2016-01-01

Sample Size Estimation for Genomics Experiments with Dependent End Points

Desmond Koomson
University of Texas at El Paso, dkoomson05@gmail.com

SAMPLE SIZE ESTIMATION FOR GENOMICS EXPERIMENTS WITH DEPENDENT END POINTS

DESMOND KOOMSON

Master's Program in Mathematical Sciences

APPROVED:

Amy E. Wagler, Ph.D., Chair

Leung Ming-Ying, Ph.D.

Germán Rosas-Acosta, Ph.D.

Charles Ambler, Ph.D.
Dean of the Graduate School

to my

FAMILY

with love

SAMPLE SIZE ESTIMATION FOR GENOMICS EXPERIMENTS WITH DEPENDENT END POINTS

by

DESMOND KOOMSON

THESIS

Presented to the Faculty of the Graduate School of

The University of Texas at El Paso

in Partial Fulfillment

of the Requirements

for the Degree of

MASTER OF SCIENCE

Master's Program in Mathematical Sciences

THE UNIVERSITY OF TEXAS AT EL PASO

August 2016

Acknowledgements

Blessed be the name of the LORD, who has not turned away my prayer, nor his mercies from me. I am overwhelmed by what He has done and grateful for His abundant grace, bringing what he begun to a successful end. I am also very grateful to my supervisor and mentor Dr. Amy Wagler, whose guidance and supervision made this work a success. You are like a mother to me, always encouraging and guiding me through my studies.

I wish to thank the other members of my committee, Dr. Ming-Ying Leung, Director, Computational Science and Dr. Germán Rosas-Acosta, Department of Biological Sciences, all at The University of Texas at El Paso. Your guidance and assistance were greatly valuable to the completion of this work.

I also wish to thank Dr. Joan Staniswalis, Dr. Panagis Moschopoulos, Dr. Naijun Sha, Dr. Xiaogang Su, Dr. Ori Rosen, and Dr. Behzad Djafari-Rouhani from the Mathematical Sciences Department at The University of Texas at El Paso who taught and mentored me as well through my studies. Additionally, I wish to thank all professors and staff of the Mathematical Sciences Department here in The University of Texas at El Paso for all their support, enabling me to complete my degree.

I appreciate my wonderful family as well, both here in the United States and Ghana. Your love and prayers has made this journey worthwhile traveled. To all my other well-wishers whose names are not mentioned, I say "God richly bless you".

Abstract

In typical genomics studies involving numerous association tests of gene mutations with a disease, error rate control via multiplicity adjustment is paramount because even if all genes were to be non-differentially associated, we would still make some false positives. Many methods exist that incorporate the control of multiplicity for normally distributed endpoints in sample size estimation, but none addresses the issue for non-normally correlated endpoints. One common practice in the literature is to assume an equal correlation among all differentially associated or expressed genes, thereby using the generalized binomial or beta-binomial model to compute the comparison-wise power of detecting these genes.

We present a fast and simple novel approach for estimating sample size which focuses on controlling the family-wise error rate using Hunter and Worsley's method for normally, t and chi-square distributed endpoints of any correlation structure. The sample size needed are computed using either a two-sample z-test or chi-square test formula depending on whether the response variable is continuous or binary at the desired comparison-wise power (using the binomial model), adjusted family-wise error rate and standardized effect size. These modifications would provide sample size estimates that are close to their exact values under more general correlation structures, where the generalized binomial or beta-binomial model may fail to perform.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Brief Insight

The simultaneous evaluation of several hypotheses in experiments gives rise to the need for multiplicity adjustments. When left unattended, ignoring multiplicity of errors can seriously undermine the reliability of any statistical inference being made, resulting in uncorroborated claims. Hence, the performance of simultaneous inferences or multiple comparisons comes into play when we make or construct several related tests or interval estimates at the same time [1].

When you perform a large number of statistical tests, some will have p-values being significant purely by chance, even if all the null hypotheses are really true. Hence, anytime we reject multiple null hypotheses because a p-value is less than a critical value, it's highly possible that we're wrong in our decision and making false positive conclusions. Besides our vital objective with multiple comparisons is to reduce the number of false positives (i.e. falsely rejected true null hypotheses).

The issue of multiple comparisons is one of how to define and control error rates. Each of the individual tests or confidence intervals (CIs) has a Type 1 error rate, say ε_i that can be controlled by an experimenter. By considering these individual tests or CIs together as a family of tests or intervals, we can also compute a combined Type 1 error rate for this family.

To illustrate use of multiple comparison procedures consider the following scenarios as follows;

1. Suppose a new way of teaching high school students on improving their writing skills is claimed to be more efficient than an existing standard way of teaching writing. Students in the two (2) groups can be compared in terms of their grammar, spelling, organization, content, and other attributes. Now assuming a recent study into this claim publishes the table of p-values below;

Table 1.1: Writing attributes with their p-value

Control Group	**Treatment Group**	**p-values from mean differences**
Grammar	Grammar	0.01802
Spelling	Spelling	0.22048
Organization	Organization	0.03200
Content	Content	0.00875

   - At an $\varepsilon_i = 0.05$ significant level, three (3) of the test results are each significant.
   - As more and more attributes get compared, the likelihood of the treatment and control groups differing on at least one attribute due to chance increases.
   - Note that each of these test was controlled at an $\varepsilon_i = 0.05$ level. This is the individual error rate or comparison wise error rate (CWER).

2. García-Arenzana et al. (2014) [2] tested associations of 25 dietary variables with mammographic density, an important risk factor for breast cancer, in Spanish women. The following results were found:

Table 1.2: Dietary variables with their p-values

Dietary Variable	**p-value**	**Dietary Variable**	**p-value**
Total calories	0.001	Eggs	0.275
Olive oil	0.008	Blue fish	0.34
Whole milk	0.039	Legumes	0.341
White meat	0.041	Carbohydrates	0.384
Proteins	0.042	Potatoes	0.569
Nuts	0.06	Bread	0.594
Cereals and pasta	0.074	Fats	0.696
White fish	0.205	Sweets	0.762
Butter	0.212	Dairy products	0.94
Vegetables	0.216	Semi-skimmed milk	0.942
Skimmed milk	0.222	Total meat	0.975
Red meat	0.251	Processed meat	0.986
Fruit	0.269		

- Five of the variables show significance (p-value<0.05).
- Again, because 25 dietary variables were tested each at an individual error rate of 0.05, you'd expect one or two variables to show a significant result purely by chance, even if diet had no effect on mammographic density.
- What happens if we assume the tests are independent?
  Let P_r(Reject each $H_0|$ each H_0 true) $= 0.05 \sim$ CWER.
  $\implies P_r$(Fail to Reject each $H_0|$ each H_0 true) $= 0.95$
  Hence for the 25 dietary test;
  $\implies P_r$(Fail to Reject all $H_0 s|$ all $H_0 s$ true) $= 0.95^{25}$
  $\implies P_r$(Rejecting at least one of 25 $H_0 s|$ all $H_0 s$ true) $= 1 - 0.95^{25} = 0.7226$

- This means that for the entire set of 25 independent tests, there is a 72.3% chance of getting at least one true null hypothesis falsely rejected. Also, increasing the number of tests only increases the error rate for the entire family, hence the chance of making at least one false rejection becomes almost certain.
- This calculated overall (combined) rate is called the experiment wise error rate (EER) or familywise error rate (FWER).

3. Also in Microarray experiments, DNA technology provides tools for studying the associations between genes and a disease simultaneously. Here, association studies are very important for uncovering disease causes in populations. As an example, consider a $(p \times n)$ dataset of $p = 1061$ genes by $n = 30$ (20 patients and 10 control groups).

These and many other situations are what motivates multiplicity adjustment, i.e. How to control various error rates when performing loads of test.

1.2 Errors & Error Rates

1.2.1 Some Notations

Let m=$\{H_{01}$,H_{02},. . . ,$H_{0m}\}$ be a set of null hypotheses to be tested, their combined or overall null hypotheses H_0 can also be defined as;

$$H_0 = \{H_{01} \cap H_{02} \cap H_{03} \cap \cdots \cap H_{0m}\}$$

- H_0 is true if and only if all $H_{0i}s$ are true and false if any of the $H_{0i}s$ is rejected.
- ε_i and ε are the Type 1 error rates for the *ith* test and combined tests respectively.

For instance in scenario 2; $m = 25$, H_{0i} is the null hypothesis that dietary i has no effect on mammographic density and H_0 is the null hypotheses that diet has no effect on mammographic density.

A Single Hypothesis

Consider the table of outcomes below for a single hypothesis test;

Table 1.3: Possible outcomes for a single hypothesis

Decision	**Reality**	
	True Null	**False Null**
Reject Null	False Positive (Type 1 Error)	True Positive
Fail to Reject Null	True Negative	False Negative (Type 2 Error)

In general classical literature, we control the probability of making a Type 1 error (ε), and among those procedures that control ε choose one that makes less Type 2 Error (β).

Multiple Hypotheses

Since the earlier scenarios presented above involves multiple hypotheses, we consider another version of Table 1.3 as shown below;

Table 1.4: Possible Decisions (Outcomes) for m hypotheses

Reality	**Decision**		
	Reject Null	**Fail to Reject Null**	Total
True Null	V(ε)	S($1-\varepsilon$)	m_0
False Null	U($1-\beta$)	T(β)	m_1
Total	R	R^c	m

- For "m" hypotheses, we have $m = V + S + U + T$. In practice, these counts are unknown but can be worked with theoretically.
- m_0 is the number of true nulls out of the set of hypotheses (unknown).
- m_1 is the number of false nulls out of the set of hypotheses (unknown).
- R is the number of rejected nulls out of the set of hypotheses (known).
- R^c is the number of non-rejected nulls out of the set of hypotheses (known).
- $\frac{V}{m_0}$ is the false positive fraction, $\frac{U}{m_1}$ is the sensitivity fraction, $\frac{U}{R}$ is the true discovery fraction, $\frac{V}{R}$ is the false discovery fraction and $\frac{U+S}{m}$ is the accuracy fraction.
- Wang and Chen (2004) proposed the use of sensitivity measure u/m_1 for gene selection.

There are several ways of defining a combined Type 1 error rate (ε) for a family tests. Because of this variety, many get confused as to which error rate to use since different error rates differ in their multiplicity procedures. Definitions of these error rates depend on the numbers or fractions of falsely rejected null hypotheses, which will never be known in practice.

1.2.2 Comparison-wise Error Rate (CWER)

This is the probability of rejecting any particular H_{0i} in a single test when that H_{0i} is true. Here, ε is controlled for each individual test ignoring all of the other tests, i.e.;

$$P_r(\text{Reject each } H_{0i} | H_{0i} \text{ true}) \leq \varepsilon$$

This error rate simply ignores multiple testing correction. Scenario 2 above controlled the comparison-wise error rate at a 5% level.

1.2.3 Familywise Error Rate (FWER)

This is the probability of rejecting a least one of the $H_{0i}s$ (hence rejecting H_0) when all $H_{0i}s$ are true. Here, the expected fraction of tests in which we reject at least one of the $H_{0i}s$ when H_0 is true is controlled at ε for all tests. FWER control is;

$$P_r(\text{Reject at least one } H_{0i}|H_0 \text{ true}) \leq \varepsilon$$

Since we usually assume all $H_{0i}s$ to be true in H_0, then $T = U = 0$ in Table 1.4 and FWER control can be written as;

$$P_r(V > 0|H_0 \text{ true}) \leq \varepsilon$$

FWER also controls CWER at no more than ε. Again in Scenario 2, the FWER is the fraction of times we would have declared at least one dietary as having an effect on mammographic density when in fact all dietary have no effect.

1.2.4 False Discovery Rate (FDR)

In real life, we cannot always assume that H_0 is true, since some of the $H_{0i}s$ will be false actually. This is where the FDR comes into play, briefly mentioned by Simes (1986). Developed in detail by Benjamini and Hochberg (1995), the FDR allows for the possibility of some $H_{0i}s$ being false. Now in reference to Table 1.4;

- Let $\dfrac{V}{R}$ be the proportion of rejected $H_{0i}s$ that are actually true. This is also known as the false discovery fraction, where; $\frac{V}{R} = 0$ when $R = 0$.
- Controlling FDR means making sure that; $E\left(\frac{V}{R}\right) \leq \varepsilon$ i.e. the expected fraction of false rejections(discoveries) is at most ε.
- Assuming all $H_{0i}s$ are true (i.e. H_0 true), then all discoveries (rejections) are false and FDR is just the FWER. So FDR also controls the FWER at ε.

Lastly, the more correct rejections you make, the more false rejections FDR lets you make but this ratio is limited.

1.2.5 Strong Familywise Error Rate (SFWER)

This is the probability of making any false rejections, i.e. $P_r\left(\frac{V}{R} > 0\right)$. It also allows the possibility of some $H_{0i}s$ being false under the null hypotheses, but doesn't take into consideration the number of correct rejections made unlike the FDR. Hence, a true rejection cannot make a false rejection more likely as seen under FDR. In reference to Table 1.4, SFWER control is given as;

$$P_r(\text{Rejecting any } H_{0i}| \text{ some } H_{0i}s \text{ are true}) \leq \varepsilon$$
$$\implies P_r(V > 0| \text{ some } H_{0i}s \text{ are true}) \leq \varepsilon$$

Therefore "ε" is controlled for every subset of H_0. In Scenario 2, assuming five of the dietaries had an effect with mammographic density, then the probability of declaring at least one of the other 20 dietaries as having an effect with mammographic density would be no more than a SFWER control at 5%.

1.3 Outline of the thesis

We will organize the remaining parts of the thesis in this manner. Chapters 2 and 3 provides a review of literature on some multiple comparison procedures and existing models for determining the comparison-wise power in multiple hypotheses settings. We also outlined the sample size formula for one and two samples z-tests, and quote that for chi-square tests. In Chapter 4, our proposed sample size methodology is presented and explained in detail. We will then use simulations to analyze and compare the performance of the proposed method to existing methods in the literature. Finally, we will present and discuss the results obtained from the simulation study and provide areas for future work in Chapters 5 and 6.

Chapter 2

Multiple Comparison Procedures (MCPs)

In this chapter we provide few methods that seek to control the various error rates introduced in the previous chapter. We would like to state that an experimenter's knowledge of the type of error rate he/she wishes to control helps in the selection of any of these MCPs to control such error. These methods are all Bonferroni based.

