EFRON'S METHOD ON LARGE SCALE CORRELATED DATA AND ITS REFINEMENTS

Asmita Ghoshal

A Dissertation

Submitted to the Graduate College of Bowling Green
State University in partial fulfillment of
the requirements for the degree of

DOCTOR OF PHILOSOPHY

August 2023

Committee:

John Chen, Committee Chair

Alexis Ostrowski,
Graduate Faculty Representative

Umar Islambekov

Riddhi Ghosh

ABSTRACT

John Chen, Committee Chair

This dissertation focuses on methodological innovation for multiple testing on hypotheses related to large-scale and correlated data, where error rate control is intrinsically critical. Research toward this goal necessitates rigorous discussions on a thorny concept, the strong control of familywise error rate (FWER). In the literature, published papers in this regard subsequently avoid this intricate issue by adapting feeble criteria such as the weak control of FWER or the false discovery rate. Different from conventional approaches, we directly tackle the problem with the strong control of FWER.

Starting with Efron's data on an inference problem related to 7128 genes of 72 patients, consisting of 47 acute lymphoblastic leukemia patients and 25 acute myeloid leukemia patients, the dissertation lays out fundamental terminologies facilitating the research on multiple inferences after discussing a method controlling the false discovery rate following the empirical approach of estimating the correlation parameter.

Following a review of the current literature, one distinct feature of the dissertation attributes to multiple testing procedures on odds ratios when several populations are of interest. When the joint distribution of a cluster of subsequent populations is approximately available, such as the utilization of the Cochran-Mantel-Haenszel statistic, a sequential testing method of strong control of FWER is proposed. The new method outperforms the traditional Holm's procedure (which also strongly controls FWER) in terms of substantiating any significant discovery that is detected by the latter.

Another feature of the dissertation explores the sequential testing procedure for the comparison of the odds ratio. It effectuates a general stepwise exact inference procedure that strongly controls the FWER. The new procedure is robust and versatile for both parametric and nonparametric settings. When the new procedure was employed with the Jonckheere-Terpstra test, it distinctly improved power performance, as shown in a simulation. The new procedure was

applied to analyze a real-life dataset from CDC regarding the age effect on binge alcoholism. It reveals the fact that the rate of binge alcoholism steadily increases in the age group of 18-34.

Finally, the dissertation shifts attention to the analysis of large-scale correlated data posted in Efron's paper. It attributes more intrinsic inference outcomes to the new procedure proposed in this dissertation research. Specifically, the new method was combined with a normality bootstrapping method. The outcome greatly enhances preceding analytic results on the gene expression data. An implementation adapting a nonparametric bootstrapping method on the data casts a new highlight on the robustness of the new procedure.

ACKNOWLEDGMENTS

I want to express my gratitude to my advisor, Dr. John Chen, for his unwavering support and guidance throughout the process of completing my dissertation. Additionally, I extend my heartfelt thanks to Dr. Alexis Ostrowski, Dr. Umar Islambekov, and Dr. Riddhi Ghosh for their valuable contributions as members of my dissertation committee.

TABLE OF CONTENTS

Page

LIST OF FIGURES

LIST OF TABLES

PREFACE

It is a common practice in industry and academia to employ multiple hypotheses in scientific research. The existing step-wise procedures for conducting simultaneous hypotheses testing are deemed inadequate for large-scale data analysis when considering strong control over the family-wise error rate ($FWER$). Holm's step-down method strongly controls the family-wise error rate at a given significance level of α. However, the procedure ends up being too punitive for large-scale simultaneous hypothesis testing as it uses the Bonferroni correction in a sequential manner . On the other hand, despite a statistically more powerful approach than Holm's step-down procedure, Hochberg's step-up procedure does not strongly control the family-wise error rate. Bradley Efron uses an estimated false discovery rate (FDR) based on an underlying distributional assumption in Efron (2010). Albeit, FDR can weakly control the $FWER$ when all null hypotheses under consideration are true. This dissertation aims to develop a robust rejective algorithm of multiple hypothesis testing to ensure strong control over the $FWER$. First, I focused on improving the existing hypothesis-testing procedures involving multiple odds ratios. Following my research, I expanded the recently suggested methodology in a non-parametric framework. In addition, I compared the statistical power of the new algorithm and that of Holm's step-down procedure under various setups. The real-life application of the newly proposed algorithm has demonstrated its effectiveness. Finally, I extended the new methodology for testing simultaneous hypotheses related to large-scale data analysis that involves significant correlation. A bootstrapped 95% confidence interval was constructed when the underlying distribution of the test statistics was unknown. This dissertation has successfully proposed a confidence procedure that meets the requirements of a robust simultaneous hypothesis technique in scientific research. It guarantees strong control over the family-wise error rate, making it highly useful in today's world of large-scale data analysis.

Asmita Ghoshal

Bowling Green, Ohio

June 22, 2023

CHAPTER 1 INTRODUCTION AND LITERATURE REVIEW

1.1 Introduction and Background

Researchers often encounter correlated data across various fields in today's scientific landscape. Conducting multiple hypothesis testing accurately presents a significant challenge. To mitigate the problem of high correlation, various methods can be used. False Discovery Rate (FDR), which is the expected value of the ratio of the number of false rejections to the total number of rejections, and the Family Wise Error Rate($FWER$), which is the probability of making at least one type I error, are two such methods that are in use. When conducting statistical analysis, it is more effective to prioritize the control of the family-wise error rate ($FWER$) over the false discovery rate (FDR) in order to minimize false positives. This is because FDR can only offer weak control over $FWER$ when all the null hypotheses under consideration are true. Holm (1979) proposed a step-wise procedure for multiple testing that can control $FWER$. However, when the number of hypotheses under consideration is large enough, Holm's procedure ends up with overly strict rejection criteria, thus can lead to an erroneous conclusion. This research aims to develop a robust methodology for controlling the family-wise error rate in the presence of significant correlations. To achieve the objective of the proposed research topic, I am beginning by examining previous studies on large-scale data analyses as well as studies on simultaneous hypothesis testing.

In large-scale hypothesis testing, Efron (2010) proposes a reliable method for calculating the empirical false discovery rate even when significant correlations are present. A confidence procedure proposed by Chen (2016) can conduct simultaneous hypothesis testing while controlling the family-wise error rate in a strong sense. In the literature review, I present a comprehensive understanding of the topics relevant to my research study in the two papers mentioned above. In addition, I have explored various definitions from Casella and Berger (2021) relevant to simultaneous hypothesis testing and confidence procedures. To start, I examine the methodology outlined in the research conducted by Efron in 2010, as referenced in the bibliography. Following this, I will provide my interpretation of the confidence procedure

discussed in Chen's non-parametric study from 2016, including any pertinent definitions.

Large-scale studies frequently involve a significant number of correlated instances. So, a practitioner must consider the correlation between different cases accurately. Without a sound methodology, a practitioner might end up underestimating the variability in the data raising severe consequences. The data set under consideration concerns a leukemia microarray study by Golub, Slonim, Tamayo, Huard, Gaasenbeek, Mesirov, Coller, Loh, Downing, Caligiuri, et al. (1999) that has been used in Efron (2010) paper for motivation and illustration. I have used the same data set to present a comparative study between Holm's Algorithm and the proposed algorithm in Chapter 4. Corresponding to each of the 7128 genes, two disease categories are being studied, acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML), with a total of 72 patients. There are 47 patients under ALL and 25 patients under AML. We aim to investigate any significant difference in gene expressions between ALL and AML for each of the $N(= 7128)$ genes simultaneously.

Problem Statement:

The objective of this dissertation is to test multiple null hypotheses, labeled as $H_{01}, ...H_{0N}$, while accounting for significant pairwise correlations in a reliable manner. The primary goal is to create a simultaneous confidence procedure that can strongly control the family-wise error rate (as defined in 2.2). The focus is on achieving exactness in the process by strongly controlling the family-wise error rate at a given significance level.

1.2 Methodology for Empirical FDR (Efron (2010))

In Efron (2010), the author has used a microarray experiment involving thousands of genes to demonstrate the idea of capturing the correlation parameter while obtaining the cumulative distribution of the summary statistics. In the microarray experiment, numerous genes were scrutinized to identify the presence of a potential disease. Corresponding to each gene, a z value is produced. Essentially, these z values represent the correlated test statistics. An important assumption is that the z-values follow normal distributions with different means and variances. In

the microarray experiment under consideration, N=7128 genes are being examined to identify the genetic difference between two forms of leukemia, namely acute lymphoblastic leukemia (ALL) and acute myeloid leukemia(AML).

The number of ALL patients and AML patients involved in the study are 47 and 25, respectively. The expression levels on 7128 genes are measured for each of these patients. The aim is to accurately calibrate the correlation structure between these 7128 gene expression levels. The main idea is to establish the sufficiency of estimating the $\frac{N(N-1)}{2}$ pairwise correlations for capturing the correlation structure of the leukemia data robustly.

The methodology in Efron (2010) computes two-sample t-statistic comparing the AML and ALL expression levels for each of the 7128 genes. After that, these t-statistic values are converted to z-values using the inverse of the cumulative distribution function for standard normal. These z-values approximate an empirical right-sided cumulative distribution function, namely $\widehat{F}$ defined next.

$$\widehat{F}(x) = \frac{\#\{Z_i > x\}}{N} \tag{1.2.1}$$

Where, z_i's denotes the z-value corresponding to the i^{th} gene.

The primary concern of Efron (2010) is accurately obtaining summary statistics like empirical cumulative distribution function. Finally, the Efron (2010) deduces the properties of those above empirical cumulative distribution functions in the presence of substantial correlation using the root-mean-square of the pairwise correlations. A simple formula is derived for calculating the pairwise correlations. Usually, the presence of correlation decreases the accuracy of statistical models. So, practitioners must understand and analyze the correlation's impact on summary statistics estimates. Additionally, it is pivotal to study the consequences of correlated variables on various statistical tests like hypothesis testing. Efron (2010) addresses all the above-noted aspects in a concise manner.

The root means square correlation parameter is estimated using the leukemia data mentioned above. Using rigorous theory and reduction techniques, in Efron (2010) the author has

developed a simple formula for capturing the variability of the test statistic corresponding to the empirical cumulative distribution function, namely $\widehat{F}$. Without the correlation in the picture that is assuming independent z_i's, the variance of $\widehat{F}$ can be given as $\frac{\widehat{F}(x)(1-\widehat{F}(x))}{N}$. However, the presence of correlation leads to an additional penalty term while computing $Var\left\{\widehat{F}(x)\right.$. The ultimate goal of Efron (2010) is to estimate the penalty term efficiently. Appendix D provides a detailed description of the methodology employed in Efron (2010).

1.2.1 Estimation of the Correlation Parameter α (Efron (2010))

Keeping in mind that the goal of Efron (2010) is to deduce a simple formula for the covariance of $\mathbf{cov}_1$, one must estimate the correlation parameter, namely α, robustly. The author has elucidated the estimation process using the data coming from the leukemia study. A brief description of the data is given next. **X** denotes the data matrix for the leukemia study with N=7128 rows with the ith row representing the two sample t-statistic comparing the expression levels on the ith gene for all the patients. And n=72 columns representing the total number of participants in the leukemia study (47 ALL and 25 AML patients). Thus x_{ij} of **X** is the expression level for ith gene on jth patient. To reduce the noise in the genetic expression, each column of the **X** is replaced by the corresponding z-values using the following transformation.
$\widetilde{x_{ij}} = \Phi^{-1}((r_{ij} - 0.5)/7128)$ where r_{ij} is the column rank of x_{ij} in the jth column.

The rms correlation parameter α in Efron (2009) is estimated separately for ALL patients, AML patients, and both. A subset of the $N \times n$ matrix **X** namely $\mathbf{X}_0$ with 7128 rows and n_0 columns for $n_0 = 47, 25,$ and 72 are used for computing the empirical distribution of the correlation distribution. Note that $N(N-1)/2$ possible pairwise correlations in each of the three cases. Using these N(N-1)/2 pairwise correlations, one can determine the mean(m) and variance (ν) of the empirical distribution of correlation estimate, namely $\widehat{\rho}$.

The distribution of possible pairwise correlations for the disease categories AML and ALL and the combined data is shown in Figure 1.1 with a total of $N(N-1)/2 = 25400628$ correlations. From Figure 1.1, it can be inferred that the pairwise correlations follow an approximately normal distribution for ALL, AML, and the combined data. Additionally, the

centers of all three normal distributions are roughly centered around zero. This is a critical observation as in Efron (2010) for estimating the correlation parameter, various moments of the distribution of pairwise correlations ($g(\rho)$) are required (.0.21).

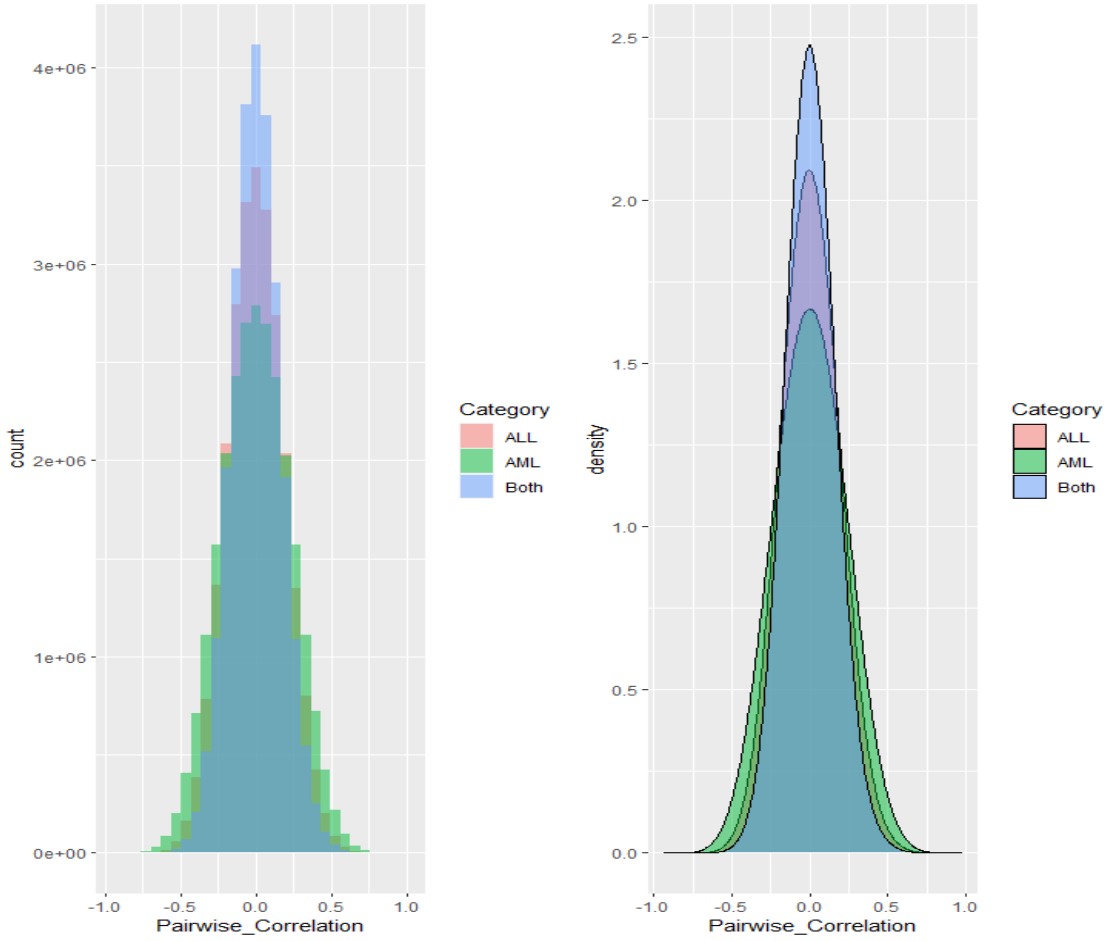

Figure 1.1 Distribution of pairwise correlations

1.2.2 Replications and Discussion of Relevant Results from Efron (2010)

This section presents replications of some results in Efron (2010) that are relevant to my research. Additionally, I discuss some critical aspects of the estimated correlation parameter. Furthermore, the distributions of the two sample t-test statistics and the z-scores after transforming the t-values are presented in figures 1.2 and 1.3, respectively.

Table 1.1 below highlights the estimates of the standard deviation of the right-sided cdf $\widehat{F}(x)$ at five different values, namely 1,2,3,4 and 5. $\widehat{F}(x)$ denotes the estimate of the empirical cdf

at any given x; $\widehat{\mathbf{sd}}$ corresponds to the estimate of standard deviation of $\widehat{F}(x)$ computed considering the presence of correlation between the z values;$\widehat{\text{sd}}_0$ is the estimate of standard deviation ignoring the correlation; $\widehat{\text{sd}}_{\text{perm}}$ represents the permutation standard deviation. The total number of replications used for computing $\widehat{\text{sd}}_{\text{perm}}$ is 2000.

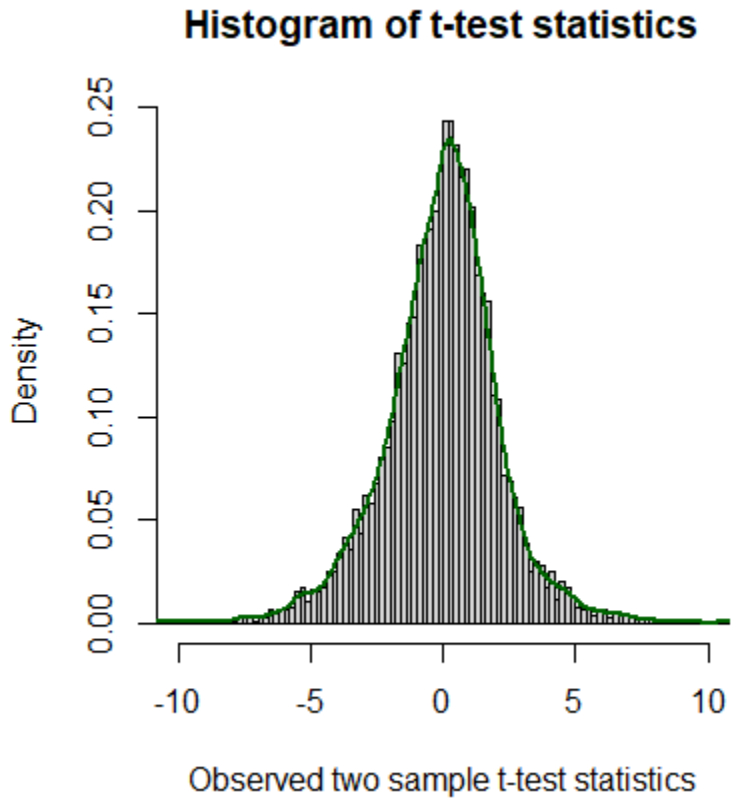

Figure 1.2 Two sample t-test statistics

$\widehat{\text{sd}}_0$ and $\widehat{\mathbf{sd}}$ captures the variability in the empirical cdf at any given x . Whereas the permutation standard deviation determines how precisely $\widehat{F}(x)$ estimates the actual value of the right-sided empirical cdf at x. A smaller value indicates a more precise estimate of the real value of the right-sided empirical cdf at x. The permutation standard deviation is computed from repeated permutations of the 72 patients under investigation, assuming the null hypothesis is true: there is no difference in gene expressions between AML and ALL.

