MULTIPLE COMPARISONS WITH A CONTROL UNDER A GROUP TESTING SCENARIO

By

YANINA GRANT

Bachelor of Science

Technical Institute of Advanced

Studies of Monterrey

Monterrey, Nuevo León, Mexico

2000

Submitted to the Faculty of the
Graduate College of the
Oklahoma State University
in partial fulfillment of
the requirements for
the Degree of
MASTER OF SCIENCE
May, 2006

MULTIPLE COMPARISONS WITH A

CONTROL UNDER A GROUP

TESTING SCENARIO

Thesis Approved:

_____

Thesis Advisor
Dr. Melinda H. McCann

_____

Dr. Stephanie A. Monks

_____

Dr. William D. Warde

_____

Dean of Graduate College
Dr. A. Gordon Emslie

ACKNOWLEDGMENTS

I would like to express my gratitude to my academic advisor, Dr. Melinda McCann for her commitment to helping see this project trough to its final completion, and her wise guidance during its development. She always provided timely and instructive comments and evaluation at every stage of the thesis process, allowing me to complete this project on schedule. My sincere appreciation extends to my other committee members Dr. William Warde and Dr. Sthepanie Monks for their kind assistance and support.

I would also like to give my special appreciation to my family, specially my father and mother who instilled in me from an early age, the desire and skills to complete this Masters Degree. Further thanks go to my special boyfriend, Ashwin Bhuvanendran, for his support and encouragement at times of difficulty.

Finally, I would like to thank the Department of Statistics for supporting during these three years of my Master's Program.

TABLE OF CONTENTS

LIST OF TABLES

Table Page

CHAPTER 1

INTRODUCTION

When a researcher is performing a study with more than two levels of a factor variable, often the purpose is to quantify the differences among the factor level means. We refer to the problem of determining simultaneously which levels of a factor differ significantly from a control as multiple comparisons with a control (MCC). Dunnett (1955) developed conservative simultaneous (1-α)100% confidence intervals to estimate these differences simultaneously under normality with independent samples. Assuming that μ_k is the mean of the control and $\mu_1,...,\mu_{k-1}$ are the means of the non-control levels of the factor variable, the parameters of interest are $\mu_i - \mu_k$, for $i = 1,...,k-1$. Sometimes, however, the response of interest is binary. For example, when determining human immunodeficiency virus (HIV) seropositivity, individuals are classified as either positive or negative, and when detecting defective items in large populations, items are classified as either defective or non-defective. In cases like these, rather than estimating the differences between factor level means, the goal of the researcher might be to estimate differences between proportions of independent binomial variables, with $p_i - p_k$ for $i = 1,...,k-1$ as the parameters of interest, where p_k is the proportion of the level considered the control and $p_1,...,p_{k-1}$ are the proportions of the non-control levels of the factor of interest. A useful application of the later design occurs in the analysis of prevalence of a given disease in humans, where the population of interest is divided into subgroups that share the same characteristic such as age, race or social status. Similarly, in plant/animal disease assessment it is of interest to know the proportion of vectors (organisms such as insects)

capable of transmitting a virus, in this case the population can be subdivided by vector species, plant variety, or environmental conditions.

Consider specifically an experiment involving determination of HIV seropositivity. The HIV virus can be transmitted in different ways, through sexual contact with others that are infected, from an infected mother to her unborn baby, and through blood; the latter occurs when people share needles or from blood transfusions in hospitals. Currently, effective methods are available for prevention of HIV transmission through blood, as a result, the risk has been reduced to low levels in the developed countries; however, screening tests can be relatively expensive for developing countries where the government faces serious financial restrictions. One possible way to reduce the cost of HIV testing involves pooling the sera from several individuals and then testing the pool for HIV infection; we refer to this methodology as group-testing.

Group-testing involves pooling individuals into groups, testing the groups and classifying each group as either positive or negative. A negative result for the group occurs when none of the individuals possesses the characteristic of interest; a positive result for the group occurs when at least one of the individuals possesses the characteristic. This type of sampling provides substantial benefits, among them the reduction of cost of classifying all individuals of a population according to whether or not they possess a certain trait when the incidence rate is very low. Benefits from group testing depend on the size of the pools. Unfortunately, optimum selection of group sizes is difficult because it requires knowledge of the unknown prevalence (Hughes-Oliver & Rosenberger 2000).

The origin of group-testing is credited to Robert Dorfman (1943); he proposed the use of a simple binomial model to reduce the number of medical tests while detecting all members of a population that have the syphilis antigen. Dorfman suggested pooling several blood samples and testing the pool for the syphilis antigen; and in the cases where a group tested positive, all the subjects from that particular group would be re-tested on an individual basis. Dorfman's goal was to classify each of the individuals in the population as infected or not while reducing the expected

number of tests, this is known as the classification problem. One can also use group testing to estimate the proportion infected in a population without necessarily identifying the infected individuals, which is known as the estimation problem. In fact, Dorfman's idea of pooling blood samples for syphilis screening was not used; but some years later the concept was analyzed again and today useful applications can be found in the literature of group-testing. Among those applications are determination of HIV seropositivity (Gastwirth & Hammick 1988, Gastwirth & Johnson 1994, Kline, Brothers, Brookmeyer, Zeger & Quinn 1989, Wein & Zenios 1996), estimation of virus infection rates in plants (Bhattacharyya, Karandinos & DeFolliart 1979, Hepworth 1996, Swallow 1985, Tebbs & Bilder 2004, Thompson 1962), leak testing of sealed radioactive sources (Thomas, Pasternack, Vacirca & Thompson 1973) and genetics (Chick 1996); while group testing has been shown to have substantial benefits, multiplicity adjustments for experiments conducted under this protocol has been limited.

McCann and Tebbs (2006) derive two simultaneous confidence interval approaches for all pairwise proportion differences when the data is obtained under a group-testing scenario. Both of their procedures, Tukey-Kramer and Jeffreys-Perks, are based on asymptotic results; the authors evaluate both approaches in terms of simultaneous coverage probability and mean interval length for small sample cases. McCann and Tebbs conclude that both procedures have good performance for small samples, recommending the Tukey-Kramer approach for large groups of about 10 individuals or more; and the Jeffreys-Perks procedure for small groups of less than 10 individuals. McCann and Tebbs also illustrate the application of both procedures using data from an HIV study involving drug users in Houston. The population of interest is classified based on two different factors, race and type of drug injected, with three levels for the race factor (Hispanic, white, and black) and four levels for the type of drug used by the individual (heroin, cocaine/heroine, cocaine, and cocaine/amphetamines). Pairwise differences are estimated between the proportion of positives at each level within each factor. This methodology allows the

researcher to rank the different levels of a factor based on HIV prevalence while also retaining the ability to assess practical significance via the interval estimates.

In some cases, however, differences with a control are of interest. The goal of the current paper is to provide, within a group-testing context, simultaneous confidence intervals for proportion differences between the control and all other levels. Both simultaneous inferential procedures presented by McCann and Tebbs (2006), will be modified for this scenario and subsequently the Dunnett and Jeffreys-Perks simultaneous confidence intervals will be generated and illustrated using data from a multiple-vector transfer design study.

For reference throughout the paper, we now consider a specific example where MCC is clearly wanted and the data are correlated via a group testing protocol.

It is well known that plant viruses are the causes of big losses in crop production and quality all over the world, therefore developing treatments to control the spread of viruses is of prime importance. Normally, organisms such as insects transmit viruses. An insect that is capable of transmitting the causative agent of diseases is known as vector. In pathology, often it is of interest to estimate the probability of virus transmission by a single vector; one way to do it consists of collecting samples of insects and caging insects individually with a healthy plant. If after a virus test the plant tests positive, then the insect is classified as a vector; and the proportion of infected plants can be used as an estimate of the proportion of vectors in the population. Usually, the populations of insects are very large, and the previous procedure may be cost prohibitive due to the required number of plants, cages and space; thus it is convenient to test several insects on each healthy plant; this is known in the literature as multiple-vector transfer design (Thompson 1962).

Consider for instance a study by Bawden and Kassanis (1946) to compare the susceptibility of five different varieties of potato to the transmission of potato virus Y. In the same research, the authors performed an experiment, which showed that tobacco is more easily infected than potato by the virus Y, and thus it can be used as a reference when measuring

infection rate. For this reason, when illustrating the Dunnett and Jeffreys-Perks approaches under the MCC modality, tobacco is used as the control level and five different varieties of potato are used as the non-control levels. Bawden and Kassanis, first used two insects to transmit the virus from infected tobacco plants to 36 healthy plants of each of the varieties under analysis and the proportion of healthy plants that showed symptoms of the disease was recorded as a measure of the susceptibility of a plant to the virus Y infection. In a second replicate of the experiment, four insects were used to transmit the virus to another 36 healthy plants of each variety and again the proportion of test plants that became infected was recorded.

The remainder of the paper is organized as described next. In Chapter 2, we review the methodology of the Tukey-Kramer and Jeffreys-Perks procedures for all pairwise proportion differences, when the data is collected under a group-testing context. In Chapter 3, we present the theoretical background that makes possible the transition from all pairwise comparisons to multiple comparisons with a control. In Chapter 4, we evaluate the performance of both the Dunnett and Jeffreys-Perks approaches based on simultaneous coverage probability for various scenarios. In Chapter 5, we illustrate the above procedures using data from a multiple-vector transfer design study. In Chapter 6, we discuss the general conclusions of this work and consider some future extensions.

