Spring 3-28-2017

Contributions to Statistical Testing, Prediction, and Modeling

John C. Pesko
University of New Mexico

John Carl Pesko

Candidate

Mathematics and Statistics

Department

This dissertation is approved, and it is acceptable in quality and form for publication:

Approved by the Dissertation Committee:

Guoyi Zhang, Chair

Ronald Christensen, Co-Chair

Huining Kang

Yan Lu

Contributions to Statistical Testing, Prediction, and Modeling

by

John Carl Pesko

B.S., Statistics, University of New Mexico, 2011
B.A., Russian, University of New Mexico, 2011
B.A., English, University of New Mexico, 2011
M.S., Statistics, University of New Mexico, 2013

DISSERTATION

Submitted in Partial Fulfillment of the
Requirements for the Degree of

Doctor of Philosophy
Statistics

The University of New Mexico

Albuquerque, New Mexico

May, 2017

Dedication

To my parents and my twin sister.

"The lyf so short, the craft so long to lerne" – Chaucer

Acknowledgments

I would like to express my gratitude toward my advisor, Dr. Guoyi Zhang. Dr. Zhang was a font of unending encouragement and guidance, always eager to discuss my research and future plans each step of the way. I'll forever appreciate his forthrightness and understanding throughout this arduous and rewarding process.

Of course, I am also beholden to my co-advisor Dr. Ron Christensen, whose frankness and wealth of experience proved as challenging as it was enlightening. Ron is a paragon in Statistics with an inspiringly deep knowledge of the philosophy behind the methodology, and his supreme humor made working together endlessly enjoyable.

I'd also like to thank Dr. Huining Kang, who has bookended my academic career as one of my first professors and my current supervisor at the UNM Cancer Research Facility. I've always appreciated his affable demeanor and our recent work together has given me confidence in my ability to perform research while cultivating in me the work ethic needed of a Biostatistician in an industrial setting. While guiding my work investigating methods for high-throughput biomarker evaluation under the case-cohort study design, Dr. Kang has imparted his programming wizardry upon me while serving as an ideal supervisor—his wealth of knowledge commands my respect, and his kindness earns my trust.

Finally, I'll express my appreciation for Dr. Lu, the last member of my advisory committee and my favorite teacher throughout the years. Dr. Lu made the daunting task of graduate school much less stressful as I've always been able to rely on her to support my scholastic and professional development. Always a pleasure to work with, Dr. Lu taught me the foundations of classical statistical theory as well as introduced me to more modern algorithmic approaches that profoundly impacted my philosophy toward data analysis and science as a whole.

I'll close by thanking Dr. Erik Erhardt for the valuable opportunities he afforded me throughout my academic career and my fellow graduate students—especially Dr. Lang Zhou and Dr. Fares Qeadan—for their emotional and academic support while we mired through the scholastic jungles, and of course my parents for supporting me from the moment they wrought me into existence, my twin sister Sara Pesko for being by my side before I could remember, and those countless others who, in the interest of brevity cannot be named here, but will always have a place in my heart.

Contributions to Statistical Testing, Prediction, and Modeling

by

John Carl Pesko

B.S., Statistics, University of New Mexico, 2011

B.A., Russian, University of New Mexico, 2011

B.A., English, University of New Mexico, 2011

M.S., Statistics, University of New Mexico, 2013

Ph.D, Statistics, University of New Mexico, 2017

Abstract

This dissertation consists of three parts that chronicle my major research as a Statistics Ph.D Candidate at the University of New Mexico. First, I present my primary research in the UNM Department of Mathematics & Statistics, which delves into the relationship between parametric bootstrap and objective Bayesian approaches to significance testing, and includes an in-depth examination of the heteroscedastic analysis of variance problem. For the one-way problem, we present tests based on the parametric bootstrap, objective Bayes, the predictive distribution, and an unweighted test statistic. These approaches are compared theoretically, with simulation studies, and with a real data application. The findings of the one-way case is extended to testing for differences in means in the RCBD with subsampling and heteroscedastic errors model. We establish variance parameter estimates, propose an objective Bayesian test, and a new unweighted test for fixed group effects. We derive the asymptotic distribution for which to compare the unweighted test statistic, conduct a simulation study, and show how to solve an applied problem using the objective Bayesian method. Lastly, we look at some general results pertaining to Bayesian significance testing, defining a Bayesian p-value and describing some of its properties.

The second part involves my work with the UNM Department of Neurology and discusses the development of a random forest algorithm for the early detection of patients with Binswanger's disease, a subgroup of vascular cognitive impairment dementia. We use cross validation to compare several methods for predicting if vascular dementia patients are of the Binswanger type or if they more likely suffer from some other small vessel disease. We investigate which biomarkers are most important for classification and see that a random forest algorithm accurately identifies Binswanger's patients earlier than clinicians are able to, which can reduce the number of patients needed for a clinical trial while improving the chance of success.

The dissertation concludes with my work in Statistical Genomics from my time at the UNM Cancer Research Facility, which involves an examination of methods for high-throughput gene expression analysis under the case-cohort study design. The case-cohort study design blends the efficiency of case control studies with the philosophical soundness of full cohort studies, and is an efficient way to analyze survival data, particularly for large cohorts with low failure rates. Using a tandem of real data examples and simulation studies, we investigate the performance of the most popular case-cohort analysis approaches in the context of high-dimensional biomarker evaluation.

Contents

List of Figures

List of Tables

Part I

Parametric Bootstrap and Objective Bayesian Testing with Applications to Heteroscedastic ANOVA

Chapter 1

Introduction

This research places a major emphasis on linear models, which are mathematical models that are linear in the parameters. The standard linear model is

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{e}, \qquad \mathrm{E}[\boldsymbol{e}] = \mathbf{0}$$

where $\boldsymbol{Y}$ is a vector of observations of a random variable, $\boldsymbol{X}$ is a known matrix that specifies the model, $\boldsymbol{\beta}$ is a vector of unknown fixed parameters that we want to estimate, and $\boldsymbol{e}$ is a vector of unobservable error terms. Linear models are used to model a non-deterministic process or phenomena in such a way that we can understand the relationships between variables involved and predict future outcomes. The assumptions behind this model are that we have no systematic bias, so that $\mathrm{E}[\boldsymbol{e}] = \mathbf{0}$, i.e. the expected value of the error terms is zero.

Analysis of variance (ANOVA) is perhaps the most commonly applied procedure in the linear modeling framework. Any time we want to predict a continuous response using categorical predictors, we have an ANOVA problem. Whether we are comparing two groups with a t test or using a more complicated design such as a randomized complete block design (RCBD), they can all be viewed as special cases of the ANOVA problem. To compare groups, we have parameters representing group means and the $\boldsymbol{X}$ matrix indicates the group each observation is in.

In many cases, we can assume normally distributed and uncorrelated error terms uncorrelated with constant variance such that $\text{Cov}(\boldsymbol{e}) = \sigma^2\boldsymbol{I}$, where $\boldsymbol{I}$ is the identity matrix. Under these conditions the ordinary least squares (OLS) procedure gives known distributional results.

To provide more flexibility for modeling less well-behaved data, we need to loosen these strict assumptions. Our primary focus will be a consideration heteroscedastic data. For ANOVA problems, heteroscedasticity means that each group has a different variance term. We allow for an unbalanced design where we have different sample sizes per group, giving us an unbalanced heteroscedastic ANOVA model.

For the one-way heteroscedastic ANOVA model, we have:

$y_{ij} = \mu_i + \epsilon_{ij}$	the linear model
y_{ij}	independent observations
$i = 1, 2, \ldots, a$	groups
$j = 1, 2, \ldots, n_i$	observations within a group
$e_{ij} \sim N(0, \sigma_i^2)$	normally distributed errors
$N = \sum_{i=1}^{a} n_i$	total observations
μ_i	population means
$\bar{y}_{i.} = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}$	sample means
σ_i^2	population variances for each group
$s_i^2 = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i.})^2$	sample variances for each group

Our goal is to perform a significance test of the hypothesis:

$$\text{H}_0 : \mu_1 = \mu_2 = \cdots = \mu_a$$

This is a problem with a rich history, and many competing approaches have been proposed. Chapter 2 includes a literature review to help establish some of the key topics related to this problem.

In Chapter 3 we explore the one-way heterANOVA problem in depth, comparing tests based on the parametric bootstrap (PB), objective Bayes (OB), the predictive distribution, and an unweighted test statistic. We establish conditions for which the PB and OB tests are equivalent, compare how the methods approximate the distribution of the test statistic, evaluate their performance with a simulation study, and show how to solve a real data problem.

Chapter 4 investigates the randomized complete block design (RCBD) with subsampling when there are heteroscedastic error terms. In this chapter, we develop an OB test and an unweighted test for testing for a difference in group means. We present the results of a simulation study that compares the type I error rate achieved by the OB test to the chi-squared test that is used when the variance terms are known as an upper bound to how well we can do. A real data example is presented, and we derive the asymptotic distribution of the unweighted test statistic.

Chapter 5 focuses on general Bayesian significance testing. We propose a general Bayesian p-value for significance testing and discuss its properties. We close this research in Chapter 6 with a brief synopsis and proposal of future research paths.

Chapter 2

Literature Review

2.1 Significance Testing

Christensen (2005) describes Fisherian significance testing as a probabilistic "proof by contradiction" in which a model is proposed and we use observed data to "examine the extent to which the data contradict the model". If the data we observe is highly unlikely under the assumed model, it implies that the null model is incorrect. In significance testing, the p-value represents the probability under the null model of observing a test statistic "as weird or weirder than you actually saw". Sufficiently low p-values lead us to conclude that we've observed enough evidence to suggest that the initial assumptions are incorrect, in which case we reject the null model.

Researchers often posit a parametric hypothesis, such as $\mathrm{H}_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$, that the true population parameter $\boldsymbol{\theta}$ is equal to some proposed value $\boldsymbol{\theta}_0$. A test statistic $T(\boldsymbol{Y})$ is calculated and compared to some reference distribution, and if a low p-value is found, people often assume that a rejection of H_0 is the same as saying there is evidence that $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$. While this is one reason for a low p-value, model misspecification is another. Examples of model misspecification are assuming normality when it isn't reasonable or assuming homoscedasticity when it is violated. For linear models, we can look at a plot of the residuals to assess these assumptions. If they are upheld, we can reasonably conclude that $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$, but when the assumptions appear inappropriate, we are forced to rethink everything we are doing.

If we have a test statistic that does not follow a tractable distribution, obtaining a p-value is not straightforward. This often occurs when nuisance parameters are present. Nuisance parameters are parameters that we are not interested in doing inference on, but cannot be ignored. The one-sample t test features a nuisance parameter in the unknown σ^2. Although we wish to test a mean μ, because we don't know the variance σ^2 we cannot use the normal distribution. Instead, we use a pivotal quantity $t = \frac{\bar{y}-\mu_0}{s\sqrt{n}}$ which follows a t_{n-1} distribution and does not depend on the unknown variance parameter. The heteroscedastic ANOVA problems also result in a test statistic that doesn't follow a known distribution, and the p-value resulting from the classical tests depend on nuisance parameters, so we have to find a workaround.

2.2 The Behrens-Fisher Problem

The Behrens-Fisher problem is another name for the two-sample t test when the groups have different variance terms. When these terms are equal, the classical test statistic follows a t distribution, a result that is not guaranteed under heteroscedasticity. In the usual two-sample t test, the assumptions of normality, constant variance, and independence are all part of the null model. When there is evidence contradicting the equal variance assumption, we can reject the null either because the means or the variances are different. A better interpretation is that we are testing whether the data comes from the same normal population.

Welch (1938) proposed an approximate answer to the Behrens-Fisher problem.

This solution has eclipsed the classical homoscedastic t test as the default approach since it is better under heteroscedasticity and the same in the balanced case when variance terms are equal. Welch's observed test statistic is:

$$t_W = \frac{\bar{y}_1 - \bar{y}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

where the difference in means follows a normal distribution. In the balanced case, t_W follows a t distribution with degrees of freedom $df = N - 2$. For the general case, Welch proposes a normal approximation when samples are large, but in the small sample case uses the Satterthwaite approximation (Satterthwaite, 1946) for the degrees of freedom of the t distribution, so $df = \frac{\left(\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1-1}\left(\frac{s_1^2}{n_1}\right)^2+\frac{1}{n_2-1}\left(\frac{s_2^2}{n_2}\right)^2}$.

2.3 Heteroscedastic One-Way ANOVA

In one-way ANOVA problems we are interested in testing if group means are different. Our null hypothesis is $\mathrm{H}_0 : \mu_1 = \mu_2 = \cdots = \mu_a$, and assumes independent and normally-distributed error terms. In the homoscedastic case, OLS gives an F test that looks at the proportion of variability explained by the differences in groups compared to the remaining variability:

$$\begin{aligned}
\text{MSGrps} &= \frac{\sum_{i=1}^{a} n_i(\bar{y}_{i.} - \bar{y}_{..})^2}{a-1} \\
MSE &= \frac{\sum_{i=1}^{a}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i.})^2}{N-a} \\
F_{Obs} &= \frac{\text{MSGrps}}{\text{MSE}} \\
F_{Obs} &\sim F_{a-1,N-a}
\end{aligned}$$

The homoscedastic F test performs poorly for heterANOVA, especially for small samples. Generalized least squares (GLS) loosens the constant variance assumption, and allows for an arbitrary covariance matrix for the error terms, i.e. $\text{Cov}(\boldsymbol{e}) = \boldsymbol{V}$.

Chapter 3, Section 2 gives the details of the test. Under H$_0$, the weighted Wald-style test statistic features group means being weighted by the inverse of their standard error terms. When the σ_i^2's are known, we get:

$$T = \sum_{i=1}^{a} \frac{n_i}{\sigma_i^2} \bar{y}_{i.}^2 - \frac{\left(\sum_{i=1}^{a} \frac{n_i}{\sigma_i^2} \bar{y}_{i.} \right)^2}{\sum_{i=1}^{a} \frac{n_i}{\sigma_i^2}}$$

and we can compare T to a χ^2_{a-1} distribution to obtain p-values for a significance test. In applied problems where the variance terms are unknown, we replace them with their respective sample variances and obtain the observed test statistic:

$$T_{Obs} = \sum_{i=1}^{a} \frac{n_i}{s_i^2} \bar{y}_{i.}^2 - \frac{\left(\sum_{i=1}^{a} \frac{n_i}{s_i^2} \bar{y}_{i.} \right)^2}{\sum_{i=1}^{a} \frac{n_i}{s_i^2}}$$

In the heteroscedastic case, the distribution of T_{Obs} under H$_0$ is not known, so we need another way to obtain a p-value. The generalized p-value approach described in Chapter 2, Section 4 allows us get a p value when nuisance parameters are present. The nuisance parameters for the one-way heteroscedastic ANOVA problem are the σ_i^2s. This typically requires us to simulate the distribution of the test statistic under H$_0$ using approaches like PB and OB.

Akritas and Papadatos (2004) have proposed an unweighted test that works well for one-way heteroscedastic ANOVA problems, even when the number of groups goes to infinity. We'll look more into this approach in Chapter 3, Section 5, and generalize it for testing fixed effects in the RCBD with subsampling with heteroscedastic errors in Chapter 4, Sections 7–9.

It is worth noting that an alternative path to solving the heterANOVA problem is by transforming the response. The famous approach by Box and Cox (n.d.) tries to find a power transformation that makes the likelihood look as good as possible by attempting to fix all problems in the residuals, including heteroscedasticity. This approach is widely implemented and readily available in software, but results in the analysis is being done on a different scale than that of the original data, resulting in an analysis that may be answering very different questions, or provides unintelligible answers. One solution to this problem is to back-transform to the original scale before interpretion, but this is reasonable when $\mathrm{E}[f(Y)] \approx f(\mathrm{E}[Y])$. Our focus is on modeling the covariance structure, so we will not explore the transformation approach any further.

2.4 Generalized P-values

Tsui and Weerahandi (1989) propose a "generalized" p-value as a solution to performing significance tests when nuisance parameters prevent a trivial solution. In our case, the nuisance parameters are the σ_i^2's. In general, suppose we wish to test a null model, say:

$$\boldsymbol{Y} \sim f_0(\boldsymbol{y}|\boldsymbol{\theta}_0, \boldsymbol{\phi}_0)$$

where $\boldsymbol{Y}$ is a random variable, $\boldsymbol{y}$ is an observed sample of $\boldsymbol{Y}$, $\boldsymbol{\theta}_0$ is a vector of the parameters of interest under H_0, and $\boldsymbol{\phi}_0$ is a vector of nuisance parameters under H_0. To find the generalized p-value for testing $\mathrm{H}_0 : \boldsymbol{\theta} \leq \boldsymbol{\theta}_0$, we need to find a generalized test variable. $T(\boldsymbol{Y}|\boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{\phi})$ is called a generalized test variable if the following hold:

1. $T(\boldsymbol{y}|\boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{\phi})$ is a pivot, free of $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$.

2. If we specify $\boldsymbol{\theta}$, the distribution of $T(\boldsymbol{Y}|\boldsymbol{y}, \boldsymbol{\theta}_0, \boldsymbol{\phi})$ is free of $\boldsymbol{\phi}$.

3. For fixed $\boldsymbol{y}$ and $\boldsymbol{\phi}_0$, $\Pr(T \leq t|\boldsymbol{\theta})$ is a monotonic function of $\boldsymbol{\theta}$ for any u.

The generalized p-value is defined as:

$$p - value = \Pr\left[T(\boldsymbol{Y}|\boldsymbol{y}, \boldsymbol{\theta}_0, \boldsymbol{\phi}) \geq T(\boldsymbol{y}|\boldsymbol{y}, \boldsymbol{\theta}_0, \boldsymbol{\phi})\right]$$

Tsui and Weerahandi used their generalized p-value idea to solve the Behrens-Fisher problem. Suppose we have:

$$\bar{Y}_1 \sim N\left(\mu_1, \frac{\sigma_1^2}{n_1}\right), \quad \bar{Y}_2 \quad \sim N\left(\mu_2, \frac{\sigma_2^2}{n_2}\right), \quad \frac{n_1 S_1^2}{\sigma_1^2} \sim \chi^2_{n_1-1}, \quad \frac{n_2 S_2^2}{\sigma_2^2} \sim \chi^2_{n_2-1}$$

where $\bar{Y}_1$, $\bar{Y}_2$, S_1^2, and S_2^2 are independent of one another. Let lower case versions of these letters represent the observed values from a sample. Our parameter of interest is $\theta = \mu_1 - \mu_2$, and the nuisance parameter is $\boldsymbol{\phi} = (\sigma_1^2, \sigma_2^2)$. A good choice for the generalized test variable is:

$$T = T(Y_1, Y_2 | y_1, y_2, \boldsymbol{\phi}) = \frac{\bar{Y}_1 - \bar{Y}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sqrt{\frac{s_1^2 \sigma_1^2}{S_1^2 n_1} + \frac{s_2^2 \sigma_2^2}{S_2^2 n_2}} = Z \sqrt{\frac{(n_1 - 1) s_1^2}{n_1 X_1^2} + \frac{(n_2 - 1) s_2^2}{n_2 X_2^2}}$$

where $Z \sim N(0,1)$, $X_1^2 \sim \chi^2_{n_1-1}$, and $X_2^2 \sim \chi^2_{n_2-1}$ are all independent random variables. $T_{Obs} = \bar{y}_1 - \bar{y}_2$, and for fixed y_1 and y_2, the distribution of T is free of $\boldsymbol{\phi}$. Since E$[T]$ is monotonically increasing with respect to $\boldsymbol{\phi}$, T is a generalized test variable and the corresponding generalized p-value is:

$$\mathcal{P}_{gen} = \Pr(|T| \geq |\bar{y}_1 - \bar{y}_2| | \theta = 0) = \Pr(T^2 \geq (\bar{y}_1 - \bar{y}_2)^2 | \theta = 0)$$

which we can evaluate via simulation.

2.5 The Parametric Bootstrap (PB)

Krishnamoorthy, Lu, and Mathew (2007) compared several approaches to the one-way heteroscedastic ANOVA problem, showing that the PB approach outperformed the Welch test, generalized F test, and James test. Unlike the more commonly implemented nonparametric bootstrap, the PB requires a model parameterization. While the PB is less flexible, it has the advantage of being more efficient, which is particularly advantageous in small sample problems.

The fundamental idea driving the PB is the "plug-in" principle. For some data model $\boldsymbol{Y} \sim f(\boldsymbol{Y}|\boldsymbol{\theta})$ we sample from $\boldsymbol{Y} \sim f(\boldsymbol{Y}|\hat{\boldsymbol{\theta}})$. This uses the naive frequentist approach to prediction that ignores the fact that we have to estimate the unknown parameters. The hope is that given a representative sample and a reasonable estimate of $\boldsymbol{\theta}$, we can replicate the mechanism generating the data. For our purposes, we'll use the PB to simulate the sampling distribution of the observed test statistic under the null hypothesis, and compare the observed test statistic to this estimated sampling distribution to obtain an estimate of the generalized p-value for testing differences in group means.

Zhang (2015a) proposed a PB approach for multiple comparisons in the one-way heteroscedastic ANOVA problem. His work includes a solution for the unbalanced case and a simulation study demonstrating its efficacy. Zhang (2015b) showed that using the PB to construct simultaneous pairwise comparisons outperforms the Tukey-Kramer procedure in two-way heteroscedastic ANOVA.

2.6 Objective Bayes (OB)

Objective Bayes (OB) uses Bayes' Theorem to obtain posterior distributions of parameters based on some "objective" prior distribution and the likelihood function. Objective priors are proposed to be appropriate in situations that we have no prior beliefs, information, or opinions about the parameters. The goal is to let the data speak for itself as much as possible.

The pioneers of Bayesianism, Thomas Bayes and Pierre-Simon Laplace, employed flat priors on the unobserved parameters in the context of their "inverse probability" approach (Bayes & Price, 1763; marquis de Laplace, 1820), but Jeffreys (1946) is widely considered the originator of OB methodology. As Kass and Wasserman (1996) discuss, despite Bayesian inference being rooted in subjectivist philosophy, the vast majority of applied Bayesian analyses are carried out using "so-called 'non-informative' priors".

There is no such thing as a truly "noninformative" prior, and examples like the paradox described in Lindley (1957) demonstrate the potentially catastrophic nature of using flat priors. In most cases, flat priors have little effect on the posterior analysis and give answers that resemble frequentist solutions, but it is worth recognizing that a flat prior for a mean parameter that places equal probability on the $(-\infty, \infty)$ line is a bit ridiculous. For any monstrously large central interval, the density is still infinitely small compared to the density outside that interval. Practically, this selection is fine for most problems, similar to how we can get away with using a normal distribution to model variables such as height despite the physical impossibility of observing negative values.

For the sake of argument, it is nigh impossible to think of a problem in which we can't come up with a prior that is less ridiculous than a flat one without influencing our posterior much. For mean parameters, one may use a bounded uniform prior with a very large range. Nevertheless, we are concerned with the relationship between the PB and OB methods, and the use of flat priors facilitates an understanding of the connection between the two approaches.

2.7 PB and OB Relationship

Tsui and Weerahandi (1989), note that their "p-value for the Behrens-Fisher problem turns out to be numerically... the same as Jeffreys's Bayesian solution and the Behrens-Fisher fiducial solution." Jeffrey's is the OB solution. The fiducial solution in Fisher (1941) arose from considering "Studentization". For Student (1908), this meant dividing the sample mean by the sample standard deviation so it only depends on the location parameter μ. Fisher proposes an extension of this argument to justify the procurement of a posterior distribution for a parameter θ without any prior. By attempting to "make a Bayesian omelet without breaking the Bayesian eggs" (Savage, 1961), Fisher's approach quickly garnered controversy. Nonetheless, the fiducial argument proposed an elegantly simple solution to the Behrens-Fisher problem. Despite being considered a dead end for awhile, fiducial inference has been resurrected with generalized confidence distributions, an extension of Weerahandi's generalized confidence intervals (Weerahandi, 1995).

From Efron (1998): assume we can construct an upper confidence limit $\alpha \times 100\%$ for θ for every possible value of α, then we define the confidence distribution of θ as:

$$\Pr[\theta < \hat{\theta}(\alpha)] = \alpha$$

and if we "interpret this as a probability distribution for θ given the data... the classic *wrong* interpretation of confidence", that "θ is in the interval $(\hat{\theta}(0.9), \hat{\theta}(0.91))$ with probability 0.01", etc., then by taking the limit, we have that the confidence distribution is the fiducial distribution. Hence, the only difference between confidence distributions and fiducial ones is the interpretation. There is a perfectly valid frequentist justification for confidence distributions if we don't confuse the repeated sampling property with actual probability (easier said than done).

Bootstrap distributions are the most common type of confidence distribution. Studentization is used to establish pivotal quantities in PB tests. For the Behrens-Fisher problem, there is a clear relationship between the PB and OB tests, and we will demonstrate that there is a strong similarity between PB and some OB approaches in the one-way heterANOVA problem.

Bayarri and Berger (2004) consider OB to be perhaps "the most promising route to the unification of Bayesian and frequentist statistics". Efron (2013) discusses the relationship between the PB and OB, specifically demonstrating their near-equivalency for the problem of estimating the correlation parameter of a bivariate normal distribution and Efron (2012) shows the existence of a "Bayes/bootstrap" conversion factor for multidimensional exponential families.

From a philosophical standpoint, the PB and OB tests are fundamentally coming from different places. The PB considers the data as random and the parameters as fixed, while OB considers the data as fixed and the parameters as random. Empirically, the PB and OB approaches tend to give similar results, so for one-way heteroscedastic ANOVA, we'll investigate an OB test that looks at how far away T_{Obs} is from the posterior distribution of:

$$\tilde{T} = \sum_{i=1}^{a} \frac{n_i}{s_i^2}(\bar{y}_{i.} - \mu_i)^2 - \frac{\left(\sum\limits_{i=1}^{a} \frac{n_i}{s_i^2}(\bar{y}_{i.} - \mu_i)\right)^2}{\sum\limits_{i=1}^{a} \frac{n_i}{s_i^2}}$$

because under H$_0$, $\tilde{T} = T_{Obs}$. We use an analogous procedure for an OB test for fixed group effects in the RCBD with subsampling model with heteroscedastic errors.

Chapter 3

One-Way Heteroscedastic ANOVA

3.1 The Problem

One-way heteroscedastic ANOVA is a class of problems where we wish to test for differences in group means when the observations for each group have differing levels of variability. The model in focus is:

$$y_{ij} = \mu_i + e_{ij}, \quad i = 1, \ldots, a,\ j = 1, \ldots, n_i, \quad e_{ij} \sim N(0, \sigma_i^2)$$

For a significance test of $\mathrm{H}_0 : \mu_1 = \mu_2 = \cdots = \mu_a$, the Wald-type weighted test statistic is given by:

$$T \equiv \sum_{i=1}^{a} \frac{n_i}{\sigma_i^2} \bar{y}_{i\cdot}{}^2 - \frac{\left(\sum_{i=1}^{a} \frac{n_i}{\sigma_i^2} \bar{y}_{i\cdot} \right)^2}{\sum_{i=1}^{a} \frac{n_i}{\sigma_i^2}}$$

and since the σ_i^2's are unknown in practice, we replace them with their respective sample variances to get the observed test statistic:

$$T_{Obs} \equiv \sum_{i=1}^{a} \frac{n_i}{s_i^2} \bar{y}_{i\cdot}{}^2 - \frac{\left(\sum_{i=1}^{a} \frac{n_i}{s_i^2} \bar{y}_{i\cdot} \right)^2}{\sum_{i=1}^{a} \frac{n_i}{s_i^2}}$$

See Appendix A for a derivation of T and T_{Obs}. The distribution of T_{Obs} is unknown, so our goal is to estimate it so that we can perform a significance test for the equality of group means.

In this chapter, we explore several approaches to testing H_0, including the PB (Section 2), OB (Section 3), predictive approach (Section 5) and the unweighted test statistic approach of Akritas & Papadatos (Section 6). Section 4 highlights the relationship between the PB and OB tests, and establishes the conditions for which they are equivalent. simulation study in Section 7. Finally, we conclude with a real data example in Section 8.

3.2 Parametric Bootstrap (PB) Approach

For a PB test, one can simulate the raw data, but for our goal of simulating T_{Obs} under H_0, we can directly sample the sufficient statistics $\bar{y}_i$ and s_i^2:

$$\bar{y}_{iB} \sim N\left(\bar{y}_{..}, \frac{s_i^2}{n_i}\right) \qquad \text{and} \qquad s_{iB}^2 \sim \frac{s_i^2 \chi^2_{n_i-1}}{n_i - 1}$$

While the plug-in principle in general will use a normal distribution with mean $\bar{y}_{..}$, under H_0 the $\bar{y}_{iB}$'s have the same mean, and based on the structure of T_{Obs}, it doesn't matter what that mean is, so without loss of generality, we take it to be 0 for simplicity, sampling $\bar{y}_{iB} \sim N\left(0, \frac{s_i^2}{n_i}\right)$.