As can be observed from table 1.1 above, disregarding the correlation between z-values

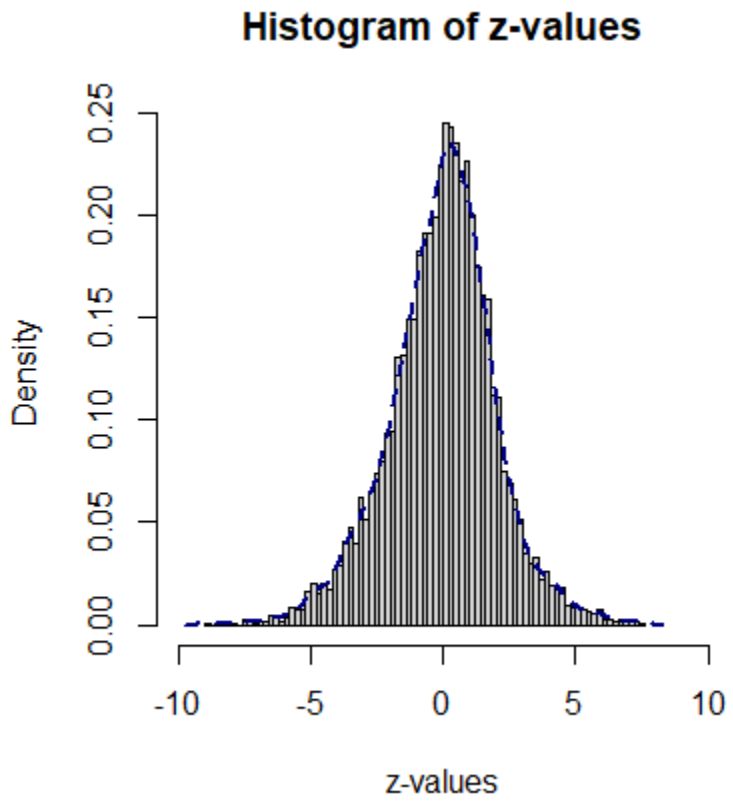

Figure 1.3 z-scores

always leads to underestimation of the variability in the right-handed empirical cdf. Additionally, the low values of $\widehat{\text{sd}}_{\text{perm}}$ suggest the applicability of the methodology in large-scale hypothesis testing.

One of the main objectives of Efron (2010) is to compute the root-mean-square (rms) correlation (.0.21) in an accurate manner using both .0.34 and .0.35. The next table represents the estimated value of the rms correlation using two different formulas as in .0.32 and in .0.33. The rightmost column of the table below provides the result corresponding to the 100 simulations of the following model with N=6000, $n_1 = n_2 = 40$, and true $\alpha = 0.10$.

$$\mu_0, \sigma_0) = (0, 1), \quad p_0 = 0.95 \quad \text{and } (\mu_1, \sigma_1) = (2.5, 1), \quad p_1 = 0.05 \tag{1.2.2}$$

The last column of the table 1.2 shows the simulated mean $\pm$ standard deviation values

Table 1.1 Estimated Standard Deviations for the Empirical Right-Sided Cumulative Distribution Function

Estimates	**x=1**	**x=2**	**x=3**	**x=4**	**x=5**
$\widehat{F}(x)$	0.289	0.128	0.055	0.026	0.011
$\widehat{\mathbf{sd}}$	0.0063	0.0068	0.0062	0.0046	0.0031
$\widehat{\text{sd}}_0$	0.0054	0.0040	0.0027	0.0019	0.0012
$\widehat{\text{sd}}_{\text{perm}}$	0.0222	0.0192	0.0119	0.0063	0.0034
$\widehat{\text{FDR}}(x)$	0.938	0.911	0.678	0.373	0.162

Table 1.2 Root-Mean-Square Estimates of the Pairwise Correlation Parameter α (Efron (2010))

Estimates of α	**ALL**	**AML**	**Both**	**Simulation**
$\widehat{\alpha}$	0.121	0.109	0.114	0.1054 ± 0.0074
$\tilde{\alpha}$	0.118	0.092	0.113	0.1045 ± 0.0075

obtained from 100 simulations of the model specified in 1.2.2. As can be seen from the table 1.2 above, the rms correlations computed using all 72 patients under the leukemia study are the same till the second place of decimal. The rms correlations for the simulation study also establish the accuracy of the estimation process. Note that the variability (sd) of the estimates only differs at the third place of decimal. So, based on the discussions above, we can conclude that the assumption of independence for the gene expression under the Leukemia study is erroneous. By considering the correlation, the statistical inference concerning the Leukemia study can be enhanced for improved accuracy.

1.3 Relevant Definitions, Theorems and Lemmas for Multiple Hypothesis Testing

Within this section, I aim to thoroughly explain the theorems and lemmas that pertain to the research question. Initially, I will outline several pertinent definitions to my research inquiry. Then I will discuss the primary theorem outlined in the research paper by Chen (2016), followed by two brief theorems related to the existing step-wise rejective algorithms.

Definition 1.1. *Confidence Procedure (Casella and Berger (2021)): If $X \sim f(x|\theta)$, where $x \in \mathbf{X}$ and $\theta \in \Theta$, then a confidence procedure is a set in the Cartesian product space $\mathbf{X} \times \Theta$, defined as $\{(x, \theta) : (x, \theta) \in \mathbf{C}\}$, where $\mathbf{C} \in \mathbf{X} \times \Theta$.*

Definition 1.2. *Confidence Set (Casella and Berger (2021)): For fixed x, the θ-section or confidence set is defined as $A(\theta) = \{x : (x, \theta) \in \mathbf{C}\}$.*

Definition 1.3. *Acceptance Region(Casella and Berger (2021)): For fixed θ, the x-section or acceptance region is defined as $C(x) = \{\theta : (x, \theta) \in \mathbf{C}\}$.*

Definition 1.4. *Boole's Inequality(Casella and Berger (2021)): If $\mathbb{P}$ is a probability function, then for $A_1, A_2, ...$ $\mathbb{P}(\cup_{i=1}^{\infty} A_i) \leq \sum_{i=1}^{\infty} \mathbb{P}(A_i)$.*

Definition 1.5. *Bonferroni Inequality(Casella and Berger (2021)): If $\mathbb{P}$ is a probability function, then for any sets $A_1, A_2, ..., A_n$, $\mathbb{P}(\cap_{i=1}^{n} A_i) \geq \sum_{i=1}^{n} \mathbb{P}(A_i) - (n-1)$.*

Theorem 1.1. *(Chen (2016)) For multiple testing problems $H_{i0} : \theta_i \in \Theta_i$ versus $H_{i1} : \theta_i \in \Theta_i^c$, $i = 1, 2, ..., k$ assume that for any nested rejection region and permissible integers i and t, there exists inverted confidence set $C_i^t(\boldsymbol{y})$ that is directed towards Θ_i^c. When screening down from the largest to the smallest ordered p-value, let m be the index that satisfies*

i) $\widehat{P}_{(m)} \geq \frac{\alpha}{k-m+1}$

ii) for any index $i : m < i \leq k$,
$\widehat{P}_{(i)} < \frac{\alpha}{k-i+1}$, *we have*

$$\mathbb{P}\Big(\boldsymbol{\theta} \in \Theta_{(k)}^c \cap ... \cap \Theta_{(m+1)}^c \cap \Theta_{(m)}^{k-m+1}\Big) \geq 1 - \alpha \tag{1.3.1}$$

where $C_0^k\big(\boldsymbol{y}\big) = \Theta$ when m=0 (In this case, all p-values are smaller than the corresponding cut-off values) and $\Theta_{(k+1)}^c = \Theta$ when m=k, for notational convenience.

The proof of Theorem 1.1 is provided in detail in Appendix E.

Holm's and Hochberg's stepwise procedures are viable options for multiple hypothesis testing. However, it has been found that Hochberg's step-up procedure is not as effective in controlling the family-wise error rate. As a result, this dissertation primarily focuses on Holm's step-down Procedure and utilizes it for any subsequent comparative study. Next, I state and present two small theorems related to existing Holm's step-down Procedure.

Theorem 1.2. *Holm's procedure strongly controls the family-wise error rate ($FWER$).*

Proof. Let $H = \{H_1, H_2, ..., H_n\}$ be a given collection of null hypotheses. Let $H^{'} \subseteq H$. Suppose $H_0 = \bigcap_{i=1}^{n}\{H_i \text{is true}\}$ and $H_0^{'} = \bigcap_{i=1}^{|H^{'}|}\{H_j^{'} \text{is true}\}$. The strong control of $FWER$ is defined as $\mathbb{P}_{H_0^{'}}(RejectingH_0^{'}) \leq \alpha \quad \forall H^{'} \subseteq H$.

Consider testing n null hypotheses simultaneously against their respective alternatives. Let I be the set of indices of true hypotheses and m be the number of elements in I. We must show that $\mathbb{P}$(Rejecting H_i for some $i \in I$)$\leq \alpha$ for a given significance level α.

$\mathbb{P}$(Rejecting $H_i,\ i \in I$) [As we can reject I even if only one of the hypotheses in I is false]

=$\mathbb{P}$($\bigcup_{i\in I}$ Rejecting H_i)

$\leq \sum_{i\in I} \mathbb{P}$(Rejecting H_i) [Using Boole's Inequality]

A hypothesis is rejected if the corresponding p-value is smaller than the threshold the sequentially rejective procedure obtained. Let p_i be the p-value corresponding to the i^{th} hypothesis test. Considering that p-values $\sim Uni(0, 1)$, we consider the following scenarios.

Case 1: Let $p_i > \alpha/m \quad \forall i \in I$.

Then $\mathbb{P}(\{i : p_i > \alpha/m\})$=1.

Therefore, $\mathbb{P}(\{i : p_i \leq \alpha/m\})$=0.

Thus $\mathbb{P}$(Rejecting H_i)=0 $\forall i \in I$.

Therefore, $\mathbb{P}$(Rejecting $H_i,\ i \in I$)=$0 < \alpha$

Case 2: $\exists i \in I$ s.t.$p_i \leq \alpha/m$.

This i can be one or more of the m hypotheses in I.

Therefore, $\mathbb{P}$(Rejecting $H_i,\ i \in I$)

=$\mathbb{P}$($\bigcup_{i\in I}$ Rejecting H_i)

$\leq \sum_{i\in I} \mathbb{P}(\{i : p_i \leq \frac{\alpha}{m}\})$

=$\sum_{i\in I} \mathbb{P}(\{p_i \leq \frac{\alpha}{m}\})$

=$\sum_{i\in I} \frac{\alpha}{m}$ [As $p_i \sim Uni(0, 1) \forall i \in I$]

=$m.\frac{\alpha}{m}$ [AS $|I| = m$]=α.

Thus, $\mathbb{P}$(Rejecting $H_i,\ i \in I$)$\leq \alpha$. □

Theorem 1.3. *The Extended Simes (1986) procedure (step-up procedure) is stronger than Holm's Procedure (step-down Procedure). In other words, the Extended Simes (1986) procedure will reject any sub-collection of hypotheses rejected by Holm's procedure.*

Proof. Let $H = \{H_1, H_2, ..., H_m\}$ be a given collection of hypotheses. Suppose we arrange the p-values of the individual hypothesis in ascending order. Let $P_{(1)}, P_{(2)}, ...P_{(m)}$ be the ordered p-values. Let $H_{(i)}$ denote the null hypothesis corresponding to the i^{th} ordered p-value $P_{(i)}$.

Then according to the **Holm's Procedure** at a significance level α if

$P_{(1)} < \alpha/m$ we reject $H_{(1)}$ and proceed to check whether

$P_{(2)} < \alpha/(m-1)$ if true, we reject $H_{(2)}$ and proceed to check whether

$P_{(3)} < \alpha/(m-2)$ ...

$\vdots$

$P_{(k)} < \alpha/(m-k+1)$if true, we reject $H_{(k)}$ and proceed to check whether

$P_{(k+1)} < \alpha/(m-(k+1)+1) = \alpha/(m-k)$...

$\vdots$

$P_{(m)} < \alpha$ Therefore, $for 1 \leq k \leq m$, we sequentially check whether $P_{(k)} < \alpha/(m-k+1)$; if true, we reject the corresponding hypothesis and proceed to the next step; else, we stop and terminate the algorithm. On the other hand, Simes (1986) Procedure sequentially rejects all $H_{(i')}$, $i' \leq i$ if $P_{(i)} \leq \alpha/(m-i+1)$ for $i = m, m-1, ..., 1$.

So, the step-up procedure starts with checking $P_{(m)} \leq \alpha$ if true, we reject all of the hypotheses in consideration, namely $H_{(1)}, H_{(2)}, ..., H_{(m)}$.

If $P_{(m)} > \alpha$, then we proceed to check whether

$P_{(m-1)} \leq \alpha/(m-(m-1)+1) = \alpha/2$ if true, then reject $H_{(1)}, H_{(2)}, ..., H_{(m-1)}$. If $P_{(m-1)} > \alpha/2$, then we proceed to check whether $P_{(m-2)} \leq \alpha/3$ ...

$\vdots$

If $P_{(1)} \leq \alpha/m$ reject $H_{(1)}$. So, if $P_{(k)}$ is rejected by Holm's procedure then by construction $P_{(k)} < \alpha/(m-k+1)$. Now if one of the $P_{(k+1)} \leq P_{(k+1)} \leq ... \leq P_{(m)}$ is less than $\alpha/(m-k+1)$, then $P_{(k)}$ is automatically rejected by the Simes (1986) procedure. If not, then as

$P_{(k)} \leq \alpha/(m-k+1)$ Simes (1986) procedure rejects $P_{(1)}, P_{(2)}, ..., P_{(k)}$. Thus Simes (1986) procedure rejects any hypothesis that Holm's procedure rejects. Moreover, Simes (1986)'s procedure can reject a hypothesis even if Holm's procedure does not reject it due to its top-down approach. □

It's essential to recognize that while Holm's procedure effectively controls the Family-Wise Error Rate ($FWER$) in a strong sense, it may not be the most effective test for simultaneous hypothesis testing. Hence, there is a need for a new algorithm capable of surpassing the statistical power of Holm's step-down Procedure and ensuring strong control over the $FWER$. In the following section, I will delve into the potential inquiries pertinent to my dissertation research, which I introduced in my literature review. These questions are based on my understanding of the existing literature in the field of simultaneous hypothesis. Additionally, I will outline the structure of my dissertation, which will conclude this chapter on literature review.

1.4 Research Plan

An area of potential expansion, enhancement, or application of Efron's work on correlated data in Efron (2010) involves the simultaneous control of FWER for 7128 hypotheses. Efron used the Golub et al. (1999) data set for a large-scale statistical estimation of empirical FDR using the False Discovery Rate. Whereas, in my research, the Golub et al. (1999) data set is used for large-scale simultaneous hypothesis testing that can strongly control the family-wise error rate ($FWER$) at a given significance level (α). For each of the 7128 genes, a two-sample t-statistic is computed under the assumption that the variability of a given gene expression is the same in ALL and AML. Under H_0, it is assumed that there is no significant difference in the gene expressions between ALL and AML. Now, Corresponding to each gene, there are 72 patients. So the degrees of freedom, in this case, is 72-2=70. Then the observed two-sample t test statistics are transformed into a standard normal variate (z-values) by probability integral transformation. $z_i = \Phi^1(F_{70}(t_i)),\ i = 1, 2, ..., N$ where Φ is the cumulative distribution function (cdf) of standard normal distribution, and F_{70} is the cdf of Student's-t distribution with 70 degrees of freedom.

Creating an algorithm that considers all 7128 z-values simultaneously and effectively controls the $FWER$ is possible.

Another approach is to create an innovative statistical method for handling vast sets of related data that can control the family-wise error rate ($FWER$) without relying on the probability integral transformation. To be precise, one can calculate two sample t-test statistics based on the hypothesis that gene expressions in ALL and AML have different variances. Then construct a bootstrap-based exact confidence procedure that can control the $FWER$ in the strong sense. While performing multiple hypothesis testing, controlling the probability of making one or more type one errors in the family is pivotal. By making the $FWER$ not exceed a given threshold α, the probability of making at least one type 1 error is controlled at level α. Below is a brief discussion on $FWER$.

Let $M = \{1, 2, ..., m\}$ be the index set associated with the null hypothesis $H_1, H_2, ..., H_m$ and $M_0 \subseteq M$ be the set of $m_0 = |M_0|$ true hypotheses. Let V denote the number of Type-I errors. Then the family-wise error rate ($FWER$) is defined as $FWER = \mathbb{P}(V > 0)$. The $FWER$ is used ubiquitously in large-scale multiple testing where strong evidence is required.

The last one is to study the robustness of the newly proposed confidence procedure by investigating various models, including Cauchy, Student's t-distribution, skew-normal, and others. Since the Efron (2010) is grounded on the normality assumption, but the approach Efron used is a general method, there is a possibility to extend that to non-parametric or empirical Bayesian analysis. Suppose we want to explore the effectiveness of the rms correlation on gene expressions that do not originate from correlated normal variates but instead from a correlated Cauchy distribution. However, we know that Cauchy does not have a mean or variance. Therefore, it can be interesting to investigate this methodology's relevance in such a scenario. Additionally, a comprehensive simulation study that includes four or more hypotheses can be presented to demonstrate the effectiveness and validity of the proposed methodology.

1.5 Dissertation Structure

In Chapter 1 on Introduction and Literature Review, I thoroughly discuss the current knowledge of existing materials on large-scale correlated data analyses by Efron (2010). As part of my research, I was able to reproduce significant findings from Efron (2010). Additionally, I provided thorough evidence and explanations for definitions and theorems pertinent to my study area, including a critical theorem from Chen (2016).

Chapter 2 comprehensively examines simultaneous inference related to multiple odds ratios. In this chapter, I use step-wise algorithms on simulated data to analyze the statistical power and computational complexity of different testing procedures concerning odds ratios to assess their applicability and aptness.

In the third chapter, I propose a simultaneous exact inference procedure that can strongly control the family-wise error rate with ample evidence. In addition to the above, I am comparing the statistical power of the recently introduced simultaneous inference algorithm with Holm's step-down procedure in different scenarios. Furthermore, I demonstrated the usability of the new algorithm through a non-parametric test setup.

In Chapter 4, I compare the applicability of Holm's step-down procedure and the newly proposed algorithm in large-scale data analysis. The analysis is based on microarray data from Golub et al. (1999) under different underlying assumptions about the data. Specifically, I compute two-sample t-test statistics assuming unequal variances between AML and ALL. I utilize these test statistics to get the bootstrapped p-values and the bootstrapped 95% confidence interval for Holm's procedure and the new algorithm, respectively.

Chapter 5 of this dissertation thoroughly summarizes all the topics discussed in the previous chapters. Furthermore, potential areas for future research are emphasized.

CHAPTER 2 ADVANCED SIMULTANEOUS INFERENCE FOR MULTIPLE ODDS RATIOS

2.1 Motivation

Multiple testing, aka simultaneous hypotheses testing, is useful to counter cognitive bias. In this chapter, we try to analyze existing procedures that can strongly control family-wise error rates for hypothesis testing related to multiple odds ratios. The odds ratios are used to check whether the odds of getting cured of a disease among patients in the treatment group (receiving treatment) is higher than those in the control group (receiving placebos). In the case of the cohort data, odds ratios can be used to check whether the odds of catching an infectious disease are higher among the people exposed to an infected individual than those who have not come in close contact with an infected person. Efron (2010) defined a robust way of finding test statistics' empirical cumulative distribution function in the presence of significant correlation. Efron's procedure estimate the empirical false discovery rate ($\widehat{FDR}$) and thus can be used to conduct multiple hypothesis testing at a given significance level. However, only when all null hypotheses are true, FDR weakly controls the family-wise error rate ($FWER$)Benjamini and Hochberg (1995).

Suppose the number of hypotheses for drawing simultaneous inference under consideration is m. Table 2.1 shows the total number of true null and true alternative hypotheses.

Table 2.1 Contingency Table of True Null and Non-true Null Hypotheses Benjamini and Hochberg (1995)

	Failed to Reject	Rejected	Total
True Null	U	V	m_0
Non-True Null	T	S	$m - m_0$
Total	$m - R$	R	m

Definition 2.1. *Let $Q = V/R$ when $V + S > 0$ and $Q = 0$ when $V + S = 0$. Then, the false discovery rate is defined as FDR=$\mathbb{E}[Q]$.*

Definition 2.2. *The family-wise error rate is defined as $FWER$=$\mathbb{P}(V \geq 1)$.*

Lemma 2.1. *Controlling FDR controls $FWER$ weakly if all the null hypotheses under consideration are true(Benjamini and Hochberg (1995)).*

Proof. We want to show that $\mathbb{E}[Q] \leq \mathbb{P}(V \geq 1)$.