CHAPTER 2

THEORETICAL BACKGROUND

For the purpose of this analysis we assume that the population of interest is divided into subgroups that share a specific characteristic; henceforth, the subgroups will be referred as strata. Suppose there are a total of *k* strata; individuals are then pooled into groups within each stratum, with the *i*th stratum containing n_i groups all of sample size s_i, $i = 1,...,k$. In group-testing, each unit is assumed to represent an independent Bernoulli variable where the probability that a randomly selected subject possesses the characteristic of interest is p_i, and the probability of randomly selecting a subject not possessing the characteristic is $(1 - p_i)$. Thus the probability of randomly selecting a group of s_i subjects all of whom do not possess the characteristic of interest is $(1 - p_i)^{s_i}$, and consequently the probability of obtaining a group of s_i subjects where at least one subject possesses the trait of interest is $\theta_i = 1 - (1 - p_i)^{s_i}$. Let $Y_{il} = 1$ if the *l*th group in the *i*th stratum possesses the characteristic of interest, and $Y_{il} = 0$ otherwise, $i = 1,2,...,k$, $l = 1,2,...,n_i$. It is assumed that the Y_{il} are also independent and identically distributed Bernoulli random variables with mean $\theta_i = 1 - (1 - p_i)^{s_i}$. Using the invariance property of maximum likelihood estimators and the usual binomial MLE for θ_i, the MLE of p_i can be written as $\hat{p}_i = 1 - (1 - \hat{\theta}_i)^{1/s_i}$ for all i; where $\hat{\theta}_i = \sum_{l=1}^{n_i} y_{il} / n_i$ corresponds to the observed proportion of groups possessing the characteristic of interest in stratum i (see Appendix A for specific details).

Recall that our goal is to obtain simultaneous confidence intervals for all $p_i - p_k$, $i = 1,...,k-1$, where p_k represents the prevalence for the subgroup deemed the control in this setting. McCann and Tebbs (2006) consider a similar problem and develop two simultaneous confidence interval approaches for all pairwise proportion differences when the data is obtained under a group-testing scenario. Both of their procedures, the Tukey-Kramer and Jeffreys-Perks, are based on asymptotic results, and both have good performance for small sample cases in terms of simultaneous coverage probability and mean interval length. McCann and Tebbs apply both procedures to cases where the researcher is interested in all pairwise differences; for example, if the purpose is to rank different races based on the proportion of HIV positives, while retaining the ability to estimate the differences in these proportions, all pairwise differences can be estimated via simultaneous confidence intervals and the ranking determined as a result of the pairwise comparisons. These procedures can be modified appropriately for our all comparisons with a control scenario. The Tukey-Kramer and Jeffreys-Perks approaches derived by McCann and Tebbs are presented next.

2.1 Tukey-Kramer confidence intervals

The Tukey-Kramer method is an appropriate multiple comparison procedure when all pairwise differences are of interest. Hochberg and Tamhane (1987) provide the details of this multiple comparison procedure for the individual testing case. This approach can be also applied to the group-testing scenario since we have independent group binary responses for all strata. Hochberg and Tamhane recommend the use of the Tukey-Kramer intervals noting that they are conservative for large sample sizes ($n_i \geq 300$), and have smaller width than other typical procedures; however, for small sample sizes the procedure exhibits poor coverage characteristics and alternative procedures should be considered.

Recall that the usual binomial MLE, $\hat{\theta}_i$, has an asymptotic normal distribution with mean θ_i and variance $\theta_i(1-\theta_i)/n_i$. As $\hat{p}_i = g(\hat{\theta}_i) = 1-(1-\hat{\theta}_i)^{1/s_i}$, the delta method yields that the asymptotic distribution of $\hat{p}_i$ is also normal with mean p_i and variance $v_i \equiv v(p_i, s_i) = [s_i^{-2}(1-(1-p_i)^{s_i})(1-p_i)^{2-s_i}]/n_i$ (see Appendix B for specific details).

By the properties of the MLE's, $\hat{p}_i$ is a consistent estimator of p_i, and since $\hat{v}_i = v(\hat{p}_i, s_i)$ is a function of $\hat{p}_i$, we can also say that $\hat{v}_i$ is a consistent estimator of v_i. Hence, it is appropriate to use $\hat{v}_i$ as an estimate of the unknown quantity v_i. Thus, for the group-testing case, the Tukey-Kramer simultaneous $100(1-\alpha)$ percent asymptotic confidence intervals for all pairwise differences $p_i - p_j$, $i < j$, are given by

$$(\hat{p}_i - \hat{p}_j) \pm |q^*_{\alpha,k,\infty}| \sqrt{\hat{v}_i + \hat{v}_j} \qquad (1)$$

where $\hat{v}_i = v(\hat{p}_i, s_i)$ and $\hat{v}_j = v(\hat{p}_j, s_j)$ are the asymptotic estimated variances of $\hat{p}_i$ and $\hat{p}_j$ respectively, and $|q^*_{\alpha,k,\infty}|$ is as described in section 5.1.1 of Hsu 1996 (McCann and Tebbs 2006).

Since this statistical procedure involves proportions, some difficulties occur when none of the groups in stratum *i* and stratum *j* have the characteristic of interest ($x_i = x_j = 0$); in this case a degenerate interval is produced. There is also a problem when all the groups possess the characteristic of interest in stratum *i* or stratum *j* ($x_i = n_i$ or $x_j = n_j$); under this scenario a non-informative interval is given, [-1,1].

The second procedure derived by McCann and Tebbs (2006), the Jeffreys-Perks confidence intervals, is presented next.

2.2 Jeffreys-Perks confidence intervals

Beal (1987) analyses the use of asymptotically-based confidence intervals for the difference between the probability of success for two binomial populations. She evaluates the

statistical behavior of five such intervals (Wald, Mee, Miettinen & Nurminen, Haldane, and Jeffreys-Perks). Except for the Jeffreys-Perks intervals, these procedures will not be illustrated in this paper but they are described in Beal (1987). Beal recommends the Jeffreys-Perks interval if one wants to compute a simple interval. The Jeffreys-Perks interval shows a considerable improvement over other options when the values of both p_i and p_j are not very small or very large.

Piegorsch (1991) considers Beal's formulation of a Jeffreys-Perks interval, but rather than estimating the difference between the proportions of two binomial populations, Piegorsh considers simultaneous intervals for all $p_i - p_j$ and for all comparisons with a control. He shows that the Jeffreys-Perks procedure exhibits generally nominal empirical coverage characteristics, and recommends it for use with small to moderate sample sizes.

Both Beal (1987) and Piegorsch (1991), only consider the case when $s_i = s_j = 1$. However, McCann and Tebbs (2006) generalise the Jeffreys-Perks procedure for the case when $s_i > 1$ and $s_j > 1$, which corresponds to the group-testing scenario. They utilise Beal's reparameterisation, $a = p_i + p_j$ and $b = p_i - p_j$. However, Beal considers a prior distribution $f_{P_i,P_j}(p_i, p_j)$ that is proportional to $\{p_i(1-p_i)p_j(1-p_j)\}^{-1/2}$ for $0 < p_i < 1$ and $0 < p_j < 1$; but McCann and Tebbs place a non-informative prior distribution on the proportion of groups possessing the characteristic of interest (θ_i, θ_j), since group responses are observed under the group-testing scenario. Thus, the prior $f_{\Theta_i\Theta_j}(\theta_i, \theta_j)$ is proportional to $\{\theta_i(1-\theta_i)\theta_j(1-\theta_j)\}^{-1/2}$, for $0 < \theta_i < 1$ and $0 < \theta_j < 1$. Consequently, if we consider $p_i = 1-(1-\theta_i)^{1/s_i}$ and $p_j = 1-(1-\theta_j)^{1/s_j}$, solve for $\theta_i = 1-(1-p_i)^{s_i}$ and $\theta_j = 1-(1-p_j)^{s_j}$ respectively, and perform

a bivariate transformation with $|J| = \frac{\partial \theta_i}{\partial p_i}.\frac{\partial \theta_j}{\partial p_j} - \frac{\partial \theta_i}{\partial p_j}.\frac{\partial \theta_j}{\partial p_i} = s_i s_j (1-p_i)^{s_i - 1}(1-p_j)^{s_j - 1}$, then we can write the prior distribution on an individual scale as

$$f_{P_i P_j}(p_i, p_j) \propto f_{\Theta_i \Theta_j}(1-(1-p_i)^{s_i}, 1-(1-p_j)^{s_j})|J| \text{ or}$$

$$f_{P_i P_j}(p_i, p_j) \propto \{1-(1-p_i)^{s_i}\}^{-1/2}(1-p_i)^{(s_i/2)-1}\{1-(1-p_j)^{s_j}\}^{-1/2}(1-p_j)^{(s_j/2)-1}, \text{ for values of}$$

$0 < p_i < 1$ and $0 < p_j < 1$.

Now, let X_i be the number of groups in stratum i that possess the trait of interest, thus $X_i = \sum_{l=1}^{n_i} Y_{il}$. The $X_1, X_2, ..., X_k$ are independent and identically distributed Binomial random variables, hence the likelihood function $f_{X_i, X_j | P_i, P_j}(x_i, x_j \mid p_i, p_j)$ for $x_i = 0,1,...,n_i$ and $x_j = 0,1,...,n_j$, is proportional to $\{1-(1-p_i)^{s_i}\}^{x_i}(1-p_i)^{s_i(n_i - x_i)}\{1-(1-p_j)^{s_j}\}^{x_j}(1-p_j)^{s_j(n_j - x_j)}$. Consequently, the posterior distribution of p_i and p_j is given by

$$f_{P_i, P_j | X_i, X_j}(p_i, p_j \mid x_i, x_j) = \frac{f_{X_i, X_j | P_i, P_j}(x_i, x_j).f_{P_i, P_j}(p_i, p_j)}{\int_0^1 \int_0^1 f_{X_i, X_j | P_i, P_j}(x_i, x_j).f_{P_i, P_j}(p_i, p_j) dp_i dp_j}.$$

First, consider the denominator where the double integral can be expressed as follows:

$$\int_0^1 \{1-(1-p_i)^{s_i}\}^{x_i - 1/2}(1-p_i)^{s_i(n_i - x_i + 1/2)-1} dp_i * \int_0^1 \{1-(1-p_j)^{s_j}\}^{x_j - 1/2}(1-p_j)^{s_j(n_j - x_j + 1/2)-1} dp_j .$$

Since p_i and p_j are independent and identically distributed, then without loss of generality we can work with p_i individually. The integral involving p_i can be rewritten as

$$\int_0^1 \{1-(1-p_i)^{s_i}\}^{x_i - 1/2}(1-p_i)^{s_i n_i - s_i x_i + \frac{1}{2}s_i}(1-p_i)^{-1} dp_i . \quad (2)$$