2.1 Ordinary Bonferroni

This is the easiest and most widely applicable MCP. It works for "m" set of null hypotheses. Now considering two sets of null hypotheses, the idea behind this technique follows as such;

$$P_r(H_{01}) = P_r(H_{02}) = \text{probability of Type 1 Error} = \varepsilon$$
$$P_r(H_{01}^c \cap H_{02}^c) = 1 - P_r(H_{01} \cup H_{02}) = 1 - P_r(H_{01}) - P_r(H_{02}) + P_r(H_{01} \cap H_{02})$$
$$P_r(H_{01}^c \cap H_{02}^c) \geq 1 - P_r(H_{01}) - P_r(H_{02}) = 1 - 2\varepsilon$$
$$P_r(\text{at least one Type 1 Error}) \leq 1 - (1 - 2\varepsilon) = 2\varepsilon$$

Hence by making $P_r(H_{01}) = P_r(H_{02}) = \frac{\varepsilon}{2}$, the probability of making at least one Type 1 error becomes at most ε, i.e.;

$$P_r(\text{at least one Type 1 Error}) \leq \varepsilon$$

Now if P_i is the p-value for testing H_{0i}, then by Bonferroni we reject H_{0i} (and thus H_0); if $P_i < \frac{\varepsilon}{m}$, i.e. each null hypothesis is evaluated at level $\frac{\varepsilon}{m}$. Sometimes an adjusted version is

also preferred, given as; Bonferroni p-value adjusted $= m * P_i < \varepsilon$.

Also to obtain simultaneous $(1 - \varepsilon)$ confidence intervals, Bonferroni says construct each individual interval with coverage $(1 - \frac{\varepsilon}{m})$. The test or (and) intervals need not be independent or related in anyway. The procedure provides a SFWER control and confidence intervals that are simultaneous.

2.2 Holm's Sequential Bonferroni

This procedure provides a simple modification to the ordinary Bonferroni method whiles controlling the SFWER, but produces no simultaneous confidence intervals (Holm, 1979). The technique follows as such;

1. First sort the p-values of the m test hypotheses in an ascending order, i.e. $P_{(1)}, P_{(2)}, \ldots, P_{(m)}$ with their associated null hypotheses, i.e. $H_{0(i)}$ for $i = 1, 2, \ldots, s$. The smallest p-value has a rank of $j = 1$, then next smallest has $j = 2$, etc.

2. Starting from the smallest p-value say $P_{(1)}$, reject $H_{0(i)}$ if;

$$P_{(j)} \leq \frac{\varepsilon}{m - j + 1}, \ldots \forall j = 1, 2, \ldots, i$$

   Otherwise, we fail to reject $H_{0(i)}$ and any other thereafter if the first non-significant p-value is reached. At this point we stop.

More power is achieved as compared to the ordinary Bonferroni since only the smallest p-value gets compared to $\frac{\varepsilon}{m}$. Sometimes an adjusted version is also preferred, given as; Holm p-value adjusted $= (m - j + 1) * P_{(j)} < \varepsilon$, where j is the rank of the p-value.

2.3 FDR Control Method

Having defined the false discovery rate as $E\left(\frac{V}{R}\right)$ for $R > 0$ in 1995, Benjamini and Hochberg (2000) proposed a control procedure for FDR as follows;

1. First sort the p-values of the m test hypotheses in an ascending order, i.e. $P_{(1)}, P_{(2)}, \ldots, P_{(m)}$ with their associated null hypotheses, i.e. $H_{0(i)}$ for $i = 1, 2, \ldots, s$. The smallest p-value has a rank of $j = 1$, then next smallest has $j = 2$, etc.

2. With a desired FDR level (ε), proceed with the largest p-value and work down. Starting from the largest p-value say $P_{(j=m)}$, we reject $H_{0(i)}$ and any other(s) having p-value(s) less than the Benjamini-Hochberg critical value, i.e.;

$$P_{(j)} \leq \frac{j * \varepsilon}{m}, \ldots for some \;\; j \geq i$$

   Otherwise, we fail to reject $H_{0(i)}$ and proceed through.

The procedure works well when the tests are statistically independent and controls FDR, but not SFWER. Also, an adjusted version is sometimes preferred given as; FDR p-value adjusted $= \frac{m * P_{(j)}}{j} < \varepsilon$, where j is the rank of the p-value.

Chapter 3

Sample Size Estimation For Testing Many Hypotheses

3.1 Sample Size Formulas

We will consider sample size formulas for both continuous (i.e. differentially expressed genes) and discrete (i.e. independent binary outcomes) response random variables.

3.1.1 Continuous Outcomes

Under a single hypothesis, the sample size problem is usually formulated in order to ensure the power $(1-\beta)$ of detecting any mean difference (standardized effect size) δ^* at a pre-specified Type 1 error rate ε (CWER). Assuming the hypothesis $H_0 : \mu = \mu_0 \ \ vrs \ \ H_1 : \mu \neq \mu_0$, we define the RR and AR as follows:

$$\text{RR (Rejection Region)} : \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} > Z_{1-\varepsilon/2} \quad or \quad \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} < Z_{\varepsilon/2}$$

$$\text{AR (Acceptance Region)} : \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} < Z_{1-\varepsilon/2} \quad or \quad \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} > Z_{\varepsilon/2}$$

In order to obtain an expression in terms of $(1-\beta)$, we calculate the sample size using the lower tail of the standard normal distribution.

$$\beta(\text{Type II error}) = P\left(\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}} > Z_{\varepsilon/2}|H_0 \ \ false\right)$$

$$\beta = 1 - P\left(\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}} < Z_{\varepsilon/2}|H_0 \ \ false\right)$$

$$P\left(Z < \frac{\mu_0-\mu^*}{\sigma/\sqrt{n}} + Z_{\varepsilon/2}\right) = 1-\beta \ \ \cdots \ \ w.l.g$$

where μ^* is the observed mean and μ_0 is the hypothesized mean.

Let $\mu_0 - \mu^* = \delta$, the mean difference. Then,

$$\phi\left(\frac{\delta}{\sigma/\sqrt{n}} + Z_{\varepsilon/2}\right) = 1-\beta \Longrightarrow \frac{\delta}{\sigma/\sqrt{n}} = Z_{1-\beta} - Z_{\varepsilon/2}, \ \ where \ \ \phi^{-1}(1-\beta) = Z_{1-\beta}$$

$$\frac{n\delta^2}{\sigma^2} = (Z_{1-\beta} - Z_{\varepsilon/2})^2$$

$$\therefore n \approx \frac{\sigma^2(Z_{\varepsilon/2} - Z_{1-\beta})^2}{\delta^2} = \frac{(Z_{\varepsilon/2} - Z_{1-\beta})^2}{\delta^{*2}}, \ \ where \ \ \delta^* = \delta/\sigma$$

Also under a two-sample z-test given ε and δ^*, the sample size needed in each group in order to achieve a desired power $(1-\beta)$ can be derived in a similar way to the one-sample test. Consider the hypothesis $H_0 : \mu_1 - \mu_2 = D_0 \ \ vrs \ \ H_1 : \mu_1 - \mu_2 \neq D_0$, we define the RR and AR as follows:

$$\text{RR (Rejection Region)} : \frac{(\bar{X}_1 - \bar{X}_2) - D_0}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} > Z_{1-\varepsilon/2} \quad or \quad \frac{(\bar{X}_1 - \bar{X}_2) - D_0}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} < Z_{\varepsilon/2}$$

$$\text{AR (Acceptance Region)} : \frac{(\bar{X}_1 - \bar{X}_2) - D_0}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} < Z_{1-\varepsilon/2} \quad or \quad \frac{(\bar{X}_1 - \bar{X}_2) - D_0}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} > Z_{\varepsilon/2}$$

Again in order to obtain an expression in terms of $(1-\beta)$, we calculate the sample size using the lower tail of the standard normal distribution. By assuming equal sample sizes;

$$\beta(\text{Type II error}) = P\left(\frac{(\bar{X}_1 - \bar{X}_2) - D_0}{\sigma\sqrt{\frac{1}{n} + \frac{1}{n}}} > Z_{\varepsilon/2} | H_0 \ \ false\right)$$

$$\beta = 1 - P\left(\frac{(\bar{X}_1 - \bar{X}_2) - D_0}{\sigma\sqrt{\frac{2}{n}}} < Z_{\varepsilon/2}\right) \ \ \cdots \ \ w.l.g$$

$$P\left(Z < \frac{D_0 - \mu^*}{\sigma\sqrt{\frac{2}{n}}} + Z_{\varepsilon/2}\right) = 1 - \beta$$

where μ^* is the observed mean difference and D_0 is the hypothesized mean difference we want to detect.

Again, let $D_0 - \mu^* = \delta$, then;

$$\phi\left(\frac{\delta}{\sigma\sqrt{\frac{2}{n}}} + Z_{\varepsilon/2}\right) = 1 - \beta \Longrightarrow \frac{\delta}{\sigma\sqrt{\frac{2}{n}}} = Z_{1-\beta} - Z_{\varepsilon/2} \quad , where \ \ \phi^{-1}(1-\beta) = Z_{1-\beta}$$

$$\frac{n\delta^2}{2\sigma^2} = (Z_{1-\beta} - Z_{\varepsilon/2})^2$$

$$\therefore n \approx \frac{2\sigma^2(Z_{\varepsilon/2} - Z_{1-\beta})^2}{\delta^2} = \frac{2(Z_{\varepsilon/2} - Z_{1-\beta})^2}{\delta^{*2}} \quad , where \ \ \delta^* = {}^{\delta}/_{\sigma} \qquad (3.1)$$

NOTE: Z_ε is the lower ε percentile of a standard normal distribution.

For the simultaneous analysis of a set of hypotheses (genes), sample size depends on ε, $(1-\beta)$ and δ^* of each individual hypothesis. We also assume an equal standardized effect size δ^* for all m_1s; thus the power $(1-\beta)$ of making any true discovery is constant.

3.1.2 Binary Outcomes

For two independent binary populations with proportion parameters p_1 and p_2, if x_1 and x_2 are the observed number of *successes* of interest in the two populations respectively with index n, then the comparison of these two populations can be cross-classified as a 2×2 contingency table and Fisher's "exact" (two-sided) test may be used to test the null hypothesis, $p_1 = p_2$, against the alternative that $p_1 \neq p_2$.

To find the minimum n such that we achieve a $100(1-\beta)$ percent comparison-wise power (CW-Power), numerous formulas have been suggested. Some of which are

1. The "Uncorrected χ^2 formula" as given in Fleiss (1973) is,

$$n = \frac{\left[Z_{1-\varepsilon/2}\sqrt{2\bar{p}\bar{q}} + Z_{1-\beta}\sqrt{(p_1q_1 + p_2q_2)}\right]^2}{(p_1 - p_2)^2}, \tag{3.2}$$

   where $\bar{p} = \frac{x_1+x_2}{n_1+n_2} = \frac{n_1p_1+n_2p_2}{n_1+n_2} = \frac{p_1+p_2}{2}$ for $n_1 = n_2 = n$ and $\bar{q} = 1 - \bar{p}$. Also, Haseman (1978) argued that sample sizes using this formula are generally too low.

2. The "Corrected χ^2 formula" as given by Kramer and Greenhouse (1959) is,

$$n = A\left[1 + \sqrt{1 + \frac{8(p_1 - p_2)}{A}}\right]^2 \Big/ \left[4(p_1 - p_2)^2\right], \tag{3.3}$$

   where $A = [Z_{1-\varepsilon/2}\sqrt{2\bar{p}\bar{q}} + Z_{1-\beta}\sqrt{(p_1q_1 + p_2q_2)}]^2$. Equation (3.3) looks like equation (3.2) with a correction factor. The formula is also said to be conservative [8].

3. J. T. Casgrande et al (1978) proposed a general form for equations (3.2 & 3.3) and an "Improved χ^2 formula" as well. This is given as

$$n = A\left[1 + \sqrt{1 + \frac{4(1 - 2c)(p_1 - p_2)}{A}}\right]^2 \Big/ \left[4(p_1 - p_2)^2\right], \tag{3.4}$$

   By setting $c = -0.5$ in (3.4), we get the Kramer-Greenhouse formula (3.3) and by doing so for $c = 0.5$, we get also (3.2). Therefore for $c = 0$, we get the improved χ^2 formula

$$n = A\left[1 + \sqrt{1 + \frac{4(p_1 - p_2)}{A}}\right]^2 \Big/ \left[4(p_1 - p_2)^2\right]. \tag{3.5}$$

   Sample sizes obtained using this formula are usually close to their exact values and lie between values obtained using equations (3.2) and (3.3).

Equations (3.1) and (3.5) are the two formulas that will be used for the calculation of sample sizes depending on whether the response is continuous or binary.

3.2 Independence Model

In relation to Table 1.4, lets assume the relationship among the m set of hypotheses are independent (for instance, the responses among m genes be independent). The outcome or decision (to reject or fail to reject) of a univariate test on a false null hypothesis (for instance, a differentially expressed gene) can be modelled by a Bernoulli X_i random variable with success probability $(1-\beta)$. By success, we mean making a true rejection (for instance, declaring a differentially expressed gene as significant).

Given m_0 and m_1, the random variable for the number of true rejections (discoveries) X, is defined as a sum of the m_1 independent Bernoulli X_i random variables each with a $(1-\beta)$ success probability. Mathematically,

$$X = \sum_{i=1}^{m_1} X_i \text{ is binomially distributed} \sim Bin(m_1, 1-\beta),$$

with mean $m_1(1-\beta)$ and variance $m_1\beta(1-\beta)$. The probability of at least k true discoveries is given as

$$\phi = \sum_{x=k}^{m_1} \binom{m_1}{x} (1-\beta)^x \beta^{m_1-x} \tag{3.6}$$

ϕ can be interpreted as the (family-wise) power of identifying at least k out of m_1 false nulls (i.e. differentially expressed genes) for a given comparison-wise power $(1-\beta)$.

Given a CWER of ε, to detect all false nulls (for instance, all differentially expressed genes) at the desired family-wise power of ϕ, the required comparison-wise power $(1-\beta)$ for each individual hypothesis (gene) can be calculated from equation (3.6) by putting $k = m_1$;

$$\phi = \sum_{x=k=m_1}^{m_1} \binom{m_1}{x} (1-\beta)^x \beta^{m_1-x} = (1-\beta)^{m_1}$$
$$\therefore 1-\beta = \phi^{1/m_1}$$

When m_1 is moderate or large, $(1-\beta)$ is close to 1, even for small ϕ. For example, when $m_1 = 200$ and $\phi = 0.05$, then $1-\beta = 0.9851$. Thus, a large sample size would be needed in order to identify all false nulls (for instance, all differentially expressed genes). This

independence model will not take into account dependencies among the test endpoints and, hence, will overestimate sample sizes in some cases. When dependencies are anticipated, it is better to use a dependence model instead.

3.3 Dependence Model

The assumption of mutual independence as seen with the previous model would not always hold in situations where the hypotheses being tested are correlated (for instance, responses among different genes may not be necessarily uncorrelated). Even if this assumption is true, the ordinary binomial distribution in practice has been shown not to provide a good fit usually, because the observed variance tends to be higher than that of this parametric model when fitted [3]. This set back calls for a revised model that accounts for the possibility of association among hypotheses being tested.