For each value of $B = 1, ..., M$, we sample $\bar{y}_{iB}$, s_{iB}^2 , for $i = 1, ..., a$, and compute:

$$T_{PB} \equiv \sum_{i=1}^{a} \frac{n_i}{s_{iB}^2} \bar{y}_{iB}^2 - \frac{(\sum\limits_{i=1}^{a} \frac{n_i}{s_{iB}^2} \bar{y}_{iB})^2}{\sum\limits_{i=1}^{a} \frac{n_i}{s_{iB}^2}}$$

With a sufficiently large number of draws M of T_{PB}, we can flesh out an estimate of the sampling distribution of T_{Obs} under H_0, and estimate the generalized p-value

$Pr(T_{PB} > T_{Obs})$ with:

$$\mathcal{P}_{PB} = \frac{1}{M}\sum_{B=1}^{M} \mathrm{I}(T_{PB} > T_{Obs})$$

where I() is an indicator variable such that:

$$\mathrm{I}(T_{PB} > T_{Obs}) = \begin{cases} 1 & \text{if } T_{PB} > T_{Obs} \\ 0 & \text{otherwise} \end{cases}$$

3.3 Objective Bayesian (OB) Approach

Empirically, the OB approach to the significance test of $\mathrm{H}_0 : \mu_1 = \mu_2 = \cdots = \mu_a$ is very similar to the PB approach. A Bayesian significance test looks at how far away T_{Obs} is from the posterior distribution of:

$$\tilde{T} = \sum_{i=1}^{a} \frac{n_i}{s_i^2}(\bar{y}_{i.} - \mu_i)^2 - \frac{\left(\sum_{i=1}^{a} \frac{n_i}{s_i^2}(\bar{y}_{i.} - \mu_i)\right)^2}{\sum_{i=1}^{a} \frac{n_i}{s_i^2}}$$

Note that when H_0 is true, $\tilde{T} = T_{Obs}$. With flat priors on the μ_is:

$$\frac{\bar{y}_{i.} - \mu_i}{s_i/\sqrt{n_i}}|\boldsymbol{Y} \sim t_{n_i-1} \therefore \bar{y}_{i.} - \mu_i|\boldsymbol{Y} \sim \frac{s_i}{\sqrt{n_i}}t_{n_i-1}$$

we repeatedly sample and compute:

$$\begin{aligned} T_{OB} &\equiv \sum_{i=1}^{a} \frac{n_i}{s_i^2}\left(\frac{s_i}{\sqrt{n_i}}t_{n_i-1}\right)^2 - \frac{\left(\sum_{i=1}^{a} \frac{n_i}{s_i^2}\frac{s_i}{\sqrt{n_i}}t_{n_i-1}\right)^2}{\sum_{i=1}^{a} \frac{n_i}{s_i^2}} \\ &= \sum_{i=1}^{a} t_{n_i-1}^2 - \frac{\left(\sum_{i=1}^{a} \frac{\sqrt{n_i}}{s_i}t_{n_i-1}\right)^2}{\sum_{i=1}^{a} \frac{n_i}{s_i^2}} \end{aligned}$$

Similar to the PB test, for the OB test, we sample T_{OB} M times, and estimate the p-value with:

$$\mathcal{P}_{OB} = \frac{1}{M}\sum_{B=1}^{M} \mathrm{I}(T_{OB} > T_{Obs})$$

3.4 PB/OB Relationship

Noting the relationship between the standard normal, chi-squared, and t distributions:

$$\frac{z_i}{\sqrt{\chi^2_{n_i-1}/(n_i-1)}} \overset{d}{=} t_{n_i-1}$$

We can rewrite T_{PB} in terms of draws from the t_{n_i-1} distribution:

$$T_{PB} = \sum_{i=1}^{a} t^2_{n_i-1} - \frac{\left(\sum_{i=1}^{a} \frac{\sqrt{n_i}}{s_i} t_{n_i-1}\right)^2}{\sum_{i=1}^{a} \frac{n_i(n_i-1)}{s_i^2 \chi^2_{n_i-1}}}$$

Hence, T_{PB} and T_{OB} will be equivalent when:

$$\hat{\psi} = \frac{\chi^2_{n_i-1}}{(n_i-1)} = 1 \qquad \text{for all } i$$

where ψ is the dispersion parameter. Also note that $T_{PB} \to T_{OB}$ asymptotically. To see this, note that a $\chi^2_{n_i-1}$ is a sum of $n_i - 1$ independent squared standard normal variables:

$$z^2 \sim \chi^2_1 \qquad \mathrm{E}[z^2] = 1$$

so by the law of large numbers:

$$\frac{\sum_{t=1}^{n_i-1} z_t^2}{n_i - 1} \to 1$$

and hence $\hat{\psi} \to 1$ for all i, and $T_{PB} \to T_{OB}$.

3.5 Predictive Distribution

The predictive distribution is the distribution for unobserved or future realizations of the response variable, given the already observed data. Ron Christensen suggested investigating a test based on the predictive distribution, based on the work of Aitchison (1975). The crux of this paper is that the predictive distribution $f(\boldsymbol{Y}_{new}|\boldsymbol{Y})$

provides a better estimate of $f(\boldsymbol{Y}|\boldsymbol{\theta})$ than the plug-in distribution $f(\boldsymbol{Y}|\hat{\boldsymbol{\theta}})$, based on the Kullback-Leibler divergence $K()$, i.e.:

$$K(f(\boldsymbol{Y}_{new}|\boldsymbol{Y}), f(\boldsymbol{Y}|\boldsymbol{\theta})) \leq K(f(\boldsymbol{Y}|\hat{\boldsymbol{\theta}}), f(\boldsymbol{Y}|\boldsymbol{\theta}))$$

Our test of $\mathrm{H}_0 : \mu_1 = \mu_2 = \cdots = \mu_a$ is a one-sided test, so performance is primarily related to the behavior of the right-tail of the sampling distribution of T_{Obs} under H_0. Hence, even if the predictive approach gives a better estimate of this sampling distribution overall, it may not lead to a better test. For one-way heteroscedastic ANOVA, under H_0, new data observations come from:

$$\frac{y_{n_i+1}}{s_i\sqrt{1+\frac{1}{n_i}}}|\boldsymbol{Y} \sim t_{n_i-1}$$

and for a new sample of size m: $y_{n_i+1}, y_{n_i+2}, ..., y_{n_i+m}$ we have:

$$\bar{y}^*_{i.} = \frac{\sum_{j=1}^{m} y_{n_i+j}}{m} = \bar{y}_{i.} + s_i\sqrt{1+\frac{1}{n_i}}\frac{\sum_{j=1}^{m} t_i}{m}$$

with:

$$\mathrm{E}[t_i] = 0, \mathrm{Var}(t_i) = \frac{n_i-1}{n_i-3} \therefore \mathrm{E}[\bar{y}^*_{i.}|\boldsymbol{Y}] = \bar{y}_{i.}, \mathrm{Var}(\bar{y}^*_{i.}|\boldsymbol{Y}) = s_i^2\left(1+\frac{1}{n_i}\right)\frac{n_i-1}{n_i(n_i-3)}$$

Implementing a normal approximation, we have:

$$\bar{y}^*_{i.}|\boldsymbol{Y} \sim N\left(\bar{y}_{i.}, \frac{(n_i-1)(n_i+1)}{n_i^2(n_i-3)}s_i^2\right)$$

$$s_i^{2*}|\boldsymbol{Y} \sim \frac{\chi^2_{n_i-1}}{n_i-1}\frac{(n_i-1)(n_i+1)}{n_i(n_i-3)}s_i^2$$

and noting that:

$$\bar{y}^*_{i.}|\boldsymbol{Y} \overset{d}{=} (z_i + \bar{y}_{i.})\frac{(n_i-1)(n_i+1)}{n_i^2(n_i-3)}s_i^2$$

we can compute draws of:

$$T_{pred} \equiv \sum_{i=1}^{a} \frac{n_i}{s_i^2\left(\frac{\chi^2_{n_i-1}}{n_i-1}\frac{(n_i-1)(n_i+1)}{n_i(n_i-3)}\right)}\left((z_i+\bar{y}_{i.})\sqrt{\frac{s_i^2(n_i-1)(n_i+1)}{n_i^2(n_i-3)}}\right)^2$$

$$-\frac{\left(\sum_{i=1}^{a} \frac{n_i}{s_i^2\left(\frac{\chi^2_{n_i-1}}{n_i-1}\frac{(n_i-1)(n_i+1)}{n_i(n_i-3)}\right)}\left((z_i+\bar{y}_{i.})\sqrt{\frac{s_i^2(n_i-1)(n_i+1)}{n_i^2(n_i-3)}}\right)\right)^2}{\sum_{i=1}^{a} \frac{n_i}{s_i^2\left(\frac{\chi^2_{n_i-1}}{n_i-1}\frac{(n_i-1)(n_i+1)}{n_i(n_i-3)}\right)}}$$

To approximate the generalized p-value $\Pr(T_{Pred} > T_{Obs})$, we take M draws of T_{Pred}, and calculate:

$$\mathcal{P}_{pred} = \frac{1}{M} \sum_{B=1}^{M} \mathrm{I}(T_{Pred} > T_{Obs})$$

3.6 The Unweighted Test Statistic

Akritas and Papadatos (2004) developed an alternative one-way ANOVA test statistic as a competitor to the usual weighted, ratio-based F statistic. Their statistic is unweighted and difference-based:

$$T_a \equiv \frac{\sum\limits_{i=1}^{a} \left[n_i \left(\bar{y}_{i.} - \bar{y}_{..} \right)^2 - s_i^2 \left(1 - \frac{n_i}{N} \right) \right]}{\sqrt{a}}$$

and is inspired by the fact that in the balanced case E[MSGrps] = E[MSE] under H_0. T_a reflects an adjustment to center MSGrps − MSE for use in the unbalanced case. For the unbalanced case, as $a \to \infty$, the asymptotic distribution of T_a is:

$$T_a \xrightarrow{d} N\left(0, 2\left(\tau^4 + \gamma^4\right)\right)$$

where $\frac{1}{a} \sum\limits_{i=1}^{a} \sigma_i^4 \to \tau^4 \in (0, \infty)$ and $\frac{1}{a} \sum\limits_{i=1}^{a} \frac{\sigma_i^4}{n_i - 1} \to \gamma^4 \in (0, \infty)$. We compare T_a to this distribution to obtain the p-value, $\mathcal{P}_a$. The unweighted test statistic is effective for one-way heteroscedastic ANOVA, and performs well even when the number of groups is very large. Appendix B includes details on how to estimate the σ_i^4s to ensure consistent estimation of τ^4 and γ^4.

3.7 Simulation Study

To evaluate and compare the PB, OB, predictive, and unweighted approaches, we use several simulation based approaches. First, we compare how the methods estimate the sampling distribution of the observed test statistic for a specific set of parameter values. Next, we evaluate the effectiveness of the tests based on their empirical type I error rate and power. We split this task into two components—a "shotgun" approach where we consider hundreds of scenarios to get a sense of which methods are effective for an arbitrary one-way ANOVA problem, and a more traditional, focused approach that looks at power curves. All simulations are carried out using R (R Core Team, 2016).

3.7.1 Density Estimation Comparison

In this section we compare how the PB, OB and predictive approaches estimate the sampling distribution of the observed test statistic. We omit the approach of Akritas & Papadatos as it is based on an entirely different test statistic. The simulation proceeds as follows:

- Select the number of groups a, sample sizes n_i, group means μ_i and group variances σ_i^2 for $i = 1, 2, \ldots, a$.
- Generate the sufficient statistics $\bar{y}_i$ and s_i^2.
- Calculate $B = 1, 2, \ldots, 1000$ draws of T_{PB}, T_{OB}, and T_{pred}.
- Plot the kernel density estimates.

We present examples of a small sample size case and large sample size case. For each, we consider $a = 3$ groups under H_0, with the μ_is all equal to zero, and $\sigma_i^2 = c(9, 4, 1)$. Figure 3.1 shows the small sample case that features group sizes $n_i = c(5, 7, 6)$, and Figure 3.2 shows the large sample case that features group sizes $n_i = c(108, 72, 69)$.

Figure 3.1: Density Estimation—Small Sample Sizes

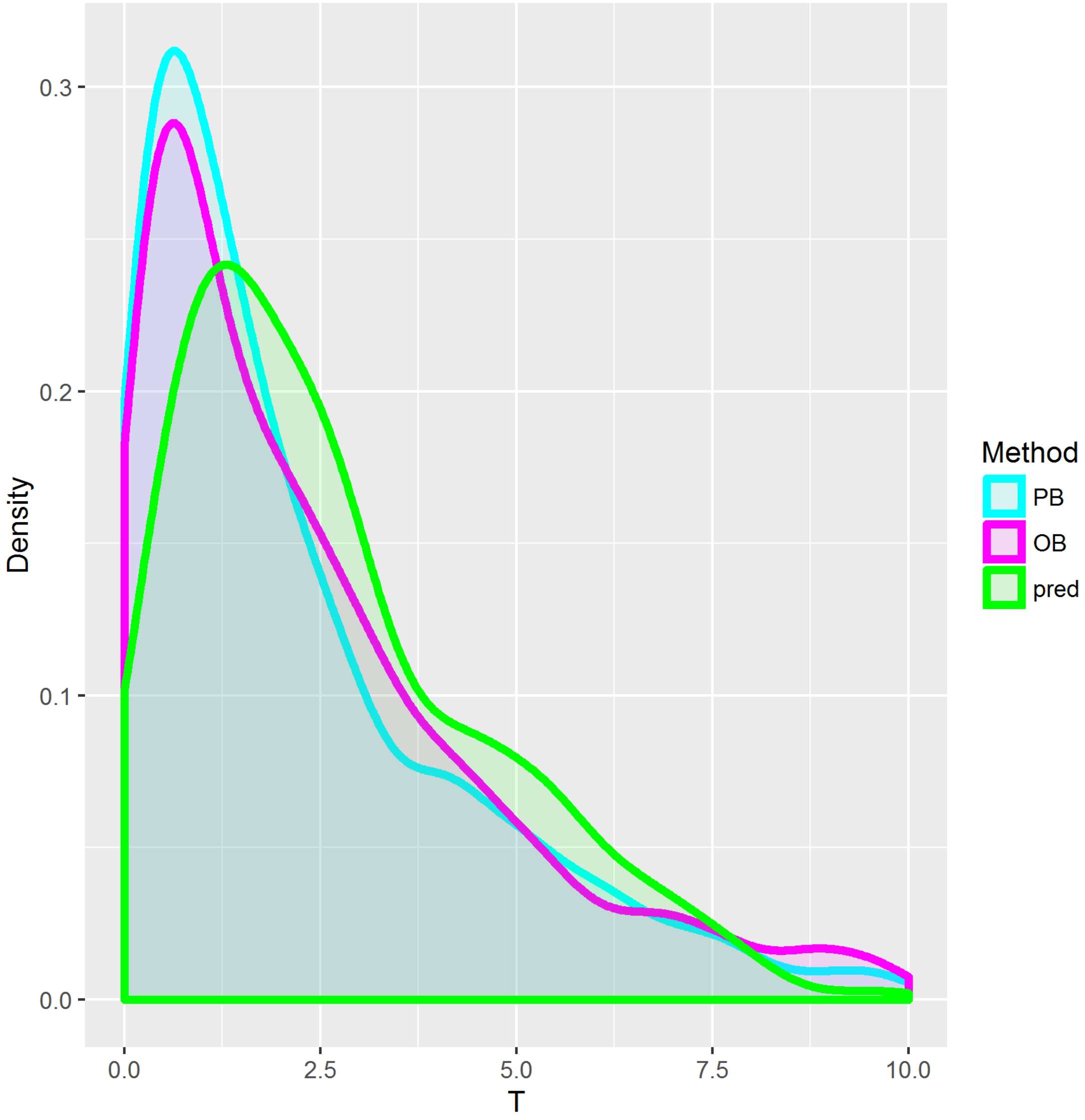
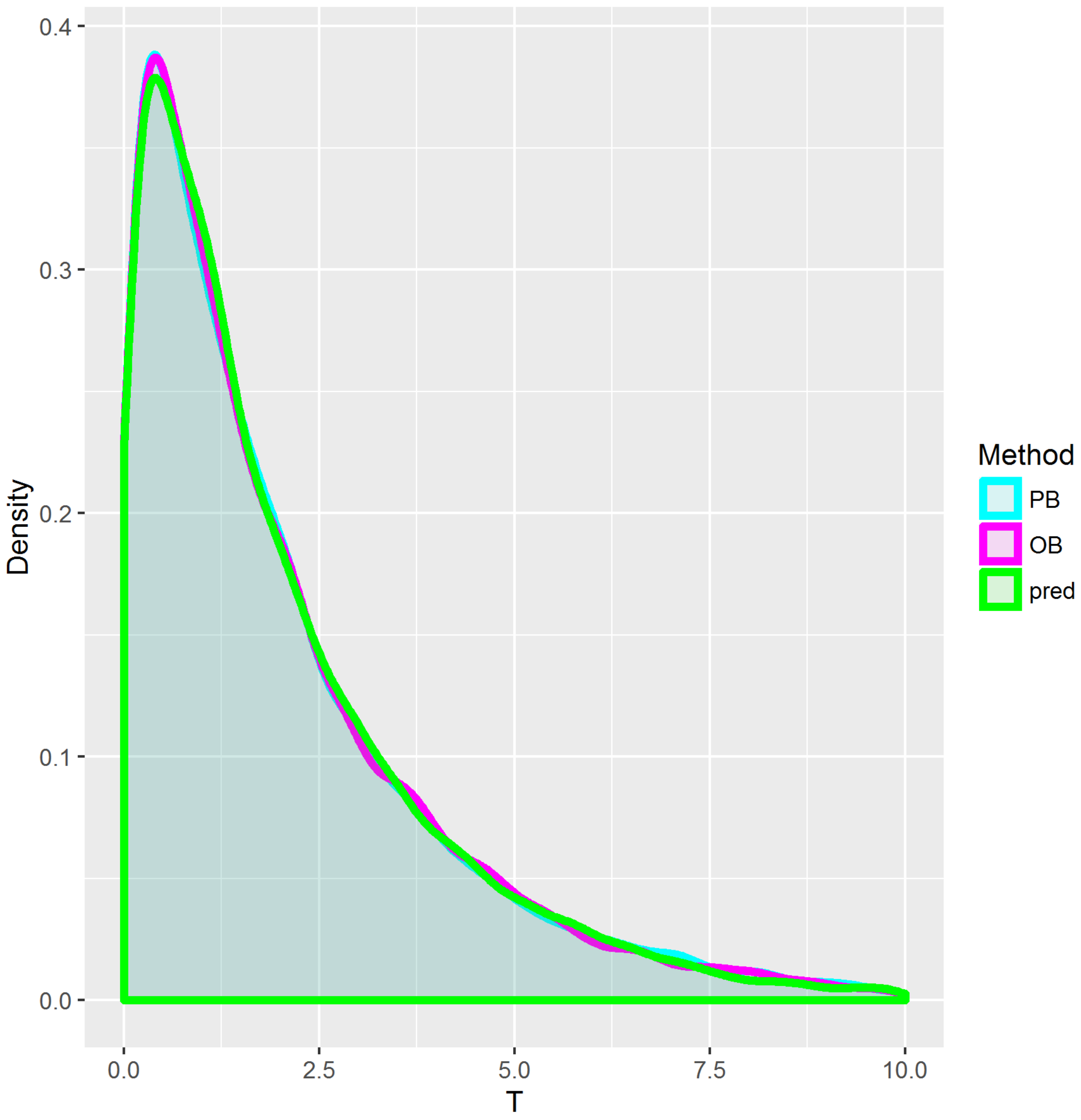


Figure 3.2: Density Estimation—Large Sample Sizes



In both cases we see the methods estimating the test statistic with a similar shape—positive, unimodal, right-skewed distributions with the bulk of the density close to 0, which coincides with the null hypothesis of equal group means. In the small sample case, all approaches perform similarly, and in the large sample case, the methods are nearly identical, suggesting the asymptotic equivalence of the methods.

3.7.2 Type I Error Rate and Power

The most common way to compare statistical tests is via power analysis. In a simulation, we generate the data, so we know the true parameter values, and hence whether the null hypothesis is true or not true. The power of a test is the probability that it correctly rejects H_0 when it is false. We estimate the power by calculating the proportion of false null hypotheses that are correctly rejected. The closer to 1, the better, with one important caveat—the method must also be able to control the type I error rate.

A type I error is committed when a true null hypothesis is incorrectly rejected. We set a nominal level of significance, which is the proportion of null hypotheses that will be rejected based on sampling variability alone, and estimate the type I error rate with the proportion of true null hypotheses that are incorrectly rejected. The closer to the nominal level, the better. Controlling the type I error rate takes precedence over power. To illustrate this, consider a test that always rejects the null hypothesis. This ridiculous test will achieve 100% power, but it will also always make type I errors when the H_0 is true.

Both type I error rate and power are rejection rates—the type I error rate is the rejection rate when H_0 is true, and power is the rejection rate when H_0 is false. The simulation proceeds as follows:

- Specify the number of groups a, sample sizes n_i, group means μ_i, group variances σ_i^2, and a nominal significance level α.
- For plotting power curves, calculate the effect size. We use T.
- For each of the R simulation repetitions:
  - Generate sufficient statistics $\bar{y}_i$ and s_i^2.

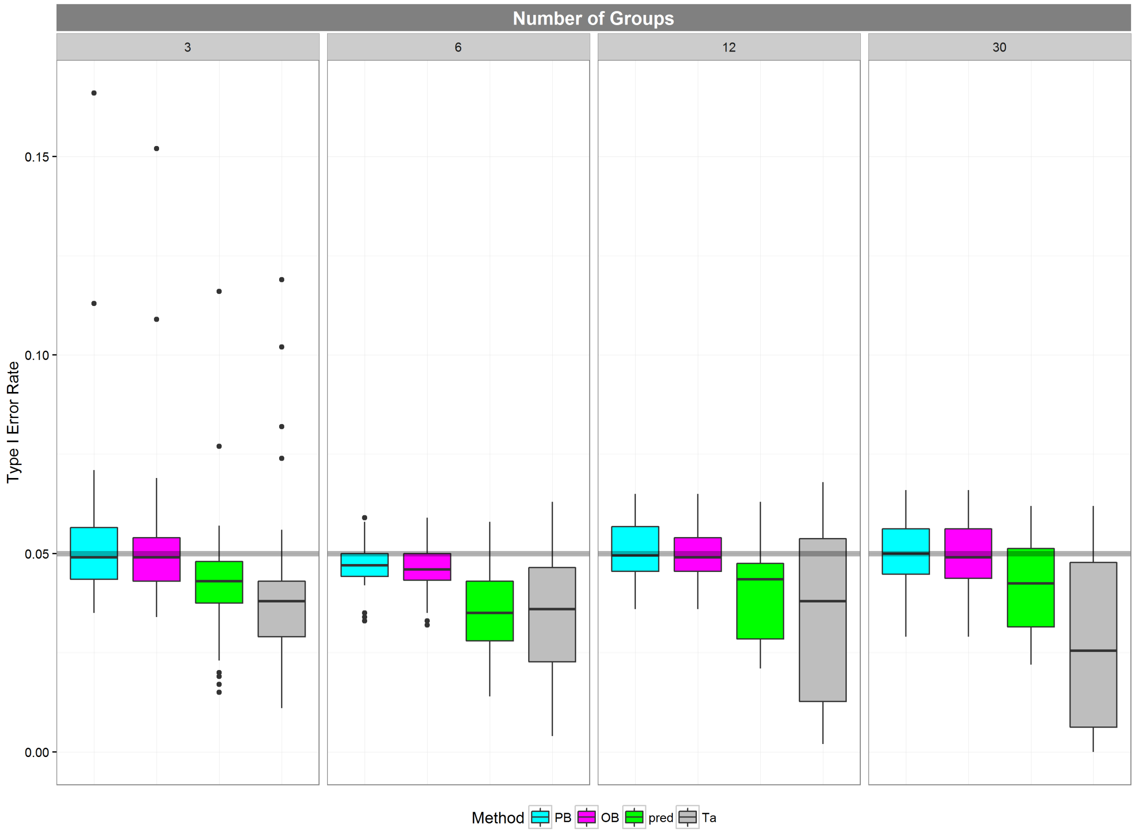
  - Calculate the observed test statistic T_{Obs} and unweighted test statistic T_a.
  - Draw B observations of T_{PB}, T_{OB}, and T_{pred}, and calculate their corresponding p-values $\mathcal{P}_{PB}$, $\mathcal{P}_{OB}$, and $\mathcal{P}_{pred}$.
  - Calculate τ^4 and γ^4 and obtain $\mathcal{P}_a$.
  - Compare the p-values from each method to the nominal α level—if $\mathcal{P} < \alpha$, add 1 to the rejection counter for that method.

- Divide the number of rejections for each method by the number of simulation repetitions R to obtain the rejection rate.
- If H$_0$ is true (all μ_is are equal), the rejection rate represents the type I error rate.
- If the H$_0$ is false (not all μ_is are equal), the rejection rate represents the power.

With so many parameters, it is difficult to assess how the methods will perform in all situations. We propose a shotgun approach where we throw a vast array of parameter combinations at each method and pool together the type I error rate and power results to compare the relative effectiveness of the methods. We also present an example of a more traditional or focused approach that looks at power curves.

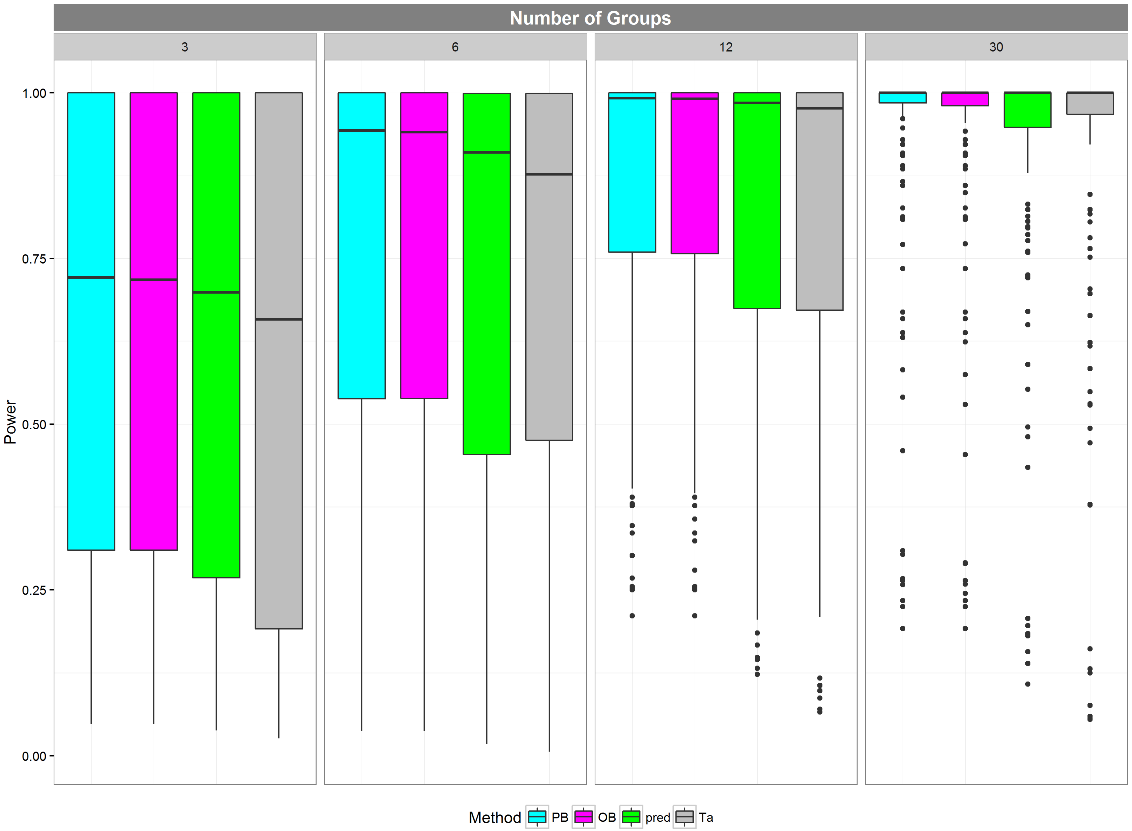
The shotgun approach sees only $\alpha = 0.05$, $R = 1000$, and $B = 1000$ fixed. Our simulation considers hundreds of combinations of the number of groups, group sizes, sample means, and sample variances. With so many combinations, we are able to get a better sense of how the methods perform for an arbitrary one-way heteroscedastic ANOVA problem, but are unable to look at each individual case in detail. We present box plots of the rejection rates broken down by the number of groups. We plot the true null scenarios (for assessing type I error rate) in Figure 3.3, and the false null scenarios (for assessing power) in Figure 3.4.

Figure 3.3: Type I Error Rate—106 Different Scenarios



In general, the PB, OB, predictive, and unweighted approaches are all able to control the type I error rate across the board. PB and OB have very similar performance, with type I error rates very close to the nominal level 0.05. The predictive approach is a bit more conservative, with type I error rate typically less than the nominal level, and the unweighted test appears to be even more conservative. For the 12 and 30 group simulations, the predictive and unweighted tests demonstrate more variability than they do in the 3 and 6 group cases.

