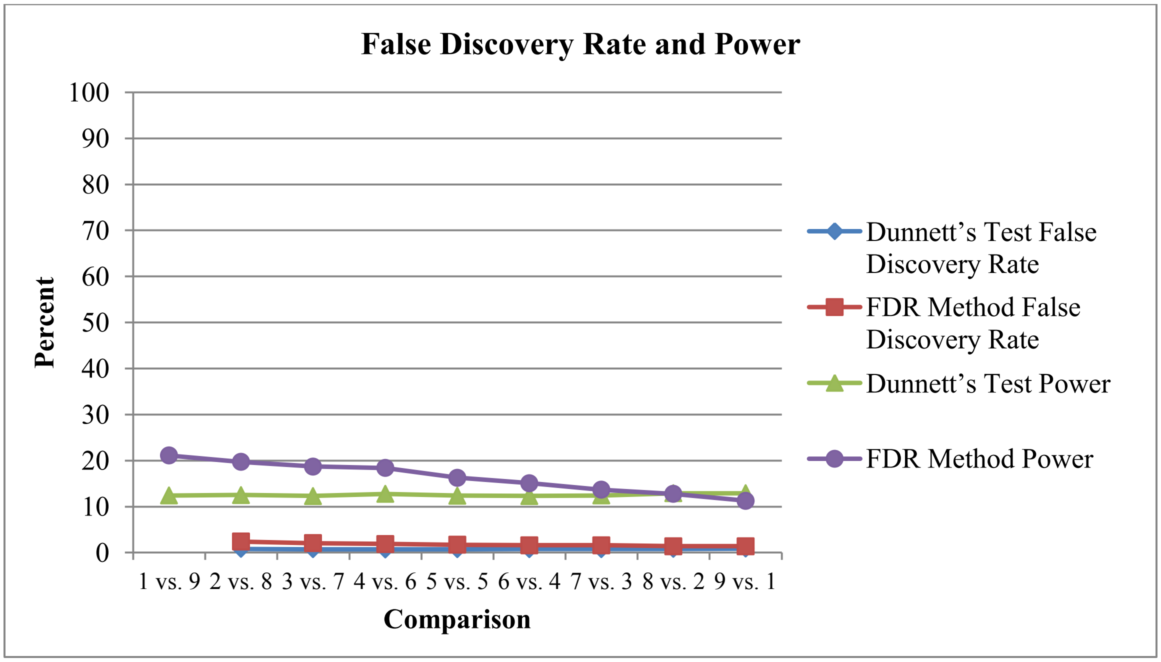

Figure 8: False Discovery Rate and Power Curves for an Effect Size of One SD

5.3. Further Analysis Conclusions

The results presented in this section support the use of the FDR test when the number of replicates and/or the effect size is small. With a small number of replicates an increase in power of up to 17 percent was found for the FDR method. The false discovery rate was controlled at a rate less than five percent for the FDR method, and the rejection rates were higher overall. Even with a high effect size, we saw the benefits of the FDR method with a small number of replicates.

When the effect size decreased from four standard deviations to three standard deviations (using five replicates), the simulation still seemed to favor Dunnett's test. When the effect size went down to two standard deviations, we could see the benefits of the FDR method. With an effect size of one standard deviation, we can also see that the FDR method is preferable.

There are many other ways to follow up with this data to address the concerns that were raised in Chapter 4 and to get a better idea of how the two post hoc tests perform. Due to the time constraints on this research we were not able to explore all of the different options for follow up, but we feel the further analysis discussed in this chapter has helped to fill in some gaps and answers some questions that came about in Chapter 4. Overall, it appears that the FDR method is advantageous in terms of power while also controlling for the type I error rate when comparing each treatment to a control, especially in situations when there are a small number of replicates and when effect size is small.

CHAPTER 6. CONCLUSIONS

Based on the results from the simulation study and follow up research, there is evidence that the FDR method performs better than Dunnett's test in realistic testing situations. While there were a few instances where this was not the case, Dunnett's test only had marginally better results, which would support using the FDR method in most cases, provided that any increase in computing power could be accounted for when using the FDR procedure. The FDR method was especially powerful with few replicates and small effect sizes.

There are many other tests that would be interesting to explore as an extension of this research. We did not look into datasets with more than 30 varieties; as our original dataset had 193 different varieties, it would be interesting to see how these two tests performed with an increased number of treatments. Due to the amount of computing time necessary for these larger comparisons, they were not feasible within the scope of this study. We speculate that with a larger number of varieties (with two replicates), the FDR method will continue to outperform Dunnett's test.

Another vein of research that might stem from this paper would be to use Storey's positive FDR and q-values. This method uses m_0 instead of m in calculations, but requires an estimation of m. See Storey (2002) for more information on these methods.

Based on the real data and simulation studies assessed in this research, we find the FDR method to perform well in both power and controlling for type I error. The FDR method outperformed Dunnett's test in many of the comparisons made here. Therefore, the FDR method might be preferable to using Dunnett's test when comparing many treatments to one control, especially in realistic data situations where there are a limited number of replicates or a small effect size.

REFERENCES

Basford, K. E., & Tukey, J. W. (1999). *Graphical analysis of multiresponse data: Illustrated with a plant breeding trial.* Boca Raton, FL: Chapman & Hall/CRC.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B, 57*(1), 289-300.

Benjamini, Y., & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics, 29*(4), 1165-1188.

Dunnett, C. W. (1955). A multiple comparison procedure for comparing several treatments with a control. *Journal of the American Statistical Association, 50*(272), 1096-1121.

Dunnett, C. W. (1964). New tables for multiple comparisons with a control. *Biometrics, 20*(3), 482-491.