Kupper and Haseman (1978) introduced a generalization of the ordinary binomial distribution called the correlated binomial (CB) distribution. This distribution is derived on the assumption that the binary responses of the hypotheses (genes) being tested are not mutually independent. Their idea stems from Bahahur [4], who showed that the correct and most general expression for $P(X = x)$ takes the form

$$P(X = x) = P_{[1]}(x) \cdot f(x_1, x_2, \cdots, x_n), \tag{3.7}$$

where $f(x_1, x_2, \cdots, x_n)$ is a "correction factor" that gets multiplied by the ordinary binomial distribution ($P_{[1]}(x)$) to "correct for" the lack of mutual independence among the $X_i s$. Bahadur showed that if X_i is standardized to $Z_i = (X_i - p)/[p(1-p)]^{1/2}$, then

$$f(x_1, x_2, \cdots, x_n) = 1 + \sum_{i<j} E(Z_i Z_j) z_i z_j + \sum_{i<j<k} E(Z_i Z_j Z_k) z_i z_j z_k + \cdots + E(Z_1 Z_2 \cdots Z_n) z_1 z_2 \cdots z_n.$$

Thus $f(x_1, x_2, \cdots, x_n)$ is a function of p and the second-order product moment, the third-order product moment, etc.,up to the nth order product moment. Because of the complexity of the general form of equation (3.7), working with approximations often helps.

One simple technique to obtaining such an approximation is to disregard correlations of order higher than are required for reasonable accuracy. For instance, suppose we ignore all higher-order correlations, then we get back the ordinary binomial distribution. Now, if all approximations higher than order two are reasonably neglected, then

$$P_{[2]}(X = x) = P_{[1]}(x)[1 + \sum_{i<j} E(Z_i Z_j) z_i z_j], \tag{3.8}$$

is a second-order approximation to $P(x)$. $P_{[2]}(X = x)$ and any $P_{[m]}(X = x)$ for $1 < m < n$, may fail to be non-negative for some values of x, even though it true that $\sum_{m=0}^{n} P_{[m]}(x) = 1$. Note that

$$E(Z_i Z_j) = Corr(X_i, X_j) = \rho = \theta / p(1-p),$$

where $Cov(X_i, X_j) = \theta$. Hence, by retaining only the $1st$ order correlation between responses (genes) and denoting θ as the covariance, the random variable X is such that

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x} \cdot \Big\{ 1 + \frac{\theta}{2p^2(1-p)^2} \big[(x - np)^2 + x(2p-1) - np^2 \big] \Big\}, \tag{3.9}$$

where p is the probability of success. The bounds for the parameters in equation (3.9) that makes it a valid probability distribution have been shown by Bahadur; see [5].
Notes to reader:

1. When $\theta = 0$, the CB distribution becomes the ordinary binomial distribution.

2. The CB model allows for the possibility of negative intra-cluster correlation (ρ).

Now $w.l.g$, let $p = 1 - \beta$ the probability of making a true discovery out of a set of false nulls (m_1). By assuming the relationship among these false nulls are equally correlated, i.e. $Corr(X_i, X_j) = \rho$ for any i, js, the mean and variance of the total number of true discoveries $X = \sum_{i=1}^{m_1} X_i$ are

$$E(X) = m_1(1-\beta) \text{ and } Var(X) = m_1 \beta (1-\beta) \big[1 + \rho(m_1 - 1) \big]. \tag{3.10}$$

When $\theta > 0$, X has a supra-binomial variation (or an over-dispersion binomial), and ρ is once again know as the intra-cluster correlation. Over-dispersion means that the data

shows evidence that the variance of the response X is "too big" in comparison to $n\beta(1-\beta)$. When this abnormality occurs, the beta-binomial model provides an appropriate correction. The distribution is developed with the assumption that the probability parameter of success $(1-\beta)$ is random (i.e. changes for each test hypothesis). The model therefore assumes as prior distribution for $(1-\beta)$ a beta distribution. It must be said that numerous researchers have contributed to the theory behind beta-binomial and its applications in various fields, notable among them are Pearson (1925), Skellam (1948), Lord (1965), Greene (1970), Massy et. al. (1970), Griffiths (1973), Williams (1975), Huynh (1979), Wilcox (1979), Smith (1983), Lee and Sabavala (1987), Hughes and Madden (1993), and Shuckers (2003).

The beta distribution is a continuous distribution on the interval $[0, 1]$, with shape parameters $a > 0$ and $b > 0$. By letting $(1-\beta)$ have a beta distribution, its probability density function is given by;

$$f[(1-\beta)|a, b] = \frac{(1-\beta)^{a-1}\beta^{b-1}}{\beta(a, b)}, \;\; 0 < \beta < 1, \;\; a > 0, \;\; b > 0 \tag{3.11}$$

where $\beta(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$ denotes the beta function with $\Gamma(\cdot)$ being the gamma function.
Notes to reader:

1. If both a and b are greater than 1, the distribution is unimodal.

2. If both a and b equals 1, the distribution is equivalent to the continuous uniform distribution on that interval.

3. If one of a or b is less than 1, the distribution is J-shaped or reverse J-shaped.

4. If both a and b are less than 1, the distribution is U-shaped.

The mean and variance are given by;

$$E(1-\beta) = \frac{a}{a+b} \;\; and \;\; Var(1-\beta) = \Big(\frac{a}{a+b}\Big)\Big(\frac{b}{a+b}\Big)\Big(\frac{1}{a+b+1}\Big).$$

By Bayes Theorem, the joint distribution of X and $(1-\beta)$ is given as

$$f(x, 1-\beta) = f(x|1-\beta) \cdot f(1-\beta|a, b)$$

$$f(x, 1-\beta) = \binom{m_1}{x}(1-\beta)^x \beta^{m_1-x} \cdot \frac{(1-\beta)^{a-1}\beta^{b-1}}{\beta(a,b)}$$

The marginal distribution of x, $P(x)$ is derived to be

$$\begin{aligned} P(X=x) &= \int_0^1 f(x,1-\beta)d\beta = \int_0^1 \binom{m_1}{x}(1-\beta)^x \beta^{m_1-x} \cdot \frac{(1-\beta)^{a-1}\beta^{b-1}}{\beta(a,b)} d\beta \\ &= \binom{m_1}{x}\frac{1}{\beta(a,b)}\int_0^1 (1-\beta)^{a+x-1}\beta^{b+m_1-x-1}d\beta \\ &= \binom{m_1}{x}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \cdot \frac{\Gamma(a+x)\Gamma(b+m_1-x)}{\Gamma(a+b+m_1)} \\ \therefore P(X=x) &= \binom{m_1}{x}\frac{\beta(a+x,b+m_1-x)}{\beta(a,b)}, \;\; x=0,1,2,\ldots,m_1. \end{aligned} \tag{3.12}$$

$P(x)$ is the beta-binomial distribution, and we write $x \sim betabin(n,a,b)$. Its mean and variance can be derived using the idea of hierarchical models in (see [6]). Since X and $(1-\beta)$ are both random variables, then

$$E[X] = E[E(X|1-\beta)], \;\; Var(X) = E[Var(X|1-\beta)] + Var[E(X|1-\beta)]],$$

provided that the expectations exist. By means of the above expression, we obtain

$$E[X] = E[m_1(1-\beta)] = \frac{m_1 a}{a+b},$$

and

$$\begin{aligned} Var[X] &= E[m_1\beta(1-\beta)] + Var[m_1(1-\beta)] \\ &= m_1 E(\beta) - m_1 E(\beta^2) + {m_1}^2 Var(1-\beta) \\ &= m_1[E(\beta) - Var(\beta) - E[\beta]^2 + m_1 Var(1-\beta)] \end{aligned}$$

but $E(\beta) = \frac{b}{a+b}$, therefore

$$\begin{aligned} Var[X] &= m_1\left[\frac{b}{a+b} - \frac{ab}{(a+b)^2(a+b+1)} - \frac{b^2}{(a+b)^2} + \frac{m_1 ab}{(a+b)^2(a+b+1)}\right] \\ &= m_1\left[\frac{ab(a+b+1+m_1-1)}{(a+b)^2(a+b+1)}\right] = m_1\left(\frac{a}{a+b}\right)\left(\frac{b}{a+b}\right)\left(1+\frac{m_1-1}{a+b+1}\right). \end{aligned}$$

Now if we compare the above mean and variance expressions of the Beta-Binomial with that of the Correlated Binomial in equation (3.10), we have $(1-\beta) \equiv \frac{a}{a+b}$ and $\rho \equiv (a+b+1)^{-1}$. Let $\rho = \theta$, then

$$E(X) = m_1(1-\beta) \text{ and } Var(X) = m_1\beta(1-\beta)\bigg(1+\theta(m_1-1)\bigg),$$

and we see again that the variance is higher than that of the $bin(m_1, 1-\beta)$, hence accounting for any possible over-dispersion. Now these results for correlated binomial outcomes may be utilized to more accurately calculate the probability of true discoveries or the family-wise power for sample size estimation.

The probability of at least k true discoveries is given as

$$\phi = \sum_{x=k}^{m_1} \binom{m_1}{x} \frac{\beta(a+x, b+m_1-x)}{\beta(a,b)} \tag{3.13}$$

where ϕ can be interpreted as the (family-wise) power of identifying at least k out of m_1 true discoveries (i.e. truly differentially expressed genes) for a given CWER ε and comparison-wise power $(1-\beta)$. Hence, to detect at least x truly non-null endpoints, we can use any of the three measures such as, sensitivity (x/m_1), true discovery (x/r) and accuracy $(m_0 - v + x)/m$ (see [7]).

Equation (3.13) is the fundamental formula needed for the estimation of sample size under the correlated model. For given ϕ and θ, the required comparison-wise power $(1-\beta) = a/(a+b)$ is obtained by solving the beta-binomial tailed probability in equation (3.13). The needed sample size can then be determined using equation (3.1) for a given CWER ε and δ^*.

As stated by C.-A. Tsai et al, 2004 (see [7]), the intra-cluster correlation θ between Bernoulli random variables X_i and X_j is usually estimated in practice, by assuming each pair of the test statistics corresponding to these random variables to be bivariate normally distributed with a common correlation ρ.

Chapter 4

Proposed Sample Size Methodology

The issue of power and sample size analyses in genomic studies has received a considerable amount of attention, but most methods do not consider the dependency structure among the test endpoints. Even when it is considered, the test endpoints from these numerous tests are usually assumed to be normally distributed and methods that take into account the correlation structure are available (e.g., Hunter, 1976; McCann & Edwards, 1996; Naiman, 1987; Sun & Loader, 1994; Worsley, 1982). But, when the tests endpoints are not normally distributed, these methods are not directly applicable.

Resampling (e.g. Westfall and Troendle, 2008) and simulation-based methods have also been considered for these chi-square dependent endpoints. All these approaches are either criticized for being sometimes "computationally demanding" or confusing to practitioners in fields other than statistics. None of the methods proposed so far in the literature have focused on incorporating the control of multiplicity specifically for chi-square distributed endpoints with arbitrary correlation structure into sample size estimation (see [10]).

In this chapter, we propose a method of sample size estimation that addresses these drawbacks. In formulating our method, we wish to infer ideas from C.-A. Tsai et al's proposed model and procedure for sample size estimation.

4.1 Proposed Method

According to C.-A. Tsai et al, 2004 [7], when the tests are correlated, the number of true discoveries can be modelled by a generalized binomial model (i.e. CB model). Even though this model exists, they proposed using the beta-binomial distribution to model the number of true discoveries because of its mathematical tractability. The beta-binomial model as reviewed in the literature assumes both an equal standardized effect size and correlation among the genes. But instead of assuming the beta-binomial model with equal dependency among test endpoints (which are usually assumed to be normally distributed) to estimate sample sizes, we propose a method that will not require such rigid assumptions. In particular, the proposed method will attempt to make sample size estimates that take into account the FWER and complex dependency structure of the test endpoints. The proposed method will be adaptable to both normally or t distributed as well as chi-square distributed test endpoints.

> *Can we control at least the* ***FWER*** *when these endpoints are t or chi-square dependent random variables with arbitrary correlation structure and better estimate sample sizes using independence?*

The answer to this question is the main concern of this thesis. Hence, the objective is to;

- Bound the probability of the union of all m sets, i.e. for all events $A_i = \{Y_i > c\}, \quad i = 1, 2, \ldots m.$ where Y_i is a random variable and c a constant.
- Develop a computationally efficient algorithm for estimating sample sizes based on this bound while controlling the FWER.

A simple approach using the Bonferroni inequality

$$P\Big(\bigcup_{i=1}^{m} A_i\Big) \leq \sum_{i=1}^{m} P\Big(A_i\Big) = \alpha,$$

generally is too conservative and more powerful bounds are needed. Hunter (1970) and Worsley (1982) improved this inequality by tightening the upper bound on $P\Big(\cup_{i=1}^{m} A_i\Big)$

using a tree spanning algorithm. The use of the tree algorithm is justified by the inequality

$$P\Big(\bigcup_{i=1}^{m} A_i\Big) \leq \sum_{i=1}^{m} P\Big(A_i\Big) - \sum_{(i,j)\in\tau} P\Big(A_i \cap A_j\Big). \tag{4.1}$$

Hunter and Worsley method involves choosing a point c such that the right-hand side of the inequality above equals α. In particular, the set of nodes for the tree are the $A_i s$ and the set of branches are the intersections of $A_i A_j s$. This provides an improved solution over Bonferroni in ALL settings, and only requires the computations of the univariate and bivariate cumulative distributions of interest. However, in order to evaluate the joint probability on the right hand side of equation (4.1), calculation of bivariate t or chi-square probabilities are necessary. Again, the goal is to find c such that $P\Big(\cup_{i=1}^{m} A_i\Big)$ is bounded by α (here $\mathbf{Y} \sim \boldsymbol{\chi^2_m(\nu, R)}$ *or* $\boldsymbol{t_m(\nu, R)}$, where m is the dimension, ν is the degrees of freedom and $\mathbf{R}$ is the correlation structure). For simulation purposes, we will also make assumptions about the correlation structure of the endpoints, such as assuming compound symmetric (CSS) and autoregressive structures (ARS) (see [10]).