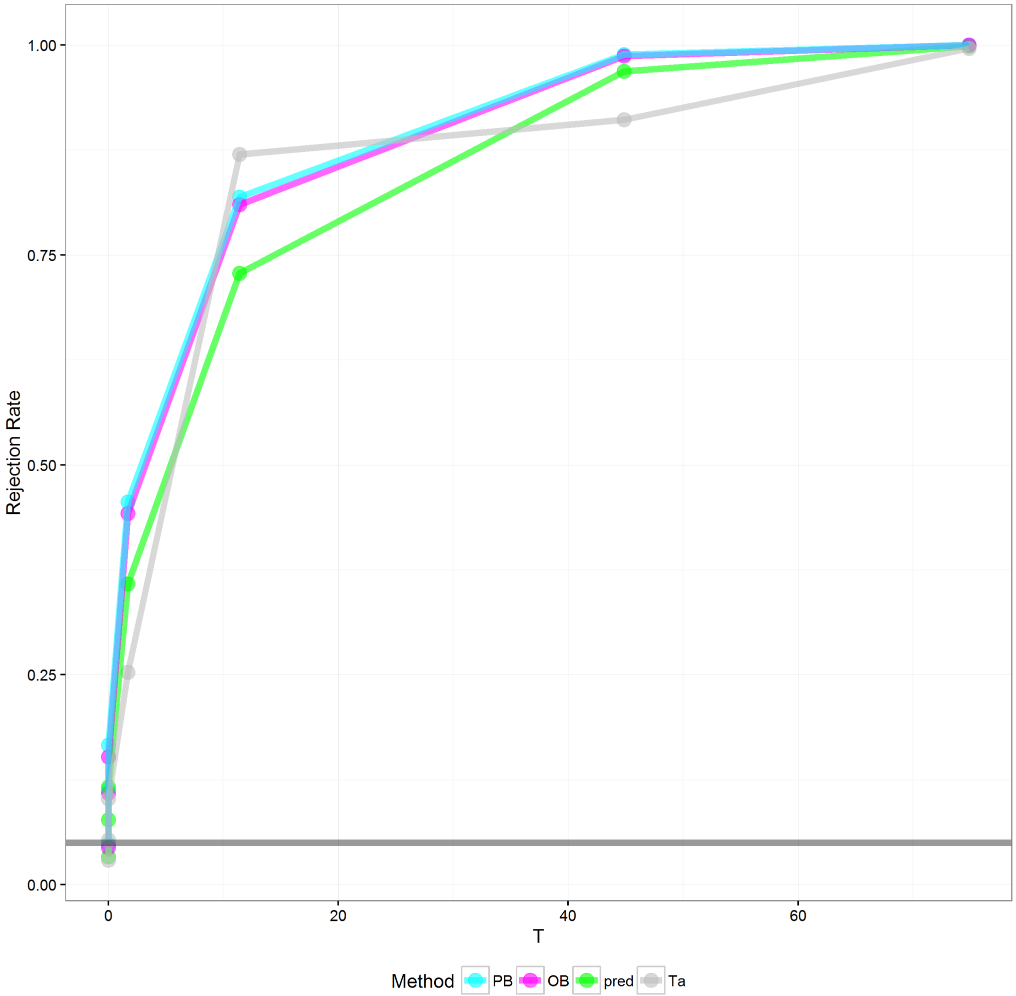
Figure 3.4: Power Analysis—712 Different Scenarios



The PB, OB, predictive, and unweighted approaches all perform fairly similarly in terms of power. Based on the median power levels and interquartile ranges, PB and OB have nearly identical performance, slightly outperforming the predictive approach, which in turn slightly outperforms the unweighted approach.

The focused approach reveals the behavior of the power curves with higher resolution at the expense of considering fewer cases. Figure 3.5 shows an example of the power curves, where we leave the number of groups fixed at $a = 3$, group sizes at $n_i = c(4, 8, 9)$, and variances at $\sigma_i^2 = c(1, 4, 9)$, while varying only the μ_is. In the example, we look at 7 sets of μ_i's, ranging from small to large effect sizes. We set a nominal significance level of $\alpha = 0.05$, perform $R = 1000$ runs at each setting, and use $B = 1000$ draws of T_{PB}, T_{OB}, and T_{pred} a density from which to find the approximate p-values associated with these methods.

Figure 3.5: Power Curves

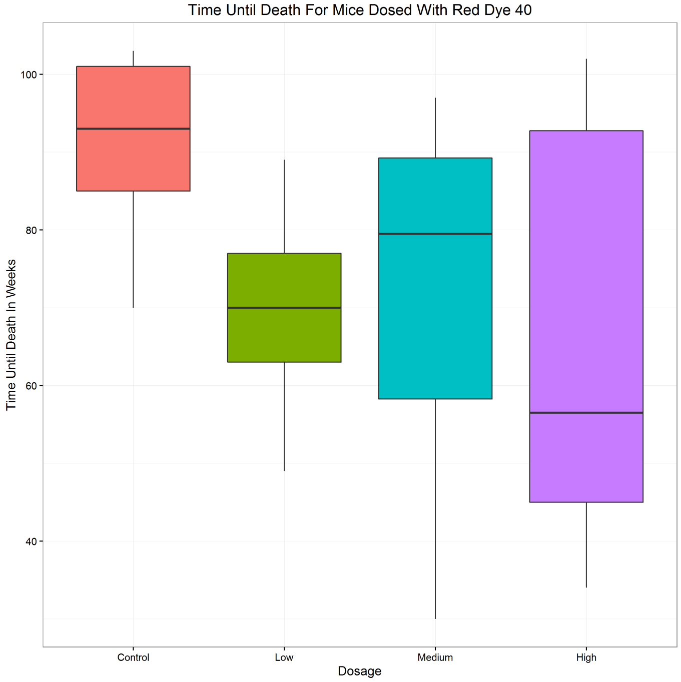


Overall, we see what we'd expect, with power increasing for each as effect size increases. The methods perform comparably, and we see that the PB and OB approaches have nearly identical performance.

3.8 Example: Red Dye Number 40

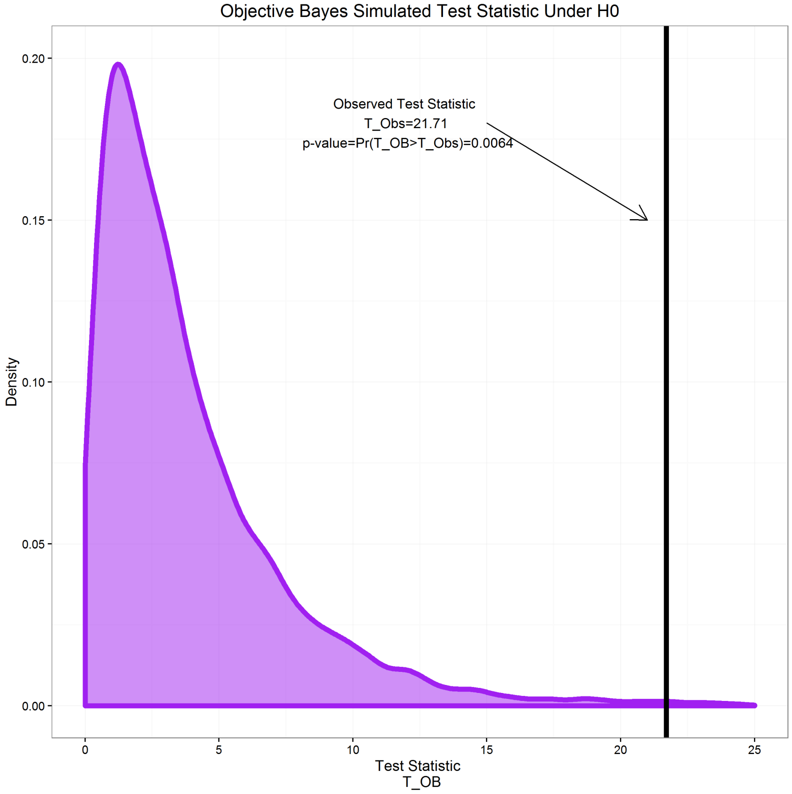
To illustrate the testing procedure in a more practical light, we demonstrate the OB approach for a real one-way heteroscedastic ANOVA problem. The dataset in focus is from Lagakos and Mosteller (1981), who explored the carcinogenic effects of Red Dye Number 40 by feeding mice various doses of the dye and then recording their time of death (in weeks). The experiment features four groups of various sizes: a control group ($n_1 = 11$), a low dosage group ($n_2 = 9$), a medium dosage group ($n_3 = 10$), and a high dosage group ($n_4 = 8$). Figure 3.6 uses box plots to show the empirical distribution of survival time for each group, and suggests that heteroscedasticity is present.

Figure 3.6: Empirical Distribution of Mouse Lifespan by Group



A Breusch-Pagan test gives a p-value of 0.0058, so we conclude that heteroscedasticity is present. Hence, the Red Dye 40 experiment is a one-way heteroscedastic ANOVA problem. To conduct the OB test for a difference of group means, first calculate group means and variances, and center group means around the grand mean. Next, calculate the observed test statistic. For this problem, we obtain $T_{Obs} = 21.71$. We then generated 10000 draws of T_{OB} under H_0. Figure 3.7 depicts the estimated distribution of the test statistic under H_0 in purple, along with a black vertical line indicating the observed test statistic:

Figure 3.7: Objective Bayes Simulated Test Statistic Distribution



Our estimate for the generalized p-value $\mathcal{P}_{OB} = \Pr[T_{OB} > T_{Obs}]$ is the proportion of T_{OB} draws that were greater than T_{Obs}, so $\mathcal{P}_{OB} = \frac{64}{10000} = 0.0064$. Hence, we reject H_0 and conclude that at least two of the groups of mice have different lifespans, on average.

Chapter 4

RCBD with Subsampling and Heteroscedastic Errors

4.1 Introduction

In this chapter we offer a more complicated special case where we have dependent data—specifically, two-way heterANOVA under the randomized complete block design (RCBD) with subsampling and heteroscedastic errors. The model in focus is:

$y_{ijk} = \mu_i + \eta_j + e_{ijk}$	the linear model
$i = 1, 2, \ldots, a$	groups
$j = 1, 2, \ldots, b$	blocks
$k = 1, 2, \ldots, n_{ij}$	subsamples
μ_i	group means
$\eta_j \sim N(0, \sigma_w^2)$	random block effects
$e_{ijk} \sim N(0, \sigma_i^2)$	normally distributed measurement errors

$\eta_j \perp\!\!\!\perp e_{ijk}$ block effects independent of measurement errors

In the RCBD, observations within each block are correlated with one another, hence our descriptor of the data as dependent. A standard RCBD features one observation at each treatment-block combination, but an obvious extension is to allow for subsampling multiple observations at each treatment-block combination.

There are two sources of error in this model, the within-block or measurement error terms σ_i^2, and the random block effect term σ_w^2. We'll allow for heteroscedastic measurement error terms. The subsampling model is incredibly useful if there is a great deal of variability within each block, but subsamples will only reduce our estimate of the within block variability. Hence, in order to reduce the total variability associated with the response, we need to increase the number of complete blocks. If between block variability is high and within block variability is low, for example, you'll gain very little additional information by collecting subsamples.

Our goal is to perform a test for fixed group effects: $\mathrm{H}_0 = \mu_1 = \mu_2 = \cdots = \mu_a$, and the observed test statistic does not follow a known distribution. Hence, we once again need to simulate the sampling distribution of the observed test statistic under H_0 to obtain a generalized p-value and conduct the test.

In Section 4.2, we describe the RCBD with subsampling and heteroscedastic errors model, and present the weighted test statistic for testing H_0. Section 4.3 details an OB approach and includes a simulation study to explore the type I error rate. An application to a real data example—the sea urchin grazing experiment—is demonstrated in Section 4.4.

Sections 4.7–4.9 define a new unweighted test statistic for the RCBD with subsampling model. We discuss how to appropriately consider the sources of error, define the expected values of the relevant mean squared terms, and conclude with a derivation of the asymptotic distribution of this unweighted test statistic.

4.2 RCBD with Subsampling

In our exploration of the RCBD with subsampling model, we consider fixed group effects and random block effects. We'll focus on testing for a difference in group means. For this purpose we can simplify the model described by averaging over the subsamples:

$$\begin{aligned}
\bar{y}_{ij.} &= \mu_i + \eta_j + \bar{e}_{ij.} \\
\bar{y}_{ij.} &= \frac{\sum\limits_{k=1}^{n_{ij}} y_{ijk}}{n_{ij}} \\
\bar{e}_{ij.} &= \frac{\sum\limits_{k=1}^{n_{ij}} e_{ijk}}{n_{ij}} \\
\eta_j &\sim N\left(0, \sigma_w^2\right) \\
\bar{e}_{ij.} &\sim N\left(0, \frac{\sigma_i^2}{n_{ij}}\right) \\
\text{Var}(\bar{y}_{ij.}) &= \sigma_w^2 + \frac{\sigma_i^2}{n_{ij}}
\end{aligned}$$

We see that an increase in subsamples only gives a reduction of within block variability. For comparing group means, we can reframe the RCBD with subsampling model as a one-way ANOVA problem. Blocks are assumed to be independent from one another, so we have:

$$\begin{aligned}
\bar{y}_{i..} &= \mu_i + \xi_i \\
\xi_i &= \frac{1}{b}\sum_{j=1}^{b}(\eta_j + \bar{e}_{ij.}) \\
v_i \equiv \text{Var}(\bar{y}_{i..}) &= \text{Var}\left(\frac{1}{b}\sum_{j=1}^{b}\bar{y}_{ij.}\right) = \frac{1}{b^2}\text{Var}\left(\bar{y}_{i1.} + \bar{y}_{i2.} + \ldots + \bar{y}_{ib.}\right) \\
&= \frac{1}{b^2}\left(\left(\sigma_w^2 + \frac{\sigma_i^2}{n_{i1}}\right) + \left(\sigma_w^2 + \frac{\sigma_i^2}{n_{i2}}\right) + \ldots + \left(\sigma_w^2 + \frac{\sigma_i^2}{n_{ib}}\right)\right)
\end{aligned}$$

$$= \frac{b\sigma_w^2 + \sum\limits_{j=1}^{b} \frac{\sigma_i^2}{n_{ij}}}{b^2} \equiv \frac{1}{\omega_i}$$

By averaging over the subsamples and blocks, it is clear that for a reduction of the overall variance terms associated with the group means, we need to increase the number of complete blocks. This approach also facilitates the derivation of a weighted Wald-type test statistic for testing H_0:

$$T = \sum_{i=1}^{a} \omega_i \bar{y}_{i..}^2 - \frac{\left(\sum\limits_{i=1}^{a} \omega_i \bar{y}_{i..}\right)^2}{\sum\limits_{i=1}^{a} \omega_i}$$

In practice, we don't know the σ_w^2 or σ_i^2 terms, so we have to estimate them, giving the observed test statistic:

$$T_{Obs} = \sum_{i=1}^{a} \hat{\omega}_i \bar{y}_{i..}^2 - \frac{\left(\sum\limits_{i=1}^{a} \hat{\omega}_i \bar{y}_{i..}\right)^2}{\sum\limits_{i=1}^{a} \hat{\omega}_i}$$

The derivation of T and the estimation of the $\hat{\omega}_i$s is presented in Appendix C.

4.3 OB Approach & Simulation Results

The OB test aims to see how far away T_{Obs} is from the posterior distribution of:

$$\tilde{T} = \sum_{i=1}^{a} \hat{\omega}_i (\bar{y}_{i..} - \mu_i)^2 - \frac{\left(\sum\limits_{i=1}^{a} \hat{\omega}_i (\bar{y}_{i..} - \mu_i)\right)^2}{\sum\limits_{i=1}^{a} \hat{\omega}_i}$$

To do this, we sample from:

$$\frac{\bar{y}_{i..} - \mu_i}{\hat{v}_i} | \boldsymbol{Y} \sim t_{a(b-1)} \therefore \bar{y}_{i..} - \mu_i | \boldsymbol{Y} \sim \hat{v}_i t_{a(b-1)}$$

and calculate:

$$T_{OB} \equiv \sum_{i=1}^{a} t_{a(b-1)}^2 - \frac{\left(\sum\limits_{i=1}^{a} \sqrt{\hat{\omega}_i} t_{a(b-1)}\right)^2}{\sum\limits_{i=1}^{a} \hat{\omega}_i}$$

With enough samples of T_{OB}, we can simulate the distribution of T_{Obs} under H_0, and estimate the generalized p-value:

$$\mathcal{P}_{OB} = \frac{1}{M}\sum_{B=1}^{M} \mathrm{I}(T_{OB} > T_{Obs})$$

Table 4.1 lists the achieved type I error rates from simulations under various parameter settings. In the table: a is the number of groups, b is the number of blocks, the σ_i^2s are the measurement errors for each group, σ_w^2 is the between block error, and the n_{ij}s are the number of subsamples for each treatment/block combination. For economy of space, the entire list of the n_{ij} values for each simulation run is not displayed, but the various values the n_{ij}s take are displayed to convey if the simulation features large or small, and balanced or unbalanced subsample sizes. We present both the type I error rate from the OB test as well as that from the χ^2 test when we know the true variance parameters as an upper-bound for how well we can do. Each simulation was run 2000 times with 5000 draws of T_{OB} used to approximate the sampling distribution of the observed test statistic under H_0. For each simulation, the nominal significance level is $\alpha = 0.05$.

Table 4.1: Objective Bayes Simulation Table

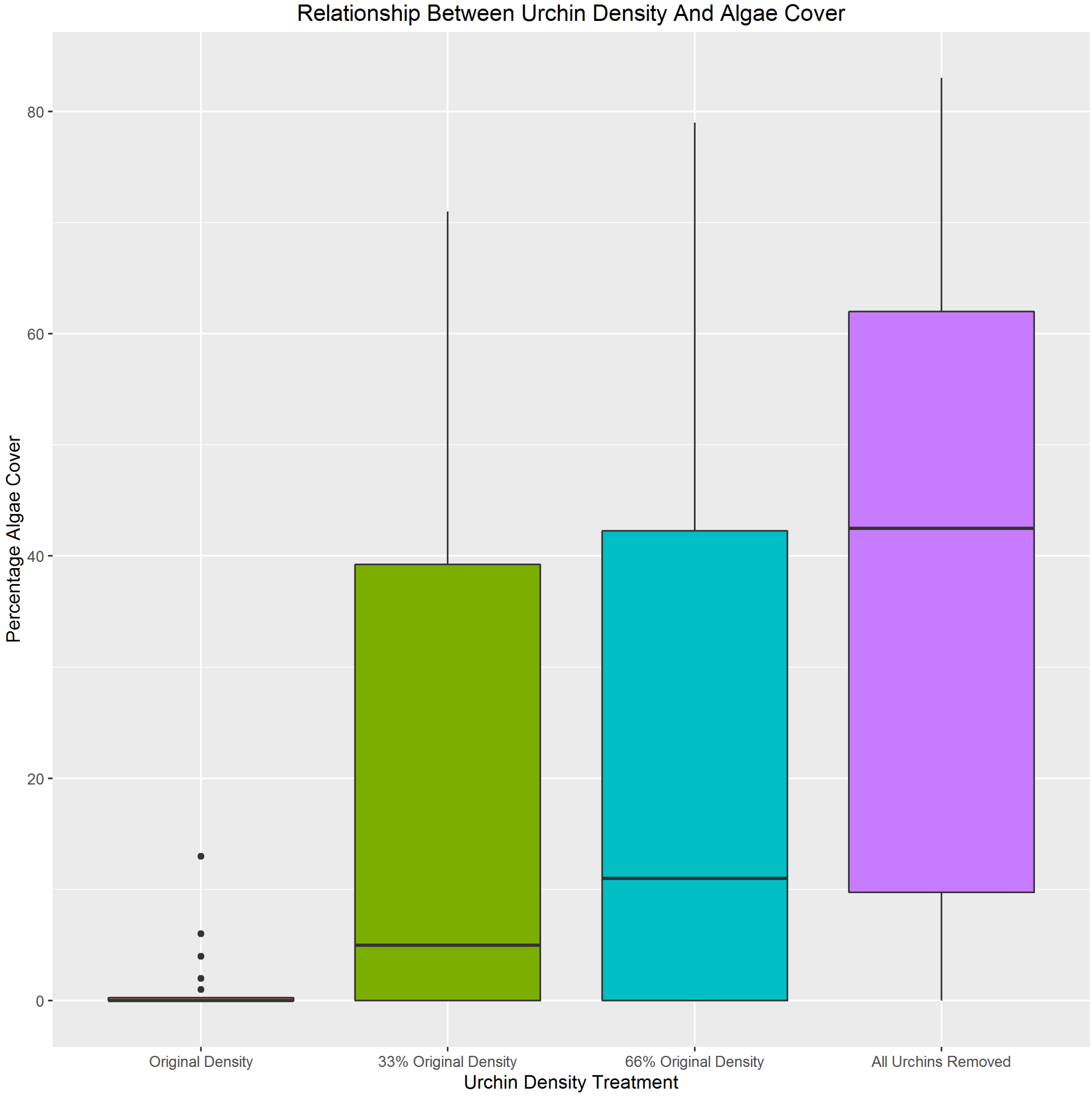
a	b	σ_i^2's	σ_w^2	n_{ij}'s	OB Type I Error Rate	Known Variance Type I Error Rate
3	4	1,1,1	16	40	0.0155	0.0515
3	4	1,1,1	0.25	40	0.02	0.059
3	10	1,1,1	16	40	0.039	0.0525
3	10	1,1,1	0.25	40	0.0355	0.0525
5	5	1,4,9,16,25	16	40	0.0305	0.0525
5	5	1,4,9,16,25	0.25	40	0.023	0.0485
5	5	1,4,9,16,25	16	40,20	0.0285	0.0485
5	5	1,5,9,16,25	0.25	40,20	0.032	0.0585
5	5	1,4,9,4,1	16	40	0.0275	0.059
5	5	1,4,9,4,1	0.25	40	0.0275	0.05
5	5	1,4,9,4,1	16	40,20	0.03	0.0575
5	5	1,4,9,4,1	0.25	40,20	0.03	0.053
3	50	1,1,1	16	40	0.0585	0.063
3	50	1,1,1	0.25	40	0.039	0.0405
3	50	1,4,9	16	40	0.465	0.048
3	50	1,4,9	0.25	40	0.049	0.0515
3	50	1,4,9	16	40,20,10	0.04	0.0445
3	50	1,4,9	0.25	40,20,10	0.0475	0.0495
3	50	1,4,9	16	5,10,15	0.048	0.048
3	50	1,4,9	0.25	5,10,15	0.0475	0.0485
3	20	1,4,9	16	20,10,15	0.0455	0.0505
3	20	1,4,9	0.25	20,10,15	0.048	0.0575
3	10	1,4,9	16	5,10,15	0.027	0.0435
3	10	1,4,9	0.25	5,10,15	0.036	0.045
30	10	1,4,9,1,4,9,...,1,4,9	16	20,10,15	0.048	0.055
30	10	1,4,9,1,4,9,...,1,4,9	0.25	20,10,15	0.052	0.0565
30	50	1,4,9,1,4,9,...,1,4,9	16	20,10,15	0.056	0.0555
30	50	1,4,9,1,4,9,...,1,4,9	0.25	20,10,15	0.051	0.0505
3	50	1,25,100	225	40,20,10	0.041	0.042
3	50	1,25,100	0.25	40,20,10	0.052	0.052
3	5	1,25,100	225	40,20,10	0.0205	0.046
3	5	1,25,100	0.25	40,20,10	0.02	0.0525
3	5	1,25,100	225	5,10,8	0.0285	0.0585
3	5	1,25,100	0.25	5,10,8	0.0245	0.0535
3	5	1,4,9	16	5,10,8	0.016	0.041
3	5	1,4,9	0.25	5,10,8	0.0175	0.0385
3	20	1,4,9	16	5,10,8	0.045	0.0515
3	20	1,4,9	0.25	5,10,8	0.043	0.054
3	20	1,4,9	16	2,1,3	0.0515	0.059
3	20	1,4,9	0.25	2,1,3	0.044	0.0495

The OB approach produces p-values close to the nominal level, regardless of the presence of heteroscedasticity. It appears to be equally effective regardless of whether whether σ_w^2 is large or small, relative to the σ_i^2's, and works well for both the balanced and unbalanced cases. The number of groups and number of subsamples has little effect on the type I error rate in the cases considered. The one situation in which the model produces unsatisfactory performance is the case where we have 5 or less blocks. For these simulations, the OB test is conservative as it tends to produce p-values less than the nominal level. This result reflects the crux of the subsampling problem—to reduce the overall variability, we need to increase the number of blocks, not just the number of subsamples, hence the poor performance in the small number of blocks case even when the number of subsamples is large. Comparatively, increasing the number of blocks results in a rapid performance boost—in the cases featuring more than a handful of blocks, the OB type I error rate is typically within around 0.01 of the nominal level of significance, which is nearly in line with the achieved type I error rate for the χ^2 test when variance terms are known. This indicates that the OB approach reliably controls the type I error rate for these cases.

4.4 Example: Sea Urchin Grazing Experiment

Andrew and Underwood (1993) performed an experiment in which they altered the density of sea urchins in various subtidal regions in New South Wales, Australia, in order to examine the effect of sea urchin grazing on the percentage cover of filamentous algae. Their experiment is an example of RCBD with subsampling. The treatment of interest is sea urchin density, and there are four groups: control (original density at the site), 66% original density, 33% original density, and all urchins removed. The experiment was performed in four patches, which are considered random blocks, with five subsamples per group/patch combination. The response of interest is the percentage of algae coverage at each site. Hence, the sea urchin grazing experiment is an example of a balanced RCBD with subsampling experiment. Figure 4.1 shows the empirical distribution of algae coverage by group, and suggests the presence of heteroscedasticity:

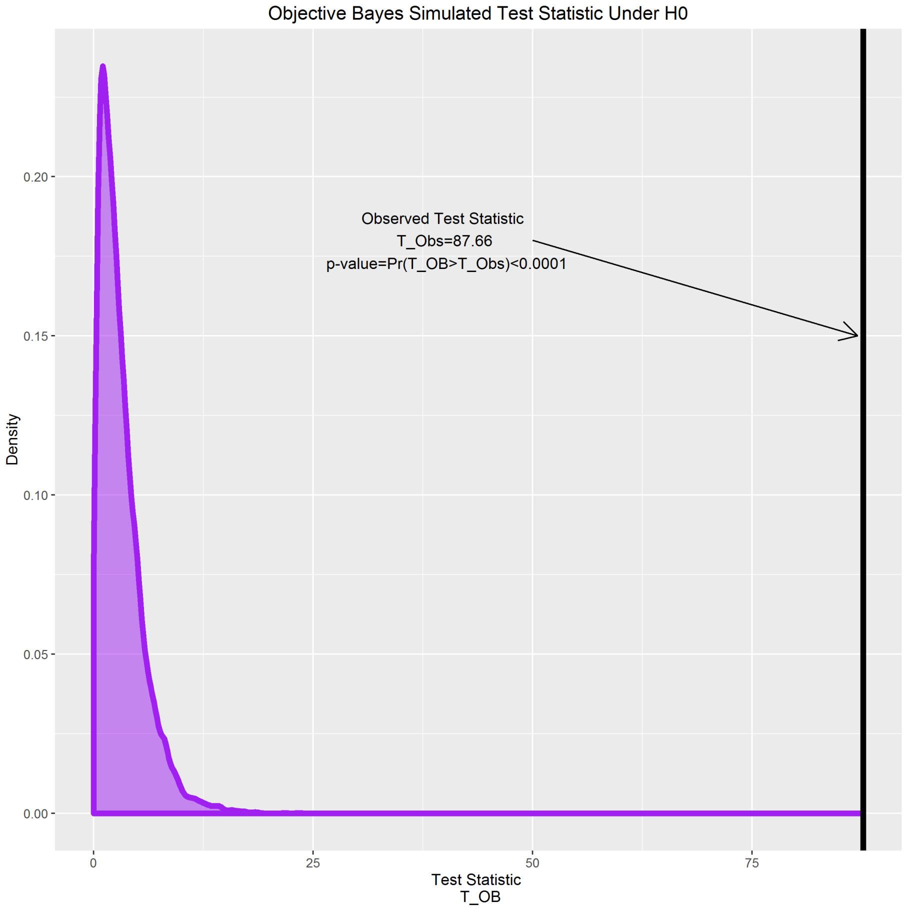
Figure 4.1: Empirical Distribution of Algae Coverage by Group



The side-by-side box plots of algae cover by urchin density tend to be right-skewed with the exception of the rather symmetric "all urchins removed" group. The control group exhibits far less variability than the others, and a Breusch-Pagan test confirms the presence of heteroscedasticity (p-value=0.0016). There may be a bit of a boundary problem since we can't have less than 0% algae coverage, but we'll proceed with our test of fixed effects using the OB approach for illustrative purposes:

Figure 4.2 depicts the distribution of the test statistic under the null model as approximated by the OB procedure. A vertical line shows the value of the observed test statistic $T_{Obs} = 87.66$.

Figure 4.2: Illustration of the OB Test



The density is based on 10,000 draws of T_{OB}, none of which were higher than the observed test statistic. This gives us a p-value of less than 0.0001, leading us to reject H_0 and conclude that at least two group means differ. The sites in which urchins were NOT removed typically have close to 0% algae coverage, since there are urchins to eat the algae. Conversely, the sites at which the urchins were completely removed typically have over 40% coverage as there are no urchins to eat the algae.