Howell, D. C. (2002). *Statistical methods for psychology* (5th ed.). Pacific Grove, CA: Duxbury.

Lazar, N. (2012). Multiplicity control in large data sets presents new challenges and opportunities. *Chance, 25*(2), 37-40.

McClean, P. E., Garden-Robinson, J., Johnson, C., Osorno, J., Slator, B., Kelly, J., …, Miklas, P. (n.d.). Project summary. Retrieved from http://www.beancap.org/_pdf/BeanCAP_Project_Description.pdf

Montgomery, D. C. (2009). *Design and analysis of experiments* (7th ed.). Hoboken, NJ: John Wiley & Sons, Inc.

SAS Institute Inc. 2011. *SAS/STAT 9.3 User's Guide*. Cary, NC: SAS Institute Inc.

Shaffer, J. P. (1995). Multiple hypothesis testing: A review. *Annual Review of Psychology, 46*, 561-584.

Singh, S.P., Gepts, P., &Debouck, D.G. (1991). Races of common bean (Phaseolus vulgaris, Fabaceae). *Economic Botany, 45*(3), 379-396.

Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B, 64*(3), 479-498.

Storey, J. D. (2003). The positive false discovery rate: A Bayesian interpretation and the q-value. *The Annals of Statistics, 31*(6), 2013-2035.

Westfall, P. H., & Young, S. S. (1993). *Resampling-based multiple testing: Examples and methods for* p-*value adjustment.* New York, NY: A Wiley Interscience Publication.

APPENDIX A. SAS CODE AND OUTPUT - REAL DATA ANALYSIS

```
title1 'Durango Model';

data durango;
  infile 'E:\Thesis\Thesis_Durango.csv' firstobs=2 obs=1601 dlm=',';
  input Block Rep Location $ ID Variety: $23. MC $ DF DM PH LG GH SW SY;
  if Variety in ('Vista' 'Eclipse' 'Zorro' 'Ensign') then delete;
  run;

proc print data=durango;
  title2 'Verify Data - Durango';
  run;

proc univariate data=durango;
       where Location="ND";
       var SY;
       histogram SY / normal midpoints=0 to 3500 by 250;
       title2 'Univariate SY';
run;
```

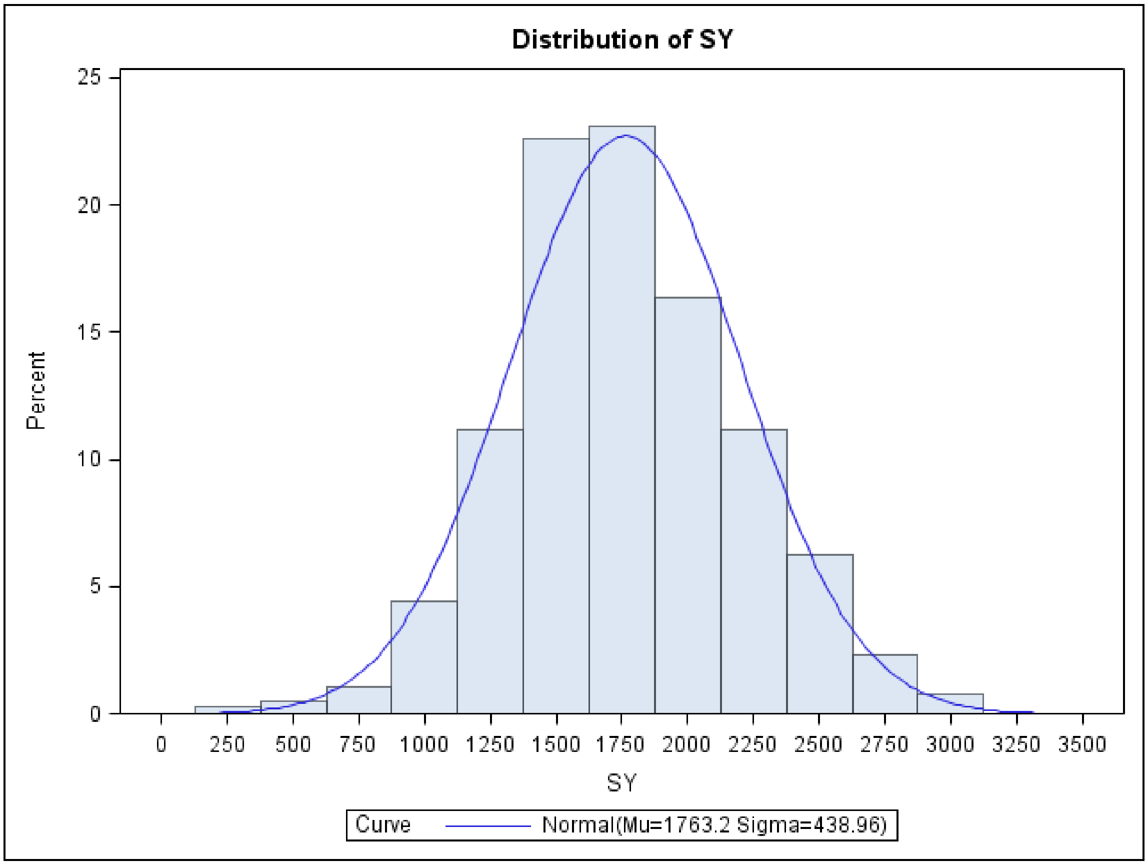

Figure A1: Histogram of Seed Yield Data for North Dakota

```
ods graphics / labelmax=56400;
ods output diff=pvalsND;
proc glm data=durango;
       where Location="ND";
       class Variety;
       model SY = Variety;
       means Variety / Dunnett ("UI-126");
       means Variety / Dunnett ("INTAPrecoz");
       lsmeans Variety / pdiff;
       title2 "ANOVA for SY - Dunnett";
run;
quit;
```