Compound Symmetric Structure

$$\begin{pmatrix} 1 & \rho_a & \rho_a & \rho_a \\ \rho_a & 1 & \rho_a & \rho_a \\ \rho_a & \rho_a & 1 & \rho_a \\ \rho_a & \rho_a & \rho_a & 1 \end{pmatrix}.$$

Autoregressive Structure

$$\begin{pmatrix} 1 & \rho_a & \rho_a^2 & \rho_a^3 \\ \rho_a & 1 & \rho_a & \rho_a^2 \\ \rho_a^2 & \rho_a & 1 & \rho_a \\ \rho_a^3 & \rho_a^2 & \rho_a & 1 \end{pmatrix}$$

Dunnett and Sobel (1954) gave an expression for the joint t density function of m variates $(Y_1, Y_2, \ldots, Y_m)$. The bivariate t density function is given by

$$P\Big(Y_1 = c_1, Y_2 = c_2; \rho_{12}\Big) = \frac{1}{2\pi\sqrt{1-\rho_{12}^2}}\left[1 + \frac{c_1^2 - 2\rho_{12}^2 c_1 c_2 + c_2^2}{\nu(1-\rho_{12}^2)}\right]^{-\frac{1}{2}(\nu+2)},$$

where $Corr(Y_1, Y_2) = \rho_{12}$.

Also, Krishnaiah (1980) derived the joint chi-square density function for (Y_1, Y_2). The cumulative function is given by

$$P\Big(Y_1 \leq c_1, Y_2 \leq c_2\Big) = \Big(1 - \rho_{12}^2\Big)^{-\nu/2} \times \sum_{j=0}^{\infty} \frac{1}{\Gamma(\nu/2 + j)\Gamma(j+1)} \times {\rho_{12}}^{2j} \gamma(\nu/2 + i, c_1^*)\gamma(\nu/2 + i, c_2^*),$$

where γ is the incomplete gamma function, $c_i^* = {c_i}/{2(1 - {\rho_{12}}^2)}$ for $i = 1, 2$ and $Corr(Y_1, Y_2) = \rho_{12}$.

The $(1-\varepsilon)$ quantile c, i.e. $P\big(Y_1 \leq c_1, Y_2 \leq c_2\big) = 1-\varepsilon$, can be obtained using numerical integration, where $c_1 = c_2 = c$. Note

1. $P(Y_i > c) = \varepsilon_\tau$, where ε_τ is an adjusted familywise-error rate $(FWER_\tau)$
2. With $FWER_\tau$, we then find the adjusted comparison-wise error (α_τ) using either
   - $\alpha_\tau = \frac{-\ln(1-FWER_\tau)}{m_0}$ (Lee and Whitmore, 2002)
   - $\alpha_\tau = \frac{FWER_\tau}{m_0}$ (Bonferroni)

Therefore with α_τ and β values, we can calculate the sample sizes using any of the sample size formulas. Note that since we account for the correlation structure among the endpoints in the calculation of α_τ, β is obtained using the independence model by setting ϕ, m_1 and x to some values in the equation

$$\phi = \sum_{x=k}^{m_1} \binom{m_1}{x} (1-\beta)^x \beta^{m_1 - x} \ ,$$

where x can be $(\lambda \times m_1)$ according to C.-A. Tsai et al, 2004 (see [7]).

Chapter 5

Results

This chapter presents the results from our simulation studies. The simulation outcomes will provide grounds for comparing the proposed method with other well known methods of sample size calculation. Also, different combination settings are used in coming up with possible sample sizes. We then compare the differences in sample sizes between the proposed method and the other methods to access performance.

5.1 Simulation Setting

The following factors corresponding to the number of genes, m, proportion of false nulls, π_1, family-wise power, ϕ, sensitivity rate, λ, intra-cluster correlation, θ and family-wise error rate, ε are considered in general during all the simulation process;

1. $m = 1000, 10000$
2. $\phi = 0.8, 0.85, 0.9, 0.95$
3. $\lambda = 0.8, 0.9, 0.95$
4. $\theta = 0.00, 0.05, 0.22, 0.5$
5. $\pi_1 = m_1/m = 0.05, 0.1, 0.2$
6. $\varepsilon = 0.05$.

Data consisting of the various combinations of the factor levels (for $m, \phi, \lambda, \theta, \pi_1$ and ε) is generated. Under the independence model, $\theta = 0.00$ hence it is ignored as a factor but

with the dependence model θ remains a factor. With various combinations of the factor levels and equations (3.6) and (3.13), we compute the comparison-wise power $(1-\beta)$ using a one-dimensional root(zero) finding method for each combination under any of the models (i.e independence, dependence and proposed). The *uniroot* function (*stats* package) in *R* is used for finding the root. For example, consider the following table

Table 5.1: Possible values for comparison-wise power at different combinations

Model	m	ϕ	λ	π_1	θ	$1-\beta$
Independence	1000	0.8	0.8	0.05	0.00	0.8336
Independence	10000	0.95	0.9	0.1	0.00	0.9143
Dependence	1000	0.8	0.8	0.05	0.22	0.8845
Dependence	10000	0.95	0.9	0.1	0.22	0.9837

Because of convergence issues in the computation of the comparison-wise power under the dependence model using the various assumed factor level combinations for m, ϕ, λ and π_1, the intra-cluster correlation (θ) can only lie between 0.00 and 0.5 inclusive (i.e. $0 < \theta \leq 0.5$). With a given FWER (ε) of 0.05, we compute the CWER (α) using $\alpha = FWER/m_0$. Having computed $(1-\beta)$ and α values, sample sizes are obtained using either equation (3.1) or (3.5). Equations (3.1) and (3.5) involves specifying values for δ^* (standardized effect size) and $p_1 \& p_2$ (proportions of successes of interest) respectively, where $p_1 > p_2$. These factors are also set as $\delta^* = 2$ and $0.05 \leq p_1 \& p_2 \leq 0.7$ when relevant in the simulations.

For the proposed model, we set the number of replications to the length of the factor level combinations generated when computing the adjusted FWER (i.e. $FWER_\tau$) under settings such as the pairwise correlation, ρ_{ij}, the degree of freedom, ν, etc. Again for simulation purposes, we only assume two correlation structures (compound symmetric and auto regressive) for the endpoints. Given an $FWER_\tau$ obtained by adjusting the FWER based on the correlation structure, we compute the $CWER_\tau$ (α_τ) using $\alpha_\tau = FWER_\tau/m_0$.

Having computed $(1-\beta)$ and α_τ values, sample sizes are again obtained using either equation (3.1) or (3.5). The proposed model makes use of the independence model in computing $(1-\beta)$.

In comparing the performance of the proposed method to both independence and dependence models, we simply examine the sample size differences between each of these existing models and the proposed at all factor level combinations to ascertain any gains or losses as well as existing trends. Here, *positive* difference means the proposed method results in savings (i.e. less sample size) whiles any *negative* difference implies the counter. The *lattice* package in R is employed in summarizing these differences as well as discovering patterns that may exist between certain factors and sample size differences. All simulations were performed in R statistical software package.

5.2 Simulation Results

We present the simulation results of our proposed method compared to the other existing methods. Table 5.2 contains the family-wise adjusted error rates using the proposed method at certain levels of m (number of genes) and ρ (correlation). Levels P_{CSS} and P_{ARS} refer to the proposed method assuming compound symmetric and autoregressive correlation structures respectively between endpoints.

Table 5.2: $FWER_\tau$ under proposed method for $\rho = 0.22, 0.5, 0.9$ at a given FWER

FWER	m	Method	$\rho = 0.22$	$\rho = 0.5$	$\rho = 0.9$
0.05	1000	P_{CSS}	0.05231	0.06329	0.11816
		P_{ARS}	0.05011	0.05303	0.09805
	10000	P_{CSS}	0.05292	0.06741	0.15996
		P_{ARS}	0.05013	0.05370	0.12135

Below are the simulations for the two (2) response variables considered.

5.2.1 Continuous Outcomes

Tables 5.3, 5.4 and 5.5 show the sample size estimates and their differences for the proposed method (P_{CSS} and P_{ARS}) and dependence model (BB) at different levels of m, π_1, λ and ϕ. From the tables, the proposed method yields smaller significant estimates than the dependence model in settings where a high level of power at given sensitivity rate is preferred. This is true for lower and moderate proportions of false nulls (i.e. π_1 at 0.05 and 0.1). The proposed method only performs poorly in settings with large number of genes having a high proportion of false nulls (i.e. m and π_1 at 10000 and 0.2 respectively).

Table 5.3: Dependence model (BB) vs proposed method for $\rho = 0.22$

m	π_1	ϕ	λ	P_{ARS}.vs.BB	P_{CSS}.vs.BB	BB	P_{ARS}	P_{CSS}
1000	0.05	0.95	0.80	4 **	4 **	18	14	14
	0.10	0.95	0.80	5 **	5 **	18	13	13
	0.20	0.95	0.80	4 **	4 **	17	13	13
		0.95	0.90	4 **	4 **	19	15	15
10000	0.05	0.90	0.80	4 **	4 **	19	15	15
		0.95	0.80	5 **	6 **	21	16	15
		0.95	0.90	5 **	5 **	23	18	18
	0.10	0.90	0.80	4 **	4 **	19	15	15
		0.95	0.80	6 **	6 **	21	15	15
		0.95	0.90	5 **	5 **	23	18	18
		0.95	0.95	4 **	4 **	24	20	20
1000	0.05	0.85	0.80	2 *	2 *	15	13	13
		0.90	0.80	3 *	3 *	16	13	13
		0.90	0.90	2 *	2 *	18	16	16
		0.95	0.90	3 *	3 *	19	16	16

Signif. codes: '**' best gain '*' moderate gain '·' loss

Table 5.4: Dependence model (BB) vs proposed method for $\rho = 0.22$

m	π_1	ϕ	λ	P_{ARS}.vs.BB	P_{CSS}.vs.BB	BB	P_{ARS}	P_{CSS}
1000	0.05	0.95	0.95	2 *	2 *	20	18	18
	0.10	0.85	0.80	2 *	2 *	15	13	13
		0.85	0.90	2 *	2 *	17	15	15
		0.90	0.80	3 *	3 *	16	13	13
		0.90	0.90	2 *	2 *	18	16	16
		0.95	0.90	3 *	3 *	19	16	16
		0.95	0.95	3 *	3 *	21	18	18
	0.20	0.85	0.80	2 *	2 *	15	13	13
		0.90	0.80	3 *	3 *	16	13	13
		0.90	0.90	3 *	3 *	18	15	15
		0.90	0.95	2 *	2 *	19	17	17
		0.95	0.95	2 *	2 *	20	18	18
10000	0.05	0.80	0.80	2 *	2 *	17	15	15
		0.85	0.80	3 *	3 *	18	15	15
		0.85	0.90	2 *	2 *	20	18	18
		0.90	0.90	3 *	3 *	21	18	18
		0.90	0.95	2 *	2 *	22	20	20
		0.95	0.95	3 *	3 *	24	21	21
	0.10	0.80	0.80	2 *	2 *	17	15	15
		0.85	0.80	3 *	3 *	18	15	15
		0.85	0.90	2 *	2 *	20	18	18
		0.90	0.90	3 *	3 *	21	18	18
		0.90	0.95	2 *	2 *	22	20	20
	0.20	0.95	0.95	2 *	2 *	22	20	20

Signif. codes: '**' best gain '*' moderate gain '·' loss

Table 5.5: Dependence model (BB) vs proposed method for $\rho = 0.22$

m	π_1	ϕ	λ	P_{ARS}.vs.BB	P_{CSS}.vs.BB	BB	P_{ARS}	P_{CSS}
10000	0.20	0.80	0.80	-6 ·	-6 ·	9	15	15
		0.80	0.90	-8 ·	-8 ·	9	17	17
		0.80	0.95	-11 ·	-11 ·	9	20	20
		0.85	0.80	-6 ·	-6 ·	9	15	15
		0.85	0.90	-8 ·	-8 ·	9	17	17
		0.85	0.95	-11 ·	-11 ·	9	20	20
		0.90	0.80	-5 ·	-5 ·	10	15	15
		0.90	0.90	-8 ·	-7 ·	10	18	17
		0.90	0.95	-10 ·	-10 ·	10	20	20
		0.95	0.80	-4 ·	-4 ·	11	15	15
		0.95	0.90	-7 ·	-7 ·	11	18	18

Signif. codes: '**' best gain '*' moderate gain '·' loss

Table 5.6 below depicts the only factor combinations that result in some kind of gains with the proposed method. At all other factor combinations, there exist no significant difference in sample size estimates obtained for the proposed method and independence model. Table 5.7 below shows examples of such cases.

Table 5.6: Independence model (IND) vs proposed method for $\rho = 0.22$

m	π_1	ϕ	λ	P_{ARS}.vs.IND	P_{CSS}.vs.IND	IND	P_{ARS}	P_{CSS}
10000	0.05	0.95	0.80	0	1	16	16	15
	0.20	0.90	0.90	0	1	18	18	17

Signif. codes: '**' best gain '*' moderate gain '·' loss

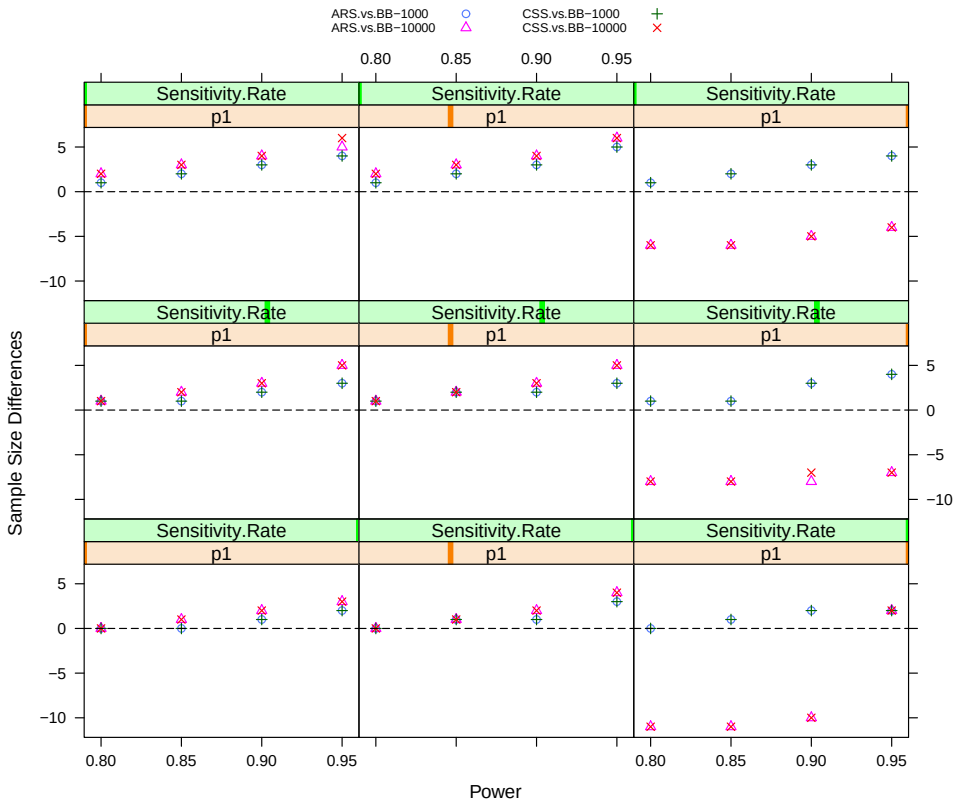
Table 5.7: Independence model (IND) vs proposed method for $\rho = 0.22$

m	π_1	ϕ	λ	P_{ARS}.vs.IND	P_{CSS}.vs.IND	IND	P_{ARS}	P_{CSS}
1000	0.05	0.80	0.80	0	0	13	13	13
		0.80	0.90	0	0	15	15	15
	0.10	0.80	0.90	0	0	15	15	15
		0.85	0.90	0	0	15	15	15
	0.20	0.90	0.95	0	0	17	17	17
		0.95	0.80	0	0	13	13	13
10000	0.05	0.80	0.80	0	0	15	15	15
		0.80	0.90	0	0	18	18	18
	0.10	0.85	0.95	0	0	20	20	20
		0.90	0.80	0	0	15	15	15
	0.20	0.95	0.80	0	0	15	15	15
		0.95	0.90	0	0	18	18	18
		0.95	0.95	0	0	20	20	20

Signif. codes: '**' best gain '*' moderate gain '·' loss

Figure 5.1 below shows plots of sample size differences between the proposed method (P_{CSS} and P_{ARS}) and dependence (a) and independence (b) models at different levels of m, π_1, sensitivity rate (λ) and power (ϕ). From the plots, we observe that at lower and moderate proportions of false nulls (i.e. π_1 at 0.05 and 0.1), there exist significant differences in sample size estimates obtained for the proposed method and dependence model as power level increases. This is not the case in settings with large number of genes having a high proportion of false nulls (i.e. m and π_1 at 10000 and 0.2 respectively). The proposed method performs poorly in those settings. Also from part (b) of the plot, we observe little or no significant difference in sample size estimates obtained for the proposed method and independence model after accounting for dependency.