4.5 An Unweighted Test Statistic

In this section, we propose a difference-based statistic for a test of fixed effects in the RCBD with subsampling and heteroscedastic errors model. With two sources of error, we cannot naively look at something like:

$$T_{wrong} = \text{MSGrps} - \text{MSE}$$

as this approach would underestimate the variability by only including the measurement error σ_i^2. Ignoring the random block effect makes the group effects "appear to be more significant than they really are" (Christensen, 2011). Instead, we need to look at the appropriate error line. First, we average over the subsamples and then proceed as in one-way ANOVA. The model is:

$$\bar{y}_{ij.} = \mu_i + \epsilon_{ij}$$

where:

$$\epsilon_{ij} = \eta_j + \bar{e}_{ij.} \qquad \text{and} \qquad \epsilon_{ij} \sim N\left(0, \sigma_w^2 + \frac{\sigma_i^2}{n_{ij}}\right)$$

Define the relevant MS terms as:

$$\text{MSGrps} = \frac{b}{a-1}\sum_{i=1}^{a}(\bar{y}_{i..} - \bar{y}_{...})^2 \qquad \text{and} \qquad \text{“MSE”} = \frac{1}{a(b-1)}\sum_{i=1}^{a}\sum_{j=1}^{b}(\bar{y}_{ij.} - \bar{y}_{i..})^2$$

where:

$$\bar{\epsilon}_{i.} = \bar{\eta}_{.} + \bar{e}_{i..} \qquad \text{and} \qquad \bar{\epsilon}_{i.} \sim N\left(0, \frac{b\sigma_w^2 + \sum_{j=1}^{b}\frac{\sigma_i^2}{n_{ij}}}{b^2}\right)$$

and

$$\bar{\epsilon}_{\cdot\cdot} = \bar{\eta}_{\cdot} + \bar{e}_{\cdot\cdot\cdot} \qquad \text{and} \qquad \bar{\epsilon}_{\cdot\cdot} \sim N\left(0, \frac{ab\sigma_w^2 + \sum_{i=1}^{a}\sum_{j=1}^{b}\frac{\sigma_i^2}{n_{ij}}}{a^2b^2}\right)$$

Under H$_0$, E(MSGrps) = E("MSE"), so it is reasonable to conduct a test based on:

$$T_a = \text{MSGrps} - \text{"MSE"}$$

See Appendix D for details. To perform a significance test for a difference in group means, we can compare T_a to its asymptotic distribution:

$$N\left(0, \frac{2b\tilde{\tau}^4}{a(b-1)}\right)$$

and obtain the p-value for the test, $\mathcal{P}_a$. Appendix E details the derivation of the asymptotic distribution of T_a.

Chapter 5

Bayesian Significance Testing

5.1 The Bayesian P-value

In this section, we propose a general framework for Bayesian significance testing. In the one-way problem, the PB and OB methods approximate both the distribution of the test statistic in a very similar way. Note that nothing about Bayesian significance testing requires us to be objective, so we could implement similar ideas using conjugate Bayesian approaches. The Bayesian p-value is based on a significance test that essentially takes the null hypothesis to be the test statistic, and we broadly define it as:

$$\mathcal{P}_B = \Pr[g(\boldsymbol{\theta}|\boldsymbol{Y}) \leq g(\boldsymbol{\theta}_0|\boldsymbol{Y})|\boldsymbol{Y}]$$

which reduces to the posterior probability of being outside the posterior ellipse that has $\boldsymbol{\theta}_0$ on it. This is equivalent to checking if the observed test statistic is too large to reasonably be from the null distribution. In other words, using Bayesian p-values to make a decision in a significance test is equivalent to looking at highest posterior probability intervals. For the one-sided tests we are interested in, to calculate the p-value, we use:

$$\mathcal{P}_B = \Pr[T(\boldsymbol{Y}, \boldsymbol{\theta}) \geq T(\boldsymbol{Y}, \boldsymbol{\theta}_0) | \boldsymbol{Y}].$$

The PB uses a sampling distribution with $\boldsymbol{\theta}$ fixed and $\boldsymbol{Y}$ random, while OB uses a sampling distribution with $\boldsymbol{Y}$ fixed and $\boldsymbol{\theta}$ random. Despite arising from different philosophical arguments, both methods tend to result in empirically similar or identical tests. In Section 2, we look at some properties of OB significance testing that are valued from a frequentist perspective—the repeated sampling property and large sample property.

5.2 Properties of OB Significance Testing

5.2.1 Repeated Sampling Property

Suppose that $T =< T_1, T_2, \ldots, T_k >$ form a sequence of test statistics $T(\boldsymbol{Y}, \boldsymbol{X\beta}, \boldsymbol{\phi})$. For a sequence of parameters $< \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_k >$ from k populations with independent sample spaces, where the $\boldsymbol{\beta}_i$'s have the regular posterior distribution with a non-informative prior on $\boldsymbol{\beta}$. Under the null hypothesis, $T_0 =< T_{01}, T_{02}, \ldots, T_{0k} >$ forms a sequence of observed values $T(\boldsymbol{Y}, \boldsymbol{X}_0\boldsymbol{\gamma})$. For a significance test where we reject H$_0$ when $\mathcal{P}_B < \alpha$, when the null hypothesis is true, an OB significance test will reject that hypothesis an average of α times with repeated sampling.

Proof. For a one-sided test, define a sequence of indicator variables as:

$$\delta_i = \begin{cases} 1 & \text{if } T_{0i} \notin [0, u_i] \\ 0 & \text{otherwise} \end{cases}$$

where u_i is the critical value from a posterior distribution such that:

$$\Pr[T_i \in [0, u_i] | \boldsymbol{Y}_i] = 1 - \alpha.$$

Then we have:

$$\Pr\left(lim_{k\to\infty}\frac{\sum\limits_{i=1}^{k}\delta_i}{k} = \alpha \right) = 1$$

If we define:

$$\Delta_i = \begin{cases} 1 & \text{if } T_{0i} \notin [0, U_i] \\ 0 & \text{otherwise} \end{cases}$$

then Δ_i is a sequence of Bernoulli random variables. Because $\Pr(T_i \in [0, u_i] | \boldsymbol{Y}_i) = 1 - \alpha$, by construction we have:

$$\Pr(\Delta_i = 1) = \alpha \text{ for } i = 1, 2, \ldots, k$$

Moreover, δ_i is an observation of the Bernoulli random variable Δ_i. Because $< \Delta_1, \Delta_2, \ldots, \Delta_k >$ is a sequence of independent Bernoulli random variables with probability of success α, the result follows from the strong law of large numbers. □

5.2.2 Large Sample Property

Under the conditions described in Section 5.2.1, u_i, the critical value from the posterior distribution of the test statistic is a consistent estimate of Q_α, the α^{th} quantile from the limiting χ^2 distribution.

Proof. As the sample size goes to infinity, the consistent estimates of the nuisance parameters converge in probability to their population parameters:

$$\hat{\boldsymbol{\phi}} \xrightarrow{p} \boldsymbol{\phi}$$

From linear model theory:

$$T(\boldsymbol{Y}, \boldsymbol{X\beta}, \hat{\boldsymbol{\phi}}) \xrightarrow{d} \chi^2_{r(\boldsymbol{X})-r(\boldsymbol{X}_0)}$$

and since the χ^2 distribution is continuous:

$$u_i \xrightarrow{p} \mathrm{Q}_\alpha$$

□

Chapter 6

Discussion & Future Work

For one-way heteroscedastic ANOVA, the PB, OB, and predictive approaches estimate the sampling distribution of the observed test statistic very similarly, even for small samples, and for large samples, they are almost identical. The PB and OB approaches only differ by a $\hat{\phi}$ term, and they are asymptotically equivalent. Regarding future work, perhaps one could find a Bayesian approach that gives exactly the same test as the PB. In all simulations, the PB and OB approaches are practically identical. The PB and OB approaches control the type I error rate very extremely effectively, while the predictive and unweighted approaches are a bit more conservative in this regard. Similarly, the PB and OB have slightly better performance than the predictive and unweighted tests, in terms of power.

For the RCBD with subsampling and heteroscedastic errors model, we proposed two new solutions for testing $\mathrm{H}_0 = \mu_1 = \mu_2 = \cdots = \mu_a$. The OB test is able to control the type I error rate close to the nominal level, except for the cases where there are only a few blocks. For future work in this area, it would be interesting to investigate the PB and predictive approaches, and see how these and the unweighted test perform in terms of type I error rate and power.

Ideally, future research will find an elegant means of unifying the PB and OB approaches to significance testing. Otherwise, one will have to investigate these approaches to testing on a case-by-case basis. ANOVA problems present a clear path to implementing the approaches we considered, but regression problems and others remain unexplored.

Appendix A

One-Way Heteroscedastic ANOVA Observed Test Statistic

For testing $\mathrm{H}_0 = \mu_1 = \mu_2 = \cdots = \mu_a$, we set up a full model: $y_{ij} = \mu_i + e_{ij}$ and reduced model: $y_{ij} = \mu + e_{ij}$. Let $\boldsymbol{J}_{n_i}$ denote a $n_i \times 1$ matrix of ones, $\boldsymbol{J} \equiv \boldsymbol{J}_N$, and $\boldsymbol{I}_{n_i}$ be an $n_i \times n_i$ identity matrix. The one-way heteroscedastic ANOVA model fits in the linear model framework thusly:

Full Model: $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$

$$\begin{pmatrix} y_{11} \\ y_{12} \\ \vdots \\ y_{an_a} \end{pmatrix} = \begin{pmatrix} \boldsymbol{J}_{n_1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{J}_{n_2} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \boldsymbol{J}_{n_a} \end{pmatrix} \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_a \end{pmatrix} + \begin{pmatrix} e_{11} \\ e_{22} \\ \vdots \\ e_{an_a} \end{pmatrix}$$

where:

$$V \equiv \mathrm{Cov}(e) = \begin{pmatrix} \sigma^2 \boldsymbol{I}_{n_1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \sigma_2^2 \boldsymbol{I}_{n_2} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \sigma_a^2 \boldsymbol{I}_{n_a} \end{pmatrix}$$

Reduced Model: $\boldsymbol{Y} = \boldsymbol{J}\mu + \boldsymbol{e}$

The test statistic takes the form $T = \boldsymbol{Y}'(\boldsymbol{A} - \boldsymbol{A}_0)'\boldsymbol{V}^{-1}(\boldsymbol{A} - \boldsymbol{A}_0)\boldsymbol{Y}$. Letting $C()$ denote the column space: $C(\boldsymbol{VX}) \subset C(\boldsymbol{X})$, and thus $\boldsymbol{AY} = \boldsymbol{MY}$, where $\boldsymbol{M} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ is the perpendicular projection operator onto $C(\boldsymbol{X})$. Hence, we establish that:

$$\begin{aligned}
\boldsymbol{AY} = \boldsymbol{MY} &= \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} \\
&= \boldsymbol{X} diag\left(\frac{1}{n_1}, \frac{1}{n_2}, \ldots, \frac{1}{n_a}\right) \begin{pmatrix} y_{1.} \\ y_{2.} \\ \vdots \\ y_{a.} \end{pmatrix} \\
&= \boldsymbol{X} \begin{pmatrix} \bar{y}_{1.} \\ \bar{y}_{2.} \\ \vdots \\ \bar{y}_{a.} \end{pmatrix} = \begin{pmatrix} \bar{y}_{1.}\boldsymbol{J}_{n_1} \\ \bar{y}_{2.}\boldsymbol{J}_{n_2} \\ \vdots \\ \bar{y}_{a.}\boldsymbol{J}_{n_a} \end{pmatrix}
\end{aligned}$$

where *diag* indicates a diagonal matrix.

In the reduced model, $C(\boldsymbol{VX}_0) \not\subset C(\boldsymbol{X})$, so we have to calculate the estimate for μ directly. For ease of notation, let $\boldsymbol{J} = \boldsymbol{J}_N$ and $D(\sigma_i^2 \boldsymbol{I}_{n_i}) = \boldsymbol{V}$ so that we have:

$$\begin{aligned}
\boldsymbol{A}_0\boldsymbol{Y} &= \boldsymbol{X}_0\left[\boldsymbol{X}_0'\boldsymbol{V}^{-1}\boldsymbol{X}_0]\right]^{-1}\boldsymbol{X}_0'\boldsymbol{V}^{-1}\boldsymbol{Y} \\
&= \boldsymbol{J}\left[\boldsymbol{J}'D(\sigma_i^2\boldsymbol{I}_{n_i})^{-1}\boldsymbol{J}\right]^{-1}\boldsymbol{J}'D(\sigma_i^2\boldsymbol{I}_{n_i})^{-1}\boldsymbol{Y} \\
&= \boldsymbol{J}\left[\sum_{i=1}^{a}\frac{n_i}{\sigma_i^2}\right]^{-1}\sum_{i=1}^{a}\sum_{j=1}^{n_i}\frac{y_{ij}}{\sigma_i^2} = \boldsymbol{J}\left[\frac{1}{\sum\limits_{i=1}^{a}\frac{n_i}{\sigma_i^2}}\right]\sum_{i=1}^{a}\frac{y_{i.}}{\sigma_i^2} \\
&= \boldsymbol{J}\left[\frac{1}{\sum\limits_{i=1}^{a}\frac{n_i}{\sigma_i^2}}\right]\sum_{i=1}^{a}\frac{n_i}{\sigma_i^2}\bar{y}_{i.} = \boldsymbol{J}\left[\frac{\sum\limits_{i=1}^{a}\frac{n_i}{\sigma_i^2}\bar{y}_{i.}}{\sum\limits_{i=1}^{a}\frac{n_i}{\sigma_i^2}}\right]
\end{aligned}$$

If we let $\bar{y}^*_{..} = \frac{\sum\limits_{i=1}^{a} \frac{n_i}{\sigma_i^2}\bar{y}_{i.}}{\sum\limits_{i=1}^{a} \frac{n_i}{\sigma_i^2}}$ be our weighted average, then we have:

$$\boldsymbol{AY} - \boldsymbol{A}_0\boldsymbol{Y} = \begin{pmatrix} \bar{y}_{1.}\boldsymbol{J}_{n_1} \\ \bar{y}_{2.}\boldsymbol{J}_{n_2} \\ \vdots \\ \bar{y}_{a.}\boldsymbol{J}_{n_1} \end{pmatrix} - \left(\bar{y}^*_{..}\boldsymbol{J}\right)$$

so that:

$$T = \boldsymbol{Y}'(\boldsymbol{A} - \boldsymbol{A}_0)'\boldsymbol{V}^{-1}(\boldsymbol{A} - \boldsymbol{A}_0)\boldsymbol{Y} = \sum_{i=1}^{a} \frac{n_i}{\sigma_i^2}\bar{y}_{i.}^2 - \frac{\left(\sum\limits_{i=1}^{a} \frac{n_i}{\sigma_i^2}\bar{y}_{i.}\right)^2}{\sum\limits_{i=1}^{a} \frac{n_i}{\sigma_i^2}}$$

and estimating the σ_i^2's with their sample variances, we get the observed test statistic:

$$T_{Obs} = \sum_{i=1}^{a} \frac{n_i}{s_i^2}\bar{y}_{i.}^2 - \frac{\left(\sum\limits_{i=1}^{a} \frac{n_i}{s_i^2}\bar{y}_{i.}\right)^2}{\sum\limits_{i=1}^{a} \frac{n_i}{s_i^2}}$$

Appendix B

U Statistic Calculation for the Unweighted Approach

For consistent estimation of τ^4 and γ^4, we need an unbiased estimate for the σ_i^4 terms. U statistics (Hoeffding, 1948), provide a method for finding unbiased estimators. If we have a random sample of independent, identically distributed random variables, and can find an unbiased estimator for a parameter based on a subset of these observations, the U statistic is defined as the arithmetic average of this unbiased estimator across all possible subsamples that can give rise to that estimator. U statistics are minimum-variance unbiased estimators, are strongly consistent, and follow a normal distribution, asymptotically. (Hoeffding, 1961).

To use U statistics to estimate the σ_i^4's, we need $n_i \geq 4$ for every group. For one parameter σ^4:

$$\begin{aligned} x_1, x_2, \ldots, x_n &\overset{iid}{\sim} \mathrm{E}[x_1] = \mu, \mathrm{Var}(x_i) = \sigma^2 \\ h(x_1, x_2, x_3, x_4) &= \frac{1}{4}(x_1 - x_2)^2 (x_3 - x_4)^2 \\ \mathrm{E}[h(x_1, x_2, x_3, x_4)] &= \sigma^4 \end{aligned}$$

$$
\begin{aligned}
U &= \frac{1}{\binom{n}{4}} \sum_{1\le i<j<k<l\le N} h(x_i, x_j, x_k, x_l) \\
&= \frac{1}{\binom{n}{4}} \sum_{i<j<k<l} \frac{1}{4}(x_i - x_j)^2(x_k - x_l)^2 \\
&= \frac{1}{\binom{n}{4}} \frac{1}{4} \frac{1}{P_4^4} \sum_i \sum_j \sum_k \sum_l (x_i - x_j)^2(x_k - x_l)^2 \\
&= \frac{\sum\limits_i \sum\limits_j \sum\limits_k \sum\limits_l \left((x_i - \bar{x})^2(x_j - \bar{x})^2\right)\left((x_k - \bar{x})^2 + (x_l - \bar{x})^2\right)}{4n(n-1)(n-2)(n-3)} \\
&= \frac{n}{(n-1)(n-2)(n-3)} \sum_i \sum_k (x_i - \bar{x})^2(x_k - \bar{x})^2 \\
&= \frac{n}{(n-1)(n-2)(n-3)} \sum_i \sum_k (x_i^2 - 2x_i\bar{x} + \bar{x}^2)(x_k^2 - 2x_k\bar{x} + \bar{x}^2) \\
&= \frac{n}{(n-1)(n-2)(n-3)} \sum_k \left(\sum_i x_i^2 - n\bar{x}^2\right)(x_k^2 - 2x_k\bar{x} + \bar{x}^2) \\
&= \frac{n}{(n-1)(n-2)(n-3)} \left(\sum_i x_i^2 - n\bar{x}^2\right)\left(\sum_k x_k^2 - n\bar{x}^2\right) \\
&= \frac{n}{(n-1)(n-2)(n-3)} \left(\sum_i x_i^2 - n\bar{x}^2\right)^2 \\
&= \frac{n(n-1)}{(n-2)(n-3)}(s^2)^2
\end{aligned}
$$

So for each σ_i^4, we'll have:

$$\hat{\sigma}_i^4 = \frac{n_i(n_i - 1)}{(n_i - 2)(n_i - 3)}(s_i^2)^2$$

Appendix C

RCBD Observed Test Statistic

We use the GLS framework to derive our weighted test statistic for a test of fixed effects under the RCBD with subsampling and heteroscedastic errors model. Our null hypothesis is $\mathrm{H}_0 : \mu_1 = \mu_2 = \cdots = \mu_a$, that all group means are equal. Thus, the test statistic takes a similar form as before. $T = \bar{\boldsymbol{Y}}'(\bar{\boldsymbol{A}} - \bar{\boldsymbol{A}}_0)'\bar{\boldsymbol{V}}^{-1}(\bar{\boldsymbol{A}} - \bar{\boldsymbol{A}}_0)\bar{\boldsymbol{Y}}$:

Full Model

$$\bar{\boldsymbol{Y}} = \boldsymbol{I}_a\boldsymbol{\beta} + \boldsymbol{\xi}$$

$$\begin{pmatrix} \bar{y}_{1..} \\ \bar{y}_{2..} \\ \vdots \\ \bar{y}_{a..} \end{pmatrix} = \boldsymbol{I}_a \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_a \end{pmatrix} + \begin{pmatrix} \xi_1 \\ \xi_2 \\ \vdots \\ \xi_a \end{pmatrix}$$

$$\bar{\boldsymbol{V}} = \mathrm{Cov}(\boldsymbol{\xi}) = \begin{pmatrix} v_1 & 0 & \cdots & 0 \\ 0 & v_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & v_a \end{pmatrix}$$

Reduced Model

$$\bar{\boldsymbol{Y}} = \boldsymbol{J}_a\boldsymbol{\mu} + \boldsymbol{\xi}$$

To construct T, note that:

$$\begin{aligned}
\boldsymbol{J}_a'\bar{\boldsymbol{V}}^{-1}\boldsymbol{J}_a &= \sum_{i=1}^{a} \frac{b^2}{b\sigma_w^2 + \sum\limits_{j=1}^{b} \frac{\sigma_i^2}{n_{ij}}} \equiv \sum_{i=1}^{a} \omega_i \\
\bar{\boldsymbol{A}}_0 &= \boldsymbol{J}_a[\boldsymbol{J}_a'\bar{\boldsymbol{V}}^{-1}\boldsymbol{J}_a]^{-1}\boldsymbol{J}_a'\bar{\boldsymbol{V}}^{-1} = \frac{1}{\sum\limits_{i=1}^{a}\omega_i}\boldsymbol{J}_a\boldsymbol{J}_a'\bar{\boldsymbol{V}}^{-1} \\
&= \frac{1}{\sum\limits_{i=1}^{a}\omega_i}(\omega_1\boldsymbol{J}_a, \omega_2\boldsymbol{J}_a, \ldots, \omega_a\boldsymbol{J}_a) \\
\begin{pmatrix} \hat{\mu} \\ \hat{\mu} \\ \vdots \\ \hat{\mu} \end{pmatrix} &= \bar{\boldsymbol{A}}_0\bar{\boldsymbol{Y}} = \bar{\boldsymbol{A}}_0 \begin{pmatrix} \bar{y}_{1..} \\ \bar{y}_{2..} \\ \vdots \\ \bar{y}_{a..} \end{pmatrix} \\
&\therefore \\
\hat{\mu} &= \frac{\sum\limits_{i=1}^{a}\omega_i\bar{y}_{i..}}{\sum\limits_{i=1}^{a}\omega_i}
\end{aligned}$$

$$\begin{aligned}
\bar{\boldsymbol{A}}\boldsymbol{Y} &= \boldsymbol{I}_a \begin{pmatrix} \bar{y}_{1..} \\ \bar{y}_{2..} \\ \vdots \\ \bar{y}_{a..} \end{pmatrix} \\
&\therefore \\
\hat{\mu}_i &= \bar{y}_{i..} \\
T &= \bar{\boldsymbol{Y}}'(\bar{\boldsymbol{A}} - \bar{\boldsymbol{A}}_0)'\bar{\boldsymbol{V}}^{-1}(\bar{\boldsymbol{A}} - \bar{\boldsymbol{A}}_0)\bar{\boldsymbol{Y}} \\
&= \sum_{i=1}^{a}\omega_i\bar{y}_{i..}^2 - \frac{\left(\sum\limits_{i=1}^{a}\omega_i\bar{y}_{i..}\right)^2}{\sum\limits_{i=1}^{a}\omega_i}
\end{aligned}$$

Since we never know the true values of the variance components in applied problems, we derive their sample estimates in furtherance of finding the observed test statistic. We estimate the σ_i^2s with the sample variance terms (the s_i^2s), and for σ_w^2, we look at the expected block effect, plug in the estimates for the σ_i^2s, and solve for $\hat{\sigma}_w^2$:

$$\begin{aligned}
\hat{\mu} = \bar{y}_{...} = \frac{1}{a}\sum_{i=1}^{a}\bar{y}_{i..} &= \frac{1}{a}\sum_{i=1}^{a}\left(\mu + \bar{\eta}_j + \frac{1}{b}\sum_{j=1}^{b}\bar{e}_{ij.}\right) \\
&= \mu + \bar{\eta}_j + \frac{1}{ab}\sum_{i=1}^{a}\sum_{j=1}^{b}\bar{e}_{ij.} \\
\bar{y}_{ij.} - \bar{y}_{...} &= \mu - \mu + \eta_j - \bar{\eta}_j + \bar{e}_{ij.} - \frac{1}{ab}\sum_{i=1}^{a}\sum_{j=1}^{b}\bar{e}_{ij.} \\
&= \frac{b}{b}\eta_j - \frac{1}{b}\eta_j - \frac{1}{b}\sum_{\substack{t=1\\t\neq j}}^{b}\eta_t + \bar{e}_{ij.} - \frac{1}{ab}\bar{e}_{ij.} - \frac{1}{ab}\sum_{\substack{u=1\\u\neq i}}^{a}\sum_{\substack{t=1\\t\neq j}}^{b}\bar{e}_{ut.} \\
\mathrm{E}\left(\bar{y}_{ij.} - \hat{\mu}\right)^2 &= \left(\frac{b-1}{b}\right)^2\sigma_w^2 + \left(\frac{b-1}{b^2}\right)\sigma_w^2 \\
&\quad + \left(\frac{ab-1}{ab}\right)^2\frac{\sigma_i^2}{n_{ij}} + \left(\frac{1}{ab}\right)^2\sum_{\substack{u=1\\u\neq i}}^{a}\sum_{\substack{t=1\\t\neq j}}^{b}\frac{\sigma_i^2}{n_{ut}} \\
&= \frac{b^2 - 2b + 1 + b - 1}{b^2}\sigma_w^2 \\
&\quad + \frac{a^2b^2 - 2ab + 1}{a^2b^2}\frac{\sigma_i^2}{n_{ij}} + \frac{1}{a^2b^2}\sum_{\substack{u=1\\u\neq i}}^{a}\sum_{\substack{t=1\\t\neq j}}^{b}\frac{\sigma_i^2}{n_{ut}} \\
&= \frac{b-1}{b}\sigma_w^2 + \frac{ab-2}{ab}\frac{\sigma_i^2}{n_{ij}} + \frac{1}{a^2b^2}\sum_{i=1}^{a}\sum_{j=1}^{b}\frac{\sigma_i^2}{n_{ij}} \\
\mathrm{E}\sum_{i=1}^{a}\sum_{j=1}^{b}\left(\bar{y}_{ij.} - \hat{\mu}\right)^2 &= \sum_{i=1}^{a}\sum_{j=1}^{b}\mathrm{E}\left(\bar{y}_{ij.} - \hat{\mu}\right)^2 \\
&= \sum_{i=1}^{a}\sum_{j=1}^{b}\left(\frac{b-1}{b}\sigma_w^2 + \frac{ab-2}{ab}\frac{\sigma_i^2}{n_{ij}} + \frac{1}{a^2b^2}\sum_{i=1}^{a}\sum_{j=1}^{b}\frac{\sigma_i^2}{n_{ij}}\right) \\
&= a(b-1)\sigma_w^2 + \frac{ab-2}{ab}\sum_{i=1}^{a}\sum_{j=1}^{b}\frac{\sigma_i^2}{n_{ij}} + \frac{1}{ab}\sum_{i=1}^{a}\sum_{j=1}^{b}\frac{\sigma_i^2}{n_{ij}}
\end{aligned}$$

$$= a(b-1)\sigma_w^2 + \frac{ab-1}{ab}\sum_{i=1}^{a}\sum_{j=1}^{b}\frac{\sigma_i^2}{n_{ij}}$$

$$\therefore \hat{\sigma}_w^2 = \begin{cases} \frac{1}{a(b-1)}\left(\sum\limits_{i=1}^{a}\sum\limits_{j=1}^{b}(\bar{y}_{ij.}-\hat{\mu})^2 - \frac{ab-1}{ab}\sum\limits_{i=1}^{a}\sum\limits_{j=1}^{b}\frac{s_i^2}{n_{ij}}\right), & \text{if } \hat{\sigma}_w^2 > 0 \\ 0, & \text{otherwise} \end{cases}$$

Plugging in these estimates, we get:

$$\hat{\omega}_i = \frac{b^2}{b\hat{\sigma}_w^2 + \sum\limits_{j=1}^{b}\frac{s_i}{n_{ij}}}$$

so the observed test statistic under the null hypothesis is:

$$T_{Obs} = \sum_{i=1}^{a}\hat{\omega}_i\bar{y}_{i..}^2 - \frac{\left(\sum\limits_{i=1}^{a}\hat{\omega}_i\bar{y}_{i..}\right)^2}{\sum\limits_{i=1}^{a}\hat{\omega}_i}$$