Table A1: Example of ANOVA Table Output from SAS

```
                                          Durango Model
                                      ANOVA for SY - Dunnett
                                        The GLM Procedure

Dependent Variable: SY

                                               Sum of
       Source                      DF         Squares     Mean Square    F Value    Pr > F

       Model                      192     58187436.91       303059.57       3.68    <.0001
       Error                      192     15804115.00        82313.10
       Corrected Total            384     73991551.91

                        R-Square     Coeff Var      Root MSE       SY Mean
                        0.786406      16.27184      286.9026      1763.184

       Source                      DF     Type III SS     Mean Square    F Value    Pr > F
       Variety                    192     58187436.91       303059.57       3.68    <.0001
```

```
/* Use all pairwise comparisons for FDR */
data pvalND (keep=loc row col raw_p);
       set pvalsND;
       array ps{193} _1-_193;
       r+1;
       do col = r to 193;
              row=r;
              raw_p=ps{col};
              output;
       end;
run;

data pvalND;
       set pvalND;
       if row=col then delete;
run;

/* FDR applied for Dunnett's with UI-126 (id 167) as control - high */
title2 'P-values for ND Highest UI-126';
data pvalNDuppercol;
       set pvalND;
       where col=167;
run;
```

```
data pvalNDupperrow;
       set pvalND;
       where row=167;
run;

data pvalNDupper;
       set pvalNDuppercol pvalNDupperrow;
run;

proc multtest inpvalues=pvalNDupper fdr out=outpNDupper;
run;

data outpNDupper;
       set outpNDupper;
       if fdr_p<=0.05 then flagup=1;
       else flagup=0;
run;

proc freq data=outpNDupper;
       tables flagup;
run;

/* FDR applied for Dunnett's with INTAPrecoz (id 83) as control - low */
       title2 'P-values for ND Lowest Yield - INTA Precoz';
data pvalNDlowercol;
       set pvalND;
       where col=83;
run;

data pvalNDlowerrow;
       set pvalND;
       where row=83;
run;

data pvalNDlower;
       set pvalNDlowercol pvalNDlowerrow;
run;

proc multtest inpvalues=pvalNDlower fdr out=outpNDlower;
run;

data outpNDlower;
       set outpNDlower;
       if fdr_p<=0.05 then flaglow=1;
       else flaglow=0;
run;

proc freq data=outpNDlower;
       tables flaglow;
run;
```

APPENDIX B. SAS CODE AND OUTPUT - STAGE ONE

B.1. Type I Error Rate Code

```
dm 'log; clear; output; clear;';
options ls=90 ps=120 formchar="|----|+|---+=|-/\<>*" nodate nonumber symbolgen mprint;
title1 'Durango Model';
options threads=yes cpucount=8;

libname Thesis 'E:\Thesis\Simulations\';

%macro typeone(seed0,n1,mu1,sd1,n2,samples,out_ds,out_f);

   title2 "Type One Error Simulation &n2 reps - &out_ds";
   data Thesis.gen_data&n2 (keep=sample variety rep Y);

     call streaminit(&seed0);  *** Initialize with desired seed. ***;

     do sample=1 to &samples;
       do variety=1 to &n1;
         do rep=1 to &n2;
             Y=rand('Normal',(&mu1),&sd1); output;
        end;
       end;
     end;
   run;

   proc sort;
       by Sample variety;
   run;

   ods listing close;            *** Toggle standard output off. ***;

   ods output diff=Thesis.&out_ds._pvals overallanova=Thesis.panova&n2 (keep=probf);
   proc glm data=Thesis.gen_data&n2 plots=none; /*Dunnett and overall output*/
        by sample;
        class variety;
        model Y= variety;
        lsmeans variety /pdiff=control("1");
   run;

   ods output diff=Thesis.&out_f;  /*FDR output*/
   proc glm data=Thesis.gen_data&n2 plots=none;
       by sample;
       class Variety;
       model Y=Variety;
       lsmeans variety / pdiff;
   run;

   ods listing;


   data Thesis.panova&n2;
       set Thesis.panova&n2;
       if ProbF >.;
       if ProbF <= 0.05 then ANOVAflag = 1;
       else ANOVAflag=0;
   run;
```

```
/*Dunnett flagging for alpha value */
   proc sort data=Thesis.&out_ds._pvals;
       by sample;
   run;

   data Thesis.&out_ds._pvals (keep=Sample Effect Dependent RowName _1);
       set Thesis.&out_ds._pvals;
       by sample;
       if RowName=1 then delete;
   run;

   data Thesis.&out_ds._pvals;
       retain Dunnflag;
       set Thesis.&out_ds._pvals;
       by sample;
       if first.sample then Dunnflag=0;
       if _1 <= 0.05 then Dunnflag=1;
       if last.sample;
   run;

/*FDR flagging for alpha value*/
   data Thesis.&out_f (keep= sample Effect Dependent RowName _1);
       set Thesis.&out_f;
   run;

   data Thesis.&out_f;
       set Thesis.&out_f;
       if RowName=1 then delete;
       rename _1=raw_p;
   run;

   proc sort data=Thesis.&out_f;
       by sample;
   run;

   proc multtest inpvalues(raw_p)=Thesis.&out_f fdr out=Thesis.out&out_f;
       by sample;
   run;

   proc sort data=Thesis.out&out_f;
       by sample;
   run;

   data Thesis.out&out_f;
       set Thesis.out&out_f;
       retain FDRflag;
       by sample;
       if first.sample then FDRflag=0;
       if fdr_p <= 0.05 then FDRflag=1;
       if last.sample;
   run;

/*Frequency tables for flagged values */
   proc freq data=Thesis.panova&n2;
       tables ANOVAflag;
   run;

   proc freq data=Thesis.&out_ds._pvals;
       tables Dunnflag;
   run;
   proc freq data=Thesis.out&out_f;
       tables FDRflag;
   run;

%mend typeone;

%typeone(0,10,1763.18,438.9605143,5,10000,ND_10VarD5rep, ND_10VarF5rep);
%typeone(0,10,1763.18,438.9605143,10,10000,ND_10VarD10rep, ND_10VarF10rep);
```