Case 1: $m_0 = m$, that is when all null hypotheses are true. As $m_0 = m$ then $T + S = 0$. Therefore $S = 0$. Thus, $V = R$ in this case. Now, if $V = R = 0$ from definition 2.1, we have $Q = 0$. If $V > 0$ then $V = R$. Thus $Q = 1$. Therefore,

$$\mathbb{P}(V \geq 1) = \mathbb{P}(Q = 1) = 1.\mathbb{P}(Q = 1) + 0.\mathbb{P}(Q = 0) = \mathbb{E}[Q]$$

Thus $FDR = FWER$ when all null hypotheses are true.

Case 2: $m_0 < m$ If $S = 0$ then we have $FDR = FWER$. Suppose, $S > 0$ then $V \leq R$. Now, when $V = 0$ we get $Q = V/R = 0$ and when $V > 0$ then $Q = V/R \leq 1$. Thus, $Q = V/R \leq 1$ for $V \geq 1$. Therefore $Q \leq 1_{\{V \geq 1\}}$. Taking the expectation of both sides, we get $\mathbb{E}[Q] \leq \mathbb{P}(V \geq 1)$.

Thus, by combining cases 1 and 2, we can say that controlling FDR implies control of $FWER$ in the weak sense. □

Holm's step-down procedure strongly controls the family-wise error rate ($FWER$) at a significance level of α (Theorem 1.2). However, Holm's procedure is conservative, using the Bonferroni correction sequentially. On the other hand, Hochberg's step-up procedure does not strongly control the family-wise error rate. Bradley Efron in Efron (2010) uses an estimated false discovery rate ($\widehat{FDR}$) based on an underlying distributional assumption of correlated z-values. Specifically, Efron (2010) defined the empirical right-handed cdf of the test statistic as an estimate of the false discovery rate and showed that a high correlation has minimal impact on the FDR. But controlling FDR can only ensure a weak control over $FWER$ when all null hypotheses under scrutiny are true. The ongoing research aims to develop an exact simultaneous confidence procedure that strongly controls the family-wise error rate ($FWER$). I start by exploring existing step-wise procedures that strongly control the $FWER$ for several odds ratios

involving multiple 2×2 tables.

2.2 Existing Step-wise Multiple Testing Algorithms and Tests

In this section, I briefly discuss Holm's (Step-Down) procedure (Holm (1979)) and Hochberg's (Step-Up) procedure (Hochberg (1988)). Then I talk about two statistical tests used for conducting hypothesis testing related to multiple odds ratios in practice. After that, I investigate the feasibility of the step-wise confidence procedure in Chen (2016) to improve the existing step-wise procedures. Let $H = \{H_1, H_2, ..., H_m\}$ be a given collection of hypotheses. We arrange the p-values of the individual hypothesis in ascending order.
Suppose,$P_{(1)}, P_{(2)}, ...P_{(m)}$ denote the ordered p-values. Let $H_{(i)}$ denote the null hypothesis corresponding to the i^{th} ordered p-value $P_{(i)}$.

2.2.1 Step-wise Rejective Algorithms

Holm's Step-Down Rejective Algorithm for Multiple Testing (Holm (1979)) :

According to Holm's Procedure at a significance level α if $P_{(1)} < \alpha/m$ we reject $H_{(1)}$ and proceed to check whether $P_{(2)} < \alpha/(m-1)$ if true, we reject $H_{(2)}$ and proceed to check whether $P_{(3)} < \alpha/(m-2)$ ...

$\vdots$

$P_{(k)} < \alpha/(m-k+1)$if true, we reject $H_{(k)}$ and proceed to check whether

$P_{(k+1)} < \alpha/(m-(k+1)+1) = \alpha/(m-k)$...

$\vdots$

$P_{(m)} < \alpha$

Therefore, $for 1 \leq k \leq m$, we sequentially check whether $P_{(k)} < \alpha/(m-k+1)$; if true, we reject the corresponding hypothesis and proceed to the next step; else we stop and terminate the algorithm.

Hochberg's Step-Up Rejective Algorithm for Multiple Testing (Hochberg (1988)):

The step-up procedure sequentially rejects all $H_{(i')},\ i' \leq i$ if $P_{(i)} \leq \alpha/(m-i+1)$ for $i = m, m-1, ..., 1$. If $P_{(m)} \leq \alpha$, we reject all the hypotheses under consideration namely

$H_{(1)}, ..., H_{(m)}$.

If $P_{(m)} > \alpha$, then we proceed to check whether $P_{(m-1)} \leq \alpha/(m-(m-1)+1) = \alpha/2$ if true, then reject $H_{(1)}, H_{(2)}, ..., H_{(m-1)}$. If $P_{(m-1)} > \alpha/2$, then we proceed to check whether $P_{(m-2)} \leq \alpha/3$ ...

$\vdots$

If $P_{(1)} \leq \alpha/m$ reject $H_{(1)}$

The step-up procedure is stronger than Holm's step-down procedure because it will reject any hypothesis rejected by Holm's algorithm due to its top-down approach (Theorem 1.3).

Step-wise Confidence Procedure (Chen (2016)):

In case of the step-wise confidence procedure if $\widehat{P}_{(m)} \geq \alpha$ then we fail to reject $H_{(m)}$ and stop and report the $(1-\alpha) \times 100$ % simultaneous confidence interval for the parameters of interest $\theta_{(1)}, ..., \theta_{(m)}$ corresponding to $H_{(1)}, ..., H_{(m)}$.

If $\widehat{P}_{(m)} < \alpha$ then we reject $H_{(m)}$ and proceed to check whether $\widehat{P}_{(m-1)} \geq \alpha/2$. If true, we fail to reject$H_{(m-1)}$ and report the $(1-\alpha/2) \times 100\%$ simultaneous confidence interval for $H_{(1)}, ..., H_{(m-1)}$. Otherwise, we reject $H_{(m-1)}$ and go to the next step...

$\vdots$

If $\widehat{P}_{(i)} \geq \alpha/(m-i+1)$then we fail to reject$H_{(i)}$ and report the $(1-\alpha/(m-i+1)) \times 100\%$ simultaneous confidence interval for $H_{(1)}, ..., H_{(i)}$. Otherwise, we reject $H_{(i)}$ and go to the next step...

$\vdots$

If $\widehat{P}_{(1)} \geq \alpha/m$ then we fail to reject $H_{(1)}$ and report $(1-\alpha/m) \times 100\%$ confidence interval for $H_{(1)}$. Else, we reject $H_{(1)}$.

The sequentially rejective step-wise confidence procedure strongly controls the $FWER$ (Theorem 1.1).

2.2.2 Tests Involving Multiple Odds Ratios

Cochran-Mantel-Haenszel (CMH) Test:

Table 2.2 Case Control Data

	Case	Control	
Exposure	a_i	b_i	m_{1i}
Unexposed	c_i	d_i	m_{2i}
	n_{1i}	n_{2i}	N_i

The Mantel-Haenszel odds ratio estimates the odds ratio for the association between exposure to an infection and having the disease caused by the infection, controlling for the possible confounding effects of the stratifying variable. Suppose we have K categories of the stratifying variable. Then the multiple hypotheses under consideration can be expressed as next.

$$H_0 : OR_1 = OR_2 = OR_3 = ... = OR_K = 1$$

vs

$$H_a : At\ least\ one\ of\ the\ OR_1, OR_2, OR_3, ..., OR_K \neq 1$$

Test Statistic for the Cochran-Mantel-Haenszel (CMH) test is given by

$$CMH = \frac{\sum_{i=1}^{K}(a_i - \frac{m_{1i}n_{1i}}{N_i})}{\sum_{i=1}^{K}(\frac{m_{1i}m_{2i}n_{1i}n_{2i}}{N_i(N_i-1)})}$$

Under the null hypothesis $CMH \sim \chi^2_{df=1}$. The null hypotheses for the CMH test are $OR_1 = OR_2 = ... = OR_K = 1$, where K is the total number of null hypotheses that must be verified simultaneously. Using the CMH test, rejecting the null hypothesis only tells us that at least one of the ORs is not equal to 1. Thus, we lack specific details about the individual null hypothesis. Thus we need an approach that does not consider all strata simultaneously. The Sequential-Mantel-Haenszel test can be considered a potential solution to our concern.

Sequential-Mantel-Haenszel Test:

The Sequential-Mantel-Haenszel test, instead of considering all of the strata together, tests $H_0 : OR = 1$ versus $H_a : OR \neq 1$ for each stratum sequentially. That is the same as testing $H_0 : p_1 = p_2 = p$ versus $H_a : p_1 \neq p_2$ for each stratum sequentially. In this section, I explain the

usage of the well-known two-proportion z-test for employing the sequentially rejective algorithm. I have also employed Holm's step-down procedure using individual χ^2 tests. The test statistic is defined below for $i = 1, 2, 3, 4$.

$$X_i^2 = \frac{(a_i - \frac{m_{1i}n_{1i}}{N_i})}{(\frac{m_{1i}m_{2i}n_{1i}n_{2i}}{N_i(N_i-1)})}$$

Under H_0,$X_i^2 \sim \chi_1^2 \; \forall i$.

Here, we are testing the equality of two proportions for each stratum. So, we can employ the two-proportion z-test for the step-wise procedures and construct a 95% confidence interval for the test statistic for the step-wise confidence procedure. The test statistic for the two proportional z-test for $i = 1, 2, 3, 4$ is defined below.

$$z = \frac{p_1 - p_2}{\sqrt{p(1-p)/m_1 + p(1-p)/m_2}}$$

Under H_0, $z \sim N(0, 1)$ asymptotically. The step-wise confidence procedure can be used for testing multiple odds ratios as it strongly controls the family-wise error rate.

2.3 Simulation

$K = 4$ strata(groups) have been chosen for the simulation study and power analysis. The hypotheses under consideration are stated below.

$$H_{01} : p_{11} = p_{12} \quad \text{vs} \quad H_{a1} : p_{11} \neq p_{12}$$

$$H_{02} : p_{21} = p_{22} \quad \text{vs} \quad H_{a2} : p_{12} \neq p_{22}$$

$$H_{03} : p_{31} = p_{32} \quad \text{vs} \quad H_{a3} : p_{31} \neq p_{32}$$

$$H_{04} : p_{41} = p_{42} \quad \text{vs} \quad H_{a4} : p_{41} \neq p_{42}$$

The chosen level of significance (α) for all the subsequent analyses is 0.05.

Definition 2.3.

$$p_1 = \mathbb{P}(D|E), \quad p_2 = \mathbb{P}(D|E^c)$$

Where E denotes the event of Exposure, and D indicates the event of being under the case group, the group of individuals afflicted by the disease.

For Group 1 we are simulating x_1 and x_2 from two random variables X_1 and X_2 that follow $Binomial(100, p_1)$ and $Binomial(100, p_2)$ respectively. Where, $p_1 = 0.9$ and $p_2 = 0.5$. For Group 2 we are simulating x_1 and x_2 from two random variables X_1 and X_2 that follow $Binomial(200, p_1)$ and $Binomial(200, p_2)$ respectively. Where, $p_1 = 0.5$ and $p_2 = 0.5$. For Group 3 we are simulating x_1 and x_2 from two random variables X_1 and X_2 that follow $Binomial(150, p_1)$ and $Binomial(150, p_2)$ respectively. Where, $p_1 = 0.5$ and $p_2 = 0.5$. For Group 4 we are simulating x_1 and x_2 from two random variables X_1 and X_2 that follow $Binomial(50, p_1)$ and $Binomial(50, p_2)$ respectively. Where, $p_1 = 0.8$ and $p_2 = 0.5$.

Next, I discuss the output of the multiple testing procedures mentioned above using the simulated data set in the statistical programming language R.

As shown in figure 2.1, the Cochran-Mantel-Haenszel test can successfully highlight that there is a significant difference between p_1 and p_2 in at least one of the four groups. However, the CMH test can not determine which of these four groups has a significant difference between p_1 and p_2 (2.3) in a specific manner. Thus, we need a step-wise algorithm to detect the groups with a significant difference between p_1 and p_2 in a precise way.

```
## data:  array(data = c(TBL.gr1, TBL.gr2, TBL.gr3, TBL.gr4), dim = c(2,     2, 4))
## CMH statistic = 1.2240e+01, df = 1.0000e+00, p-value = 4.6769e-04, MH
## Estimate = 1.5506e+00, Pooled Odd Ratio = 1.5544e+00, Odd Ratio of
## level 1 = 1.2264e+01, Odd Ratio of level 2 = 7.8638e-01, Odd Ratio of
## level 3 = 1.1427e+00, Odd Ratio of level 4 = 7.2112e+00
```

Figure 2.1 Output of the CMH test

Figure 2.2 shows the output of the step-wise rejective algorithm using the χ^2 test. According to Holm's step-down procedure, the two groups with a significant difference between p_1 and p_2 (2.3) are group 1 and group 4. As Holm's Step-Down Algorithm terminates after the second step in this case. Thus, Holm's step-down procedure can detect the specific groups with

non-true null hypotheses.

```
##          gr.names     chi.pval      alpha   rej
## TBL.gr1      gr1 2.381729e-10 0.01250000  TRUE
## TBL.gr4      gr4 6.048692e-05 0.01666667  TRUE
## TBL.gr2      gr2 2.713082e-01 0.02500000 FALSE
## TBL.gr3      gr3 6.441599e-01 0.05000000 FALSE
```

Figure 2.2 Output of Step-Wise Procedure using χ^2 test

I ran the step-wise rejective algorithm using the two proportional$z$-test to verify the soundness of the conclusion of Holm's step-down procedure. And we end up with the same conclusion as using the χ^2 test. The output of the two proportional$z$-test is shown in figure 2.3.

```
##          gr.names       z.pval      alpha   rej
## TBL.gr1      gr1 2.381729e-10 0.01250000  TRUE
## TBL.gr4      gr4 6.048692e-05 0.01666667  TRUE
## TBL.gr2      gr2 2.713082e-01 0.02500000 FALSE
## TBL.gr3      gr3 6.441599e-01 0.05000000 FALSE
```

Figure 2.3 Output of Step-Wise Procedure using proportional z- test

In the case of the step-up procedure, as soon as we fail to reject the two hypotheses corresponding to the most significant and second-largest p-values (figure 2.2), respectively, we proceed to the third step and compare the second smallest p-values with respective step α and eventually end up rejecting the hypotheses corresponding to the smallest and second smallest p-values. Thus, the Step-Up Procedure also declares groups 1 and 4 with the non-true null hypotheses.

2.3.1 Step-Wise Confidence Procedure

This section employs the step-wise confidence procedure (as described in section 2.2) using the two-proportion z test. The 95% simultaneous confidence interval for the z-test statistic is reported as shown in figure 2.5 as an output of the Step-Wise Confidence Procedure. Figure 2.4 shows that the result of the simultaneous hypothesis testing is inconclusive. Thus, we do not have

any particular group highlighted as the one with an untrue null. However, it is possible to create a simultaneous confidence interval of 95% that corresponds to the test statistics.

```
##          gr.names       z.pval      alpha          rej
## TBL.gr1      gr1 2.381729e-10 0.01250000 Inconclusive
## TBL.gr4      gr4 6.048692e-05 0.01666667 Inconclusive
## TBL.gr2      gr2 2.713082e-01 0.02500000 Inconclusive
## TBL.gr3      gr3 6.441599e-01 0.05000000 Inconclusive
```

Figure 2.4 Output of the Step-Wise Confidence Procedure

As can be noted from figure 2.5 for groups 1 and 4, the corresponding confidence intervals do not contain 0, indicating that these two groups might have a significant difference between p_1 and p_2 (definition 2.3).

```
[1] "group name gr1"
[1] 0.2598834 0.5601166
[1] "group name gr4"
[1] 0.165488 0.634512
[1] "group name gr2"
[1] -0.18965402  0.06965402
[1] "group name gr3"
[1] -0.1174551  0.1841218
```

Figure 2.5 The 95% simultaneous C.I. for the z-test statistic

2.3.2 Bootstrapped Distribution of Odds Ratios

Here we focus on the bootstrapped distribution of odds ratios (ORs) as we don't know the shape of the underlying sampling distribution of ORs. We generate 10,000 bootstrapped samples, with replacements from the available data, for both the exposed and unexposed categories. The sample sizes used for resampling corresponding to group 1, group 2, group 3, and group 4 are 100, 200, 150, and 50, respectively. Then we compute the $\widehat{OR}_{boot_1}, \widehat{OR}_{boot_2}, ..., \widehat{OR}_{boot_{10,000}}$ corresponding to all four hypotheses under consideration. After that, we report the 95% simultaneous confidence interval based on the bootstrapped distributions. Specifically, by utilizing Bonferroni correction, we cap the simultaneous type I error rate at 0.05. A null hypothesis is rejected if the corresponding confidence interval does not contain 1.

Figure 2.6 showcases the 95% simultaneous confidence interval based on the bootstrapped distributions of odds ratios and the four observed values of the odds ratios. Here we are capping the overall type I error rate at 0.05. Therefore, at an individual level, we get a 98.75% confidence interval for each of the four strata under scrutiny. The confidence intervals corresponding to group 1 and group 4 do contain 1. Therefore the null hypotheses corresponding to groups 1 and 4 are rejected, and we claim that for groups 1 and 4, p_1 (definition 2.3) is significantly different than p_2 (definition 2.3).

The simultaneous bootstrapped 95% confidence set for the ORs is given by

```
##           g1.gr1    g1.gr2    g1.gr3    g1.gr4
## 0.625%   4.63521 0.4729345 0.6339363  2.363636
## 99.375% 64.95349 1.2974492 2.0148322 42.666667
```

And the observed ORs are

```
##    TBL.gr1    TBL.gr2    TBL.gr3    TBL.gr4
## 12.2637363  0.7863818  1.1426941  7.2111801
```

Figure 2.6 The 95% simultaneous C.I. based on the bootstrapped distribution of odds ratios

2.4 Power Analysis

In this section, I compare three statistical methods: Bonferroni Correction, Holm's step-down procedure, and simultaneous confidence interval using the bootstrapped distribution of the odds ratios. When computing the overall power of a statistical test procedure, it's important to note that here the instances are independent. Therefore, to determine the statistical power of a procedure, we can multiply the proportion of rejections corresponding to Group 1 and Group 4, as these two groups are where the alternative hypothesis is true. And by definition, power is the probability of rejecting the null hypothesis when the alternative hypothesis is true. The well-known Fisher's exact test and Wilcoxon's rank sum test are utilized on the simulated data to employ Bonferroni Correction and Holm's step-down procedure.

For Fisher's test and Wilcoxon rank sum test, we compare the $p-$value with the corresponding significance level, and we reject the null hypothesis if the $p-$ value is smaller than

the step-wise threshold. For the bootstrap method, the null hypothesis is rejected if the corresponding confidence interval does not contain 1. The number of simulations used for power analysis is 2000; for the bootstrapped C.I., the number of bootstrapped samples (with replacement) used is 10,000. Below I present a comparative study of the powers of the statistical tests mentioned above.

Table 2.3 Statistical Powers of Different Step-Wise Procedures

Test Name	Bonferroni(Fisher)	Bonferroni(Wilcoxon)	Step Down(Fisher)	Step Down(Wilcoxon)	Bootstrapped C.I.
Power	70%	74%	72%	81%	89%
Time(mins)	0.14	0.09	0.15	0.011	51.53

As seen from table 2.3, the bootstrapped-based confidence procedure has a much higher power over the well-known Bonferroni correction and Holm's Step-Down Algorithm. However, the time complexity of the bootstrapped-based procedure is much higher than the rest. I did not report the power corresponding to the step-up approach because we want to concentrate on the step-wise methods that can strongly control the $FWER$.

CHAPTER 3 STEP-WISE PROCEDURE AND IMPROVEMENT

3.1 Introduction

Suppose we want to test the association between exposure and disease, controlling for the possible confounding effects of the stratifying variable. Say, we have K 2×2 tables, each corresponding to a category of the stratifying variable under consideration.