Appendix D

Expected Value of MSGrps and "MSE" for the RCBD Model

The expected values of MSGrps and "MSE" reveal that a test of H_0 based on $T_a =$ MSGrps $-$ "MSE" is reasonable in the RCBD with subsampling and heteroscedastic errors model. To find the expected value of "MSE", note that:

$$\begin{aligned}
\bar{y}_{ij.} - \bar{y}_{i..} &= \mu_i + \epsilon_{ij} - \mu_i - \bar{\epsilon}_{i.} \\
&= \epsilon_{ij} - \bar{\epsilon}_{i.} = \epsilon_{ij} - \frac{1}{b}\sum_{j=1}^{b} \epsilon_{ij} \\
&= \frac{b}{b}\epsilon_{ij} - \frac{1}{b}\epsilon_{ij} - \frac{1}{b}\sum_{\substack{t=1 \\ t\neq j}}^{b} \epsilon_{it} \\
&= \left(\frac{b-1}{b}\right)\epsilon_{ij} - \frac{1}{b}\sum_{\substack{t=1 \\ t\neq j}}^{b} \epsilon_{it} \\
\mathrm{E}(\bar{y}_{ij.} - \bar{y}_{i..})^2 &= \left(\frac{b-1}{b}\right)^2 \left(\sigma_w^2 + \frac{\sigma_i^2}{n_{ij}}\right) + \frac{1}{b^2}\sum_{\substack{t=1 \\ t\neq j}}^{b} \left(\sigma_w^2 + \frac{\sigma_i^2}{n_{it}}\right)
\end{aligned}$$

$$
\begin{aligned}
&= \left(\frac{b-1}{b}\right)^2 \sigma_w^2 + \frac{b-1}{b^2}\sigma_w^2 + \left(\frac{b-1}{b}\right)^2 \frac{\sigma_i^2}{n_{ij}} + \frac{1}{b^2}\sum_{\substack{t=1\\t\neq j}}^{b} \frac{\sigma_i^2}{n_{ij}} \\
&= \frac{b^2-2b+1+b-1}{b^2}\sigma_w^2 + \frac{b^2-2b}{b^2}\frac{\sigma_i^2}{n_{ij}} + \frac{1}{b^2}\sum_{j=1}^{b}\frac{\sigma_i^2}{n_{ij}} \\
&= \frac{b-1}{b}\sigma_w^2 + \frac{b-2}{b}\frac{\sigma_i^2}{n_{ij}} + \frac{1}{b^2}\sum_{j=1}^{b}\frac{\sigma_i^2}{n_{ij}}
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{E}\sum_{i=1}^{a}\sum_{j=1}^{b}(\bar{y}_{ij.}-\bar{y}_{i..})^2 &= \sum_{i=1}^{a}\sum_{j=1}^{b}\left(\frac{b-1}{b}\sigma_w^2 + \frac{b-2}{b}\frac{\sigma_i^2}{n_{ij}} = \frac{1}{b^2}\sum_{j=1}^{b}\frac{\sigma_i^2}{n_{ij}}\right) \\
&= a(b-1)\sigma_w^2 + \frac{b-2}{b}\sum_{i=1}^{a}\sum_{j=1}^{b}\frac{\sigma_i^2}{n_{ij}} + \frac{1}{b}\sum_{i=1}^{a}\sum_{j=1}^{b}\frac{\sigma_i^2}{n_{ij}} \\
&= a(b-1)\sigma_w^2 + \frac{b-1}{b}\sum_{i=1}^{a}\sum_{j=1}^{b}\frac{\sigma_i^2}{n_{ij}}
\end{aligned}
$$

and hence:

$$
\mathrm{E}(\text{“MSE”}) = \sigma_w^2 + \frac{1}{ab}\sum_{i=1}^{a}\sum_{j=1}^{b}\frac{\sigma_i^2}{n_{ij}}
$$

To find the expected value of MSGrps, note that:

$$
\begin{aligned}
\bar{y}_{i..} - \bar{y}_{...} &= \mu_i - \bar{\mu}_{.} + \bar{\epsilon}_{i.} - \bar{\epsilon}_{..} \\
&= (\mu_i - \bar{\mu}_{.}) + \frac{1}{b}\sum_{j=1}^{b}\epsilon_{ij} - \frac{1}{ab}\sum_{i=1}^{a}\sum_{j=1}^{b}\epsilon_{ij} \\
&= (\mu_i - \bar{\mu}_{.}) + \frac{a}{ab}\sum_{j=1}^{b}\epsilon_{ij} - \frac{1}{ab}\sum_{j=1}^{b}\epsilon_{ij} - \frac{1}{ab}\sum_{\substack{u=1\\u\neq i}}^{a}\sum_{j=1}^{b}\epsilon_{uj}
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{E}(\bar{y}_{i..} - \bar{y}_{...})^2 = \mathrm{E}(\bar{y}_{i..} - \bar{y}_{...})^2 &= (\mu_i - \bar{\mu}_i)^2 + \left(\frac{a-1}{ab}\right)^2 \sum_{j=1}^{b}\left(\sigma_w^2 + \frac{\sigma_i^2}{n_{ij}}\right) \\
&+ \left(\frac{1}{ab}\right)^2 \sum_{\substack{u=1\\u\neq i}}^{a}\sum_{j=1}^{b}\left(\sigma_w^2 + \frac{\sigma_{s_u}^2}{n_{uj}}\right) \\
&= (\mu_i - \bar{\mu}_i)^2 + \left(\frac{a-1}{ab}\right)^2\left(b\sigma_w^2 + \sum_{j=1}^{b}\frac{\sigma_i^2}{n_{ij}}\right)
\end{aligned}
$$

$$
\begin{aligned}
&+\left(\frac{1}{ab}\right)^2\left((a-1)b\sigma_w^2+\sum_{\substack{u=1\\u\neq 1}}^{a}\sum_{j=1}^{b}\frac{\sigma {s_u}^2}{n_{uj}}\right)\\
&=(\mu_i-\bar{\mu}_{.})^2+\frac{a^2-2a+1}{a^2b^2}b\sigma_w^2+\frac{a-1}{a^2b^2}b\sigma_w^2\\
&\quad+\frac{a^2-2a+1}{a^2b^2}\sum_{j=1}^{b}\frac{\sigma_i^2}{n_{ij}}+\frac{1}{a^2b^2}\sum_{\substack{u=1\\u\neq i}}^{a}\sum_{j=1}^{b}\frac{\sigma_{s_u}^2}{n_{uj}}\\
&=(\mu_i-\bar{\mu}_{.})^2+\frac{a^2-2a+1+a-1}{a^2b^2}b\sigma_w^2\\
&\quad+\frac{a^2-2a}{a^2b^2}\sum_{j=1}^{b}\frac{\sigma_i^2}{n_{ij}}+\frac{1}{a^2b^2}\sum_{i=1}^{a}\sum_{j=1}^{b}\frac{\sigma_i^2}{n_{ij}}\\
&=(\mu_i-\bar{\mu}_{.})^2+\frac{a-1}{ab}\sigma_w^2\\
&\quad+\frac{a-2}{ab^2}\sum_{j=1}^{b}\frac{\sigma_i^2}{n_{ij}}+\frac{1}{a^2b^2}\sum_{i=1}^{a}\sum_{j=1}^{b}\frac{\sigma_i^2}{n_{ij}}
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{E}\sum_{i=1}^{a}(\bar{y}_{i..}-\hat{\mu})^2&=\sum_{i=1}^{a}(\mu_i-\bar{\mu}_{.})^2+\frac{a-1}{b}\sigma_w^2\\
&\quad+\frac{a-2}{ab^2}\sum_{i=1}^{a}\sum_{j=1}^{b}\frac{\sigma_i^2}{n_{ij}}+\frac{1}{ab^2}\sum_{i=1}^{a}\sum_{j=1}^{b}\frac{\sigma_i^2}{n_{ij}}\\
&=\sum_{i=1}^{a}(\mu_i-\bar{\mu}_{.})^2+\frac{a-1}{b}\sigma_w^2+\frac{a-1}{ab^2}\sum_{i=1}^{a}\sum_{j=1}^{b}\frac{\sigma_i^2}{n_{ij}}
\end{aligned}
$$

and hence:

$$\mathrm{E(MSGrps)}=\frac{b}{a-1}\sum_{i=1}^{a}(\mu_i-\bar{\mu}_{.})^2+\sigma_w^2+\frac{1}{ab}\sum_{i=1}^{a}\sum_{j=1}^{b}\frac{\sigma_i^2}{n_{ij}}$$

Under $\mathrm{H}_0:\mu_1=\mu_2=\cdots=\mu_a$, all μ_is are equal, so $\mu_i=\bar{\mu}_{.}$ and:

$$\mathrm{E(MSGrps)}=\sigma_w^2+\frac{1}{ab}\sum_{i=1}^{a}\sum_{j=1}^{b}\frac{\sigma_i^2}{n_{ij}}=\mathrm{E[\text{“MSE”}]}$$

Hence, it would be reasonable to do a test based on:

$$T_a=\mathrm{MSGrps}-\text{“MSE”}$$

where we see how far away T_a is from zero.

Appendix E

Asymptotic Distribution of T_a for the RCBD Model

We consider the RCBD with subsampling and heteroscedastic errors model averaged over the subsamples so we can following the proof for the one-way balanced case in Akritas and Papadatos (2004) to find the asymptotic distribution of T_a. The number of blocks $b \geq 2$ is fixed, with $\bar{y}_{ij.}$ independent, $\mathrm{E}[\bar{y}_{ij.}] = \mu$ under H_0, and $0 < \nu_i^2 \equiv \mathrm{Var}(\bar{y}_{ij.}) = \sigma_w^2 + \frac{\sigma_i^2}{n_{ij}} < \infty$ so the random variables $\bar{y}_{ij.}$ have the same distribution for each row i. Now, assume:

$$\frac{1}{a}\sum_{i=1}^{a} \nu_i^4 \xrightarrow{a\to\infty} \tilde{\tau}^4 \in (0,\infty)$$

and that for some $\delta > 0$,

$$\sup_{a\geq 1} \frac{1}{a}\sum_{i=1}^{a} \left(\mathrm{E}\left|z_{i1}\right|^{2+\delta}\right)^2 < \infty$$

then we have:

$$T_a \xrightarrow[a\to\infty]{d} N\left(0, \frac{2b\tilde{\tau}^4}{a(b-1)}\right)$$

Proof. Under H_0, the μ_is are all equal, so without loss of generality, assume that they are equal to zero, and hence $\mathrm{E}[\bar{y}_{ij.}] = 0$. Define:

$$\mathcal{A} \equiv \begin{pmatrix} \boldsymbol{B}_1 & -c_3\boldsymbol{J}_b\boldsymbol{J}_b' & \cdots & -c_3\boldsymbol{J}_b\boldsymbol{J}_b' \\ -c_3\boldsymbol{J}_b\boldsymbol{J}_b' & \boldsymbol{B}_2 & \cdots & -c_3\boldsymbol{J}_b\boldsymbol{J}_b' \\ \vdots & \vdots & \vdots & \vdots \\ -c_3\boldsymbol{J}_b\boldsymbol{J}_b' & -c_3\boldsymbol{J}_b\boldsymbol{J}_b' & \cdots & \boldsymbol{B}_a \end{pmatrix}$$

with $\boldsymbol{B}_i$ being $b \times b$ matrices having elements b_{ij}:

$$b_{ij} = \begin{cases} \frac{c_1}{b} - c_2 - c_3, & \text{if } i = j \\ \frac{c_1}{b} - c_3, & \text{if } i \neq j \end{cases}$$

where

$$c_1 = \frac{N-1}{(N-a)(a-1)} \qquad c_2 = \frac{1}{N-a} \qquad c_3 = \frac{1}{N(a-1)}$$

If we define $\boldsymbol{Z} = \begin{pmatrix} \bar{y}_{11.} \\ \bar{y}_{12.} \\ \cdots \\ \bar{y}_{ab.} \end{pmatrix}$, then:

$$T_a = \text{MSGrps} - \text{“MSE”} = \boldsymbol{Z}'\mathcal{A}\boldsymbol{Z}$$

and

$$\mathrm{E}[\boldsymbol{Z}'\mathcal{A}\boldsymbol{Z}] = \mathrm{E}[\boldsymbol{Z}'\mathcal{A}_{\mathcal{D}}\boldsymbol{Z}]$$

where $\boldsymbol{A}_D = \begin{pmatrix} \boldsymbol{B}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{B}_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \boldsymbol{B}_a \end{pmatrix}$. We can establish the limiting distribution of T_a by by looking at that of:

$$\sqrt{a}\boldsymbol{Z}'\mathcal{A}_{\mathcal{D}}\boldsymbol{Z} = \sqrt{a}(\text{MSGrps} - \text{“MSE”})$$

as

$$a\mathrm{E}[\boldsymbol{Z}'\mathcal{A}\boldsymbol{Z} - \boldsymbol{Z}'\mathcal{A}_{\mathcal{D}}\boldsymbol{Z}] \leq \frac{2}{a(a-1)^2}\left(\sum_{i=1}^{a} \nu_i^2\right)^2 \xrightarrow{a\to\infty} 0.$$

The U statistic:

$$U = \frac{1}{\sqrt{a}(b-1)} \left[\left(\sum_{j=1}^{b} \bar{y}_{ij.} \right)^2 - \sum_{j=1}^{b} \bar{y}_{ij.}^2 \right]$$

has $\mathrm{E}[U] = 0$ and $\mathrm{Var}(U) = \frac{2b\sigma_i^4}{a(b-1)}$. Because $\sqrt{a}\boldsymbol{Z}'\mathcal{A}_{\mathcal{D}}\boldsymbol{Z} = \sum_{i=1}^{a} U_{a,i}$ we can show:

$$\mathrm{Var}(\sqrt{a}\boldsymbol{Z}'\mathcal{A}_{\mathcal{D}}\boldsymbol{Z}) = \frac{2b}{a(b-1)} \sum_{i=1}^{a} \nu_i^2 \xrightarrow{a\to\infty} \frac{2b\tilde{\tau}^4}{b-1}$$

where $\sum_{i=1}^{a} v_i^2 \xrightarrow{a\to\infty} \tilde{\tau}^4$, and since Lyapunov's condition holds, as for some $\delta > 0$:

$$\sum_{i=1}^{a} \mathrm{E}\left|U\right|^{2+\delta} = (a(b-1)^2)^{-1-\delta/2} \sum_{i=1}^{a} \left| \left(\sum_{j=1}^{b} \bar{y}_{ij.} \right)^2 - \sum_{j=1}^{b} \bar{y}_{ij.}^2 \right|^{2+\delta}$$

$$\leq b^{2+\delta} a^{-1-\delta/2} \sum_{i=1}^{a} \left(\mathrm{E}[|z_{i1}|^{2+\delta}] \right)^2 \xrightarrow{a\to\infty} 0$$

then by the central limit theorem:

$$T_a \xrightarrow[a\to\infty]{d} N\left(0, \frac{2b\tilde{\tau}^4}{a(b-1)}\right)$$

□

Part II

Early Identification of Binswanger's Disease Patients Using Random Forests

Chapter 1

Introduction

Vascular disease has a major impact on all forms of dementia, and reducing vascular risk factors is now recognized as a way to reduce the worldwide burden of dementia (Gorelick et al., 2011; Hachinski et al., 2006; Snyder et al., 2015). Vascular cognitive impairment dementia (VCID) has a heterogeneous nature, which impacts epidemiological studies and interferes with drug testing (Pantoni, 2010; Román et al., 2010). VCID can be dichotomized into small vessel disease (SVD) and large vessel disease (LVD). While LVD is characterized by strokes and can be diagnosed by neuroimaging, SVD has slowly progressive symptoms and findings on neuroimaging that overlap with neurodegenerative diseases and normal aging. Binswanger's disease (BD) is a SVD with extensive demyelination secondary to vascular disease that is characterized by hyperreflexia, gait imbalance, incontinence, and executive dysfunction (Bennett, Wilson, Gilley, & Fox, 1990; Caplan, 1995; Miller Fisher, 1989; Olszewski, 1962; Román, Erkinjuntti, Wallin, Pantoni, & Chui, 2002; Rosenberg, Kornfeld, Stovring, & Bicknell, 1979). The BD group is optimal for treatment trials since the natural history is more apparent, which is often difficult to discern in patients with random strokes (Erkinjuntti, Roman, Gauthier, Feldman, & Rockwood, 2004). One approach to this dilemma is to obtain a large number of biomarkers at the time of entry into the study in a group of suspected VCID patients and to follow them for two to five years to obtain a clinical diagnosis. Earlier, we showed that a Binswanger Disease Score (BDS) derived from multiple biomarkers could predict the BD diagnosis with over 80% accuracy (Rosenberg et al., 2015), but these results were not cross validated and thus overoptimistic. Recent improvements in computers enable a multimodal, data-driven approach, using increased numbers of factors. In this report we used Random Forests (RF) to calculate the probability that an individual patient belongs in the BD group. Several recent studies have used RF methods to diagnose Alzheimers disease (AD) based on MRI, PET and blood-based biomarkers (DeMarshall et al., 2016; Gray et al., 2013; Lebedev et al., 2014). We hypothesize that the RF algorithm is an improvement over BDS, exploratory factor analysis (EFA), and logistic regression.

Chapter 2

Methods

2.1 Data Sets

2.1.1 Established Diagnosis Dataset

62 patients with suspected VCI that were recruited from 2007–2010 formed the dataset used to train and test the statistical methods. Patients were seen in the Neurology Clinics at University of New Mexico Hospital and Albuquerque Veterans Medical Center. A test of competency was performed to assure that patients understood and consented to all study procedures. The University of New Mexico Human Research Review Committee approved the study. All patients underwent neurological examination, a full battery of neuropsychological tests, MRI, and lumbar puncture to obtain cerebrospinal fluid (CSF). They were followed for multiple years to ascertain the best clinical diagnosis. The diagnoses used in the study were: 1. multiple or single cerebral infarcts (MI), including lacunar infarcts limited to the basal ganglia; 2. BD or subcortical ischemic vascular disease (SIVD) when diffuse white matter (WM) involvement on MRI was associated with imbalance, hyperreflexia, and exec-

utive dysfunction; 3. leukoaraiosis (LA) when the etiology of WM changes on MRI could not be determined; and 4) AD based on elevated phosphorylated tau_{181} (PTau) in the CSF.

2.1.2 Provisional Diagnosis Dataset

A second cohort of 23 patients with suspected VCI recruited from 2012–2014 forms the “provisional diagnosis” dataset. They underwent the same tests as the first set, except all of the MRI studies were done on a 3T MRI (Siemens Corp.) rather than the 1.5T used in the MRI studies in the original dataset. The long-term clinical diagnoses are not known, but we determined a provisional diagnosis using the RF trained on the established diagnosis dataset.

2.2 Biomarkers

2.2.1 Neuropsychological Test Batteries

Cognitive tests were administered by trained research psychologists and scored according to standard procedures. Standardized (T) scores were calculated for each test. Averaged composite T-scores were calculated for executive function.

2.2.2 Magnetic Resonance Studies

Proton magnetic resonance spectroscopy imaging (^{1}H-MRSI) was performed on a 1.5T or 3.0T MRI scanner (Siemens Corp.) with a phase-encoded version of a point-resolved spectroscopy sequence (PRESS) with or without water pre-saturation (TR/TE=1500/135ms, FOV=220x220mm, slice thickness=15mm, circular k-space sampling (radius=24), total scan time=9min42s). The WM concentrations of total N-acetyl-containing compounds (NAA and N-acetylglutamylaspartate, together referred as NAA), choline-containing metabolites (CHO), and creatine + phosphocreatine (CR) are reported (Gasparovic et al., 2013). MR blood-brain barrier (BBB) measurements were performed with the dynamic contrast-enhanced MRI (DCEMRI), using Gadolinium-diethylenetriaminepentaacetic acid (Gd-DTPA; Magnevist, Bayer Corp.), as previously described (Taheri et al., 2011).

2.2.3 CSF and Blood

White matter lesions due to multiple sclerosis were ruled by measurements of albumin index, myelin basic protein and oligoclonal bands in the CSF. The inflammatory biomarkers, matrix metalloproteinases-2 (MMP-2) and MMP-9 were measured in the CSF and plasma by gelatin-substrate zymography (Candelario-Jalil et al., 2011). MMP-2 and MMP-9 indexes were calculated (Liuzzi et al., 2002). Measurements of AD proteins, amyloidβ_{142} (Aβ42), total tau, and PTau were made using assay kits (INNO-BIA AlzBio3, Innogenetics, Gent, Belgium) with the LUMINEX instrument (Luminex Corp. Austin, TX) in a laboratory that was part of the AD consortium.

2.3 Prediction Methods

2.3.1 Binswanger's Disease Score (BDS)

BDS was calculated from a heuristic combination of clinical, imaging, and CSF characteristics to indicate the likelihood that patients with clinical symptoms may have BD. The 10 items used in the original BDS are shown in Table 3.2. The BDS ranges from 0 to 10, with 10 indicating the highest expression of characteristics that are associated with BD. BDS is similar to the Delphi method as it is based exclusively on expert opinion.

2.3.2 Logistic Regression (GLM)

A generalized linear modeling approach to the classification problem that fits a regression model with a categorical response variable, logistic regression has the benefit of familiarity and interpretability (Christensen, 2006).

2.3.3 Exploratory Factor Analysis (EFA)

EFA is a variable reduction technique used with high-dimensional data. EFA produced a small set of latent factors which are linear combinations of the predictors, and these factor loadings are listed in Table 3.4.

2.3.4 Random Forests (RF)

RF is a supervised ensemble learning algorithm that is based on classification trees (Breiman, 2001; Breiman et al., 2001). Many classification trees (a “forest”) are fit (or “grown”) on bootstrapped samples of the original data. Each tree partitions the data based on a random subset of predictor variables in such a way as to try to get optimal separation between the BD and Other SVD groups. RF shows how much each variable contributes to classification accuracy by comparing how well the trees that include a variable predict compared to those that do not.

The RF algorithm is a powerful and efficient means of diagnosing patients early and it allows for the assessment of the value of biomarkers via variable importance. It can perform multiclass prediction, including other diagnoses, such as AD, with no additional burden of interpretation; this differs from logistic regression where proportional odds must be assumed for the multiclass case. RF is an attractive classification method because: 1. it automatically employs external CV by predicting a patient diagnosis based on the trees that did not include that patient in their construction, 2. it does not make assumptions other than that the sample data are representative of the population of interest, and 3. it is easy to implement.

2.3.5 Cross Validation (CV)

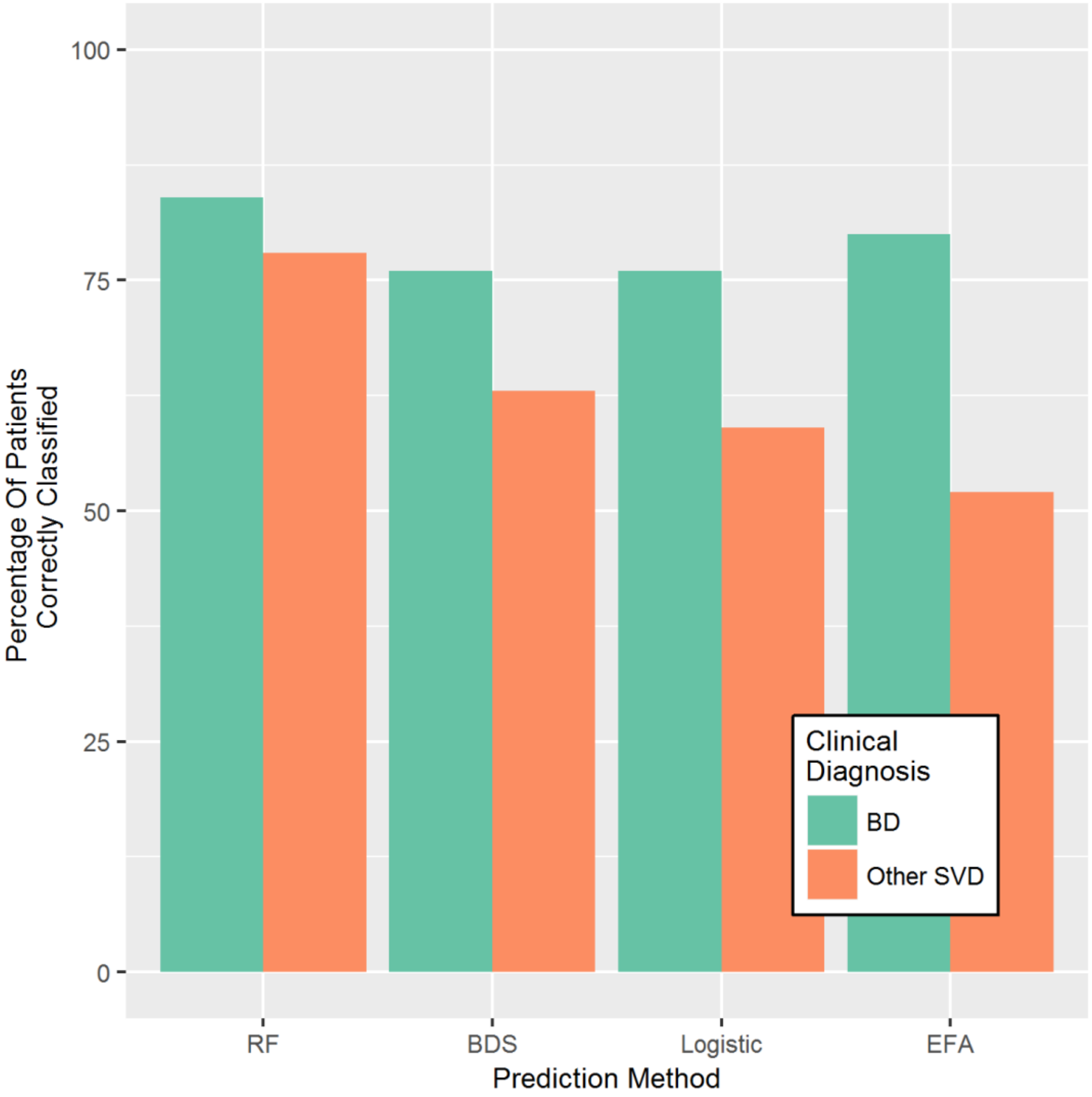
CV was used with logistic regression and EFA to better assess prediction accuracy by iteratively leaving out a proportion of observations when "training" a model or algorithm and then testing how well it predicts the left-out data by comparing the external predictions to the "true" clinical diagnosis. RF has CV built in and estimates the prediction error by comparing the so-called "out-of-bag" predictions to their respective known outcomes.

Chapter 3

Results

Comparing the classification accuracy of RF, BDS, logistic regression, and EFA methods, using the original 1.5T dataset, we found that RF and BDS are the two best methods. Figure 3.1 presents a visual representation of the confusion matrices for each classification method.

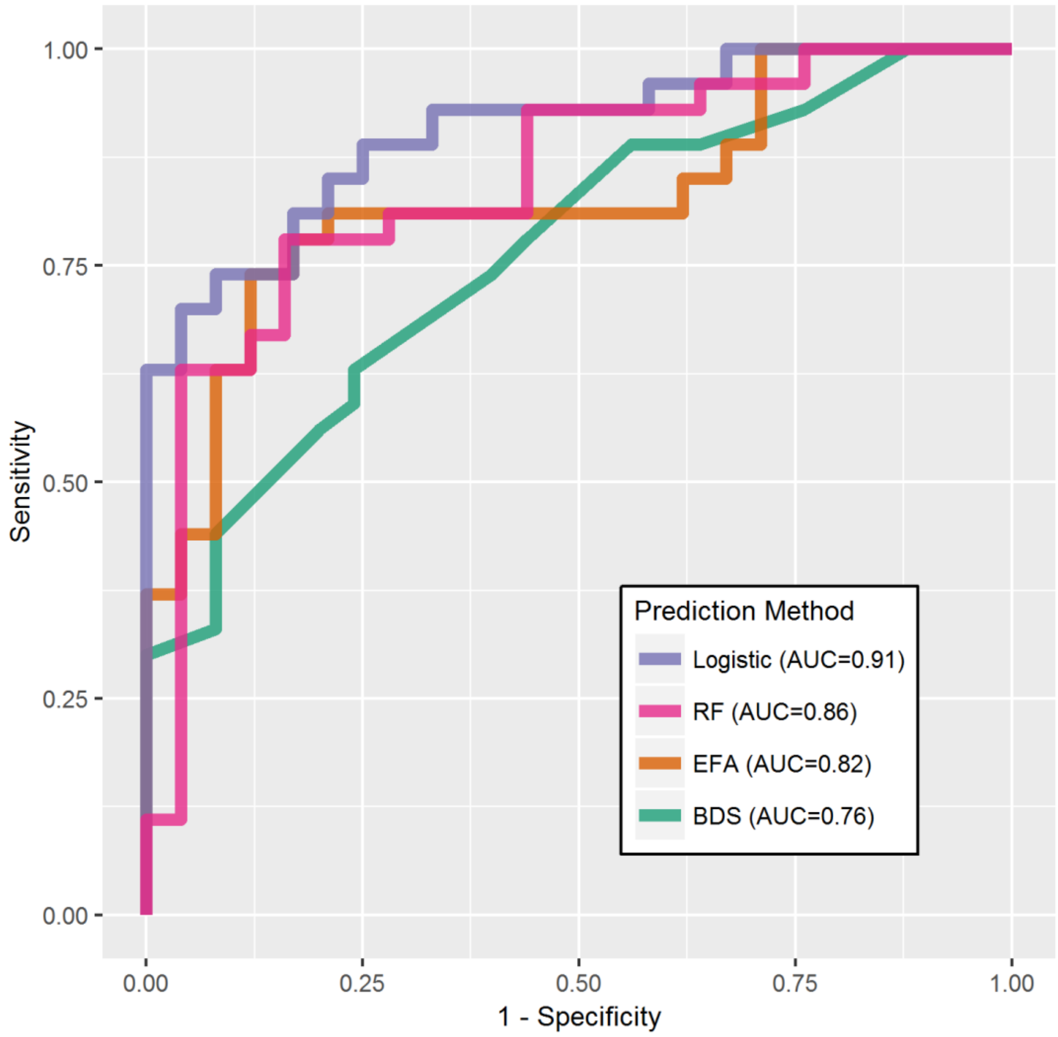
Figure 3.1: BD Classification Accuracy by Method



The methods are ordered from left to right by best to worst overall classification accuracy. Results are based on external predictions, where applicable. The green bars indicate how well each method classified individuals with a clinical diagnosis of BD while the orange bars show the classification accuracy for those with a clinical diagnosis of Other SVD. A method with perfect prediction would have both of these bars at 100%. Since we have binary prediction, subtracting the percentage of patients correctly classified from 100

Prediction accuracy for the BD class was superior to those diagnosed with some Other SVD, which is reasonable, as there is more heterogeneity in the Other SVD class. To determine the optimal cutoff value for classification with each method, ROC curves were used. Figure 3.2 provides a comparison of ROC curves:

Figure 3.2: ROC Curve Comparison



The ROC curves summarize the predictive ability for a range of thresholds by plotting sensitivity vs. one minus the specificity for each method. They can be used to assess classification ability, using area under the curve (AUC). A perfectly predicting model would have an AUC of 1 while a coin flip would have an AUC of 0.5. Note that the results here are not based on CV, except in the case of RF, for which out-of-bag predictions naturally accomplishes this. Each point on the ROC curve corresponds to a potential Pr(BD) cutoff, and the optimal cutoff is the one corresponding to the top-left most point on the ROC curve.