Table B1: Example Type I Error Results Tables

```
         Type One Error Simulation 5 reps - ND_10VarD5rep

                                          Cumulative    Cumulative
ANOVAflag    Frequency      Percent        Frequency      Percent
--------------------------------------------------------------
        0        9522        95.22            9522        95.22
        1         478         4.78           10000       100.00
```

```
          Type One Error Simulation 5 reps - ND_10VarD5rep

                                          Cumulative    Cumulative
 Dunnflag    Frequency      Percent        Frequency      Percent
-------------------------------------------------------------
        0        9520        95.20            9520        95.20
        1         480         4.80           10000       100.00
```

```
          Type One Error Simulation 5 reps - ND_10VarD5rep

                                           Cumulative    Cumulative
  FDRflag    Frequency      Percent        Frequency      Percent
------------------------------------------------------------
        0        9599        95.99            9599        95.99
        1         401         4.01           10000       100.00
```

B.2. Rejection Rate Code

```
dm 'log; clear; output; clear;';
options ls=90 ps=120 formchar="|----|+|---+=|-/\<>*" nodate nonumber;
title1 'Durango Model';
options threads=yes cpucount=8;

libname Thesis 'E:\Thesis\Simulations\';

%macro power(data,seed0,n1,mu1,sd1,n2,mu2,sd2,samples,n3,out_d,out_f, power);

title2 "Power Analysis - &out_d., &out_f";

   data Thesis.&data (keep=sample variety rep Y);

     call streaminit(&seed0);  *** Initialize with desired seed. ***;
     do sample=1 to &samples;
       do variety=1 to &n1;  *** Simulate data for high mean group ***;
         do rep= 1 to &n3;
             Y=rand('Normal',(&mu1),&sd1); output;
         end;
       end;

       do variety=&n1+1 to &n2; *** Simulate data for North Dakota mean group ***;
         do rep=1 to &n3;
             Y=rand('Normal',(&mu2),&sd2); output;
         end;
       end;
     end;
   run;

   ods listing close;            *** Toggle standard output off. ***;


   ods output diff=Thesis.&out_d overallanova=Thesis.&power (keep=probf);
   proc glm data=Thesis.&data plots=none;  /* Dunnett output plus power p-values */
        by sample;
        class variety;
        model Y= variety;
        lsmeans variety /pdiff=control("1");
   run;

   ods output diff=Thesis.&out_f; /* FDR output */
   proc glm data=Thesis.&data plots=none;
       by sample;
       class variety;
       model Y=variety;
       lsmeans variety / pdiff;
   run;
   ods listing;

   data Thesis.&power;
       set Thesis.&power;
       if ProbF >.;
       if ProbF <= 0.05 then ANOVAflag = 1;
       else ANOVAflag=0;
   run;

/* Dunnett flags for significant differences from the control */
   data Thesis.&out_d;
       set Thesis.&out_d;
       if RowName=1 then delete;
       by sample;
       if _1 <= 0.05 then flagdunn=1;
       else flagdunn=0;
   run;

/* FDR procedure and flags for significant differences from the control */
   data Thesis.&out_f (keep=sample Effect Dependent RowName _1);
       set Thesis.&out_f;
   run;

   data Thesis.&out_f;
       set Thesis.&out_f;
       if RowName=1 then delete;
       rename _1=raw_p;
   run;
   proc sort data=Thesis.&out_f;
       by sample;
   run;

   proc multtest inpvalues=Thesis.&out_f fdr out=Thesis.out&out_f;
       by sample;
   run;

   data Thesis.out&out_f;
       set Thesis.out&out_f;
       if fdr_p <= 0.05 then flagfdr = 1;
       else flagfdr = 0;
   run;

/* Frequency tables for all flags */
   proc freq data=Thesis.&power;
       tables ANOVAflag;
   run;

   proc freq data=Thesis.&out_d;
       tables flagdunn;
   run;

   proc freq data=Thesis.out&out_f;
       tables flagfdr;
   run;

%mend power;
```

B.2.1. Macro Calls for Five Replicates

```
%power(Power19,0,1,3519,438.9605143,10,1763.18,438.9605143,10000,5,NDDunnett1vs9,
NDFDR1vs9, ANOVAPower19);
%power(Power91,0,9,3519,438.9605143,10,1763.18,438.9605143,10000,5,NDDunnett9vs1,
NDFDR9vs1, ANOVAPower91);
%power(Power55,0,5,3519,438.9605143,10,1763.18,438.9605143,10000,5,NDDunnett5vs5,
NDFDR5vs5, ANOVAPower55);
%power(Power28,0,2,3519,438.9605143,10,1763.18,438.9605143,10000,5,NDDunnett2vs8,
NDFDR2vs8, ANOVAPower28);
%power(Power82,0,8,3519,438.9605143,10,1763.18,438.9605143,10000,5,NDDunnett8vs2,
NDFDR8vs2, ANOVAPower82);
%power(Power37,0,3,3519,438.9605143,10,1763.18,438.9605143,10000,5,NDDunnett3vs7,
NDFDR3vs7, ANOVAPower37);
%power(Power73,0,7,3519,438.9605143,10,1763.18,438.9605143,10000,5,NDDunnett7vs3,
NDFDR7vs3, ANOVAPower73);
%power(Power46,0,4,3519,438.9605143,10,1763.18,438.9605143,10000,5,NDDunnett4vs6,
NDFDR4vs6, ANOVAPower46);
%power(Power64,0,6,3519,438.9605143,10,1763.18,438.9605143,10000,5,NDDunnett6vs4,
NDFDR6vs4, ANOVAPower64);
```

