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

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

By

Kayéromi Donoukounmahou Gomez

In Partial Fulfillment of the Requirements
for the Degree of
DOCTOR OF PHILOSOPHY

Major Department:
Statistics

April 2015

Fargo, North Dakota

North Dakota State University
Graduate School

Title

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

By

Kayéromi Donoukounmahou Gomez

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

DOCTOR OF PHILOSOPHY

SUPERVISORY COMMITTEE:

Dr. Rhonda Magel
Co-Chair

Curt Doetkott
Co-Chair

Dr. Megan Orr

Dr. Yarong Yang

Dr. Juan Osorno

Approved:

04/10/2015	Dr. Rhonda Magel
Date	Department Chair

ABSTRACT

It has become quite common nowadays to perform multiple tests simultaneously in order to detect differences of a certain trait among groups. This often leads to an inflated probability of at least one Type I Error, a rejection of a null hypothesis when it is in fact true. This inflation generally leads to a loss of power of the test especially in multiple testing and multiple comparisons.

The aim of the research is to use simulation to address what a researcher should do to determine which treatments are significantly different from the control when there is a large number of treatments and the number of replicates in each treatment is small. We examine two situations in this simulation study: when the number of replicates per treatment is 3 and also when it is 5 and in each of these situations, we simulated from a normal distribution and in mixture of normal distributions. The total number of simulated treatments was progressively increased from 50 to 100 then 150 and finally 300. The goal is to measure the change in the performances of the False Discovery Rate method and Dunnett's test in terms of type I error and power as the total number of treatments increases.

We reported two ways of examining type I error and power: first, we look at the performances of the two tests in relation to all other comparisons in our simulation study, and secondly per simulated sample. In the first assessment, the False Discovery Rate method appears to have a higher power while keeping its type I error in the same neighborhood as Dunnett's test and in the latter, both tests have similar powers and the False Discovery Rate method has a higher type I error. Overall, the results show that when the objective of the researcher is to detect as many of the differences as possible, then FDR method is preferred. However if error is more detrimental to the outcomes of the research, Dunnett's test offers a better alternative.

ACKNOWLEDGEMENTS

I want to first and foremost acknowledge the utmost dedication of my co-chair, professor and boss (even though he doesn't like these titles) Curt Doetkott to this project. Not only for your 24/7 availability but also for unconditionally taking me under your wings and showing me how to become a good statistical consultant. I remember those moments when you spent your entire weekend helping me. I also remember those emails from me that you replied to in the middle of the night. Thank you for everything that you have done for me over these past years; not a paragraph, a page or even a whole book is enough to write everything that you have done for me and I am forever grateful. I am very grateful. My appreciation also goes to your wife Dawn Doetkott for her support. You will forever be in my heart and in my mind.

I also want to acknowledge my entire dissertation committee beginning with my chair Dr. Rhonda Magel for your constant pressure to keep me on track and your vision for this project. I remember those early brainstorming meetings in your old office. Thank you Dr. Megan Orr for engaging me into looking at pertinent details of my research. Your feedbacks have been very rewarding for me personally and also for this project. I also want to thank Dr. Yarong Yang for your help on this project. Your thoughts have guided me in having my audience in mind while writing this paper. My appreciation also goes to Dr. Juan Osorno for always making me focus on the big picture and the practical aspect of the findings of this dissertation.

Finally, I want to thank all my colleagues, friends, encounters and all those I met in Fargo, America, Europe and Africa while working on this project and who have kept encouraging me to finish and finish strong. Though you might not all understand details of what my topic is about, your inputs have been instrumental in keeping me on track. And for all those whose name might have been omitted here, please accept my gratitude. You all have helped shape a great doctor and a great leader and may Almighty God abundantly bless you.

DEDICATION

To Almighty God for making it happen in my life.

To my three angels Sèna, Djahou and Sèdjro (Mr. Legrand) Gomez. You are the reason why I keep pressing on despite all the obstacles and I want you to know that you are my stars. I also want you to grow up and be proud of Daddy and know that Daddy is always proud of you.

To my mother Victoire Savi who, though never attended school in her life, understood the value of education and early on in life, invested her life savings so that I could grow up and do what she was not able to do. By virtue of the power invested in me through this doctorate degree, I pronounce you Dr. Victoire Savi!

To my Carbon-copy mom Euneita Clark for your undeniable support throughout all these years. You did it mom! You are Dr. Euneita Clark too!

To my Chicago-grandma Arzella Clark. Thanks so much Grandma! Without you, it wouldn't have happened. Another Dr. Arzella Clark!

To Agnes Kakamor, my entire family, my father, my sisters and my brothers, thank you all for being there. This degree is for all of us!

To my adopted sister Mercy Asare, and my friends Terry Enadeghe, Jojo Midley and Emmanuel Hounguevou, your persistent encouragement and support have taken me through. You rock!

To my Chicago-mom's big family of supporters and all my friends. Thank you so much for your encouragements and celebrations every step of the way. We did it!

And to everyone I love and might have not mentioned. This is yours as well!

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

MCP ..... Multiple Comparison Procedures

FWER ..... Family-Wise Error Rate

FDR ..... False Discovery Rate

AE ..... Actual Error

AP ..... Actual Power

RE ..... Relative Error

RP ..... Relative Power

LIST OF APPENDIX TABLES

LIST OF APPENDIX FIGURES

CHAPTER 1. INTRODUCTION

1.1. Background

In recent years, the world of research has evolved so much so that the general public has grown to have either a negative, positive or neutral attitude towards the statistics that popular media has bombarded them with. Claims of relationships between a disease and a certain behavior, or the association between a certain political result and some societal change, in addition to the many genetic associations studies that turn out to be contradicted months later by yet another research in the same discipline, are at the root of this polarized society. While many of these could be explained by incorrect study designs, some of the blame ought to be shared with the statistical inference methods that statisticians have been accustomed to and the derived conclusions that tend to be mostly dependent on a certain significant P-value threshold of 0.05.

With the recent revolution of massive data in social sciences, genomics, medicine, proteomics and astrophysics as well as biology and plant sciences, many researchers have come to understand that the argument ought to be made not simply by a mere rejection of some hypothesis if a p-value is less than a certain threshold, but rather explore alternatives to the traditional inference methods available especially when there are many tests being performed either simultaneously of individually. The somewhat-newly published method of the False Discovery Rate (FDR), offered an alternative that slowly revolutionized outcomes of modern scientific research, especially in areas of genetics. In a nutshell, the FDR Method controls the expected number of false rejections out of all the null hypotheses that are rejected.

There are two distinct, yet similar, scenarios that use this method. One is the Multiple Comparison scenario which is a two-stage analysis of one dataset and the other is Multiple Testing of many individual tests simultaneously to detect, say, the association of a particular gene to a specific disease. In multiple comparisons problems, we examine differences in treatment effects after rejecting an overall null hypothesis in an analysis of variance (Lazar, 2012). It enables for further investigation of the differences among the treatments means. Post-hoc procedures have been developed in the past to help researchers detect where those differences lie. Some of the more common ones, the Sheffe's procedure,

Fisher's Least Significant Difference, Tukey's Honestly Significant Difference and the Bonferroni Correction, developed a follow-up procedure to compare all the individual treatments to each other. One of the tests being studied in this paper, the Dunnett's test, developed further a procedure that compares all other treatments against a set control, reducing the total number of comparisons in the experiment giving yet another alternative to the researcher. This second stage of the testing can easily turn into a multitude of tests. For instance, with six treatments, there are $\binom{6}{2} = \frac{6!}{4!2!} = 15$ comparisons to make and controlling the error rate for this family of comparisons is critical to avoiding misleading conclusions.

The testing procedure for multiple testing research is different. Although type I error control is still of interest, there have been no prior test conducted to reject an overall null hypothesis of difference among the mean treatments. In genomics, for instance, where many individual tests are simultaneously conducted to detect the association of a particular gene to a particular disease, no preliminary test is conducted to make a first conclusion as to a rejection of an overall null hypothesis. Lazar (2012) explains that multiple testing has to do with first stage hypothesis testing versus post hoc exploration of possible differences for multiple comparisons, with the knowledge that some differences exist. Despite the differences in the two procedures, they still have a common goal of controlling some kind of type I error rate.

Oftentimes referred to as MCPs, Multiple Comparison Procedures will be used in this paper to mean any procedure in which simultaneous inferences are made, and we want to simply assess the equality of means and may conclude inequality as our null identifies what we are actually testing. Therefore no reference will be made to confidence intervals for means differences, an area in which MCPs have well thrived. The ability of these tests to keep the error rate for the entire family of comparisons at some reasonably low level constitute an ultimate protection against false positives. In practice, since we do not know which null hypotheses are true, we work under the assumption that all the null hypotheses are true in order to fully control the family-wise error (FWER) at a specified rate. But in this study, we will have the ability to vary the number of true null hypotheses to ascertain the performances of FDR method and Dunnett's test.

In its simplest definition, FWER controls the number of false positives among all comparisons. We will use the following table to explain this and other kinds of type I error made in statistical testing when a total number of ***m*** Null Hypotheses are being tested.

Table 1.1. Random Variables Representing the Number of Errors Committed when Testing m Hypothesis

	Declared non – significant	Declared Significant	Total
True null hypothesis	U	V	m_0
Non – true null hypothesis	T	S	$m - m_0$
Total	$m - R$	R	m

Using this table, (Dudoit et al, 2003) reorganized Type I error rates in four categories: the per-comparison error rate, the per-family error rate, the family-wise error rate, and the false discovery rate. Dudoit et. al. (2003) defines the per-comparison error rate as the expected value of the number of Type I errors divided by the number of hypotheses, {E(V)/m}. Per-family error rate {E(V)} is the expected number of Type I errors or the expected number of true null hypotheses rejected. The family-wise error rate is defined as the probability of at least one Type I error; FWER=P(V≥1). Its control is set for the overall family of comparisons at stake. The false discovery rate (FDR) of Benjamini and Hochberg (1995) is also useful when in presence of a family of comparisons. It is defined as the expected proportion of Type I errors among the rejected hypotheses. In other words, FDR is the expectation of a variable T such that T=V/R when R > 0 and T=0 when R=0 (this condition is included to account for extremely rare cases arising from the fact that FDR is an expected proportion and not a probability). FDR = E(V/R | R > 0).

In general, in a multiple testing of simultaneous hypotheses where m_0 hypotheses are true in a total of m being tested, FWER = P (reject at least one of the m_o hypotheses | m_o hypotheses are all true). Specifically, suppose we are testing 50 null hypotheses using a Bonferroni type of control of the

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

Controlling Family-wise error rate therefore, aims at controlling the probability of committing a single type-I error or more within the tested family of hypotheses. When testing only one null hypothesis, we often define Type I error as the probability of a false rejection of that null hypothesis. In situations where multiple hypotheses are being tested simultaneously, a MCP method will be controlling Type I error at an average level less than or equal to α, which is a FWE rate when this control encompasses all the family of null hypotheses under study (Dudoit et al, 2003). In the same manner, for FDR type of control, we will say that it is controlling type I error on average at a rate of α if FDR< α when FDR method is applied to all the rejected null hypotheses.

Though we do not expect every significant result to be true, a high proportion of the true differences detected is desired, hence the use of a more liberal rate to control in terms of the number of type I errors, the False Discovery Rate. In MCP, FWER is set to be less than or equal to an alpha level for the entire family of comparisons, but this alpha can also become too strict especially when the number of tests is large and FDR becomes a more appropriate error rate to control.

1.2. The FDR Controlling Procedure

The False Discovery Rate Controlling Procedure was developed by Yoav Benjamini and Yosef Hochberg and published in 1995. It is defined as the expected proportion of false "discoveries" or

rejections of null hypotheses among all rejections. In the authors' own words, it is "the expected ratio of erroneous rejections to the number of rejected hypotheses." This turns out to be an appropriate error rate to control in many problems especially since a method that strongly controls FWER also strongly controls FDR and a method that strongly controls FDR weakly control FWER.

In practice, FDR method follows a simple process. It starts by ordering the P-values for the m tests in ascending order; it then finds the largest k such that p (k) ≤ k α / m and rejects all tests with p-value less than p (k) where p(k) is the k^{th} smallest P-value . In the end, FDR procedure estimates the rejection region for an average FDR ≤ α while estimating the largest k that enables a rejection of k p-values with m initial ordered p-values. In this study, the performance of FDR controlling procedure is compared to that of Dunnett's test, a FWER control procedure that has been of great use among researchers for many years when the objective of the study is to compare many treatments to a control.

1.3. Dunnett's Test

Developed by the Canadian statistician Charles Dunnett and published in 1955, Dunnett's test was designed to be used either to test the significance of the differences between each of the treatments and a set control or to establish confidence limits on the true values of the treatment differences from the control (Dunnett, 1964). In its original form, Dunnett's test has the ability to control the family-wise error rate rather than the comparison-wise error rate. Its statistic is defined by Conagin et al., (2011) as:

$$d_i = \frac{m_i - m_c}{\sqrt{\frac{2}{n} MSE}}$$

Where $m_i = mean\ of\ the\ i^{th}\ teatment$

$m_c = mean\ of\ the\ control$

$n = sample\ size\ for\ each\ treatment$

$MSE = Mean\ Square\ Error\ from\ ANOVA$

$d_i = value\ to\ be\ compared\ to\ Dunnett's\ table\ values\ to\ enable\ a\ rejecion\ or\ not\ of\ Ho$

The test was historically used in multiple comparisons experiments where the number of treatments to be compared with the control were up to nine, the number allowed in the d_i values in the original tables provided in 1955. An updated table was provided by Dunnett in 1964 to expand the test's

magnitude and power such that today, with Dunnett's test, there is lesser chances of making one false discovery since there are fewer tests of significance within the set (Conagin et al., 2011). Dunnett's test is also preferable when testing paired differences between means and a control after an analysis of variance. Dunnett's test was preferable in another study by Mukerjee et al. (1987) where the other treatments in the experiments are known to be at least as effective as the control and the objective of the research is to find which of these other treatments were significantly better than the control. Dunnett's test was found to have good power when treatment means are equal but different from the control and it provides some protection when they differ.

CHAPTER 2. LITERATURE REVIEW

2.1. Introduction

Since the publication of 'Controlling the false discovery rate: a new and powerful approach to multiple comparisons' by Benjamini and Hochberg in the Journal of the Royal Statistical Society in 1995, the False Discovery Rate (FDR) Controlling Method has been gradually capturing the attention of researchers, until recently when the challenges of massive data analysis propelled scientists to familiarize with the new method. By the year 2010, close to 5,000 citations of the original paper were recorded according to the Web of Science. Most of the researches using the method were in life sciences and in disciplines such as genetics, biochemistry etc. (Benjamini, 2010), as the field of statistics is in itself a center field that helps explain data from all other fields. With the development of health care and the new challenges risen from the numerous health problems faced by our generation, scientists have sought new avenues to a multitude of hypotheses testing. It is perhaps fair to say that the greatest successes of the FDR controlling method have been in genomics research where thousands of genes are often simultaneously under study and their relationship to particular diseases are investigated. Scientists confronted by this problem of multiplicity often find the FDR to be an appealing quantification of error (Osborne 2006).

2.2. The Evolution of FDR Method

Evidently, the emerging era of multiple treatments comparison designs is perhaps ubiquitous most especially with the increasing amount of data this and future generations are confronted with. Gone are the days when simulations studies involving a few dozen treatments are criticized because no one used multiple comparisons for problems with 50 or 100 tested hypotheses (Benjamini, 2010). But today, the method is not only at the center of many MCPs, it continues to develop extensively and new additions have been made over the past several years to make FDR method more useful in practice. Furthermore, "whenever the FWER is small and the number of hypotheses to be tested is large (e.g., of order 10^4 as in microarray studies), the ability of any FWER controlling Multiple Testing Procedures to detect false null hypotheses is inevitably limited." (Gordon et al., 2007). The method has gotten a makeover through the years with multiple additional works by the original authors and others such as Yekutieli, Efron, and

Storey. In 2001, the inventors of the FDR method made it stronger with an addition on the specific characteristics of the distribution of the test statistics being tested. They proved that when the joint distribution of the test statistics are known to have a positive regression dependency on each one from a predetermined null hypothesis subset, "the Benjamini Hochberg procedure controls the FDR at level less than or equal to q (m_0/m)." (Benjamini et al., 2001)

In Efron (2004) addressed the choice of an appropriate null hypothesis in large-scale testing situations, and how this choice affects well-known inference methods such as the false discovery rate, using what he termed Local False Discovery Rate. Local FDR, he suggested, scrutinizes the histograms of the non-null test statistics in an empirical Bayesian version without the need of strong Bayesian assumptions while focusing on density functions rather than tail probability areas. For Osborne (2006), the FDR provides an alternative quantification of error under multiplicity of comparisons and its interpretation is simple enough for scientists across disciplines. Meanwhile, Benjamini and Hochberg continued to make noticeable improvements to their original procedure. Yekutieli et al. (2006) revamped the procedure further by introducing a two-stage step through which a linear step-up procedure is used in stage one to estimate m_0, providing a new level α' which is used in the linear step-up procedure in the second stage to controls the false discovery rate at the desired level α.

Gordon and Glazko (2007) attempted a comparison of the Benjamini-Hochberg (BH) procedure with the conservative FWER of Bonferroni. Their findings were published in the Annals of Applied Statistics under the title "Control of the mean number of false Discoveries, Bonferroni and Stability of Multiple Testing." An Extended Bonferroni procedure using the traditional Bonferroni and changing its Per Family Error Rate (PFER) control ability was used in this study. They concluded that "If, for example, the practitioner decides that, on the average, he/she can afford two false positives per experiment, then it is natural to use the Bonferroni procedure with the nominal level of the PFER equaling 2. On the other hand, when the researcher wants the average proportion of false positives among all positives not to exceed 10%, he/she can use the BH procedure with the nominal level of the FDR equaling 0.1 (Gordon et al., 2007).

Efron (2007) explained how a correlation among test statistics can have a remarkable effect on the false discovery rate technique and why accuracy can be compromised in high-correlation situations. He recommended in those situations that the researcher makes a decision as to which of the m null hypotheses could be considered non-null hypotheses in order to affect the correlation among the test statistics. In 2008, he developed a process that combines and separates the tested hypotheses using a Bayesian approach where the choice of trading off variance for a bias in estimating FDR is at stake in the original partially nonparametric framework of Benjamini and Hochberg (Efron 2008).

2.3. FDR Method Going Forward

Despite all these adjustments to the originally-published method, many researchers are still investigating the main assumptions that the FDR Controlling method exploit. In that regard, underlying assumptions of the method should be checked as much as possible, and they call for the development of diagnostic procedures (Benjamini, 2010). In a review of their original paper in 2010, Benjamini wrote: "Benjamini and Hochberg (1995) introduced two things: a new error rate and an algorithm that under certain conditions controls that error rate. If we control the false discovery rate FDR at level α, it means that this procedure when applied many times to this problem on average rejects a fraction α (or less) incorrectly. The BH method does not estimate FDR. It selects the level α and adopts a rejection False Discovery Rate procedure such that FDR $\leq \alpha$" (Benjamini, 2010).

This cleared the confusions surrounding the original estimates of the method. However, the puzzle of whether FDR control is a manifestation of another principle, be it empirical Bayes, decision theoretic, minimum description length and such, or whether it is a principle of its own, which in some setting coincides with another principle, remains to be explored (Benjamini, 2010). Hence the need for more works to be done in order to fully understand all contours of the FDR Controlling method and to make the procedure more powerful. Furthermore, the algorithm behind the procedure ought to be measured with existing algorithms of multiple comparisons in order to offer scientists the necessary tools to make an informed decision as to the choice of the method to be used to analyze their data while

enhancing statistical analysis of real world data and a greater confidence in the FDR Method. This study intends to fill part of that vacuum.

CHAPTER 3. METHODS OF THE SIMULATION STUDY

3.1. Simulation Set-up

This research is a simulation study using SAS version 9.3 (SAS Institute Inc., 2011). SAS version 9.4 was used for the graphs in order to take advantage of its advanced graphing tools. The original data, described in Kubat (2013), consist of over 300 different genotypes of edible dry beans whose Seed Yield (response variable) were being compared. Two genotypes with different means were identified. The first has a mean of $\mu = 1763.18$ and the second a mean of $\mu = 2641.1$. These two means were referred to as the low and high mean respectively. The standard deviation was recorded as $\sigma = 438.96$ and normality of the data was also verified in Kubat (2013). For a complete description including preliminary assumptions verification of the original data see (Kubat, 2013). The goal of the current study is to compare the performances of the False Discovery Rate method and Dunnett's test when the number of genotypes that are the same increases progressively in varying settings in cases where 50, 100, 150 and 300 treatments are simulated. Using Dunnett's test and then FDR, we want to measure the number of rejections made by each of these tests when the null hypothesis is true (type I error) and when the null hypothesis is false (power).

In order to perform Dunnett's test, we first run an ANOVA to reject an overall null hypothesis of no difference among the treatment means. In a post hoc attempt to determine where the differences lie between pairs of treatments, Dunnett's test is then applied using SAS if overall null hypothesis is rejected. Since the first treatment is by default set as the reference treatment against which all other treatments are being compared, only the tests where the other treatments are compared to the first will be considered for the FDR Method using PROC MULTTEST in SAS (SAS Institute Inc., 2011). Dunnett's test FWER and FDR were assessed using the counts where the null hypotheses of equality to the reference was rejected.

Since Dunnett's test has a strong control of family-wise error which is controlled for all comparisons within a sample, it makes sense to look at the error rate for FDR Method also for all comparisons within the simulated sample in order to have a comparable ground for assessing the power

of FDR Method and Dunnett's test. That way, we will be having the same baseline criteria upon which we can compare FDR and Dunnett's.

For a comprehensive assessment of the performances of the two tests, the total number of simulated treatments are progressively increased from 50 to 100, 150 and then 300. This gradual increase in total number of treatments is intended to make the findings of this study useful to various disciplines and various sizes of research. It also allows for an assessment of the two tests in varying total number of true null hypotheses. In each simulation, 10,000 samples were simulated, taking into account the basic assumptions of homogeneity of variance of ANOVA, homogeneous variance of treatment population with control populations, and the characteristics of simple False Discovery Rate embedded in PROC MULTTEST with FDR option.

The power of each test is computed as the proportion of false null hypotheses which are correctly rejected when we know the two populations are different. Type I error is assessed on the proportions of true null hypotheses rejected when the two samples are simulated from the same population. In all, four measures of comparisons will be used to address the similarities and disparities between the two tests: relative power, actual power, relative type I error rate, actual type I error rate (also referred to as the family-wise error rate). The following figure 3.1 presents a complete view of the entire study.

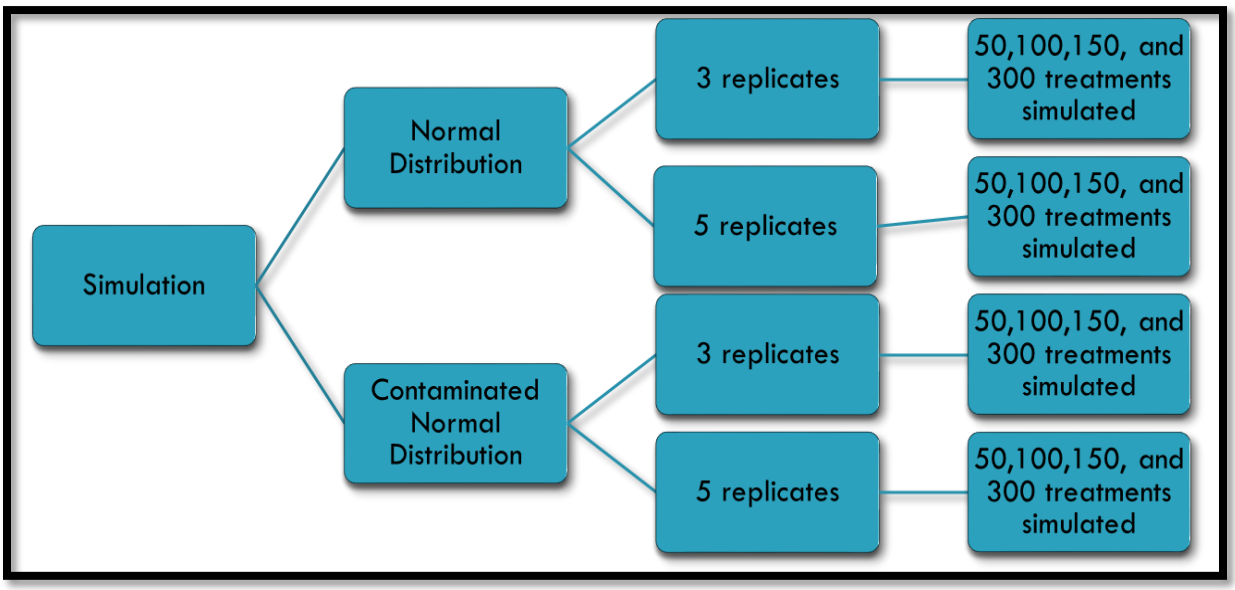

Figure 3.1. Complete View of the Study

3.2. The Measures of Comparison

To enable a comprehensive assessment of Dunnett's test with FDR method, we define four measures of comparisons upon which we will base our discussions.

3.2.1. Relative Power (RP)

RP counts all the rejections made by each method for all tests comparing the control group to the treatment groups when H_0 is false. In the first scenario of 50 simulated treatments for instance, we count the total number of rejections made by each method over a total of 490,000 comparisons (49 comparisons multiplied by 10,000 simulations). We call this ***Relative Power. Power***, since it represents the ability of the method to detect the fact that there is a known difference among the simulated treatments, and ***Relative***, since the number of rejections is per the total number of comparisons in the overall pool of all the simulated comparisons. The following formula will therefore be used to calculate RP for this and all other scenarios in the study.

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

3.2.2. Actual Power (AP)

In assessing the Actual Power, we count the number of at-least-one rejections made by each test per simulated sample when H_0 is false; and we divide this number by the total number of simulations. Each time either Dunnett's or FDR Method is able to detect a difference when there is a true difference, we count it as a true rejection. In counting the rejections in this manner, we want to measure the ability of the test to detect a true difference in a unique data since in most real life situations, we are presented with one sample and not 10,000 samples. The number of times a test is able to detect at-least-one difference is a good measure of the actual power of that test. We therefore call this the ***Actual Power,*** since it is not relative to the other simulated comparisons in other samples, but rather in relation to the comparisons within a simulated sample. Hence the denominator of the proportion representing the percentage of ***Actual Power*** in all scenarios is fixed and is the total number of simulations, 10,000.

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

While checking how the simple t-test (which we will call Raw in this study) is performing with respect to type I error rate and power in all the comparisons in which Dunnett's test and FDR are being compared, we realized that Raw had a control of about 5% on type I error and its powers are higher than those of Dunnett's test and FDR method. If this were true, then post-hoc procedures make no sense as we usually want to be able to make a relatively small type I error while having a good power. And if Raw was already controlling type I error at about 5% and at the same time having a higher power than the other two tests, then there would be no need to perform a post-hoc test.

Therefore there was a need to look closer into the rejections of each of the tests per simulation. Counting the number of rejections made per simulated sample allows for an actual assessment of the performances of the FDR method and Dunnett's test. We then tally all the at-least-one rejections made by each test to count toward the actual performance of the test. These counts are completely different from the first ones and offer a different way of comparing the performances of the two tests.

3.2.3. Relative Type I Error Rate (RE)

This ratio accounts for the total number of rejections made by either test when the treatments are known to have been simulated from the same population as the reference or control. In other words, it counts all the rejections when H_0 is true and divides it by the total number of true null hypotheses. This total number of true null hypotheses varies per setting. It is 10,000 (the total number of simulated treatments) multiplied by the total number of treatments simulated from the same population as the control. For setting 25:25 for instance, 24 samples are simulated from the same population as the control, hence the denominator for RE for this case will be 240,000 (24 × 10,000). Table 1 shows the denominator for RE for all settings in all cases.

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

3.2.4. Actual Type I Error Rate or the FWE Rate (FWER)

This is the actual type I error rate consisting of the total number of at-least-one rejection of true H_0 per simulated sample divided by the total number of simulations (10,000 for all settings).

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

All these four measures of comparing the False Discovery Rate method to Dunnett's test are assessed in different simulated settings as described below. A case will represent the total number of treatments simulated. We have four cases: 50, 100, 150 and 300. We also have two stages under study: the first stage is the simulation from a normal population and the second stage is from a population with a contaminated normal distribution (we will explain this further below). A case number will therefore be comprised of the total number of simulated treatments, the distribution of the population and the number of replicates per simulated treatment. (e.g. 50N5 is the case number for 50 simulated treatments from a normal population with 5 replicates per treatment. Case 100CN3 will be the case number for 100 simulated treatments from a contaminated normal distribution with 3 replicates per treatment).