Table 3.1 Case Control Data

	Case	Control	
Unexposed	a_i	b_i	m_{1i}
Exposure	c_i	d_i	m_{2i}
	n_{1i}	n_{2i}	N_i

$i = 1, 2, ..., K$

In conjunction with the above scenario, consider the following. Say, we are interested in comparing the efficacy of different treatments for a particular disease. Now, the task at hand is to compare the outcome of each treatment against the control group and find out which of the K treatments works best for curing the disease under consideration. In this case, we are interested in whether the odds of getting cured are the same for all of the K treatments or if any specific treatment is significantly efficient in curing the disease.

Table 3.2 Cohort Data

	Cured	Not Cured	
Treatment	a_i	b_i	m_{1i}
Control	c_i	d_i	m_{2i}
	n_{1i}	n_{2i}	N_i

$i = 1, 2, ..., K$

In the latter scenario, we are interested in the simultaneous comparisons of the treatments with a control group. Specifically, we want to test whether the odds of getting cured by treatment i are equal to that of getting cured by treatment j, $\forall i \neq j$ simultaneously. The null and alternative hypotheses can be stated below.

$$H_{0_{ij}} : \frac{p_i/(1-p_i)}{p_j/(1-p_j)} = 1 \quad vs \quad H_{a_{ij}} : \frac{p_i/(1-p_i)}{p_j/(1-p_j)} \neq 1 \quad i,j = 1,2,...,K \quad \text{and } i \neq j$$

Where $p_i = P(Cured|Treatment\ i) \quad$ for $i = 1,2,...,K$ and $p_0 = P(Cured|Control)$.Now the above can be simplified as

$$\begin{aligned} &\frac{p_i/(1-p_i)}{p_j/(1-p_j)} = 1,\ i \neq j \\ \iff &\frac{\frac{p_i/(1-p_i)}{p_0/(1-p_0)}}{\frac{p_j/(1-p_j)}{p_0/(1-p_0)}} = 1,\ i \neq j \\ \iff &\frac{p_i/(1-p_i)}{p_0/(1-p_0)} = \frac{p_j/(1-p_j)}{p_0/(1-p_0)},\ i \neq j \end{aligned}$$

A special case for the above is $\frac{p_i/(1-p_i)}{p_0/(1-p_0)} = \frac{p_j/(1-p_j)}{p_0/(1-p_0)} = 1$ for $i \neq j$ which can be re-written as below.

$$H_{0i} : \frac{p_i/(1-p_i)}{p_0/(1-p_0)} = 1 \quad vs \quad H_{ai} : \frac{p_i/(1-p_i)}{p_0/(1-p_0)} \neq 1, \quad i = 1,2,...,K$$

Lemma 3.1. *If* $OR_i = \frac{p_i/(1-p_i)}{p_0/(1-p_0)} = 1$ *then* $p_i = p_0$.

Proof.

$$\begin{aligned} &\frac{p_i/(1-p_i)}{p_0/(1-p_0)} = 1 \\ \Rightarrow &p_i/(1-p_i) = p_0/(1-p_0) \\ \Rightarrow &p_i(1-p_0) = p_0(1-p_i) \\ \Rightarrow &p_i - p_i p_0 = p_0 - p_0 p_i \\ \Rightarrow &p_i = p_0 \end{aligned}$$

□

Thus we need to test whether treatment i works as well as the placebo (control group) for $i = 1,2,...,K$ simultaneously. Ideally, we would like to develop simultaneous confidence

intervals for the difference between treatment i (p_i) and placebo (p_0) and gain some insight into the efficacy of the treatments under consideration. Finally, we would like to draw a meaningful inference in the sense that we would like to identify the most effective treatment for curing the underlying disease.

3.2 Refined Step-Down Algorithm

Let $H = \{H_1, H_2, ..., H_K\}$ be a collection of hypotheses examining the plausibility of a particular claim for K different strata. Suppose $x_1, x_2, ..., x_K$ are the observed values of the test statistics corresponding to $H_1, H_2, ...,$ and H_K, respectively. Say, $x_{(i)}$ denotes the i^{th} ordered test statistic and $H_{(i)}$ denotes the hypothesis corresponding to $x_{(i)}$.

Let T_{K-i+1} be the test statistic representing the combined data from $H_{(1)}, H_{(2)}, ...$ and $H_{(K-i+1)}$ for $i = 1, 2, ..., K$ and R_{K-i+1} be the rejection rejection region corresponding to T_{K-i+1} such that $\mathbb{P}_{\cap_{j=1}^{i} H_{(K-i+1)}}(T_{K-i+1} \in R_{K-i+1}) = \alpha$ for all $i = 1, ..., K$ where α denotes the significance level. Then the Refined Step-Down Algorithm can be employed by following the steps provided below.

- **Step 1:** If $T_K \notin R_K$, we stop and declare that there is no significant statistical evidence against any of the K hypotheses under consideration. Otherwise, we go to the next step.
- **Step 2:** If $T_{K-1} \notin R_{K-1}$ we stop and declare $H_{(K)}$ false. Else, we go to the next step.

  $\vdots$

- **Step K-1:** If $T_2 \notin R_2$ we stop and declare $H_{(K)}, H_{(K-1)}, ...,$, and $H_{(3)}$ false. Else, we go to the next step.
- **Step K:** If $T_1 \notin R_1$ we stop and declare that all hypotheses except $H_{(1)}$ are false. Else, we declare all hypotheses under consideration are false.

Theorem 3.1. *Let $H = \{H_1, H_2, ..., H_K\}$ be a collection of hypotheses examining the plausibility of a particular claim for K different strata. Suppose $x_1, x_2, ..., x_K$ are the observed values of the*

test statistics corresponding to $H_1, H_2, ...,$ and H_K, respectively. Then the Refined Step-Down Algorithm is more powerful than Holm's step-down procedure.

Proof. Let $H' = \{H_{t_1}, ..., H_{t_{n_0}}\} \subseteq H$ where $n_0 \leq K$ be a collection of hypotheses rejected by Holm's step-down procedure.. Without loss of generality, let $t_1 < t_2 < ... < t_{n_0}$. As Holm's procedure follows a step-down approach therefore $H_{t_i} = H_{(i)}$ for $i = 1, ..., n_0$ and the corresponding p-values are $p_{(1)}, ..., p_{(n_0-1)}$ and $p_{(n_0)}$.

Keeping in mind that the hypotheses under consideration are related to the very same claim across K strata and more extreme values of the test statistics lead to smaller p-values, the test statistics corresponding to $p_{(i)}$ is $x_{(K-i+1)}$ for $i = 1, 2, ..., K$. Thus the test statistics corresponding to $H_{(1)}, ..., H_{(n_0)}$ are $x_{(K)}, ..., x_{(K-n_0+1)}$.

Now by construction, T_{K-i+1} represents the test statistic for the combined data set corresponding to $H_{(1)}, H_{(2)}, ..., H_{(K-i+1)}$ for $i = 1, 2, ..., K$. Therefore the presence of extreme observations in at least one of the data sets corresponding to $H_{(1)}, H_{(2)}, ..., H_{(K-i+1)}$ will lead to $T_{K-i+1} \in R_{K-i+1}$ for $i = 1, 2, ..., K$.

As $x_{(K)} > x_{(K-1)} > \ldots > x_{(K-n_0+1)}$ representative of the extreme observations corresponding to the untrue hypotheses, therefore, the combined test statistics $T_K, T_{K-1}, ..., T_{K-n_0+1}$ will fall in their respective rejection regions by construction. In other words, $T_K \in R_K, T_{K-1} \in R_{K-1}, ..., T_{K-n_0+1} \in R_{K-n_0+1}$.

- Case 1: If $T_{K-n_0} \notin R_{K-n_0}$ then the hypotheses $H_{(1)}, ..., H_{(n_0)}$ are declared untrue by the Refined Step Down Algorithm
- Case 2: If $T_{K-n_0} \in R_{K-n_0}$ and $T_{K-n_0-1} \notin R_{K-n_0-1}$ then the hypotheses $H_{(1)}, ..., H_{(n_0)}$ and $H_{(n_0+1)}$ are declared untrue by the Refined Step Down Algorithm.

So, the Refined Step-Down Algorithm rejects either H' or a collection of hypotheses containing H'. Hence, the Refined Step-Down Algorithm is proven to be more powerful than Holm's step-down procedure. □

Lemma 3.2. *Let k be the number of categories of a chosen stratifying variable. Suppose we are interested in testing the following hypotheses simultaneously at a selected significance level α.*

$$H_{0i} : OR_i = 1 \quad vs \quad H_{ai} : OR_i \neq 1 \quad \textit{for at least one } i \in \{1, 2, ..., K\}$$

*Where $p_{1i} = \mathbb{P}(D_i|E_i)$ and $p_{2i} = \mathbb{P}(D_i|E_i^c)$ and $OR_i = \frac{p_{1i}}{(1-p_{1i})} / \frac{p_{2i}}{(1-p_{2i})}$ for $i = 1, 2, ..., K$. Then **Procedure A** strongly controls the family-wise error rate at α.*

Precedure A:

Suppose a set of multiple 2×2 tables is given. Say, there are K 2×2 tables.

Step 1: Compute the observed value of ORs for the given K 2×2 tables

Step 2: Arrange the the ORs in an ascending order $\widehat{OR}_{(1)} \leq \widehat{OR}_{(2)} \leq ... \leq \widehat{OR}_{(K)}$, where $\widehat{OR}_{(i)}$ denotes the i^{th} ordered $\widehat{OR}$

Step 3: Next, we employ the Cochran-Mantel-Haenszel (CMH) test for testing the following null and alternative hypotheses at the given significance level of α.

$$H_0 : OR_{(1)} = OR_{(2)} = ... = OR_{(K)} = 1 \quad vs \quad H_a : OR_{(i)} \neq 1 \quad \text{for at least one } i \in \{1, 2, ..., K\}$$

If we fail to reject the null hypothesis, we stop the procedure. Else, we go to the next step.

Step 4: We employ the CMH test for testing the following null and alternative hypotheses at the given significance level of α.

$$H_0 : OR_{(1)} = OR_{(2)} = ... = OR_{(K-1)} = 1 \quad vs \quad H_a : OR_{(i)} \neq 1 \text{ for at least one } i \in \{1, 2, ..., K-1\}$$

If we fail to reject the null hypothesis, we stop the procedure. Else, we go to the next step.

$\vdots$

Step m: We employ the CMH test for testing the following null and alternative

hypotheses at the given significance level of α.

$$H_0 : OR_{(1)} = OR_{(2)} = ... = OR_{(m)} = 1 \quad vs \quad H_a : OR_{(i)} \neq 1 \text{ for at least one } i \in \{1, 2, ..., m\}$$

If we fail to reject the null hypothesis, we stop the procedure. Else, we go to the next step.

$\vdots$

Step K-1: We employ the CMH test for testing the following null and alternative hypotheses at the given significance level of α.

$$H_0 : OR_{(1)} = OR_{(2)} = 1 \quad vs \quad H_a : OR_{(i)} \neq 1 \text{ for at least one } i \in \{1, 2\}$$

If we fail to reject the null hypothesis, we stop the procedure. Else, we go to the next step.

Step K: We employ the CMH test for testing the following null and alternative hypotheses at the given significance level of α.

$$H_0 : OR_{(1)} = 1 \quad vs \quad H_a : OR_{(1)} \neq 1$$

If we fail to reject the null hypothesis, we stop the procedure and conclude we have significant statistical evidence that except $OR_{(1)}$, the rest of the ORs are not equal to 1. Otherwise, we conclude that all ORs are significantly different than 1.

It is worth noting that Holm's procedure control the overall significance level by utilizing a progressively increasing significance level $\alpha/K, \alpha/(K-1), ..., \alpha/2$, and α. However, the newly proposed **Procedure-A** tests for all $(n-j)$ hypotheses simultaneously at a significance level α, resulting in a more effective test procedure as compared to the use of $\alpha/(n-j)$, for $j = n-K, n-K+1, ..., n-2, n-1$ in Holm's step down procedure. Note that we can either use the Cochran-Mantel-Haenszel (CMH) or the Chi-Square goodness of fit test at each stage, based on the relevant dataset, and compare the p-values to the designated significance level (α).

Theorem 3.2. *Procedure A strongly controls Family Wise Error Rate (FWER)*

Proof. The hypothesis test considered under procedure A is as below.

$$H_{0i} : OR_i = 1 \quad vs \quad H_{ai} : OR_i \neq 1 \quad \text{for at least one } i \in \{1, 2, ..., K\}$$

Given K, 2×2 tables the observed values (test statistics) of the ORs, $\widehat{OR}_i, i = 1, ..., K$ are computed. Then, the observed test statistics are arranged in an ascending order $\widehat{OR}_{(1)} \leq \widehat{OR}_{(2)} \leq ... \leq \widehat{OR}_{(K)}$ and hypotheses are re-written as below.

$H_0 : OR_{(1)} = OR_{(2)} = ... = OR_{(K)}$ vs $H_a : OR_{(i)} \neq 1$ for at least one $i \in \{1, 2, ..., K\}$.

Now, we employ the Cochran–Mantel–Haenszel (CMH) test stepwise. Say, T_m denotes the test statistic representing the data from all of the m 2×2 tables corresponding to $OR_{(1)}, OR_{(2)}, ...,$ and $OR_{(m)}$ respectively, $m \in \{1, 2, ..., K\}$. And R_m denotes the rejection region for the m^{th} step, $m \in \{1, 2, ..., K\}$.

Suppose $T_K \in R_k$ and $T_{K-1} \notin R_{K-1}$. Then

$\mathbb{P}_{H_0}(OR_{(K)} \neq 1) = \mathbb{P}(\text{At least one } OR_i \neq 1 | H_0) = \mathbb{P}(T_K \in R_K) = \alpha$.

As we are testing every level at a significance level α. Thus $\mathbb{P}(T_i \in R_i) = \alpha \; \forall i$.

Let $K_0 \subseteq \{1, ..., K\}$ be an arbitrary subset such that $OR = 1$ for $i \in K_0$ and $OR \neq 1$ for $i \notin K_0$. Suppose $|K_0| = n_0$.

Say, $K_0 = \{t_1, ..., t_{n_0}\}$ and without loss of generality suppose $t_1 < t_2 < ... < t_{n_0}$. As we arrange the observed ORs in ascending order, we are considering $H_0 : OR_{t_1} = OR_{t_2} = ... = OR_{t_{n_0}} = 1$ vs $H_a : OR_i \neq 1$ for at least one $i \in K_0$. Now, rejecting at least one $H_i,\ i \in K_0$ only happens when the test statistic involving that H_i belongs to the

corresponding rejection region. Thus, in this case, the

$$
\begin{aligned}
&\text{FWER} \\
&= \mathbb{P}(\text{Reject at least one } H_i, i \in \{1, 2, ..., K\} | H_i,\ i \in K_0) \\
&= \mathbb{P}(T_{K-n_0+1} \in R_{K-n_0+1}) \\
&\leq \alpha \quad \text{as at every step of the procedure the level of significance is } \alpha
\end{aligned}
$$

□

Lemma 3.3. *Ordering $\widehat{p}$s is the same as ordering $\widehat{OR}$s.*

Proof.

$$
\begin{aligned}
&\widehat{OR_i} \leq \widehat{OR_j} \quad i \neq j \\
&\Leftrightarrow \frac{\widehat{p_i}/(1-\widehat{p_i})}{\widehat{p_0}/(1-\widehat{p_0})} \leq \frac{\widehat{p_j}/(1-\widehat{p_j})}{\widehat{p_0}/(1-\widehat{p_0})} \quad i \neq j \\
&\Leftrightarrow \frac{\widehat{p_i}}{1-\widehat{p_i}} \leq \frac{\widehat{p_j}}{1-\widehat{p_j}} \quad i \neq j \\
&\Leftrightarrow \widehat{p_i}(1-\widehat{p_j}) \leq \widehat{p_j}(1-\widehat{p_i}) \quad i \neq j \\
&\Leftrightarrow \widehat{p_i} \leq \widehat{p_j} \quad i \neq j
\end{aligned}
$$

□

3.3 Power Analysis

This section thoroughly compares the statistical power exhibited by the newly Refined Step Down Algorithm and Holm's step-down procedure. under various scenarios. It is important to note that in this chapter, if the alternative hypothesis is true, the observed odds ratios for the four considered hypotheses are no longer independent. This is because they all share the same control group. Therefore, one cannot simply multiply the power corresponding to the individual untrue null hypotheses to get the overall power for the testing procedure. This section provides a comprehensive power analysis using simulation to compare the statistical powers of the suggested algorithm and Holm's step-down procedure. A detailed description of the power computation has been given for every scenario. Table 3.3 summarizes the power comparison by varying the

parameter value for the control group (p_0) for the simultaneous hypotheses involving proportionality tests.

Definition 3.1. *If $\mathbb{P}$ is a probability function,then Power=$\mathbb{P}$(Rejecting at least one H_i, $i \in I \mid$ I is the collection of untrue null hypothesis)*

In the context of non-parametric analysis, table 3.4 provides a comparison between Procedure B (the suggested algorithm) and Holm's step-down procedure, showcasing their statistical powers. The comparison is based on varying the parameter for the control group (μ_0). Based on the statistical power analysis, it is evident that the Refined Step-Down Algorithm outperforms the step-down procedure by a significant margin.

Table 3.3 Power Comparison between Holm's Step-Down Procedure and the Refined Step-Down Algorithm Using Proportionality-tests

p_0	Holm's Step-Down Procedure (χ^2)	Refined Step-Down Algorithm(χ^2)	Holm's Step-Down Procedure (χ^2)	Refined-Step Down Algorithm (CMH)
0.40	76%	96%	86%	98%
0.41	73%	95%	84%	97%
0.42	71%	95%	80%	96%
0.43	69%	94%	77%	94%
0.44	67%	94%	74%	93%
0.45	65%	93%	70%	91%
0.46	63%	93%	66%	88%
0.47	62%	93%	62%	86%
0.48	61%	93%	58%	83%
0.49	60%	92%	53%	79%
0.50	60%	92%	49%	75%

Table 3.4 Power Comparison between Holm's Step-Down Procedure and Refined Step-Down Algorithm Using Non-parametric Tests

μ_0	0.25	0.28	0.30	0.33	0.35	0.38	0.40	0.43	0.45	0.48	0.50
Holm's Step-Down Procedure	64%	62%	62%	60%	59%	57%	56%	55%	53%	52%	51%
Refined Step Down Algorithm	95%	95%	94%	94%	93%	93%	93%	92%	91%	91%	90%

3.3.1 Power Comparison Holm's Procedure versus Refined Step-Down Algorithm Using Chi-square Goodness of Fit Test

- **Step-1:** Suppose we have one control group and four treatment groups and the corresponding success probabilities $\mathbb{P}(cured|group)$ are p_0, $p_1 = 0.3$, $p_2 = 0.7$, $p_3 = p_0$,

and $p_4 = p_0$ respectively. The corresponding sample sizes are $n_0 = 50$, $n_1 = 50$, $n_2 = 50$, $n_3 = 50$, and $n_4 = 50$ respectively.

- **Step-2:** We generate n_i observations from $Bernoulli(p_i)$ for $i = 0, 1, 2, 3, 4$
- **Step-3:** Run the step-wise algorithms. We employ Holm's Step down procedure using the chi-square goodness of fit test for the pairwise comparison. For the Refined Step-Down Algorithm, we start by applying chi-square goodness of fit test on the combined data set of 5 groups. Then as described in the Refined Step-Down Algorithm, we remove one column at a time starting with the treatment column corresponding to the table with largest observed odds ratio ($\widehat{OR}$)
- **Step-4:** If the algorithm rejects the null hypothesis for either group 1 or group 2, we consider that as a valid rejection of the hypotheses simultaneously
- **Step-5:** Finally we run steps 2, 3, and 4 $N = 50,000$ times and report the average as the power
- **Step-6:** By varying p_0 from 0.4 to 0.6 with a step size of 0.01, we generate the power curve

The power curve shown in figure 3.1 is generated by using the algorithm specified above.