B.2.2. Macro Calls for Ten Replicates

```
%power(Power19,0,1,3519,438.9605143,10,1763.18,438.9605143,10000,10,NDDunnett1vs9,
NDFDR1vs9, ANOVAPower19);
%power(Power91,0,9,3519,438.9605143,10,1763.18,438.9605143,10000,10,NDDunnett9vs1,
NDFDR9vs1, ANOVAPower91);
%power(Power55,0,5,3519,438.9605143,10,1763.18,438.9605143,10000,10,NDDunnett5vs5,
NDFDR5vs5, ANOVAPower55);
%power(Power28,0,2,3519,438.9605143,10,1763.18,438.9605143,10000,10,NDDunnett2vs8,
NDFDR2vs8, ANOVAPower28);
```

```
%power(Power82,0,8,3519,438.9605143,10,1763.18,438.9605143,10000,10,NDDunnett8vs2,
NDFDR8vs2, ANOVAPower82);
%power(Power37,0,3,3519,438.9605143,10,1763.18,438.9605143,10000,10,NDDunnett3vs7,
NDFDR3vs7, ANOVAPower37);
%power(Power73,0,7,3519,438.9605143,10,1763.18,438.9605143,10000,10,NDDunnett7vs3,
NDFDR7vs3, ANOVAPower73);
%power(Power46,0,4,3519,438.9605143,10,1763.18,438.9605143,10000,10,NDDunnett4vs6,
NDFDR4vs6, ANOVAPower46);
%power(Power64,0,6,3519,438.9605143,10,1763.18,438.9605143,10000,10,NDDunnett6vs4,
NDFDR6vs4, ANOVAPower64);
```

Table B2: Example Rejection Rate Results Table

```
            Power Analysis - NDDunnett9vs1, NDFDR9vs1

                                         Cumulative    Cumulative
ANOVAflag    Frequency     Percent       Frequency      Percent
-----------------------------------------------------------------
        0            1        0.01              1          0.01
        1         9999       99.99          10000        100.00

            Power Analysis - NDDunnett9vs1, NDFDR9vs1

                                         Cumulative    Cumulative
 flagdunn    Frequency     Percent       Frequency      Percent
 ----------------------------------------------------------------
        0        79421       88.25          79421         88.25
        1        10579       11.75          90000        100.00

            Power Analysis - NDDunnett9vs1, NDFDR9vs1

                                         Cumulative    Cumulative
  flagfdr    Frequency     Percent       Frequency      Percent
  ---------------------------------------------------------------
        0        78790       87.54          78790         87.54
        1        11210       12.46          90000        100.00
```

B.3. False Discovery Rate and Power Analysis

```
dm 'log; clear; output; clear;';
options ls=120 ps=120 formchar="|----|+|---+=|-/\<>*" nodate nonumber;
options threads=yes cpucount=8;

%macro DunnettFreq(comp, n, n2);
  libname Thesis "E:\Thesis\Simulations\NewPower 10 Var &n Rep &n2 SD\";

  proc freq data=Thesis.nddunnett&comp;
       tables flagdunn*rowname;
       title1 "Two-way table for Dunnett flag vs. comparison for &comp, &n
        reps, &n2 SD effect size";
  run;
%mend DunnettFreq;

%DunnettFreq(1vs9,5,4); %DunnettFreq(2vs8,5,4); %DunnettFreq(3vs7,5,4);
%DunnettFreq(4vs6,5,4); %DunnettFreq(5vs5,5,4); %DunnettFreq(6vs4,5,4);
%DunnettFreq(7vs3,5,4); %DunnettFreq(8vs2,5,4); %DunnettFreq(9vs1,5,4);
%DunnettFreq(1vs9,10,4); %DunnettFreq(2vs8,10,4); %DunnettFreq(3vs7,10,4);
%DunnettFreq(4vs6,10,4); %DunnettFreq(5vs5,10,4); %DunnettFreq(6vs4,10,4);
```

```
%DunnettFreq(7vs3,10,4); %DunnettFreq(8vs2,10,4); %DunnettFreq(9vs1,10,4);

%macro FDRfreq(comp, n);
  libname Thesis "E:\Thesis\Simulations\NewPower 10 Var &n Rep &n2 SD\";

  proc freq data=Thesis.outndfdr&comp;
       tables flagfdr*rowname;
       title1 "Two-way table for FDR flag vs comparison for &comp, &n reps,
         &n2 SD effect size";
  run;
%mend FDRfreq;

%FDRfreq(1vs9,5,4);  %FDRfreq(2vs8,5,4);  %FDRfreq(3vs7,5,4);
%FDRfreq(4vs6,5,4);  %FDRfreq(5vs5,5,4);  %FDRfreq(6vs4,5,4);
%FDRfreq(7vs3,5,4);  %FDRfreq(8vs2,5,4);  %FDRfreq(9vs1,5,4);

%FDRfreq(1vs9,10,4);  %FDRfreq(2vs8,10,4);  %FDRfreq(3vs7,10,4);
%FDRfreq(4vs6,10,4);  %FDRfreq(5vs5,10,4);  %FDRfreq(6vs4,10,4);
%FDRfreq(7vs3,10,4);  %FDRfreq(8vs2,10,4);  %FDRfreq(9vs1,10,4);
```