3.3. Illustration of the Measures of Comparison

For illustration purposes, suppose we have only one sample in our simulated data of 50 treatments, the rejections are summarized in table 3.1 below. In this table, (+) signifies a rejection and (-) signifies a non-rejection. This is for illustration purposes only and does not reflect in any fashion the results in this study. To illustrate definitions of RP, RE, AP and AE, the values in the mean column are not shown.

Table 3.1. Sketch of a Sample of the Simulated Data for 50 Treatments

Sample 1					
Treatment	**Parent Population**	**Mean**	**Dunnett's test**	**FDR Method**	**Measure**
2	Control	**---**	-	-	Type I error
3	Control	**---**	-	+	Type I error
4	Control	**---**	+	-	Type I error
5	Control	**---**	-	+	Type I error
6	Control	**---**	-	+	Type I error
7	Control	**---**	+	-	Type I error
8	Control	**---**	+	-	Type I error
9	Control	**---**	-	-	Type I error
10	Control	**---**	-	-	Type I error
11	Control	**---**	-	-	Type I error
12	Control	**---**	-	-	Type I error
13	Control	**---**	-	-	Type I error

(Continued)

Table 3.1. Sketch of a Sample of the Simulated Data for 50 Treatments (continued)

Sample 1					
Treatment	**Parent Population**	**Mean**	**Dunnett's test**	**FDR Method**	**Measure**
14	Control	---	-	-	Type I error
15	Control	---	+	-	Type I error
16	Control	---	+	-	Type I error
17	Control	---	-	-	Type I error
18	Control	---	-	-	Type I error
19	Control	---	+	-	Type I error
20	Control	---	-	-	Type I error
21	Control	---	-	-	Type I error
22	Control	---	-	-	Type I error
23	Control	---	-	-	Type I error
24	Control	---	-	-	Type I error
25	Control	---	-	-	Type I error
26	Treatment	---	-	+	Power
27	Treatment	---	+	-	Power
28	Treatment	---	-	+	Power
29	Treatment	---	+	+	Power
30	Treatment	---		-	Power
31	Treatment	---	-	-	Power
32	Treatment	---	-	-	Power
33	Treatment	---	-	-	Power
34	Treatment	---	-	-	Power
35	Treatment	---	-	-	Power
36	Treatment	---	-	+	Power
37	Treatment	---	-	-	Power
38	Treatment	---	-	-	Power
39	Treatment	---	-	-	Power
40	Treatment	---	-	-	Power
41	Treatment	---	+	-	Power
42	Treatment	---	-	-	Power
43	Treatment	---	-	-	Power
44	Treatment	---	-	+	Power
45	Treatment	---	-	-	Power
46	Treatment	---	-	-	Power
47	Treatment	---	-	-	Power
48	Treatment	---	-	-	Power
49	Treatment	---	+	-	Power
50	Treatment	---	+	+	Power
Totals			+(Ho true)=6 +(HoFalse)=5	+(Ho true)=3 +(HoFalse)=6	

Using equations (1), (2), (3), and (4) above, the four measures of comparison are computed as follow:

For Dunnett's test:

RP=$\frac{5}{25} \times 100 = 20\%$ (using equation 1)

RE=$\frac{6}{24} \times 100 = 25\%$ (using equation 3)

For AP and AE, since this illustration uses only one sample, the number of at-least-one (true or false) rejection in the sample is 1, and AE and AP are both equal to 1.

AP = $\frac{1}{1} \times 100 = 100\%$ (using equation 2)

AE = $\frac{1}{1} \times 100 = 100\%$ (using equation 4)

And for FDR method, we will have

RE=$\frac{3}{24} \times 100 = 12.50\%$ (using equation 1)

RP=$\frac{6}{24} \times 100$=24% (using equation 3)

For AP and AE, since this illustration uses only one sample, the number of at-least-one (true or false) rejection in the sample is 1, and AE and AP are both equal to 1.

AP=$\frac{1}{1} \times 100 = 100\%$ (using equation 2)

AE=$\frac{1}{1} \times 100 = 100\%$ (using equation 4)

3.4. The Settings

We define the settings to represent a variation in the number of true null hypotheses for each increase in the total number of simulated treatments. Under a fixed number of total simulated treatments (called cases), the number of true null hypotheses is set prior to initiating the simulation in this step of the study. The number of true null hypotheses is always less than the total number of simulated treatments. The notation of a setting will be the number of true null hypotheses and the number of simulated treatments different from the control, separated by a semicolon. Several settings are examined per case and they also vary by case. In 50 simulated treatments for instance, the settings are 1:49,

10:40, 25:25, 40:10 and 49:1. The various settings per case are enumerated in the column named setting in table 3.2 below. The first number represents the control or reference group which is the control treatment and all other treatments that are not different from the control; and the second number is the treatment group which consists of the treatments that are different from the control. The tables summarizing the results will also be denoted with the case number to enable easy flow of this paper. With these settings, we are able to measure the performances of the two tests as the number of true null hypotheses increases, as we progressively increase the number of true null hypotheses from one setting to the next.

It is worth noting here that in all first settings (1:49, 1:99, 1:149 and 1:299), all comparisons are made between a single treatment (reference or control) and group of other treatments different from the control; making all rejections here true rejections, a measure of power. RP, RE, AP and FWER will be assessed in each of the settings throughout the study and in both cases (when the simulated treatments are generated for a normal population, and they are generated using a contaminated normal population) and in both number of replicates per simulated treatment (3 and 5 replicates).

3.4.1. Settings for 50 Simulated Treatments

Setting 1:49 – We simulate a dataset with 50 treatments and we set the first simulated treatment as the control population while the remaining 49 are from the treatments population. The control treatment is simulated from the low mean population with $\mu = 1763.18$ and standard deviation 438.96 while the treatment population has a means of $\mu = 2641.1$ ($1763.18 + 2 \times 438.96$) and a standard deviation 438.96. We wanted to measure an effect size of two standard deviations of the mean in order to stay closest as possible to reality. These mean and standard deviation will be used for all simulations in the first stage of this study.

Setting 10:40 – Here we simulate one treatment from the population with low mean ($\mu = 1763.18$) and nine other treatments from the same population. We then simulate forty other treatments from the population with mean 2641.1. As stated above, the control treatments will always be simulated from the low mean population while the other treatments will be from the high mean population. We keep our standard deviation the same since don't want to complicate results by changing the standard

deviation. In assessing Type I error, we count the total number of rejections for those comparisons when all treatments are the same as the reference. In this setting, we have nine comparisons out of the 49 comparisons in which the treatments are the same as the reference or control. And for the entire 10,000 simulations, we have 90,000 comparisons when the null hypotheses are known to be true and will be used to account for type I error. Forty other treatments are simulated from the high mean population, therefore, we will have 400,000 tests (40 multiplied by 10,000) where the null hypotheses are false and they will be used to account for power. The count of number of detections of differences made from the 400,000 will enable an assessment of RP.

Setting 25:25 – The number of true null hypotheses is increased to 240,000. 24 treatments are simulated here from the same population as the control while 25 treatments are simulated to be different from the control group of treatments. The denominator for RE therefore will be 240,000 while the denominator for relative power will be 250,000. All other formulas will be used as above.

Setting 40:10 – In this setting, the number of true null hypotheses over the entire population of simulations is increased to 390,000. 39 treatments are simulated from the same population as the control population while 10 treatments are from the other population. The number of comparisons when Ho is true here will be 390,000 which represents the denominator for RE and RP will have a denominator of 100,000.

Setting 49:1 – Here, 48 treatments are simulated from the same population as the control and one from the high mean population. There is a total of 480,000 of true null hypotheses, the highest in the case of 50 simulated treatments. If the direction of comparison did not matter, we would expect to see the same rejection proportions for this setting as we have seen for 1:49 setting. This is however not the case. A partial explanation to this can be found in the fact that the number of treatments that are like the reference is none in the first setting and is almost equal to the total number of simulated treatment in the last setting (49:1). The denominator for RE rate is 480,000 here while the number of rejections when Ho is false will be divided by 10,000 and multiplied by 100 for RP. As stated above, in all these settings, the denominator of the relative proportions change while those of the actual proportion do not change as the number of simulations is fixed at 10,000.

These various settings are developed in order to assess the strength of the performances of the two tests in some real life situations when we have multiple different treatments while having the knowledge of some of the treatments being similar to others in some fashion. In other words, we are trying to "trick" the two procedures under study while assessing their ability to detect differences among the means. We also do want to generalize the findings here knowing that the direction of comparison matters in the way the procedures behave. A test procedure that rejects the null hypothesis when it is false more often is said to be more powerful than another that fails to reject it while all other conditions are set to same. A test that fails to reject the null hypothesis when it is true more often will be controlling type I error at a better rate than that which rejects it.

3.4.2. Settings for 100, 150 and 300 Treatments Comparisons

For the remaining simulated cases, the number of treatments is progressively increased and the settings are set for each case. The following table (Table 3.2) summarizes the different simulated settings for each of the four cases in this study: 50, 100, 150 and 300 simulated treatments. The column "Number of true simulated null hypotheses multiplied by 10,000" contains the denominator for computing RE while the "Number of comparisons when Ho is false multiplied by 10,000" column contains the denominator for computing RP.

Table 3.2. Summary of Simulated Settings for all Cases

Simulated Treatments	Settings	Number of true simulated null hypotheses multiplied by 10,000	Number of comparisons when Ho is false multiplied by 10,000
50	1:49	0	490,000
	10:40	90,000	400,000
	25:25	240,000	250,000
	40:10	390,000	100,000
	49:1	480,000	10,000
100	1:99	0	990,000
	25:75	240,000	750,000
	50:50	490,000	500,000
	75:25	740,000	250,000
	99:1	980,000	10,000
150	1:149	0	1,490,000
	25:125	240,000	1,250,000
	50:100	490,000	1,000,000
	75:75	740,000	750,000
	100:50	990,000	500,000
	125:25	1,240,000	250,000
	149:1	1,480,000	10,000
300	1:299	0	2,990,000
	50:250	490,000	2,500,000
	100:200	990,000	2,000,000
	150:150	1,490,000	1,500,000
	200:100	1,990,000	1,000,000
	250:50	2,490,000	500,000
	299:1	2,980,000	10,000

3.5. Sample Size Increase

In the first stage of the study, we measure the performance of the two tests when the constant variance and normality assumptions are met and the number of replicates per simulated treatment is 3. In order to assess the impact of an increase in sample size, we repeat all these settings and increase the sample size of the simulated treatments to 5 while the assumptions are still met. Type I error is based on the comparisons where we know that the null hypotheses are true while power is assessed on comparisons where we know that the treatments are different from the reference as in the case of 3 replicates. We expect both Dunnett's test and FDR method to perform better in terms of power across all settings when the sample size is increased from 3 to 5. So for each of the 50, 100, 150 and 300 number

of simulated treatments in each stage, we generate two tables, one with the results when the sample size is 3 and a second table when the sample size is 5.

In the second stage of the study we changed the distribution of the simulated population from normal to a contaminated normal distribution and again each setting will be simulated with 3 replicates first and again with 5 replicates per treatment.

3.6. Simulation from a Contaminated Normal Distribution

In real life data, we oftentimes do not come across normally distributed data. We design this part of the study to accommodate that. We chose a contaminated normal distribution for the final data to undergo the two procedures under study. This distribution is also referred to as a mixture of normal distributions. We mix two normal distributions at a proportion of 90/10, with a 90% chance that the first normal distribution (to be mixed) be derived from the original population used in our normal distribution scenario and the remaining 10 percent chance is to allow a different population to enter the sample. In this case the latter is a normal distribution with the same mean of the original data and a new standard deviation that is twice that of the original population. The mixture of these two distributions forms the population on which the second stage of our study is performed to assess the impact of such a mixture on the performance of FDR method and Dunnett's test. In those situations where our real data contain outliers that clearly affect the normality of the data and we are still interested in the appropriate method to use to compare treatments means, this second stage of this study may be of help. We expect both tests to have less power across all settings and all cases in this second stage.

We will refer to this second stage of our simulation the Contaminated Normal Stage. We will again study the four cases of simulated treatments and each of these cases (50, 100, 150, and 300) will have the same set of settings as in the normal distribution stage of the study. The performances of the two tests will also be assessed using 3 replicates per simulated treatment and also using 5 replicates. And we will again expect the latter to offer more power to both tests under all settings. Any change observed here relative to the same situation in the first stage will therefore be attributed to the contamination of the normal distributions. How much of the power lost due to the contamination will be recovered by the increase in sample size is unclear at this point.

CHAPTER 4. RESULTS OF SIMULATION STUDY

4.1. Simulation from a Normal Distribution with 3 Replicates per Treatment

4.1.1. Case 50N3 – 50 Simulated Treatments

A summary of the results from this case is in Table 50N3 below. 50 treatments were simulated with 3 replicates. RE holds at roughly 5% for Raw and 0.20% for Dunnett's. Raw here represents the percentages where no attempts were made to control for the inflation of the alpha level with multiple testing. Only FDR method's RE varies when the number of true null hypotheses increases from the first to the last setting. Dunnett's test controls FWE from a very low rate of below 1.59% and increases progressively to reach a high of 5% with setting 49:1. FDR method's AE is the only one that decreases over time, i.e. as the number of true null hypotheses increases. It started at a 12.22% and increased to a peak at 14.69% for the mid-setting (25:25) before decreasing to a low of 4.37% at setting 49:1. The RP of Dunnett's test does not change as the number of true null hypotheses increases. On the other hand, FDR method's RP decreases as the number of true null hypotheses increases. The AP of both tests stay about the same for both tests until the last setting when they differ and are each equal to their RP respectively. In setting 1:49, the number of true null hypotheses is zero, since all treatments are simulated from a different population than the control. All rejections in this case therefore represent true rejections and account for power. This is also the case for all first settings in all cases and is the reason why the first two cells in the results tables are empty.

Table 4.1. Summary of Rejections Percentages in Case 50N3

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

The relative percentages do not change much for Dunnett's test which keeps roughly 0.20% type I error for all settings with about 23% RP. Dunnett's test FWER increases from 1.56% to 5.06% as the number of true null hypotheses increases while its AP decreases from 85% to 23.01% (RP at the last setting). At the highest number of true null hypotheses, Dunnett's test is still controlling FWER at about 5% confirming its firm grip on type I error control. Dunnett's RP is lower than FDR method's in setting 1:49. By the last setting, Dunnett's test RP is higher than FDR method's. However, FDR's AE started at 12.22% in the first setting before settling down at 4.37% which is lower than Dunnett's AE when the number of true Ho is at its highest, although, Dunnett's test performs consistently better overall than FDR method in AP and AE across all settings. In addition, the percentage difference of AP between both tests appears to be negligible compared to the percentage difference in RPs for all settings. The following two figures present a complete visual of the trends in the rejections percentages for both tests.

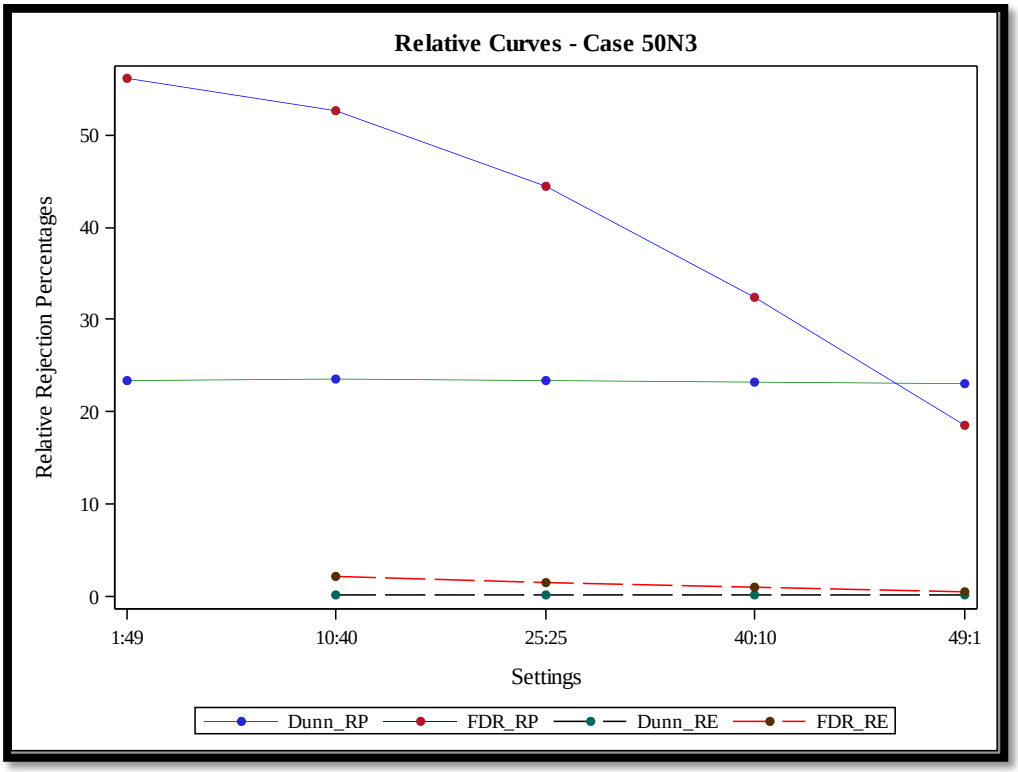

Figure 4.1. RP and RE Curves for FDR Method and Dunnett's Test in Case 50N3

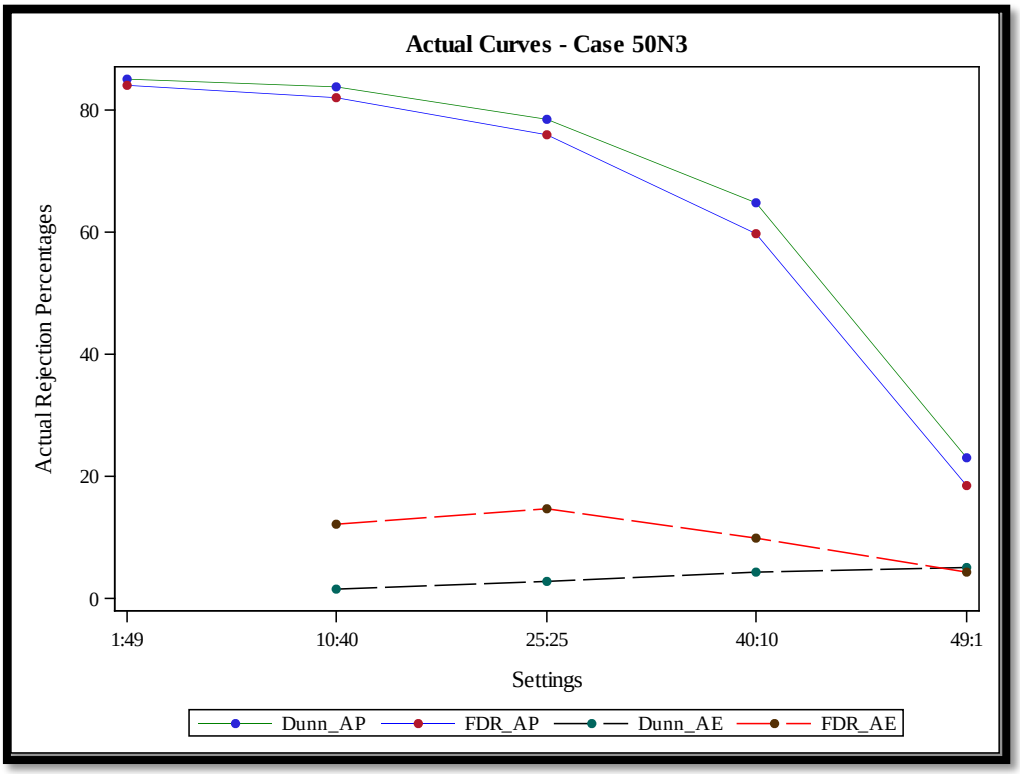

Figure 4.2. AP and AE Curves for FDR Method and Dunnett's Test in Case 50N3

4.1.2. Case 100N3 – 100 Simulated Treatments

Table 4.2. Summary of Rejections Percentages in Case 100N3

		Type I Error		**Power**	
SETTINGS	**TESTS**	Relative (RE)	Actual (AE/FWER)	Relative (RP)	Actual (AP)
1:99	Raw	___	___	68.99	99.85
	Dunnett's test			19.39	87.61
	FDR Method			57.34	85.97
25:75	Raw	5.05	45.11	68.40	99.73
	Dunnett's test	0.10	1.76	19.38	84.73
	FDR Method	1.97	19.04	51.42	82.18
50:50	Raw	5.08	60.82	68.62	99.59
	Dunnett's test	0.11	3.03	19.43	81.13
	FDR Method	1.47	19.76	45.08	77.85
75:25	Raw	4.99	70.14	68.67	99.18
	Dunnett's test	0.10	4.45	19.39	73.97
	FDR Method	0.93	14.67	35.00	68.24
99:1	Raw	4.97	75.72	68.71	68.71
	Dunnett's test	0.11	5.11	19.45	19.45
	FDR Method	0.41	4.04	14.84	14.84

Table 4.2 represents the relative and actual type I error rate and power when 100 treatments are simulated from a normal population with 3 replicates per treatment. The number of true null hypotheses is increased here in increment of 25. At double the number of simulated treatments (from 50N3 to 100N3), the impact on the performances of the two tests are not proportional to the increase in simulated treatments. Dunnett's test RPs decrease slightly compared to case 50N3 while FDR method's RP has a slight increase. This trend is observed until the number of true null hypotheses is at its highest in the last setting (99:1) where we observe a decline in the RP and AP for both tests. Also at this setting, FDR method has an AP of 14.84% (18.56 for 50N3) and Dunnett's test has an AP of 19.45% (against 23.01% in 50N3). A more significant increase in AE is observed in FDR method while Dunnett's test AE remains unchanged.

The increase in the total number of simulated treatments affects positively FDR method's AP while its impact is negative on its AE. For Dunnett's the increases in the total number of simulated treatment has a negative effect on its ability to detect significant differences in relation to the complete

family of simulated comparisons while its error level experiences a small change. So, Dunnett's test still controls FWER strongly. The observed trends in the relative percentages compared to the actual percentages and in relation to the progressive increase in the number of true null hypotheses are same here compared to 50N3. At 50:50, FDR's RP is roughly about the same as in case number 50N3. The following two figures present a complete visual of the trends in the rejections percentages for both tests.

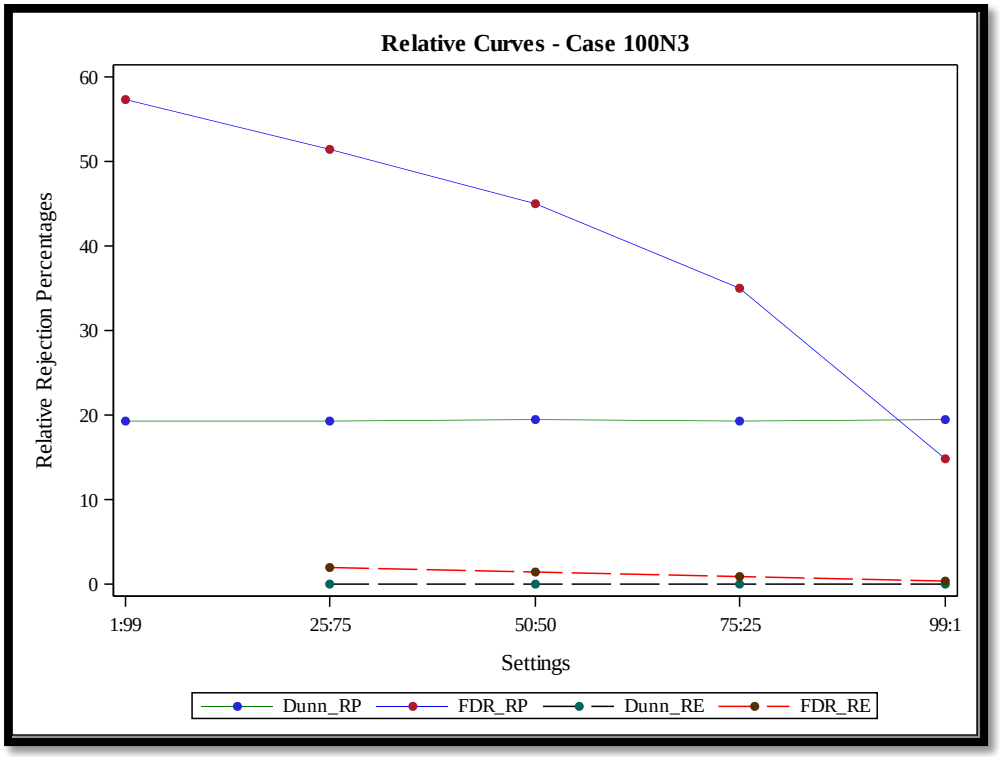

Figure 4.3. RP and RE Curves for FDR Method and Dunnett's Test in Case 100N3

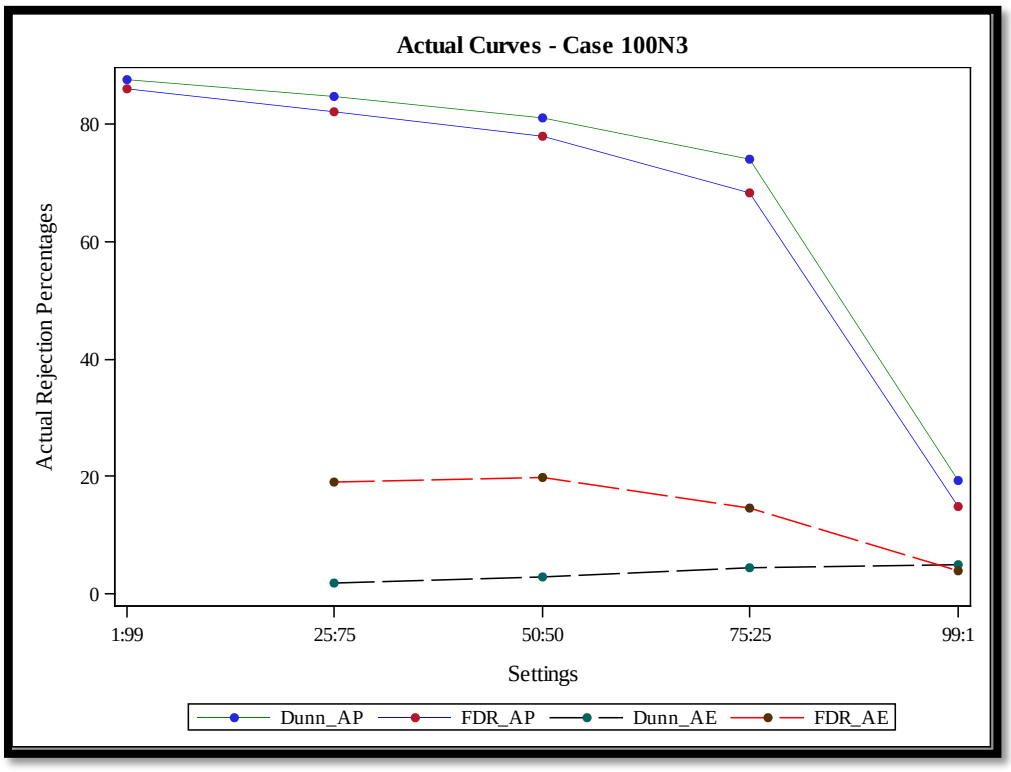

Figure 4.4. AP and AE Curves for FDR Method and Dunnett's Test in Case 100N3

4.1.3. Case 150N3 – 150 Simulated Treatments

Table 4.3. Summary of Rejections Percentages in Case 150N3

		Type I error		Power	
SETTINGS	**TESTS**	Relative (RE)	Actual (AE/FWER)	Relative (RP)	Actual (AP)
1:149	Raw	___	___	68.09	99.90
	Dunnett's test			17.07	87.93
	FDR Method			55.74	85.75
25 : 125	Raw	4.97	45.38	68.35	99.87
	Dunnett's test	0.06	1.12	17.09	86.72
	FDR Method	2.13	20.19	53.04	84.27
50 : 100	Raw	5.01	61.03	68.77	99.84
	Dunnett's test	0.07	2.16	17.38	85.49
	FDR Method	1.85	23.80	49.95	81.97
75 : 75	Raw	5.02	71.06	68.32	99.77
	Dunnett's test	0.07	3.03	17.16	82.33
	FDR Method	1.40	22.73	44.64	78.77
100 : 50	Raw	5.04	76.53	68.56	99.62
	Dunnett's test	0.08	3.85	17.01	79.26
	FDR Method	1.11	18.99	38.38	74.21
125 : 25	Raw	4.97	81.82	68.58	99.11
	Dunnett's test	0.07	4.61	17.24	70.09
	FDR Method	0.68	13.33	29.82	63.84
149:1	Raw	4.92	84.05	69.40	69.40
	Dunnett's test	0.07	4.71	16.72	16.72
	FDR Method	0.31	3.67	12.18	12.18

Dunnett's test RE is closer to case 100N3 than case 100N3 is to case 50N3. Dunnett's test also lost some percentage of RP here compared 100N3. FDR's RP started again around the 57% rate as in previous first settings with no true null hypotheses. As before, when the number of simulated treatments from the same population as the control is almost the number of simulated treatments from the other population, the RP of FDR is again at around 45%, same as in the previous mid-settings for cases 50N3 and 100N3. In other words, FDR has not seen a big variation in RP when the number of simulated treatments from the same population as the control is one short of the total number of simulated treatments from the other population. It has also kept its RE at around 1.50%. Both tests again have same RP and AP. This is the same trend we observed in the previous cases. The following two figures present a complete visual of the trends in the rejections percentages for both tests.