The same algorithm generates the power curve shown in figure 3.2 mentioned above except in **step-3** we remove one column at a time starting with the treatment column corresponding to the largest observed sample proportion among the treatment groups

3.3.2 Power Comparison Holm's Procedure versus Refined Step-Down Algorithm Using the CMH Test

- **Step-1:** Suppose we have one control group and four treatment groups and the corresponding success probabilities $\mathbb{P}(cured|group)$ are p_0, $p_1 = 0.6$, $p_2 = 0.7$, $p_3 = 0.6$, and $p_4 = 0.7$ respectively. The corresponding sample sizes are $n_0 = 50$, $n_1 = 50$, $n_2 = 50$, $n_3 = 50$, and $n_4 = 50$ respectively.

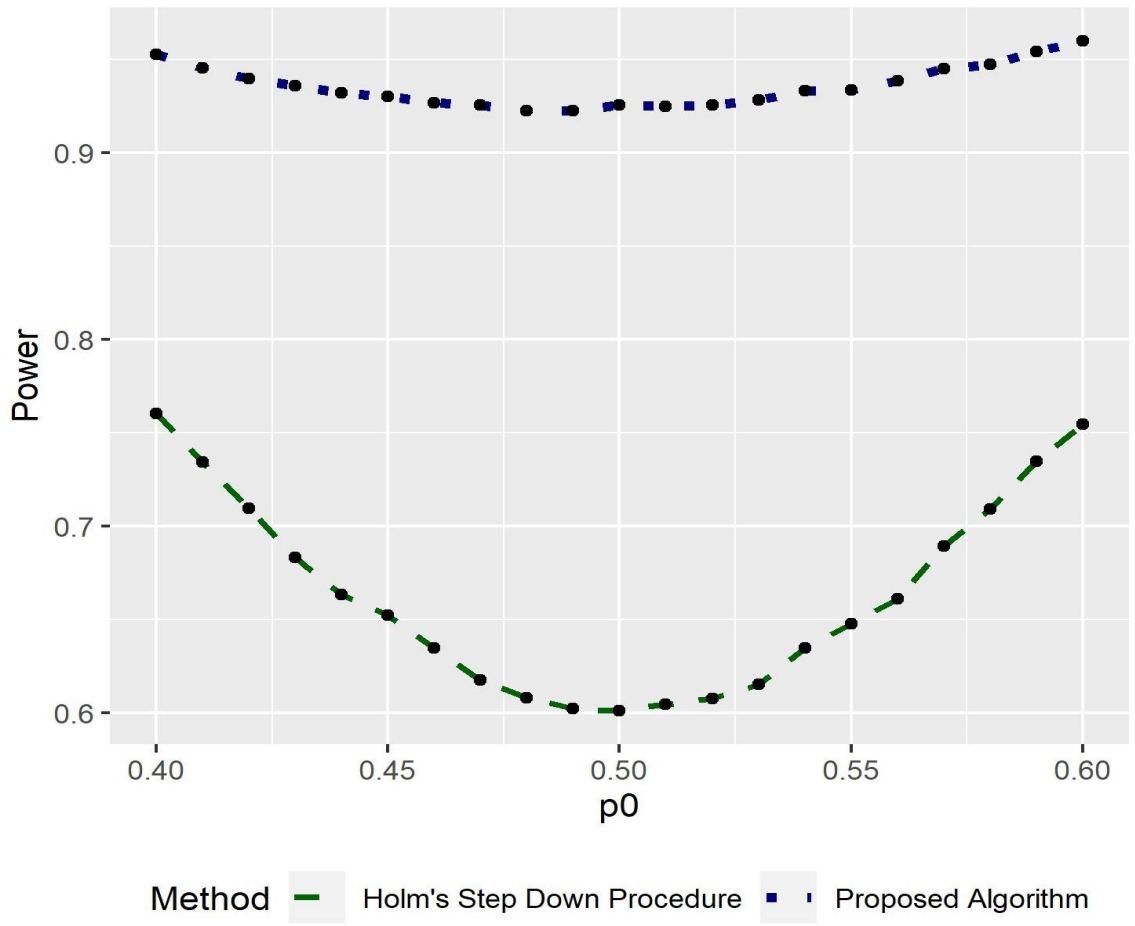

Figure 3.1 Power Comparison between Holm's Step-Down Procedure and the Refined Step-Down Algorithm (proposed algorithm) using chi-square goodness of fit test, with fixed alternative p=0.3, 0.7 (arranging observed ORs)

- **Step-2:** We generate n_i observations from $Bernoulli(p_i)$ for $i = 0, 1, 2, 3, 4$
- **Step-3:** Run the step-wise algorithms. We employ Holm's Step down procedure using the chi-square goodness of fit test for the pairwise comparison. For the Refined Step-Down Algorithm, we use Cochran-Mantel-Haenszel (CMH) test on the combined data set by assuming the presence of a stratifying factor. Then as described in the Refined Step-Down Algorithm, we remove one table at a time, starting with the the table with largest observed odds ratio ($\widehat{OR}$)
- **Step-4:** If the algorithm rejects the null hypothesis for either group 1 or group 2 or group 3,

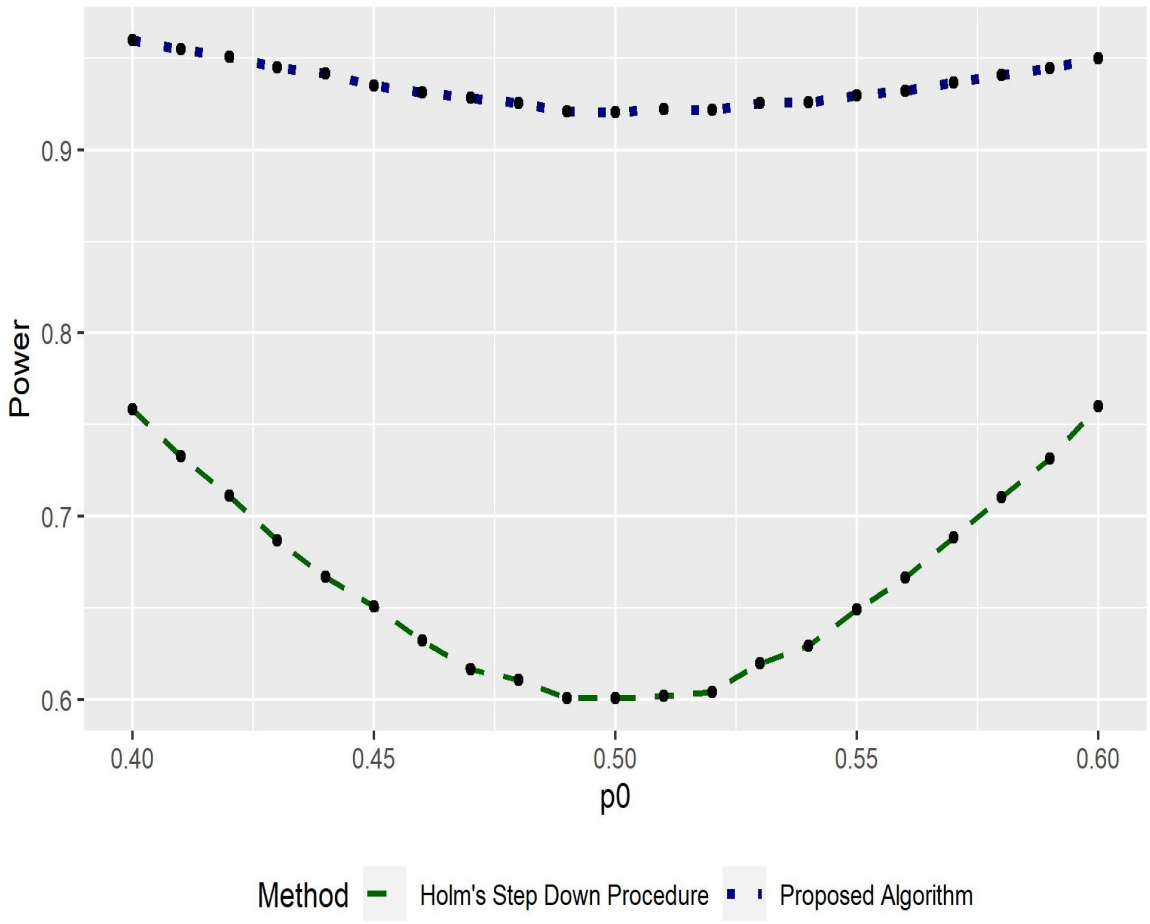

Figure 3.2 Power comparison between Holm's Step-Down Procedure and the Refined Step-Down Algorithm (proposed algorithm) using chi-square goodness of fit test, with fixed alternative p=0.3, 0.7 (arranging sample proportions)

or group 4, we consider that as a valid rejection of the hypotheses simultaneously

- **Step-5:** Finally we run steps 2, 3, and 4 $N = 50,000$ times and report the average as the power

- **Step-6:** By varying p_0 from 0.4 to 0.6 with a step size of 0.01, we generate the power curve

Figure 3.3 represents the power curve generated by using the algorithm specified above.

3.4 Holm's Procedure versus Refined Step Down Algorithm Using Non-parametric Tests

Say, We have patients' blood glucose data for 4 medications besides the control group, the group without any medication. And we are interested in testing the efficacy of the medications

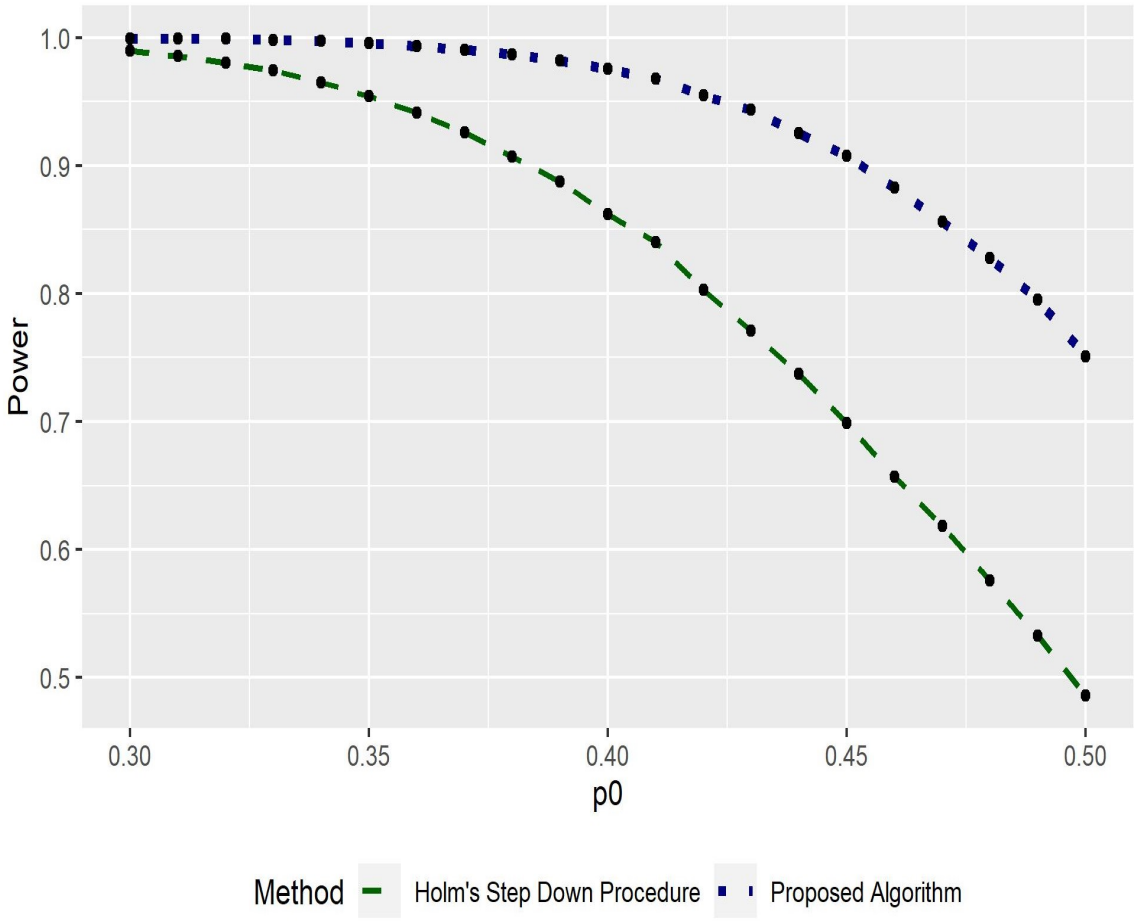

Figure 3.3 Power comparison between Holm's Step-Down Procedure and the Refined Step-Down Algorithm (proposed algorithm) by using Cochran-Mantel-Haenszel test for the proposed algorithm, with fixed alternative p=0.6, 0.7

against the control group. One can interpret that as comparing the center of 4 random variables $X_1, ..., X_4$ with the center of the baseline random variable X_0. Suppose we want to test whether the centers of the distributions of K different random variables are the same as that of the baseline random variable. However, we don't know the shape of the distributions. Say, μ_0 denotes the center of the distribution of the baseline random variables; for the example above, K=4. Then one of the many possible ways of formulating the null and alternative hypotheses could be $H_0: \mu_0 = \mu_1 = ... = \mu_{K-1} = \mu_K$ vs $H_a: \mu_0 \leq \mu_1 \leq ... \leq \mu_{K-1} \leq \mu_K$, with at least one strict ineqality

Where μ_i denotes the centers of the underlying unknown distributions for $i = 0, 1, ..., K$.

Procedure B is an exact simultaneous testing algorithm for the testing above. We employ the Jonckheere Terpstra test at every step and compare the p-values with the given significance level (α).

Precedure B:

Step 1: Compute the observed value of the sample median ($\widehat{median}$) for the given K data sets.

Step 2: Arrange the the sample medians in an ascending order $\widehat{median}_{(1)} \leq \widehat{median}_{(2)} \leq ... \leq \widehat{median}_{(K)}$, where $\widehat{median}_{(i)}$ denotes the i^{th} ordered $\widehat{median}$

Step 3: Next, we employ the Jonckheere Terpstra test for testing the following null and alternative hypotheses at the given significance level of α.

$H_0 : \mu_0 = \mu_{(1)} = ... = \mu_{(K)}$ vs $H_a : \mu_0 \leq \mu_{(1)} ... \leq \mu_{(K)}$ with at least one strict inequality.

If we fail to reject the null hypothesis, we stop the procedure. Else, we go to the next step.

Step 4: We employ the Jonckheere Terpstra test for testing the following null and alternative hypotheses at the given significance level of α.

$H_0 : \mu_0 = \mu_{(1)} = ... = \mu_{(K-1)}$ vs $H_a : \mu_0 \leq \mu_{(1)} ... \leq \mu_{(K-1)}$ with at least one strict inequality.

If we fail to reject the null hypothesis, we stop the procedure. Else, we go to the next step.

$\vdots$

Step m: We employ the Jonckheere Terpstra test for testing the following null and alternative hypotheses at the given significance level of α.

$H_0 : \mu_0 = \mu_{(1)} = ... = \mu_{(m)}$ vs $H_a : \mu_0 \leq \mu_{(1)} ... \leq \mu_{(m)}$ with at least one strict inequality.

If we fail to reject the null hypothesis, we stop the procedure. Else, we go to the next step.

$\vdots$

Step K-1: We employ the Jonckheere Terpstra test for testing the following null and alternative hypotheses at the given significance level of α.

$H_0 : \mu_0 = \mu_{(1)} = \mu_{(2)}$ versus $H_a : \mu_{(0)} \leq \mu_{(1)} \leq \mu_{(2)}$ with at least one strict inequality.

If we fail to reject the null hypothesis, we stop the procedure and conclude we have significant statistical evidence that significant differences exist between μ_0 and $\mu_{(i)}, i = 3, ..., K$. Otherwise, we conclude that at least one of the $\mu_{(1)}$ and $\mu_{(2)}$ along with $\mu_{(i)}, i = 3, ..., K$ is strictly greater than μ_0.

Holm's procedure control the overall significance level by using $\alpha/K, \alpha/(K-1), ..., \alpha/2, \alpha$, which can be improved by **Procedure-B** in which the test is for all $(n-j)$ hypotheses simultaneously, instead of $\alpha/(n-j)$.

Theorem 3.3. *Procedure B strongly controls Family Wise Error Rate ($FWER$)*

Proof. The hypothesis test considered under procedure B is as below.

$$H_0 : \mu_0 = \mu_1 = ... = \mu_{K-1} = \mu_K \quad \text{vs} \quad H_a : \mu_0 \leq \mu_1 \leq ... \leq \mu_{K-1} \leq \mu_K, \text{with at least one strict inequality}$$

Where μ_i denotes the centers of the underlying unknown distributions for $i = 0, 1, ..., K$.

K sample medians are computed as estimates of the centers the $\widehat{\mu_i}, i = 1, ..., K$ are computed. Then, the observed test statistics are arranged in an ascending order $\widehat{median}_{(1)} \leq \widehat{median}_{(2)} \leq ... \leq \widehat{median}_{(K)}$ and hypotheses are re-written as below.

$H_0 : \mu_0 = \mu_{(1)} = ... = \mu_{(K)}$ vs $H_a : \mu_0 \leq \mu_{(1)} ... \leq \mu_{(K)}$ with at least one strict inequality.

Now, we employ the Jonckheere Terpstra test stepwise. Say, T_m denotes the test statistic representing the data from all of the m hypothesis corresponding to $\mu_{(1)}, \mu_{(2)}, ...,$ and $\mu_{(m)}$ respectively together with the control group that is the data corresponding to μ_0, $m \in \{1, 2, ..., K\}$. And R_m denotes the rejection region for the m^{th} step, $m \in \{1, 2, ..., K\}$.

Suppose $T_K \in R_k$ and $T_{K-1} \notin R_{K-1}$. Then

$\mathbb{P}(At\ least\ one\ \mu_i, i \in \{1, ..., K\} > \mu_0 | H_0) = \mathbb{P}(T_K \in R_K) = \alpha.$

As we are testing every level at a significance level α. Thus $\mathbb{P}(T_i \in R_i) = \alpha\ \forall i$.

Let $K_0 \subseteq \{1, ..., K\}$ be an arbitrary subset such that $\mu_0 = \mu_i, i \in K_0$. Suppose $|K_0| = n_0$.

Say, $K_0 = \{t_1, ..., t_{n_0}\}$ and without loss of generality suppose $t_1 < t_2 < ... < t_{n_0}$. As we

arrange the observed sample medians in ascending order, we are considering $H_0 : \mu_0 = \mu_{t_1} = \mu_{t_2} = ... = \mu_{t_{n_0}}\ vs\ H_a : \mu_0 \leq \mu_{t_1} \leq \mu_{t_2} \leq ... \leq \mu_{t_{n_0}}$ with at least one strict inequality. Now, rejecting at least one $H_i,\ i \in K_0$ only happens when the test statistic involving that H_i belongs to the corresponding rejection region. Thus, in this case, the

$$
\begin{aligned}
&\text{FWER} \\
&= \mathbb{P}(\text{Reject at least one } H_i, i\ \in \{1, 2, ..., K\} | H_i,\ i \in K_0) \\
&= \mathbb{P}(T_{K-n_0+1} \in R_{K-n_0+1}) \\
&\leq \alpha \quad \text{as at every step of the procedure the level of significance is } \alpha
\end{aligned}
$$

□

I have also analyzed the power of the Refined Step-Down Algorithm against the power of Holm's Step Down Procedure to cover scenarios where underlying Distributional assumptions do not hold. The null and alternative hypotheses under consideration are as below.

$H_0 : \mu_0 = \mu_1 = \mu_2 = \mu_3 = \mu_4 \quad \text{vs} \quad H_a : \mu_0 \leq \mu_1 \leq \mu_2 \leq \mu_3 \leq \mu_4$, with at least one strict inequality

Where μ_i denotes the centers of the underlying unknown distributions for $i = 0, 1, 2, 3, 4$.