Now, utilize the change of variable $t_i = (1-p_i)^{s_i}$ with $dt_i = -s_i(1-p_i)^{s_i - 1} dp_i$. Thus the integral (2) is equivalent to

$$-\frac{1}{s_i}\int_1^0 (1-t_i)^{x_i-\frac{1}{2}} t_i^{(n_i-x_i-\frac{1}{2})} dt_i = \frac{1}{s_i}\int_0^1 (1-t_i)^{x_i-\frac{1}{2}} t_i^{(n_i-x_i-\frac{1}{2})} dt_i \ . \tag{3}$$

Let $a = n_i - x_i + 1/2$ and $b = x_i + 1/2$. After adding the appropriate constants, equation (3) can be expressed as

$$\frac{1}{s_i}\frac{\Gamma(b)\Gamma(a)}{\Gamma(a+b)}\int_0^1 \frac{\Gamma(a+b)}{\Gamma(b)\Gamma(a)}(1-t_i)^{b-1} t_i^{a-1} dt_i \tag{4}$$

where $\Gamma(.)$ symbolizes the regular gamma function. Now the solution of the integral in equation (4) is one, and thus the posterior distribution of p_i can be written as

$$\frac{\{1-(1-p_i)^{s_i}\}^{x_i-\frac{1}{2}}(1-p_i)^{s_i(n_i-x_i+\frac{1}{2})-1}}{\frac{1}{s_i}\frac{\Gamma(b)\Gamma(a)}{\Gamma(a+b)}} = \frac{s_i\Gamma(n_i+1)}{\Gamma(x_i+1/2)\Gamma(n_i-x_i+1/2)} \times \{1-(1-p_i)^{s_i}\}^{x_i-\frac{1}{2}}(1-p_i)^{s_i(n_i-x_i+\frac{1}{2})-1}.$$

Consequently the posterior distribution of p_i and p_j is given by

$$c_{ij} \times \{1-(1-p_i)^{s_i}\}^{x_i-\frac{1}{2}}(1-p_i)^{s_i(n_i-x_i+\frac{1}{2})-1} \times \{1-(1-p_j)^{s_j}\}^{x_j-\frac{1}{2}}(1-p_j)^{s_i(n_j-x_j+\frac{1}{2})-1}$$

for $0 < p_i < 1$ and $0 < p_j < 1$, where the constant is

$$c_{ij} = \frac{s_i s_j \Gamma(n_i+1)\Gamma(n_j+1)}{\Gamma(x_i+1/2)\Gamma(n_i-x_i+1/2)\Gamma(x_j+1/2)\Gamma(n_j-x_j+1/2)}.$$

Now, let $\hat{p}_i$, $\hat{p}_j$ and $\hat{b}$ be the maximum likelihood estimates of p_i, p_j and b respectively. Then the Jeffreys-Perks simultaneous $100(1-\alpha)$ percent confidence intervals for all pairwise differences $p_i - p_j$, $i < j$, are constructed by solving

$$(b-\hat{b})^2 = | q^*_{\alpha,k,\infty} | V(\hat{b}; a = \tilde{a}) \tag{5}$$

for b, where $| q^*_{\alpha,k,\infty} |$ is described in equation (1), and $V(\hat{b}; a = \tilde{a})$ is the asymptotic variance of $\hat{b} \equiv \hat{p}_i - \hat{p}_j$, with $\tilde{a}$, the posterior mean of $a \equiv p_i + p_j$, inserted as an estimate of a.

The posterior mean of a is calculated by taking the expected value with respect to the posterior distribution $f_{P_i,P_j|X_i,X_j}(p_i, p_j \mid x_i, x_j)$ as follows

$$\tilde{a} = E_{P_i,P_j|X_i,X_j}(p_i + p_j \mid X_i = x_i, X_j = x_j) = \int_0^1 \int_0^1 (p_i + p_j) f_{P_i,P_j|X_i,X_j}(p_i, p_j \mid x_i, x_j) dp_i dp_j \ .$$

Again, because of independence and identical distributions we can work with p_i and p_j separately. The posterior mean of p_i can be estimated by solving the following equation

$$\int_0^1 p_i \frac{s_i \Gamma(n_i + 1)}{\Gamma(x_i + 1/2)\Gamma(n_i - x_i + 1/2)} (1 - (1 - p_i)^{s_i})^{x_i - 1/2} (1 - p_i)^{s_i(n_i - x_i + 1/2) - 1} dp_i \ .$$

If we use the same change of variables as before, $t_i = (1 - p_i)^{s_i}$ with $dt_i = -s_i (1 - p_i)^{s_i - 1} dp_i$, then we have

$$\int_0^1 (1 - t_i^{1/s_i}) \frac{\Gamma(n_i + 1)}{\Gamma(x_i + 1/2)\Gamma(n_i - x_i + 1/2)} (1 - t_i)^{x_i - 1/2} t_i^{(n_i - x_i - 1/2)} dt_i$$

$$= 1 - \int_0^1 \frac{\Gamma(n_i + 1)}{\Gamma(x_i + 1/2)\Gamma(n_i - x_i + 1/2)} (1 - t_i)^{x_i - 1/2} t_i^{(n_i - x_i - 1/2) + (1/s_i)} dt_i \ . \qquad (6)$$

Now let $a' = n_i - x_i + 1/2 + (1/s_i)$ and $b' = x_i + 1/2$, then equation (6) becomes

$$E_{p_i|x_i}[p_i \mid X_i = x_i] = 1 - \frac{\Gamma(n_i + 1)\Gamma(a')}{\Gamma(n_i - x_i + 1/2)\Gamma(a' + b')} \int_0^1 \frac{\Gamma(a' + b')}{\Gamma(a')\Gamma(b')} (1 - t_i)^{b' - 1} t_i^{a' - 1} dt_i \ .$$

The integral in this expression equals one and consequently the posterior mean of p_i is given by

$$1 - \frac{\Gamma(n_i + 1)\Gamma(n_i - x_i + 1/2 + (1/s_i))}{\Gamma(n_i - x_i + 1/2)\Gamma(n_i + 1 + (1/s_i))} .$$

Thus the posterior mean of $a = p_i - p_j$ can be written as

$$\tilde{a} = 2 - \frac{\Gamma(n_i + 1)\Gamma(n_i - x_i + 1/2 + (1/s_i))}{\Gamma(n_i - x_i + 1/2)\Gamma(n_i + 1 + (1/s_i))} - \frac{\Gamma(n_j + 1)\Gamma(n_j - x_j + 1/2 + (1/s_j))}{\Gamma(n_j - x_j + 1/2)\Gamma(n_j + 1 + (1/s_j))} .$$

When $\tilde{a}$ is used as an estimate of a, the estimated asymptotic variance of $\hat{b} \equiv \hat{p}_i - \hat{p}_j$ is as follows

$$V(\hat{b}; a = \tilde{a}) = \frac{\{1 - g_1(\tilde{a}, b)^{s_i}\} g_1(\tilde{a}, b)^{2-s_i}}{s_i^2 n_i} + \frac{\{1 - g_2(\tilde{a}, b)^{s_j}\} g_2(\tilde{a}, b)^{2-s_j}}{s_j^2 n_j},$$

where $g_1(\tilde{a}, b) = (2 - \tilde{a} - b)/2$ and $g_2(\tilde{a}, b) = (2 - \tilde{a} + b)/2$ (see Appendix C for specific details).

As with the Tukey-Kramer intervals, some difficulties can be found with calculation of these intervals when all the groups within strata *i* and *j* possess the characteristic of interest or when none of the groups possess such a trait. When $x_i \cong n_i$ and $x_j \cong n_j$, equation (5) may not have any solution, in this case a non-informative interval [-1,1] is set for $p_i - p_j$. When $x_i = x_j = 0$, often only one solution is present. In this case a one-sided interval will be provided.

Both procedures, Tukey-Kramer and Jeffreys-Perks, as presented by McCann and Tebbs, include a critical value $|q^*_{\alpha,k,\infty}|$ which is the upper α critical point described in Section 5.1.1 of Hsu (1996); this critical value is appropriate when all pairwise comparisons are of interest; however, multiple comparisons with a control are of concern in the current paper and therefore a Dunnett critical value $|d^*_{\alpha,k,\infty}|$ is required. Next we describe how this Dunnett critical point is estimated when the parameter of interest is $p_i - p_k$, where *k* represents the control level, and we also present the Dunnett and Jeffreys-Perks confidence intervals utilizing a Dunnett critical point which is appropriate for simultaneous inferences with a control for the group testing scenario.

CHAPTER 3

MULTIPLE COMPARISONS WITH A CONTROL

When a researcher is interested in comparing all treatment groups with a control group, the multiple comparison procedure due to Dunnett is often appropriate. Dunnett (1955) provides the details on calculating the appropriate critical value for simultaneous confidence intervals for the differences between all non-control level factor means with the mean for the control level. The parameters of interest in this case are $\mu_i - \mu_k$, where μ_k represents the mean of the control factor and μ_i is the mean of the *i*th non-control factor for *i*=1,...,*k*-1. The point estimate of $\mu_i - \mu_k$ is the difference between the sample means $\bar{u}_i - \bar{u}_k$, where $\bar{u}_k$ is the sample mean of the control factor and $\bar{u}_i$ is the sample mean of the *i*th non-control factor, *i*=1,…,*k*-1. Under the assumption that $\bar{u}_i$ and $\bar{u}_k$ have independent normal distributions with constant variance σ^2, $\bar{u}_i - \bar{u}_k$ is normally distributed with mean $\mu_i - \mu_k$ and variance $\frac{\sigma^2}{n_i} + \frac{\sigma^2}{n_k}$, where n_k and n_i are the sample sizes of the control factor and the *i*th non-control factor, *i*=1,…,*k*-1, respectively. If we let

$$z_i = \frac{\bar{u}_i - \bar{u}_k - (\mu_i - \mu_k)}{\sigma\sqrt{\frac{1}{n_i} + \frac{1}{n_k}}}$$

then z_i has a normal distribution with mean zero and variance one. As we are concerned with simultaneous intervals, we need to examine the joint distribution of $u_i^* = \bar{u}_i - \bar{u}_k$, *i*=1,...,*k*-1. The joint distribution of the u_i^*, *i*=1,...,*k*-1, is a multivariate normal distribution with mean vector