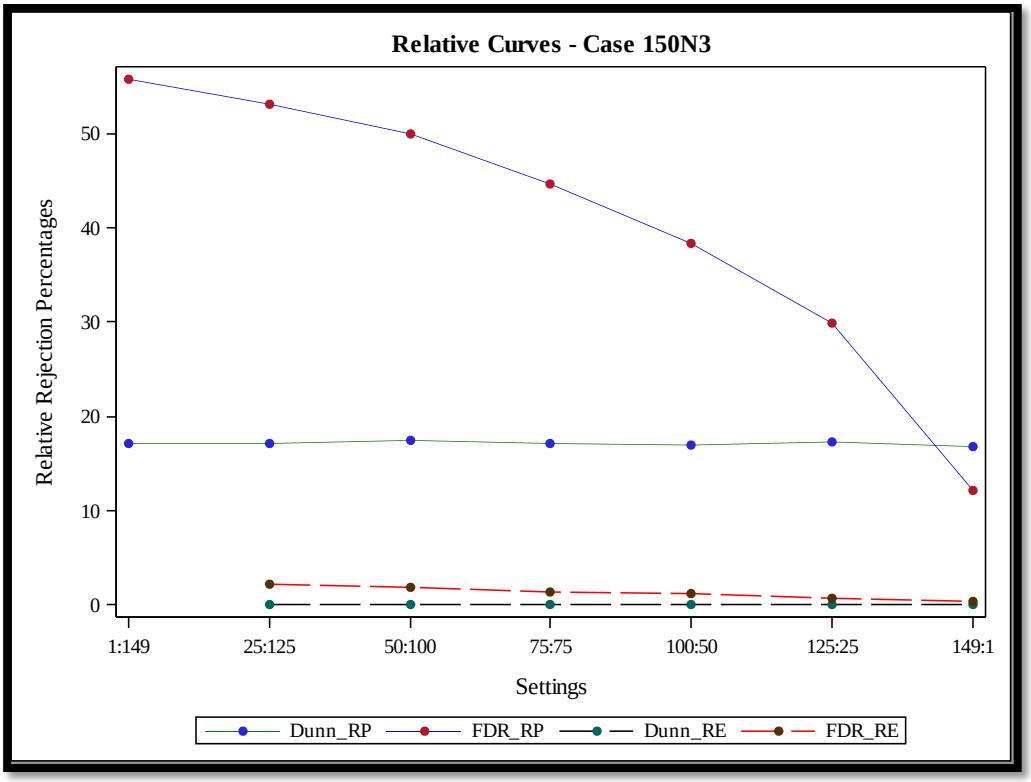

Figure 4.5. RP and RE Curves for FDR Method and Dunnett's Test in Case 150N3

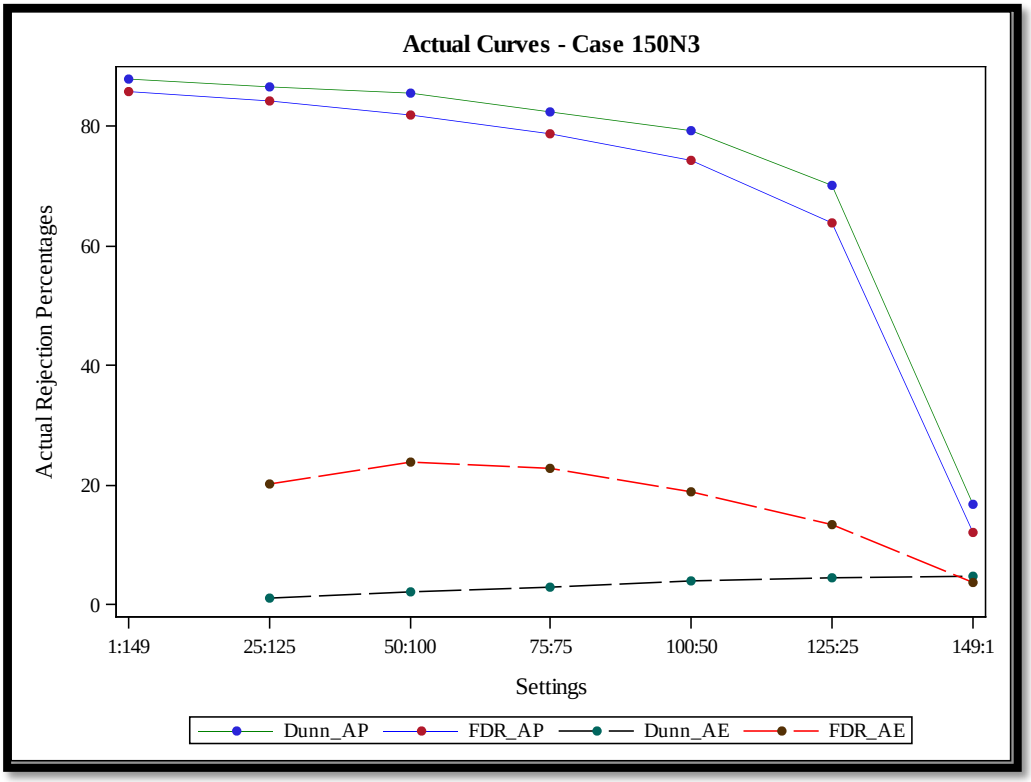

Figure 4.6. AP and AE Curves for FDR Method and Dunnett's Test in Case 150N3

4.1.4. Case 300N3 – 300 Simulated Treatments

Table 4.4. Summary of Rejections Percentages in Case 300N3

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

FDR method's RP stayed about the same except for the last treatment when the number of true null hypotheses is at its highest. In fact, both relative and actual power of FDR method slightly increased from 150 to 300 simulated treatments, again with the exception of the last setting. The APs of both tests are still high and at the same level as case 50N3. Dunnett's RE is lower while its FWER remains unchanged. Its RP is however affected in about the same fashion with the doubling of the total number of simulated treatments from 50 to 100. Both tests lost RP and AP at the last setting, but FDR method's AP remains higher than in case 150N3 until setting 250:50 before decreasing to a low of 9.13% at 299:1. The following two figures present a complete visual of the trends in the rejections percentages for both tests.

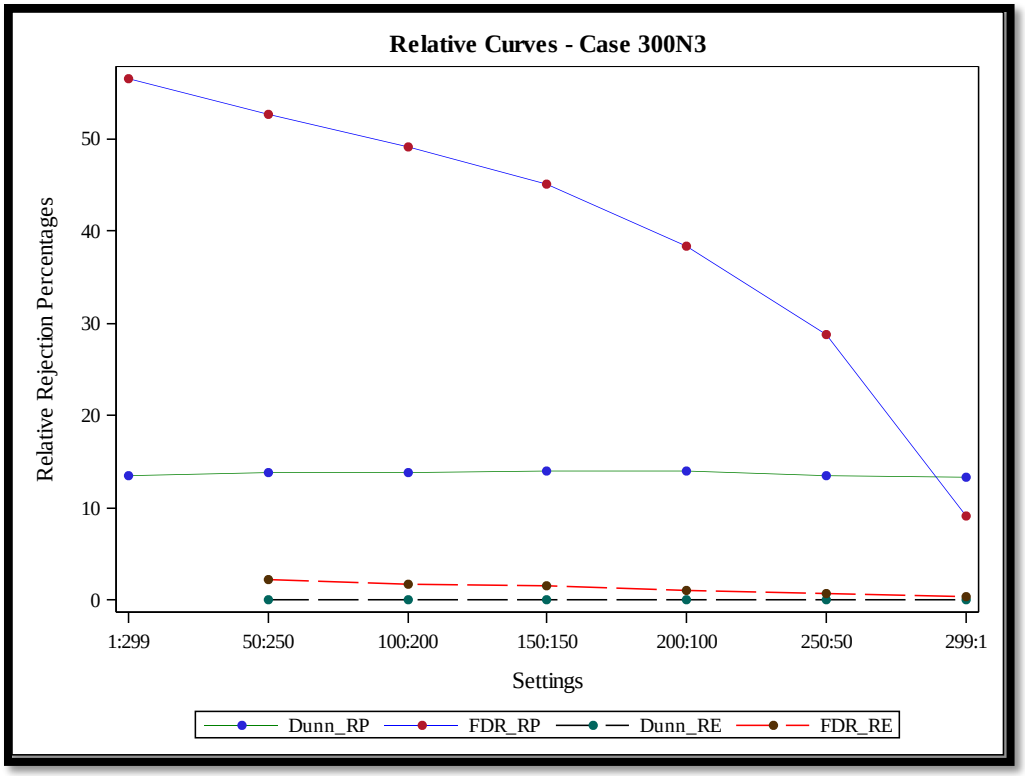

Figure 4.7. RP and RE Curves for FDR Method and Dunnett's Test in Case 300N3

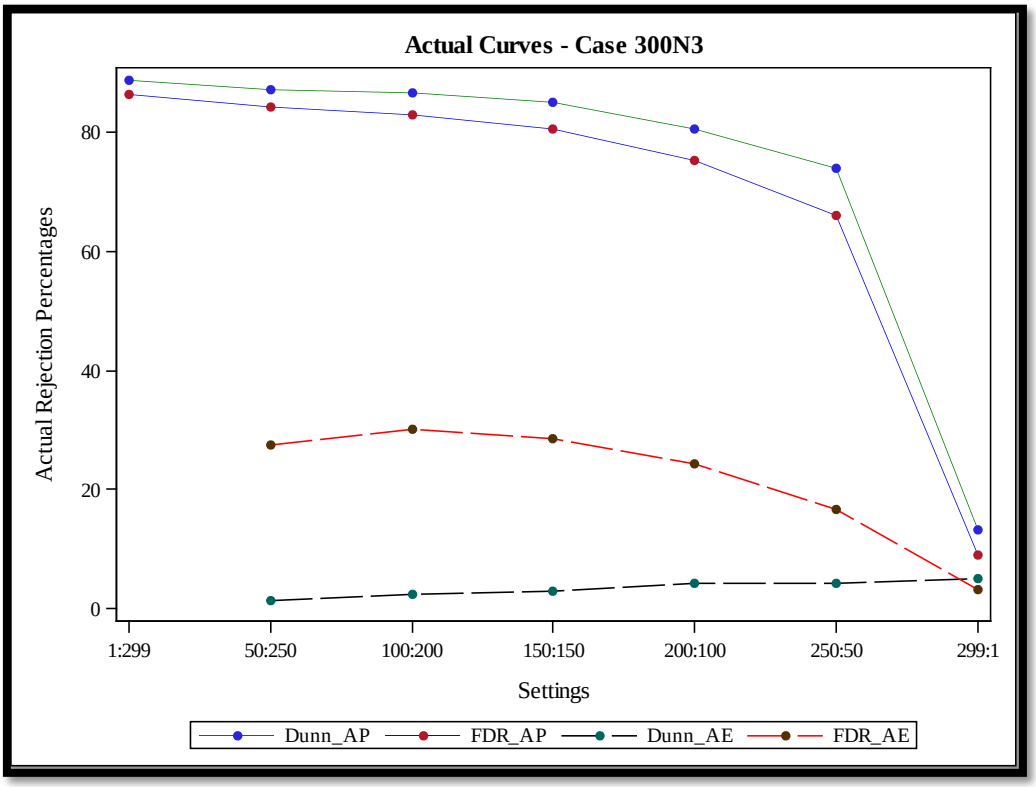

Figure 4.8. AP and AE Curves for FDR Method and Dunnett's Test in Case 300N3

4.2. Simulation from a Normal Distribution with 5 Replicates per Treatment

4.2.1. Case 50N5 – 50 Simulated Treatments

When the number of replicates is increased to 5, both tests perform better in power. Their gain in power is also seen in an increase in the AE percentages. FDR method's RPs are higher than Dunnett's up to setting 40:10 before plummeting to 6% lower than Dunnett's in the last setting (49:1). It is worth noting here that there is no noticeable difference between APs of both tests until the last setting when the number of true null hypotheses is at its highest and RP=AP is still verified at this point for each test. Dunnett's test detects at least 50% of differences in any of the setting both relative and actual still with its tight hold on type I error.

Compared to case 50N3, the RP of Dunnett's test more than doubled and Dunnett's test still maintains a constant RP irrespective of the number of true null hypotheses being simulated. Dunnett's test's type I error also does not change from case 50N3. In other words, the RE in 50N3 is same as in 50N5. Sample size increase therefore does not have any impact on the ability of Dunnett's test to control type I error. This is not the case for FDR method since FDR does not explicitly control type I error. The increase in sample size does not affect the type I error rate (relative or actual) of either test when the number of true Ho is at its highest.

Table 4.5. Summary of Rejections Percentages in Case 50N5

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

Dunnett's test keeps a constant RP across all settings as discussed earlier. FDR method's RE decreases progressively to reach the same level as Dunnett's RE. FDR method's RP decreases as the number of true null hypotheses increases. The decline in RP has a smaller slope before the mid-setting (25:25) than after the mid-setting when that slope decreases rapidly to an intersection point where both tests are performing the same in terms of RP and that intersection happens just before the last setting when the number of true null hypotheses is close to its highest. The following two figures present a complete visual of the trends in the rejections percentages for both tests.

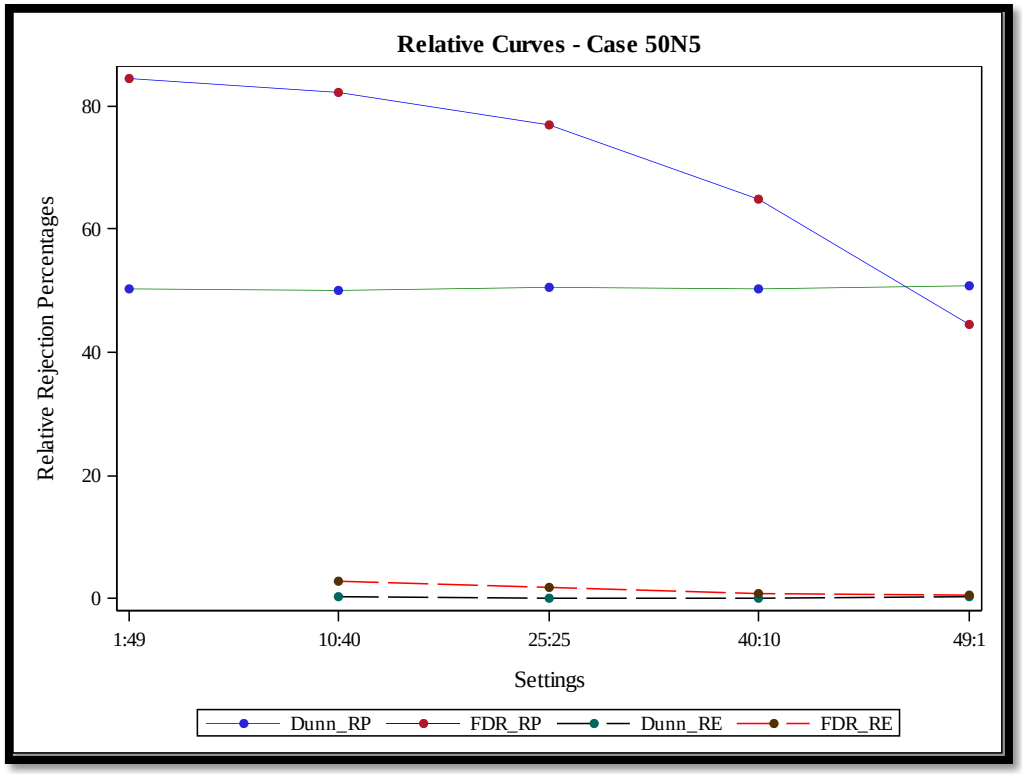

Figure 4.9. RP and RE Curves for FDR Method and Dunnett's Test in Case 50N5

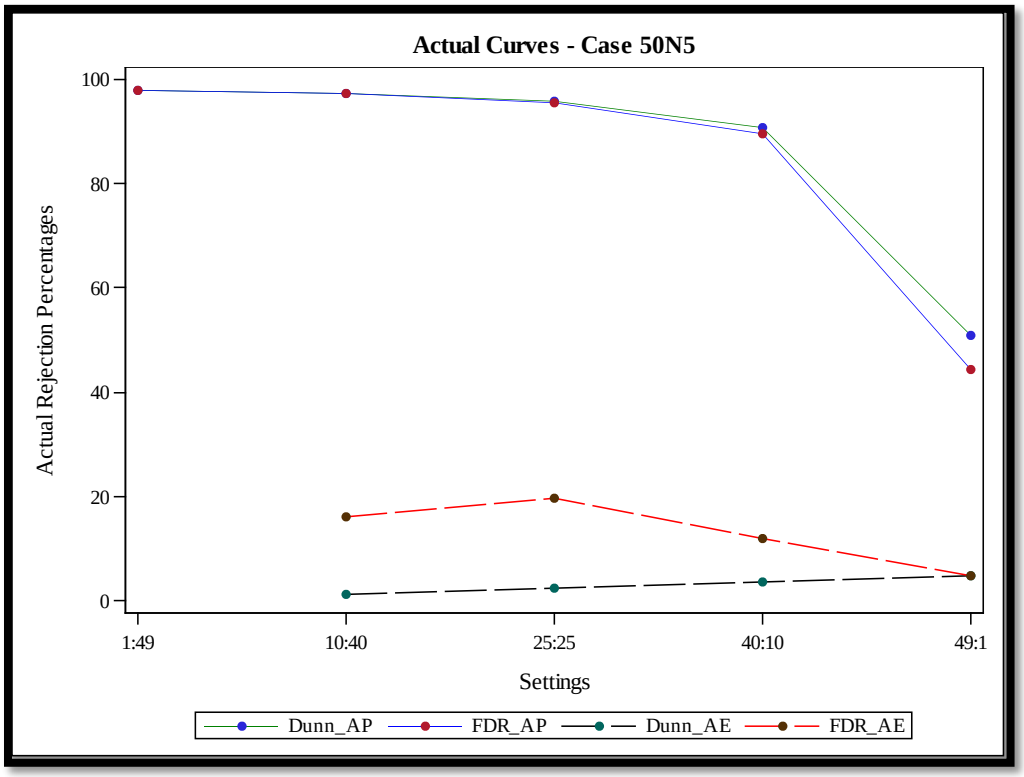

Figure 4.10. AP and AE Curves for FDR Method and Dunnett's Test in Case 50N5

4.2.2. Case 100N5 – 100 Simulated Treatments

Table 4.6. Summary of Rejections Percentages in Case 100N5

		Type I error		**Power**	
SETTINGS	**TESTS**	Relative (RE)	Actual (AE/FWER)	Relative (RP)	Actual (AP)
1:99	Raw	___	___	88.23	99.98
	Dunnett's test			44.13	98.06
	FDR Method			84.08	97.80
25:75	Raw	4.88	44.93	88.57	100.00
	Dunnett's test	0.11	1.86	44.45	97.59
	FDR Method	2.39	27.82	81.65	97.30
50:50	Raw	5.23	61.33	88.49	99.98
	Dunnett's test	0.13	3.50	44.48	96.63
	FDR Method	1.88	28.64	76.41	96.57
75:25	Raw	4.99	71.12	88.28	99.98
	Dunnett's test	0.09	3.95	43.88	94.05
	FDR Method	0.98	20.11	67.04	93.00
99:1	Raw	4.98	76.93	88.30	88.30
	Dunnett's test	0.10	4.89	**43.37**	**43.37**
	FDR Method	0.37	4.24	**35.76**	**35.76**

Again, Dunnett's test has a lower type I error rate almost half the rate when 50 treatments were simulated. The RP for Dunnett's test has also doubled and remained the same regardless of the number of true null hypotheses. The half-setting RP for FDR method remains at around 76%, same as in the case of 50 simulated treatments. This number was around 45% when the sample size was 3. As in the previous case when the number of simulated treatment is 50, FDR Method's RP doubled compared to the case 100N3 only when the number of true null hypotheses is at its highest in the 99:1 setting. The RE for FDR method at the highest number of simulated treatments is slightly lower here than when sample size was 3. Both tests again start at high and relatively same AP when there is no true null hypotheses. At setting of 25:75, FDR Method has an AP of 97.30% with an AE of 27.82%. At the same setting, Dunnett's test has a 97.59% AP for only 1.86% of FWER. This makes Dunnett's test attractive to researchers. The biggest difference in AP between both tests is again at the highest number of true null hypotheses with Dunnett's test leading with 43.37% against 35.76% for FDR and at the same setting, while both have

relatively same type I error rate below 5%. FDR Method still increases in type I error rate until the half-setting where that rate plummets to 20.11 and then 4.24 for the last setting.

The same trend is observed here on the two curves. However, Dunnett's test green curve (RP) is flatter here than in the previous case. The slope after the mid-setting here is also greater than that of the slope after the mid-setting in the previous case. The following two figures present a complete visual of the trends in the rejections percentages for both tests.

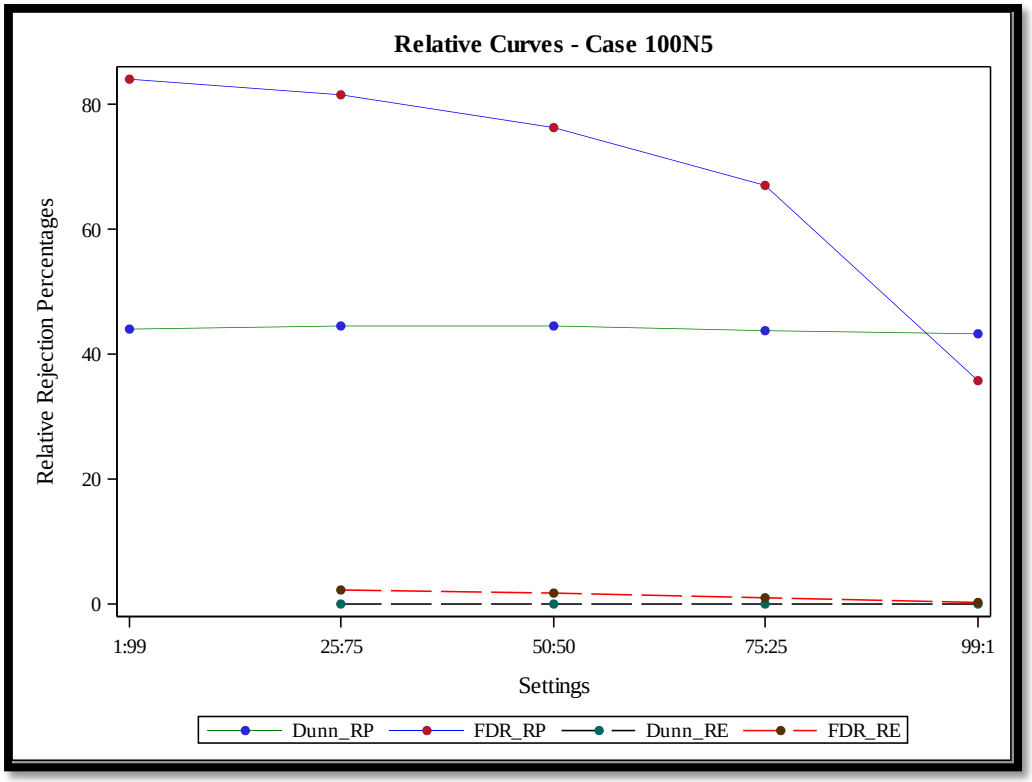

Figure 4.11. RP and RE Curves for FDR Method and Dunnett's Test in Case 100N5

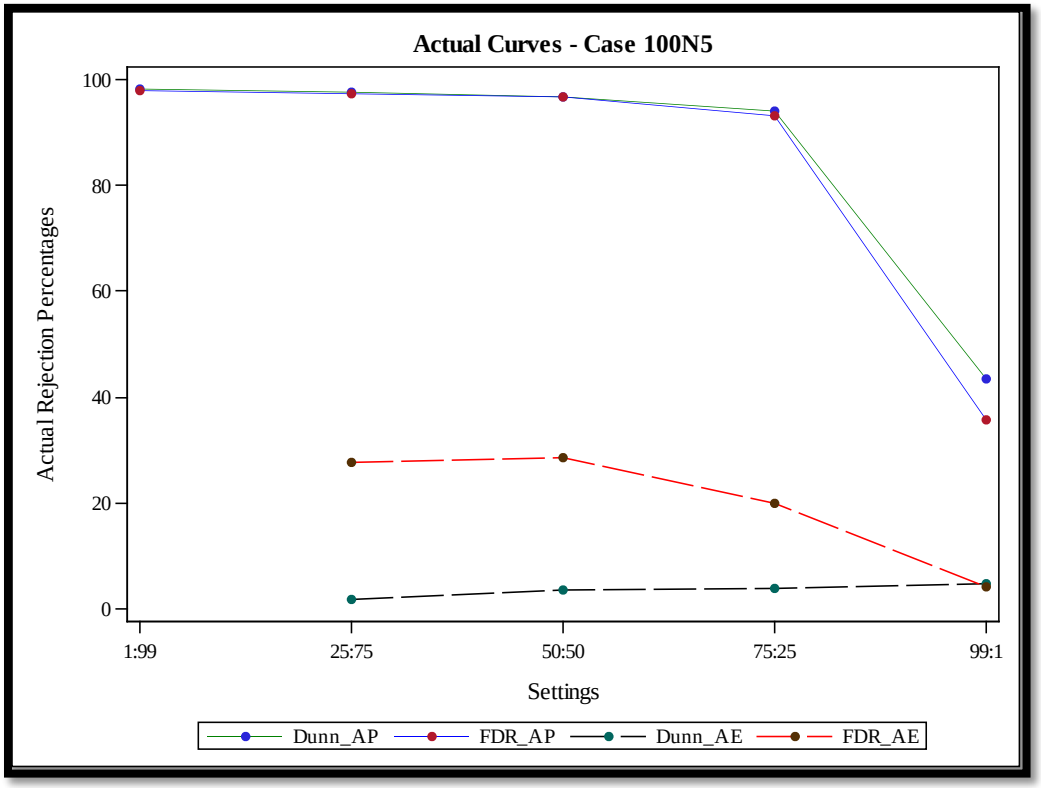

Figure 4.12. AP and AE Curves for FDR Method and Dunnett's Test in Case 100N5

It is also interesting to see that both FDR method and Dunnett's test keep their AP above 50% for the most part with the exception of the last settings. Considering the performance of both tests, Dunnett's test offers an attractive or better package of comparison of treatments. Looking at the AP power curves, FDR_AE increases as the total number of treatments increases while DUNN_AE stays relatively the same. This is expected because Dunnett's test controls FWER while FDR does not. The AP curves present Dunnett's test as a better alternative than FDR method. Its AP are continuously the same as FDR method's and at the same time, its AE is much lower than that of Dunnett's test. Even at the last setting when both tests' AP plummeted, Dunnett's test AP is still higher than that of FDR method in all cases. The RP of Dunnett's test doubles when compared to the same case when the number of replicates is 5. FDR's RP also has a significant increases from 3 replicates to 5 replicates. Both tests RPs still follow the same trend (constant for Dunnett's test and decreasing slope for FDR method).

4.2.3. Case 150N5 – 150 Simulated Treatments

Table 4.7. Summary of Rejections Percentages in Case 150N5

		Type I error		**Power**	
SETTINGS	**TESTS**	Relative (RE)	Actual (AE/FWER)	Relative (RP)	Actual (AP)
1:149	Raw	___	___	88.39	100.00
	Dunnett's test			40.84	98.34
	FDR Method			84.44	97.96
25 : 125	Raw	4.95	45.11	88.78	100.00
	Dunnett's test	0.08	1.36	41.29	98.10
	FDR Method	2.75	30.57	83.07	97.89
50 : 100	Raw	4.98	60.86	88.45	100.00
	Dunnett's test	0.07	2.13	40.70	97.65
	FDR Method	2.18	36.69	80.04	97.46
75 : 75	Raw	4.96	70.95	88.51	99.99
	Dunnett's test	0.07	3.03	40.46	97.15
	FDR Method	1.67	34.13	76.25	96.70
100 : 50	Raw	5.08	77.41	88.09	99.99
	Dunnett's test	0.07	3.99	40.44	95.83
	FDR Method	1.17	27.17	70.45	95.02
125 : 25	Raw	4.91	81.20	88.34	99.96
	Dunnett's test	0.07	4.37	40.10	92.50
	FDR Method	0.75	17.07	61.11	90.58
149:1	Raw	5.12	84.75	88.23	88.23
	Dunnett's test	0.08	5.31	**40.64**	**40.64**
	FDR Method	0.37	4.39	**33.19**	**33.19**

When there are no true null hypotheses (settings 1:49, 1:99, 1:149 and 1:299), FDR Method has been fairly constant (around 84%) for 50, 100 as well as 150 simulated treatments. This was not apparent when the number of replicates was set to 3. A bigger sample helped both tests to improve their power and error rates. This doubling of their performances is consistent across all cases. The following two figures present a complete visual of the trends in the rejections percentages for both tests.