- **Step-1:** Suppose we have one control group and four treatment groups, and the outcome variable follows unknown distributions centered around μ_0, $\mu_1 = 1$, $\mu_2 = 1.5$, $\mu_3 = 1.5$, and $\mu_4 = 2$ respectively. The corresponding sample sizes are $n_0 = 20$, $n_1 = 20$, $n_2 = 20$, $n_3 = 20$, and $n_4 = 20$ respectively.

- **Step-2:** We generate n_i observations from $Cauchy(\mu_i, 1)$ for $i = 0, 1, 2, 3, 4$

- **Step-3:** Run the stepwise algorithms. We employ Holm's Step- Down procedure using the Wilcoxon test for the pairwise comparison ($H_{0i} : \mu_0 = \mu_i$ versus $H_{0i} : \mu_0 \neq \mu_i$ for i=1, 2, 3, 4). Figure3.5 shows the power curve using the Wilcoxon test for pairwise comparison with

one-sided alternative ($H_{0i} : \mu_0 = \mu_i$ versus $H_{0i} : \mu_0 < \mu_i$ for i=1, 2, 3, 4). For the Refined Step-Down Algorithm, we use the Jonckheere Terpstra test on the combined data set. Then as described in the Refined Step-Down Algorithm, we remove one data set at a time, starting with the data set with the largest sample median ($\widehat{median}$)

- **Step-4:** If the algorithm rejects the null hypothesis for either group 1 or group 2 or group 3, or group 4, we consider that as a valid rejection of the hypotheses simultaneously
- **Step-5:** Finally we run steps 2, 3, and 4 $N = 50,000$ times and report the average as the power
- **Step-6:** By varying μ_0 from 0 to 0.5 with a step size of 0.025, we generate the power curve

Figure 3.4 represents the power curve generated by using the algorithm specified above.

As can be seen from figures 3.1, 3.2, 3.3, and 3.4, the statistical power of the new simultaneous inference exact procedure can be substantially higher than that of the Holm's step down procedure for the scenarios especially where the differences between the null-values and non-null values of the parameter of interest are negligible.

Next, I demonstrate the applicability of the newly Refined Step Down Algorithm using a real-life example. Appendix C has the R code for implementation.

3.5 Real Life Example

The Refined Step Down Algorithm is demonstrated by implementation using the binge alcohol use data available on cdc.gov for the year 2019.

Table 3.5 Percentages of Binge Alcoholism in 2019 across Different Age Groups

Age Group	2019
12-13	0.5%
14-15	3.2%
16-17	10.8%
18-25	34.3%
26-34	37.4%

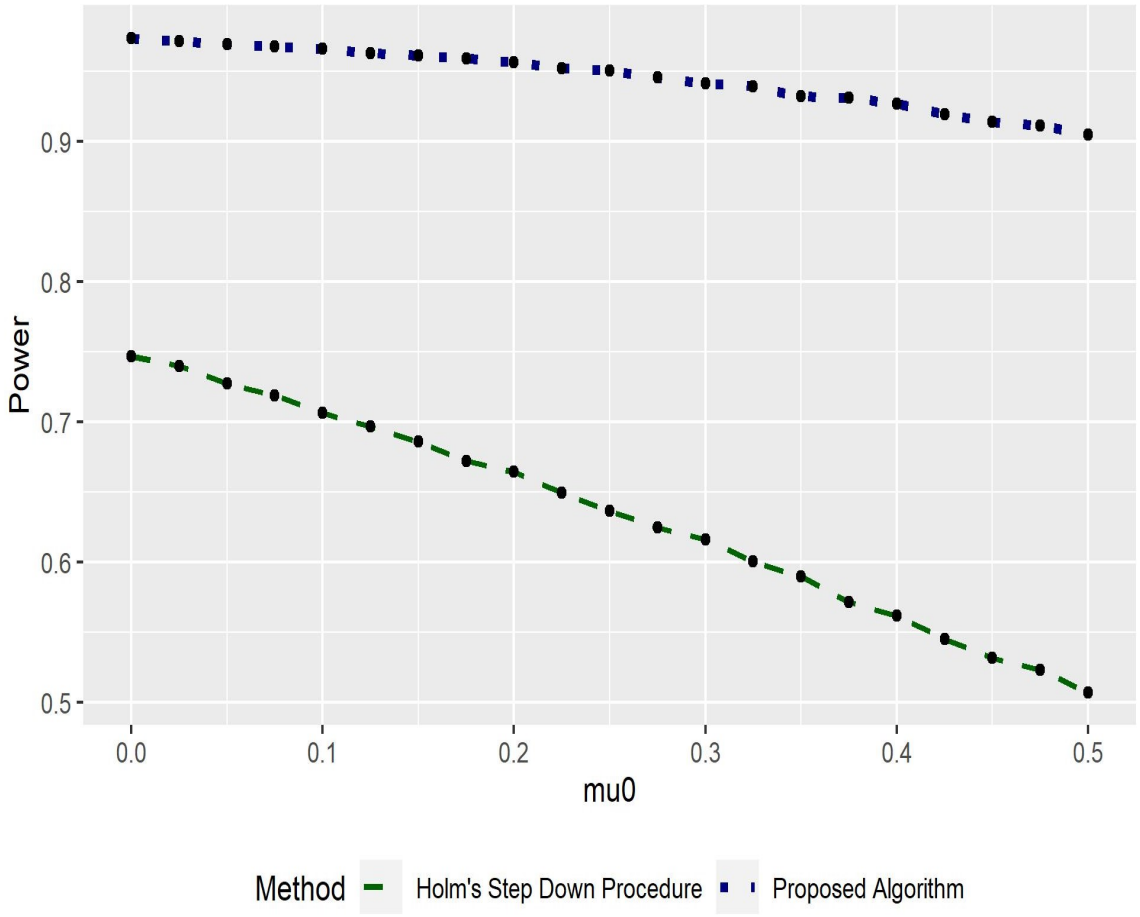

Figure 3.4 Power comparison between Holm's Step-Down Procedure and the Refined Step-Down Algorithm (proposed algorithm) using non-parametric tests with fixed alternative $\mu = 1, 1.5, 1.5, 2$

Here we are trying to analyze the trend of binge alcoholism in the past month among people aged between 12 and 34. Binge Alcoholism among youth can lead to detrimental health effects. So, in this study, alcoholism in young adults is simultaneously compared with that in teenagers and adults over 25 to identify whether binge alcoholism spikes among people aged between 18 and 25. Specifically, we are looking at the prevalence of binge alcoholism in four age groups, namely 12-13, 14-15, 16-17, and 26-34, compared to the age group 18-25. Thus null and alternative hypotheses for this particular example could be formulated as follows.

$H_{0i} : p_i = p_0 - 0.2$ vs $H_{ai} : p_i < p_0 - 0.2$ for $i = 1, 2, 3$ and $H_{04} : p_4 = p_0 + 0.2$ vs

$H_{a4} : p_4 > p_0 + 0.2$.

Where p_0 denotes the percentage of binge alcoholism in young adults, that

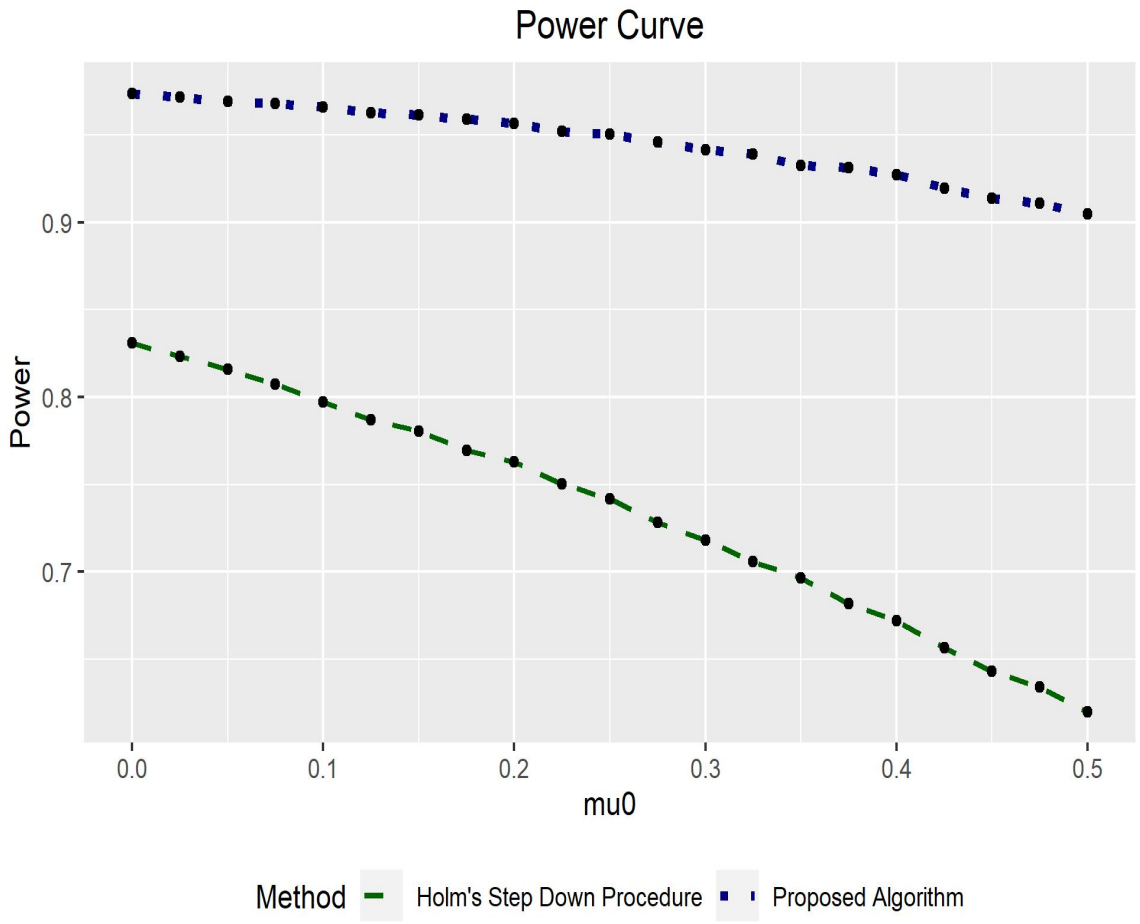

Figure 3.5 Power comparison between Holm's Step-Down Procedure and the Refined Step-Down Algorithm (proposed algorithm) using non-parametric tests with fixed alternative $\mu = 1, 1.5, 1.5, 2$ using Wilcoxon test with one-sided alternative

is, people aged between 18-25 years, and p_1, p_2, p_3 and p_4 denote the percentage of binge alcoholism in age groups 12-13, 14-15, 16-17, and 26-34, respectively. The sample size for each of the age groups is $n = 100$.

For Holm's Step Down Procedure, we compute the test statistics for four simultaneous comparisons, namely 12-13 versus 18-25, 14-15 versus 18-25, and 16-17 versus 18-25, defined next. $z = \frac{(\widehat{p}_i - \widehat{p}_0)+0.2}{\sqrt{(\widehat{p}(1-\widehat{p})(2/200))}}$ where $\widehat{p} = \frac{(\widehat{p}_i \times 100 + \widehat{p}_0 \times 100)}{200}$ for $i = 1, 2, 3$. And for 26-34 versus 18-25, the test statistic is $z = \frac{(\widehat{p}_4 - \widehat{p}_0)-0.2}{\sqrt{(\widehat{p}(1-\widehat{p})(2/200))}}$ where $\widehat{p} = \frac{(\widehat{p}_4 \times 100 + \widehat{p}_0 \times 100)}{200}$. Now, $z \overset{H_0}{\sim} N(0, 1)$. So, for Holm's procedure, we can compute the tail probabilities accordingly. For the Refined Step Down Algorithm, we sort the observed z-test statistics w.r.t. the ascending order of the absolute

differences between $\widehat{p}_i$s and $\widehat{p}_0$. Then the test statistic is computed as $\sum_{i=1}^{k} z_i^2$ for k=4,3,2,1. As the observations are independent when the null hypothesis is true, the test statistics used at every step for the new method follow a χ^2 distribution with degrees of freedom 4, 3, 2, and 1, respectively.

Table 3.6 Conclusion Based on Holm's Step-Down Procedure for the Binge Alcoholism Data

Age Groups	$p-value$	step-wise α	Rejected	Rejected by the Refined Step Down Algorithm
12-13 vs. 18-25	0.005	0.0125	YES	YES
14-15 vs. 18-25	0.022	0.0167	NO	YES
16-17 vs. 18-25	0.277	0.0250	NO	YES
26-34 vs. 18-25	0.994	0.0500	NO	YES

In this example, Holm's Step Down Procedure can detect that binge alcoholism in only one group, namely the young teens (aged 12-13 years), significantly differs from that of young adults (aged 18-25 years). However, it would be highly alarming if there is no difference in binge alcoholism between teens and young adults, as the legal drinking age in most states is 21 years. Thus, in this case, we cannot draw meaningful conclusions based on the output of Holm's step-down procedure. The reason for such an erroneous conclusion is the punitive nature of the step-down approach. As can be noted from table 3.6 for the comparison of the rate of binge alcoholism between the age groups of 14-15 and 18-25, we get a p-value of 0.022. It can be argued that the small p-value provides compelling evidence that the null hypothesis of binge alcoholism being the same in the age groups of 14-15 and 18-25 is false. However, Holm's step-down approach fails to reject the null hypothesis as we compare the p-value with the step-wise significance level of $\alpha = 0.0167$.

On the other hand, the Refined Step-Down Algorithm can identify significant differences between all groups and the baseline group of young adults regarding binge alcoholism. In this case, the finding made sense as we see a gradual increase in binge alcoholism amongst the youth. We would generally expect people between 12-18 years to not have access to alcoholic beverages as college-going young adults. Moreover, the legal drinking age in most states is 21, which falls within the age group 18-25. In addition to that, the Refined Step-Down Algorithm gives an

essential insight that between 18-34 years, the rate of binge alcoholism might steadily increase, which makes sense as most of us pursue higher studies, try to find a stable career, try to buy a home, etc. between the age of 18-34 years which can be pretty stressful sometimes and might lead to an unhealthy lifestyle choice.

CHAPTER 4 LARGE SCALE STATISTICAL ANALYSIS

4.1 Brief Description of the Data Set in Use

The data set under consideration concerns a leukemia microarray study by Golub et al. (1999) that has been used in Efron (2010) for motivation and illustration. I have used the same to present a comparative study between Holm's step-down procedure and the Refined Step-Down Algorithm. There are two disease categories: acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML). There are 47 patients under ALL and 25 patients under AML. The goal is to investigate the difference between the gene expressions of ALL and AML for each of the $N(=7128)$ genes simultaneously. Efron (2010), used this data set for large-scale statistical estimation using the False Discovery Rate. Since the false discovery rate does not control the error of incorrectly rejecting non-difference genes between ALL and AML patients, which means targeting a wrong gene location for the difference, we use the Golub et al. (1999) data set for large-scale simultaneous hypothesis testing that controls the family-wise error rate ($FWER$) at a given significance level $\alpha = 0.05$.

A two-sample t-statistic is calculated for each of the 7128 genes, assuming equal variances. This is based on 47 observations of gene expressions for ALL and 25 observations of AML. Thus we end up having 7128 values of a random variable that follows a Student's $t-$ distribution. Under H_0, it is assumed that there is no significant difference in the gene expressions between ALL and AML. Now, for each gene, we have 72 patients. So the degrees of freedom, in this case, is 72-2=70. Then the observed two sample t-statistics are transformed into $z-$ values by probability integral transformation.$z_i = \Phi^1(F_{70}(t_i))$, $i = 1, 2, ..., N$ where Φ is the cdf of standard normal distribution and F_{70} is the cdf of Student's-t distribution with 70 degrees of freedom.

Under two scenarios, I employed the Refined Step-Down Algorithm on the Golub et al. (1999) microarray data set and compared its performance against Holm's Step Down Procedure for simultaneous hypothesis testing. First, by utilizing the fact that when the null hypotheses are true, the z_i follows a normal distribution with mean 0 and variance 1 for $i = 1, 2, ...7128$. In the

second scenario, we disregard the normality assumption. 4.1 shows how the z-values mostly clustered around zero.

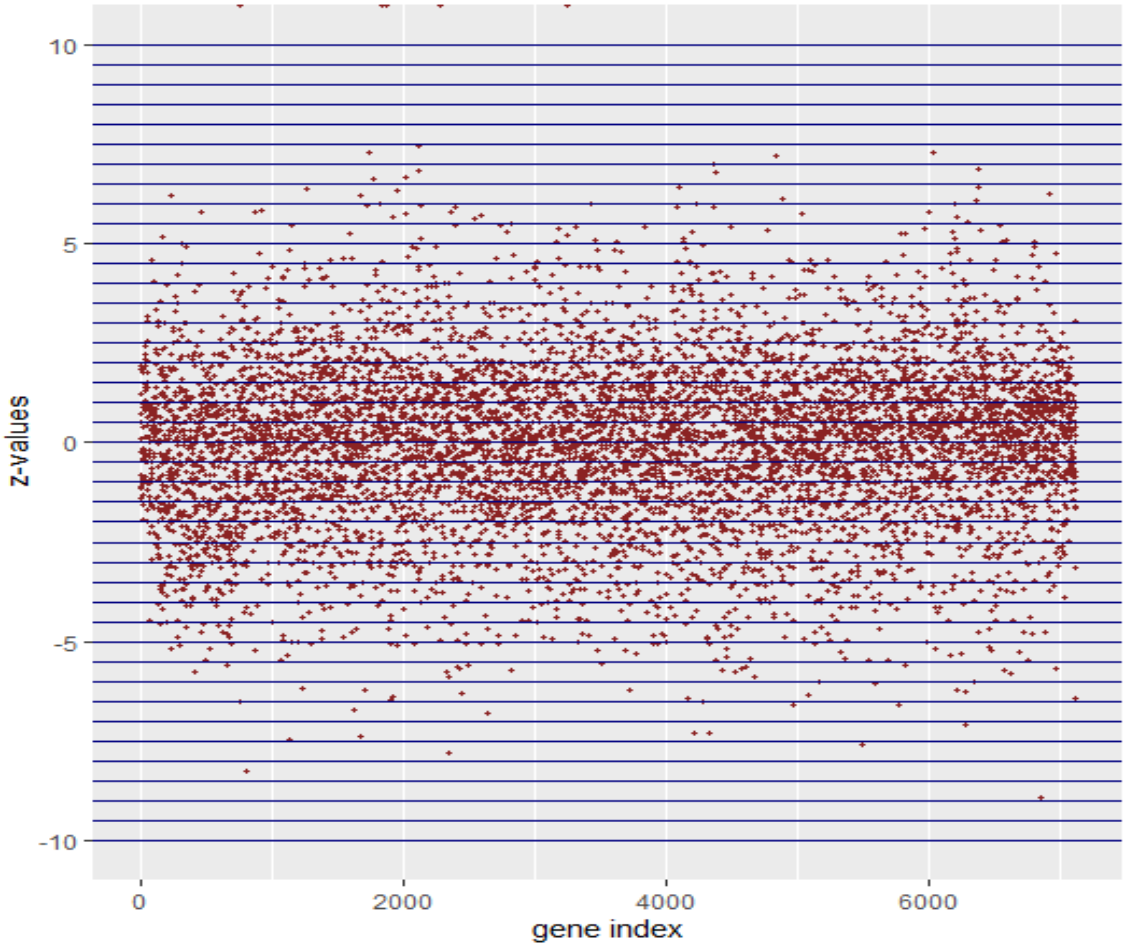

Figure 4.1 z-values against gene index

4.2 Resampling Assuming Normality

The null and alternative hypotheses get transformed to $H_{0i}: z_i \sim N(0, 1)$ vs $H_{ai}: z_i \nsim N(0, 1)$ for $i = 1, 2, ..., 7128$. And we want to perform simultaneous hypotheses testing using all of these 7128 scenarios for which the $FWER$ is controlled at the significance level α. This section covers the implementation of Holm's Step Down Procedure and the Refined Step-Down Algorithm, assuming normality.