Examples of the frequency tables used for the false discovery analysis can be found on the following two pages. The first is using Dunnett's test and the second is using the FDR method. They are both for the 2 vs. 8 comparison. Here we looked at the number of 1 flags in the column for variety 2. In the Dunnett table there are 76 flags. We divide this by 10,000 to find the false discovery rate, since variety two is the only variety other than the control coming from the high mean group. In the FDR method table, this number is 472 which is again divided by 10,000 to find the false discovery rate. To find the power, we add up the 1 flags for columns 3-10 and divide this number by 80,000. For Dunnett's test, this sum is 79,970 and for the FDR method, this sum is 79,999.

Table B3: Frequency Table Used for False Discovery Analysis, Dunnett's Test for 2 vs. 8

```
                           Durango Model - Frequency Tables - 5 reps
                      Two-way table for Dunnett flag vs. comparison for 2vs8

                                     The FREQ Procedure

                                  Table of flagdunn by RowName

flagdunn     RowName(i/j)

Frequency|
Percent  |
Row Pct  |
Col Pct  |   2    |   3    |   4    |   5    |   6    |   7    |   8    |   9    |  10    |  Total
---------+--------+--------+--------+--------+--------+--------+--------+--------+--------+
       0 |   9924 |      4 |      3 |      5 |      5 |      0 |      4 |      6 |      4 |   9955
         |  11.03 |   0.00 |   0.00 |   0.01 |   0.01 |   0.00 |   0.00 |   0.01 |   0.00 |  11.06
         |  99.69 |   0.04 |   0.03 |   0.05 |   0.05 |   0.00 |   0.04 |   0.06 |   0.04 |
         |  99.24 |   0.04 |   0.03 |   0.05 |   0.05 |   0.00 |   0.04 |   0.06 |   0.04 |
---------+--------+--------+--------+--------+--------+--------+--------+--------+--------+
       1 |     76 |   9996 |   9997 |   9995 |   9995 |  10000 |   9996 |   9994 |   9996 |  80045
         |   0.08 |  11.11 |  11.11 |  11.11 |  11.11 |  11.11 |  11.11 |  11.10 |  11.11 |  88.94
         |   0.09 |  12.49 |  12.49 |  12.49 |  12.49 |  12.49 |  12.49 |  12.49 |  12.49 |
         |   0.76 |  99.96 |  99.97 |  99.95 |  99.95 | 100.00 |  99.96 |  99.94 |  99.96 |
---------+--------+--------+--------+--------+--------+--------+--------+--------+--------+
Total       10000    10000    10000    10000    10000    10000    10000    10000    10000    90000
            11.11    11.11    11.11    11.11    11.11    11.11    11.11    11.11    11.11   100.00
```

Table B4: Frequency Table Used for False Discovery Analysis, FDR Method for 2 vs. 8

```
                             Durango Model - Frequency Tables - 5 reps
                         Two-way table for FDR flag vs comparison for 2vs8

                                       The FREQ Procedure

                                  Table of flagfdr by RowName

   flagfdr     RowName(i/j)

   Frequency|
   Percent  |
   Row Pct  |
   Col Pct  |   2    |   3    |   4    |   5    |   6    |   7    |   8    |   9    |  10    |  Total
   ---------+--------+--------+--------+--------+--------+--------+--------+--------+--------+
          0 |   9528 |      0 |      0 |      1 |      0 |      0 |      0 |      0 |      0 |   9529
            |  10.59 |   0.00 |   0.00 |   0.00 |   0.00 |   0.00 |   0.00 |   0.00 |   0.00 |  10.59
            |  99.99 |   0.00 |   0.00 |   0.01 |   0.00 |   0.00 |   0.00 |   0.00 |   0.00 |
            |  95.28 |   0.00 |   0.00 |   0.01 |   0.00 |   0.00 |   0.00 |   0.00 |   0.00 |
   ---------+--------+--------+--------+--------+--------+--------+--------+--------+--------+
          1 |    472 |  10000 |  10000 |   9999 |  10000 |  10000 |  10000 |  10000 |  10000 |  80471
            |   0.52 |  11.11 |  11.11 |  11.11 |  11.11 |  11.11 |  11.11 |  11.11 |  11.11 |  89.41
            |   0.59 |  12.43 |  12.43 |  12.43 |  12.43 |  12.43 |  12.43 |  12.43 |  12.43 |
            |   4.72 | 100.00 | 100.00 |  99.99 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 |
   ---------+--------+--------+--------+--------+--------+--------+--------+--------+--------+
    Total       10000    10000    10000    10000    10000    10000    10000    10000    10000    90000
                11.11    11.11    11.11    11.11    11.11    11.11    11.11    11.11    11.11   100.00
```