(a)



(b)

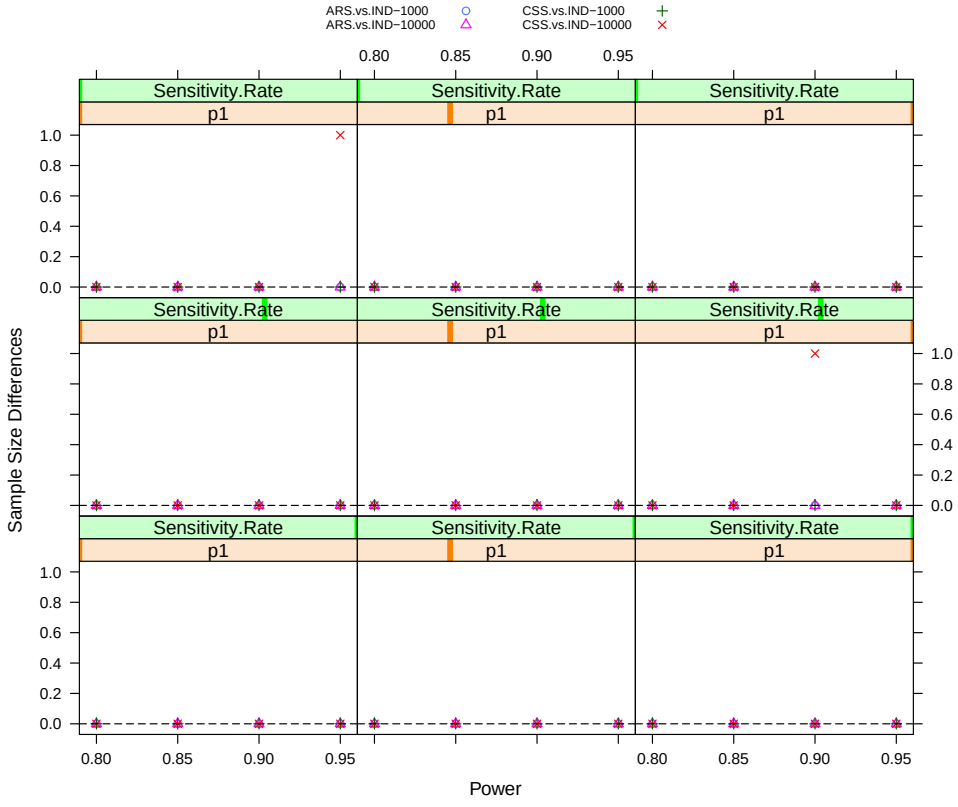


Figure 5.1: (a) & (b) shows effect of factor combinations on sample size difference

5.2.2 Binary Outcomes

Tables 5.8 and 5.9 show the sample size estimates and their differences for the proposed method (P_{CSS} and P_{ARS}) and independence model (IND) at different levels of m, π_1, p_1, p_2, λ and ϕ. From the tables, the proposed method yields small significant estimates than the independence model in settings where p_1 and p_2 gets closer (i.e., at smaller dissimilarities). However, at wider dissimilarities (between p_1 and p_2), there exist no significant difference in sample size estimates obtained for the proposed method and independence model after accounting for dependency.

Table 5.8: Independence model (IND) vs proposed method for $\rho = 0.22$

m	π_1	ϕ	λ	p_1	p_2	P_{ARS}.vs.IND	P_{CSS}.vs.IND	IND	P_{ARS}	P_{CSS}
1000	0.20	0.80	0.95	0.60	0.55	2 *	24 **	6483	6481	6459
		0.95	0.95	0.40	0.35	2 *	24 **	6490	6488	6466
		0.95	0.95	0.20	0.15	1	14 *	4011	4010	3997
	0.05	0.85	0.95	0.50	0.45	2 *	25 **	6703	6701	6678
		0.85	0.95	0.55	0.50	2 *	25 **	6703	6701	6678
		0.80	0.90	0.40	0.35	2 *	22 **	5618	5616	5596
		0.80	0.90	0.65	0.60	2 *	22 **	5618	5616	5596
10000	0.05	0.95	0.95	0.55	0.50	2 *	31 **	8066	8064	8035
		0.95	0.95	0.45	0.40	2 *	30 **	7905	7903	7875
		0.90	0.95	0.35	0.30	1	27 **	7020	7019	6993
		0.90	0.95	0.70	0.65	1	27 **	7020	7019	6993
	0.10	0.90	0.90	0.40	0.35	1	27 **	6542	6541	6515
		0.80	0.80	0.35	0.30	1	23 **	5194	5193	5171
		0.95	0.90	0.45	0.40	1	27 **	6865	6864	6838
		0.95	0.90	0.20	0.15	1	16 *	4071	4070	4055

Signif. codes: '**' best gain '*' moderate gain '·' loss

Table 5.9: Independence model (IND) vs proposed method for $\rho = 0.22$

m	π_1	ϕ	λ	p_1	p_2	P_{ARS}.vs.IND	P_{CSS}.vs.IND	IND	P_{ARS}	P_{CSS}
10000	0.20	0.95	0.95	0.25	0.20	1	21 **	5463	5462	5442
		0.95	0.95	0.20	0.15	0	17 *	4529	4529	4512
1000	0.05	0.80	0.80	0.15	0.10	1	9 *	2236	2235	2227
		0.80	0.80	0.20	0.05	0	1	255	255	254
		0.80	0.80	0.55	0.05	0	0	44	44	44
		0.90	0.95	0.25	0.15	1	4 *	1111	1110	1107
		0.90	0.95	0.30	0.05	0	0	161	161	161
	0.10	0.85	0.80	0.60	0.50	0	5 *	1255	1255	1250
		0.85	0.80	0.65	0.05	0	1	33	33	32
	0.20	0.95	0.90	0.45	0.35	1	6 *	1452	1451	1446
		0.95	0.90	0.50	0.05	0	0	59	59	59
10000	0.05	0.80	0.80	0.50	0.40	1	7 *	1486	1485	1479
		0.80	0.80	0.55	0.05	0	1	52	52	51
	0.10	0.90	0.80	0.10	0.05	0	7 *	1680	1680	1673
		0.90	0.80	0.40	0.05	0	1	88	88	87
	0.20	0.95	0.95	0.70	0.10	0	0	50	50	50
		0.95	0.95	0.70	0.60	1	7 *	1786	1785	1779

Signif. codes: '**' best gain '*' moderate gain '·' loss

Also, Tables 5.10 and 5.11 show the sample size estimates and their differences for the proposed method (P_{CSS} and P_{ARS}) and dependence model (BB) at different levels of m, π_1, p_1, p_2, λ and ϕ. From the tables, the proposed method yields smaller significant estimates than the dependence model in settings where p_1 and p_2 gets closer (i.e at smaller dissimilarities). However, at some smaller dissimilarities (and even larger ones) where ϕ (power) is at a lower level, the dependence model yields smaller significant estimates than

the proposed method after accounting for dependency.

Table 5.10: Dependence model (BB) vs proposed method for $\rho = \theta = 0.22$

m	π_1	ϕ	λ	p_1	p_2	P_{ARS}.vs.BB	P_{CSS}.vs.BB	BB	P_{ARS}	P_{CSS}
1000	0.20	0.95	0.80	0.40	0.35	1698 **	1717 **	6425	4727	4708
		0.95	0.80	0.45	0.25	101 **	102 **	393	292	291
	0.10	0.95	0.80	0.50	0.45	1691 **	1711 **	6892	5201	5181
		0.95	0.80	0.55	0.35	102 **	103 **	430	328	327
	0.05	0.80	0.80	0.35	0.30	413 **	431 **	4859	4446	4428
		0.80	0.80	0.40	0.30	106 **	111 **	1267	1161	1156
10000	0.05	0.80	0.80	0.20	0.15	419 **	433 **	3881	3462	3448
		0.80	0.80	0.25	0.15	115 **	119 **	1082	967	963
	0.10	0.95	0.90	0.65	0.60	1815 **	1841 **	8400	6585	6559
		0.95	0.90	0.70	0.50	114 **	115 **	539	425	424
	0.20	0.95	0.95	0.65	0.60	683 **	710 **	8013	7330	7303
		0.95	0.95	0.70	0.60	165 **	171 **	1950	1785	1779
1000	0.05	0.80	0.80	0.30	0.20	87 **	91 **	1047	960	956
		0.80	0.80	0.45	0.05	4 *	5 *	66	62	61
	0.10	0.90	0.95	0.60	0.45	46 **	49 **	822	776	773
		0.90	0.95	0.70	0.05	1	1	37	36	36
	0.20	0.85	0.80	0.55	0.10	8 *	8 *	63	55	55
		0.85	0.80	0.55	0.35	52 **	53 **	362	310	309

Signif. codes: '**' best gain '*' moderate gain '.' loss

Table 5.11: Dependence model (BB) vs proposed method for $\rho = \theta = 0.22$

m	π_1	ϕ	λ	p_1	p_2	P_{ARS}.vs.BB	P_{CSS}.vs.BB	BB	P_{ARS}	P_{CSS}
10000	0.05	0.80	0.80	0.35	0.05	12 *	13 *	122	110	109
		0.80	0.80	0.35	0.20	63 **	66 **	600	537	534
	0.10	0.85	0.80	0.15	0.05	94 **	96 **	642	548	546
		0.85	0.80	0.50	0.05	9 *	9 *	69	60	60
	0.20	0.95	0.95	0.60	0.45	80 **	83 **	951	871	868
		0.95	0.95	0.70	0.05	2 *	2 *	43	41	41
1000	0.05	0.80	0.95	0.65	0.60	-68 ·	-47 ·	6113	6181	6160
		0.80	0.95	0.70	0.15	-1 ·	-1 ·	51	52	52
	0.10	0.80	0.95	0.65	0.45	-6 ·	-5 ·	414	420	419
		0.80	0.95	0.65	0.60	-92 ·	-70 ·	6229	6321	6299
	0.20	0.80	0.95	0.25	0.20	-17 ·	-1 ·	4618	4635	4619
		0.80	0.95	0.65	0.60	-23 ·	-1 ·	6195	6218	6196
10000	0.20	0.80	0.80	0.30	0.10	-97 ·	-96 ·	143	240	239
		0.80	0.80	0.40	0.05	-32 ·	-32 ·	54	86	86
		0.85	0.80	0.60	0.55	-2279 ·	-2254 ·	3441	5720	5695
		0.85	0.80	0.65	0.05	-12 ·	-12 ·	26	38	38
		0.85	0.90	0.25	0.05	-102 ·	-101 ·	122	224	223
		0.85	0.90	0.25	0.20	-2313 ·	-2294 ·	2476	4789	4770

Signif. codes: '**' best gain '*' moderate gain '·' loss

In addition, Tables 5.8, 5.9, 5.10 and 5.11 show that at wider proportional dissimilarities between p_1 and p_2, smaller sample sizes are required in other to detect differentially associated genes. The reverse is true for smaller proportional dissimilarities. These results are consistent with J. T. Casgrande et al ([8], pg 485). Tables 5.12 below show results for sample size estimates and their differences using a threshold of $n = 50$.