$\boldsymbol{\mu} = (\mu_1 - \mu_k, \mu_2 - \mu_k, ..., \mu_{k-1} - \mu_k)'$ and variance-covariance matrix given by

$$\begin{bmatrix} \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_k} & \frac{\sigma^2}{n_k} & . & \frac{\sigma^2}{n_k} \\ \frac{\sigma^2}{n_k} & \frac{\sigma^2}{n_2} + \frac{\sigma^2}{n_k} & & . \\ . & & . & \frac{\sigma^2}{n_k} \\ \frac{\sigma^2}{n_k} & . & \frac{\sigma^2}{n_k} & \frac{\sigma^2}{n_{k-1}} + \frac{\sigma^2}{n_k} \end{bmatrix}.$$

Thus the correlation coefficient between u_i^* and u_j^* is obtained as follows

$$\rho_{ij} = \frac{\frac{\sigma^2}{n_k}}{\sqrt{\frac{\sigma^2}{n_i} + \frac{\sigma^2}{n_k}}\sqrt{\frac{\sigma^2}{n_j} + \frac{\sigma^2}{n_k}}} = \frac{\frac{\sigma^2}{n_k} \cdot \frac{n_k}{\sigma^2}}{\sqrt{\frac{\sigma^2}{n_i} \cdot \frac{n_k}{\sigma^2} + \frac{\sigma^2}{n_k} \cdot \frac{n_k}{\sigma^2}}\sqrt{\frac{\sigma^2}{n_j} \cdot \frac{n_k}{\sigma^2} + \frac{\sigma^2}{n_k} \cdot \frac{n_k}{\sigma^2}}}$$

$$= \frac{1}{\sqrt{\frac{n_k}{n_i} + 1}\sqrt{\frac{n_k}{n_j} + 1}} = \left(\frac{n_k}{n_i} + 1\right)^{-1/2}\left(\frac{n_k}{n_j} + 1\right)^{-1/2} = \lambda_i \lambda_j,$$

where $\lambda_i = \left(\frac{n_k}{n_i} + 1\right)^{-1/2}$, for i=1,…,k-1.

Dunnett further assumes that an estimate of σ^2 is available which is independent of $\bar{u}_1, ..., \bar{u}_k$; for example he proposes using $s^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)^2 / n$ where $n = (\sum_{i=1}^{k} n_i) - k$; hence $\frac{ns^2}{\sigma^2}$ has a chi-square distribution with n degrees of freedom and $t_i = \frac{z_i}{\sqrt{\frac{ns^2}{\sigma^2}/n}} = \frac{z_i}{s/\sigma}$ follows a student's t-distribution with n degrees of freedom.

The required critical values d_i needed to construct simultaneous confidence intervals for $\mu_i - \mu_k$, i=1,…,k-1, are then chosen to satisfy $\mathrm{P}(|t_i| \le d_i \forall i) = \mathrm{P}(|z_i| \le \frac{sd_i}{\sigma} \forall i) = P$, where P is the

joint confidence coefficient. The solutions are found by solving the following equation for d_i, i=1,…,k-1

$$\int_{-\infty}^{\infty} F\left((d_1 \frac{s}{\sigma}, d_2 \frac{s}{\sigma}, ..., d_{k-1} \frac{s}{\sigma}\right) p(s) ds$$

where $F(z_1, z_2, ..., z_{k-1})$ is the multivariate normal c.d.f. of the $|z_i|$ and *p(s)* is the probability density function of *s*. Standard practice is to require $d_1 = d_2 = ... = d_{k-1} = d$. Then Dunnett's procedure sets the above integral equal to *P* and solves for d.

However, for the purpose of the current paper, the parameters of interest are not $\mu_i - \mu_k$, i=1,…,k-1, but rather $p_i - p_k$, i=1,…,k-1, as defined in Chapter 1. Recall that in Chapter 2 the MLE of p_i, $\hat{p}_i = 1 - (1 - \hat{\theta}_i)^{1/s_i}$, and its asymptotic distribution were derived. As the strata are independent, the joint asymptotic distribution of the vector $\hat{\boldsymbol{p}} = (\hat{p}_1, ..., \hat{p}_k)'$ is multivariate normal with mean $\boldsymbol{p} = (p_1, ..., p_k)'$ and variance-covariance matrix given by

$$\Sigma = \begin{bmatrix} v_1 & 0 & 0 & 0 & 0 \\ 0 & v_2 & 0 & 0 & 0 \\ 0 & 0 & . & 0 & 0 \\ 0 & 0 & 0 & . & 0 \\ 0 & 0 & 0 & 0 & v_k \end{bmatrix}$$

where $v_i = s_i^{-2}\{(1-(1-p_i)^{s_i})(1-p_i)^{2-s_i}\}/n_i$. (By arguments similar to those in Appendix B).

Since our goal is to estimate simultaneous confidence intervals for each of the differences $p_i - p_k$, we need to determine the joint distribution of ($g_1(\hat{\boldsymbol{p}}) = \hat{p}_1 - \hat{p}_k$, $g_2(\hat{\boldsymbol{p}}) = \hat{p}_2 - \hat{p}_k$,…, $g_{k-1}(\hat{\boldsymbol{p}}) = \hat{p}_{k-1} - \hat{p}_k$)', where *k* again denotes our control level. By the multivariate delta method $\boldsymbol{g}(\hat{\boldsymbol{p}}) = (g_1(\hat{\boldsymbol{p}}), ..., g_{k-1}(\hat{\boldsymbol{p}}))'$ has an asymptotic normal distribution with mean $\boldsymbol{g}(\boldsymbol{p})$ and variance $\boldsymbol{V} = \boldsymbol{C}\Sigma\boldsymbol{C}'$, where $\boldsymbol{C}$ is a (*k-1*) by *k* matrix of partial derivatives with elements given by

$$\boldsymbol{C} = \begin{bmatrix} \frac{dg_1(\hat{\boldsymbol{p}})}{\partial p_1} & . & . & . & \frac{dg_1(\hat{\boldsymbol{p}})}{\partial p_k} \\ . & & . & & . \\ . & & . & & . \\ \frac{dg_{k-1}(\hat{\boldsymbol{p}})}{\partial p_1} & & & & \frac{dg_{k-1}(\hat{\boldsymbol{p}})}{\partial p_k} \end{bmatrix} = \begin{bmatrix} 1 & 0 & . & . & -1 \\ 0 & 1 & 0 & . & -1 \\ . & . & . & 0 & . \\ 0 & . & 0 & 1 & -1 \end{bmatrix}.$$

Thus $\boldsymbol{V}$ is a *(k-1)* by *(k-1)* matrix given by

$$\boldsymbol{V} = \begin{bmatrix} v_1 + v_k & v_k & . & v_k \\ v_k & v_2 + v_k & & \\ . & & . & v_k \\ v_k & . & v_k & v_{k-1} + v_k \end{bmatrix}$$

where $v_i = s_i^{-2}\{1-(1-p_i)^{s_i}\}(1-p_i)^{2-s_i} / n_i$, i=1,…,k.

The correlation between $g_i(\hat{\boldsymbol{p}})$ and $g_j(\hat{\boldsymbol{p}})$ can be calculated via the following equation:

$$\rho_{ij} = \frac{v_k}{\sqrt{v_i + v_k}\sqrt{v_j + v_k}} = \left(1 + \frac{v_i}{v_k}\right)^{-1/2}\left(1 + \frac{v_j}{v_k}\right)^{-1/2} = \lambda_i \lambda_j \tag{7}$$

where

$$\lambda_i = \left(1 + \frac{n_k}{n_i}.\frac{s_i^2\{1-(1-p_i)^{s_i}\}(1-p_i)^{2-s_i}}{s_k^2\{1-(1-p_k)^{s_k}\}(1-p_k)^{2-s_k}}\right)^{-1/2}, \; i = 1,...,k-1.$$

Now unlike Dunnett (1955), we do not have an independent estimate for the variance component, and consequently we cannot utilize a student's t-distribution. Instead we can construct test statistics with an asymptotic multivariate normal distribution with a mean vector of zeros and a covariance matrix with one's on the diagonal via standardization; i.e.

$$\frac{(\hat{p}_i - \hat{p}_k) - (p_i - p_k)}{\sqrt{v_i + v_k}} \xrightarrow{d} z_i \sim \mathrm{N}(0,1),\ i=1,\ldots,k\text{-}1. \tag{8}$$

Consequently infinity degrees of freedom are used in the previous Dunnett formulation, as a t_∞ distribution corresponds to a z distribution.

Since the v_i, $i = 1,...,k$, are unknown quantities we cannot obtain the exact critical value d. (Even if we knew v_i the corresponding d would still yield only asymptotically correct intervals as $\hat{p}_i - \hat{p}_k$ is only asymptotically normal.) However, as $\hat{v}_i$ is a consistent estimator of v_i (by the properties of the MLE's); we can use $\hat{v}_i$ in equations (7) and (8) for v_i and obtain a consistent estimator of the asymptotically correct critical value d. Consequently $\hat{\lambda}_i = \lambda(\hat{v}_i)$ are used as estimates of λ_i, $i = 1,...,k-1$.

In order to estimate the appropriate Dunnett critical values required to construct simultaneous confidence intervals for $p_i - p_k$ where k is the control level, a Fortran program provided by Hsu (1996) was adapted for our specific group testing scenario. The inputs for this computer program are the significance level, the degrees of freedom and $\hat{\boldsymbol{\lambda}} = (\hat{\lambda}_1, \hat{\lambda}_2, ..., \hat{\lambda}_{k-1})$.

After calculating the appropriate Dunnett critical point, the simultaneous confidence intervals are estimated. These intervals involve adjusting the simultaneous inference procedures derived by McCann and Tebbs (2006) for the MCC scenario and are used to simultaneously estimate proportion differences between a control group and several non-control groups.

Thus, for the group-testing case, the Dunnett simultaneous 100(1-α) percent confidence intervals for the differences $p_i - p_k$, where k represents the group that acts as control, are given by

$$(\hat{p}_i - \hat{p}_k) \pm | d^*_{\alpha,k,\infty} | \sqrt{\hat{v}_i + \hat{v}_k} \qquad (9)$$

where $\hat{v}_i = v(\hat{p}_i, s_i)$ and $\hat{v}_k = v(\hat{p}_k, s_k)$ are the asymptotic variances of $\hat{p}_i$ and $\hat{p}_k$ respectively, and $| d^*_{\alpha,k,\infty} |$ is the Dunnett critical value calculated using $\hat{\boldsymbol{\lambda}}$.