A tabular summary of method performance can be found in Table 3.1, where false positive, true positive ("sensitivity"), false negative, and true negative ("specificity") rates are delineated:

Table 3.1: ROC Prediction Accuracy Summary by Method

Method	%FP	%TP	%FN	%TN
RF	22	84	16	78
BDS	37	76	24	63
GLM	41	76	24	59
EFA	48	80	20	52

Method represents the classification method used, %FP is the false positive rate, %TP is the true positive rate or "sensitivity", %FN is the false negative rate, and %TN is the true negative rate or "specificity".

In the BDS, the characteristics most common to those with BD were: gait imbalance (93%), albumin index >6 (90%), MMP-2 index >0.01 (86%), executive function >45 (83%), hyperreflexia (79%), mean permeability >0.0018 (79%), and hypertension (76%). Table 3.2 breaks down the frequency of attributes contributing to BDS in the original 29 BD patients, where "proportion" is the relative frequency of patients exhibiting a specified characteristic:

Table 3.2: BDS Variable Importance

Characteristic	Frequency	Proportion
Gait Imbalance	27	0.93
Albumin >6	26	0.90
MMP-2 Index <0.01	25	0.86
Executive Function <45	24	0.83
Hyperreflexia	23	0.79
Mean Permeability >0.0018	23	0.79
Hypertension	22	0.76
NAA <12	20	0.69
Aβ42log(P$_{\tau 181}$) >150	15	0.52
Diabetes Mellitus	10	0.34

The logistic fit was determined by first fitting the full model to the original 13 predictors, and then performing stepwise model selection based on the Bayesian information criterion, a metric that balances fit and parsimony with the deviance (a measure of how well the model fits the data) versus a penalty for complexity (more predictors means a higher penalty) (Schwarz et al., 1978). The logistic fit is is presented in Table 3.3:

Table 3.3: Logistic Regression Model After Variable Selection

Parameter	Estimate	Std. Error	z value	Pr(>$\|z\|$)
(Intercept)	11.97	4.34	2.76	0.01
Executive Function	-0.13	0.06	-2.1	0.04
NAA	-0.96	0.37	-2.61	0.01
Permeability	1994.01	762.4	2.62	0.01

The reduced logistic model includes only 3 predictors: Executive Function, NAA, and mean permeability. Interpretation of logistic model coefficients is on the scale of the log-odds of having BD, a monotonic transformation of Pr(BD). For example, if there is no multicollinearity, then for each unit increase in NAA we expect a -0.96 increase in the log-odds of BD, in other words, the probability of BD decreases as NAA increases.

EFA extracted four factors from the nine variables considered. Factor loadings range from -1 to 1 with high positive values indicative that high values of the observed variable are related to high values of the factor, while high negative values indicate that low values of that observed variable are related to high values of the factor. Table 3.4 illustrates the factor loadings:

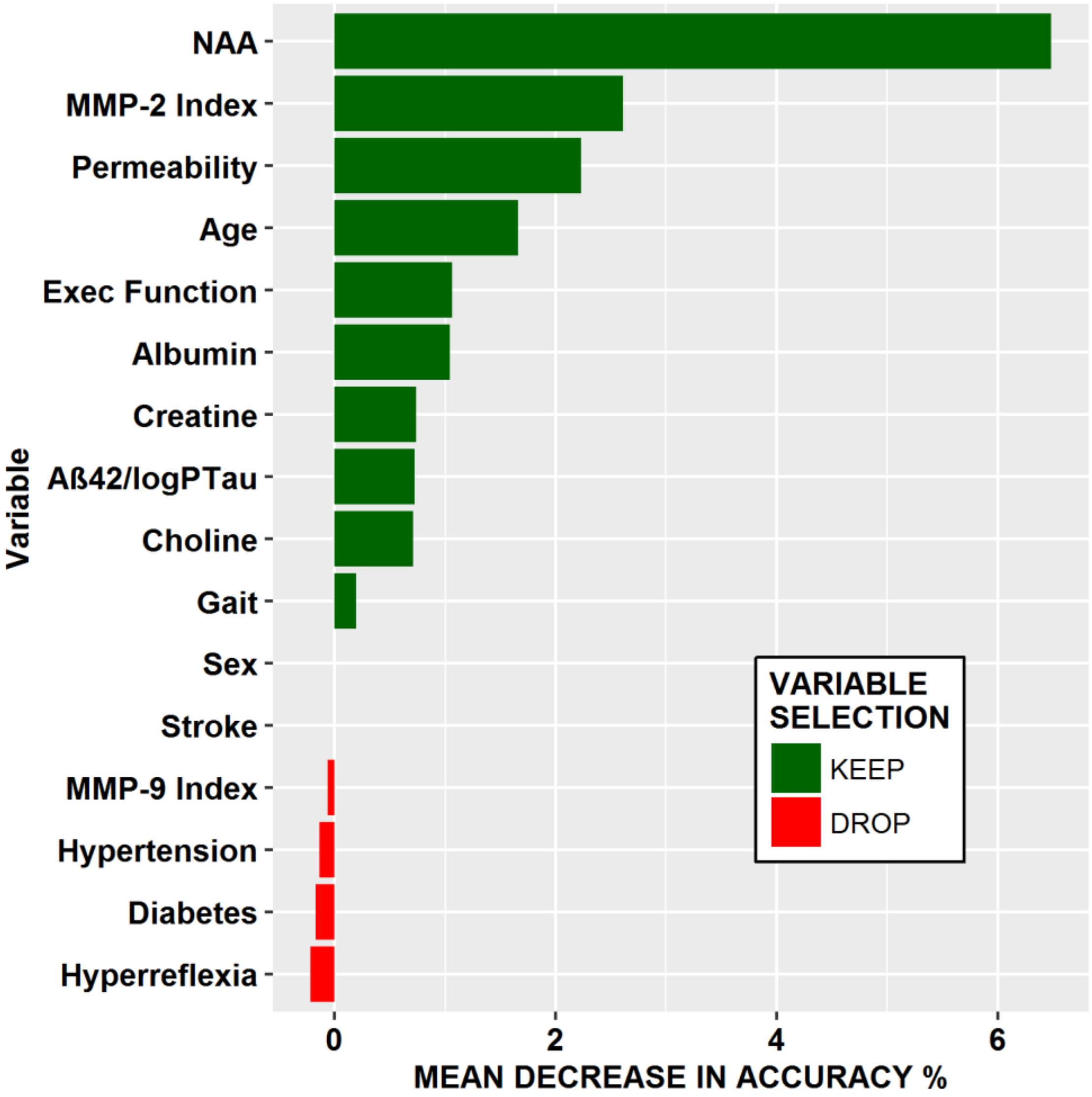
Table 3.4: EFA Factor Loadings.

Variable	F1	F2	F3	F4
Executive Function	0.41	-0.05	-0.21	0.1
NAA	**0.71**	0.26	0.43	0.17
Choline	**0.71**	0.27	0.35	-0.1
Creatine	**0.94**	0.13	0.01	0.2
Albumin Index	-0.22	**-0.74**	0.11	-0.47
Mean Permeability	-0.02	**-0.54**	-0.09	0.17
MMP-2 Index	0.12	**0.66**	-0.01	0.14
MMP-9 Index	0.1	0.04	0.02	0.57
Aβ42log(P$_{\tau 181}$)	0.06	-0.02	**0.73**	0.01

High values of Factor 1 (F1) are associated with high values of NAA, choline, and creatine. Factor 2 (2) will take high values when albumin index and mean permeability are low, or when the MMP-2 index is high. Factor 3 (F3) is positively associated with Aβ42log(P$_{\tau 181}$), and Factor 4 (F4) isn't related to much at all.

Variable importance derived from RF is shown in Figure 3.3, which shows the average drop in accuracy when a variable is left out of trees. Variables are ordered top to bottom from most to least important. Along the horizontal axis we have the average decrease in classification accuracy that occurs when a variable is left out of trees. The higher this decrease is, the more useful the variable is to prediction.

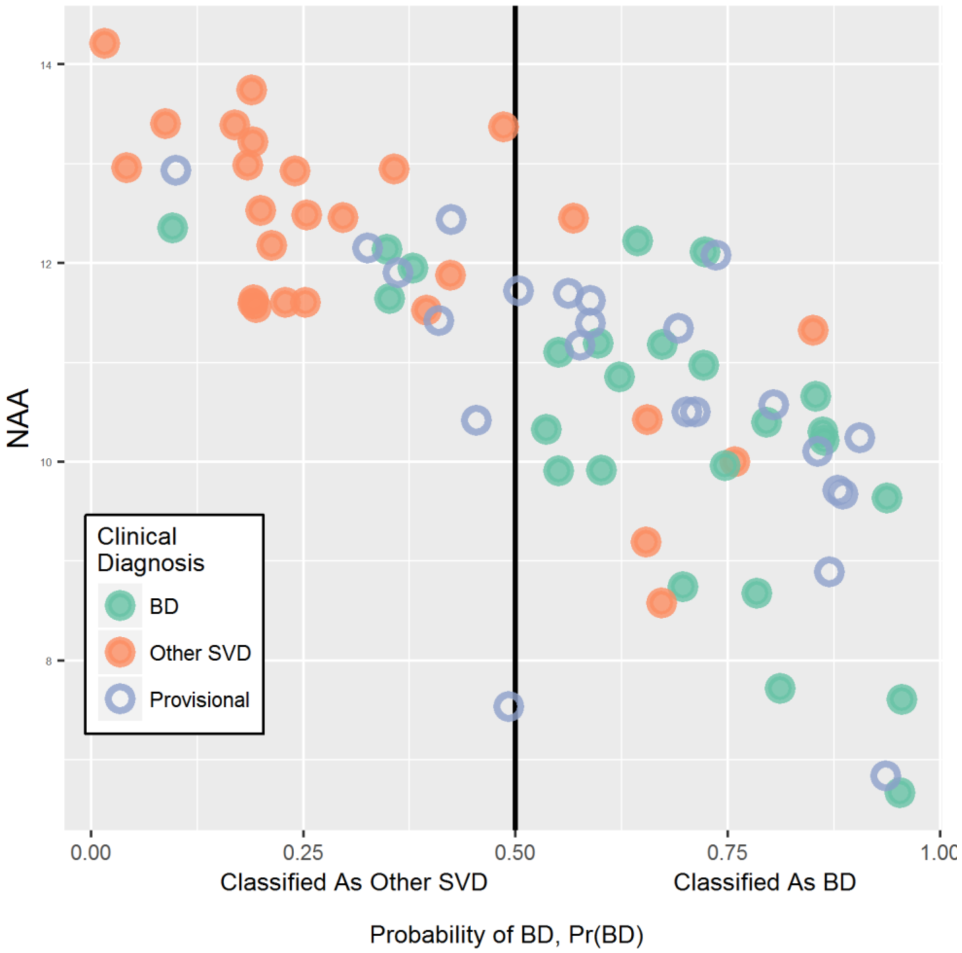
Figure 3.3: RF Variable Importance Plot



RF indicates that NAA is the most useful predictor, while the variables that hurt our ability to predict (sex, hypertension, hyperreflexia, stroke, diabetes, and MMP-9 Index) were dropped. After variable selection, the RF algorithm predicts BD vs Other SVD with 81% accuracy, implying we expect that 4 out of 5 diagnoses based only on the patients biomarkers obtained at the initial visit will match the clinicians diagnosis after a multi-year follow-up.

We used RF to predict the provisional diagnoses for the second cohort of 23 patients (those without long-term follow-up diagnoses). Although sometimes criticized as a "black box" method, with a little creativity, RF can be used show the relationship between individual biomarkers and outcome. For example, NAA, our most important predictor, has a strong, negative, linear association with Pr(BD), as shown in Figure 3.4:

Figure 3.4: Relationship Between NAA and the Probability of a Diagnosis of Binswangers Disease Pr(BD)



Since RF is not a model and can not be fit with a line as in logistic regression, we can still get an idea of how the predictor variables are related to our predicted probability of a BD diagnosis. Here we see that those with higher Pr(BD) tend to have lower NAA values. The aforementioned strong, negative linear association between the RF predicted Pr(BD) and NAA has a Pearsons r correlation coefficient of -0.75. This plot also shows the predictions for the new cohort of patients. Closed green circles represent patients with a clinical diagnosis of BD, closed orange circles represent those with a clinical diagnosis of Other SVD, and open purple circles represent the new cohort of patients with provisional diagnoses. The vertical line drawn at Pr(BD) = 0.50 is an optimal cut point (based on ROC analysis) for classifying patients as BD or Other SVD. Using this approach, four BD patients would be misclassified as Other SVD and six Other SVD patients misclassified as BD. In this new cohort, we see seven patients predicted to have Other SVD and 16 predicted to have BD, suggesting that this new cohort may have a higher proportion of BD patients than the original cohort. Individuals close to the vertical line are those for whom a diagnosis is nebulous, while the purple circle provisional patients farthest to the right are likeliest to have BD, making them the prime candidates for early inclusion in a clinical trial.

Chapter 4

Discussion

Our results show that RF can be used with multimodal biomarkers to predict the likelihood of a diagnosis of BD several years prior to a clinical diagnosis. This overcomes one of the principal impediments to treatment trials in VCID—the heterogeneous nature of the patients. BD, the small vessel progressive form, is the optimal form of VCID for clinical trials because it has a more predictable course than seen with multiple strokes, which tend to occur sporadically. RF prediction can be used to eliminate patients with white matter lesions of uncertain significance, which we have labeled as leukoaraiosis, and RF improved diagnostic accuracy in the subgroup of VCID patients with BD (Rosenberg et al., 2016; Snyder et al., 2015).

Approaches like BDS are static and do not allow for adaptability. In this report we compared the BDS with the RF algorithm and other data-driven approaches. Using a prior group of patients that had undergone long-term follow-up to form our original dataset, we showed that the set of biomarkers obtained at study entry could be used in an RF framework to classify a new group of patients as BD vs Other SVD with improved accuracy without the need for long-term follow-up. This study was another confirmation that RF could be used for personalized medicine, allowing early diagnosis in patients with VCID of the BD type and permitting treatment trials at an earlier stage of the illness.

Several of the biomarkers emerged as the most important in establishing the BD diagnosis, including structural damage to WM as shown by decreased levels of NAA, executive dysfunction on neuropsychological testing, disruption of the blood-brain barrier as indicated by increased albumin index and raised DCEMRI, and neuroinflammation as shown by reduced MMP-2 index. These factors provided the basis for the predictions and were derived using reports on BD in the literature. The heuristic BDS classifies patients well, but it is a less flexible process which cannot improve as new features or patients are included for analysis. In contrast, RF and other data-driven approaches open novel ways of analyzing large data sets such as those commonly found in medicine. RF has many advantages in prediction because it is not dependent on a parametric data model, affording both flexibility and convenience, as there is no need to impose assumptions beyond having observations that are representative of the process under study, which is inherent to any inferential procedure (Lebedev et al., 2014). RF avoids the pitfall of over fitting by implementing CV to ensure generalization to new observations, and it overcomes the problem of dimensionality by using random subsets of features in the construction of each tree that composes the ensemble forest. Without the need to estimate parameters as in a linear model, the number of predictors does not pose a problem, even in the extreme case where the number of predictors is greater than the sample size. Finally, RF can be used in an iterative approach—the established diagnosis dataset can be updated with members whose diagnoses become known in the future, improving the reliability of prediction for new patients.

In summary, this predictive strategy provides some important advantages for clinical treatment trials: 1. an individual patient can be sub-grouped on the basis of the pathophysiology, improving the chance of a successful trial in a smaller population, using an agent designed specifically for that pathophysiology, and 2. a patient can be diagnosed at an earlier stage of the illness, allowing the agent to have a better chance of success.

While our goal was to compare these methods as if they were in competition, it is also important to note the benefits of trying multiple reasonable approaches in any statistical analysis: if the methods agree, it gives more credibility to the result, as all paths lead to the same conclusion. In this data, for example, each method suggests that NAA and permeability are highly informative in determining the diagnosis of BD vs Other SVD.

In the original cohort, 43% of patients were eventually found to have BD. If Pr(BD) was used to select patients, it would be possible to reduce the inclusion of extraneous Other SVD patients in a clinical trial while enriching the sample with BD patients, improving the efficiency and the chance of obtaining a positive result.

The goal is to have an ongoing, iterative process in which initial RF predictions are made before clinical diagnosis is revealed through long-term follow-up. Our initial findings are promising in that RF agreed with clinicians provisional diagnoses in roughly 4 out of 5 patients. The RF will get stronger once a diagnosis for a provisionally diagnosed patient is confirmed and that patient is added to the established diagnosis dataset, in turn allowing an update to the RF.

An important consideration for personalized medicine is that no single biomarker was sufficient for classification. Although ^{1}H-MRS measurements of NAA were an excellent biomarker for structural damage, and more likely to be low in the BD group, a number of patients with other diagnoses had low NAA. Using the white matter hyperintensities (WMH) on FLAIR MRI, which are used in many large population-based studies, as a surrogate biomarker for BD, would also have failed because of the high percentage of normal elderly people with WMH. However, adding the BBB opening, which is indicated by the elevation of the albumin index and the increased permeability with DECMRI, improved diagnostic accuracy. Other biomarkers that proved useful in the diagnosis of BD included abnormal MMPs in the CSF, which indicates neuroinflammation, and impairment in executive functions.

A caveat for the present report is that diagnoses of patients in the provisional diagnosis group will not be known with the same level of certainty as in the established diagnosis group for several years. However, once they reach that level, they will be used to validate predictions in a refined RF. Because this is an early attempt to classify VCID patients, it is possible that the initial group of biomarkers will need to be modified for use in larger cohorts of patients required for collaborative studies. In the initial phases of the clinical trials, the patients with the highest probability from the RF method can be used in small, carefully selected patient groups. If there is success in these initial trials, the studies can be expanded to multiple centers with greater certainty of success. This would reduce the cost by reducing the number of patients that need to be studied.

Chapter 5

Acknowledgments

The work in this section was from a collaboration between the UNM Department of Mathematics and Statistics, Department of Neurology, and MIND Research Network. As the primary statistical consultant on this project, I was supervised by Erik Erhardt, who also contributed to writing and the analysis. Gary Rosenberg conceived the study as the primary investigator, participated in patient recruitment, and contributed to the writing. Jillian Prestopnik was the study coordinator, Jeffrey Thompson ran the CSF studies and assisted in the MRI and CSF data analysis, and Arvind Caprihan collected and analyzed MRI data. J.C. Adair and Janice Knoefel, study neurologists, helped recruit patients, and the work was funded by grants from the NIH RO1 NS 052305 and the US-Israeli Binational Research Foundation to GAR, with the study funded by NIH (OR 123456789). Analysis was carried out using R with extensive use of the random forest and ggplot2 packages (Liaw & Wiener, 2002; R Core Team, 2016; Wickham, 2009).

Appendix F

Table of Considered Variables

This project serves as an improvement on an initial attempt to classify Binswanger's patients (Rosenberg et al., 2015). The original analysis only looked at BDS and EFA, and it is worth noting that for unknown reasons the original analyst did not utilize all variables at their disposal for the BDS or EFA approaches. Table B.1 lists which variables are used in each method, as well as which were deemed to be of importance. As a key: "u" indicates that the feature was used in the model/algorithm and "M" indicates that the feature was both used and deemed important by the method.

Table F.1: Features Used to Select Patients Most Likely to have BD

Features	BDS	EFA	Logistic Regression	RF
I. Clinical Features				
1. Hypertension	u		u	u
2. Diabetes Mellitus	u		u	u
3. Hyperreflexia	u		u	u
4. Gait Imbalance	u		u	M
5. Stroke			u	u
6. Sex			u	u
7. Age			u	M
II. Neuropsychological Testing				
8. Executive Function	u	u	M	M
III. Metabolites in WM (H-MRSI)				
9. NAA	u	u	M	M
10. Choline		u	u	M
11. Creatine and Phosphocreatine		u	u	M
IV. Inflammation and BBB				
12. Albumin Index	u	u	u	M
13. Mean Permeability	u	u	M	M
14. MMP-2 Index	u	u	u	M
15. MMP-9 Index		u	u	u
V. Alzheimer's Biomarkers				
16. Aβ42log(P$_{\tau 181}$)	u	u	u	M

Part III

High-Throughput Gene Expression Analysis Under the Case-Cohort Study Design

Chapter 1

Background

1.1 The Case-Cohort Design

The case-cohort (CCH) study design is a prospective observational study design that blends the economy of case-control studies with the philosophical soundness of cohort studies. Proposed by Prentice (1986), CCH studies are similar to full cohort studies in that the exposures (predictor variables) are measured before outcomes (response variables). Used for survival analysis, cases are defined as observations that had an event, while controls are those that are censored at the time of the analysis.

CCH designs only consider a sample of the full cohort, which is referred to as the “subcohort”. The subcohort is augmented with all incident cases at the time of analysis. The CCH design is far more efficient than a full cohort analysis as it saves substantial time and money while sacrificing very little power. Especially when dealing with gene expression studies with rare outcomes, it makes little sense to measure an unnecessarily high number of controls.

A CCH analysis is best suited to data that is cheap to collect, but expensive to analyze or process. Taking a blood or tissue sample is quick and easy, but fully genotyping an individual from such a sample requires considerably more resources.

All of the philosophical benefits of a full cohort study are preserved in a CCH analysis. The prospective nature offers a clear narrative since the predictors are measured before an endpoint is reached. This affords several benefits—more reliable covariate information, the ability to calculate true incident rates, a reduction of selection bias and confounding, and permits the investigation of multiple outcomes on the same individuals. The CCH design is above a case-control study in the hierarchy of evidence, but lower than a clinical trial as it is still an observational study.

1.2 Survival Analysis Under the CCH Design

Researchers seek biomarkers that can inform if an individual is more or less likely to suffer negative health outcomes, or whether they will be receptive to treatment. The prevailing approach to survival analysis is the Cox proportional hazards (CPH) model (Cox, 1972):

$$h(t|\boldsymbol{x}_i) = h_0(t)e^{\boldsymbol{x}_i\boldsymbol{\beta}}$$

where t is time, $\boldsymbol{x}_i$ is a row vector of predictors for observation i, $\boldsymbol{\beta}$ is a vector of coefficients, and $h_0(t)$ is the baseline hazard function. Cox (1975) showed that we can do maximum likelihood estimation, significance testing, and interval estimation for $\boldsymbol{\beta}$ via the partial likelihood:

$$\mathcal{L}(\boldsymbol{\beta}) = \prod_{i=1}^{n} \left\{ \frac{e^{\boldsymbol{x}_i\boldsymbol{\beta}}}{\sum\limits_{j\in R(t_i)} e^{\boldsymbol{x}_j\boldsymbol{\beta}}} \right\}^{\delta_i}$$

where δ_i indicates the status of individual i (0 if they are censored, 1 if they had an event), and $R(t_i)$ indicates the risk set, which includes all individuals in the study at time t_i that are still liable to have an event.

In a CCH study, we weight the likelihood to account for the sampling scheme:

$$\mathcal{L}(\boldsymbol{\beta}) = \prod_{i=1}^{n} \left\{ \frac{e^{\boldsymbol{x}_i \boldsymbol{\beta}}}{\omega_i e^{\boldsymbol{x}_i \boldsymbol{\beta}} + \sum\limits_{\substack{j \neq i \\ j \in R(t_i) \cap S}} \omega_j e^{\boldsymbol{x}_j \boldsymbol{\beta}}} \right\}^{\delta_i}$$

where ω_i is the weight for individual i and the summation in the denominator only includes individuals at risk in the subcohort. Individual i can either be a case from inside or outside of the subcohort.

Barlow, Ichikawa, Rosner, and Izumi (1999) outline three popular methods for CCH analysis—that of Prentice (1986), Self and Prentice (1988), and Barlow himself. Table 1.1 is from Barlow's paper and describes the ω_i's under the three approaches, where π is the proportion of the full cohort taken as the subcohort:

Table 1.1: CCH Weighting Schemes

Outcome type and timing	**Prentice**	**Self-Prentice**	**Barlow**
Case outside subcohort before failure	0	0	0
Case outside subcohort at failure	1	0	1
Case in subcohort before failure	1	1	$1/\pi$
Case in subcohort at failure	1	1	1
Subcohort control	1	1	$1/\pi$

Barlow's method appears most sensible as the weights are proportional to the subcohort size. Figure 1.1 offers a visual representation of the example of Barlow's weighting when the subcohort is $\pi = 10\%$ of the full cohort.

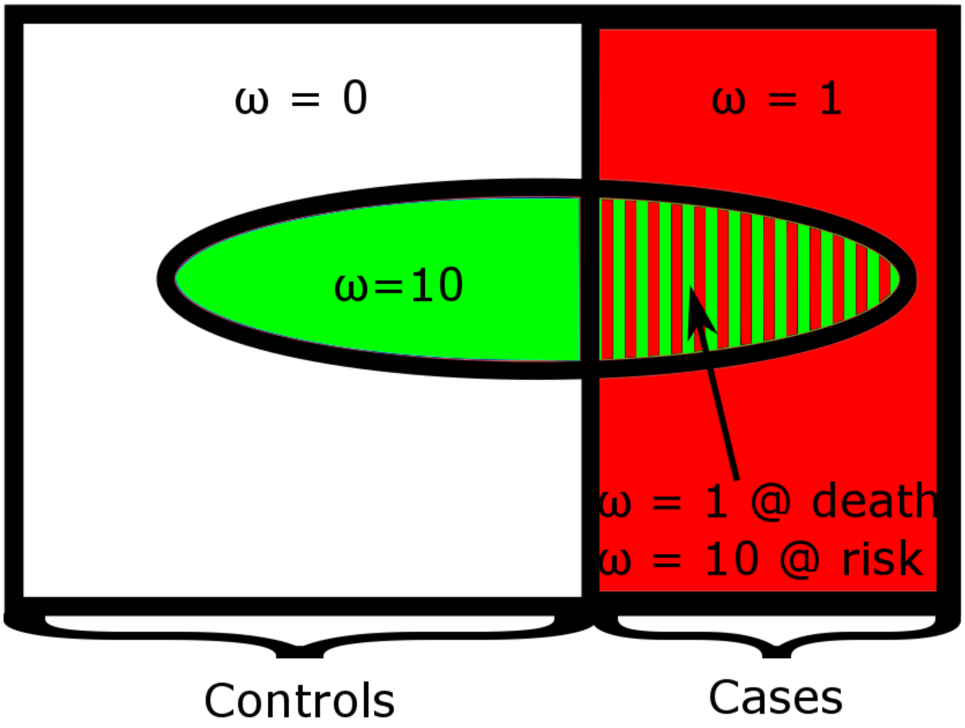


Figure 1.1: Example of Barlow's Weighting Method ($\pi = 0.10$)

In this example, Barlow's method weighs controls in the subcohort as if they are worth 10 people. Since all cases are included in this design, cases outside the subcohort are given a weight of 1. Subcohort cases are treated in two ways—before their event they are weighted with a factor of 10, just like the controls, but at the time of their event, they are treated like the cases outside the subcohort, with the weight of 1 individual.

Lin and Ying (1993) include the CCH design as a special case of the CPH model where we have incomplete covariate measurements. They show that their approximation to the likelihood score function reduces to that of Self & Prentice for the CCH problem.

Aside from weighting, the methods differ in how they estimate the variance of $\hat{\boldsymbol{\beta}}$. Prentice proposed a complicated estimate to account for the correlation caused by cases outside the subcohort being in the likelihood at their own failure time, but not beforehand, and Self & Prentice feature a covariance matrix that is asymptotically equal to Prentice's. In the other camp, Barlow and Lin & Ying both use a robust jackknife estimator.

1.3 Genomic Data

Our goal is to discover genes associated with survival. We call these "differentially expressed genes" (DEGs). Genomic data is high-dimensional in the sense that the number of features is usually far greater than the number of observations. While not without its issues, analysts typically fit a CPH model for each gene, one at a time, while adjusting for confounding variables. For each gene, we record the estimated hazard ratio (HR) and its associated p-value. We adjust these p-values based on the Benjamini-Hochberg procedure (Benjamini & Hochberg, 1995), which is designed to control the false discovery rate (FDR). Genes with corresponding adjusted p-values less than a pre-specified threshold α are deemed to be significant.

Genomic data is typically contained in an *ExpressionSet*, which is a complicated data structure that consists of three major components:

- Gene expression levels—a matrix of gene expression assay values for each gene/individual combination.
- Sample annotations—a data frame of phenotypic metadata for the samples. This includes the survival times, censoring status, and covariate information such as age, sex, cohort membership, etc.

- Gene annotations—a list of feature metadata that includes gene names, symbols, chromosomal location, etc.

A visualization of the *ExpressionSet* is presented as Figure 1.2.

