CHAPTER 1

INTRODUCTION

1.1 Statement of the Problems and Importance of the Study

To compare the average effects of three or more "treatments" in order to decide which treatments are better, which ones are worse, and by how much, while controlling the probability of making an incorrect decision simultaneously, we use multiple comparison procedures (MCPs), called mean separation tests (SAS Institute Inc. 1999). However, if we are interested merely in whether (or by how much) our treatment is, on average, different from a control or baseline average, we can make our tests more powerful by restricting the family of comparisons to include only comparisons of each of the groups with a control, rather than all possible pair-wise comparisons. For example if we have seven treatment groups, which includes a control, our family will have $k = 6$ comparisons, but if we compare an average of other treatment effects to a control, our family will have $k = 7 \times 6/2 = 21$ comparisons when we consider all possible pair-wise comparisons. By using the Bonferroni method to control FWE (Family-wise Error Rate is the probability of incorrectly rejecting at least one of the null hypothesis that makes up the family) within $\alpha = 0.05$ (Dudoit et al, 2003) the CER (Comparison-wise Error Rate is the probability the interval estimate does not contain the parameter when the null hypothesis is true) should be 0.05/6 = 0.0083 when we compare an average of a treatment effect with a control, but CER equals 0.05/21 = 0.0024 when we consider all possible pair-wise comparisons.

Recently the area of multiple comparisons has been focused on "closed testing" (Grechanovsky and Hochberg, 1999; Koch and Gansky, 1996; Zhang et al., 1997 referred in Westfall and Wolfinger, 2000). Closed testing methods typically result in a "step-wise" type method, which is more powerful than its "single step" counterparts. The simple previously mentioned Bonferroni method is an example of a "single step" method, wherein each of the multiple comparisons must use significance level α/k, where k denotes the number of distinct comparisons and where α denotes the maximum allowable Type I error rate over the set of k comparisons. A step-wise method often uses critical significance levels with larger than α/k, allowing significance differences more often, meaning that methods are more powerful (Westfall and Wolfinger, 2000).

In single step procedures, equivalent multiplicity adjustment is performed for all hypotheses, regardless of the ordering of the test statistics or unadjusted p-value, so that each hypothesis is evaluated by using critical value which is independent of the result of tests from other hypotheses. On the other hand, improvement in power, while preserving Type I error rate control may be achieved by step-wise procedures. For step-wise procedures, the rejection of the particular hypothesis is based not only on the total number of hypotheses, but also on the outcome of the test for other hypotheses (Dudoit et al, 2003).

Step-wise procedures could be classified into two categories; step-down and step-up procedures. In step-down procedures, the hypotheses that correspond to the most significant test statistics (i.e., smallest unadjusted p-value or largest absolute test statistics) are ordered and considered successively, with each subsequent test decision dependent on the outcomes of earlier ones. In case of one fails to reject a null hypothesis, no further hypotheses are rejected. In contrast, for step-up procedures, the hypotheses that correspond successively, again with each subsequent test decision dependent on the outcomes of earlier ones. When one hypothesis is rejected, all remaining hypotheses are consequently rejected.

Examples of step-up procedures include one that applies Simes inequality to control FWE, step-up Hochberg adjusted p-values (Simes, 1986 and Hochberg, 1988 referred in Dudoit et al, 2003) and Rom's method (Rom, 1990 referred in Westfall, 1999). Step-down procedures include Holm's procedure, step-down Šidák adjusted p-values, Simes modified Bonferroni, Westfall and Young step-down min P adjusted p-values and step-down max T adjusted p-values.

Step-up procedures have been found to be more powerful than step-down procedures; however, it is very important to remember that a procedure based on the Simes inequality $\left(\Pr\left(P_{(i)} > \frac{\alpha i}{k}, \forall i = 1,...,k \mid H_0^C\right) \geq 1-\alpha\right)$ relies on the assumption that the result was shown under independence, but it is a conservative procedure for dependent tests. In 1996, Troendle proposed a permutation-based step-up multiple testing procedure which takes into account the dependence structure among the test statistics and is related to the Westfall and Young (1993) step-down max T procedure (Dudoit et al, 2003).

In many situations when the joint distribution of the test statistics is unknown, then a resampling method (e.g., permutation, bootstrap) could be used to estimate unadjusted and adjusted p-values while avoiding parametric assumptions about the joint distribution of the test statistics.

As demonstrated by Westfall and Young (1993), the bootstrap resampling methodology provides valid tests of composites when the data come from a "location shift" model, which does not require normal distributions. The location shift model requires that the distributions of the data in each group must be identical, which possibly are in different locations (medians), but allows that the distributions may not be normal. In such cases the PROC MULTTEST of SAS® provides valid and closed multiple testing procedures especially for comparing several groups against a common control to develop a screening process for differences.

PROC MULTTEST is one of the major procedures for multiple tests and comparisons in SAS®, which was written specifically for the purpose of doing multiple inferences. It offers p-value adjustments using Bonferroni, Šidák, bootstrap resampling and permutation resampling, all with single-step or step-down procedures.