Additionally, the Jeffreys-Perks simultaneous 100(1-α) percent confidence intervals for the differences $p_i - p_k$, where k is again the group that acts as control, are constructed by solving

$$(b - \hat{b})^2 = | d^*_{\alpha,k,\infty} | V(\hat{b}; a = \tilde{a}) \qquad (10)$$

for b (recall $b = p_i - p_k$), where $| d^*_{\alpha,k,\infty} |$ is as in equation (9), and $V(\hat{b}; a = \tilde{a})$ is the asymptotic variance of $\hat{b} \equiv \hat{p}_i - \hat{p}_k$, with $\tilde{a}$, the posterior mean of $a \equiv p_i + p_k$, inserted as an estimate of a. (See Chapter 2 for a detailed derivation of the Jeffreys-Perks approach).

CHAPTER 4

SIMULATIONS

The Dunnett and Jeffreys-Perks approaches derived in Chapter 3 for multiple comparisons with a control under a group testing scenario, are now evaluated and compared in terms of simultaneous coverage probability.

Usually, when working with proportions, some complications arise when either all responses are positive or when all are negative. For example, in our setting, suppose that the number of groups possessing the trait of interest in stratum *k,* is zero; that suggests that the observed proportion of groups with the characteristic of interest in stratum *k* is zero, $\hat{\theta}_k = \frac{x_k}{n_k} = 0$, and consequently $\hat{p}_k = 1-(1-\hat{\theta}_k)^{1/s_k} = 0$, where $\hat{p}_k$ and s_k are as defined in previous Chapters. When $\hat{p}_k = 0$, then the estimated variance $\hat{v}_k = s_k^{-2}\{1-(1-\hat{p}_k)^{s_k}\}(1-\hat{p}_k)^{2-s_k}/n_k = 0$, and as a result the correlation between $g_i(\hat{\boldsymbol{p}}) = \hat{p}_i - \hat{p}_k$ and $g_j(\hat{\boldsymbol{p}}) = \hat{p}_j - \hat{p}_k$ is non-estimable; recall that this correlation is estimated by $\hat{\rho}_{ij} = \left(1+\frac{\hat{v}_i}{\hat{v}_k}\right)^{-1/2}\left(1+\frac{\hat{v}_j}{\hat{v}_k}\right)^{-1/2} = \hat{\lambda}_i\hat{\lambda}_j$. (See equation (7) in Chapter 3 for more details). If $\hat{\rho}_{ij} = \hat{\lambda}_i\hat{\lambda}_j$ cannot be estimated, then we also cannot estimate the Dunnett critical point and therefore the confidence intervals are not available. A similar problem occurs when $x_k = n_k$.

Having all responses positive is not only a problem for the control level, but also for the non-control levels. In order to calculate the Dunnett critical values we need to solve an equation that involves the factor $c_i = 1/(1-\hat{\lambda}_i^2)^{1/2}$; however recall that $\hat{\theta}_i = 1$ implies that $\hat{v}_i = 0$ and

$\hat{\lambda}_i = 1$, and consequently the factor c_i is non-estimable, and once again the confidence intervals would not be available.

To avoid these problems, we make some adjustments every time the observed proportion of positives is either zero or one. When $\hat{\theta}_i = 0$, we set $\hat{\theta}_i = \frac{0.5}{n_i}$ and when $\hat{\theta}_i = 1$, we set $\hat{\theta}_i = \frac{(n_i - 0.5)}{n_i}$. While these estimates are not technically the MLE's, notice that as $n \to \infty$ the probability of making such adjustments, provided $0 < p_i < 1$, i=1,…,k, will converge to zero. Thus, the adjustments should not significantly change the overall behavior of our estimator for large samples. The percentage of times such adjustments were made is included in the simulation results.

To run the simulations, random draws were taken from a binomial distribution with parameters θ_i and n_i; recall from Chapter 2 that x_i, i=1,…,k, is a binomial random variable. A significance level of 0.95 was used throughout. Now, let k be the number of strata, $\boldsymbol{n}$=(n_1,..., n_k) the number of groups within each stratum, $\boldsymbol{p}$=(p_1,...,p_k) the vector of proportions, and $\boldsymbol{s}$=(s_1,...,s_k) the respective group sizes. In our setting the first stratum is considered the control level. Now, using different values of $\boldsymbol{n}$, $\boldsymbol{s}$, $\boldsymbol{p}$ and k, and keeping for simplicity n_i and s_i constant in all strata; we ran 10,000 simulations for each set of data and estimated the simultaneous coverage for the confidence intervals derived in equations (9) and (10) in Chapter 3.

In order to estimate this simultaneous coverage probability, we simply counted the number of times the intervals simultaneously contained the true differences $p_i - p_k$, i=1,...,k-1, and divided that number by 10,000. We also kept track of the number of times we made the adjustment previously discussed to avoid $\hat{\theta}_i = 0$ or $\hat{\theta}_i = 1$. The percentage of times the adjustments were made to avoid $\hat{\theta}_i = 0$ for each set of data is provided in Table 2. The

percentage of times the adjustments were made to avoid $\hat{\theta}_i = 1$ for each set of data is provided in Table 3, and the estimated simultaneous coverage probability results are shown in Table 1.

Since the nominal coverage is 0.95, we expect that when the number of groups within each stratum, n_i, increases, then the estimated coverage probability converges to the 0.95 nominal level. In Table 1, the estimated coverage probabilities in bold are those within two estimated standard deviations using 0.95 as the true proportion value.

From Table 1, we notice that in the case of individual testing, both the Dunnett and Jeffreys-Perks procedures are very conservative as the coverage probabilities are more than two standard deviations above 0.95 for most of the cases. (This is probably due to the large percentage of times we make the adjustments in these cases.) We also notice that when $s_i > 1$, the Jeffreys-Perks procedure is often slightly more conservative than the Dunnett procedure, since more of the time the coverage probability is larger for the Jeffreys-Perks case.

From Table 1, we can also observe that when s_i=15, both procedures perform fairly well as the estimated coverage probability is very close to the nominal coverage of 0.95. When s_i=5 or s_i=10, the Jeffreys-Perks procedure tends to perform better than the Dunnett procedure, as it is often closer to the nominal coverage level. In the situations where none of the approaches is close to the nominal level, Jeffreys-Perks provides conservative confidence intervals while the Dunnett procedure is anticonservative, this also suggests that it is better to use the Jeffreys-Perks approach. In conclusion, both the Dunnett and Jeffreys-Perks approaches perform well for large group sizes ($s_i \geq 15$). For smaller groups sizes ($5 < s_i < 10$) the Jeffreys-Perks methodology is recommended.

In Tables 2 and 3, we present the number of times an adjustment was made due to $\sum_{i=1}^{n_i} x_i = 0$ and $\sum_{i=1}^{n_i} x_i = n_i$, i=1,...,k, respectively. As expected, the first type of adjustment when $\hat{\theta}_i = 0$, occurred much more often than the second type of adjustment, when $\hat{\theta}_i = 1$. This is because the true values of the p_i's in our simulations are typically very small (0.01-0.13),

consequently the θ_i's are also small and it is reasonable to obtain an observed number of groups with the characteristic of interest of zero.

CHAPTER 5

APPLICATION

In this Chapter a practical application of the Dunnett and Jeffreys-Perks procedures will be presented using data from a multiple-vector transfer design study.

In order to control the spread of plant diseases in Pathology, often it is of interest to estimate the proportion of a population that is infected with a plant virus. A multiple-vector transfer design is widely used for this purpose. This procedure consists of collecting a sample of aphids or insects, feeding them on an infected plant, and then separating these insects into groups and caging each group with a healthy or test plant. Subsequently, the researchers observe if the test plant develops the symptoms of the disease or not, and this is how the proportion of infected plants is estimated.

To illustrate the Dunnett and Jeffreys-Perks confidence intervals, data obtained from a multiple-vector transfer design was used. The objective of this study was to compare the susceptibility of different varieties of potato to the transmission of potato virus Y (Bawden and Kassanis 1946). The different varieties of potato analyzed were arran banner, majestic, arran consul, arran pilot and may queen.

Bawden and Kassanis wanted to identify the resistance of various potato varieties to the potato virus Y. For that purpose, they performed field and glasshouse experiments comparing several potato diversities. The former type of analysis involved plots of different varieties, which were exposed equally to chances of infection. The plots consisted of five rows of five plants each, where the plant located at the center was infected with virus Y. The results of this experiment are not provided in the current paper, however they can be found on Bawden and Kassanis (1946). The glasshouse experiments are described next.

Bawden and Kassanis analyzed the infection rate of different potato varieties using first an inoculation procedure and then using insects to transmit the virus Y. The inoculation method consists of taking sap from infected plants and rubbing healthy plants with the substance. Bawden and Kassanis obtained sap from plants infected with the virus Y and used it as inoculum after dilution with water. Ten plants of each variety of potato and tobacco were inoculated. Two leaflets of approximately the same size were rubbed over their surfaces with the inoculum. The results showed that 100% of the tobacco plants were infected, but the observed proportion that was infected was much lower with the potato varieties. From this study, Bawden and Kassanis concluded that tobacco is more easily infected than potato. Therefore, under our multiple comparisons with a control setting, tobacco is used as the control level.

When Bawden and Kassanis analyzed the infection rate of the varieties of potato under a multiple-vector transfer design protocol, they used the aphid Myzus persicae Sulz. to transmit the virus. The aphids were raised on turnips or radishes and starved for 4 hours. Then they were fed for 3 or 4 minutes on the undersurface of an infected tobacco leaf and immediately transferred to the healthy potato and tobacco plants. Thirty-six plants of each variety were colonized with two aphids and the number of infected plants was counted. To assess the differences among the varieties, Bawden and Kassanis simply observed the proportion of plants per variety that became infected after exposure to the virus Y; these observed proportions are given in Table 4. The researchers did not adjust in their analyses for either the group-testing setting or multiplicity, probably because these procedures were formally derived years later. (Dunnett's procedure for means appeared in 1955 and formal analyses of group testing experiments had only just began to appear by 1946). Regardless, simply reporting the percentage of infected plants does not give accurate estimates of the percentage of aphids that transmit the disease nor does it allow for accurate determination of the potato varieties whose transmission rates differ significantly from tobacco, the control.