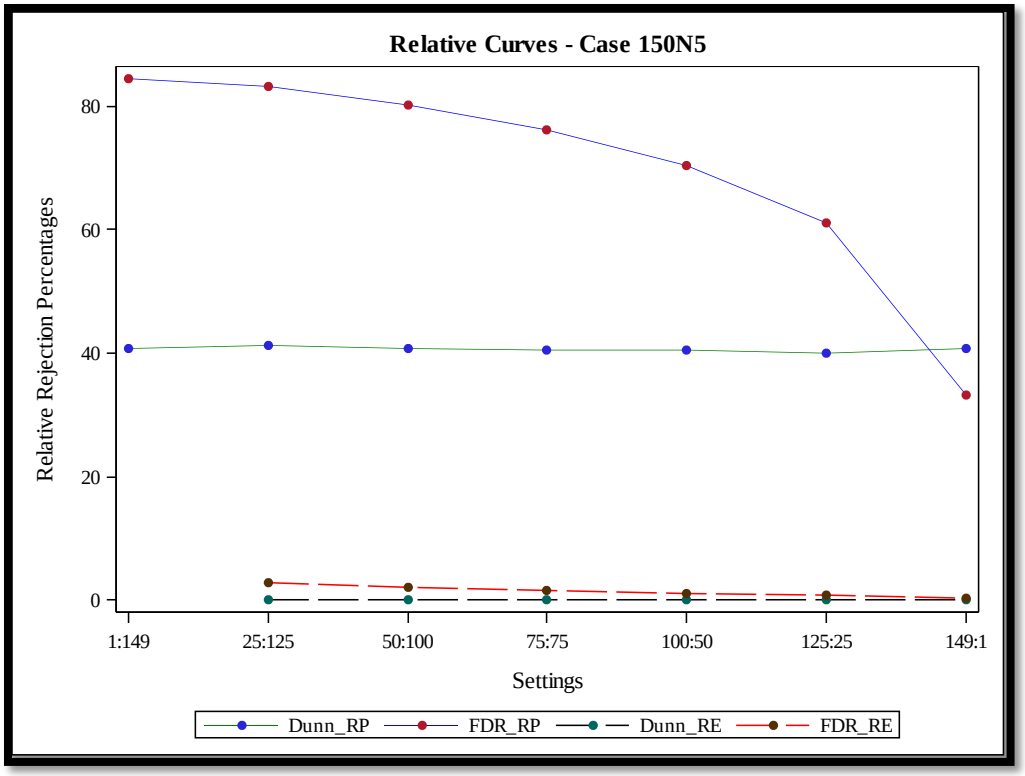

Figure 4.13. RP and RE Curves for FDR Method and Dunnett's Test in Case 150N5

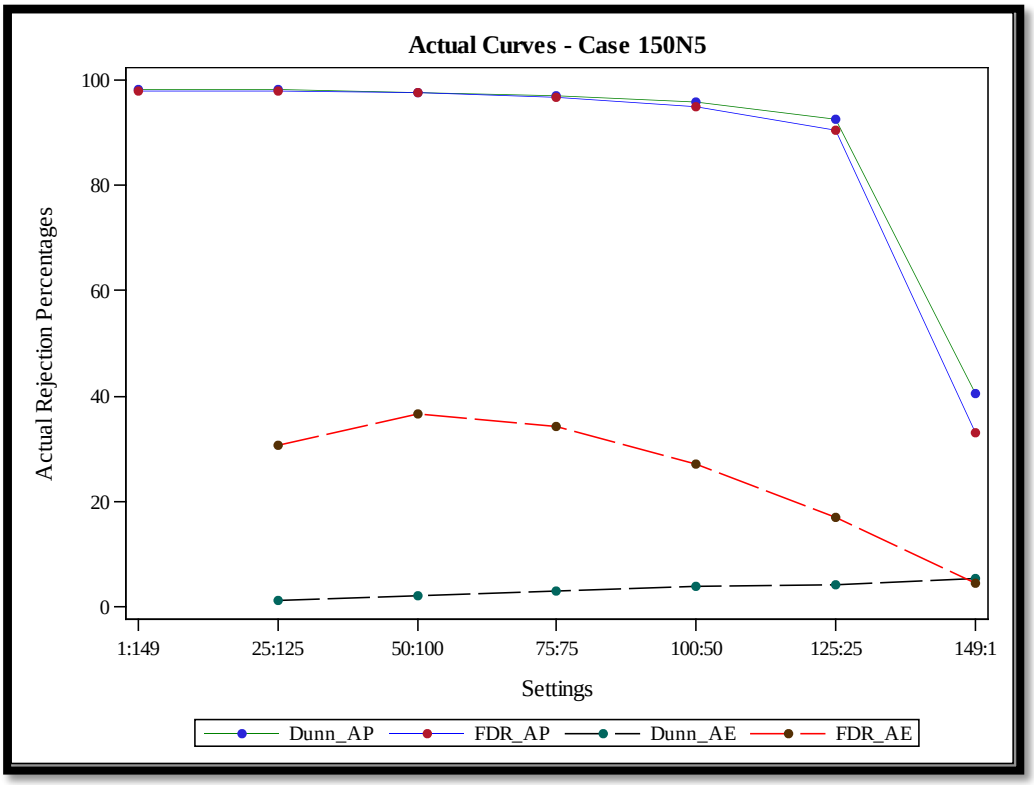

Figure 4.14. AP and AE Curves for FDR Method and Dunnett's Test in Case 150N5

FDR method's RP still starts around the same point (about 84%) at the first setting, while Dunnett's RP gets lower as the total number of simulated treatments increases. Here Dunnett's RP is sitting at a flat of about 40% throughout all settings. The difference between the type I errors is not big enough to convince a researcher to use Dunnett's test in lieu of FDR method here especially when no null hypothesis is true.

4.2.4. Case 300N5 – 300 Simulated Treatments

The same trend in performance is observed in 300N5. Table 300N5 summarizes the results.

Table 4.8. Summary of Rejections Percentages in Case 300N5

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

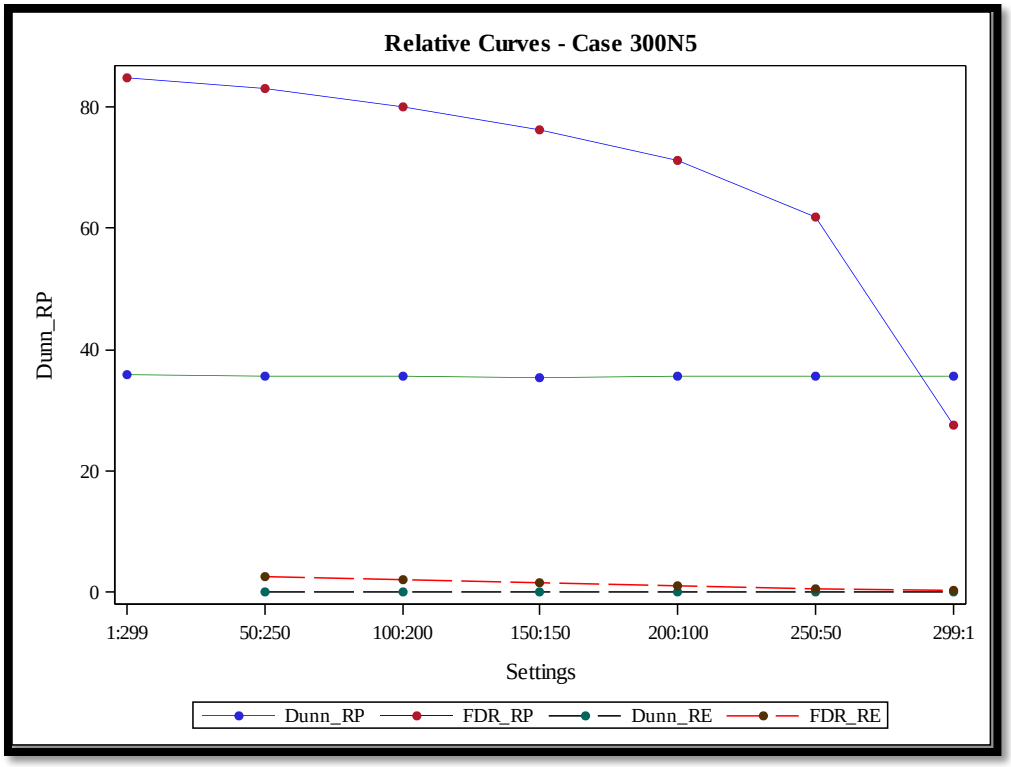

Figure 4.15. RP and RE Curves for FDR Method and Dunnett's Test in Case 300N5

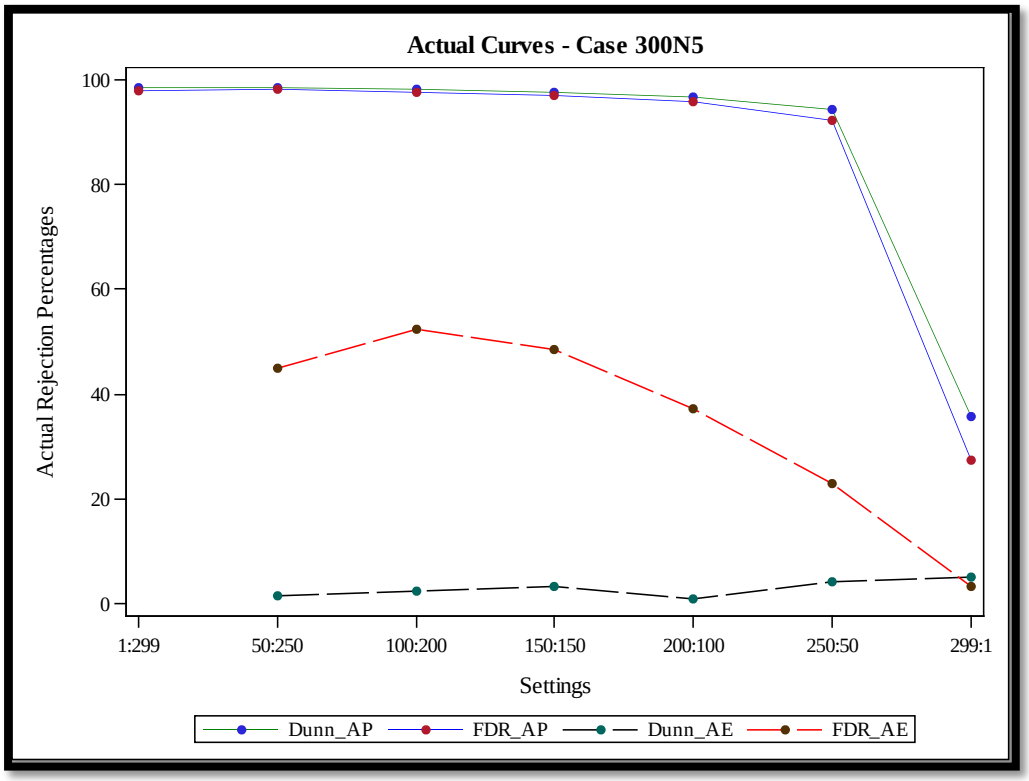

Figure 4.16. AP and AE Curves for FDR Method and Dunnett's Test in Case 300N5

For Dunnett's test, there is a decreasing trend across setting and by case for RP while RE remains low, though lower than FDR method. A high number of treatments has a direct impact on the ability of Dunnett's test to detect true differences among the treatments while FDR method's RP does not seem to be affected by an increment in the total number of treatments. In addition, the only time a decreasing trend is recorded for FDR method from the smallest to the highest total number of simulated treatments is at the last setting when the total number of true null hypotheses is at its highest. The scale of this decrease is disproportional to the increase in treatments (18.56% at 50N3 to 9.13% at 300N3). FDR method's RP has the same trend as its RE both across treatments, starting at its highest and decreases as the number of treatments same as to the control increases. Furthermore, FDR method's RP remains higher than that of Dunnett's test for all but the last setting. Somewhere just before the last setting, the two RPs cross each other. That is also seen in the distributions of the rejections displayed in the above panels. This makes FDR method more attractive to researches with large comparisons especially since both tests' RE are reasonably low across settings and across treatments.

4.3. Results from the Contaminated Normal Distribution

To enable a convenient browsing of the pages of this paper, we will assemble all the tables in the second stage in appendix A. The trends observed in the first stage of the study are similar to the trends observed in the performances of FDR method and Dunnett's test in the contaminated normal stage. Overall the contamination has a negative impact on both tests. They are both less likely to detect true differences and they also make more type I errors both familywise and comparison-wise. The main remarks from the contamination can be summarized in the following points:

- RE is not as affected by the mixture as much as RP. This makes sense as power is adversely affected by the increase in variance buy type I error is minimally affected. For 150 simulated treatments for instance, Dunnett's test has a RE of 0.08% at setting 25:125 of 150N5 while that rate is 0.10% for the same setting in case 150CN5. RP for Dunnett's test is at 41.29% at the same setting for N5 against 26.18% in CN5. For setting 50:100, FDR method's RE is 2.18% for N5 versus 2.16% in CN5; RP is at 80.04% for N5 and 65.18% for CN5.

- The effect of the contamination is not disastrous on the APs of both tests. Only a slight decrease is observed. For instance, from 300N3 to 300CN3, FDR method' s AP started from 86.36% to 9.13% from the first setting to the last setting for N3 and from 82.29% to 5.79% for CN3. Dunnett's test is from 88.76% to 13.33% for N3 and 86.30% to 8.08% for CN3.
- FWER also does not seem to suffer considerably from the contamination. AE for FDR method is from 27.40% to 3.13% for N3 and 26.08% to 7.35% for CN3.
- Overall, RP is the one measure of comparison that seems to be most affected by the contamination. In the following chapter, we will present some graphs of the rejections made by the two tests in order to offer a visualization of the differences between the two tests.

4.4. Distribution of the Rejections

In order to present a visualization (in figure 4.17 and figure 4.18) of the rejection trends for both FDR Method and Dunnett's test over the various number of treatments simulated, we plot the number true rejections made by the methods across setting for each case with normal distribution and 5 replicates. The trends in these graphs are the same with 3 sample size and also for the contaminated normal cases. Dunnett's test makes 0 rejection in about 350 samples, 1 rejection in about 300 samples, 2 and 3 rejections in about 290 samples each. The sum of all the samples is the total number of simulated samples in our study which is 10,000. The remainder of these graphs for 5 replicates with the normal case can be found in Appendix B. Clearly, FDR finds all the differences about 21% of the times while Dunnett's test finds all of the differences about 2% of the times. Both tests detect none of the differences agout the same number of time (around 3.2%).

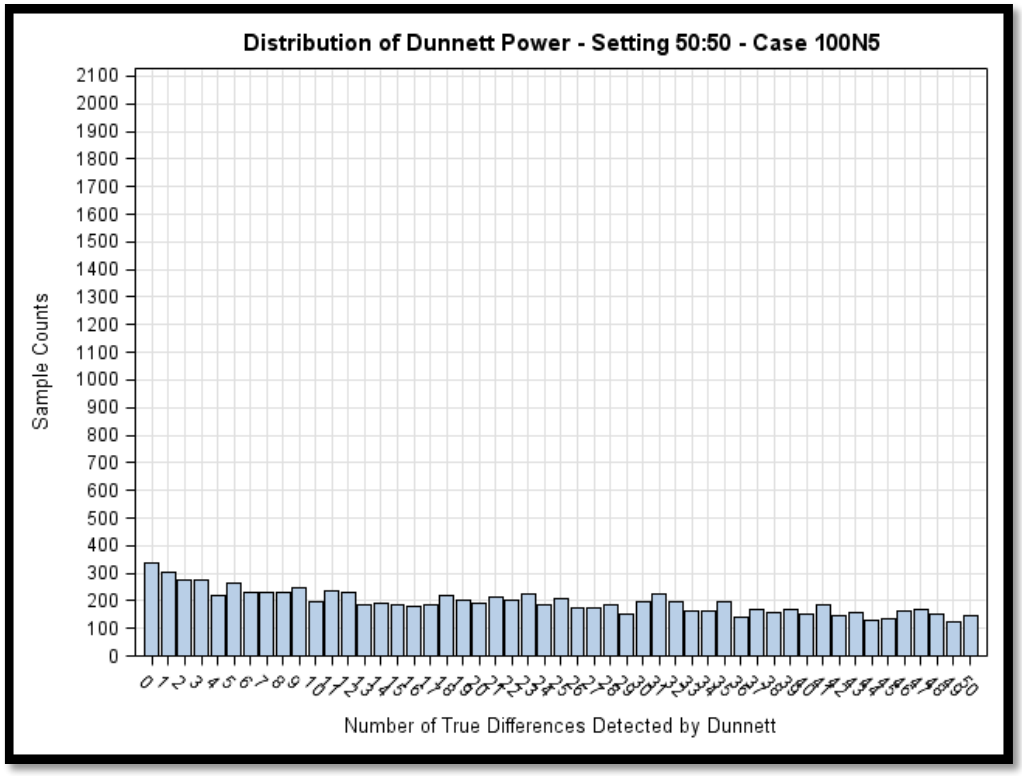

Figure 4.17. Distribution of Rejections by Dunnett's Test in Setting 50:50 for Case 100N5

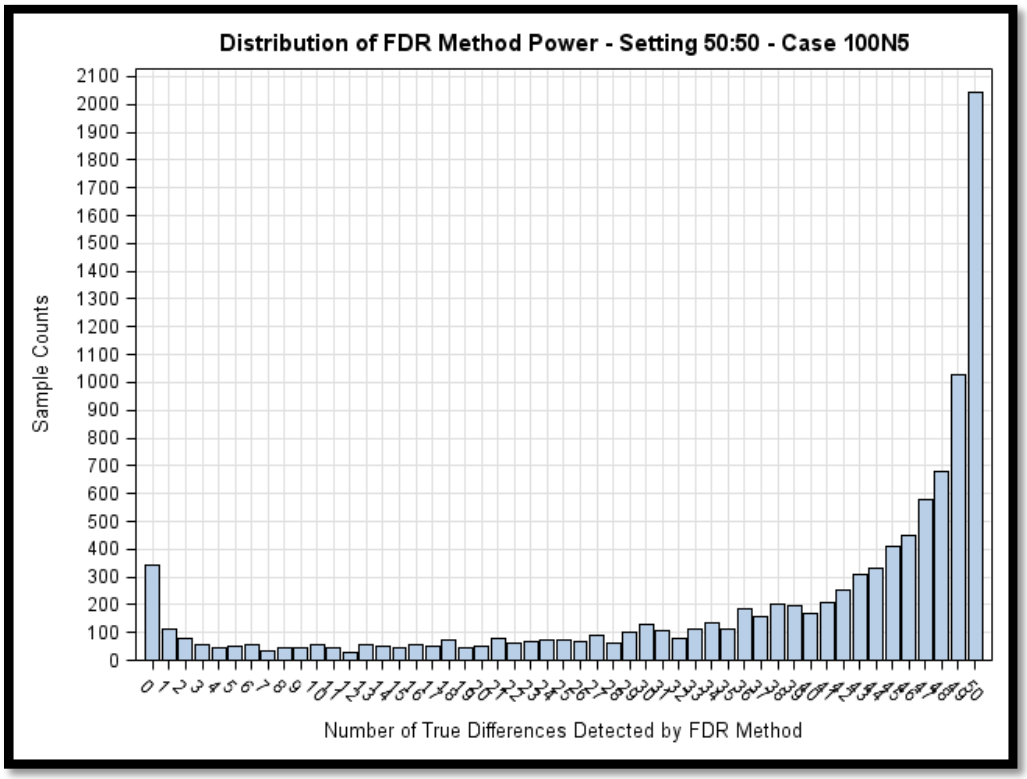

Figure 4.18. Distribution of Rejections by FDR Method in Setting 50:50 for Case 100N5

CHAPTER 5. DISCUSSION

The results displayed in the tables and graphs above present a variety of information as to the behavior of Dunnett's test and the False Discovery Rate method both individually and in relation to each other. In this chapter, we will analyze those behaviors in various sections. The general trend of the results show that in the contaminated distribution stage of the study, both tests lost some of their relative as well as actual powers compared to the same cases in the normal distribution stage. A significant increase in powers was also observed when the sample size was increased to 5 replicates. The RE and AE show a negligible impact from the contamination and the trends in normal distribution simulations are the same as in the contaminated normal distribution cases. Therefore the analyses discussed here will not focus as much on the impact of the contamination of the distributions as they will on the behavior of the two tests across settings and across cases. In the event that some contaminated distribution offers a different trend than expected, an emphasis would be made and appropriate conclusions would be drawn.

5.1. Relative Type I Error Comparison

Overall, the RE of Dunnett's test remains below 0.30% throughout the entire study. This was expected since Dunnett's test is known to have a strong control of its FWER. Records as low as 0.04% were recorded when the total number of simulated treatments reached 300. Dunnett's test REs also remain relatively constant across settings within cases. Though FDR method does not have such strong control of the FWER, the REs observed for FDR method were not alarming. Its highest RE were recorded in the first settings and a downward trend in RE was observed as the number of true null hypotheses increases from one setting to the next. This is because FDR method emphasis lies in the fact that it controls a proportion of false rejections. Thus as R increases, so will V, resulting in a higher FWER. Overall, the lowest RE recorded for FDR method was still higher than the highest RE recorded for Dunnett's test at any given point in time during this entire study. In other words, FDR method did not have a RE of 0.30% at any setting within any case for the entire study while none of Dunnett's test RE is higher than 0.30%. The variations in RE for FDR method are not consistent enough to notice a trend in either a decrease or an increases in RE across cases; i.e., we do not see a consistent pattern of an

increase or a decrease in RE as the number of simulated treatments increases for FDR method; although there seems to be a slight increase in RE when the number of replicates was increased to 5. Dunnett's test on the other hand shows a consistent decrease in RE as the number of simulated treatments increases. All RE for both tests are in the single digits and most of them are around 2% for FDR method. Thus both FDR method and Dunnett's test control RE at about 5% in this simulation study.

5.2. Actual Type I Error Comparison

Contrary to the trend observed in Dunnett's test REs, its AE varies across settings within cases. For the first case (50N3) for instance, they start at 1.56% for the first setting and increase progressively to 5.06% for last setting when the number of true null hypotheses reach its highest.

A statistical test that keeps a fixed constant on its control of FWER regardless of the setting may be missing important information to incorporate into computing its statistics. AEs for Dunnett's test are also all higher than RE from the first case to the last. Dunnett's test highest AE was recorded in case 300CN3 at 11.41% when the number of true null hypotheses was at its highest in setting 299:1 and its lowest is around 1.5% and occurs mostly in the first settings. A different trend is observed in the AE of FDR method which increases progressively till the mid settings (25:25, 50:50, 75:75, and 150:150) before beginning a downward trend till the last setting when the number of true hull hypotheses reach its highest.

At one point, FDR method's AE reached a high of 52.37% at setting 100:200 in case 300N5 (power was at 97.73% at that point). Most of FDR method's AEs are in the double digit while Dunnett's are in the single digit. FDR method's AEs also increase as the total number of treatments increases and an increment in the number of replicates also led to an increase in the number of actual type I error FDR method makes. This makes sense since higher sample size gives more power to a test as it is able to make more rejections and more type I error. The opposite is true and we can see that the contamination also has a decreasing trend on FDR method's AE.

5.3. Relative Power Comparison

With the same trend as its RE, Dunnett's test RP remains constant across settings within cases. This RP doubles when the number of replicates is increased to 5 for all cases. For 50 treatments,

Dunnett's RP jumps from 23% (N3) to 50% (N5) and stays constant for all settings. In the case of 100 simulated treatments, that number goes from 19% to 44% while for 150 simulated treatments, Dunnett's RP increases from 17% to 40%. 300N jumped from about 14% to almost 36% for 3N and 5N respectively. A similar trend is also observed in the contaminated stage of the study. However, the RP from a normal to a contaminated normal case registered a slight decrease in RP.

The increase in the number of replicates came as expected with a rise in power for both tests and also allowed Dunnett's test to keep a modest RP level across cases reaching a low of 8% when the total number of simulated treatments reaches 300 in a contaminated normal case with 3 replicates. The highest RP recorded in all cases for Dunnett's test was with case 50N5 with 50% of the true differences detected. Its power decreases progressively from 50 to 300 treatments across distribution. FDR Method on the other hand offers a completely difference picture from that of Dunnett's test. At the lowest number of true null hypotheses (1:49, 1:99, 1:149 and 1:299), FDR's RP remain relatively constant (56%) for N3 and 84% for N5.

This same trend in FDR method relative power is also observed for the second settings across cases on one hand and across distribution on the other. This goes on till the setting before the last. This is an attractive package for researchers especially with these low REs. The REs are much higher when the total number of true null hypotheses is at its lowest for the first settings of all cases, since as R increases, V also increases. The relatively higher RPs of FDR when the total number of true null hypotheses increases will also be preferable in situations where the researcher does not have a clear knowledge of a fixed control treatment.

5.4. Actual Power Comparison

It is interesting to see that both tests are comparable in their APs. Though a downward trend in slope is observed in the actual powers curves of the two tests, the difference between the numbers of at-least-one difference detected by either test is smaller than the trends that have been observed so far. Both tests start at similar APs and the observed increase from 3 replicates to 5 replicates is negligible relative to the observed differences in the RPs between the two tests. In other words, APs did not increase as fast as RP from the first stage (3 replicates) to the second stage of the study. One

explanation we could offer for this is the fact that both tests start their APs at above 85% in most cases which is a high percentage compared to most of the detections we have observed so far. For 50N3, AP is 85.05% for Dunnett's test and 83.93% for FDR method.

For 50N5, these powers become 98.04% and 97.82%. Looking across all simulated treatments, it is fair to say that the APs for both tests stay at least around 85% for the first setting of all cases and the gap of AP between the two tests is almost constant across settings in all cases. So the variation in cases do not have any significant impact on the performances of the two tests. Even the contamination of normal does not seem to have a severe impact on the performance of the two tests, at least not the sudden decrease that we observed in RPs and REs above.

Overall, the APs decrease very slowly as the number of true null hypotheses increases up to the last setting where their performances plummet to as low as 8.08% for Dunnett's test 5.79% for FDR method in setting 299:1 for 300CN3. Both tests perform very poorly in AP when the number of true null hypotheses is at its highest in any of the cases.

5.5. A Ratio-based Comparison

Here we present two cases where the ratio of true null hypotheses to the total number of simulated treatments can guide a researcher as to which of the tests to use when in presence of such situations. This comparison serves as a guide to a researcher who has a clear idea about the number of true null hypotheses that his data might have. Two cases are presented here: the first one is when the researcher knows that no true null hypotheses exists in the data; and the second is when the number of true null hypotheses is about half of the total number of null hypotheses. We use the simulated data with 3 replicates.