Table 5.12: Dependence model (BB) vs proposed method for $\rho = \theta = 0.22$

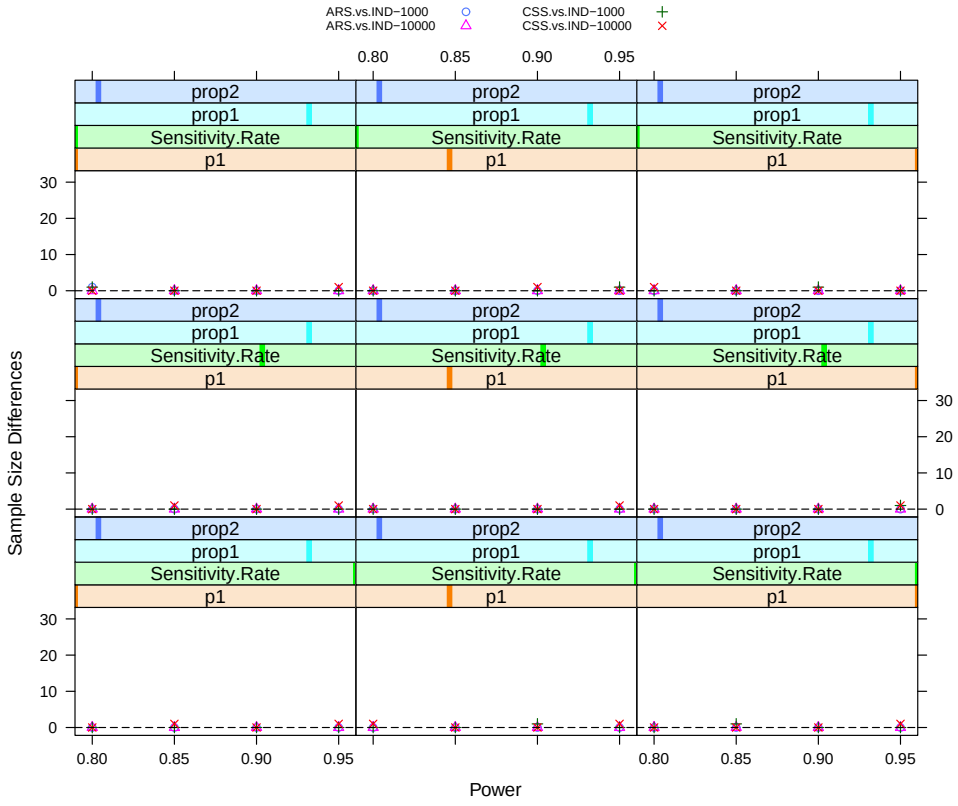
m	π_1	ϕ	λ	p_1	p_2	P_{ARS}.vs.BB	P_{CSS}.vs.BB	BB	P_{ARS}	P_{CSS}
1000	0.05	0.90	0.80	0.70	0.10	5 *	5 *	40	35	35
		0.95	0.80	0.70	0.10	8 *	8 *	44	36	36
	0.10	0.90	0.80	0.60	0.05	6 *	6 *	44	38	38
		0.95	0.80	0.60	0.05	9 *	10 *	48	39	38
	0.20	0.85	0.80	0.60	0.05	5 *	5 *	42	37	37
		0.90	0.80	0.60	0.05	7 *	7 *	44	37	37
		0.95	0.80	0.60	0.05	10 *	11 *	48	38	37
10000	0.05	0.90	0.80	0.65	0.05	7 *	7 *	46	39	39
		0.95	0.80	0.65	0.05	10 *	10 *	49	39	39
	0.10	0.85	0.80	0.65	0.05	5 *	5 *	43	38	38
		0.90	0.80	0.65	0.05	7 *	7 *	45	38	38
		0.95	0.80	0.65	0.05	10 *	11 *	49	39	38
1000	0.05	0.85	0.80	0.70	0.05	3 *	3 *	32	29	29
		0.90	0.80	0.70	0.05	4 *	4 *	33	29	29
	0.10	0.80	0.80	0.70	0.05	2 *	2 *	30	28	28
		0.80	0.80	0.70	0.15	4 *	4 *	45	41	41
		0.90	0.80	0.70	0.05	4 *	4 *	33	29	29
	0.20	0.80	0.80	0.70	0.05	2 *	2 *	30	28	28
		0.80	0.80	0.70	0.15	4 *	4 *	45	41	41
10000	0.20	0.80	0.80	0.60	0.05	-15	-14	29	44	43
		0.85	0.80	0.60	0.05	-14	-13	30	44	43
		0.90	0.80	0.60	0.05	-13	-13	31	44	44
		0.95	0.80	0.60	0.05	-11	-11	33	44	44
		0.80	0.80	0.65	0.05	-12	-12	26	38	38
		0.80	0.90	0.65	0.05	-16	-16	26	42	42

Signif. codes: '**' best gain '*' moderate gain '-' loss

Figures 5.2 and 5.3 shows plots of sample size differences between the proposed method (P_{CSS} and P_{ARS}) and independence model (IND) at different levels of genes (m), π_1, p_1, p_2, sensitivity rate (λ) and power (ϕ). From the plots, we observe that at wider proportional dissimilarities (between p_1 and p_2), there exist no significant difference in sample size estimates obtained for the proposed method and independence model. However, as p_1 and p_2 gets closer (i.e at smaller proportional dissimilarities), the proposed method yields smaller estimates than the independence model in all settings after accounting for dependency.

Figures 5.4 and 5.5 shows plots of sample size differences between the proposed method (P_{CSS} and P_{ARS}) and dependence model (BB) at different levels of genes (m), π_1, p_1, p_2, sensitivity rate (λ) and power (ϕ). From the plots again, we observe that at wider proportional dissimilarities (between p_1 and p_2), there exist no significant difference in sample size estimates obtained for the proposed method and dependence model. However, as p_1 and p_2 gets closer (i.e at smaller proportional dissimilarities), the proposed method yields smaller estimates than the dependence model in settings where the proportion of differentially associated genes is small after accounting for dependency. At such levels, the effect of increasing power becomes more evident as the proposed method results in more sample size savings.

(a)



(b)

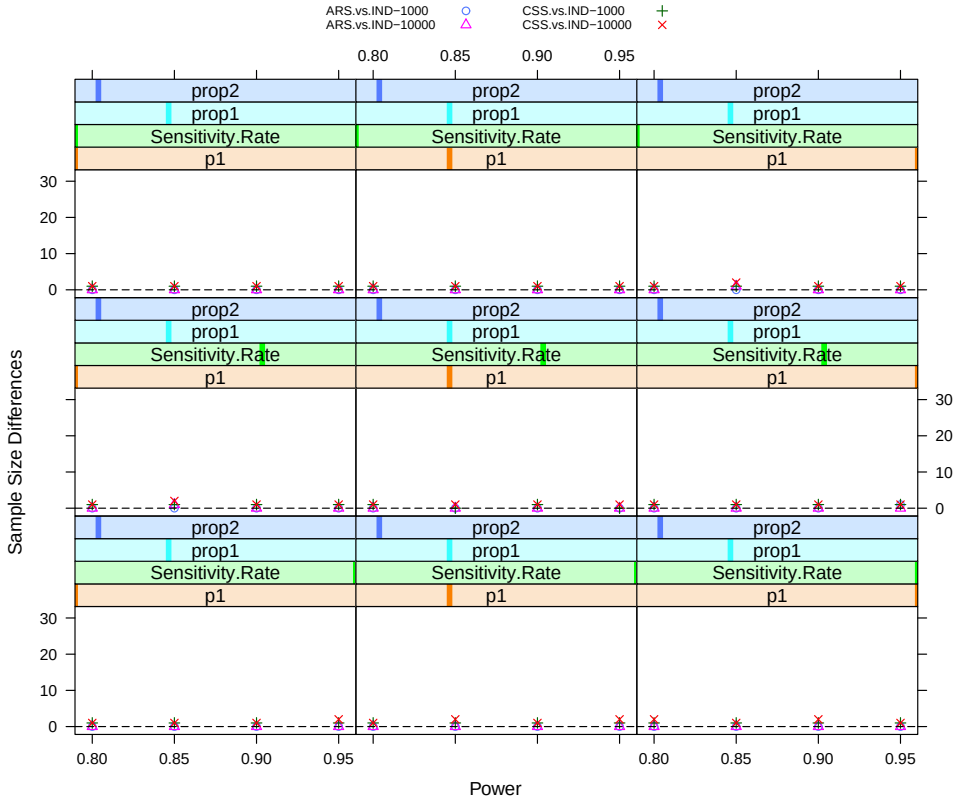
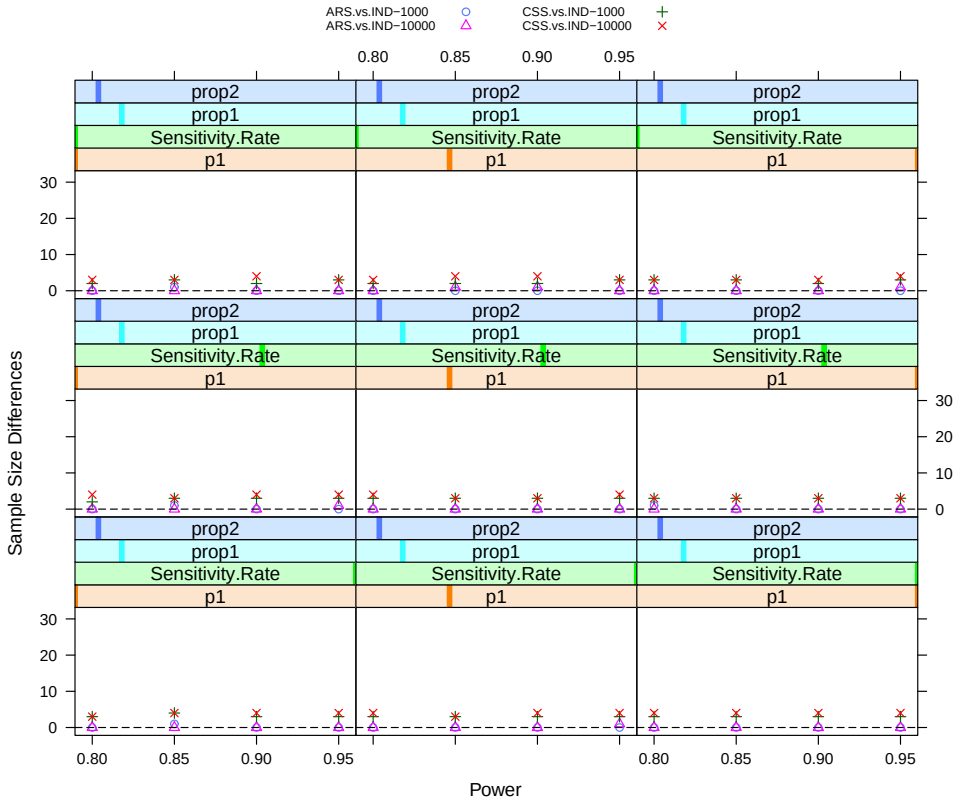


Figure 5.2: (a) & (b) shows effect of factor combinations on sample size difference

(c)



(d)

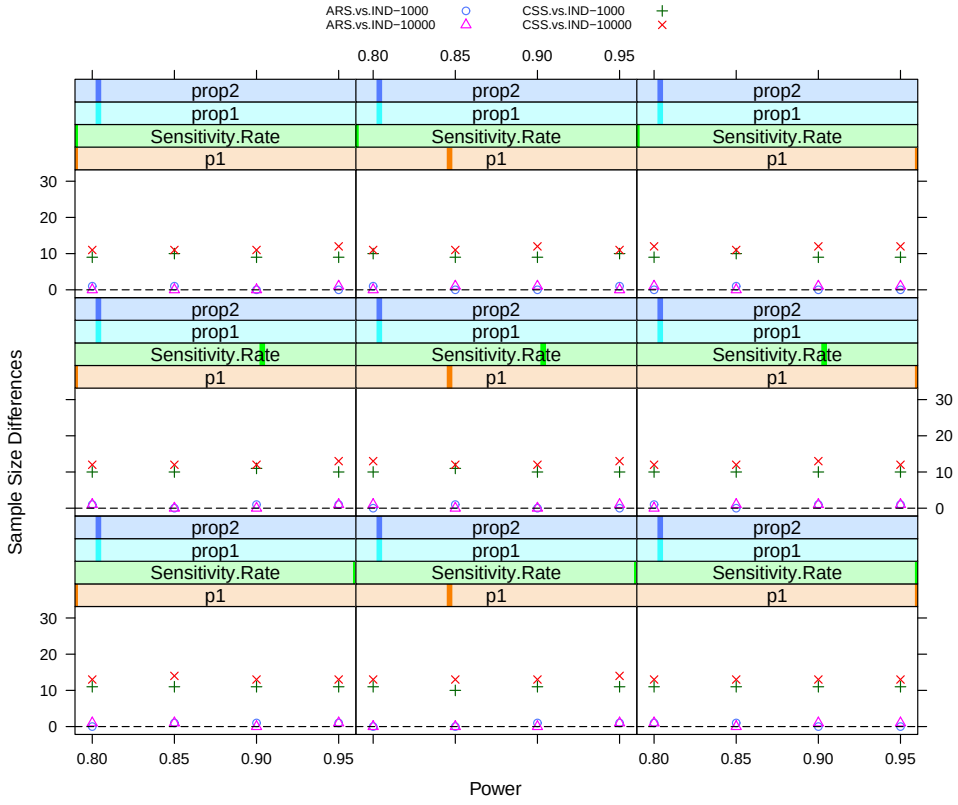
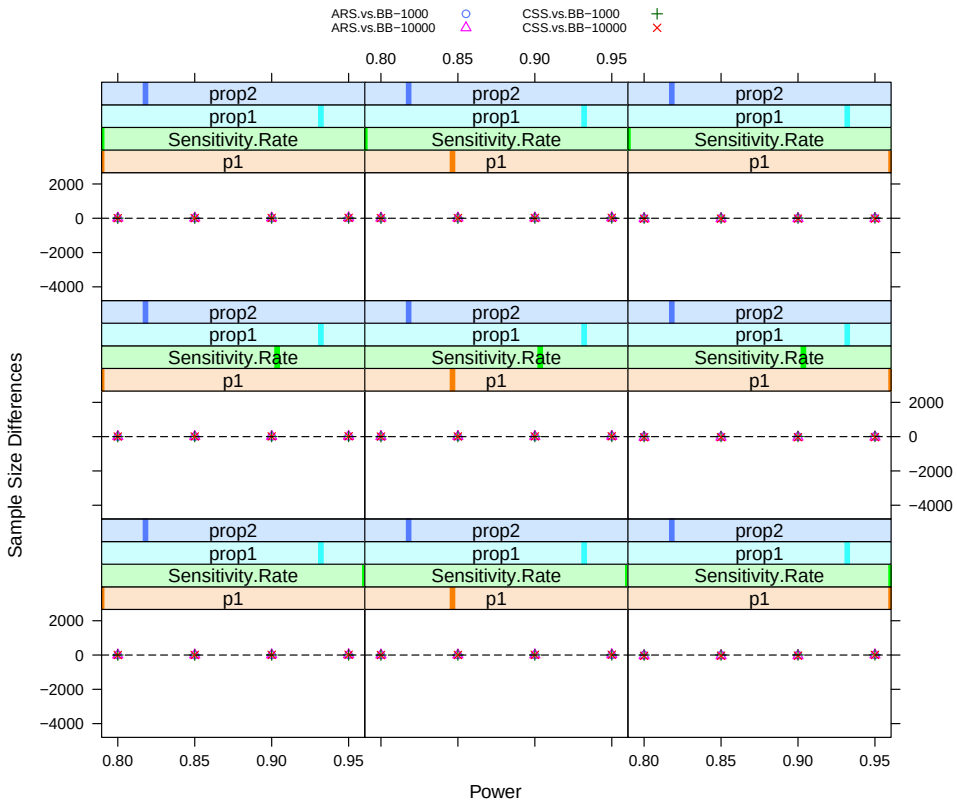


Figure 5.3: (c) & (d) shows effect of factor combinations on sample size difference

(a)



(b)