In our current paper, the data from the Bawden and Kassanis studies were extended for our MCC case under a group-testing protocol. The differences between the proportion of infected plants of each of the varieties of potato and the proportion of infected tobacco plants were simultaneously estimated using the Dunnett and Jeffreys-Perks procedures with 95% confidence. The estimated confidence intervals are shown in Table 5.

From Table 5 we notice that the Dunnett and the Jeffreys-Perks approaches generated very similar results. We can conclude that the proportion of infected tobacco plants is significantly greater than the proportion of infected plants for the following potato varieties: arran banner, majestic, arran consul, and arran pilot. We did not find a significant difference between tobacco and the may queen potato variety.

Bawden and Kassanis again performed the same general experiment using insects, but this time they colonized the healthy plants with four aphids rather than two; the other conditions remain the same. In Table 4 we present the observed proportion of test plants that were infected for the different varieties when s_i=4. As we expected, increasing the number of aphids per test plant (group size), resulted in an increased observed proportion of infected plants. Since tobacco is very sensitive to the Y virus; with four aphids per plant, all of the tobacco test plants acquired the disease. Now the difference between tobacco and each of the potato varieties, as far as infected plants is concerned, is larger; which agrees with the main conclusion provided by Bawden and Kassanis.

Table 6 shows the Dunnett and Jeffreys-Perks 95% confidence intervals when s_i=4. Since there is a larger observed difference between tobacco and each of the potato varieties, the confidence intervals are wider than in the previous case, when s_i=2. However, from this study we can conclude that tobacco is significantly more susceptible to infection than all of the potato varieties considered in this study. We also observed that the Jeffreys-Perks procedure found only one root for all intervals within the [-1,1] range and therefore the upper limit was set to 1; consequently the Jeffreys-Perks approach generated wider intervals than Dunnett's. (This is not

surprising as the Jeffreys-Perks procedure can yield only one root when the observed proportion for one group is at or near zero or one. Recall that here $\hat{\theta}$ for tobacco is one.)

In this section we presented the results of two general experiments performed by Bawden and Kassanis (1946). Their goal was to analyze the susceptibility of potato varieties to the virus Y. The experiments were conducted under a group-testing setting and used tobacco as a control level to compare several potato varieties with respect to the transmission rate. The authors did not use any statistical procedures that adjusted for the group testing or multiplicity involved in this experiment. In contrast, our analysis of the work of Bawden and Kassanis used the MCC setting under a group-testing protocol and conclusions that are more accurate are provided.

CHAPTER 6

SUMMARY AND CONCLUSIONS

McCann and Tebbs (2006) presented two approaches, the Tukey-Kramer and Jeffreys-Perks, for obtaining pairwise simultaneous confidence intervals for all comparisons of proportions under a group-testing protocol. In some cases, however, differences with a control rather than all pairwise comparisons are of interest. The purpose of the current paper is to provide, within a group-testing scenario, simultaneous confidence intervals for proportion differences between the control and all other levels. Consequently, the McCann and Tebbs approaches were adjusted for the MCC case; with the main difference here being that a Dunnett critical value is required to construct the confidence intervals. This critical point depends on the correlation between $g_i(\hat{\boldsymbol{p}}) = \hat{p}_i - \hat{p}_k$ and $g_j(\hat{\boldsymbol{p}}) = \hat{p}_j - \hat{p}_k$, where the sub index k represents the control level.

The two procedures discussed in this paper for the MCC case, Dunnett and Jeffreys-Perks, were derived in Chapter 3 and evaluated and compared in terms of coverage probability in Chapter 4. Ten thousand simulations were ran for each specific scenario which consisted of specified values for $\boldsymbol{p}=(p_1,\ldots,p_k)$, $\boldsymbol{n}=(n_1,\ldots,n_k)$, and $\boldsymbol{s}=(s_1,\ldots,s_k)$. Three, four and five strata were also considered with a confidence level of 0.95. The coverage probability was then estimated for both procedures in these scenarios. The results show that for large group sizes, $s_i \geq 15$, the Dunnett and the Jeffreys-Perks approaches both perform very well. As the Dunnett procedure is much less intense computationally, we recommend Dunnett for these situations. For smaller group sizes, the Jeffreys-Perks methodology is recommended since it is closer to the nominal level of 1-α and is a conservative procedure.

Both procedures, Dunnett and Jeffreys-Perks, were also used to evaluate real data from a multiple-vector transfer design study. Comparing the susceptibility of tobacco and five different potato varieties to the virus potato Y. When a group size of two was used, tobacco proved more sensitive than four of the potato varieties. When a group size of four was used, tobacco was more susceptible to infection than all of the potato varieties considered in this study.

The ideas presented in this paper can also be extended to other cases. For example if the investigator wants to select treatments or strata from a set such that the best treatment or population is included in a winning subset; then multiple comparisons with the best treatment or best strata level could be of interest. A Dunnett critical point is also utilized in this scenario, so the results of our setting could easily be adapted. In addition, although the calculations are much more intense, it is also possible to extend these results to other sets of defined contrasts, which may be required to answer research questions specific to a particular experiment. In this case, we would use a critical value from a conservative method such as Bonferroni, Sidak, McCann-Edwards or Hunter-Worsley, again with infinity degrees of freedom as we are analyzing proportional data.

BIBLIOGRAPHY

Bhattacharyya, M., Karandinos, M. & DeFoliart, G. (1979). Point estimates and confidence intervals for infection rates using pooled organisms in epidemiologic studies. *American Journal of Epidemiology*, **109**, 124-131.

Bawden, F. & Kassanis, B. (1946). Varietal susceptibility to potato virus Y. *Annals of Applied Biology*, **33**, 46-50.

Beal, S. (1987). Asymptotic confidence intervals for the difference between two binomial parameters for use with small samples. *Biometrics*, **43**, 941-950.

Chick, S. (1996). Bayesian models for limiting dilution assay and group test data. *Biometrics*, **52**, 1055-1062.

Dorfman, R. (1943). The detection of defective members of large populations. *Annals of Mathematical Statistics*, **14**, 436-440.

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

Gastwirth, J. & Hammick, A. (1988). Estimation of the prevalence of a rare disease, preserving the anonymity of the subjects by group testing: application to estimating the prevalence of AIDS antibodies in blood donors. *Journal of Statistical Planning and Inference*, **22**, 15-27.

Gastwirth, J. & Johnson, W. (1994). Screening with cost-effective quality control: potential applications to HIV and drug testing. *Journal of the American Statistical Association*, **89**, 972-981.

Hepworth, G. (1996). Exact confidence intervals for proportions estimated by group testing. *Biometrics*, **52**, 1134-1146.

Hochberg, Y. & Tamhane, A. (1987). Multiple Comparison Procedures. New York: Wiley.

Hsu, J. (1996). Multiple Comparisons: Theory and Methods. Boca Raton: Chapman & Hall/CRC.

Hughes-Oliver, J. & Rosenberger, W. (2000). Efficient estimation of the prevalence of multiple rare traits. *Biometrika*, **87**, 315-327.

Kline, R., Brothers, T., Brookmeyer, R., Zeger, S. & Quinn, T. (1989). Evaluation of human immunodeficiency virus seroprevalence in population surveys using pooled sera. *Journal of Clinical Microbiology*, **27**, 1449-1452.

Lehmann, E. & Casella, G. (1998). Theory of point estimation. Springer Texts in Statistics. Second Edition.

McCann, M. & Tebbs, J. (2006). Pairwise comparisons for proportions estimated by pool testing. *Journal of Statistical Planning and Inference*, in press.

Piegorsch, W. (1991). Multiple comparisons for analyzing dichotomous response. *Biometrics*, **47**, 45-52.

Swallow, W. (1985). Group testing for estimating infection rates and probabilities of disease transmission. *Phytopathology*, **75**, 882-889.

Tebbs, J. & Bilder, C. (2004). Confidence interval procedures for the probability of disease transmission in multiple-vector-transfer designs. *Journal of Agricultural, Biological, and Environmental Statistics*, **9**, 1-16.

Thomas, J., Pasternack, B., Vacirca, S. & Thompson, D. (1973). Application of group testing procedures in radiological health. *Health Physics Pergamon Press*, 259-266.

Thompson, K. (1962). Estimation of the proportion of vectors in a natural population of insects. *Biometrics*, **18**, 568-578.

Wein, L. & Zenios, S. (1996). Pooled testing for HIV screening: capturing the dilution effect. *Operations Research*, **44**, 543-569.

Table 1: Estimated simultaneous coverage probabilities for the Dunnett (D) and Jeffreys-Perks (JP) procedures with α=0.05 where the number of tests, n_i, and the group size, s_i, are constant for all strata and the first stratum is considered the control level.
*The estimated coverage probabilities in bold are those within two estimated standard deviations using 0.95 as the true proportion.

	p=(.01,.02,.03)		p=(.01,.03,.05)		p=(.01,.04,.07)		
s_i	n_i=25	n_i=50	n_i=25	n_i=50	n_i=25	n_i=50	
1	1.000	0.999	1.000	0.991	0.999	0.961	D
	1.000	1.000	1.000	0.998	0.999	0.973	JP
5	0.957	0.937	0.921	0.934	0.922	0.942	D
	0.979	0.963	0.964	**0.950**	0.937	**0.954**	JP
10	**0.947**	**0.950**	0.942	0.942	0.943	0.943	D
	0.964	0.956	0.956	**0.952**	**0.951**	**0.946**	JP
15	**0.947**	**0.952**	**0.946**	**0.946**	0.944	**0.947**	D
	0.961	**0.953**	**0.951**	**0.953**	**0.953**	**0.951**	JP
	p=(.01,.02,.03,.04)		p=(.01,.03,.05,.07)		p=(.01,.04,.07,.10)		
s_i	n_i=25	n_i=50	n_i=25	n_i=50	n_i=25	n_i=50	
1	0.999	0.999	0.999	0.965	0.998	0.934	D
	1.000	0.999	1.000	0.972	0.943	0.937	JP
5	0.933	0.940	0.916	0.937	0.901	0.941	D
	0.972	0.962	0.956	**0.951**	**0.954**	**0.948**	JP
10	0.943	**0.946**	0.935	0.941	0.940	0.943	D
	0.963	0.957	**0.949**	**0.952**	**0.947**	**0.951**	JP
15	**0.948**	**0.949**	0.943	**0.948**	**0.946**	**0.949**	D
	0.954	0.959	**0.954**	**0.950**	**0.951**	**0.951**	JP
	p=(.01,.02,.03,.04,.05)		p=(.01,.03,.05,.07,.09)		p=(.01,.04,.07,.10,.13)		
s_i	n_i=25	n_i=50	n_i=25	n_i=50	n_i=25	n_i=50	
1	1.000	0.999	0.999	0.917	0.964	0.911	D
	1.000	0.999	0.999	0.936	0.999	0.909	JP
5	0.922	0.933	0.921	0.932	0.884	0.933	D
	0.972	0.956	0.943	0.943	0.945	0.943	JP
10	0.940	0.944	0.929	0.939	0.933	0.934	D
	0.958	0.955	**0.948**	**0.950**	**0.947**	**0.949**	JP
15	0.943	**0.949**	**0.946**	0.944	0.945	**0.950**	D
	0.960	0.958	**0.953**	**0.948**	0.960	**0.947**	JP