At first, Holm's Step Down Procedure is employed. We compute $p_i = 2 \times \mathbb{P}(Z > |z_i|)$ where $Z \sim N(0, 1)$ using the observed z_i's, $i = 1, 2, ..., 7128$. Suppose, $p_{(i)}$ denotes the i^{th}

ordered p-value.

Then if $p_{(1)} < \alpha/7128$, we reject $H_{(1)}$ and proceed to check whether $p_{(2)} < \alpha/7127$ if true, we reject $H_{(2)}$ and proceed to check whether

$p_{(3)} < \alpha/7126$ ...

$\vdots$

$p_{(k)} < \alpha/(7128 - k + 1)$if true, we reject $H_{(k)}$ and proceed to check whether

$p_{(k+1)} < \alpha/(7128 - (k + 1) + 1) = \alpha/(m - k)$...

$\vdots$

$p_{(7128)} < \alpha$

Therefore, for $1 \leq k \leq 7128$, we sequentially check whether $p_{(k)} < \alpha/(7128 - k + 1)$; if true, we reject the corresponding hypothesis and proceed to the next step; else we stop and terminate the algorithm.

Holm's step-down algorithm identified 315 genes out of the 7128 to be significant. In other words, for 4.42% of the genes, there are substantial differences in the gene expressions between the two disease groups AML and ALL. This is due to the over-punitive nature of Holm's Step Down Procedure. Table 4.1 shows the top 10 significant genes that Holm's step-down algorithm failed to identify. This phenomenon is more evident when we run Holm's algorithm without any underlying distributional assumption.

Table 4.1 Output of Holm's Procedure Assuming Normality (Top 10 Unidentified Significant Genes are Reported)

Gene Index	$z - value$	$p - value$	step-wise α	Rejected	Rejected by Procedure C
1113	4.332	7.38E-06	7.34E-06	NO	YES
3504	4.325	7.63E-06	7.34E-06	NO	YES
833	-4.319	7.84E-06	7.34E-06	NO	YES
4831	-4.312	8.10E-06	7.34E-06	NO	YES
5931	4.307	8.28E-06	7.34E-06	NO	YES
5122	4.302	8.47E-06	7.34E-06	NO	YES
2945	4.301	8.52E-06	7.35E-06	NO	YES
4925	-4.297	8.64E-06	7.35E-06	NO	YES
794	-4.286	9.08E-06	7.35E-06	NO	YES
3692	4.286	9.09E-06	7.35E-06	NO	YES

Procedure C below implements the Refined Step-Down Algorithm with the z-scores computed using the Golub et al. (1999) data set. The critical task here is determining the genes that significantly differ between the two groups, ALL and AML. Under the null hypotheses, we assume no particular gene is significant in detecting differences between the two disease groups, namely ALL and AML. Therefore, assuming all the null hypotheses are true, it makes sense to consider the 7128 z-values as realized values from a standard normal distribution. Rejecting the null hypothesis will indicate that the pool of z-values is not generated from normal a variate with mean 0 and standard deviation 1. In other words, we can infer the presence of extreme cases from rejecting the null hypothesis. Therefore, we remove the extreme observations sequentially until the algorithm terminates.

Precedure C:

- **Step-1:** Take a random sample of size $B = 2000$ from the collection of 7128 z-scores with replacement
- **Step-2:** Compute the mean score based on the bootstrapped sample $\bar{X}_B$
- **Step-3:** Repeat the previous two steps N=10,000 times to get $\bar{X}_{B_1}, \bar{X}_{B_2}, ..., \bar{X}_{B_{10,000}}$
- **Step-4:** Compute the test statistic as the mean of N sample means, $\bar{X} = \frac{1}{N}\sum_{i=1}^{N} \bar{X}_{B_i}$
- **Step-5:** $p - value = 2 \times \mathbb{P}(Z > |\bar{X}|)$ where $Z \sim \mathbf{N}(0, 1/\sqrt{B.N})$
- **Step-6:** If $p - value < \alpha$ exclude the smallest 5 and the largest 5 $z-$values. Then repeat steps 1 to 6. If $p - value > \alpha$, then stop
- **Step-7:** Declare the genes corresponding to the excluded $z-$values as the significant ones

Proposition 4.1. *Procedure C strongly controls $FWER$.*

Proof. Let $K_0 \subseteq \{1, ..., 7128\}$ be an arbitrary subset such that $\mu_i = 0, i \in K_0$. Suppose $|K_0| = n_0$.

Say, $K_0 = \{t_1, ..., t_{n_0}\}$ and without loss of generality suppose $t_1 < t_2 < ... < t_{n_0}$. As we arrange the observed z_i's in ascending order, we are considering $H_0 : \mu_0 = \mu_{t_1} = \mu_{t_2} = ... = \mu_{t_{n_0}} = 0 \; vs \; H_a : \mu_i \neq 0$ for at least one $i \in K_0$. Now, rejecting at least one $H_i,\ i \in K_0$ only happens when the test statistic involving that H_i belongs to the corresponding rejection region. By construction, the test statistic T, defined in procedure C, follows a normal distribution with mean 0 and variance 1/(2000*10,000). Thus, in this case, the rejection is well-defined as $R =$ for every step of the newly proposed rejective algorithm as $R = \{T : |T| > Z_{\alpha/2}\}$ where $Z \sim N(0, 1/\sqrt{(2000 * 10,000)})$. We must remember that after every step where the algorithm rejects the null hypothesis, the initial pool of z-values got reduced by 10. Thus, after every step of rejection, the test statistic represents fewer hypotheses in the subsequent step.

FWER

$= \mathbb{P}(\text{Reject at least one } H_i, i \ \in \{1, 2, ..., 7128\} | H_i,\ i \in K_0)$

$= \mathbb{P}(T_{K-n_0+1} \in R)$

$= 2 \times \mathbb{P}(Z > Z_{\alpha/2}), \quad Z \sim N(0, 1/\sqrt{(2000 * 10,000)})$

$\leq \alpha$

□

The Refined Step-Down Algorithm identifies 980 genes out of the 7128 to be significant. This means around 13.75% of the genes are worthy of further exploration. As highlighted in section 3.3, we can see that the Refined Step-Down Algorithm is statistically more powerful than Holm's step-down procedure since Holm's method is based on the Bonferroni upper bound if Holm's procedure rejects a null hypothesis, it means that the joint probability will be smaller than α/k for $k = K, (K-1), ..., 1$, thus by controlling the exact value of the simultaneous probability for the occurrence of those events, we reject the hypothesis that is rejected by the Holm's procedure. Therefore, the Refined Step-Down Algorithm is expected to detect more non-null

cases pragmatically.

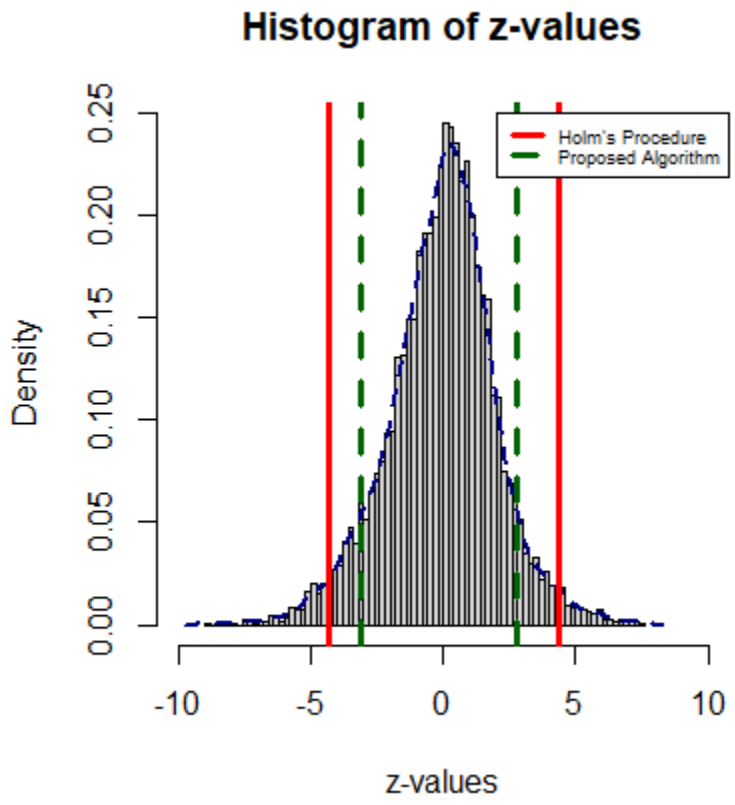

Figure 4.2 Bounds by the Holm's Procedure and by the Refined Step-Down Algorithm (proposed algorithm)

Figure 4.2 highlights the lower tail and upper tail boundaries detected by both Holm's step-down procedure and the Refined Step-Down Algorithm.

4.3 Bootstrapping without Normality

The main objective of the algorithm is to identify genes that exhibit a substantial disparity between AML and ALL. As a result, the null and alternative hypotheses can be expressed in the following manner.

$H_{0i}: \mu_{ALL_i} = \mu_{AML_i}$ versus $H_{ai}: \mu_{ALL_i} \neq \mu_{AML_i}$ for $i = 1, 2, ..., 7128$. Where μ_{ALL_i} and μ_{AML_i} denote the mean of the random variable of gene expression corresponding to the i^{th} gene for ALL and AML, respectively. Under null, we assume no significant gene detects the

difference between AML and ALL. Thus we can treat the test statistics computed under the null hypothesis as observations from a random variable centered around zero. It is important to note that when the test statistics show extreme observations, it's improbable that the corresponding null hypotheses are true.

The step-wise procedures are conducted in this section without converting the two-sample t-test statistics into a standard normal variate. First, I will compute two-sample t statistics for each of the 7128 genes by assuming unequal variances. The computation will be based on 47 gene expressions for ALL and 25 gene expressions for AML. Then I calculate the bootstrapped p-values corresponding to each of the 7128 test statistics under consideration using the observed two samples t-test statistics for Holm's procedure as below.

To obtain the bootstrapped p-values, we will apply the following procedures corresponding to each of the 7128 observed values of the test statistic.

- **Step 1:** Take a random sample of size $n = 7128$ with replacement from the 7128 observed values of the test statistic $x_1^*, ..., x_{7128}^*$
- **Step 2:** Compute $prob = \frac{\#\{|x_i^*| > |teststatistic_{observed}|\}}{7128}$
- **Step 3:** Repeat step 1 and step 2, $N = 10,000$ times and obtain $prob_1, ..., prob_{10,000}$
- **Step 4:** Compute the bootstrapped $p - value = \frac{\sum_{i=1}^{N} prob_i}{N}$

To construct a confidence interval for the Refined Step-Down Algorithm, we used bootstrapping and obtained a 95% level of confidence. Here is the process for building a 95% bootstrapped confidence interval for the Refined Step-Down Algorithm, which we will refer to as Procedure D.

Precedure D:

- **Step 1:** Take a random sample of size $n = 7128$ with replacement from the 7128 observed values of the test statistic $x_1^*, ..., x_{7128}^*$

- **Step 2:** Construct a 95% confidence interval, say given by (C_L, C_U)
- **Step 3:** Repeat step 1 and step 2, $N = 10,000$ times and obtain $(C_{L_1}, C_{U_1}), ..., (C_{L_{10,000}}, C_{U_{10,000}})$
- **Step 4:** Compute the bootstrapped confidence interval (C_L^*, C_U^*) where $C_L^* = \frac{\sum_{i=1}^{N} C_{L_i}}{N}$ and $C_U^* = \frac{\sum_{i=1}^{N} C_{U_i}}{N}$

Proposition 4.2. *Procedure D strongly controls $FWER$.*

Proof. In Procedure D, the observed test statistics are treated as a sample from the underlying distribution of the test statistic of non-significant genes under the null hypotheses. In this scenario, we can declare a significant gene if it corresponds to a test statistic value outside the bootstrapped 95% confidence interval $(-4.265061, 4.726258)$. Therefore, in this case

FWER

$= \mathbb{P}(\text{Reject at least one } H_i, i \in \{1, 2, ..., 7128\} | H_i, \ i \in K_0)$ where $K_0$is the collection of true null hypotheses

$= \mathbb{P}(T \notin (C_L^*, C_U^*))$

$= 1 - \mathbb{P}(T \in (C_L^*, C_U^*))$

$= 1 - 0.95$

$= 0.05$

$= \alpha$

□

Thus, we can achieve an exact procedure for conducting simultaneous hypothesis testing for the large-scale gene study under consideration using the bootstrapped confidence intervals.

Figure 4.3 shows the tail boundaries detected by the Refined Step-Down Algorithm. Using the bootstrapped 95% confidence interval $(-4.265061, 4.726258)$, 355 out of the 7128 genes are labeled significant. Thus, noteworthy differences between the two disease groups, AML and ALL, are recognized in around 4.98% of the genes.

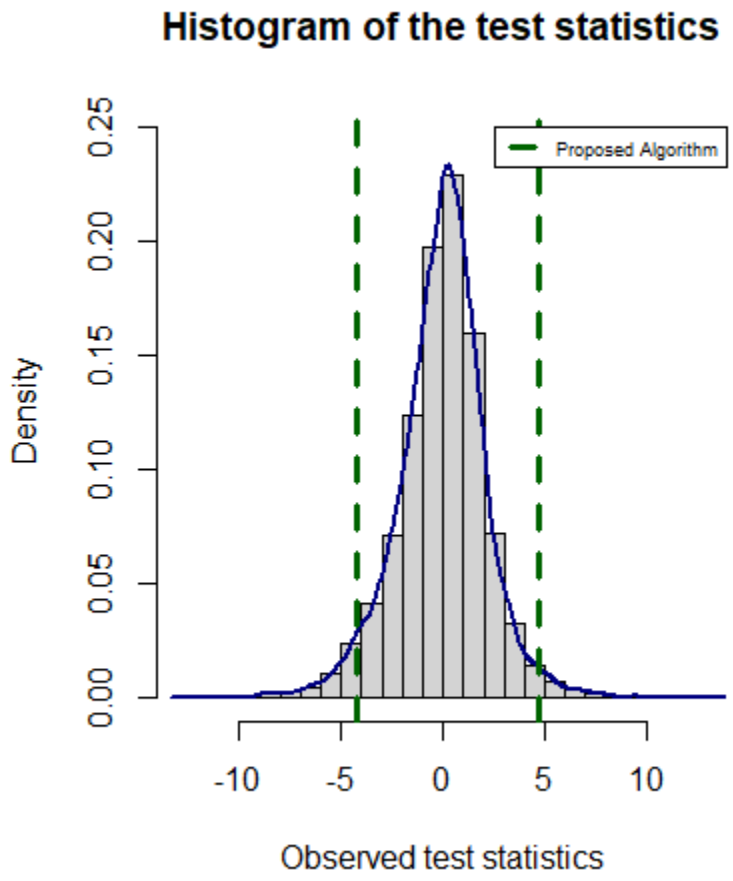

Figure 4.3 Bounds by the Refined Step-Down Algorithm (proposed algorithm); se($C^*_{0.025}$)=0.13, se($C^{0.025}$)=0 .11

Without the normality assumption, Holm's step-down procedure fails to highlight anyone out of the 7128 genes as significant in having notable differences between the AML and ALL disease groups. The bootstrapped p-values are always more significant than the corresponding step-wise threshold of the significance level ($\alpha = 0.05/k$, $k = 7128, 7127, ..., 1$).

Table 4.2 shows the observed test statistics corresponding to the smallest five bootstrapped p-values along with their respective step-wise thresholds as an output to Holm's step-down procedure. As can be noted from table 4.2, the step-wise thresholds are too small to reject even those hypotheses with an observed test statistic as significant as 12.558 or as small as -12.985. This is due to the over-corrective nature of Holm's step-down procedure while conducting

Table 4.2 Output of Holm's Procedure using Bootstrapped p-values (Top 5 Significant Genes are Reported)

Gene Index	test-statistic	$p - value$ (se)	step-wise α	Rejected	Rejected by Procedure D
3252	-12.985	1.38E-04 (1E-04)	7.01E-06	NO	YES
6854	12.558	2.84E-04 (2E-04)	7.02E-06	NO	YES
1882	-12.383	4.24E-04 (2E-04)	7.02E-06	NO	YES
4847	-11.591	5.58E-04 (3E-04)	7.02E-06	NO	YES
1834	-11.249	7.02E-04 (3E-04)	7.02E-06	NO	YES

large-scale simultaneous hypotheses testing. On the contrary, the new method can highlight 355 genes as necessary for further explorations in identifying crucial differentiating factors between the two disease groups under study, namely AML and ALL.

Through this example, we can showcase the effectiveness of the recently introduced algorithm for extensive statistical data analysis, specifically in testing multiple hypotheses simultaneously. In addition, this new approach can be considered an enhancement to the current method that utilizes both Holm's step-down algorithm and Efron's empirical FDR approach for estimation.

CHAPTER 5 CONCLUSION AND FUTURE DIRECTIONS

5.1 Conclusion

The primary objective of this dissertation was to develop a reliable approach for performing simultaneous hypothesis testing that can strongly control the family-wise error rate ($FWER$). The driving force behind this was the inadequacy of Holm's step-down procedure, which can strongly control the $FWER$ but lacks statistical power in large-scale data analysis due to its strict step-down approach. However, it is essential to note that Hochberg's step-up procedure is more robust than Holm's step-down procedure. Despite this, it may not be as effective in controlling the Family-Wise Error Rate ($FWER$). In large-scale data analysis with significant correlation, Efron (2010) provides a reliable method to determine the empirical False Discovery Rate (FDR). However, relying on the empirical FDR with a q threshold (Benjamini and Hochberg (1995)) of 0.05 for hypothesis testing was found to be insufficient. After conducting thorough research, it was discovered that only 22 out of the 7128 genes held significance when using the $q = 0.05$ (Benjamini and Hochberg (1995)) on the empirical FDR. However, this approach left numerous extreme values of the test statistics unexplored and required further investigation. Furthermore, it's important to note that controlling the FDR is only possible when all the null hypotheses being considered are true, which is a significant constraint. One of the primary obstacles involved finding a way to handle the relationship between test statistics for various hypotheses while simultaneously devising a new methodology capable of confidently addressing the problem of family-wise error rate (FWER). After reviewing the existing literature on correlated simultaneous hypotheses testing methods, I started exploring procedures for conducting multiple hypotheses testing related to correlated contingency tables in Chapter 2.

When testing multiple odds ratios in Chapter 2, the step-up algorithm was more robust than the step-down algorithm. A step-wise confidence procedure was deemed valid, while step-wise methods were inconclusive. Moreover, the step-wise confidence procedure strongly controls the family-wise error rate. The simultaneous confidence interval based on the bootstrapped distribution of odds ratios was found to be a prudent way of drawing meaningful

inferences while testing multiple hypotheses involving odds ratios. However, its time complexity is too high compared to the existing step-wise procedures. Based on my understanding of the analyses presented in Chapter 2, a novel methodology to conduct simultaneous hypothesis testing was developed and discussed in the subsequent chapters. The Refined Step-Down Algorithm is less punitive compared to Holm's step-down procedure. Additionally, it can ensure strong control over the $FWER$.

The third chapter introduces a meticulous method for testing multiple hypotheses. The effectiveness of the newly developed algorithm was demonstrated under both the parametric and non-parametric frameworks. After a comprehensive analysis of statistical powers using simulation, it has been determined that the new method is more effective than Holm's step-down algorithm. In addition to the above, I have analyzed a real-life data set from cdc.gov on binge alcoholism among the youth and demonstrated the effectiveness of the Refined Step-Down Algorithm compared to Holm's step-down procedure. Next, I broadened my knowledge of analyzing a substantially correlated data set using the latest technique.