APPENDIX C. SAS CODE AND OUTPUT - STAGE TWO

The following macro calls can be used with the code in Appendix B, to find the type I error rate, rejection rate, false discovery rate, and power for stage two.

```
%typeone(0,30,1763.18,438.9605143,2,10000,ND_30VarD2rep, ND_30VarF2rep);


%power(Power129,0,1,3519,438.9605143,30,1763.18,438.9605143,10000,2,NDDunnett1vs29,
NDFDR1vs29, ANOVAPower129);
%power(Power291,0,29,3519,438.9605143,30,1763.18,438.9605143,10000,2,NDDunnett29vs1,
NDFDR29vs1, ANOVAPower291);
%power(Power1515,0,15,3519,438.9605143,30,1763.18,438.9605143,10000,2,NDDunnett15vs15,
NDFDR15vs15, ANOVAPower1515);


%DunnettFreq(1vs29,2,4);   %DunnettFreq(15vs15,2,4);   %DunnettFreq(29vs1,2,4);

%FDRfreq(1vs29,2,4);       %FDRfreq(15vs15,2,4);       %FDRfreq(29vs1,2,4);
```

APPENDIX D. SAS CODE AND OUTPUT - FURTHER ANALYSIS

D.1. Two Replicates

The following macro calls can be used with the code in Appendix B for type I error, rejection rate, power, and false discovery analysis for the further analysis using two replicates.

```
%typeone(0,10,1763.18,438.9605143,2,10000,ND_10VarD2rep, ND_10VarF2rep);
```

```
%power(Power19,0,1,3519,438.9605143,10,1763.18,438.9605143,10000,2,NDDunnett1vs9,
NDFDR1vs9, ANOVAPower19);
%power(Power91,0,9,3519,438.9605143,10,1763.18,438.9605143,10000,2,NDDunnett9vs1,
NDFDR9vs1, ANOVAPower91);
%power(Power55,0,5,3519,438.9605143,10,1763.18,438.9605143,10000,2,NDDunnett5vs5,
NDFDR5vs5, ANOVAPower55);
%power(Power28,0,2,3519,438.9605143,10,1763.18,438.9605143,10000,2,NDDunnett2vs8,
NDFDR2vs8, ANOVAPower28);
%power(Power82,0,8,3519,438.9605143,10,1763.18,438.9605143,10000,2,NDDunnett8vs2,
NDFDR8vs2, ANOVAPower82);
%power(Power37,0,3,3519,438.9605143,10,1763.18,438.9605143,10000,2,NDDunnett3vs7,
NDFDR3vs7, ANOVAPower37);
%power(Power73,0,7,3519,438.9605143,10,1763.18,438.9605143,10000,2,NDDunnett7vs3,
NDFDR7vs3, ANOVAPower73);
%power(Power46,0,4,3519,438.9605143,10,1763.18,438.9605143,10000,2,NDDunnett4vs6,
NDFDR4vs6, ANOVAPower46);
%power(Power64,0,6,3519,438.9605143,10,1763.18,438.9605143,10000,2,NDDunnett6vs4,
NDFDR6vs4, ANOVAPower64);
```

```
%DunnettFreq(1vs9,2,4); %DunnettFreq(2vs8,2,4); %DunnettFreq(3vs7,2,4);
%DunnettFreq(4vs6,2,4); %DunnettFreq(5vs5,2,4); %DunnettFreq(6vs4,2,4);
%DunnettFreq(7vs3,2,4); %DunnettFreq(8vs2,2,4); %DunnettFreq(9vs1,2,4);
```

```
%FDRfreq(1vs9,2,4);  %FDRfreq(2vs8,2,4);  %FDRfreq(3vs7,2,4);
%FDRfreq(4vs6,2,4);  %FDRfreq(5vs5,2,4);  %FDRfreq(6vs4,2,4);
%FDRfreq(7vs3,2,4);  %FDRfreq(8vs2,2,4);  %FDRfreq(9vs1,2,4);
```

D.2. Effect Size of Three SDs, Five Replicates

The following macro calls can be used with the rejection rate, power, and false discovery analysis code found in Appendix B.

```
%power(Power19,0,1,3080,438.9605143,10,1763.18,438.9605143,10000,5,NDDunnett1vs9,
NDFDR1vs9, ANOVAPower19);
%power(Power91,0,9,3080,438.9605143,10,1763.18,438.9605143,10000,5,NDDunnett9vs1,
NDFDR9vs1, ANOVAPower91);
%power(Power55,0,5,3080,438.9605143,10,1763.18,438.9605143,10000,5,NDDunnett5vs5,
NDFDR5vs5, ANOVAPower55);
%power(Power28,0,2,3080,438.9605143,10,1763.18,438.9605143,10000,5,NDDunnett2vs8,
NDFDR2vs8, ANOVAPower28);
%power(Power82,0,8,3080,438.9605143,10,1763.18,438.9605143,10000,5,NDDunnett8vs2,
NDFDR8vs2, ANOVAPower82);
```

```
%power(Power37,0,3,3080,438.9605143,10,1763.18,438.9605143,10000,5,NDDunnett3vs7,
NDFDR3vs7, ANOVAPower37);
%power(Power73,0,7,3080,438.9605143,10,1763.18,438.9605143,10000,5,NDDunnett7vs3,
NDFDR7vs3, ANOVAPower73);
%power(Power46,0,4,3080,438.9605143,10,1763.18,438.9605143,10000,5,NDDunnett4vs6,
NDFDR4vs6, ANOVAPower46);
%power(Power64,0,6,3080,438.9605143,10,1763.18,438.9605143,10000,5,NDDunnett6vs4,
NDFDR6vs4, ANOVAPower64);


%DunnettFreq(1vs9,5,3);   %DunnettFreq(2vs8,5,3);   %DunnettFreq(3vs7,5,3);
%DunnettFreq(4vs6,5,3);   %DunnettFreq(5vs5,5,3);   %DunnettFreq(6vs4,5,3);
%DunnettFreq(7vs3,5,3);   %DunnettFreq(8vs2,5,3);   %DunnettFreq(9vs1,5,3);

%FDRfreq(1vs9,5,3);   %FDRfreq(2vs8,5,3);   %FDRfreq(3vs7,5,3);
%FDRfreq(4vs6,5,3);   %FDRfreq(5vs5,5,3);   %FDRfreq(6vs4,5,3);
%FDRfreq(7vs3,5,3);   %FDRfreq(8vs2,5,3);   %FDRfreq(9vs1,5,3);
```