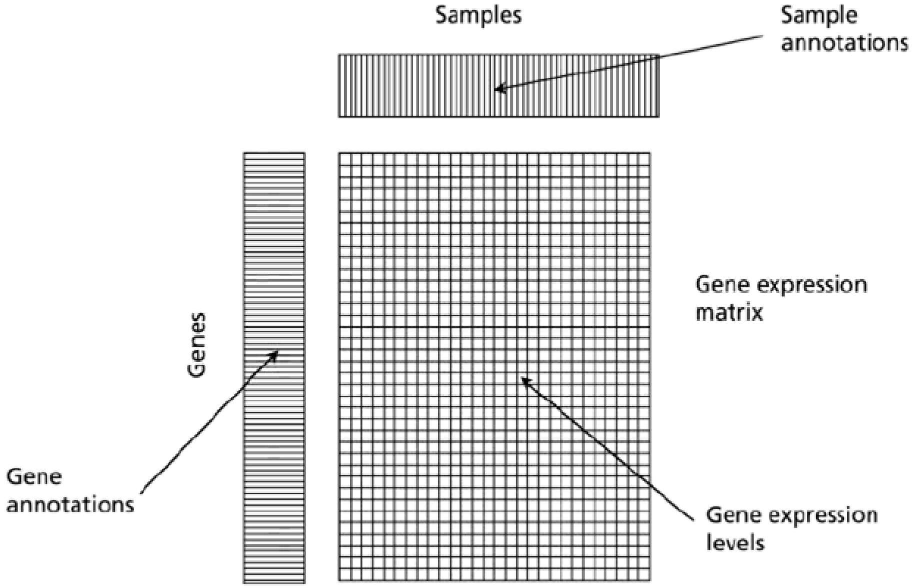


Figure 1.2: *ExpressionSet* data structure

No exploration of CCH method performance exists in the published literature for applications to high-dimensional data like that seen in gene expression studies. Such studies aim to discover "differentially expressed genes" (DEGs)—genes that are associated with survival outcome. We investigate CCH method performance when applied to real data and with a simulation study.

With real data, we never know the truth, but for existing cohort studies we can investigate how well a CCH analysis captures the results from a full cohort analysis. In simulation studies, we do know which genes are true DEGs, so we estimate the power of a method with the proportion of DEGs that are found to be significant, and estimate the FDR with the proportion of significant genes that are not DEGs.

Chapter 2

Results & Discussion

2.1 Childhood Leukemia Data

We first compare the four methods by analyzing a gene expression microarray dataset from a study of patients with high-risk pediatric B-cell precursor acute lymphoblastic leukemia (BCP-ALL) (Harvey et al., 2013; Kang et al., 2010). The dataset includes 207 children with BCP-ALL from the Childrens Oncology Group (COG) clinical study P9906 and 594 pediatric BCP-ALLs from the COG clinical study AALL0232. The original study data can be accessed at `https://clinicaltrials.gov`, using the identification codes NCT00005603 and NCT00075725. The combined dataset features 54,675 probe set expression levels of each patient's pretreatment leukemia cells, which were measured using the Affymetrix HG U133 Plus 2.0 platform. After removing the probe sets associated with sex-related genes, globins, and Affymetrix internal controls, we were left with 54,504 probe sets for the analysis. Without ambiguity, we refer to these probe sets as genes. The Robust Multi-array Average (RMA) algorithm was used to generate and normalize the gene expression levels. The data have been deposited in the National Center for Biotechnology Information

Gene Expression Omnibus (`http://www.ncbi.nlm.nih.gov/geo`) and are accessible through series accession numbers GSE68735 and GSE68790. The data can also be accessed at the NCI TARGET Initiative website (`https://ocg.cancer.gov/programs/target`).

One of the study goals is to identify genes whose expression levels are associated with event-free survival (EFS). "Event" can refer to relapse, death, or secondary malignancies. To identify the DEGs, we fit the CPH model to the EFS data with the expression level of a gene and a set indicator as the predictor variables. The set indicator identifies whether patients were from the P9906 or AALL0232 trials, and it is included in the model to adjust for the set effect between the two trials. Note that we fit as many CPH models as there are genes (in this case, 54,504), as each is assessing the association of only one gene with EFS.

Of the 801 individuals in the leukemia dataset, 213 had an event, giving an observed incident rate of around 26.6%. Table 2.1 gives a breakdown of the expected number of subjects that are cases or controls, by whether they are inside or outside the subcohort, in a CCH design. Each row gives the expected sample size breakdown for a given subcohort fraction π. Expected sample sizes are reported since the sampling variability inherent to the CCH design means that the number of cases/controls that are in/out of the subcohort is not guaranteed to be constant across samples.

Table 2.1: Expected Sample Size Breakdown for the Leukemia Data

Sampling fraction π	**Subcohort size**	**Cases in subcohort**	**Cases outside subcohort**	**Total cases**	**Controls in subcohort**	**Total CCH sample size**
0.1	80.1	21.3	191.7	213	58.8	271.8
0.2	160.2	42.6	170.4	213	117.6	330.6
0.3	240.3	63.9	149.1	213	176.4	389.4
0.4	320.4	85.2	127.8	213	235.2	448.2
0.5	400.5	106.5	106.5	213	294	507
0.6	480.6	127.8	85.2	213	352.8	565.8
0.7	560.7	149.1	63.9	213	411.6	624.6
0.8	640.8	170.4	42.6	213	470.4	683.4
0.9	720.9	191.7	21.3	213	529.2	742.2
1.0	801	213	0	213	588	801

The final row in Table 2.1 shows the characteristics for the full cohort, since $\pi = 1$. A key feature of the CCH design is that it includes all cases. Hence, the "total cases" column remains constant across π. Given the relatively high incidence rate for the leukemia data, CCH samples will tend to have a higher number of cases than controls when $\pi = 0.4$ or lower, so we can expect to see only modest gains in efficiency.

Figure 2.1 shows the "pseudo-FDR" achieved by each CCH method for the childhood leukemia dataset. Each boxplot represents the 5-number summary over 100 samples for various levels of π, with black dots indicating outlying samples. The smooth lines represent the mean pseudo-FDR for each method.

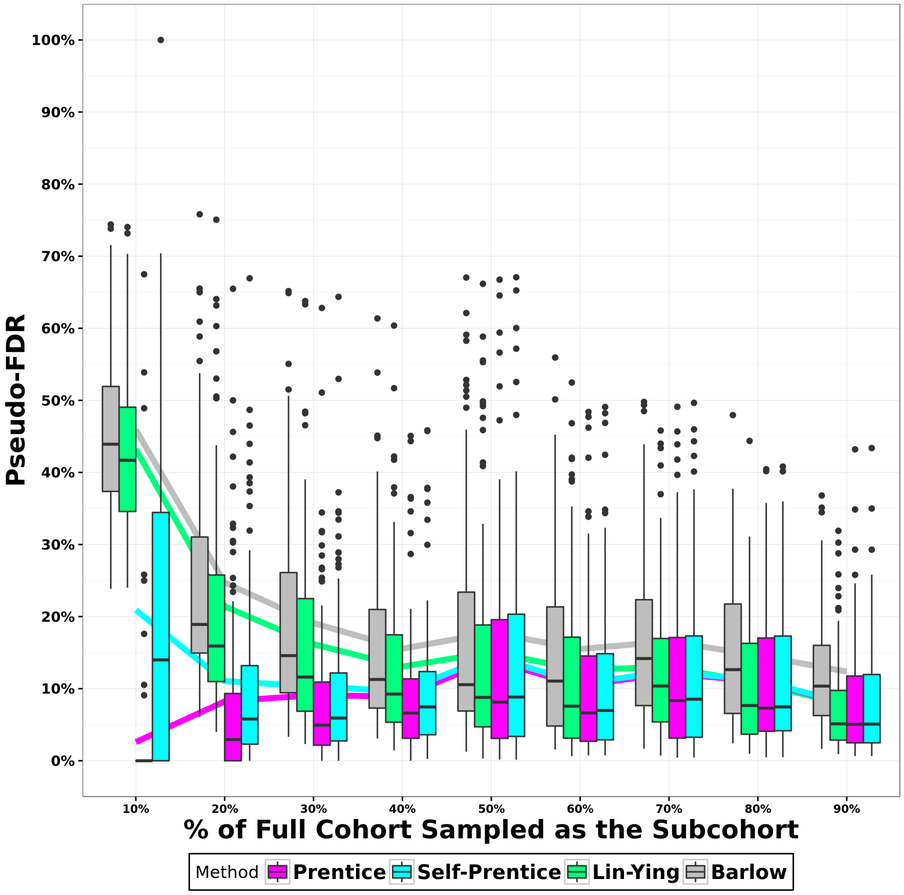


Figure 2.1: Pseudo-FDR for the Childhood Leukemia Data

Barlow and Lin-Ying are relatively more liberal than the other methods, particularly for small subcohort sizes. The distribution of pseudo-FDR for each method quickly approaches a similar shape and typical level of around 5%–15%, with Barlow's method consistently a bit more liberal than the others, followed by Lin-Ying, Self-Prentice, and Prentice's method, which was the most conservative.

Figure 2.2 displays the "pseudo-power" achieved by each CCH method for the childhood leukemia data. Each boxplot represents the 5-number summary from 100 CCH samples for various levels of π, with black dots indicating outlying samples. The smooth lines intersect the mean pseudo-power for each method.

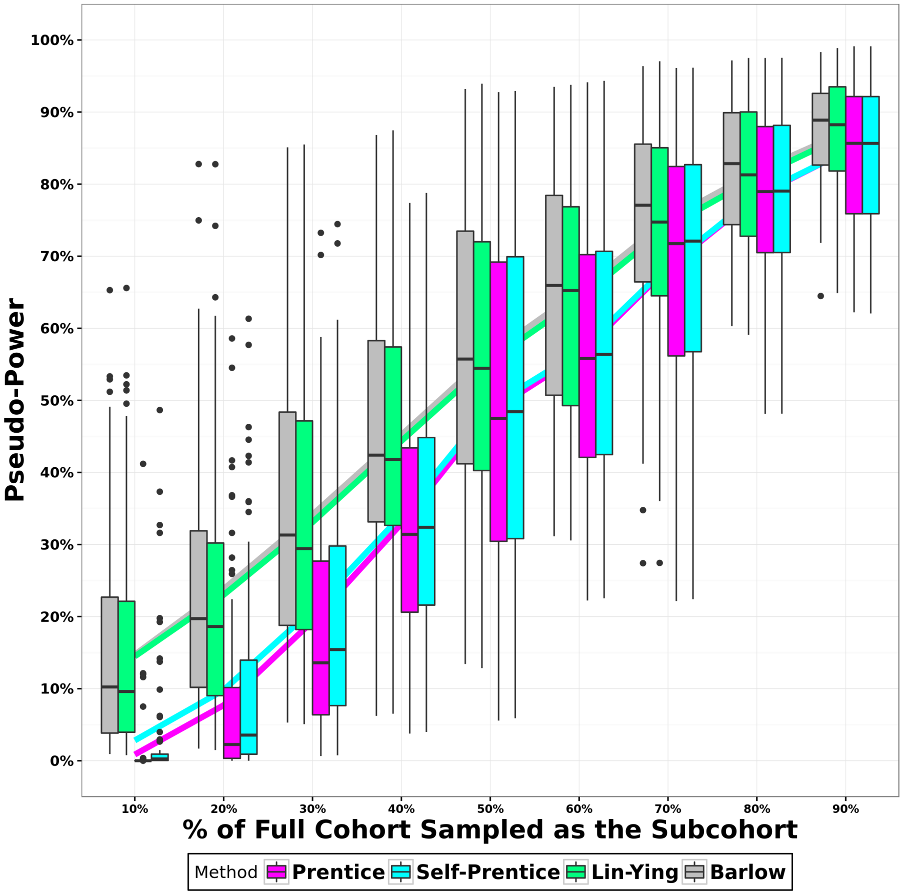


Figure 2.2: Pseudo-power for the Childhood Leukemia Data

As π increases, so too does the pseudo-power. Note that the CCH methods are equivalent to a full cohort analysis when $\pi = 100\%$. Barlow's method appears more "powerful" than the others as it captures a greater proportion of genes deemed significant by the full cohort analysis, followed by Lin-Ying and more distantly by the nearly-identical Self-Prentice and Prentice methods.

2.2 Breast Cancer Data

The breast cancer data combines the NKI (Van De Vijver et al., 2002) and TransBIG (Desmedt et al., 2007) data sets, which contain information related to the survival rates of 493 breast cancer patients. The gene expression data is from microarrays and has 10,566 features. The pseudo-FDR for the CCH methods is presented in Figure 2.3.

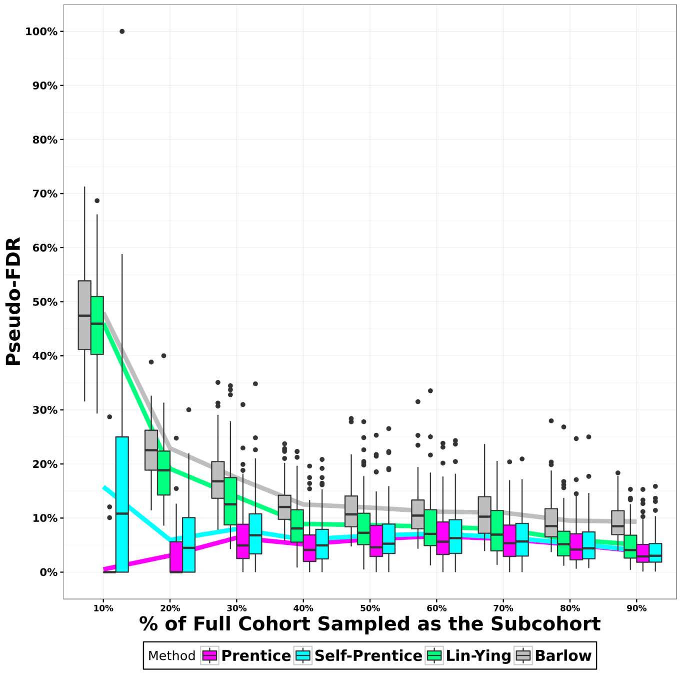


Figure 2.3: Pseudo-FDR for the Breast Cancer Data

The methods exhibit very similar behavior for the breast cancer, with pseudo-FDR quickly converging to a similar shape. Barlow and Lin-Ying are too liberal for the smallest subcohort size, but quickly approach similar pseudo-FDR levels as Prentice and Self-Prentice as π increases. Self-Prentice has a lot of variability at $\pi = 10\%$, but is comparable to Prentice's method for the other levels of π. Uniformly across π, from most conservative to most liberal we have Prentice, Self-Prentice, Lin-Ying, and Barlow.

Figure 2.4 shows the pseudo-power for the four CCH methods.

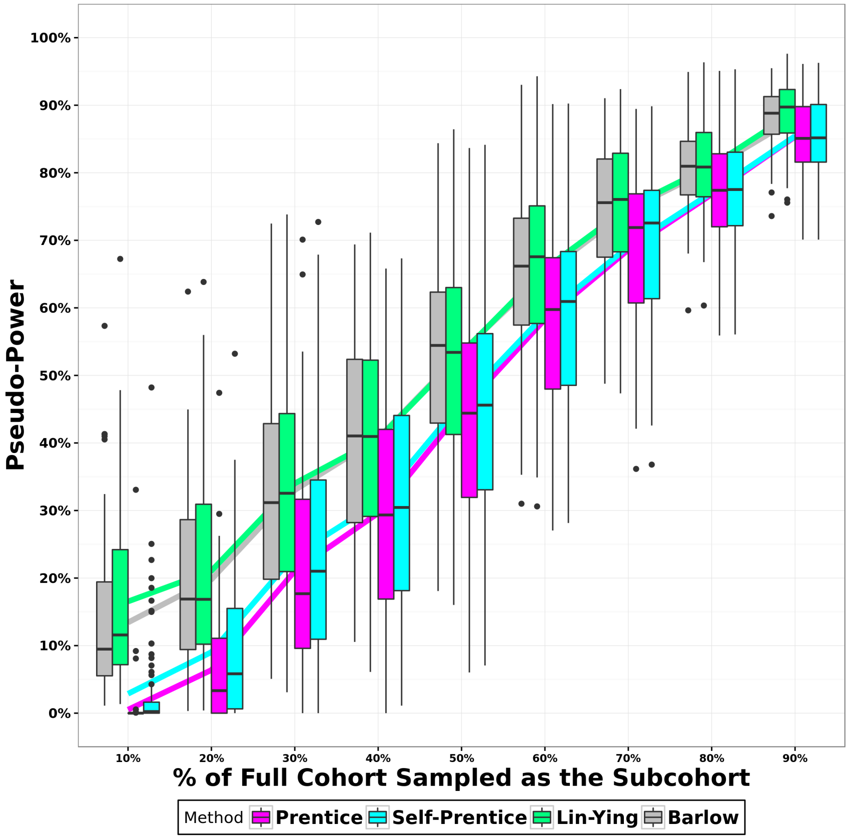


Figure 2.4: Pseudo-power for the Breast Cancer Data

Once again, we see a roughly linear increase in pseudo-power as π increases, regardless of method. The approaches of Barlow and Lin & Ying perform comparably and are uniformly more powerful across π than the approaches of Prentice and Self & Prentice, which also perform very similarly. The breast cancer data analysis suggests a tradeoff—the methods of Barlow and Lin & Ying appear to be more powerful at the cost of being too liberal for small samples, while the approaches of Prentice and Self & Prentice are better at limiting the pseudo-FDR at the cost of having a lower pseudo-power.

2.3 Simulation

Using simulated data to compare the effectiveness of the CCH approaches offers the advantage of knowing which genes are true DEGs. First, we generated gene expression data under various settings. Then, for each CCH method, we fit CPH models for each gene and extracted the estimated HRs and their associated p-values. P-values were adjusted using the BH procedure, and genes were called significant if their adjusted p-values were less than 0.05. We then estimated the FDR and power for each method. The estimated FDR was recorded as the proportion of genes deemed significant that were actually null genes, and the estimated power was recorded as the proportion of DEGs correctly identified as such. To account for sampling variability, the simulation was repeated 100 times at each setting. Since the empirical distributions of FDR and power appeared skewed, we used their median values for comparing the CCH methods.

Survival times were generated to give an incidence rate near 10%. For more details on how the gene expression data was simulated, please refer to the Methods section. Regarding sample size, we left the full cohort size fixed while adjusting the sampling fraction. Table 2.2 offers a breakdown of the expected sample sizes by sampling fraction.

Table 2.2: Expected Sample Size Breakdown for the Simulation Study

Sampling fraction π	**Subcohort size**	**Cases in subcohort**	**Cases outside subcohort**	**Total cases**	**Controls in subcohort**	**Total CCH sample size**
0.05	200	20	380	400	180	580
0.10	400	40	360	400	360	760
0.15	600	60	340	400	540	940

With a full cohort size of 4000, we expect around 400 total cases in a given sample. Our total CCH sample sizes tend to range from around 580 to 940, depending on the sampling fraction π.

Aside from investigating the effect of the sampling fraction, we examined how the methods behaved for different proportions of DEGs. The total number of genes was fixed at 1000, and we investigated the cases where 0%, 4%, 8%, 12%, 16%, 20%, 40%, 60%, 80%, and 100% of these genes were DEGs. We explored the 0%–20% window with finer resolution as it is more realistic range of DEGs.

Different HRs were used to generate the DEGs to see how the methods behave as effect size changes. HRs were chosen uniformly from 1.3–1.4, 1.4–1.5,1.5–1.6, and 1.6–1.7. This range was selected because in preliminary testing, DEGs with HRs less than 1.3 were undetectable, while those with HRs greater than 1.7 resulted near-perfect performance for each method. As a rule of thumb, small, medium, and large effect sizes are related to HRs around 1.3, 1.5, and 2.0, respectively.

2.4 FDR and Power

Using FDR and power to evaluate the methods, a method is considered superior if it has higher power, with one major caveat—FDR has to be near or below 5% in accordance with our nominal $\alpha = 0.05$ significance level.

Figure 2.5 shows the median FDR achieved by each method, broken down by the range of hazard ratios used to generate the DEGs, sampling fraction, and the percentage of genes that are DEGs.

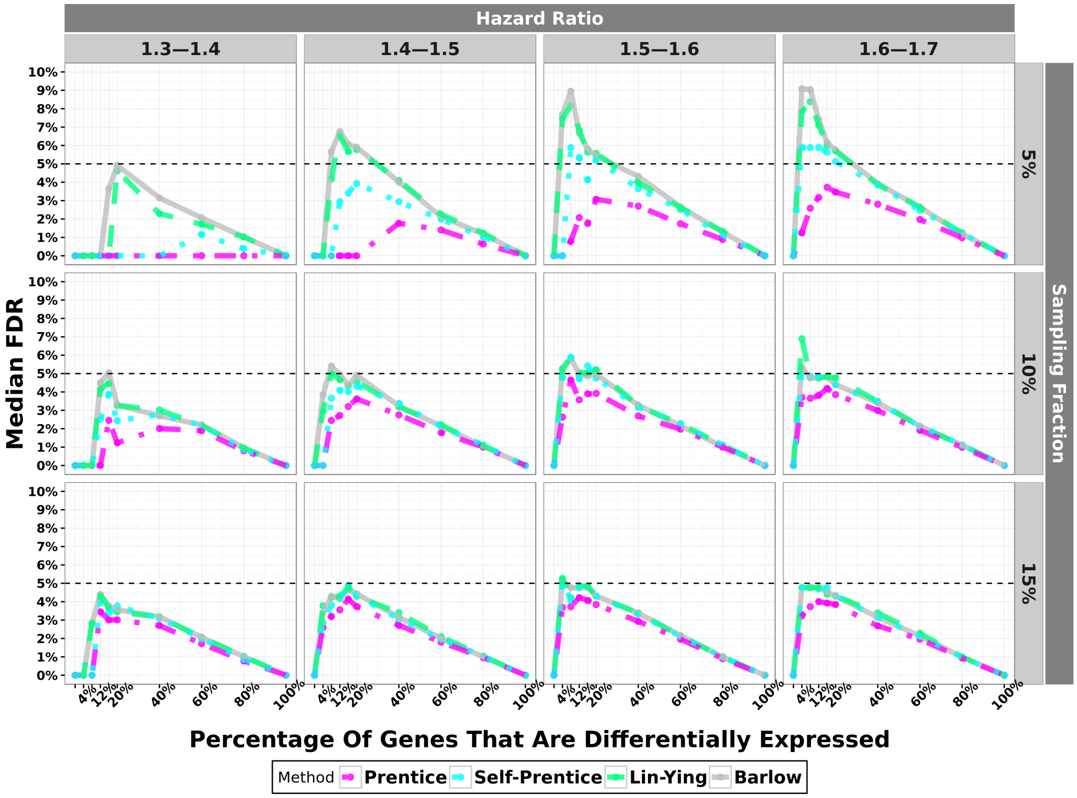


Figure 2.5: Median FDR for the Simulated Data

All methods control the FDR for sampling fractions of 10% and 15%, but the approaches of Barlow and Lin & Ying have trouble controlling FDR for the smaller samples associated with a sampling fraction of 5%.

Figure 2.6 displays the median power achieved by each method, broken down by the range of hazard ratios used to generate the DEGs, sampling fraction, and the percentage of genes that are DEGs.

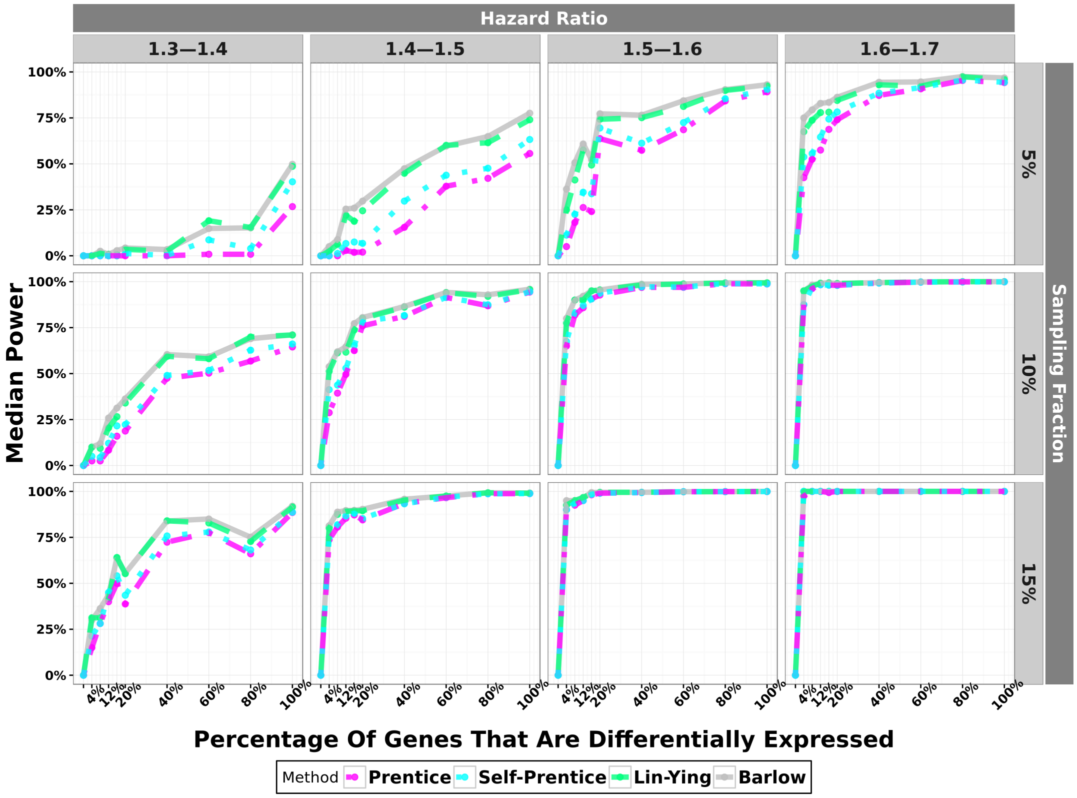


Figure 2.6: Median Power for the Simulated Data

As HR or sampling fraction increases, so too does median power. The methods of Barlow and Lin & Ying are uniformly more powerful than those of Prentice and Self & Prentice, but no method is sensitive enough to detect small HRs when the proportion of DEGs is in the realistic range. Conversely, all methods are effective for large samples or when HRs are at least medium.

2.5 Method Agreement

In addition to FDR and power, we consider method concordance. We use two approaches—a Venn diagram to display how many of the genes identified as significant are the same across methods, and a scatter plot matrix to compare how similarly the top genes are ranked by each method. Note that Figures 5 and 6 are based on the case where $\pi = 0.15$, the proportion of genes that were DEGs was 20%, and the DEGs were simulated based on HRs that were sampled from a $\mathcal{U}(1.5, 1.6)$.

Figure 2.7 depicts the Venn diagram approach, which shows the overlap between the top 200 most significant genes identified by each method, ordered from most to least significant based on adjusted p-value (lower p-value=more significant). Counts in overlapping regions indicate the number of shared genes.

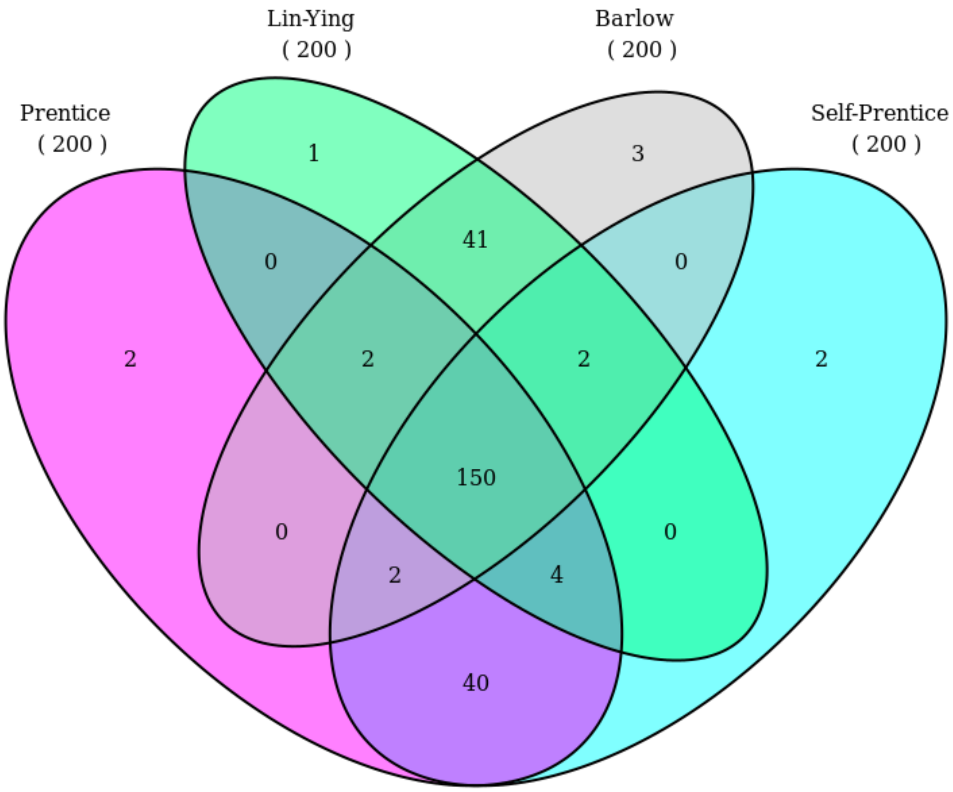


Figure 2.7: Venn Diagram for CCH Method Agreement

In this example, 150 of the top 200 genes are the same for each method. This large overlap implies high concordance among the four methods. The overlap between Prentice and Self-Prentice is 196, and the overlap between Lin-Ying and Barlow is 195, indicating almost perfect concordance between Prentice and Self-Prentice, and between Lin-Ying and Barlow.