To decide which test or procedure in PROC MULTTEST to use, three main components of the problem need to be identified: (1) the assumptions of the statistical model, (2) the comparison or testing objectives of the study and (3) the set of comparisons we want to consider; for example, all pair-wise comparisons in the ANOVA, all pair-wise comparisons with a control and multiple comparisons with the best.

PROC MULTTEST also has some procedures where little is assumed about distributions, correlation, etc. These procedures were designed specifically to allow for nonnormality using bootstrap and permutation resampling methods. The general structure of the data is that the observations (vectors) are assumed independent and the covariance matrices are assumed constant. However the distributional form is not specified. The commands are BOOTSTRAP, PERM, STEPBOOT and STEPPERM. BOOTSTRAP is a command used to perform bootstrap resampling. PERM is a command used to perform permutation resampling. STEPBOOT is a command used to perform bootstrap resampling for adjusting p-values by step-down procedures. STEPPERM is a command used to perform permutation resampling for adjusting p-values by step-down procedures (Westfall and Tobias, 1999; SAS Institute Inc. 2002).

In 2001, Smith and Taylor introduced the notion of the dependent bootstrap procedure and some important properties were also established. In addition, they investigated confidence intervals using dependent bootstrap and found that dependent bootstrap confidence intervals achieved similar coverage to the traditional (independent) bootstrap and normal theory confidence intervals while retaining slightly smaller lengths.

Hence, step-up procedures are more powerful than step-down procedures, when Simes inequality $\left(\Pr\left(P_{(i)} > \frac{\alpha i}{k}, \forall i = 1,\ldots,k \mid H_0^C \right) \geq 1-\alpha \right)$ relies on the assumption. Because it is difficult to assume Simes inequality in practice, this research is mainly focused on step-down procedures using bootstrap resampling. The bootstrap resampling may be classified into two categories based on the random method; resampling the sample data with replacement (independent bootstrap) and resampling c copies of the sampled data without replacement (dependent bootstrap). So the researcher would like to extend the idea of dependent bootstrap to the Westfall and Young bootstrap min P for comparing several means of treatment with a control. The efficiency of the new procedure will be compared with the traditional step-down bootstrap min P and Dunnett's t statistic, in the case that the sample data is normally distributed with homogeneous variance.

1.2 Research Objectives

The objectives of this research are the following;

1.2.1 To present the new step-down dependent bootstrap min P procedures for comparing several means with a control.

1.2.2 To compare the efficiency of the new step-down dependent bootstrap min P with the traditional step-down bootstrap min P and Dunnett's *t* statistic.

1.3 Research Hypothesis

The efficiency of the new step-down dependent bootstrap min P is not worse than the traditional step-down bootstrap min P procedure and Dunnett's *t* statistic for comparing several means with a control.

1.4 Research Scope

The scope of the study is summarized by the following;

1.4.1 There are 3 treatment groups and 1 control group.

1.4.2 Each group has the equal sample size.

1.4.3 The sample size (r) is 3, 4, 5, …, 10.

1.4.4 The error term is normally distributed with equal variance.

1.4.5 The significance level of the two sided test is set to be 0.05.

1.4.6 For the dependent bootstrap procedure, let c (the number of copies of the sample data) be equal to 2 and 4.

1.4.7 The number of bootstrap resamples (B) is set to be 100, 1,000 and 10,000.

1.4.8 The number of Monte Carlo simulations is set to be 1,000 repetitions.

1.5 Comparison Criterion

The criteria for comparisons of the efficiency of each procedure are determined by a pre-specified and acceptable level α for the Type I error rate. In addition, it was also determined by minimizing the Type II error rate which is equivalent to maximizing power of the test, in the class of all tests with Type I error rate at most α.

1.6 Research Advantage

The advantage of the study is to establish that the new step-down dependent bootstrap min P procedure is as good as Dunnett's *t* statistic for comparing several treatment means with a control, when the error term of balanced one-way ANOVA model is normally distributed with equal variances.

1.7 Definition of Terms

1.7.1 Closed Testing is a general method to perform more than one hypothesis test simultaneously.

1.7.2 A Step-wise procedure is a procedure for which rejection of a particular hypothesis is based not only on the total number of hypotheses but also on the outcome of the test of other hypotheses.

1.7.3 Type I error is the error of rejecting a null hypothesis when it is actually true.

1.7.4 Type II error is the error of failing to reject a null hypothesis when the alternative hypothesis is the true state of nature.

1.7.5 FWE (Family-wise error rate) is the probability of making one or more false discoveries, or type I errors among all the hypotheses when performing multiple pair-wise tests.

1.7.6 MEER (Maximum Experiment-wise Error Rate) is the maximum probability of at least one false rejection of the null hypothesis over is entire experiment.

1.7.7 A Raw p-value or Unadjusted p-value is a p-value of a test statistic without an adjustment by any procedure to control the error rate.

1.7.8 An Adjusted p-value is a p-value of a test statistic after an adjustment by a procedure to control the error rate.

1.7.9 The Power of a test is the probability of rejecting null hypothesis when it is not true. It can be written as $1-\beta$; where β denotes the Type II error rate.