D.3. Effect Size of Two SDs, Five Replicates

The following macro calls can be used with the rejection rate, power, and false discovery analysis code found in Appendix B.

```
%power(Power19,0,1,2641,438.9605143,10,1763.18,438.9605143,10000,5,NDDunnett1vs9,
NDFDR1vs9, ANOVAPower19);
%power(Power91,0,9,2641,438.9605143,10,1763.18,438.9605143,10000,5,NDDunnett9vs1,
NDFDR9vs1, ANOVAPower91);
%power(Power55,0,5,2641,438.9605143,10,1763.18,438.9605143,10000,5,NDDunnett5vs5,
NDFDR5vs5, ANOVAPower55);
%power(Power28,0,2,2641,438.9605143,10,1763.18,438.9605143,10000,5,NDDunnett2vs8,
NDFDR2vs8, ANOVAPower28);
%power(Power82,0,8,2641,438.9605143,10,1763.18,438.9605143,10000,5,NDDunnett8vs2,
NDFDR8vs2, ANOVAPower82);
%power(Power37,0,3,2641,438.9605143,10,1763.18,438.9605143,10000,5,NDDunnett3vs7,
NDFDR3vs7, ANOVAPower37);
%power(Power73,0,7,2641,438.9605143,10,1763.18,438.9605143,10000,5,NDDunnett7vs3,
NDFDR7vs3, ANOVAPower73);
%power(Power46,0,4,2641,438.9605143,10,1763.18,438.9605143,10000,5,NDDunnett4vs6,
NDFDR4vs6, ANOVAPower46);
%power(Power64,0,6,2641,438.9605143,10,1763.18,438.9605143,10000,5,NDDunnett6vs4,
NDFDR6vs4, ANOVAPower64);


%DunnettFreq(1vs9,5,2);   %DunnettFreq(2vs8,5,2);   %DunnettFreq(3vs7,5,2);
%DunnettFreq(4vs6,5,2);   %DunnettFreq(5vs5,5,2);   %DunnettFreq(6vs4,5,2);
%DunnettFreq(7vs3,5,2);   %DunnettFreq(8vs2,5,2);   %DunnettFreq(9vs1,5,2);

%FDRfreq(1vs9,5,2);   %FDRfreq(2vs8,5,2);   %FDRfreq(3vs7,5,2);
%FDRfreq(4vs6,5,2);   %FDRfreq(5vs5,5,2);   %FDRfreq(6vs4,5,2);
%FDRfreq(7vs3,5,2);   %FDRfreq(8vs2,5,2);   %FDRfreq(9vs1,5,2);
```

D.4. Effect Size of One SD, Five Replicates

The following macro calls can be used with the rejection rate, power, and false discovery analysis code found in Appendix B.

```
%power(Power19,0,1,2202,438.9605143,10,1763.18,438.9605143,10000,5,NDDunnett1vs9,
NDFDR1vs9, ANOVAPower19);
%power(Power91,0,9,2202,438.9605143,10,1763.18,438.9605143,10000,5,NDDunnett9vs1,
NDFDR9vs1, ANOVAPower91);
%power(Power55,0,5,2202,438.9605143,10,1763.18,438.9605143,10000,5,NDDunnett5vs5,
NDFDR5vs5, ANOVAPower55);
%power(Power28,0,2,2202,438.9605143,10,1763.18,438.9605143,10000,5,NDDunnett2vs8,
NDFDR2vs8, ANOVAPower28);
%power(Power82,0,8,2202,438.9605143,10,1763.18,438.9605143,10000,5,NDDunnett8vs2,
NDFDR8vs2, ANOVAPower82);
%power(Power37,0,3,2202,438.9605143,10,1763.18,438.9605143,10000,5,NDDunnett3vs7,
NDFDR3vs7, ANOVAPower37);
%power(Power73,0,7,2202,438.9605143,10,1763.18,438.9605143,10000,5,NDDunnett7vs3,
NDFDR7vs3, ANOVAPower73);
%power(Power46,0,4,2202,438.9605143,10,1763.18,438.9605143,10000,5,NDDunnett4vs6,
NDFDR4vs6, ANOVAPower46);
%power(Power64,0,6,2202,438.9605143,10,1763.18,438.9605143,10000,5,NDDunnett6vs4,
NDFDR6vs4, ANOVAPower64);
```

```
%DunnettFreq(1vs9,5,1);   %DunnettFreq(2vs8,5,1);   %DunnettFreq(3vs7,5,1);
%DunnettFreq(4vs6,5,1);   %DunnettFreq(5vs5,5,1);   %DunnettFreq(6vs4,5,1);
%DunnettFreq(7vs3,5,1);   %DunnettFreq(8vs2,5,1);   %DunnettFreq(9vs1,5,1);

%FDRfreq(1vs9,5,1);   %FDRfreq(2vs8,5,1);   %FDRfreq(3vs7,5,1);
%FDRfreq(4vs6,5,1);   %FDRfreq(5vs5,5,1);   %FDRfreq(6vs4,5,1);
%FDRfreq(7vs3,5,1);   %FDRfreq(8vs2,5,1);   %FDRfreq(9vs1,5,1);
```