Figure 2.8 shows the scatter plot matrix of gene ranking by method. The scatter plot matrix allows us to compare how similarly the methods rank the genes in terms of significance. The lower triangle of the matrix features the bivariate scatter plots, and the upper triangle displays the corresponding correlation coefficients (Pearson's r). A ranking is needed from each method for each gene involved in a given scatter plot, so we consider only the genes in the four-way intersection of the Venn diagram. Hence, each of the scatter plots in the lower triangle feature 150 points, although their rankings still range from 1 to 200. Note that a correlation coefficient of 1 would indicate perfect agreement between two methods, as would the points falling on a perfect line in the scatter plot.

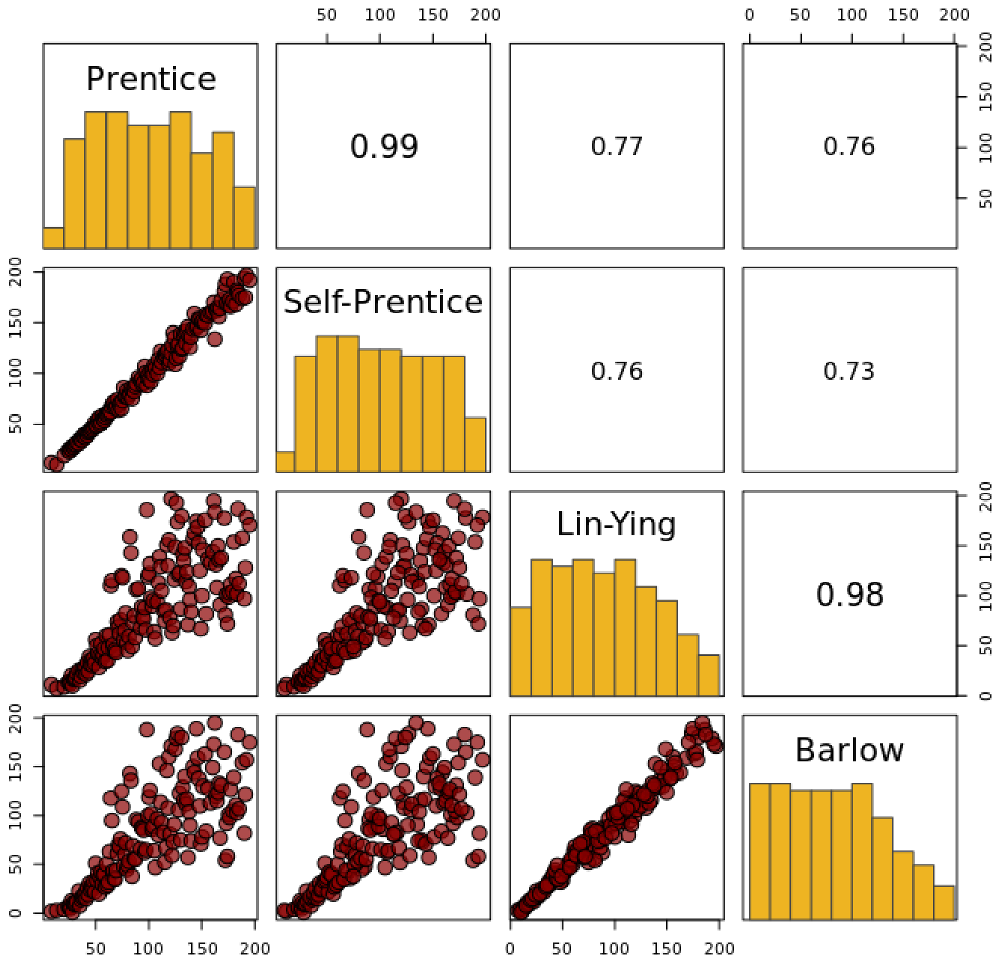


Figure 2.8: Scatter Plot Matrix of Gene Ranking.

A sampling fraction of $\pi = 100\%$ means the subcohort is equal to the full cohort, and we get absolute agreement between the CCH methods. As the sampling fraction decreases, there is gradually more and more disagreement between the methods, with counts moving from the center of the Venn diagram toward the periphery, and correlation decreasing in magnitude toward 0.

For real and simulated examples, we universally observed that the highest number of counts occurred in the four-way intersection, followed by the intersections between Prentice and Self-Prentice, and between Barlow and Lin-Ying. Similarly, in terms of gene rankings, there is a nearly perfect positive correlation between Prentice and Self-Prentice, as well as between Barlow and Lin-Ying.

Chapter 3

Methods

3.1 Assessing Method Performance for Real Data

For real data problems such as our leukemia dataset, we treat the full cohort analysis as "the truth" since we don't actually know which genes are truly DEGs. We consider the genes deemed significant by the full cohort analysis to be "DEGs" so we can compare the CCH methods to some baseline. We define pseudo-FDR and pseudo-power measurements to behave similarly to FDR and power for situations in which the truth is known.

If we let $\mathcal{F}$ and $\mathcal{C}$ be the sets of genes called significant by the full cohort and CCH analyses, respectively, and $\mathrm{N}(\cdot)$ be a function that returns the number of elements in a set, define pseudo-FDR as:

$$\text{pseudo-FDR} \equiv \frac{\mathrm{N}(\mathcal{C} \cap \mathcal{F}^{\mathsf{c}})}{\mathrm{N}(\mathcal{C})}$$

where $^{\mathsf{c}}$ indicates the complement of a set, and define pseudo-power as:

$$\text{pseudo-power} \equiv \frac{\mathrm{N}(\mathcal{C} \cap \mathcal{F})}{\mathrm{N}(\mathcal{F})}$$

Pseudo-FDR is the proportion of significant genes in a CCH analysis that are not also detected by the full cohort analysis, while pseudo-power is the proportion of significant genes from the full cohort analysis that are detected by a CCH analysis. For a CCH method to be considered effective in a real data example, it should have low pseudo-FDR and high pseudo-power.

3.2 Simulating Survival Times

Bender et al. Bender, Augustin, and Blettner (2005) used the inverse probability integral transform (PIT) to generate survival data under the CPH model, which has the survival function:

$$S(\boldsymbol{t}|\boldsymbol{x}) = exp[-H_0(\boldsymbol{t})e^{\boldsymbol{x}\boldsymbol{\beta}}]$$

where $H_0(\boldsymbol{t})$ is the cumulative hazard function. By the inverse PIT we have time $\boldsymbol{T}$ as a random variable:

$$\boldsymbol{T} = S^{-1}(\boldsymbol{U}|\boldsymbol{x}) = H_0^{-1}\left(-\frac{\log(\boldsymbol{U})}{e^{\boldsymbol{x}\boldsymbol{\beta}}}\right)$$

where $\boldsymbol{U}$ is a random variable following a continuous $\mathcal{U}(0,1)$ distribution. Draw a random set of $\boldsymbol{u}$s and plug them into the distribution of $\boldsymbol{T}$ to get a set of observed latent survival times $\boldsymbol{t} = S^{-1}(\boldsymbol{u}|\boldsymbol{x})$. Select $\boldsymbol{\beta}$ to be the desired log(HR) associated with the predictor $\boldsymbol{x}$. Draw censoring times $\boldsymbol{C} \sim Exp(\lambda_{\text{cens}})$ and compare them to their corresponding latent survival times. If censoring occurs before the latent survival time for an individual, they are right-censored, and hence in the control group. Otherwise, they were observed to have experienced an event during the time frame of the study and are considered cases.

Using a Weibull distribution with scale λ and shape ρ, survival times are generated using:

$$\boldsymbol{t} = \left(-\frac{\log(\boldsymbol{u})}{\lambda e^{\boldsymbol{x}\boldsymbol{\beta}}}\right)^{\frac{1}{\rho}}$$

This works for generating one gene expression vector, but to generate lots of genes related to a set of survival times, we have to go a bit further.

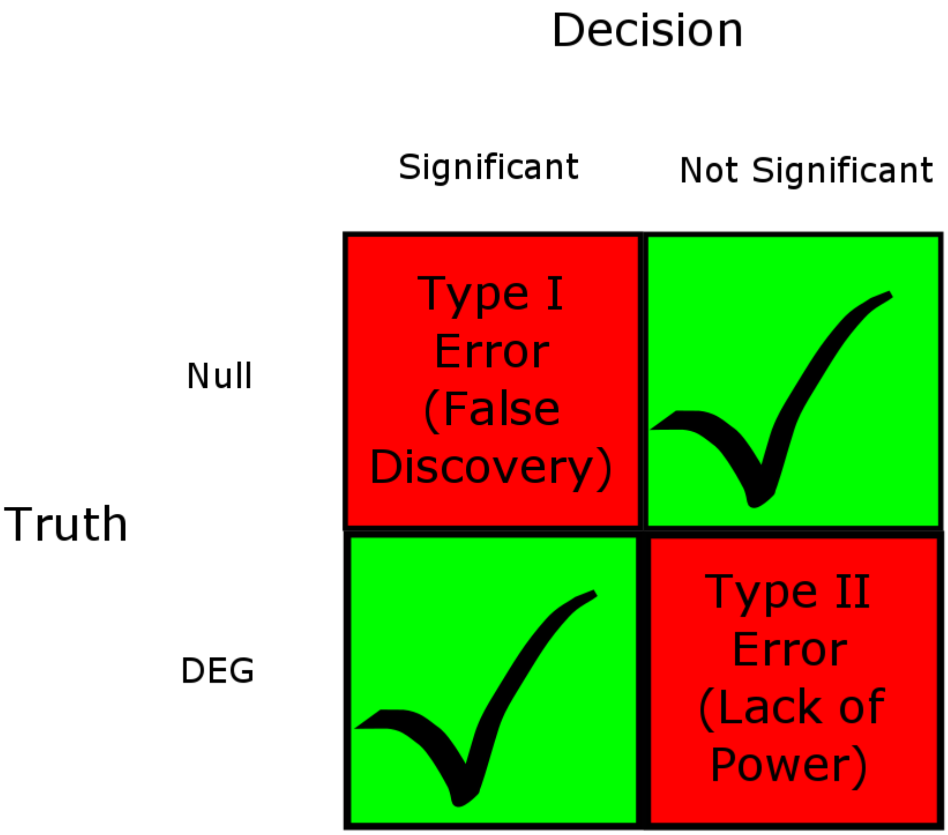
To generate many DEGs, we first generate survival times and censoring information using the method described above, with $\boldsymbol{x}^* \sim \mathrm{N}(0,1)$ as our "seed" gene used to generate the survival times (since standardization will allow us to use HR as a measure of relative effect size). For our simulations, the shape, scale, and censoring rate parameters were fixed at $\rho = 1$, $\lambda = 1$, and $\lambda_{\text{cens}} = 10$ to give an incidence rate near 10%. Rewriting the survival time equation in terms of x_{ij}, the expression level of gene i for individual j is:

$$x_{ij} = \frac{-log\left(\frac{-\lambda t_j^{\rho}}{log(u_j)}\right)}{\beta_i} + e_{ij}$$

where the e_{ij}'s are $\mathrm{N}(0,1)$ perturbations. We choose $\beta_i = \log(\mathrm{HR}_i)$ and draw a set of perturbations to generate expression levels for each DEG. For the null genes, we can draw random numbers since we only care that they are unrelated to survival.

3.3 Assessing Method Performance for Simulated Data

For each gene i, our significance test is $\mathrm{H}_0^{(i)} : \mathrm{HR} = 1$, as this indicates no relationship between gene i and survival time. We know the truth as to whether genes are DEGs or null genes, and we can either decide a gene is significant or not significant, so we are left with four possible scenarios, two of which are correct decisions, and two of which are incorrect. These are summarized in Figure 3.1.



Truth \ Decision	Significant	Not Significant
Null	Type I Error (False Discovery)	✓
DEG	✓	Type II Error (Lack of Power)

Figure 3.1: Summary of Possible Outcomes

We hope to identify all DEGs as significant and all null genes as not significant (the correct decisions). We also aim to not say that null genes are significant (type I error) or fail to call DEGs significant (type II error). Hence, to evaluate the performance of our four methods in focus, we will look at median FDR and power for each method under various settings.

To estimate FDR, we look at the proportion of rejected null hypotheses that were incorrectly rejected. If we have $i = 1, 2, \ldots, p$ total genes:

$$\widehat{\text{FDR}} = \frac{\text{false rejections}}{\text{total rejections}} = \frac{\sum_{i=1}^{p} \text{I}\left(\text{H}_0^{(i)}\text{rejected} \mid \text{H}_0^{(i)}\text{true}\right)}{max\left[\sum_{i=1}^{p} \text{I}\left(\text{H}_0^{(i)}\text{rejected}\right), 1\right]}$$

where I(·) is an indicator function, and we take the maximum between the number of rejections and 1 to avoid dividing by 0 in the denominator in the event that there are no rejections. The FDR approach stands as an alternative to approaches that aim to control the family-wise error rate (FWER). FDR preserves power while still accounting for the multiple tests being performed. This is really the only reasonable option for high-throughput data, as attempting to control the FWER usually leads to a cripplingly conservative cutoff for significance that completely eradicates any ability to detect DEGs.

To estimate power, we calculate the proportion of DEGs correctly identified as such:

$$\widehat{\text{Power}} = \frac{\text{correct rejections}}{\text{total DEGs}} = \frac{\sum_{i=1}^{p} \text{I}\left(\text{H}_0^{(i)}\text{rejected} \mid \text{H}_0^{(i)}\text{false}\right)}{\sum_{i=1}^{p} \text{I}\left(\text{H}_0^{(i)}\text{false}\right)}$$

where higher power indicates better performance. An important caveat is that this is only true when the type I error rate or FDR is being controlled—one could obviously use a test that calls every gene significant in order to achieve 100% power, but that test would be useless!

Chapter 4

Conclusions

The theory, real data example, and simulation study all lead to the same overwhelming conclusions. Prentice and Self & Prentice, with their similar weighting schemes and asymptotically-identical covariance matrices for $\hat{\beta}$, exhibit a great deal of agreement when evaluating biomarkers in a high-throughput setting. Similarly, the approaches of Barlow and Lin & Ying have strong agreement, due in large part to their shared use of robust variance estimation.

If we base our decision exclusively on performance metrics like FDR and power, we are left with essentially two options. The approaches of Prentice and Self & Prentice staunchly control the FDR, but are less sensitive in their ability to detect DEGs than the approaches of Barlow and Lin & Ying, which may have issues controlling FDR in some cases. For most practical cases, all methods will give similar results, but it is important to also consider philosophical soundness. In this regard, Barlow's method has the clear advantage as it has the most intuitive weighting scheme.

Chapter 5

Future Work

Although each of the analyses appear to lead to the same inevitable conclusions, there is possible room to improve our investigation. Perhaps most notable is the fact that genes were generated independently in our simulation study. In real data examples, we typically see distinct clustering patterns, and therein lies a clear avenue for more nuanced and realistic simulations. However, the testing procedure treats genes as if they are independent, so this point may be inconsequential.

High-throughput data is fraught with many unexplored complications. The popular approach of fitting many CPH models will understandably raise some eyebrows in regard to the inability to effectively evaluate the underlying model assumptions, in particular, the proportional hazards assumption, for all of the models being fit. Hence, it worth investigating a permutation test like the significance analysis of microarrays (SAM) (Tusher, Tibshirani, & Chu, 2001). Given the case-cohort sampling design, such a test is not a straightforward application. My proposed avenue of analysis is to essentially use a technique like bootstrapping to augment the CCH sample with additional controls outside the subcohort to effectually "fake" a full cohort, and then proceed to use the permutation test.

Chapter 6

Acknowledgments

The work in this part of the dissertation was from my research at the UNM Cancer Research Facility. I (John Pesko) contributed to the study design, analysis, interpretation and presentation of the results, and the writing of the manuscript. All work was done under the supervision of Huining Kang, who devised the research concept, and our work was funded by Gabrielle's Angel Foundation.

References

tchison1975goodness Aitchison, J. (1975). Goodness of prediction fit. *Biometrika*, *62*(3), 547–554.

2004heteroscedastic Akritas, M. G., & Papadatos, N. (2004). Heteroscedastic one-way ANOVA and lack-of-fit tests. *Journal of the American Statistical Association*, *99*(466), 368–382.

andrew1993density Andrew, N., & Underwood, A. (1993). Density-dependent foraging in the sea urchin *centrostephanus rodgersii* on shallow subtidal reefs in New South Wales, Australia. *Marine Ecology Progress Series*, *99*, 89–98.

barlow1999analysis Barlow, W. E., Ichikawa, L., Rosner, D., & Izumi, S. (1999). Analysis of case-cohort designs. *Journal of Clinical Epidemiology*, *52*(12), 1165–1172.

ayarri2004interplay Bayarri, M. J., & Berger, J. O. (2004). The interplay of Bayesian and frequentist analysis. *Statistical Science*, *19*(1), 58–80.

mr1763essay Bayes, M., & Price, M. (1763). An essay towards solving a problem in the doctrine of chances. by the late Rev. Mr. Bayes, FRS communicated by Mr. Price, in a letter to John Canton, AMFRS. *Philosophical Transactions (1683–1775)*, *53*, 370–418.

ender2005generating Bender, R., Augustin, T., & Blettner, M. (2005). Generating survival times to simulate Cox proportional hazards models. *Statistics in Medicine*, *24*(11), 1713–1723.

mini1995controlling Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, *57*(1), 289–300.

bennett1990clinical Bennett, D. A., Wilson, R. S., Gilley, D. W., & Fox, J. H. (1990). Clinical diagnosis of Binswanger's disease. *Journal of Neurology, Neurosurgery & Psychiatry*, *53*(11), 961–965.

box1964analysis Box, G. E., & Cox, D. R. (n.d.). An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, *26*.

breiman2001random Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

eiman2001statistical Breiman, L., et al. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, *16*(3), 199–231.

andelario2011matrix Candelario-Jalil, E., Thompson, J., Taheri, S., Grossetete, M., Adair, J. C., Edmonds, E., … Rosenberg, G. A. (2011). Matrix metalloproteinases are associated with increased blood-brain barrier opening in vascular cognitive impairment. *Stroke*, *42*(5), 1345–1350.

aplan1995binswanger Caplan, L. R. (1995). Binswanger's disease–revisited. *Neurology*, *45*(4), 626–633.

istensen2005testing Christensen, R. (2005). Testing Fisher, Neyman, Pearson, and Bayes. *The American Statistician*, *59*(2), 121–126.

christensen2006log Christensen, R. (2006). *Log-linear models and logistic regression* (1st ed.). Springer Science & Business Media, New York, NY.

hristensen2011plane Christensen, R. (2011). *Plane answers to complex questions: the theory of linear models* (4th ed.). Springer Science & Business Media, New York, NY.

cox1992regression Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, *34*(2), 187–220.

cox1975partial Cox, D. R. (1975). Partial likelihood. *Biometrika*, *62*(2), 269–276.

rshall2016detection DeMarshall, C. A., Nagele, E. P., Sarkar, A., Acharya, N. K., Godsey, G., Goldwaser, E. L., … others (2016). Detection of Alzheimer's disease at mild cognitive impairment and disease progression using autoantibodies as blood-based biomarkers. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, *3*, 51–62.

desmedt2007strong Desmedt, C., Piette, F., Loi, S., Wang, Y., Lallemand, F., Haibe-Kains, B., … others

(2007). Strong time dependence of the 76-gene prognostic signature for node-negative breast cancer patients in the TRANSBIG multicenter independent validation series. *Clinical Cancer Research*, *13*(11), 3207–3214.

efron1998ra Efron, B. (1998). R.A. Fisher in the 21st century. *Statistical Science*, *13*(2), 95–114.

efron2012bayesian Efron, B. (2012). Bayesian inference and the parametric bootstrap. *The Annals of Applied Statistics*, *6*(4), 1971.

efron2013250 Efron, B. (2013). A 250-year argument: belief, behavior, and the bootstrap. *Bulletin of the American Mathematical Society*, *50*(1), 129–146.

njuntti2004emerging Erkinjuntti, T., Roman, G., Gauthier, S., Feldman, H., & Rockwood, K. (2004). Emerging therapies for vascular dementia and vascular cognitive impairment. *Stroke*, *35*(4), 1010–1017.

isher1941asymptotic Fisher, R. A. (1941). The asymptotic approach to Behrens's integral, with further tables for the d test of significance. *Annals of Eugenics*, *11*(1), 141–172.

gasparovic20131h Gasparovic, C., Prestopnik, J., Thompson, J., Taheri, S., Huisa, B., Schrader, R., ... Rosenberg, G. A. (2013). 1h-mr spectroscopy metabolite levels correlate with executive function in vascular cognitive impairment. *Journal of Neurology, Neurosurgery & Psychiatry*, *84*(7), 715–721.

orelick2011vascular Gorelick, P. B., Scuteri, A., Black, S. E., DeCarli, C., Greenberg, S. M., Iadecola, C., ... others (2011). Vascular contributions to cognitive impairment and dementia. *Stroke*, *42*(9), 2672–2713.

gray2013random Gray, K. R., Aljabar, P., Heckemann, R. A., Hammers, A., Rueckert, D., Initiative, A. D. N., et al. (2013). Random forest-based similarity measures for multi-modal classification of Alzheimer's disease. *NeuroImage*, *65*, 167–175.

chinski2006national Hachinski, V., Iadecola, C., Petersen, R. C., Breteler, M. M., Nyenhuis, D. L., Black, S. E., ... others (2006). National Institute of Neurological Disorders and Stroke–Canadian stroke network vascular cognitive impairment harmonization standards. *Stroke*, *37*(9), 2220–2241.

rvey2013development Harvey, R. C., Kang, H., Roberts, K. G., Chen, I.-M. L., Atlas, S. R., Bedrick, E. J.,

. . . others (2013). Development and validation of a highly sensitive and specific gene expression classifier to prospectively screen and identify B-precursor acute lymphoblastic leukemia (ALL) patients with a Philadelphia chromosome-like (Ph-like or BCR-ABL1-Like) signature for therapeutic targeting and clinical intervention. *Blood*, *122*(21), 826–826.

hoeffding1948class Hoeffding, W. (1948). A class of statistics with asymptotically normal distribution. *The Annals of Mathematical Statistics*, *19*(3), 293–325.

hoeffding1961strong Hoeffding, W. (1961). The strong law of large numbers for u-statistics. *Institute of Statistics Mimeo Series*, *302*.

effreys1946invariant Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, *186*(1007), 453–461.

kang2010gene Kang, H., Chen, I.-M., Wilson, C. S., Bedrick, E. J., Harvey, R. C., Atlas, S. R., . . . others (2010). Gene expression classifiers for relapse-free survival and minimal residual disease improve risk classification and outcome prediction in pediatric B-precursor acute lymphoblastic leukemia. *Blood*, *115*(7), 1394–1405.

kass1996selection Kass, R. E., & Wasserman, L. (1996). The selection of prior distributions by formal rules. *Journal of the American Statistical Association*, *91*(435), 1343–1370.

orthy2007parametric Krishnamoorthy, K., Lu, F., & Mathew, T. (2007). A parametric bootstrap approach for anova with unequal variances: Fixed and random models. *Computational Statistics & Data Analysis*, *51*(12), 5731–5742.

lagakos1981case Lagakos, S., & Mosteller, F. (1981). A case study of statistics in the regulatory process: the FD&C Red No. 40 experiment. *Journal of the National Cancer Institute*, *66*(1), 197–212.

lebedev2014random Lebedev, A., Westman, E., Van Westen, G., Kramberger, M., Lundervold, A., Aarsland, D., . . . others (2014). Random forest ensembles for detection and prediction of Alzheimer's disease with a good between-cohort robustness. *NeuroImage: Clinical*, *6*, 115–125.

randomforest2002 Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R News*, *2*(3), 18-22. Retrieved from `http://CRAN.R-project.org/doc/Rnews/`

lin1993cox Lin, D., & Ying, Z. (1993). Cox regression with incomplete covariate measurements. *Journal of the American Statistical Association*, *88*(424), 1341–1349.

dley1957statistical Lindley, D. V. (1957). A statistical paradox. *Biometrika*, *44*(1/2), 187–192.

uzzi2002intrathecal Liuzzi, G., Trojano, M., Fanelli, M., Avolio, C., Fasano, A., Livrea, P., & Riccio, P. (2002). Intrathecal synthesis of matrix metalloproteinase-9 in patients with multiple sclerosis: implication for pathogenesis. *Multiple Sclerosis Journal*, *8*(3), 222–228.

marquis1820theorie marquis de Laplace, P. S. (1820). *Théorie analytique des probabilités.* V. Courcier, Paris, France.

iller1989binswanger Miller Fisher, C. (1989). Binswanger's encephalopathy: a review. *Journal of Neurology*, *236*(2), 65–79.

wski1962subcortical Olszewski, J. (1962). Subcortical arteriosclerotic encephalopathy. Review of the literature on the so-called Binswanger's disease and presentation of two cases. *World Neurology*, *3*, 359.

pantoni2010cerebral Pantoni, L. (2010). Cerebral small vessel disease: from pathogenesis and clinical characteristics to therapeutic challenges. *The Lancet Neurology*, *9*(7), 689–701.

prentice1986case Prentice, R. L. (1986). A case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika*, *73*(1), 1–11.

Rcitation R Core Team. (2016). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `https://www.R-project.org/`

oman2002subcortical Román, G. C., Erkinjuntti, T., Wallin, A., Pantoni, L., & Chui, H. C. (2002). Subcortical ischaemic vascular dementia. *The Lancet Neurology*, *1*(7), 426–436.

roman2010randomized Román, G. C., Salloway, S., Black, S. E., Royall, D. R., DeCarli, C., Weiner, M. W., . . . Posner, H. (2010). Randomized, placebo-controlled, clinical trial of donepezil in vascular dementia. *Stroke*, *41*(6), 1213–1221.

berg1979subcortical Rosenberg, G. A., Kornfeld, M., Stovring, J., & Bicknell, J. M. (1979). Subcortical arteriosclerotic encephalopathy (binswanger) computerized tomography. *Neurology*, *29*(8), 1102–1102.

nberg2015validation Rosenberg, G. A., Prestopnik, J., Adair, J. C., Huisa, B. N., Knoefel, J., Caprihan, A., . . . Schrader, R. (2015). Validation of biomarkers in subcortical ischaemic vascular disease of the Binswanger type: approach to targeted treatment trials. *Journal of Neurology, Neurosurgery & Psychiatry*, *86*(12), 1324–1330.

enberg2016consensus Rosenberg, G. A., Wallin, A., Wardlaw, J. M., Markus, H. S., Montaner, J., Wolfson, L., . . . others (2016). Consensus statement for diagnosis of subcortical small vessel disease. *Journal of Cerebral Blood Flow & Metabolism*, *36*(1), 6–25.

aite1946approximate Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin*, *2*(6), 110–114.

vage1961foundations Savage, L. J. (1961). *The foundations of statistics reconsidered.* University of Calif Press, Berkeley, CA.

hwarz1978estimating Schwarz, G., et al. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464.

self1988asymptotic Self, S. G., & Prentice, R. L. (1988). Asymptotic distribution theory and efficiency results for case-cohort studies. *The Annals of Statistics*, *16*(1), 64–81.

snyder2015vascular Snyder, H. M., Corriveau, R. A., Craft, S., Faber, J. E., Greenberg, S. M., Knopman, D., . . . others (2015). Vascular contributions to cognitive impairment and dementia including Alzheimer's disease. *Alzheimer's & Dementia*, *11*(6), 710–717.

student1908probable Student. (1908). The probable error of a mean. *Biometrika*, *6*(1), 1–25.

taheri2011blood Taheri, S., Gasparovic, C., Huisa, B. N., Adair, J. C., Edmonds, E., Prestopnik, J., . . . others (2011). Blood-brain barrier permeability abnormalities in vascular cognitive impairment. *Stroke*, *42*(8), 2158–2163.

tsui1989generalized Tsui, K.-W., & Weerahandi, S. (1989). Generalized p-values in significance testing of hypotheses in the presence of nuisance parameters. *Journal of the American Statistical Association*, *84*(406), 602–607.

her2001significance Tusher, V. G., Tibshirani, R., & Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences*, *98*(9), 5116–5121.

van2002gene Van De Vijver, M. J., He, Y. D., van't Veer, L. J., Dai, H., Hart, A. A., Voskuil, D. W., … others (2002). A gene-expression signature as a predictor of survival in breast cancer. *New England Journal of Medicine*, *347*(25), 1999–2009.

andi1995generalized Weerahandi, S. (1995). Generalized confidence intervals. In *Exact statistical methods for data analysis* (pp. 143–168). Springer Science & Business Media, New York, NY.

lch1938significance Welch, B. L. (1938). The significance of the difference between two means when the population variances are unequal. *Biometrika*, *29*(3/4), 350–362.

ggplot22009 Wickham, H. (2009). *ggplot2: Elegant graphics for data analysis.* Springer-Verlag New York. Retrieved from `http://ggplot2.org`

zhang2015parametric Zhang, G. (2015a). A parametric bootstrap approach for one-way ANOVA under unequal variances with unbalanced data. *Communications in Statistics-Simulation and Computation*, *44*(4), 827–832.

ang2015simultaneous Zhang, G. (2015b). Simultaneous confidence intervals for pairwise multiple comparisons in a two-way unbalanced design with unequal variances. *Journal of Statistical Computation and Simulation*, *85*(13), 2727-2735.