Table 2: Percentage of times the adjustment for $\hat{\theta}_i = 0$ was made during the simulations for Dunnett (D) and Jeffreys-Perks (JP).

	p=(.01,.02,.03)		p=(.01,.03,.05)		p=(.01,.04,.07)		
s_i	n_i=25	n_i=50	n_i=25	n_i=50	n_i=25	n_i=50	
1	0.9559	0.8062	0.9158	0.7063	0.8837	0.6706	D
	0.9527	0.8062	0.9118	0.7123	0.8862	0.6693	JP
5	0.3598	0.0882	0.3009	0.0768	0.2911	0.0773	D
	0.3507	0.0906	0.2996	0.0804	0.2870	0.0801	JP
10	0.0863	0.0070	0.0811	0.0070	0.0805	0.0068	D
	0.0877	0.0074	0.0856	0.0071	0.0826	0.0063	JP
15	0.0239	0.0003	0.0232	0.0005	0.0244	0.0004	D
	0.0239	0.0004	0.0244	0.0003	0.0226	0.0004	JP
	p=(.01,.02,.03,.04)		p=(.01,.03,.05,.07)		p=(.01,.04,.07,.10)		
s_i	n_i=25	n_i=50	n_i=25	n_i=50	n_i=25	n_i=50	
1	0.9685	0.8320	0.9320	0.7282	0.8885	0.6713	D
	0.9709	0.8358	0.9283	0.7184	0.8883	0.6621	JP
5	0.3622	0.0857	0.2976	0.0822	0.2841	0.0779	D
	0.3523	0.0895	0.2990	0.0835	0.2809	0.0821	JP
10	0.0860	0.0068	0.0835	0.0050	0.0763	0.0080	D
	0.0891	0.0066	0.0804	0.0057	0.0801	0.0077	JP
15	0.0230	0.0004	0.0212	0.0003	0.0250	0.0005	D
	0.0231	0.0006	0.0227	0.0008	0.0251	0.0005	JP
	p=(.01,.02,.03,.04,.05)		p=(.01,.03,.05,.07,.09)		p=(.01,.04,.07,.10,.13)		
s_i	n_i=25	n_i=50	n_i=25	n_i=50	n_i=25	n_i=50	
1	0.9793	0.8394	0.9352	0.7216	0.8927	0.6678	D
	0.9774	0.8441	0.9354	0.7291	0.8899	0.6682	JP
5	0.3580	0.0886	0.3057	0.0842	0.2891	0.0796	D
	0.3602	0.0860	0.2992	0.0825	0.2896	0.0844	JP
10	0.0902	0.0060	0.0822	0.0063	0.0812	0.0065	D
	0.0907	0.0054	0.0774	0.0053	0.0811	0.0055	JP
15	0.0207	0.0004	0.0219	0.0001	0.0248	0.0005	D
	0.0214	0.0003	0.0234	0.0008	0.0216	0.0007	JP

Table 3: Percentage of times the adjustment for $\hat{\theta}_i = 1$ was made during the simulations for Dunnett (D) and Jeffreys-Perks (JP).

	p=(.01,.02,.03)		p=(.01,.03,.05)		p=(.01,.04,.07)		
s_i	n_i=25	n_i=50	n_i=25	n_i=50	n_i=25	n_i=50	
1	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	D
	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	JP
5	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	D
	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	JP
10	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	D
	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	JP
15	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	D
	0.0000	0.0000	0.0000	0.0000	0.0001	0.0000	JP
	p=(.01,.02,.03,.04)		p=(.01,.03,.05,.07)		p=(.01,.04,.07,.10)		
s_i	n_i=25	n_i=50	n_i=25	n_i=50	n_i=25	n_i=50	
1	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	D
	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	JP
5	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	D
	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	JP
10	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	D
	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	JP
15	0.0000	0.0000	0.0002	0.0000	0.0033	0.0001	D
	0.0000	0.0000	0.0000	0.0000	0.0037	0.0000	JP
	p=(.01,.02,.03,.04,.05)		p=(.01,.03,.05,.07,.09)		p=(.01,.04,.07,.10,.13)		
s_i	n_i=25	n_i=50	n_i=25	n_i=50	n_i=25	n_i=50	
1	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	D
	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	JP
5	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	D
	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	JP
10	0.0000	0.0000	0.0000	0.0000	0.0009	0.0000	D
	0.0000	0.0000	0.0000	0.0000	0.0012	0.0000	JP
15	0.0000	0.0000	0.0010	0.0000	0.0404	0.0012	D
	0.0000	0.0000	0.0010	0.0000	0.0410	0.0019	JP

Table 4: Bawden and Kassanis (1946). Observed proportion of infected plants with the potato virus Y for different varieties.

*Using two aphids to colonize the healthy plants

Test plant	Proportion of infected plants
Arran Banner	(3/36)=0.083
Majestic	(5/36)=0.139
Arran Consul	(10/36)=0.278
Arran Pilot	(12/36)=0.333
May Queen	(15/36)=0.417
Tobacco	(25/36)=0.694

*Using four aphids to colonize the healthy plants

Test plant	Proportion of infected plants
Arran Banner	(10/36)=0.278
Majestic	(17/36)=0.472
Arran Consul	(19/36)=0.528
Arran Pilot	(25/36)=0.694
May Queen	(21/36)=0.583
Tobacco	(36/36)=1.000

Table 5: Analysis of the multiple-vector transfer design study of Bawden & Kassanis (1946). Dunnett and Jeffreys-Perks 95% simultaneous confidence intervals for p_k-p_i, where p_k=proportion of tobacco infected plants and p_i=proportion of infected plants for the diverse varieties of potato.

*Two aphids were used to colonize each test plant.

Difference between	Point Estimate	Confidence Intervals	
varieties		Dunnett	Jeffreys-Perks
Tobacco - Arran Banner	0.4044	(0.2278 , 0.5815)	(0.2204 , 0.5729)
Tobacco - Majestic	0.3747	(0.1921 , 0.5583)	(0.1856 , 0.5500)
Tobacco - Arran Consul	0.2965	(0.0993 , 0.4950)	(0.0948 , 0.4876)
Tobacco - Arran Pilot	0.2635	(0.0604 , 0.4670)	(0.0567 , 0.4604)
Tobacco - May Queen	0.2103	(-0.0004 , 0.4224)	(-0.0028 , 0.4165)

Table 6: Analysis of the multiple-vector transfer design study of Bawden & Kassanis (1946). Dunnett and Jeffreys-Perks 95% simultaneous confidence intervals for p_k-p_i, where p_k=proportion of tobacco infected plants and p_i=proportion of infected plants for the diverse varieties of potato.

*Four aphids were used to colonize each test plant.

Difference between	Point Estimate	Confidence Intervals	
varieties		Dunnett	Jeffreys-Perks
Tobacco - Arran Banner	0.5786	(0.3018 , 0.8553)	(0.3422 , 1.0000)
Tobacco - Majestic	0.5090	(0.2272 , 0.7909)	(0.2658 , 1.0000)
Tobacco - Arran Consul	0.4857	(0.2020 , 0.7694)	(0.2397 , 1.0000)
Tobacco - Arran Pilot	0.4002	(0.1090 , 0.6914)	(0.1421 , 1.0000)
Tobacco - May Queen	0.4601	(0.1743 , 0.7459)	(0.2110 , 1.0000)

APPENDIX A: The maximum likelihood estimator of $\hat{p}_i$

Recall that $\theta_i = 1-(1-p_i)^{s_i}$ is the probability that a group of size s_i subjects has at least one subject possessing the trait of interest. The usual binomial MLE of θ_i is estimated with the following procedure.

The likelihood is given by:

$$L(\theta_i) = \frac{n!}{(n-x_i)!x_i!}\theta_i^{x_i}(1-\theta_i)^{n_i-x_i};$$

where $X_i = \sum_{l=1}^{n_i} Y_{il}$ with Y_{il} defined in Chapter 2. Thus, the natural logarithm of the likelihood is given by

$$\ln L(\theta_i) = \ln\left(\frac{n!}{(n-x_i)!x_i!}\right) + x_i \ln(\theta_i) + (n_i - x_i)\ln(1-\theta_i),$$

and the first derivative of the $\ln L(\theta_i)$ is given by

$$\frac{d\ln L(\theta_i)}{d\theta_i} = \frac{x_i}{\theta_i} - \frac{(n_i - x_i)}{(1-\theta_i)}.$$

To find the MLE $\hat{\theta}_i$, we set the first derivative of $\ln L(\theta_i)$ to zero and solve for $\hat{\theta}_i$

$$\frac{x_i}{\hat{\theta}_i} = \frac{(n_i - x_i)}{(1-\hat{\theta}_i)} \Rightarrow \frac{(1-\hat{\theta}_i)}{\hat{\theta}_i} = \frac{(n_i - x_i)}{x_i} \Rightarrow \frac{1}{\hat{\theta}_i} = \frac{n_i}{x_i} \Rightarrow \hat{\theta}_i = \frac{x_i}{n_i} = \bar{x}_i.$$

To ensure that we have a maximum, we need to find the second derivative of $\ln L(\theta_i)$, evaluate this for $\theta_i = \hat{\theta}_i$, and verify that the result is negative. Here

$$\frac{d^2\ln L(\theta_i)}{d\theta_i^2} = -\frac{x_i}{\theta_i^2} - \frac{(n_i - x_i)}{(1-\theta_i)^2},$$

and $\left.\frac{d^2\ln L(\theta_i)}{d\theta_i^2}\right|_{\theta_i=\hat{\theta}_i} = -\frac{x_i}{\bar{x}} - \frac{(n_i-x_i)}{(1-\bar{x})^2} < 0$ since $x_i \geq 0$ and $n_i \geq x_i$; hence, $\hat{\theta}_i = \frac{x_i}{n_i}$ is a maximum.