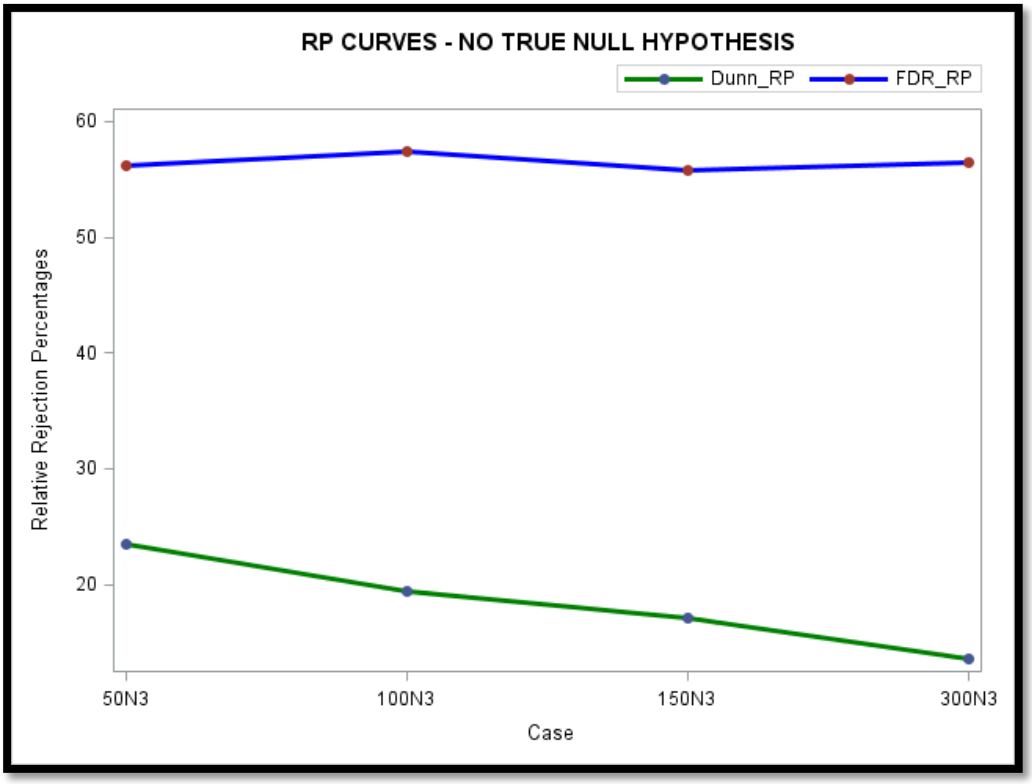

Figure 5.1. Relative Power for First Settings (1:49, 1:99, 1:149 and 1:299) in all Simulated Treatments

FDR method is able to detect at least one-half of the false null hypotheses and keeps its RP at a relatively high level above 55% irrespective of the total number of simulated treatments.

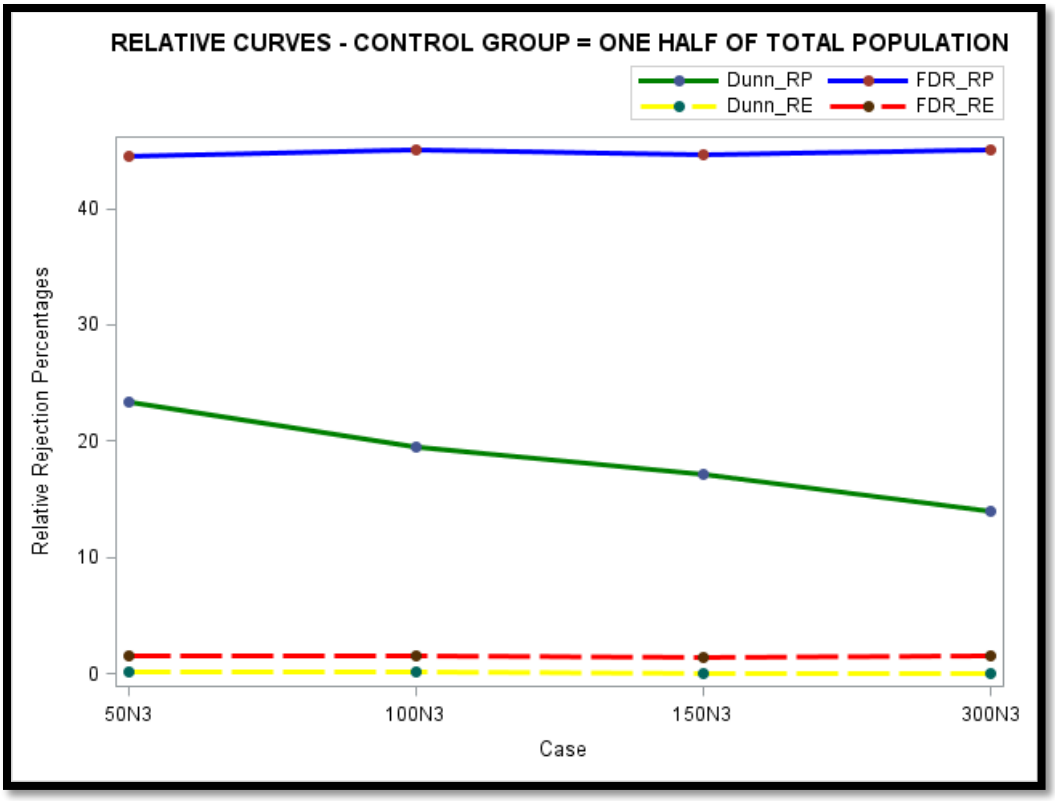

Figure 5.2. Relative Power of Mid-settings (25:25, 50:50, 75:75 and 150:150) in all Simulated Treatments

With these two figures, FDR method's RP holds or even improves with higher number of treatments while Dunnett's RP declined. A powerful attribute of the FDR method can be induced here. Dunnett's test RPs decreases as the number of simulated treatments increases. So as new researches involving many treatments being considered at the same time, a relative power of 57% is preferable to a 23% that decreases as the testing scope gets larger more treatments are added to the study, it makes sense that FDR method be used. When it is not clear how many of the treatments in a study are the same as the control, FDR method still offer a better approach than Dunnett's test with an RP of about 45% keeping a FWER very close to that of Dunnett's when as much as half of the treatments are similar to the control in the characteristics being measured in the study.

Furthermore, Dunnett's test RP decreases as the total number of treatments increases and such trend will reduce its RP to an insignificant level with 500, 800 or 1000 treatments under study. In conclusion, the two most important assets of FDR method here are the stability of its RP and RE as the

total number of treatments increases; and its higher ability to detect more differences than Dunnett's test. This is because Dunnett's test controls FWER and FDR method controls FDR.

5.6. The Contamination Effect on FDR Method and Dunnett's Test

The following graphs show the effect of the contamination when 300 treatments are simulated with 5 replicates. The orange dotted curve for Dunn_RE_N overlaps Dunn_RE_CN curve and is the reason why there is only one visible dotted line on figures 5.3 and 5.4

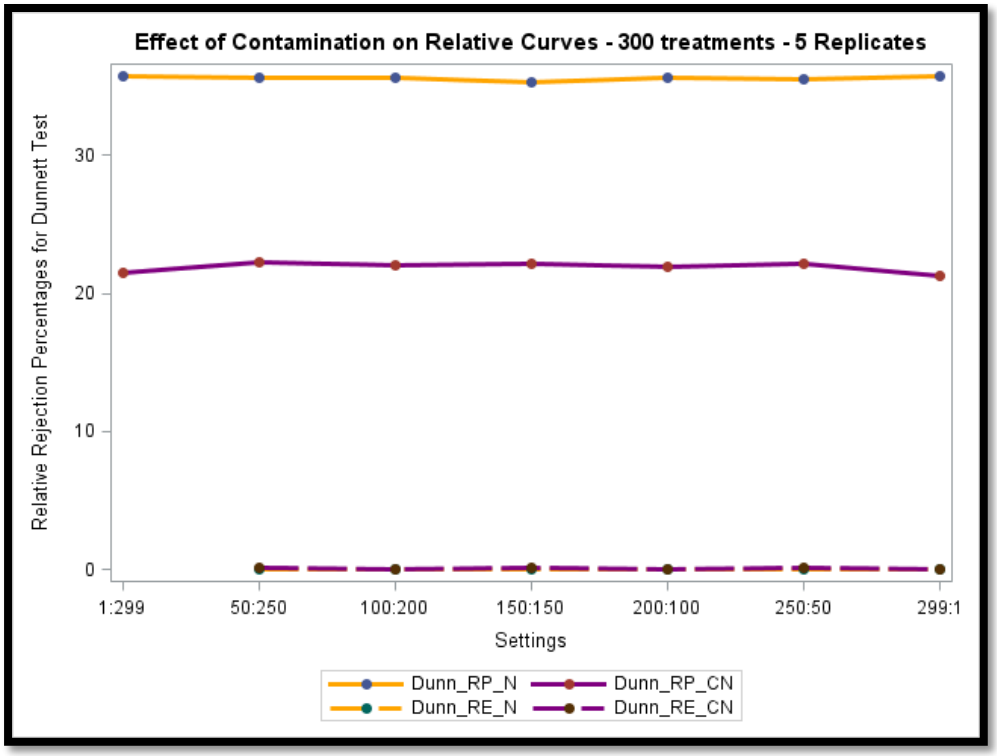

Figure 5.3. Contamination Effect on Relative Measures for Dunnett's Test for 300 Treatments with 5 Replicates

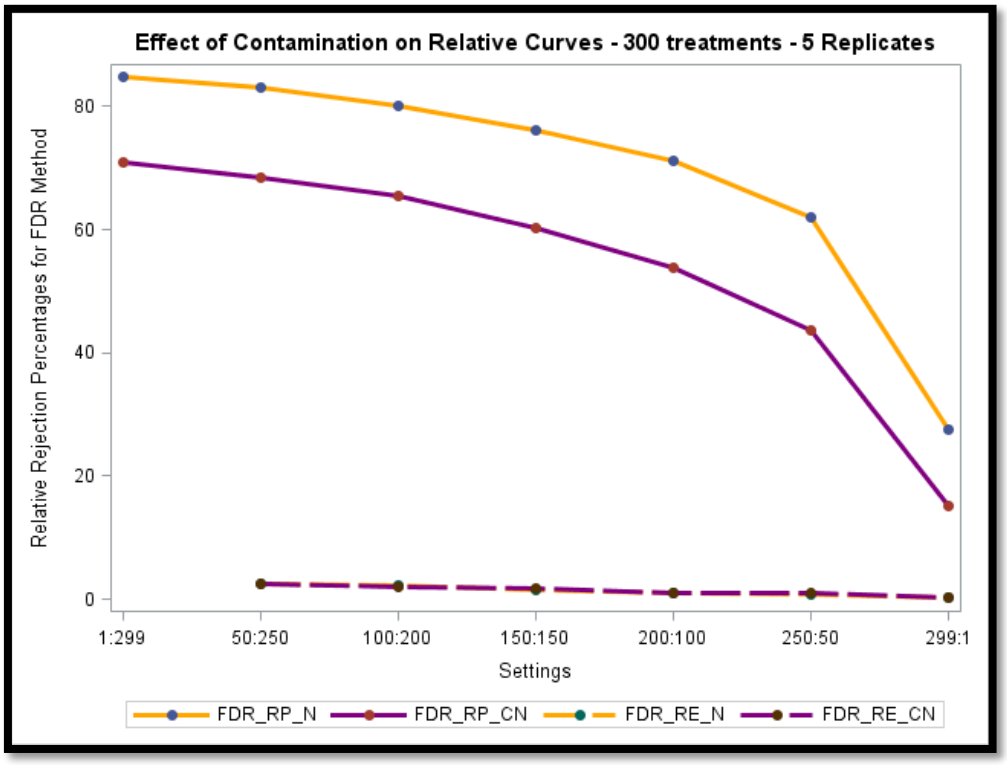

Figure 5.4. Contamination Effect on Relative Measures for FDR Method for 300 Treatments with 5 Replicates

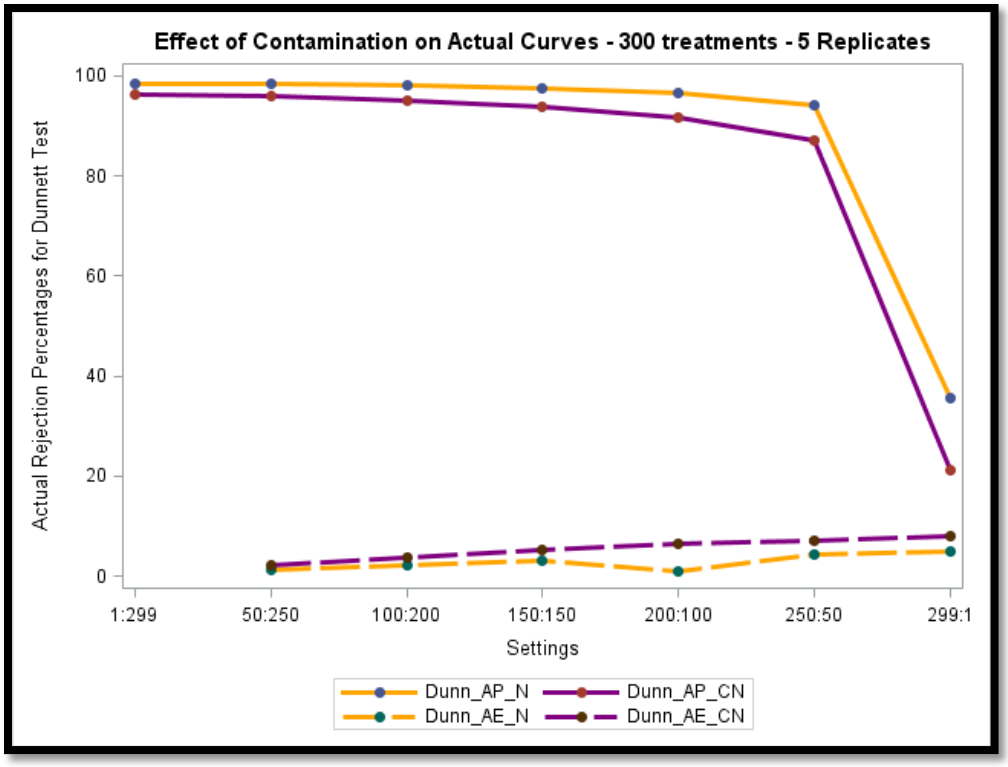

Figure 5.5. Contamination Effect on Actual Measures for Dunnett's Test for 300 Treatments with 5 Replicates

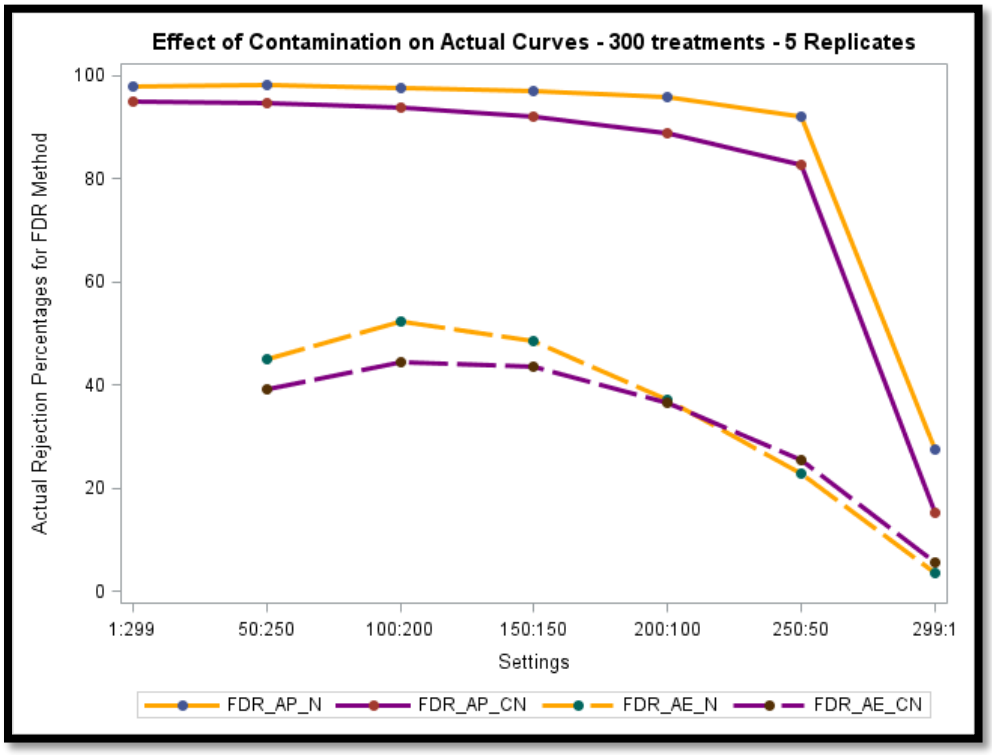

Figure 5.6. Contamination Effect on Actual Measures for FDR Method for 300 Treatments with 5 Replicates

The expected effect of the contamination is the same as the observed. With such a contamination, we expected Dunnett's test and FDR method to make more errors and lose some of their power. Actual errors made by FDR method in the contaminated case is lower than in the normal case until setting 200:100 before the normal case makes less error than the contaminated case. The loss of power is constant across setting.

In summary, the relative power of the FDR method is larger than that of Dunnett's in most settings and across cases and as the relative power of FDR method decreases and as the number of treatments increases, hence Dunnett's test is less powerful than FDR method in cases when we have high number of treatments and the number of replicates per sample size is small. FDR power is not affected by the number of treatments being compared. FDR's actual error is much higher than that of Dunnett's FWER as the number of treatments is increased from 50 to 100, 150 and 300. The strongest part of Dunnett's test is perhaps its strong control of type I error while FDR's strength lies in its ability to detect more of the differences.

CHAPTER 6. AN APPLICATION TO MEAT SCIENCE DATA

Using the data from a recent study of ß-phenylethylamine as a novel nutrient treatment to reduce bacterial contamination due to Escherichia coli O157:H7 on beef meat data, we tested the performance of the False Discovery Rate method and Dunnett's test. The data was collected using the Phenotype Microarray (PM) technology developed for the determination of bacterial phenotypes. In this study, PM was used to test bacterial phenotypes in a 96 well format where the individual nutrients were dried to the base of each well. There were two plates of wells: the carbon sources plates and the nitrogen sources plates. When used with the tetrazolium dye that is provided by the manufacturer, the bacterial phenotype that is measured is respiration, which is indicative of growth. Each well has a different nutrient and we are interested in comparing the response variable RLU, a measure of biofilm amounts in each cell of the isogenic mutants in the nitrogen sources plates. A regulator that controls biofilm amounts in E.coli was used in this study since it has been proved that it also controls cell division in the same bacteria. The measurements were taken for three bacterial phenotypes in the carbon as well as the nitrogen sources, leading to six panels of data. We used one panel of data and considered only the mutant phenotype in this comparison between FDR method and Dunnett's test.

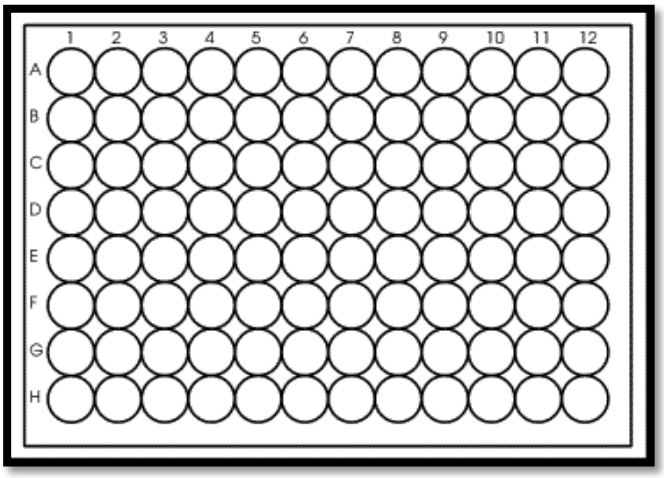

Figure 6.1. Image of a 96-cell Plate Containing the Bacterial Phenotypes

We compared the response variable (RLU) across the wells to identify how many of them are flagged by either test as significantly different from the lowest means response on one side and the highest means response on the other side. For a complete detail of the data, see Lynnes, T. et al (2014).

(*** signifies a rejection and – a non-rejection at a significance level of 0.05). Column comparisons contain the coordinates of the 96 cells plate. Here are the comparison results in table 6.1.

Table 6.1. Output of the Results of FDR Method and Dunnett's Test on Meat Science Data

Comparison with Lowest Means			Comparison with Highest Means		
Comparisons with G10	**Dunnett's Test**	**FDR Method**	**Comparisons with E8**	**Dunnett's Test**	**FDR Method**
E8 - G10	***	***	**B12 - E8**	***	***
F12 - G10	***	***	**A3 - E8**	***	***
F4 - G10	***	***	**C4 - E8**	***	***
F7 - G10	***	***	**D10 - E8**	***	***
E12 - G10	***	***	**F6 - E8**	***	***
F3 - G10	***	***	**A4 - E8**	***	***
F11 - G10	***	***	**F2 - E8**	***	***
G1 - G10	***	***	**D4 - E8**	***	***
F9 - G10	***	***	**G10 - E8**	***	***
C1 - G10	***	***	**H9 - E8**	***	***
H9 - G10	***	***	**H8 - E8**	***	***
H8 - G10	***	***	**B10 - E8**	***	***
B10 - G10	***	***	**H1 - E8**	***	***
H1 - G10	***	***	**E5 - E8**	***	***
E5 - G10	***	***	**E11 - E8**	***	***
E11 - G10	***	***	**E7 - E8**	***	***
E7 - G10	***	***	**H3 - E8**	***	***
H3 - G10	***	***	**E6 - E8**	***	***
E6 - G10	***	***	**H2 - E8**	***	***
H2 - G10	***	***	**E9 - E8**	***	***

Table 6.1. Output of the Results of FDR Method and Dunnett's Test on Meat Science Data (continued)

Comparison with Lowest Means

Comparisons with G10	Dunnett's Test	FDR Method
H4 - G10	***	***
G4 - G10	***	***
E1 - G10	***	***
E3 - G10	***	***
E4 - G10	***	***
D5 - G10	***	***
A10 - G10	***	***
F5 - G10	***	***
G2 - G10	***	***
H12 - G10	***	***
G3 - G10	***	***
H10 - G10	***	***
A9 - G10	***	***
H7 - G10	***	***
G5 - G10	***	***
E2 - G10	***	***
F1 - G10	***	***
B5 - G10	***	***
A12 - G10	***	***
H5 - G10	***	***
G6 - G10	***	***
A11 - G10	***	***

Comparison with Highest Means

Comparisons with E8	Dunnett's Test	FDR Method
G4 - E8	***	***
E1 - E8	***	***
E3 - E8	***	***
E4 - E8	***	***
D5 - E8	***	***
A10 - E8	***	***
F5 - E8	***	***
G2 - E8	***	***
H12 - E8	***	***
G3 - E8	***	***
H10 - E8	***	***
A9 - E8	***	***
H7 - E8	***	***
G5 - E8	***	***
E2 - E8	***	***
F1 - E8	***	***
B5 - E8	***	***
A12 - E8	***	***
H5 - E8	***	***
G6 - E8	***	***
A11 - E8	***	***
B1 - E8	***	***

Table 6.1. Output of the Results of FDR Method and Dunnett's Test on Meat Science Data (continued)

Comparison with Lowest Means			Comparison with Highest Means		
Comparisons with G10	**Dunnett's Test**	**FDR Method**	**Comparisons with E8**	**Dunnett's Test**	**FDR Method**
D8 - G10	***	***	**G8 - E8**	***	***
G8 - G10	***	***	**H11 - E8**	***	***
H11 - G10	***	***	**B11 - E8**	***	***
B11 - G10	***	***	**B7 - E8**	***	***
B7 - G10	***	***	**D3 - E8**	***	***
D3 - G10	***	***	**H6 - E8**	***	***
H6 - G10	***	***	**D2 - E8**	***	***
D2 - G10	***	***	**G9 - E8**	***	***
G9 - G10	***	***	**G11 - E8**	***	***
G11 - G10	***	***	**G12 - E8**	***	***
G12 - G10	***	***	**D7 - E8**	***	***
D7 - G10	***	***	**A7 - E8**	***	***
A7 - G10	***	***	**D1 - E8**	***	***
D1 - G10	***	***	**E10 - E8**	***	***
E10 - G10	***	***	**A2 - E8**	***	***
A2 - G10	***	***	**B8 - E8**	***	***
B8 - G10	***	***	**C12 - E8**	***	***
C12 - G10	***	***	**C8 - E8**	***	***
C8 - G10	***	***	**B9 - E8**	***	***
B9 - G10	***	***	**C11 - E8**	***	***
C11 - G10	***	***	**C2 - E8**	***	***
C2 - G10	***	***	**F10 - E8**	***	***

Table 6.1. Output of the Results of FDR Method and Dunnett's Test on Meat Science Data (continued)

Comparison with Lowest Means			Comparison with Highest Means		
Comparisons with G10	**Dunnett's Test**	**FDR Method**	**Comparisons with E8**	**Dunnett's Test**	**FDR Method**
B3 - G10	***	***	**D11 - E8**	***	***
C6 - G10	***	***	**A8 - E8**	***	***
A8 - G10	***	***	**B6 - E8**	***	***
B6 - G10	***	***	**G7 - E8**	***	***
G7 - G10	***	***	**A5 - E8**	***	***
A5 - G10	***	***	**C3 - E8**	***	***
C3 - G10	***	***	**D9 - E8**	***	***
D9 - G10	***	***	**B4 - E8**	***	***
B4 - G10	***	***	**C10 - E8**	***	***
C10 - G10	***	***	**C7 - E8**	***	***
C7 - G10	***	***	**A1 - E8**	***	***
A1 - G10	***	***	**C5 - E8**	***	***
C5 - G10	***	***	**B2 - E8**	***	***
B2 - G10	***	***	**F8 - E8**	***	***
F8 - G10	***	***	**D12 - E8**	***	***
D12 - G10	***	***	**C9 - E8**	***	***
C9 - G10	***	***	**A6 - E8**	***	***
A6 - G10	***	***	**C1 – E8**	–	***
B12 - G10	–	***	**F9 - E8**	–	***
A3 - G10	–	***	**F12 - E8**	–	–
C4 - G10	–	***	**F4 - E8**	–	–
D10 - G10	–	***	**F7 - E8**	–	–

Table 6.1. Output of the Results of FDR Method and Dunnett's Test on Meat Science Data (continued)

Comparison with Lowest Means			Comparison with Highest Means		
Comparisons with G10	**Dunnett's Test**	**FDR Method**	**Comparisons with E8**	**Dunnett's Test**	**FDR Method**
A4 - G10	–	***	**F3 - E8**	–	–
F2 - G10	–	***	**F11 - E8**	–	–
D4 - G10	–	–	**G1 - E8**	–	–

Absence of stars represents no significant difference detected between cells. When all other treatments are compared with the lowest means, FDR method and Dunnett's test flagged 88 of the tests as significant. FDR method flagged 7 more as significant and only one of the tests was insignificant for FDR method. Dunnett's test flagged the remaining 8 of the 96 tests as insignificant. For the comparisons with the highest means, Dunnett's test flagged 9 of the tests as insignificant while FDR method only counts 7 of the tests as insignificant.

CHAPTER 7. CONCLUSION

In the end, the FDR method is a compromise between using multiple t-tests and using Dunnett's test. Overall, the choice of an appropriate test will depend on the objectives at stake in each research. When type I error is detrimental to the outcomes of the research, Dunnett's test offers a better alternative. Though the power of Dunnett's test decreases as the number of treatments increases, its Family Wise Error Rate is maintained at a constant level across all treatments. This option might be of value in a marketing situation in which a type I error could cause the loss of a considerable amount of money and the power of the test is not as important as severely limiting the probability of a type I error. However, when the objective of the researcher is to detect as many of the differences as possible, and a type I error although not desired is not as detrimental, then the FDR method is preferred. The probability of a type I error increases as the number of treatments get higher using the FDR method, but not as much as using multiple t-tests. The strength in the FDR method is that it can find considerably more actual differences than Dunnett's test.

Consider the plant breeder data used in our simulation. When there are a total of 300 treatments with 49 treatments that are the same as the control and 250 treatments different from the control. Dunnett's test will generally not find any of the treatments that are the same as the control as being significantly different. However, Dunnett's test will only find approximately 56 of the 250 treatments that are different from the control as being significantly different. The FDR method will find approximately 1 of the 49 treatments the same as the control as being significantly different. However, the FDR method will find approximately 171 of the 250 treatments different from the control as being significantly different. Hence, by using the FDR method over Dunnett's test there will be 115 additional differences found where there are differences (over Dunnett's test) and only 1 extra difference marked as significantly different where no actual difference exists.

When nearly half of the treatments are the same as the control, the FDR method finds approximately 91 of the 150 treatments that are different from the control as being significantly different from the control while finding 3 or fewer treatments that are the same as the control to be significantly different from the control. In this situation, Dunnett's test is only able to identify 33 out of the 150

differences that are significantly different from the control as being significantly different from control; while finding at most 1 treatment that is the same as the control as significantly different from the control. Hence by using the FDR method, over Dunnett's test, there will be 58 additional differences found where there are differences and only 2 or less extra differences marked as significantly different where no actual difference exists.

REFERENCES

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

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

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

Benjamini, Yoav. (2010). Discovering the false discovery rate. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 72*(4), 405-416. doi: 10.1111/j.1467-9868.2010.00746.x

Conagin, A., Barbin, D., and Demetrio, C. G. B. (2011). Modified Dunnett's Test for a Randomized Complete Block Design. *Bras and Biom, Sao Paula, 29*(4), 599-610.

Dudoit, S., Shaffer, J. P., and Boldrick, J. C. (2003). Multiple Hypothesis Testing in Microarray ExperimentsStatistical Science. *Statistical Science, 18*(1), 71-103.

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

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

Efron, Bradley. (2004). Large-Scale Simultaneous Hypothesis Testing: The Choice of a Null Hypothesis. *Journal of the American Statistical Association, 99*(465), 96-104.