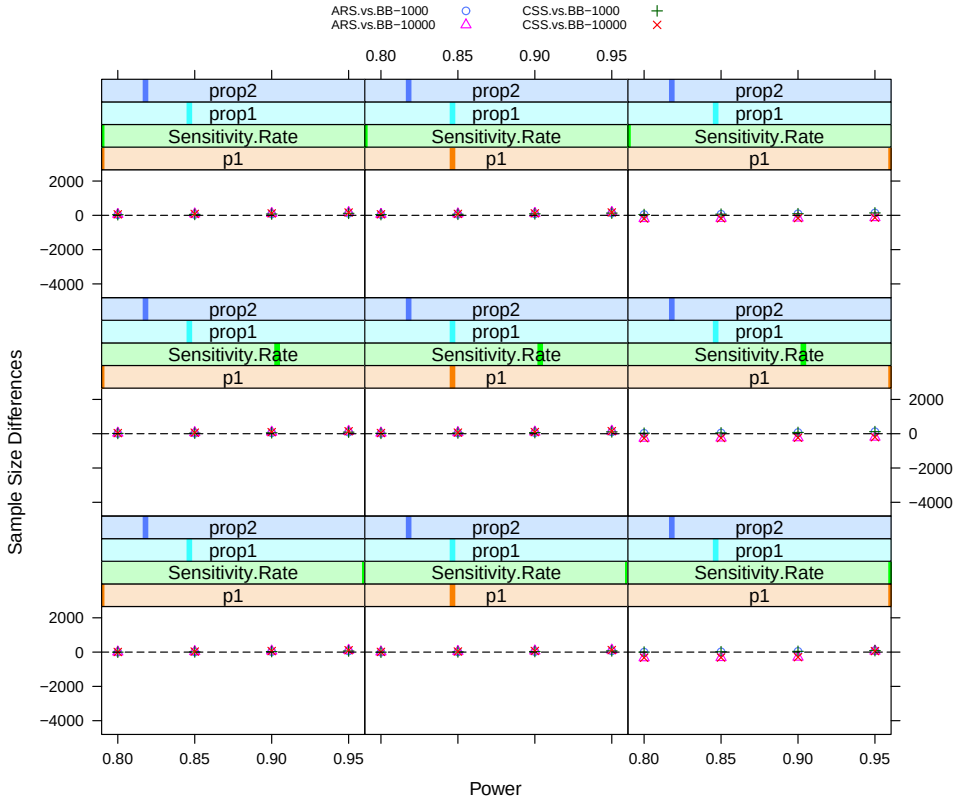
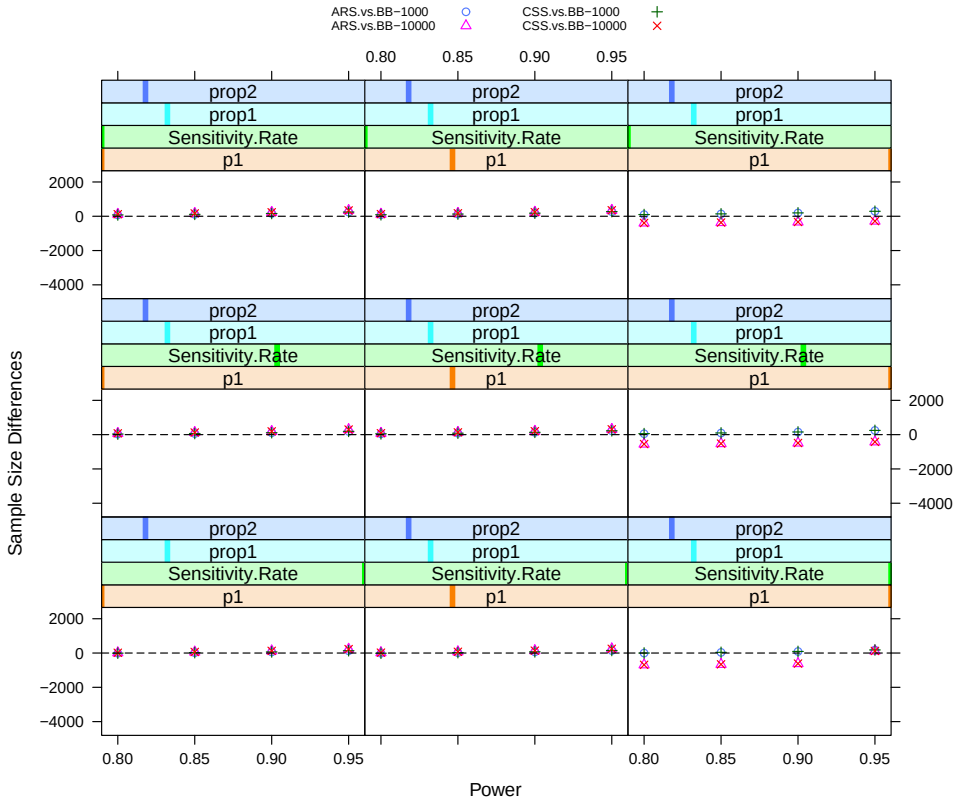


Figure 5.4: (a) & (b) shows effect of factor combinations on sample size difference

(c)



(d)

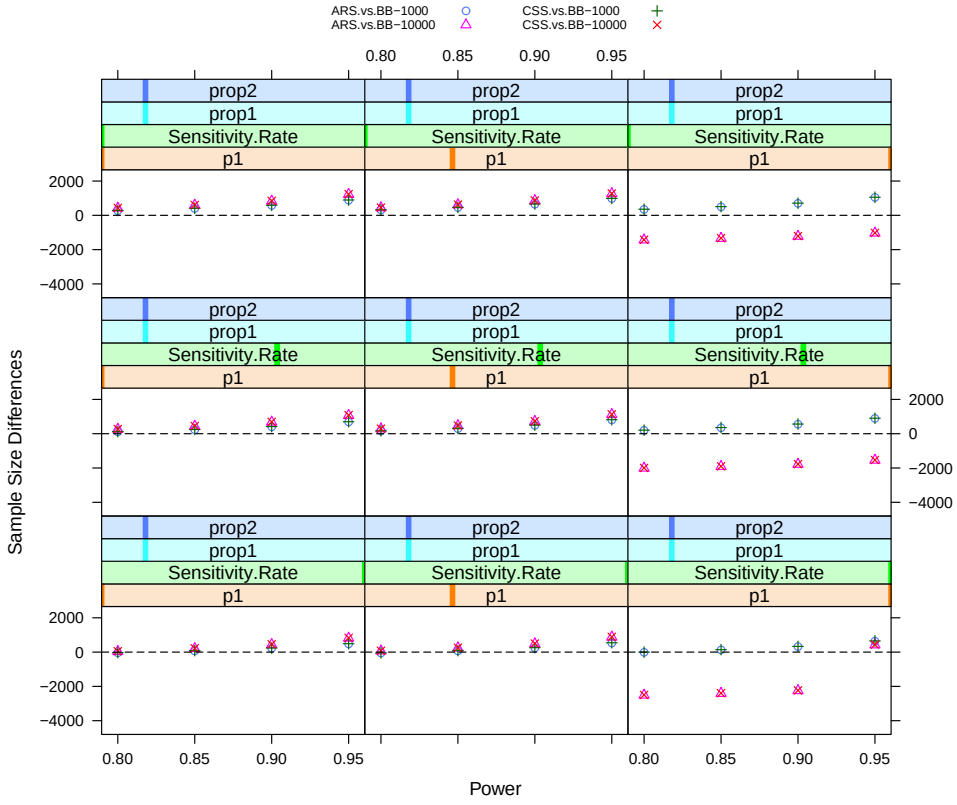


Figure 5.5: (c) & (d) shows effect of factor combinations on sample size difference

5.3 Data Example

For demonstrative purposes, we present the different methods of sample size calculation as applied to a real life dataset. The dataset is an Onco-Miner Pipeline Data about some gene mutations that may or may not be present in cancerous (treatment group) and non cancerous (control group) patients. In all, there are 20 treatment patients and 10 control groups with 33 clinically selected gene mutations. The response variable is binary, measuring the present or absent of a particular gene mutation among the two groups.

Our goal here is to compute sample sizes based on these three methods (i.e. the proposed, independence and dependence methods) and compare results. In the analysis, the association structure of the gene mutations are estimated using the 33 gene mutations. The *polychor* function (*polycor* package) in R is used calculating the sample polychoric correlations between genes. The estimated correlation matrix is used in the proposed method for calculating the $FWER_{\tau}$. We also estimate θ (the intra-cluster correlation) to be 0.334025 by taking the average of all pairwise correlations. Below are the computational results for all methods using an $FWER$ of 0.05, $m = 33$, $\phi = 0.8, 0.85, 0.9, 0.95$, $\lambda = 0.9$, p_1 (treatment) $= 0.5$ and p_2 (control) $= 0.05$.

From Tables 5.13, 5.15 and 5.14 below, we observe that our proposed method yields smaller significant estimates in comparison with the independence and dependence methods. Also, because of the wider proportional dissimilarity between p_1 and p_2, smaller sample sizes are estimated as been required in other to detect the differentially associated genes. The tables' round bracket values are the comparison-wise powers $(1 - \beta)$.

Table 5.13: Proposed Method for $FWER_\tau = 0.06673$ (PCP)

	Familywise Power (ϕ)			
π_1	$\phi = 0.8$	$\phi = 0.85$	$\phi = 0.9$	$\phi = 0.95$
0.05	25	26	29	32
	(0.5528)	(0.6127)	(0.6838)	(0.7764)
0.1	29	31	33	36
	(0.7129)	(0.7556)	(0.8042)	(0.8647)
0.2	37	38	40	43
	(0.8805)	(0.8999)	(0.9212)	(0.9466)

Table 5.14: Dependence Method for $\theta = 0.33403$ (PCP)

	Familywise Power (ϕ)			
π_1	$\phi = 0.8$	$\phi = 0.85$	$\phi = 0.9$	$\phi = 0.95$
0.05	30	32	36	41
	(0.6920)	(0.7552)	(0.8248)	(0.9041)
0.1	32	34	37	42
	(0.7424)	(0.7956)	(0.8538)	(0.9199)
0.2	36	38	41	46
	(0.8417)	(0.8786)	(0.9167)	(0.9568)

PCP : comparison-wise power.

Table 5.15: Independence Method for $\theta = 0.00$ (PCP)

	Familywise Power (ϕ)			
π_1	$\phi = 0.8$	$\phi = 0.85$	$\phi = 0.9$	$\phi = 0.95$
0.05	26	28	30	33
	(0.5528)	(0.6127)	(0.6838)	(0.7764)
0.1	31	32	34	38
	(0.7129)	(0.7556)	(0.8042)	(0.8647)
0.2	38	40	42	45
	(0.8805)	(0.8999)	(0.9212)	(0.9466)

PCP : comparison-wise power.

Chapter 6

Discussion and Conclusion

6.1 Summary

In genomic studies involving sample size and power analyses, the numerous methods that have been proposed to our knowledge come with some restrictions such as assuming equal correlation among differentially associated genes, lack of multiplicity control when test endpoints are non-normally distributed with arbitrary correlation structure, etc. This thesis has led to the development of a novel approach for sample size estimation in genomics settings, taking into account the dependency structure of test endpoints whiles controlling for multiplicity.

From the simulation results, by considering both continuous and binary outcomes in a whole, our proposed method performs relatively better compared to the dependence model in almost all settings. Even though our method yields nearly or exactly the same results as the independence model, the latter ignores dependency even when they exist.

6.2 Recommendations

To an experimenter who seeks the best course of action under situations of sample size estimation in genome-wide association study, we propose the following

1. When small to moderate proportions of genes are differentially expressed or associated (i.e. π_1 at low and moderate levels), capitalizing on the dependency structure using the proposed method yields more gains.

2. Also, when a large proportion of genes are differentially expressed or associated (i.e. π_1 and m at high levels), capitalizing on the dependency structure would only yield gains if higher power and sensitivity are desire.

References

[1] Gary W. Oehlert, University of Minnesota. *A First Course in Design and Analysis of Experiments*, 2010.

[2] García-Arenzana, N., E.M. Navarrete-Muñoz, V. Lope, P. Moreo, S. Laso-Pablos, N. Ascunce, F. Casanova-Gómez, C. Sánchez-Contador, C. Santamariña, N. Aragonés, B.P. Gómez, J. Vioque, and M. Pollán. *Calorie intake, olive oil consumption and mammographic density among Spanish women. International journal of cancer 134: 1916-1925.*, 2014.

[3] Altham, P.M.E. (1978). *Two generalizations of the binomial distribution.* Applied Statistics 27, 162-167.

[4] Bahadur, R.R. (1961). *A representation of the joint distribution of responses to n dichotomous items. In studies in Item Analysis and Prediction, H. Solomon (ed.)*, Stanford University Press, Stanford, California.

[5] Kupper, L.L. and Haseman, J.K. (1978). *The use of a correlated binomial model for the analysis of certain toxicological experiments.* Biometrics 34, 69-76.

[6] Casella, G and Berger, R (2001). *Statistical Inference.* Brooks/Cole, Boston, MA, USA, second edition, pg 188-194.

[7] Tsai, C. A., Wang, S. J., Chen, D. T., & Chen, J. J. (2005). *Sample size for gene expression microarray experiments.* Bioinformatics, 21(8), 1502-1508.

[8] Casagrande, J. T., Pike, M. C., & Smith, P. G. (1978). *An improved approximate formula for calculating sample sizes for comparing two binomial distributions.* Biometrics, 483-486.

[9] Dunnett, C. W., & Sobel, M. (1954). *A bivariate generalization of Student's t-distribution, with tables for certain special cases.* Biometrika, 41(1-2), 153-169.

[10] Wagler, A. & McCann, M. (in review). *A multiplicity adjustment for chi-square distributed endpoints.* In review at Journal of Statistical Computation and Simulation.