Thus, by the invariance properties of maximum likelihood estimators, the MLE of $p_i = 1-(1-\theta_i)^{1/s_i}$ is $\hat{p}_i = 1-(1-\hat{\theta}_i)^{1/s_i}$ where $\hat{\theta}_i = \frac{x_i}{n_i}$.

APPENDIX B: The distribution of $\hat{p}_i$

As the binomial likelihood is a member of the regular exponential class (REC) the regularity conditions regarding the limiting distribution of the MLE's are satisfied, and the distribution of $\hat{\theta}_i$ is asymptotically normal with asymptotic mean θ_i and asymptotic variance given by $\dfrac{1}{-E\left[\dfrac{d^2}{d\theta_i^2}\ln L(\theta_i)\right]}$.

To calculate the asymptotic variance we first consider the denominator, where the likelihood is given by

$$L(\theta_i) = \frac{n!}{(n-x_i)!x_i!}\theta_i^{x_i}(1-\theta_i)^{n_i-x_i};$$

and the natural logarithm of the likelihood is

$$\ln L(\theta_i) = \ln\left(\frac{n!}{(n-x_i)!x_i!}\right) + x_i\ln(\theta_i) + (n_i - x_i)\ln(1-\theta_i) \quad .$$

The first and second derivatives of $\ln L(\theta_i)$ are given by the following equations:

$$\frac{d\ln L(\theta_i)}{d\theta_i} = \frac{x_i}{\theta_i} - \frac{(n_i - x_i)}{(1-\theta_i)}, \text{ and } \frac{d^2\ln L(\theta_i)}{d\theta_i^2} = -\frac{x_i}{\theta_i^2} - \frac{(n_i - x_i)}{(1-\theta_i)^2}.$$

The expected value of the second derivative of $\ln L(\theta_i)$ is given by

$$E\left[\frac{d^2\ln L(\theta_i)}{d\theta_i^2}\right] = -\frac{E[x_i]}{\theta_i^2} - \frac{E[n_i - x_i]}{(1-\theta_i)^2}$$

$$= -\frac{n_i\theta_i}{\theta_i^2} - \frac{n_i}{(1-\theta_i)^2} + \frac{n_i\theta_i}{(1-\theta_i)^2} \quad (\text{since } E[x_i] = n_i\theta_i)$$

$$= -\frac{n_i\theta_i}{\theta_i^2} - \frac{n_i(1-\theta_i)}{(1-\theta_i)^2} = -\frac{n_i}{\theta_i} - \frac{n_i}{(1-\theta_i)} = -\frac{n_i}{\theta_i(1-\theta_i)}$$

after multiplying by -1 the denominator of the asymptotic variance becomes

$$-E\left[\frac{d^2\ln L(\theta_i)}{d\theta_i^2}\right] = \frac{n_i}{\theta_i(1-\theta_i)}.$$

Consequently, the asymptotic variance of $\hat{\theta}_i$ is $\dfrac{1}{-E\left[\dfrac{d^2}{d\theta_i^2}\ln L(\theta_i)\right]} = \dfrac{\theta_i(1-\theta_i)}{n_i}$

Now, the delta-method states:

"If $\sqrt{n}(T_n - \theta) \to N(0, \sigma^2)$, then $\sqrt{n}\{h(T_n) - h(\theta)\} \to N(0, \sigma^2\{h'(\theta)\}^2)$, provided $h'(\theta)$ exists and it is not zero".[1]

Recall that $\hat{p}_i = g(\hat{\theta}_i) = 1 - (1 - \hat{\theta}_i)^{1/s_i}$, and since above we derived the asymptotic normal distribution of $\hat{\theta}_i$, then applying the delta-method we can derive the distribution of $\hat{p}_i$ as follows:

$$h'(\hat{\theta}_i) = \frac{1}{s_i}(1 - \theta_i)^{1/s_i - 1}$$

$$\left[h'(\hat{\theta}_i)\right]^2 = \frac{1}{s_i^2}(1 - \theta_i)^{2/s_i - 2}$$

$$\sigma^2\left[h'(\hat{\theta}_i)\right]^2 = \frac{\theta_i(1 - \theta_i)}{n_i}\frac{1}{s_i^2}(1 - \theta_i)^{2/s_i - 2}$$

$$= \frac{\theta_i(1 - \theta_i)(1 - \theta_i)^{2/s_i - 2}}{n_i s_i^2}$$

$$= \frac{(1 - (1 - p_i)^{s_i})(1 - p_i)^{s_i}((1 - p_i)^{s_i})^{2/s_i - 2}}{n_i s_i^2}, \text{ (since } \theta_i = 1 - (1 - p_i)^{s_i})$$

$$= \frac{(1 - (1 - p_i)^{s_i})(1 - p_i)^{s_i}(1 - p_i)^{2 - 2s_i}}{n_i s_i^2}$$

$$= \frac{s_i^{-2}(1 - (1 - p_i)^{s_i})(1 - p_i)^{2 - s_i}}{n_i}$$

Thus $\hat{p}_i$ also follows an asymptotic normal distribution with mean p_i and variance $v_i = v(p_i, s_i) = \frac{s_i^{-2}(1 - (1 - p_i)^{s_i})(1 - p_i)^{2 - s_i}}{n_i}$.

[1] Lehmann, E. & Casella, G. (1998)

APPENDIX C: The estimated asymptotic variance of $\hat{b} = \hat{p}_i - \hat{p}_j$

In appendix B, we showed that $\hat{p}_i$ has an asymptotic normal distribution with mean p_i and variance $v_i = [s_i^{-2}(1-(1-p_i)^{s_i})(1-p_i)^{2-s_i}]/n_i$. Since $\hat{b} = \hat{p}_i - \hat{p}_j$ is a linear combination of variables with independent asymptotic normal distributions; then its distribution is also asymptotic normal with mean $p_i - p_j$ and variance

$$v_i + v_j = \frac{(1-(1-p_i)^{s_i})(1-p_i)^{2-s_i}}{s_i^2 n_i} + \frac{(1-(1-p_j)^{s_j})(1-p_j)^{2-s_j}}{s_j^2 n_j}.$$

Recall that $a = p_i + p_j$ and $b = p_i - p_j$, and note that

$$\frac{2-a-b}{2} = 1 - \frac{p_i}{2} - \frac{p_j}{2} - \frac{p_i}{2} + \frac{p_j}{2} = 1 - p_i, \text{ and also that}$$

$$\frac{2-a+b}{2} = 1 - \frac{p_i}{2} - \frac{p_j}{2} + \frac{p_i}{2} - \frac{p_j}{2} = 1 - p_j.$$

Now, using $\tilde{a}$, the posterior mean of a derived in Chapter 2, as an estimate of a and letting $g_1(\tilde{a}, b) = \frac{(2-\tilde{a}-b)}{2}$ and $g_2(\tilde{a}, b) = \frac{(2-\tilde{a}+b)}{2}$, we can rewrite the asymptotic variance of $\hat{b} = \hat{p}_i - \hat{p}_j$ as:

$$V(\hat{b}; a = \tilde{a}) = \frac{\{1-g_1(\tilde{a},b)^{s_i}\}g_1(\tilde{a},b)^{2-s_i}}{s_i^2 n_i} + \frac{\{1-g_2(\tilde{a},b)^{s_j}\}g_2(\tilde{a},b)^{2-s_j}}{s_j^2 n_j}.$$

VITA

Yanina Grant

Candidate for the Degree of

Master of Science

Thesis: MULTIPLE COMPARISONS WITH A CONTROL UNDER A GROUP TESTING SCENARIO

Major Field: Statistics

Biographical:

Personal Data: Born in Torreón, Coahuila, Mexico.

Education: Received Bachelor of Science degree in Economics with a minor in Finance from the Technical Institute of Advanced Studies of Monterrey, Monterrey, Nuevo León, Mexico, in May 2000. Completed the requirements for the Master of Science degree with a major in Statistics at Oklahoma State University, Department of Statistics, in May 2006.

Experience: Employed as Marketing Analyst by the Department of Marketing, Villacero Group, Monterrey, Nuevo León, Mexico, from January 2001 to April 2002. Interned at the Neuropharmacology Department, Scripps Research Institute, from June 2005 to August 2005.

Professional Memberships: American Statistical Association, Institute of Mathematical Statistics.

Name: Yanina Grant Date of Degree: May, 2006

Institution: Oklahoma State University Location: Stillwater, Oklahoma

Title of Study: MULTIPLE COMPARISONS WITH A CONTROL UNDER A GROUP TESTING SCENARIO

Pages in Study: 41 Candidate for the Degree of Master of Science

Major Field: Statistics

Scope and Method of Study: The objective of this project is to derive two simultaneous inference approaches, Dunnett and Jeffreys-Perks, to quantify proportion differences between a control and several non-control levels under a group testing scenario. Both procedures were evaluated in terms of estimated coverage probability via simulations. Ten thousand simulations were run for diverse data scenarios and the coverage probability was estimated by calculating the proportion of times the confidence intervals simultaneously contained the real proportion difference.

Findings and Conclusions: It was found that when using large group sizes of about 15 or more observations both the Dunnett and Jeffreys-Perks approaches perform fairly well. For small sample sizes, between 5 and 10 observations, the Jeffreys-Perks intervals are recommended as their coverage is closer to the nominal 1-α level, and it also is a conservative approach while the Dunnett procedure is anticonservative. We also found that group testing provides more accurate results than individual testing since more coverage probabilities are closer to the nominal when using group sizes greater than one.

ADVISOR'S APPROVAL: Dr. Melinda McCann.