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

Efron, Bradley. (2008). Simultaneous Inference: When Should Hypothesis testing problems be combined? *The Annals of Applied Statistics, 2*(1), 197-223.

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

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

Lazar, Nicole. (2012). The Big Picture: Multiplicity Control in Large Data Sets Presents New Challenges and Opportunites. *CHANCE, 25*(2), 37-40. doi: 10.1080/09332480.2012.685368.

Lynnes T, Horne SM, Prüβ BM. (2014). ß-phenylethylamine as a novel nutrient treatment to reduce bacterial Contamination due to Escherichia coli O157:H7 on beef meat. *Meat Science, 96*(1), 165-172.

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

Ozkaya, Guven and Ercan, Ilker. (2012). Examining Multiple Comparison Procedures According to Error Rate, Power Type and False Discovery Rate. *Journal of Modern Applied Statistical Methods, 11*(2), 348-360.

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

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

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

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

APPENDIX A. TABLES AND FIGURES FROM THE CONTAMINATED NORMAL SIMULATION WITH 3 REPLICATES

Table A1. Summary of Rejections Percentages in Case 50CN3

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

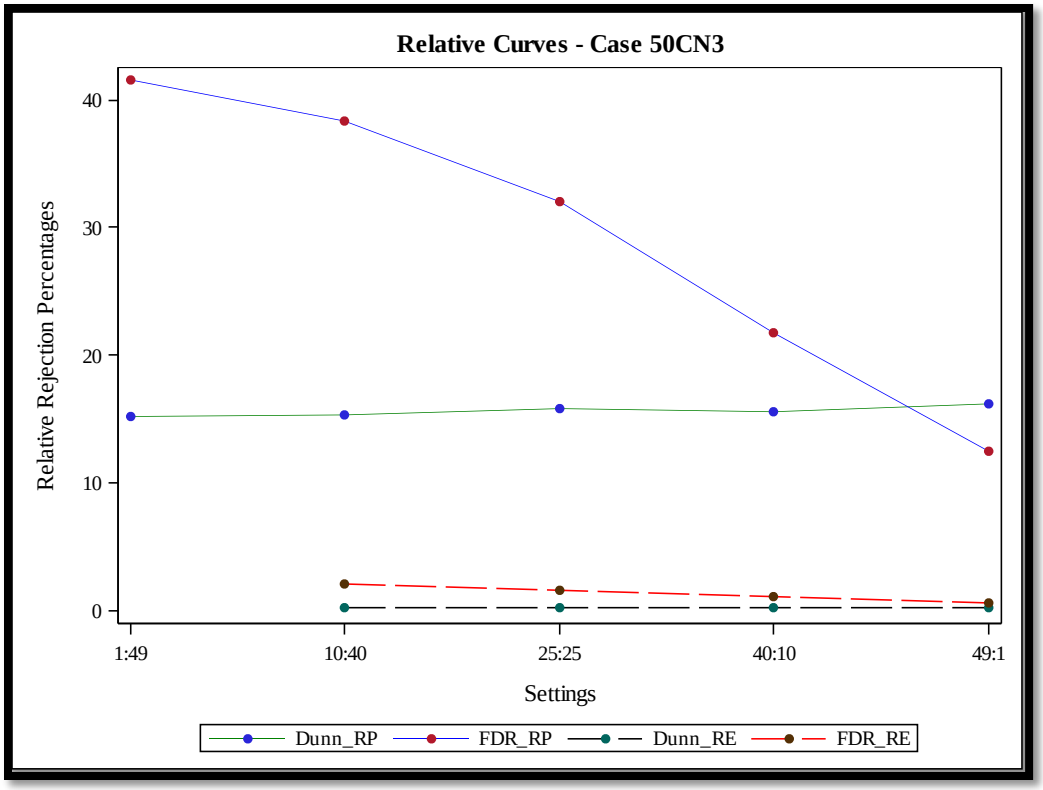

Figure A1. RP and RE Curves for FDR Method and Dunnett's Test in Case 50CN3

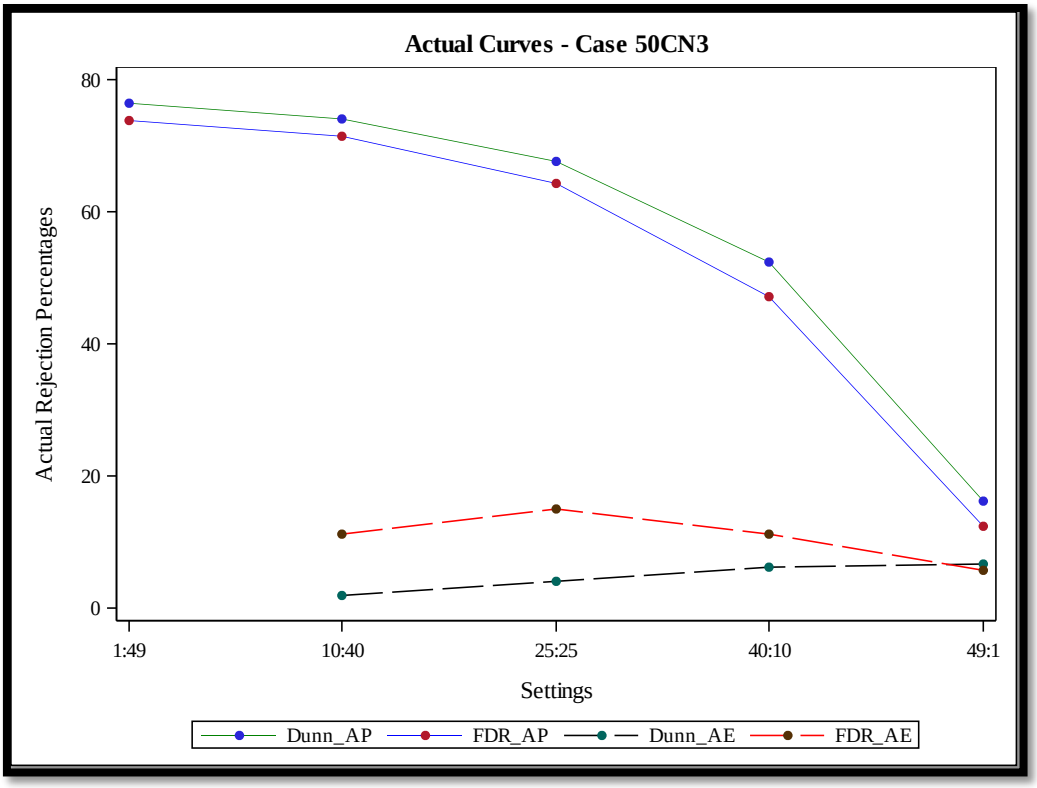

Figure A2. AP and AE Curves for FDR Method and Dunnett's Test in Case 50CN3

Table A2. Summary of Rejections Percentages in Case 100CN3

		Type I error		Power	
SETTINGS	**TESTS**	Relative (RE)	Actual (AE/FWER)	Relative (RP)	Actual (AP)
1:99	Raw	___	___	57.52	99.43
	Dunnett's test			12.29	81.49
	FDR Method			42.33	77.97
25:75	Raw	5.05	47.67	57.50	99.36
	Dunnett's test	0.18	2.91	11.83	77.74
	FDR Method	1.95	18.17	37.03	73.98
50:50	Raw	4.93	65.84	57.98	99.10
	Dunnett's test	0.18	4.87	12.22	73.35
	FDR Method	1.56	19.74	31.48	68.64
75:25	Raw	5.03	75.67	57.57	97.90
	Dunnett's test	0.19	6.64	11.84	61.34
	FDR Method	1.06	14.91	22.30	55.41
99:1	Raw	5.04	83.16	57.91	57.91
	Dunnett's test	0.18	8.66	11.95	11.95
	FDR Method	0.57	6.30	9.00	9.00

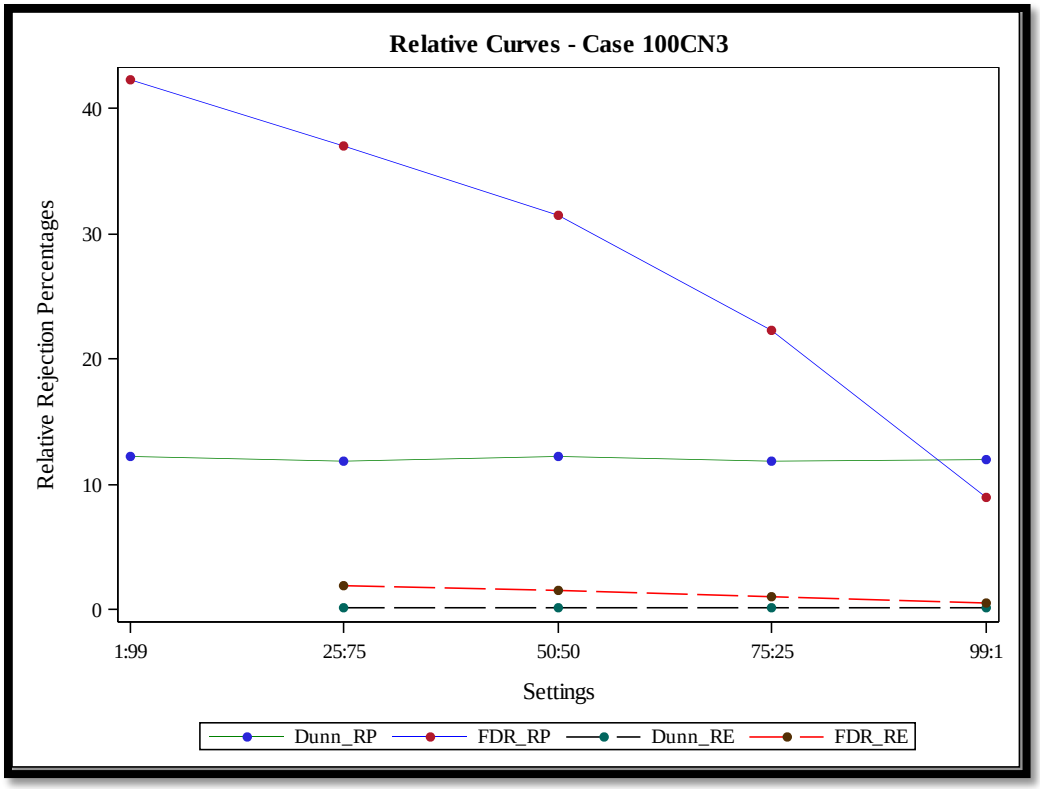

Figure A3. RP and RE Curves for FDR Method and Dunnett's Test in Case 100CN3

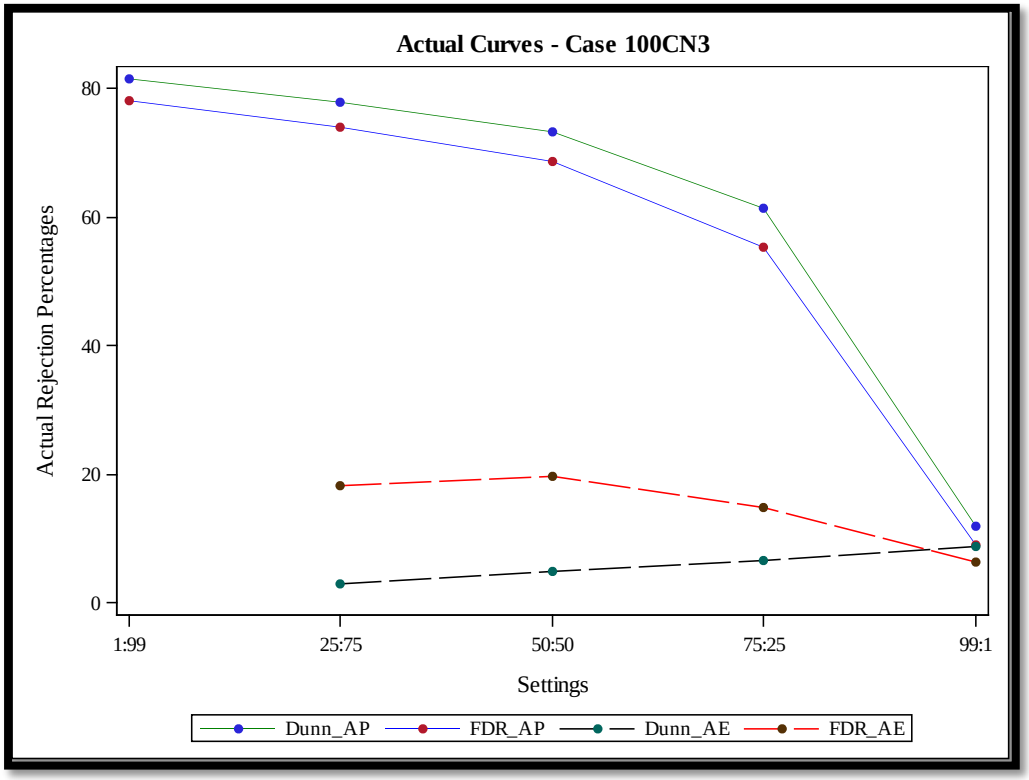

Figure A4. AP and AE Curves for FDR Method and Dunnett's Test in Case 100CN3

Table A3. Summary of Rejections Percentages in Case 150CN3

		Type I error		**Power**	
SETTINGS	**TESTS**	Relative (RE)	Actual (AE/FWER)	Relative (RP)	Actual (AP)
1:149	Raw	___	___	57.63	99.74
	Dunnett's test			10.65	83.14
	FDR Method			42.09	79.37
25 : 125	Raw	5.07	48.35	56.83	99.83
	Dunnett's test	0.16	2.26	10.27	80.14
	FDR Method	2.12	19.33	37.97	76.03
50 : 100	Raw	5.49	67.24	57.53	99.58
	Dunnett's test	0.17	5.17	10.78	77.50
	FDR Method	2.07	23.93	35.34	73.80
75 : 75	Raw	5.17	76.23	56.84	99.39
	Dunnett's test	0.14	5.87	10.05	73.51
	FDR Method	1.44	23.11	29.85	68.48
100 : 50	Raw	5.03	83.08	57.68	98.90
	Dunnett's test	0.13	6.99	10.19	68.70
	FDR Method	1.08	20.38	25.24	63.30
125 : 25	Raw	5.23	87.36	57.46	97.46
	Dunnett's test	0.15	8.55	10.48	57.28
	FDR Method	0.89	16.06	18.69	50.29
149:1	Raw	5.21	90.75	58.58	58.58
	Dunnett's test	0.15	9.40	10.46	10.46
	FDR Method	0.89	6.75	7.78	7.78

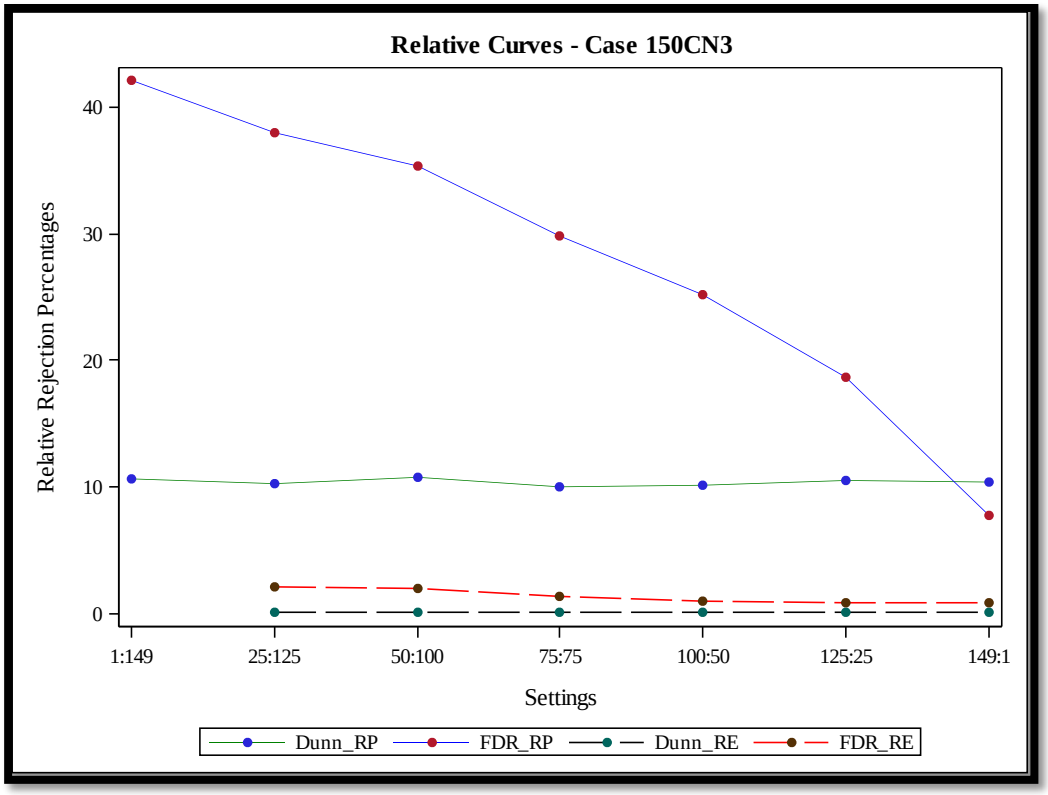

Figure A5. RP and RE Curves for FDR Method and Dunnett's Test in Case 150CN3

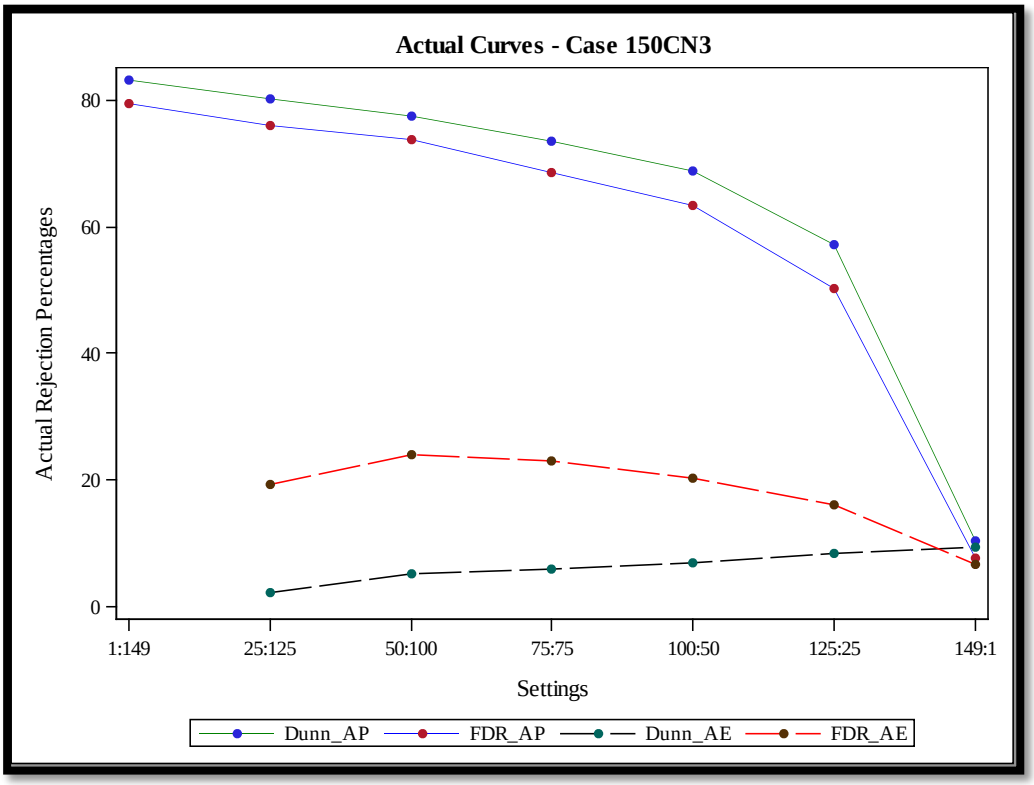

Figure A6. AP and AE Curves for FDR Method and Dunnett's Test in Case 150CN3

Table A4. Summary of Rejections Percentages in Case 300CN3

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

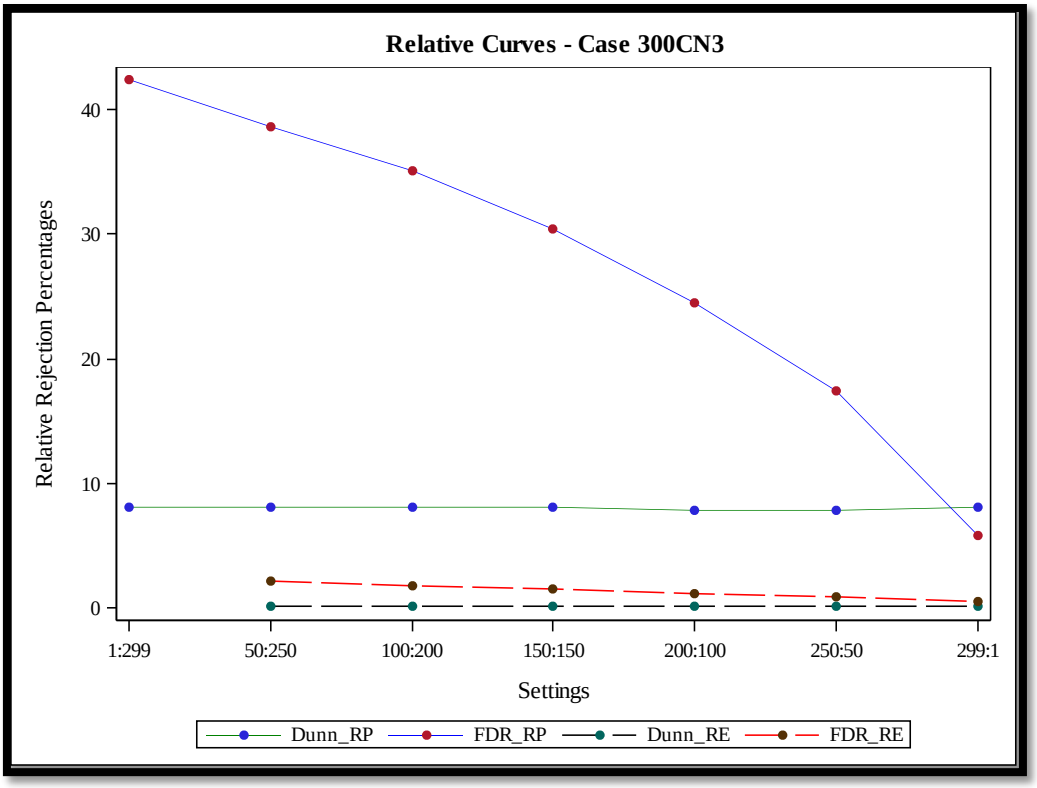

Figure A7. RP and RE Curves for FDR Method and Dunnett's Test in Case 300CN3

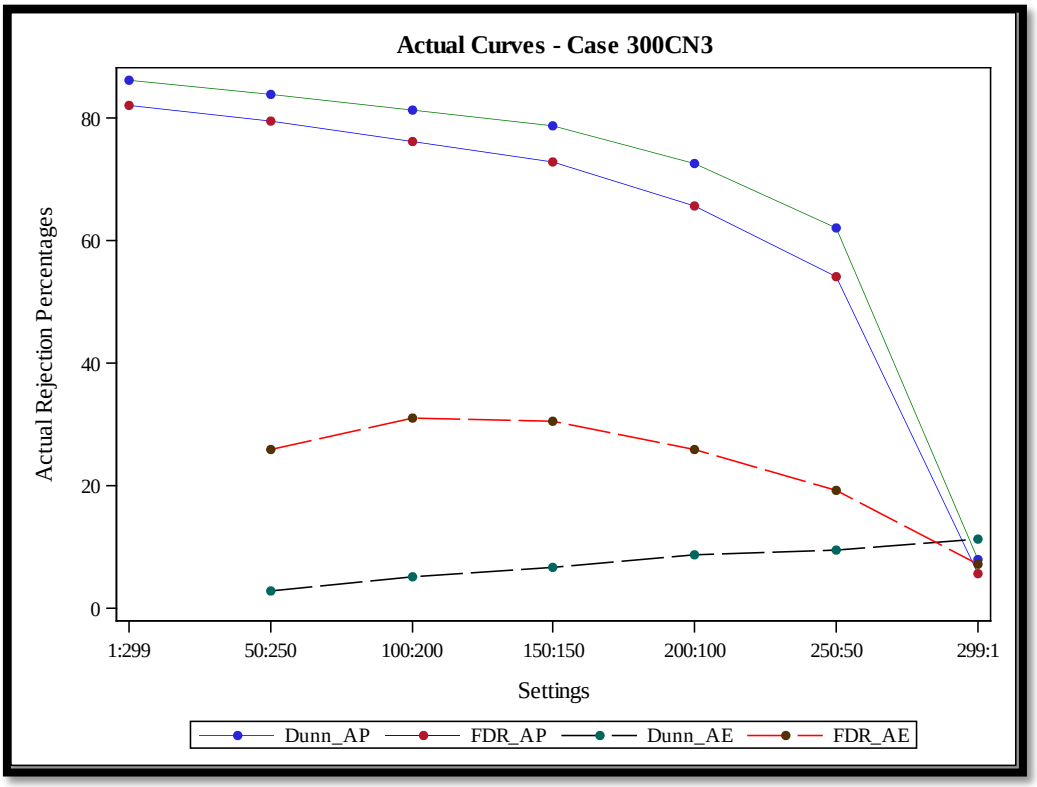

Figure A8. AP and AE Curves for FDR Method and Dunnett's Test in Case 300CN3

APPENDIX B. TABLES AND FIGURES FROM THE CONTAMINATED NORMAL SIMULATION WITH 5 REPLICATES

Table B1. Summary of Rejections Percentages in Case 50CN5

		Type I error		**Power**	
SETTINGS	**TESTS**	Relative (RE)	Actual (AE/FWER)	Relative (RP)	Actual (AP)
1:49	Raw	___	___	79.38	99.85
	Dunnett's test			36.07	94.12
	FDR Method			71.76	93.83
10:40	Raw	4.87	26.41	78.85	99.84
	Dunnett's test	0.23	1.67	34.60	93.23
	FDR Method	2.37	13.72	67.68	92.49
25:25	Raw	4.98	46.14	79.74	99.72
	Dunnett's test	0.23	3.67	35.89	90.09
	FDR Method	1.81	18.69	61.90	89.32
40:10	Raw	5.22	60.51	79.00	98.79
	Dunnett's test	0.24	5.50	34.90	79.81
	FDR Method	1.04	13.79	47.59	76.51
49:1	Raw	5.10	64.27	**78.81**	**78.81**
	Dunnett's test	0.23	6.42	**35.30**	**35.30**
	FDR Method	0.57	5.98	**29.63**	**29.63**