Appendix

R CODES

```
# Load packages & set working directory

# -------------------------------------

library(igraph)
library(cubature)
library("ape")
library("RBGL")
#library("graph")
library("MSBVAR")
library("dlm")
library("Matrix")
library("nlme")
library(MASS)
library("mvtnorm")

options(digits=22)
setwd("C:/Users/dkoomson/Desktop/New_SampSize/ARS")



#require("TailRank")
#require(Biobase)
```

```
#library("gmp")


# BINARY RESPONSE
# ------------------------------------------------

# ================================================
# Numerical Integraion Functions
# ================================================


# calculate HW prob points

f.chi=function(d,m,rho) {#m=100;d=1
  d=c(d,d)
  j=seq(0,150,by=1)
  part1=as.double((1-rho^2)^(-(m/2))/(gamma(m/2)*2^(m)))
  part2=as.double(lgamma((m/2+j))+log(factorial(j)))
  part3=as.double(log((rho/(2*(1-rho^2)))^(2*j)))
  part4=as.double(((log(d[1])+log(d[2]))*(m/2+j-1)))
  part5=as.double(exp(-(d[1]+d[2])/(2*(1-rho^2))))
  temp2=as.double((part3)+part4-part2+log(part5))
  p.t=as.double(part1*sum(exp(temp2)))
  return(p.t)
}

F.c=function(d,m,rho) {
  j=seq(0,30,by=1)
```

```
  part1=(1-rho^2)^(m/2)
  tempsum=0
  tempsumr=0
  part2.1=gamma((m/2)+j)
  part2.2=factorial(j)*gamma(m/2)
  part2.3=rho^(2*(j))
  part2.4=part2.1/part2.2
  part2=(part2.4*part2.3)^2
  d.s=d/(2*(1-rho^2))
  part3=pgamma(d.s/(1-rho^2),(m/2)+j)
  part4=pgamma(d.s/(1-rho^2),(m/2)+j)
  part5=part3*part4
  part6=part2*part5
  p.t=part1*sum(part2*part3*part4);p.t
  r.t=1-part1*sum(part2);r.t
  F=p.t+r.t
  return(F)
}



secant <- function(fun, x0, x1, tol=1e-04, niter=500){
  for ( i in 1:niter ) {
    x2 <- x1-fun(x1)*(x1-x0)/(fun(x1)-fun(x0))
    if (abs(fun(x2)) < tol)
      return(x2)
    x0 <- x1
    x1 <- x2
  }
```

```
  stop("exceeded allowed number of iteractions")
}#k;n;nu;rho;h


hw.chisq=function(V,nu,l,u,alpha){
  #V=v
  k=nrow(V)
  R=cov2cor(V)
  acosR<-acos(R)

  ####### Find minimum spanning tree ####################
  dis<-graph.adjacency(acosR,mode="max",weighted=TRUE)
  mst=minimum.spanning.tree(dis)
  #plot(mst)
  phi=E(mst)$weight
  F.c2=Vectorize(F.c,vectorize.args="rho")
  ineq=function(d){return(alpha-(k)*(1-pchisq(d,nu))+
      sum(1-2*pchisq(d,nu)+F.c2(d,nu,cos(phi)/(2*pi))))}

  ############### root finding secant method ##############################

  crit=secant(ineq, x0=l, x1=u)
  return(list(Pointwise=l,hw=crit,Bonferroni=u,phi=phi))}
```

```
# ================================================
# Sample Size Simulation Functions
# ================================================


# fhi is the familywise power under the binomial model
# x is the minimum number of TRUE DISCOVERIES
# m1 is the number of differentially associated genes (FALSE NULLS)
# thau is sensitivity rate
# prop1 & prop2 are the proportions of differentially associated
# genes in the two groups (treatment and control).


myfunc2SR <- function (m,p1,fhi,thau,bita) {
  m1<-(p1*m);
  x<-min(c(floor(m1*thau),m1),na.rm=TRUE);  ### Sensitivity rate (thau)
  fwise<-fhi-pbinom(x-1,m1,1-bita,lower.tail = FALSE)
}


outA <- function (out2,out3,out4,out5,prop1,prop2,cfac,fwer_adj,alpha1,alpha2) {

  bita1<-uniroot(myfunc2SR,interval = c(0, 1),m=out2,p1=out3,fhi=out4,thau=out5);

  pbar=(prop1+prop2)/2
  A1=((qnorm(1-(alpha1/2))*sqrt(2*pbar*(1-pbar)))+
        (qnorm(1-bita1$root)*sqrt((prop1*(1-prop1))+(prop2*(1-prop2)))))^2;
```

```
  size1<-(A1*(1+sqrt(1+(4*(1-(2*cfac))*
          abs(prop1-prop2))/A1))^2)/(4*(prop1-prop2)^2)

  A2=((qnorm(1-(alpha2/2))*sqrt(2*pbar*(1-pbar)))+
        (qnorm(1-bita1$root)*sqrt((prop1*(1-prop1))+(prop2*(1-prop2)))))^2;
  size2<-(A2*(1+sqrt(1+(4*(1-(2*cfac))*abs(prop1-prop2))/A2))^2)/(4*(prop1-prop2)^2)

  return(matrix(c(1-bita1$root,size1,size2),ncol=3,
                dimnames=list("Value",c("CW_POWER","SS1","SS2"))))
}

sampout<-function (Xmat) {

  for (i in (1:nrow(Xmat))){
    x2<-Xmat[i,10];x3<-Xmat[i,11];x4<-Xmat[i,12];x5<-Xmat[i,13];
    x6<-Xmat[i,14];x7<-Xmat[i,15];x8<-Xmat[i,16];x9<-Xmat[i,17]
    x10<-Xmat[i,18];x11<-Xmat[i,19]
    matnew<-try(outA(out2=x2,out3=x3,out4=x4,out5=x5,prop1=x6,prop2=x7,cfac=x8,
                     fwer_adj=x9,alpha1=x10,alpha2=x11))
    Xmat[i,c("CW_POWER","SS1","SS2")]<-matnew[1:3]
  }
  return(Xmat)
}


#mean.res=NULL
# p=1/beta and beta spans 0 to 1 and hence p>1 - p=1
# is a complettely non-sparse sequence
```

```
# for n=2^10, larger p yields very sparse sequences-values between p=5/4 (25% non-zero)
# 6/4 (9.96% non-zero) 7/4 (5%), 8/4 (3.13%), 9/4 (2.15%), 10/4 (1.56%)


master1<-function (pl,rhol,nul,fwerl,nl,p1l,fhil,thaul,prop1l,prop2l,cfacl){

  #   nl,p1l,fhil,thaul,deltal are all vectors
  #   ml is levels of the number test hypotheses/genes
  #   p1l is levels of the proportion of False Nulls
  #   fhil is levels of desired family-wise power
  #   thaul is levels of sensitivity rates
  #   deltal is levels of standardized effect size

for (p in pl){#p=5/4
#,6/4,7/4,8/4,9/4,10/4)){
(beta=1/p)
#rho=0,.5,.9 perhaps?

for (rho in rhol){#rho=0
for (nu in nul){#nu=1
for (alp in fwerl) {
alpha=alp
#vary n from 2^9 to 2^12??

for (n in nl){#nl=10^3,n=2^9
sig=1
#also run nu=10
k=1:n
```

```
(n0=round(n**(1/p)))
#for FWERk control I am assuming the rounded value of 1% of n
(s=min(round(.01*n),n0))

reps=(length(p1l)*length(fhil)*length(thaul)*
      length(prop1l)*length(prop2l)*length(cfacl))
master.res=matrix(NA,nrow=reps,ncol=9)

for (i in 1:reps){ #i=1
#generate weak lp ball vector
m.vec=(n0*k^(-1/p))
mu=c(rep(mean(m.vec[1:n0]),times=n0),rep(0,times=n-n0))

#generate covariance structure
times=seq(1,(2*n),2)
H <- abs(outer(times, times, "-")); ## auto regressive structure(ARS)
#H=matrix(1,nrow=n,ncol=n);diag(H)=0  ## compound symmetric structure (CSS)

#this is the correlation matrix generated
v <-  rho^H
#cv=chol(t(cov2cor(sig*v)))
#X=m.vec+cv^2%*%matrix(rchisq(ncol(v),df=nu,ncp=mu),nrow=n,ncol=1)
Z=rmvnorm(nu,mean=rep(0,nrow(v)),sigma=v,method="chol")
dim(Z)=c(nrow(v),nu)
sumsq=function(x) {sum(x^2)}
X=apply(Z,c(1),sumsq)

#generate critical value HW
```

```
l=as.double(qchisq(1-alpha,nu));l
u=as.double(qchisq(1-alpha/n,nu));u
options(digits=22)
comps=hw.chisq(v,nu,l,u,alpha)
(crit=comps$hw)

#compute augmented significance level
pval=sort(1-pchisq(sort(X,decreasing=T),nu),decreasing=T)
(a.adj=(nrow(v)*(1-pchisq(crit,nu))))

# record results-obs sig levels should be close to n0/n, ideally
# this just saves the proportion of any signal
# detections for ONE iteration of the simulations

master.res[i,]=c(n=n,s=s,rho=rho,n0=n0,p=p,nu=nu,
               sig=sig,true=n0/n,q=a.adj)
            }

colnames(master.res)=c("n","s","rho","n0","p","nu","sig","true","FWER_adj")

mymat3<-matrix(nrow=reps,ncol=10)
colnames(mymat3)<-c("m","p1","FW-Power","Sensitivity Rate",
                "prop1","prop2","cfactor","FWER_adj","alpha1","alpha2")
i=0
m<-n
for (p1 in p1l){
m1<-(p1*m);
#         m=1000; m1=50
```

```
# number of false rejections divided by
# the number of expected rejections E(R)=m1
# a.adj is FWER

alpha1=a.adj/(m-m1) #(m*a.adj/m1)/(m-m1) ## using bonferronni
alpha2=-log(1-a.adj)/(m-m1) ## using standard approx for v

for (a1 in fhil) {
for (a2 in thaul) {
for (prop1 in prop1l) {
for (prop2 in prop2l) {
for (cfac in cfacl) {
mymat3[i+1,1:ncol(mymat3)]<-c(m,p1,a1,a2,prop1,prop2,
                              cfac,a.adj,alpha1,alpha2)
i=i+1
}}}}}}


master.res1<-cbind(master.res,mymat3)
CW_POWER<-NA;SS1<-NA;SS2<-NA;
master.res1<-data.frame(master.res1,CW_POWER,SS1,SS2)
master.res1<-sampout(Xmat=master.res1)
write.csv(master.res1,
          paste("n=",n,"rho=",rho,"p=",p,"nu=",nu,"sig=",sig,".csv",sep=""))

}}}}}}
```

```
pl=c(8/4)
rhol=c(0.1,0.22,.5,.9)
nul=c(1)
nl=c(10^3, 10^4) # c(2^9,2^11)
fwerl=c(0.05)

p1l<-c(0.05,0.1,0.2)
fhil<-c(0.8,0.85,0.90,0.95)
thaul<-c(0.8,0.9,0.95) ######### sensitivity rates
prop1l<-seq(0.05,0.8,0.05)
prop2l<-seq(0.05,0.8,0.05)
cfacl<-c(0)

master1(pl,rhol,nul,fwerl,nl,p1l,fhil,thaul,prop1l,prop2l,cfacl)



# CONTINUOUS RESPONSE
# ------------------------------------------------

#################### Hunter-Worsley ###################

hw<-function(C,alpha,nu,V){
  #V=V_hat
  R=cov2cor(C%*%V%*%t(C))
  acosR<-acos(R)

####### Step One: Find minimum spanning tree ####################
```

```
dis<-graph.adjacency(acosR,mode="max",weighted=TRUE)
mst <- minimum.spanning.tree(dis)
lay <- layout.reingold.tilford(dis, mode="all")
#plot(mst, layout=lay)



####### Step Two: Numerical Integration Method######
p<-nrow(C)
r=qr(C)$rank

intg<- function(d)
  {

fx <- function(x){

  B<-p*pf(((d*x)^(-2)-1)/(r-1), r-1, 1)

  if (r==2) {for (i in 1:(p-1)) {
    phi<-E(mst)$weight[i]
    B<-B-max(-phi/pi+2*acos(x*d)/pi,0)}
}

else {
for (i in 1:(p-1)) {
phi<-E(mst)$weight[i]

gx<-function(w)
```

```
{acos(d*x*sqrt(1/(1-w)))*w^(r/2-2)}
#print(d)
gx<-Vectorize(gx)

if (((cos(phi/2)/(x*d))^2-1) >0 ) {
B<-B+phi/pi*pf(2*max(0,(cos(phi/2)/(x*d))^2-1)/
  (r-2),r-2,2)-(r-2)/pi*(integrate(gx,
  0,1-1/(1+max(0,(cos(phi/2)/(x*d))^2-1)))$value)  }

else B<-B+phi/pi*pf(2*max(0,(cos(phi/2)/(x*d))^2-1)/(r-2),r-2,2)

 }
}

return (B*df(r*x^2, nu, r)*2*r*x)

}

fx<-Vectorize(fx)
gf <- function(k) {integrate(fx, lower=0, upper=1/k)$value}
#d=qt(1-.05,nu);d=sqrt(r*qf(1-alpha,r, nu))
return (gf(d)-alpha)}

######### Step Three:  Root Finding Algorithm #############
secant <- function(fun, x0, x1, tol=1e-4, niter=500){
for ( i in 1:niter ) {#fun=intg
  x2 <- x1-fun(x1)*(x1-x0)/(fun(x1)-fun(x0))
if (abs(fun(x2)) < tol)
```

```
  return(x2)
  x0 <- x1
  x1 <- x2
  print(c(x0,x1,x2))
  }
stop("exceeded allowed number of iteractions")

}

  secant(intg, x0=qt(1-alpha/2,nu), x1=qt(1-alpha/(2*choose(k,2)),nu))
}

# ================================================
# Sample Size Simulation Functions
# ================================================

outA <- function (out2,out3,out4,out5,delta,alpha1,alpha2) {

  bita1<-uniroot(myfunc2SR,
                 interval = c(0, 1),m=out2,p1=out3,fhi=out4,thau=out5);
  # alpha<-out1/(out2*(1-out3)); #change to simulation code's alpha2 & alpha2 values
  size1<- (2*((qnorm(1-bita1$root)-qnorm(alpha1/2))^2))/((delta)^2); #alpha1 sample size
  size2<- (2*((qnorm(1-bita1$root)-qnorm(alpha2/2))^2))/((delta)^2); #alpha2 sample size
  return(matrix(c(1-bita1$root,size1,size2),ncol=3,
                dimnames=list("Value",c("CW-POWER","SS1","SS2"))))
}
# The rest of the functions mimics that in the binary response case.
```

Curriculum Vitae

Desmond Koomson, born on March 15, 1989 is the second son of Daniel Kwesi Koomson and Elizabeth Arthur. After completing Pope John's Senior High School, Koforidua in June 2007, he continued his college education a year later in Kwame Nkrumah University of Science and Technology (KNUST), Kumasi where he pursued a bachelor's degree in Actuarial Science, graduating with First Class Honors. At KNUST, Desmond showed active involvement in student fellowships which resulted in him serving as a student leader in various positions.

In Fall 2014, he entered the Graduate School of The University of Texas at El Paso (UTEP) as a graduate student pursuing a master's degree in Statistics. While in UTEP, he worked as a Teaching Assistant until Fall 2015 when he became a Research Assistant. He is currently working under the supervision and mentorship of Dr. Amy Wagler, conducting research in Sample Size Estimation for Genomics Experiments with Dependent End Points. He plans to continue his studies in Ph.D Computational Science program at The University of Texas at El Paso in the near future.

Present address: 806 West Yandel Dr, Apt # 7,
El Paso, Texas 79902.