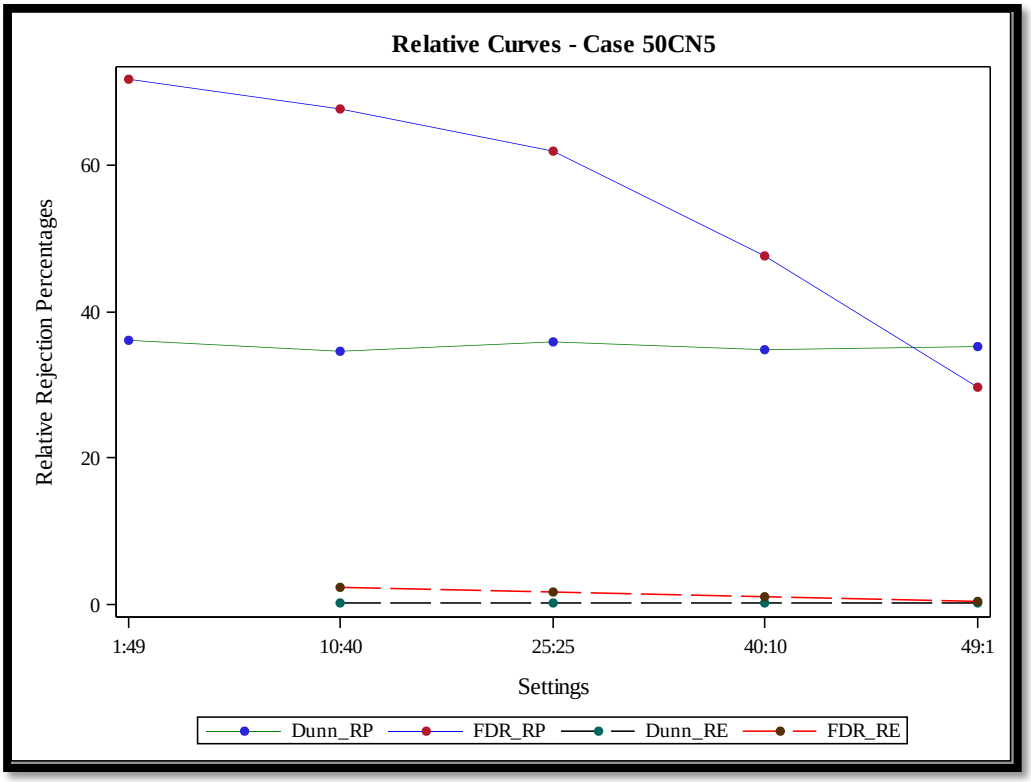

Figure B1. RP and RE Curves for FDR Method and Dunnett's Test in Case 50CN5

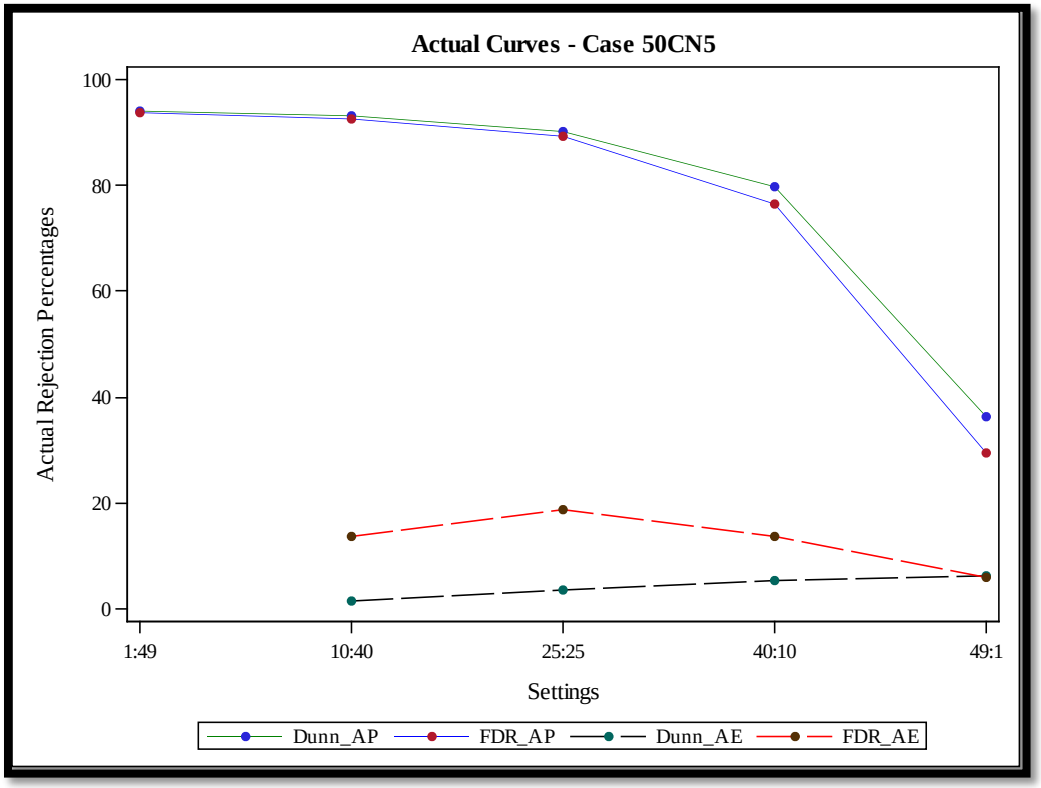

Figure B2. AP and AE Curves for FDR Method and Dunnett's Test in Case 50CN5

Table B2. Summary of Rejections Percentages in Case 100CN5

		Type I error		**Power**	
SETTINGS	**TESTS**	Relative (RE)	Actual (AE/FWER)	Relative (RP)	Actual (AP)
1:99	Raw	___	___	79.35	99.98
	Dunnett's test			29.63	95.12
	FDR Method			71.64	94.55
25:75	Raw	5.09	48.05	78.92	99.98
	Dunnett's test	0.14	2.46	29.49	93.50
	FDR Method	2.21	23.87	66.65	93.00
50:50	Raw	5.24	65.42	79.13	99.92
	Dunnett's test	0.14	4.19	29.55	91.35
	FDR Method	1.70	26.07	60.69	89.75
75:25	Raw	5.04	74.52	79.14	99.52
	Dunnett's test	0.13	5.49	29.23	85.77
	FDR Method	1.07	19.54	49.83	82.92
99:1	Raw	5.03	80.44	79.38	79.38
	Dunnett's test	0.14	7.30	29.49	29.49
	FDR Method	0.45	5.90	23.86	23.86

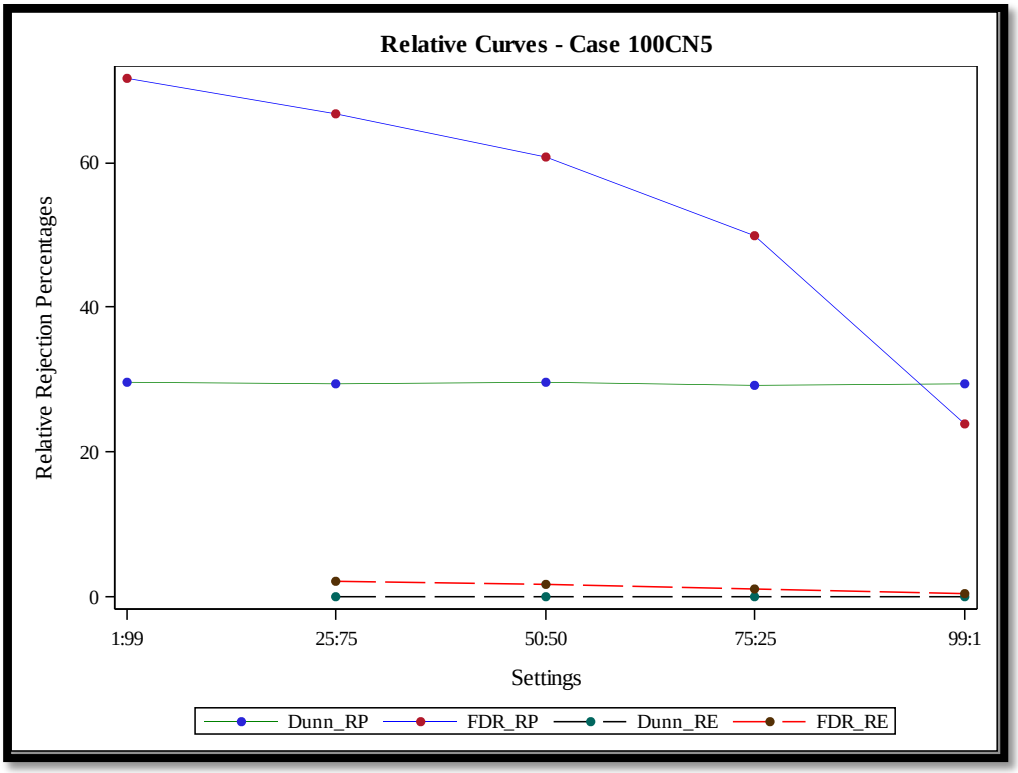

Figure B3. RP and RE Curves for FDR Method and Dunnett's Test in Case 100CN5

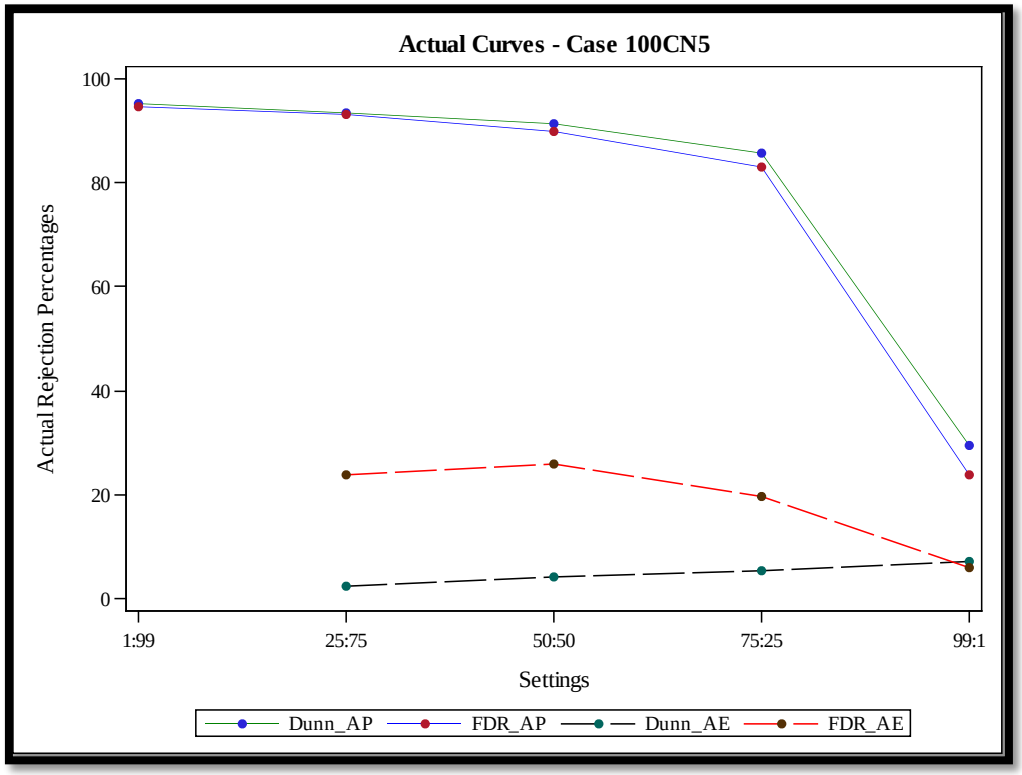

Figure B4. AP and AE Curves for FDR Method and Dunnett's Test in Case 100CN5

Table B3. Summary of Rejections Percentages in Case 150CN5

		Type I error		Power	
SETTINGS	**TESTS**	Relative (RE)	Actual (AE/FWER)	Relative (RP)	Actual (AP)
1:149	Raw	___	___	79.40	99.92
	Dunnett's test			26.80	95.56
	FDR Method			71.49	94.37
25 : 125	Raw	4.93	45.50	79.36	99.98
	Dunnett's test	0.10	1.80	26.18	94.79
	FDR Method	2.44	25.29	68.61	93.95
50 : 100	Raw	5.17	64.70	79.30	99.95
	Dunnett's test	0.12	3.75	26.65	93.96
	FDR Method	2.16	32.69	65.18	92.83
75 : 75	Raw	5.19	74.91	79.37	99.96
	Dunnett's test	0.12	4.58	26.65	92.28
	FDR Method	1.69	32.53	60.59	90.69
100 : 50	Raw	4.80	80.27	79.59	99.90
	Dunnett's test	0.09	5.43	26.18	90.63
	FDR Method	1.09	25.38	54.12	88.54
125 : 25	Raw	5.10	85.83	79.57	99.66
	Dunnett's test	0.11	6.57	26.74	84.11
	FDR Method	0.89	19.42	44.48	79.81
149:1	Raw	5.21	89.03	79.99	79.99
	Dunnett's test	0.12	7.82	26.83	26.83
	FDR Method	0.51	6.14	21.02	21.02

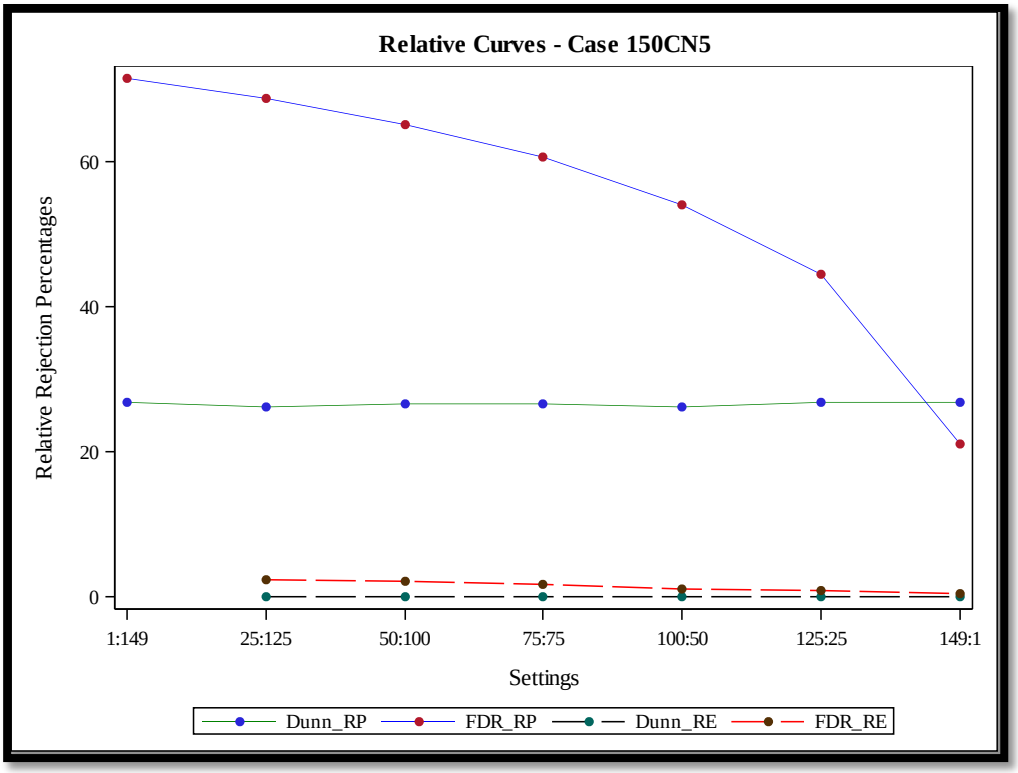

Figure B5. RP and RE Curves for FDR Method and Dunnett's Test in Case 150CN5

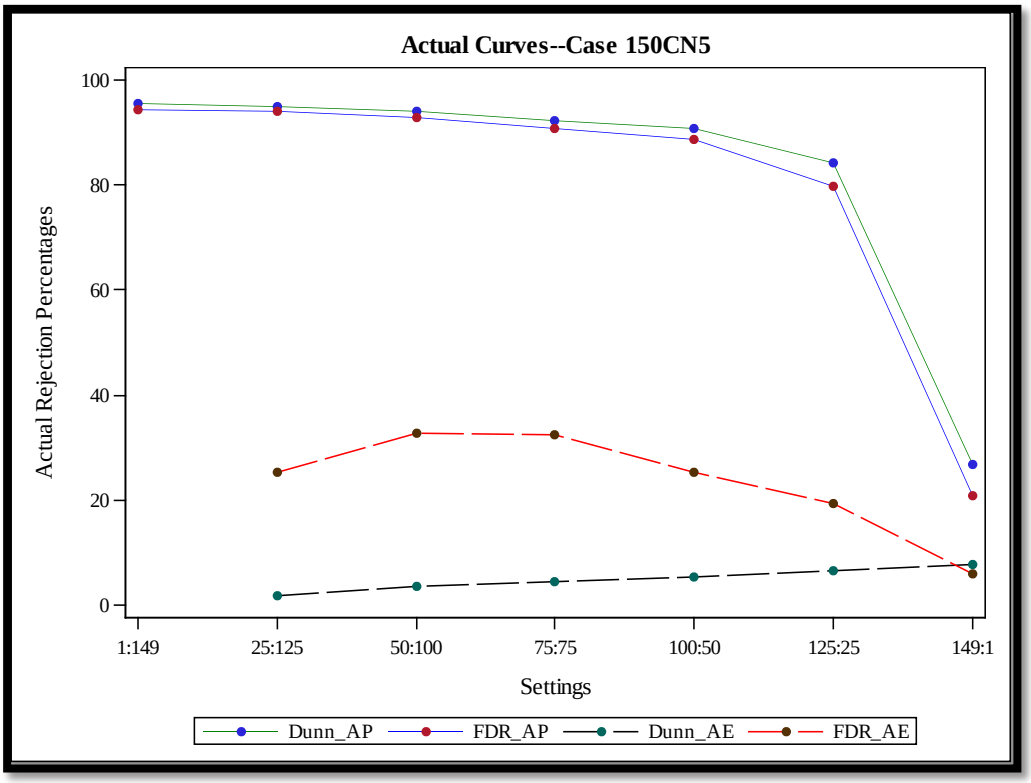

Figure B6. AP and AE Curves for FDR Method and Dunnett's Test in Case 150CN5

Table B4. Summary of Rejections Percentages in Case 300CN5

		Type I error		Power	
SETTINGS	**TESTS**	Relative (RE)	Actual (AE/FWER)	Relative (RP)	Actual (AP)
1:299	Raw			78.95	100.00
	Dunnett's test	___	___	21.55	96.15
	FDR Method			70.94	94.91
50:250	Raw	5.24	66.02	79.28	99.99
	Dunnett's test	0.08	2.24	22.22	95.96
	FDR Method	2.49	39.12	68.46	94.66
100:200	Raw	4.95	81.08	79.72	99.99
	Dunnett's test	0.07	3.71	22.07	95.17
	FDR Method	1.97	44.46	65.53	93.79
150:150	Raw	5.09	89.78	79.40	99.98
	Dunnett's test	0.08	5.21	22.11	93.86
	FDR Method	1.67	43.45	60.37	92.01
200:100	Raw	5.01	93.44	79.41	99.97
	Dunnett's test	0.07	6.56	21.91	91.85
	FDR Method	1.15	36.49	53.81	88.96
250:50	Raw	5.03	95.86	79.64	99.95
	Dunnett's test	0.08	6.98	22.16	86.98
	FDR Method	0.92	25.37	43.63	82.80
299:1	Raw	4.93	97.87	**79.16**	**79.16**
	Dunnett's test	0.07	8.09	**21.23**	**21.23**
	FDR Method	0.39	5.70	**15.26**	**15.26**

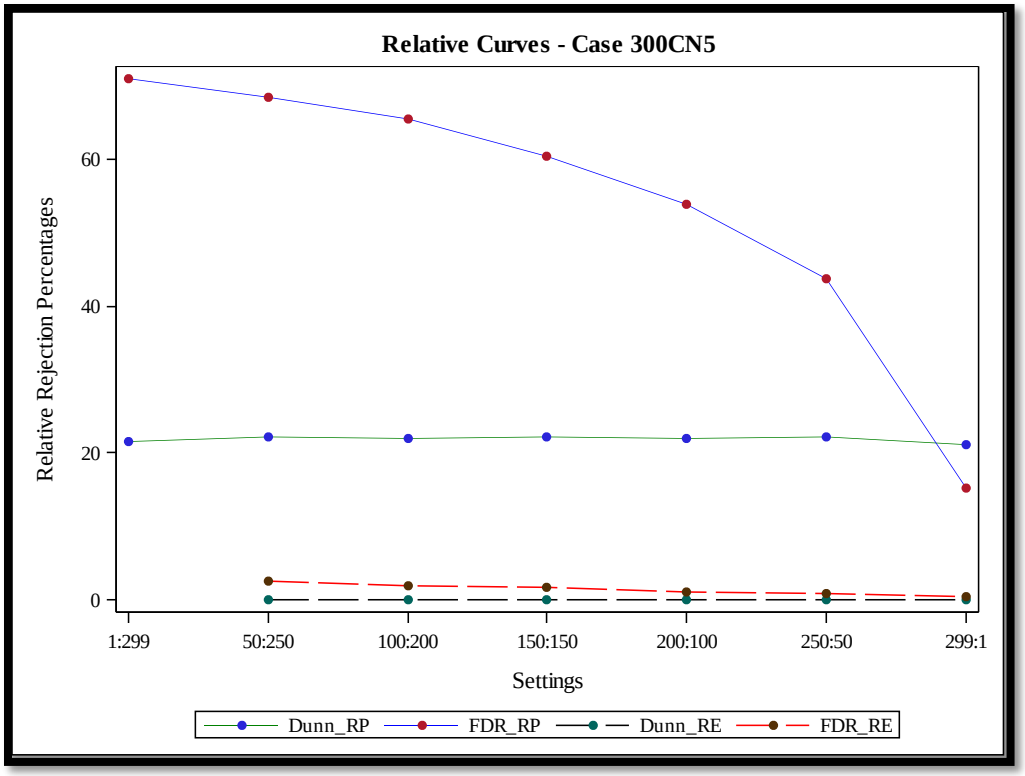

Figure B7. RP and RE Curves for FDR Method and Dunnett's Test in Case 300CN5

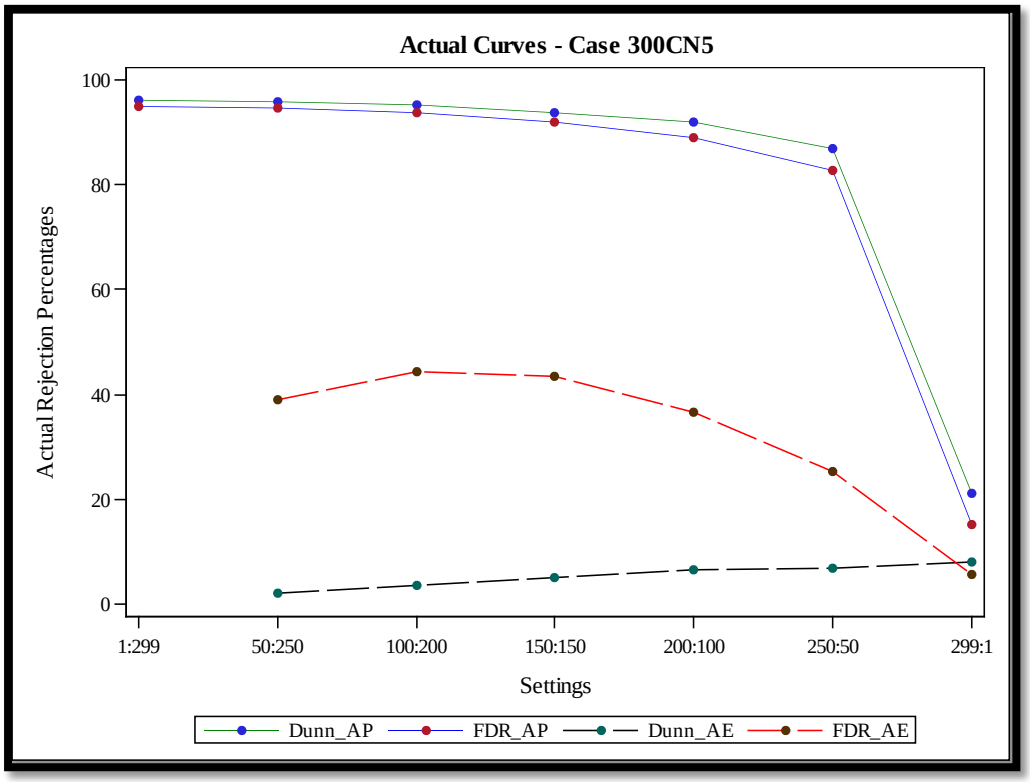

Figure B8. AP and AE Curves for FDR Method and Dunnett's Test in Case 300CN5

APPENDIX C. FIGURES SHOWING THE DISTRIBUTION OF THE REJECTIONS MADE BY BOTH TESTS

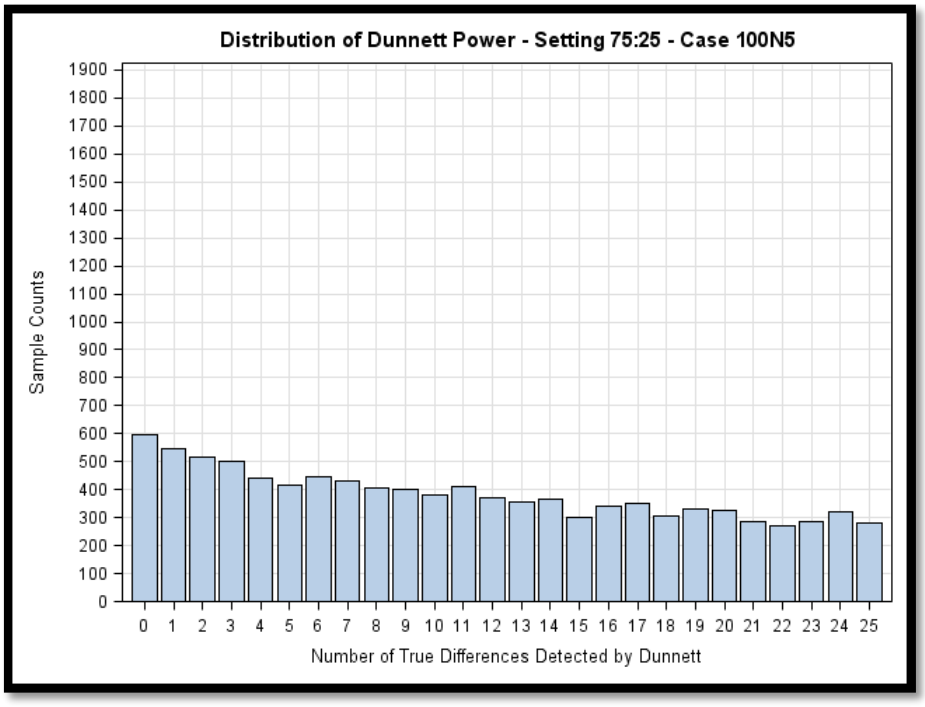

Figure C1. Distribution of the Rejections by Dunnett's Test in Setting 75:25 Case 100N5

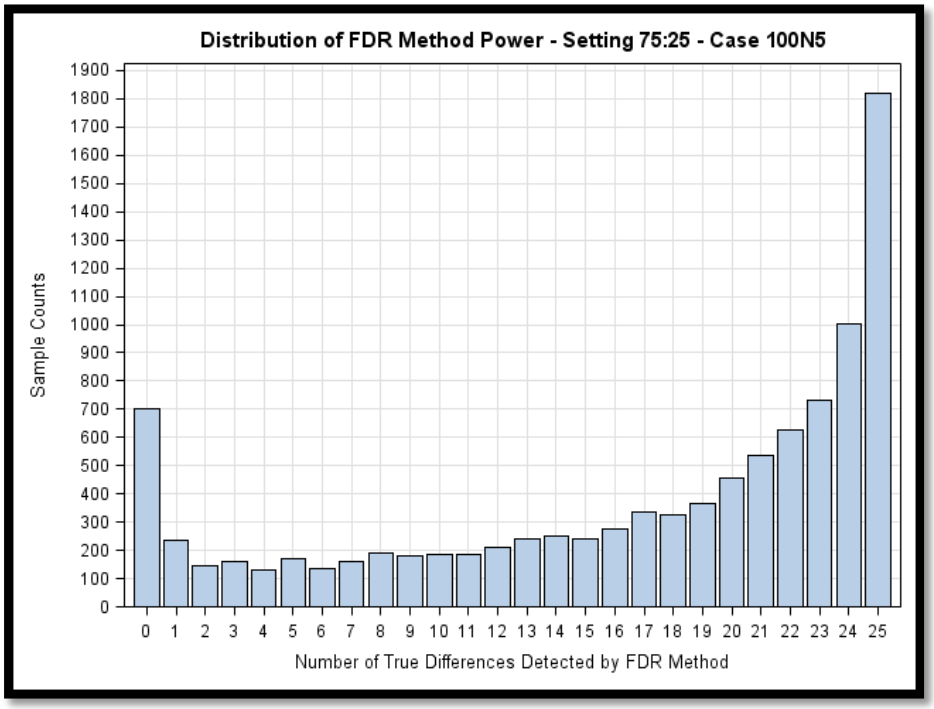

Figure C2. Distribution of the Rejections by FDR Method in Setting 75:25 Case 100N5

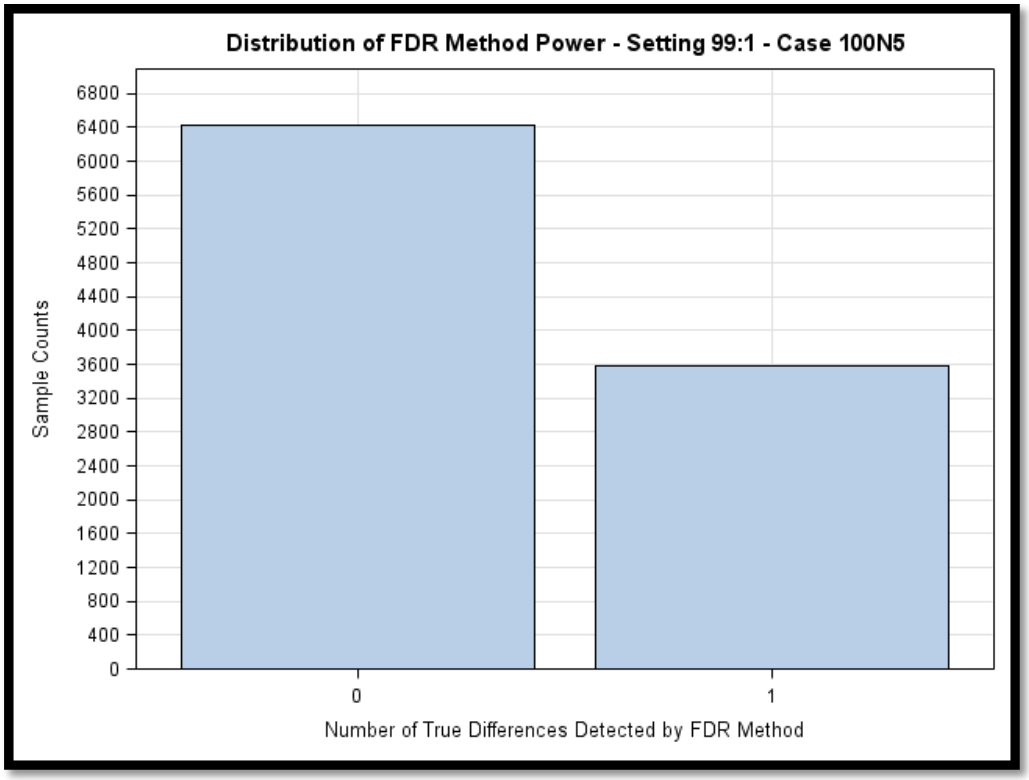

Figure C3. Distribution of the Rejections by FDR Method in Setting 99:1 Case 100N5

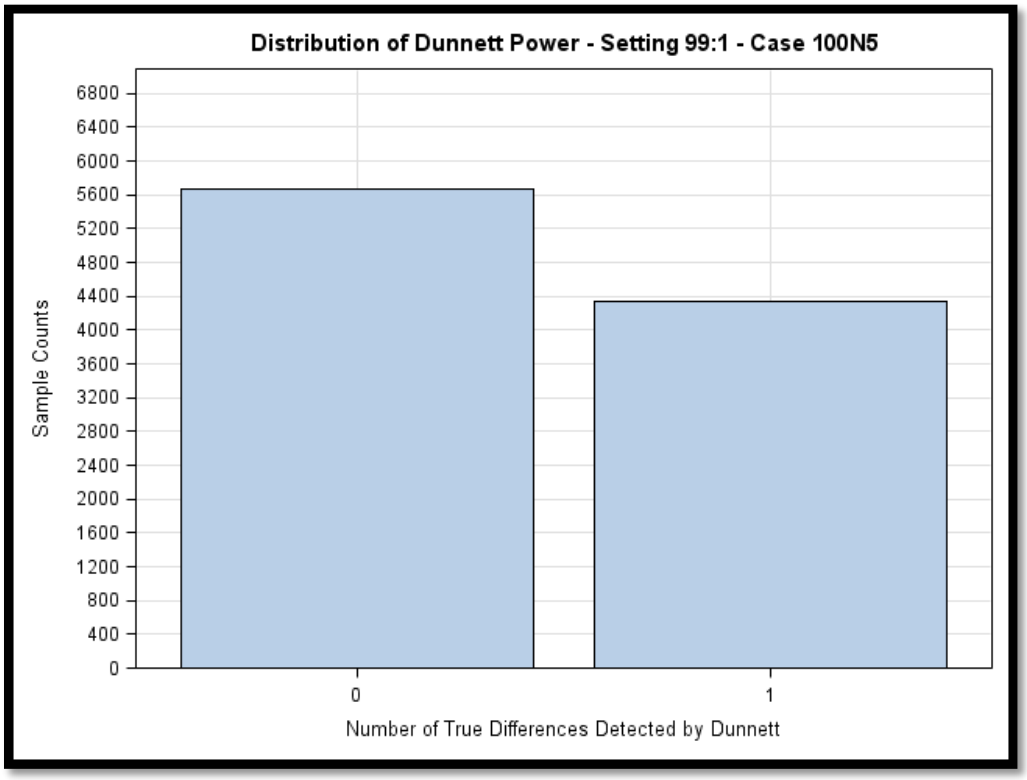

Figure C4. Distribution of the Rejections by Dunnett's Test in Setting 99:1 Case 100N5

APPENDIX D. SAS CODES FOR THE STUDY

The data generation portion of the SAS codes are presented in two parts due to the length of time it takes for each the data to be generated and the ANOVA test. For the large number of treatments (150 and 300), the code in appendix C1 can take on average 48 to 72 hours to generate outputs using only one processor or core. The codes in Appendix D2 and D3 have been designed to reduce the average time to run each simulation by distributing the simulation across available cores in modern day computers, reducing the average time to 12 to 15 hours (most PCs in 2015 have 8 cores to distribute across).

D.1. Non-Distributed Processing Code – Normal Distribution

```
options linesize=80 pagesize=60 formchar="|----|+|---+=|-/\<>*"
        mprint;

title1 'CRD: Dunnett vs. FDR -- Determining Relative Power';
libname KG 'C:\Users\kayeromi.gomez\Documents\newcoderuns\';

*** MVs parms: Samples,Reps,Mu,Sigma,ES,HoLines,N_Lines,seed0;

************************************************************
*** The generate macro accepts 8 parameters as follows:     ***
***                                                         ***
***    Samples = # of samples to generate                   ***
***    Reps  = sample size for each treatment (Equal Sizes) ***
***    Mu    = population mean                              ***
***    Sigma = population standard deviation                ***
***    ES = Effect Size for diff between Reference Line     ***
***         Population and Non-Reference Line Population    ***
***         Expressed as multiple of Sigma                  ***
***    HoLines = # like the Reference Line +1 (Line #1)     ***
***       Note: the Reference Line by default is Line #1    ***
***             All comparisons for Dunnett and t will be   ***
***             to Line #1 ... FDR uses the t results so    ***
***             also compares everything to Line #1         ***
***    N_Lines = # of Lines total                           ***
***              (Line #1 + # of Lines like Reference Line  ***
***               + # Lines different from Line #1)         ***
***    Seed0 = starting seed for simulation                 ***
***                                                         ***
*** Here the populations are from the Normal distribution.  ***
*** The intent is to mimic large plant breeding study with  ***
*** potentially hundreds of genetic lines - thus the use of ***
*** Lines in the names.                                     ***
```

```
************************************************************;

%macro generate(Samples,Reps,Mu,Sigma,ES,HoLines,N_Lines,seed0);

  data mc (keep=Sample Reps Size Line Y);

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

    Size=&Reps;
    do Sample=1 to &Samples;
        do Line=1 to &HoLines;
          do Reps=1 to &Reps;
             *** Distribution of Reference Line Population. **;
             y=rand('Normal',&Mu,&Sigma);
             output;
          end;
        end;
        do Line=(&HoLines+1) to &N_Lines;
          do Reps=1 to &Reps;
             *** Distribution of Non-Ref Line Population.  ***;
             *** The Effect Size is a multiple of sigma.   ***;
             y=rand('Normal',(&Mu+&ES*&Sigma),&Sigma);
             output;
          end;
        end;
    end;
  run;

  proc sort data=mc; *** Facilitate By-Sample Processing. ****;
    by Sample Line;
    run;

  ods output diff=DT_Raw(rename=(RowName=Line)
                         keep=Sample RowName _1 _2 _3);
  ods listing close;
  proc glm outstat=AoV_F_Results;
    by Sample;
    class Line;
    model Y = Line;
    lsmeans Line / adjust=dunnett;
    lsmeans Line / pdiff;
    title2 "Completely Randomized Design for Simulated Yields";
    run;
  ods listing;

  ************************************************************
  *** AoV_F_Results has the p-values from the overall ANOVA ***
  *** F-tests for each simulated sample. They are used only ***
  *** to make sure the simulation seems to be specified OK  ***
  *** in terms of Type I error or Power scenarios.          ***
  ************************************************************;
  data AoV_F_Results;
```

```
  set AoV_F_Results;
  if _type_='SS3'; *** Count rejections using Type III SS. **;
  F_flag=0;                    *** Initialize F_flag to 0. ***;
  if PROB <= 0.05 then F_flag=1; *** Flag small p-values. ***;
  rename PROB=F_p;
  drop _NAME_ _SOURCE_ _TYPE_ DF SS;
  run;

************************************************************
*** DT_Raw contains p-values from Dunnett and t-tests     ***
*** for each simulated sample.  If type=d pulls Dunnett   ***
*** results and type=t pulls the simple t-tests comparing ***
*** each Line to the Reference Line (#1).                 ***
************************************************************;
data DT_Raw;
  set DT_Raw;
    format _1 8.4;
    Line_n=input(Line,8.); * Convert char Line to Number. ***;
    if Line_n=1 then delete; * Line 1 is ref ... no stats. **;
    if _2=. or _3=. then type='d'; *** No Dunn for _2, _3. **;
        else type='t'; *** These are simple t-tests.      ***;
              *** Get rid of char version of Line, _2, _3 ***;
    drop Line _2 _3; *** from pairwise t ... not needed.  ***;
    rename _1=Raw_P; *** Raw p-values used in MultTest.   ***;
  run;
 *proc print data=DT_Raw(obs=50);
 * title2 "Raw p-Values for Dunnett and t";
 * run;

************************************************************
*** Dunnetts results - D_flag is a binary set to 1 if     ***
*** any p-value comparing a Line to Line #1 is <=.05 ...  ***
*** does not check to see if is incorrect rejection. Used ***
*** as cross-check against the ANOVA F-test results.      ***
************************************************************;
*******  Rename Line_n back to Line, Raw_p to Dunn_p.  ******;
data Dunnett(rename=(Line_n=Line Raw_P=Dunn_p));
  set DT_Raw;
  by Sample;
  if type='d' ;   *** Subset raw p-vals - only Dunnetts.  ***;
  retain D_flag;   *** Manage setting flags for Dunnetts. ***;
  ***** Lines <= HoLines are grouped in Ho true group.  *****;
  if Line_n<=&HoLines then Group='Ho True';
     else if Line_n>&HoLines then Group='HoFalse';
  if first.Sample then D_flag=0; * Initialize D_flag to 0. **;
  if Raw_P <= 0.05 then D_flag=1; *** Flag small p-values. **;
  run;
*proc print data=Dunnett(obs=50);
*  title2 "Raw p-Values for Dunnett";
*  run;

data AoV_Dunn_Results;
```

```
  set Dunnett;
  if Line=&N_Lines;                  *** Only last.Sample. ***;
  drop Line Type Group Dunn_p;
  run;

************************************************************
*** FDR results - FDR_flag is binary that is set to 1 if  ***
*** any p-value comparing a Line to Line #1 is <=.05 ...  ***
*** does not check to see if is incorrect rejection. Used ***
*** as cross-check against the ANOVA F-test results.      ***
************************************************************;
ods output Pvalues=FDR_p;
ods listing close;
proc multtest inpvalues=DT_Raw fdr;
  where type='t'; *** Pull only simple t-tests - not Dunn. **;
  by sample;
  title2 "FDR Adjustments Using t Results";
  run;
ods listing;
*proc print data=FDR_p(obs=50);
*  title2 "Raw p-Values for FDR t";
*  run;

data FDR_p;
  set FDR_p;
  retain FDR_flag;
  by sample;
  Line=Test+1;     *** Change Test variable back to Line. ***;
  if Line<=&HoLines then Group='Ho True';
     else if Line>&HoLines then Group='HoFalse';
  if first.sample then FDR_flag=0;
  if FalseDiscoveryRate <= 0.05 then FDR_flag=1;
  run;

data AoV_FDR_Results;
  set FDR_p;
  if Line=&N_Lines;                  *** Only last.Sample. ***;
  drop Test Line Raw FalseDiscoveryRate Group;
  run;

************************************************************
*******   Combine AoV Data Sets, All_Pvals Data Sets  *******
************************************************************;
data AoV_Results;
  merge AoV_F_Results AoV_Dunn_Results AoV_FDR_Results;
  by Sample;
  run;
*** Create permanent SAS Data Set with Parameters in name. **;
data KG.AoV_Results_&HoLines._&N_Lines._&ES._&Reps;
  set AoV_Results;
    Mu=&Mu;
    Sigma=&Sigma;
```

```
    run;

  data All_pvals(drop=type D_flag Test FDR_flag);
    merge Dunnett FDR_p;
    by Sample Line;
    label Dunn_p='Dunn_p';
    run;
  *** Create permanent SAS Data Set with Parameters in name. **;
  data KG.All_pvals_&HoLines._&N_Lines._&ES._&Reps;
    set All_pvals;
      Mu=&Mu;
      Sigma=&Sigma;
    run;

proc format;
  value pt 0-.05='Reject'
          .05<-1='DNR';
  value flagf  1='Reject'
               0='DNR';
  run;

*proc print data=AoV_Results(obs=50);
  run;
proc freq data=AoV_Results;
  tables F_flag D_flag FDR_flag;
  format F_flag D_flag FDR_flag flagf.;
  title2 "Compare to Overall ANOVA F, Dunnett and FDR Results";
  title3 "Mean=&Mu, SD=&Sigma, &HoLines,&N_Lines, Effect=&ES,
R=&Reps";
  run;

*proc print data=All_pvals(obs=50);
  run;
proc freq data=All_pvals;
  tables Group*(Raw Dunn_p FalseDiscoveryRate) / nopct;
  tables Group*Dunn_p*Raw / nopct;
  tables Group*FalseDiscoveryRate*Raw / nopct;
  tables Group*Dunn_p*FalseDiscoveryRate / nopct;
  format Dunn_p FalseDiscoveryRate Raw pt.;
  title2 "Raw p, Dunnett and FDR Results - Relative Power";
  title3 "Mean=&Mu, SD=&Sigma, &HoLines,&N_Lines, Effect=&ES,
R=&Reps";
  run;

%mend generate;

*************************************************************
*** Run the simulation to generate data sets, results.      ***
*************************************************************;
%generate(10000, 3, 1763.18, 438.9605143, 2, 1, 50, 1234567);
```

D.2. Distributed Processing Code - Normal Population

The dispatching codes work with %Distribute macro provided by SAS Institute Inc. It was set to dispatch 250 samples to each of the 8 cores of the computer simultaneously.

```
options source2 ls=80 ps=60 cpucount=8 fullstimer formchar="|----|+|---+=|-/\<>*" ;
ods graphics off;
ods html close;
ods listing;
ods noresults;

filename grdbx 'C:\Users\kayeromi.gomez\Documents\';

%include grdbx(GridDistribute);

%macro Macros;
  %syslput rem_seed0=&seed0;
  %syslput rem_reps=&reps;
  %syslput rem_mu=&mu;
  %syslput rem_esmult=&esmult;
  %syslput rem_sigma=&sigma;
  %syslput rem_HoTrt=&HoTrt;
  %syslput rem_NumTrt=&NumTrt;
%mend;


%macro RInit;
  rsubmit remote=Host&iHost wait=yes macvar=Status&iHost;
  /*
  / This macro initializes the host, preparing it to run
  / the fundamental task multiple times. In this case,
  / all you need to do is initialize the data set in which
  / all the results for this host are to be collected.
  /-----------------------------------------------------*/
  %macro FirstRSub;
    options nonotes;
    data Sample; /* Create an empty dataset */
    if (0);
    run;
  %mend;
  /*
  / This macro defines the fundamental task. SAS/IML is
  / used to perform the basic simulation the number of
  / times (&rem_NIter) required. Each chunk of results
  / is created in a work data set called Sample, and all
  / the results for this host are collected in
  / Results.Sample&rem_iHost.
  /-----------------------------------------------------*/
```

```
  %macro TaskRSub;
    data mc (keep=sample iter size trt y);

    call streaminit(&rem_seed0);  *** Initialize with desired seed.
***;

    size=&rem_reps;
    do sample=1 to &rem_Niter;
        do trt=1 to &rem_hotrt;
          do iter=1 to &rem_reps;
             y=rand('Normal',(&rem_mu),&rem_sigma);
             output;
          end;
        end;
        do trt=(&rem_hotrt+1) to &rem_NumTrt;
          do iter=1 to &rem_reps;

y=rand('Normal',(&rem_mu+&rem_esmult*&rem_sigma),&rem_sigma);
             output;
          end;
        end;
    end;
  run;

  proc sort data=mc;
    by sample trt;
    run;

  ods output diff=cert(rename=(RowName=trt) keep=sample RowName _1 _2
_3);
  ods listing close;
  proc glm outstat=goal;
    by sample;
    class trt;
    model y = trt;
    lsmeans trt / adjust=dunnett;
    lsmeans trt / pdiff;
    title2 "Completely Randomized Design (CRD) for Simulated Yields";
    run;
  ods listing;

  data cert;
    set cert;
    format _1 8.4;
    if compress(trt)='1' then delete;
    if _2=. or _3=. then type='d';
        else type='t';
    rename _1=Raw_P;
  run;

  /*
  / data iSample sets up an incremental data set to be collected
```

```
  / by %RCollect so simply copies the cert data set.
  /-------------------------------------------------------------*/
  data iSample;
    set cert;
    run;

  /*
  / data Sample combines the incremental data sets.
  /-------------------------------------------------------------*/
  data Sample;
    set Sample iSample;
    run;

  %mend;
  endrsubmit;
%mend;

/*
/ Arguments for the distribution of parcels to cores.
/ NIter is the increment to be sent to a core at one time.
/ NIterAll is the total number of samples across all cores.
/-------------------------------------------------------------*/
%let NIter = 250;
%let NIterAll = 10000;
%SignOn;

/*
/ Generate(samples, reps, mu, sigma, esmult, HoTrt, NumTrt, seed0)
/ e.g., Generate(10000, 5, 1763.18, 438.9605143, 2, 100, 300, 0)
/-------------------------------------------------------------*/
%let Reps = 5;
%let Mu = 1763.18;
%let Sigma = 438.9605143;
%let ESmult = 2;
%let HoTrt = 100;
%let NumTrt = 300;
%let seed0 = 0;


%Distribute;
%RCollect(Sample,Sample / view=Sample);


/*
/ Code below does multtest adjustments and aggregates results.
/-------------------------------------------------------------*/
title1 'FDR vs Dunnett in CRD ANOVA';
libname KG 'C:\Users\kayeromi.gomez\Documents\';

data KG.Sampleup_&HoTrt._&NumTrt._&ESmult._&Reps;
  set sample;
    ctr+1;
```

```
    sampcnt=int((ctr-1)/((&Numtrt-1)*2))+1;
    ***** Trt <= HoTrt are grouped in Ho true group.  *****;
    if Trt<=&HoTrt then Group='Ho True';
       else if Trt>&HoTrt then Group='HoFalse';
    Mu=&Mu;
    Sigma=&Sigma;
   run;

proc format;
  value pt 0-.05='Reject'
           .05<-1='DNR';
  value flagf  1='Reject'
               0='DNR';
  run;

proc freq data=KG.Sampleup_&HoTrt._&NumTrt._&ESmult._&Reps;
  where type='d';
  tables Trt*Raw_P / nopct nocol;
  format Raw_P  pt.;
  title2 "Raw p Results for &hotrt Controls";
  run;

%SignOff;
```

D.3. Distributed Processing Code - Contaminated Normal Population

```
options source2 ls=80 ps=60 cpucount=8 fullstimer formchar="|----|+|---+=|-/\<>*" ;
ods graphics off;
ods html close;
ods listing;
ods noresults;

filename grdbx 'C:\Users\kayeromi.gomez\Documents\';

%include grdbx(GridDistribute);

%macro Macros;
  %syslput rem_seed0=&seed0;
  %syslput rem_reps=&reps;
  %syslput rem_mu=&mu;
  %syslput rem_esmult=&esmult;
  %syslput rem_sigma=&sigma;
  %syslput rem_HoTrt=&HoTrt;
  %syslput rem_NumTrt=&NumTrt;
%mend;


%macro RInit;
```

```
  rsubmit remote=Host&iHost wait=yes macvar=Status&iHost;
  /*
  / This macro initializes the host, preparing it to run
  / the fundamental task multiple times. In this case,
  / all you need to do is initialize the data set in which
  / all the results for this host are to be collected.
  /-----------------------------------------------------*/
  %macro FirstRSub;
    options nonotes;
    data Sample; /* Create an empty dataset */
    if (0);
    run;
  %mend;
  /*
  / This macro defines the fundamental task. SAS/IML is
  / used to perform the basic simulation the number of
  / times (&rem_NIter) required. Each chunk of results
  / is created in a work data set called Sample, and all
  / the results for this host are collected in
  / Results.Sample&rem_iHost.
  /-----------------------------------------------------*/
  %macro TaskRSub;
    data mc (keep=sample iter size trt y);

    call streaminit(&rem_seed0);  *** Initialize with desired seed.
***;

    size=&rem_reps;
    do sample=1 to &rem_Niter;
        do trt=1 to &rem_hotrt;
          do iter=1 to &rem_reps;

***  Distribution of Reference Line Population. ***;
***  Bernoulli to select distribution            ***;
***          for reference line population       ***;
***          The effect size of the mixture is used***;

               D1=RAND('Bernoulli',0.90);
          if D1=1 then X1=RAND('Normal',&rem_mu,&rem_sigma);
               else if D1=0 then
X1=RAND('Normal',&rem_mu,(&rem_esmult*&rem_sigma));

           y=X1 ;
            output;
          end;
        end;
        do trt=(&rem_hotrt+1) to &rem_NumTrt;
          do iter=1 to &rem_reps;
*** Distribution of Non-Ref Line Population.  ***;
*** Bernoulli to select distribution         ***;
*** The Effect Size is a multiple of sigma.   ***;
***  The effect size of the mixture is used   ***;
```

```
                D2=RAND('Bernoulli',0.90);
            if D2=1 then
X2=RAND('Normal',(&rem mu+&rem esmult*&rem_sigma),&rem_sigma);
                else if D2=0 then
X2=RAND('Normal',(&rem_mu+&rem_esmult*&rem_sigma),(&rem_esmult*&rem_si
gma));

              y=X2;
              output;
           end;
         end;
    end;
  run;

  proc sort data=mc;
    by sample trt;
    run;

  ods output diff=cert(rename=(RowName=trt) keep=sample RowName _1 _2
_3);
  ods listing close;
  proc glm outstat=goal;
    by sample;
    class trt;
    model y = trt;
    lsmeans trt / adjust=dunnett;
    lsmeans trt / pdiff;
    title2 "Completely Randomized Design (CRD) for Simulated Yields";
    run;
  ods listing;

  data cert;
    set cert;
    format _1 8.4;
    if compress(trt)='1' then delete;
    if _2=. or _3=. then type='d';
        else type='t';
    rename _1=Raw_P;
  run;

  /*
  / data iSample sets up an incremental data set to be collected
  / by %RCollect so simply copies the cert data set.
  /-------------------------------------------------------------*/
  data iSample;
    set cert;
    run;

  /*
  / data Sample combines the incremental data sets.
  /-------------------------------------------------------------*/
```

```
  data Sample;
    set Sample iSample;
    run;

  %mend;
  endrsubmit;
%mend;

/*
/ Arguments for the distribution of parcels to cores.
/ NIter is the increment to be sent to a core at one time.
/ NIterAll is the total number of samples across all cores.
/-------------------------------------------------------------*/
%let NIter = 250;
%let NIterAll = 10000;
%SignOn;

/*
/ Generate(samples, reps, mu, sigma, esmult, HoTrt, NumTrt, seed0)
/ e.g., Generate(10000, 5, 1763.18, 438.9605143, 2, 200, 300, 0)
/-------------------------------------------------------------*/
%let Reps = 5;
%let Mu = 1763.18;
%let Sigma = 438.9605143;
%let ESmult = 2;
%let HoTrt = 200;
%let NumTrt = 300;
%let seed0 = 0;


%Distribute;
%RCollect(Sample,Sample / view=Sample);


/*
/ Code below does multtest adjustments and aggregates results.
/-------------------------------------------------------------*/
title1 'FDR vs Dunnett in CRD ANOVA';
libname KG 'C:\Users\kayeromi.gomez\Documents\';

data KG.Sampleup_&HoTrt._&NumTrt._&ESmult._&Reps;
  set sample;
    ctr+1;
    sampcnt=int((ctr-1)/((&Numtrt-1)*2))+1;
    ***** Trt <= HoTrt are grouped in Ho true group.  *****;
    if Trt<=&HoTrt then Group='Ho True';
       else if Trt>&HoTrt then Group='HoFalse';
    Mu=&Mu;
    Sigma=&Sigma;
   run;

proc format;
```

```
  value pt 0-.05='Reject'
          .05<-1='DNR';
  value flagf  1='Reject'
               0='DNR';
  run;

proc freq data=KG.Sampleup_&HoTrt._&NumTrt._&ESmult._&Reps;
  where type='d';
  tables Trt*Raw_P / nopct nocol;
  format Raw_P  pt.;
  title2 "Raw p Results for &hotrt Controls";
  run;

%SignOff;
```

D.4. Analysis Code

This part of the code is a continuation of the Distributed Processing codes. It can be run after the data has been generated with normal or contaminated normal distribution. Notice the %let calling the data set already generated with the Distributed Processing codes.

```
options linesize=80 pagesize=60 formchar="|----|+|---+=|-/\<>*"
        mprint;

title1 'CRD: Dunnett vs. FDR -- Determining Relative Power';
libname KG 'C:\Users\kayeromi.gomez\Documents\';

%let dsn=sampleup_100_300_2_5;

proc format;
  value pt 0-.05='Reject'
          .05<-1='DNR';
  value fctrf  1-1000='1+ Rej'
                  0='DNR';
  run;

*** Copy from permanent SAS Data Set with Parameters in name. **;
*** Name parms: HoLines, N_Lines, ES, Reps                    ***;

************************************************************
*** DT_Raw contains p-values from Dunnett and t-tests     ***
*** for each simulated sample.  If type=d pulls Dunnett   ***
*** results and type=t pulls the simple t-tests comparing ***
*** each Line to the Reference Line (#1).                 ***
************************************************************;
data DT_Raw;
```

```
  set KG.&dsn;       *** Copy from permanent SAS data set. ***;
    Line_n=input(Trt,8.); ** Convert char Line to Number. ***;
    if Line_n=1 then delete; * Line 1 is ref ... no stats. **;
    if _2=. or _3=. then type='d'; *** No Dunn for _2, _3. **;
        else type='t'; *** These are simple t-tests.      ***;
              *** Get rid of char version of Trt,  _2, _3 ***;
    drop Trt  _2 _3; *** from pairwise t ... not needed.  ***;
  run;
 proc print data=DT_Raw(obs=50);
  title2 "Raw p-Values for Dunnett and t";
  run;

************************************************************
*** Dunnetts results :                                     ***
*******  Rename Line_n back to Line, Raw_p to Dunn_p.  ******
************************************************************;
data Dunnett(rename=(Line_n=Line Raw_P=Dunn_p));
  set DT_Raw;
  if type='d' ;   *** Subset raw p-vals - only Dunnetts.  ***;
  run;
proc print data=Dunnett(obs=50);
  title2 "Raw p-Values for Dunnett";
  run;

************************************************************
*** FDR results:                                           ***
************************************************************;
ods output Pvalues=FDR_p;
ods listing close;
proc multtest inpvalues=DT_Raw fdr;
  where type='t'; *** Pull only simple t-tests - not Dunn. **;
  by sampcnt;
  title2 "FDR Adjustments Using t Results";
  run;
ods listing;
proc print data=FDR_p(obs=50);
  title2 "Raw p-Values for FDR t";
  run;

data FDR_p;
  set FDR_p;
  Line=Test+1;     *** Change Test variable back to Line. ***;
  run;

data All_pvals(drop=type Test);
  merge Dunnett FDR_p;
  by sampcnt Line;
  label Dunn_p='Dunn_p';
  run;

data Eval_Ps;
  set All_pvals;
```

```
  by sampcnt group;
  if first.group then do;
     Raw_Ctr=0;
     Dun_ctr=0;
     FDR_Ctr=0;
     end;
  if Raw<=.05 then Raw_Ctr+1;
  if Dunn_p<=.05 then Dun_Ctr+1;
  if FalseDiscoveryRate<=.05 then FDR_Ctr+1;
  if last.group;
  run;

proc print data=Eval_Ps;
  where sample<11;
  run;

ods rtf file='100 300 r5_ES2.rtf';
proc freq data=Eval_Ps;
  tables group*(Raw_Ctr Dun_ctr FDR_ctr) / nopct;
  format Raw_Ctr Dun_ctr FDR_ctr fctrf.;
  title2 'Distributions of Rejections per Sample';
  title3 "Settings were: &dsn";
  run;

proc freq data=All_pvals;
  tables Group*(Raw Dunn_p FalseDiscoveryRate) / nopct;
  format Dunn_p FalseDiscoveryRate Raw pt.;
  title2 'Pairwise Rejections - Not Simultaneous Inference';
  title3 "Settings were: &dsn";
  run;

proc freq data=Eval_Ps;
  tables group*(Raw_Ctr Dun_ctr FDR_ctr) / nopct;
  format Raw_Ctr Dun_ctr FDR_ctr fctrf.;
  title2 'Distributions of Rejections per Sample';
  title3 "Settings were: &dsn";
  run;
```