Golub et al. (1999) data on 7128 gene expressions with 25 acute myeloid leukemia (AML) patients and 47 acute lymphoblastic leukemia (ALL) patients is thoroughly analyzed in Chapter 4. This study aimed to identify any significant genes that show contrasting gene expressions between AML and ALL patients. Compared to Holm's step-down procedure, a comparative study was conducted to showcase the efficiency of the recently proposed algorithm for large-scale multiple hypotheses testing. The proposed exact confidence procedure was proven more effective when conducting simultaneous hypothesis testing for large-scale data analysis.

In summary, this dissertation accomplished its objective of creating a reliable methodology that exhibits greater statistical power than Holm's procedure while effectively managing the Family-Wise Error Rate ($FWER$). In addition, this research has effectively confirmed the suitability of the new confidence approach of simultaneous hypothesis testing across a range of scientific studies and frameworks. The following section delves into the significant research extensions that intrigued me and that I plan to explore in the near future.

5.2 Future Work

In the future, I plan to investigate the applicability of rms correlation in scenarios where gene expressions are not derived from correlated normal variables but rather from a correlated Cauchy distribution. Since Cauchy distributions lack both mean and variance, exploring the usefulness of the recently proposed methodology in such cases would be an intriguing exercise.

Another critical step is to conduct a sensitivity analysis by altering the correlation value from -1 to +1 to gauge the methodology's dependence on correlation α. Furthermore, the computational time efficiency can be assessed by analyzing a range of N (e.g., number of genes) values and varying values of n_1 (number of patients in ALL) and n_2 (number of patients in AML).

The current methodology focuses on the specific scenario when we want to test the plausibility of a particular claim for K different strata. So, the types of underlying hypotheses are the same. For example, when we want to check whether all income groups' mean expenditure is the same. An exciting task would be extending this to a generic situation where the type of underlying hypotheses might differ. Some may be parametric some may be non-parametric. For example, one might be interested in simultaneously testing different attributes of an individual or subject of an experiment. Some features can be quantifiable (age, height, weight, blood pressure, etc.), and some can be qualitative (blood group, lifestyle, smoking habits, etc.).

- **Simultaneous Confidence Set:** Chen (2016) discusses an approach for constructing simultaneous confidence regions for multiple hypothesis testing. In other words, one can build a k-dimensional rejection region considering all hypotheses simultaneously. A simultaneous rejection is declared when the data in hand fall into the aforementioned rejection region. Here, one is constructing the rejection region based on the joint distribution of the test statistics corresponding to all the k hypotheses under scrutiny. As mentioned in theorem 1.1 for multiple testing problems $H_{i0} : \theta_i \in \Theta_i$ versus $H_{i1} : \theta_i \in \Theta_i^c \; i = 1, 2, ..., k$ there exists inverted confidence set that is directed towards θ_i^c such that the family-wise error rate can be strongly controlled at the given significance level.

To construct such a simultaneous confidence set based on the joint probability distribution of k test statistics, one must accurately capture the pairwise correlation structure of the test statistics. By utilizing the joint distribution of the test statistics, it is possible to accurately establish the confidence region for evaluating multiple hypotheses simultaneously.

- **Estimation of the unknown correlation:** As highlighted in Appendix D (section .0.4), a robust way of estimating the pairwise correlation for the gene expression data set is defined in Efron (2010). It essentially computes the value of the root mean square correlation based on the distribution of the observed pairwise correlations. However, it would be helpful to develop an algorithm for estimating the complete variance-covariance matrix corresponding to the joint distribution of the test statistics. Once the entire correlation structure of the test statistics corresponding to all k hypotheses is captured, one can pragmatically define the simultaneous rejection region.

Apart from those mentioned above, it is possible to examine the effect of altering means and standard deviations individually while keeping the correlation fixed on the methodology. It is worth noting that the methodology's efficacy can be observed in a model with a low correlation value and considerably high standard deviations.

BIBLIOGRAPHY

Benjamini, Y. and Y. Hochberg (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.

Casella, G. and R. L. Berger (2021). *Statistical inference*. Cengage Learning.

Chen, J. T. (2016). A nonparametric coherent confidence procedure. *Communications in Statistics-Theory and Methods 45*(11), 3397–3409.

Efron, B. (2009). Are a set of microarrays independent of each other? *The annals of applied statistics 3*(3), 922.

Efron, B. (2010). Correlated z-values and the accuracy of large-scale statistical estimates. *Journal of the American Statistical Association 105*(491), 1042–1055.

Golub, T. R., D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, et al. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *science 286*(5439), 531–537.

Hochberg, Y. (1988). A sharper bonferroni procedure for multiple tests of significance. *Biometrika 75*(4), 800–802.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, 65–70.

Simes, R. J. (1986). An improved bonferroni procedure for multiple tests of significance. *Biometrika 73*(3), 751–754.

APPENDIX A SELECTED R PROGRAMS FROM CHAPTER 2

- This code is used to implement the step-wise procedures using the χ^2 test in Chapter 2.

```
step.alpha<-alpha/seq(4,1,-1)
gr.names<-c("gr1","gr2","gr3","gr4")
chi.pval<-sapply(TBL.list,chi_pval)
df.chi<-data.frame(cbind(gr.names,chi.pval))
df.chi<-df.chi[order(as.numeric(chi.pval),decreasing = FALSE),]
df.chi$alpha<-as.numeric(step.alpha)
df.chi$chi.pval<-as.numeric(df.chi$chi.pval)
m=length(gr.names)
i=1
while (i<=m){
        if (df.chi$chi.pval[i]<df.chi$alpha[i]){
                df.chi$rej[i]<-TRUE
                }else{
                        break
                }
        i=i+1
}
if (i<m){
        df.chi$rej[i:m]<-FALSE
        }
df.chi
```

- The function Boot.rej defines the rejection criterion based on the bootstrapped confidence interval in Chapter 2.

```
Boot.rej<-function(alpha=0.05){
        r1.gr1 = r2.gr1 = 100;  p1.gr1 = .9;  p2.gr1 = .5
```

```
        x1.gr1<-rbinom(r1.gr1, 1, p1.gr1)
        x2.gr1<-rbinom(r2.gr1, 1, p2.gr1)

        r1.gr2 = r2.gr2 = 200;  p1.gr2 = .5;  p2.gr2 = .5
        x1.gr2<-rbinom(r1.gr2, 1, p1.gr2)
        x2.gr2<-rbinom(r2.gr2, 1, p2.gr2)

        r1.gr3 = r2.gr3 = 150;  p1.gr3 = .5;  p2.gr3 = .5
        x1.gr3<-rbinom(r1.gr3, 1, p1.gr3)
        x2.gr3<-rbinom(r2.gr3, 1, p2.gr3)

        r1.gr4 = r2.gr4 = 50;  p1.gr4 = .8;  p2.gr4 = .5
        x1.gr4<-rbinom(r1.gr4, 1, p1.gr4)
        x2.gr4<-rbinom(r2.gr4, 1, p2.gr4)

        ### Defining Bootstrap rejection region
        nsim<-10000
        OR.boot<-replicate(nsim,stat.fun(x1.gr1,x2.gr1,x1.gr2,x2.gr2,
                                         x1.gr3,x2.gr3,x1.gr4,x2.gr4))
        m<-nrow(OR.boot)
        CI.boot<-apply(OR.boot,1,quantile,
                       probs=c(alpha/(2*m),1-alpha/(2*m)))

        rej<-1-c(CI.boot[1,1]<1 && 1<CI.boot[2,1],
                 CI.boot[1,4]<1 && 1<CI.boot[2,4])

        rej
}
```

APPENDIX B SELECTED R PROGRAMS FROM CHAPTER 3

- This code is used to implement the newly proposed algorithm for simultaneous hypotheses testing with contingency tables in Chapter 3.

```
while(m>0){
        dat=props.list[names(sorted_stats)[1:m]]
        teststat<-sum(sapply(dat,z.sq))
        p.value<-pchisq(teststat,df=m,lower.tail = FALSE)
        if(p.value<alpha){
                rej_gr<-c(rej_gr,names(sorted_stats[m]))
        }else{
                break
        }
        m<-m-1
}
```

- This code is used for implementing the newly proposed algorithms in the non-parametric paradigm.

```
gr.list.jt<-list(x0,x1,x2,x3,x4)
        names(gr.list.jt)<-c("gr0","gr1","gr2","gr3","gr4")
        medians<-sapply(gr.list.jt,median)

        sorted_stats<-medians[order(medians,decreasing = FALSE)]

        m<-length(medians)
        rej_gr<-c()

        while(m>=2){
                dat=gr.list.jt[names(sorted_stats)[1:m]]
```

```
                group<-c()
                for (k in seq(1,length(dat),1)){
                        group<-c(group,rep(k,n))
                }
                space<-c()
                for (k in seq(1,length(dat),1)){
                        space<-c(space,dat[[k]])
                }
                test_jt<-jonckheere.test(space,group,
                                alternative="increasing")
                if(test_jt$p.value<alpha){
                        rej_gr<-c(rej_gr,names(sorted_stats[m]))
                }else{
                        break
                }
                m<-m-1
        }

        if("gr1" %in% rej_gr | "gr2" %in% rej_gr){
                rej_gc<-1
        }else{
                rej_gc<-0
        }
```

- This code is used for generating the power curve in figure 3.3

```
N<-50000 ## Number of simulations
## Setting the sample sizes
n<-50
```

```
## Defining the level of significance
alpha<-0.05

cmh_chi_pval<-function(TBL.list){

        diff<-function(TBL){
                TBL[1,1]/sum(TBL[1,])-TBL[2,1]/sum(TBL[2,])
        }

        wt<-function(TBL){
                (sum(TBL[1,])*sum(TBL[2,]))/(sum(TBL[1,])+sum(TBL[2,]))
        }

        wtd_mn_dif<-function(TBL.list){
               w<-sapply(TBL.list,wt)
               d<-sapply(TBL.list,diff)
               sum(w*d)/sum(w)

        }
        marginals<-function(TBL){
                w<-wt(TBL)
                p_hat_mar<-sum(TBL[,1])/sum(TBL)
                w*p_hat_mar*(1-p_hat_mar)
        }

        SE<-function(TBL.list){
                w<-sapply(TBL.list,wt)
                sqrt(sum(sapply(TBL.list,marginals)))/sum(w)
        }
```

```
        w<-sapply(TBL.list,wt)
        d<-sapply(TBL.list,diff)
        stat<-sum(w*d)/(sqrt(sum(sapply(TBL.list,marginals))))

        pchisq(stat^2,df=1,lower.tail = FALSE)
}

OR.stat<-function(tab){
        (tab[1,1]*tab[2,2])/(tab[1,2]*tab[2,1])
}

# Defining Rejections for Holm's and the proposed algorithm

rej.holm.gc<-function(alpha){

        n0=n1=n2=n3=n4=n

        x0<-rbinom(n0,1,prob=p0)
        x1<-rbinom(n1,1,prob=p1)
        x2<-rbinom(n2,1,prob=p2)
        x3<-rbinom(n3,1,prob=p3)
        x4<-rbinom(n4,1,prob=p4)

        r1.gr1=n1
        r2.gr1=n1
        r1.gr2=n2
        r2.gr2=n2
        r1.gr3=n3
        r2.gr3=n3
```

```
r1.gr4=n4
r2.gr4=n4

## Tab1 p0 vs p1

x1.gr1<-x0
x2.gr1<-x1
g1.gr1<-c(sum(x1.gr1), r1.gr1 - sum(x1.gr1))
g2.gr1<-c(sum(x2.gr1), r2.gr1 - sum(x2.gr1))
TBL.gr1<-rbind(g1.gr1, g2.gr1)
colnames(TBL.gr1)<-c("Disease","Control")
rownames(TBL.gr1)<-c("Unexposed","Exposure")

## Tab2 p0 vs p2

x1.gr2<-x0
x2.gr2<-x2
g1.gr2<-c(sum(x1.gr2), r1.gr2 - sum(x1.gr2))
g2.gr2<-c(sum(x2.gr2), r2.gr2 - sum(x2.gr2))
TBL.gr2<-rbind(g1.gr2, g2.gr2)
colnames(TBL.gr2)<-c("Disease","Control")
rownames(TBL.gr2)<-c("Unexposed","Exposure")

# Tab3 p0 vs p3

x1.gr3<-x0
x2.gr3<-x3
g1.gr3<-c(sum(x1.gr3), r1.gr3 - sum(x1.gr3))
g2.gr3<-c(sum(x2.gr3), r2.gr3 - sum(x2.gr3))
```

```
TBL.gr3<-rbind(g1.gr3, g2.gr3)
colnames(TBL.gr3)<-c("Disease","Control")
rownames(TBL.gr3)<-c("Unexposed","Exposure")

# Tab4 p0 vs p4

x1.gr4<-x0
x2.gr4<-x4
g1.gr4<-c(sum(x1.gr4), r1.gr4 - sum(x1.gr4))
g2.gr4<-c(sum(x2.gr4), r2.gr4 - sum(x2.gr4))
TBL.gr4<-rbind(g1.gr4, g2.gr4)
colnames(TBL.gr4)<-c("Disease","Control")
rownames(TBL.gr4)<-c("Unexposed","Exposure")

gr1<-cbind(x1.gr1,x2.gr1);gr2<-cbind(x1.gr2,x2.gr2)
gr3<-cbind(x1.gr3,x2.gr3);gr4<-cbind(x1.gr4,x2.gr4)
gr.list<-list(gr1,gr2,gr3,gr4)
TBL.list<-list(TBL.gr1,TBL.gr2,TBL.gr3,TBL.gr4)
names(TBL.list)<-c("TBL.gr1","TBL.gr2","TBL.gr3","TBL.gr4")

step.alpha<-alpha/seq(4,1,-1)
gr.names<-c("gr1","gr2","gr3","gr4")
chi.pval<-sapply(TBL.list,chi_pval)
df.chi<-data.frame(cbind(gr.names,chi.pval))
df.chi<-df.chi[order(as.numeric(chi.pval),decreasing = FALSE),]
df.chi$alpha<-as.numeric(step.alpha)
df.chi$chi.pval<-as.numeric(df.chi$chi.pval)

# Defining the rejection crtiteria for Holm's
```

```
m=length(TBL.list)
i=1
while (i<m){
        if (df.chi$chi.pval[i]<df.chi$alpha[i]){
                df.chi$rej[i]<-TRUE
        }else{
                break
        }
        i=i+1
}
if (i<m){
        df.chi$rej[i:m]<-FALSE
}
df.chi<-df.chi[df.chi$rej==1,]
if("gr1" %in% df.chi$gr.names | "gr2" %in% df.chi$gr.names |
"gr3" %in% df.chi$gr.names | "gr4" %in% df.chi$gr.names  ){
        rej_holm<-1
}else{
        rej_holm<-0
}

## Rejection for proposed algorithm

OR.stats<-sapply(TBL.list,OR.stat)
sorted_stats<-OR.stats[order(OR.stats,decreasing = FALSE)]
m<-length(OR.stats)
rej_gr<-c()

while(m>=2){
```

```
                dat=TBL.list[names(sorted_stats)[1:m]]
                # dat=unlist(TBL.list[names(sorted_stats)[1:m]])
                if(cmh_chi_pval(dat)<alpha){    # #cmh.pval(dat,m)<alpha
                        rej_gr<-c(rej_gr,names(sorted_stats[m]))
                }else{
                        break
                }
                m<-m-1
        }

        if("TBL.gr1" %in% rej_gr | "TBL.gr2" %in% rej_gr|
        "TBL.gr3" %in% rej_gr|"TBL.gr4" %in% rej_gr){
                rej_gc<-1
        }else{
                rej_gc<-0
        }

        c(rej_holm,rej_gc)
}

## Computing power for Holm's and Proposed Algorithm
pow.holm.gc<-function(alpha=0.05){
        rp<-replicate(N,rej.holm.gc(alpha))
        apply(rp, 1,mean)
}

# Power curve ## fixed p-alts 0.3,0.7
p1<-0.6;p2<-0.7
p0.seq<-seq(0.3,0.5,0.01)#c(0.4,0.5,0.6)
```

```
holm.pow.seq<-c()
gc.pow.seq<-c()
for(p0 in p0.seq){
        p3=p1;p4=p2
        res<-pow.holm.gc(alpha=0.05)
        holm.pow.seq<-c(holm.pow.seq,res[1])
        gc.pow.seq<-c(gc.pow.seq,res[2])
}

df<-data.frame(cbind(p0.seq,holm.pow.seq,gc.pow.seq))
names(df)<-c("p0","Holm's Step Down Procedure","Proposed Algorithm")
library(reshape2)
library(ggplot2)
df<-melt(df,id=c("p0"))
names(df)<-c("p0","Method","Power")
p<-ggplot(df,aes(x=p0,y=Power,group=Method))+
        geom_line(aes(linetype=Method, color=Method, size=Method))+
        geom_point()+
        scale_linetype_manual(values=c("dashed", "dotted"))+
        scale_color_manual(values=c('darkgreen','navyblue'))+
        scale_size_manual(values=c(1, 1.5))+
        theme(legend.position="bottom")+
        ggtitle("Power Curve")+
        theme(plot.title = element_text(hjust = 0.5))

p
}
```

- This code is used for section 3.5

```
props.gr1<-c(0.343,0.005)
props.gr2<-c(0.343,0.032)
props.gr3<-c(0.343,0.108)
props.gr4<-c(0.343,0.374)
# Applying Holm's Algorithm
## Defining the level of significance
alpha<-0.05
thres<-0.2
pval.left<-function(props,n=100){
        p<-(props[1]*100+props[2]*100)/(2*n)
        z<-((props[2]-props[1])+thres)/(sqrt(p*(1-p[1])*(1/n+1/n)))
        pnorm(z,lower.tail=TRUE)
}
pval.right<-function(props,n=100){
        p<-(props[1]*100+props[2]*100)/(2*n)
        z<-((props[2]-props[1])-thres)/(sqrt(p*(1-p[1])*(1/n+1/n)))
        pnorm(z,lower.tail=FALSE)
}
props.list<-list(props.gr1,props.gr2,props.gr3)
step.alpha<-alpha/seq(4,1,-1)
gr.names<-c("gr1","gr2","gr3","gr4")
pval<-sapply(props.list,pval.left)
pval<-c(pval,pval.right(props.gr4))

df<-data.frame(cbind(gr.names,pval))
df<-df[order(as.numeric(pval),decreasing = FALSE),]
df$alpha<-as.numeric(step.alpha)
df$pval<-as.numeric(df$pval)
# Defining the rejection crtiteria for Holm's
```

```
m=4
i=1
while (i<m){
        if (df$pval[i]<df$alpha[i]){
                df$rej[i]<-TRUE
        }else{
                break
        }
        i=i+1
}
if (i<m){
        df$rej[i:m]<-FALSE
}
map_names<-data.frame(cbind(c("12-13","14-15","16-17","26-34"),gr.names))
colnames(map_names)<-c("Age","Group")
map_names[map_names$Group %in% df[df$rej==1,1],"Age"]

# Implementing the Proposed Algorithm
p_hats<-c(props.gr1[2],props.gr2[2],props.gr3[2],props.gr4[2])
names(p_hats)<-c("gr1","gr2","gr3","gr4")
sorted_stats<-p_hats[order(abs(p_hats-props.gr1[1]),decreasing = FALSE)]
z.left<-function(props,n=100){
        p<-(props[1]*100+props[2]*100)/(2*n)
        z<-((props[2]-props[1])+thres)/(sqrt(p*(1-p[1])*(1/n+1/n)))
        z

}
z.right<-function(props,n=100){
        p<-(props[1]*100+props[2]*100)/(2*n)
```

```
        z<-((props[2]-props[1])-thres)/(sqrt(p*(1-p[1])*(1/n+1/n)))
        z
        }
props.list<-list(props.gr1,props.gr2,props.gr3)
z_vals<-sapply(props.list, z.left)
z_vals<-c(z_vals,z.right(props.gr4))
names(z_vals)<-c("gr1","gr2","gr3","gr4")

m<-4
rej_gr<-c()
while(m>0){
        dat=z_vals[names(sorted_stats)[1:m]]
        teststat<-sum(dat^2)
        p.value<-pchisq(teststat,df=m,lower.tail = FALSE)
        if(p.value<alpha){
                rej_gr<-c(rej_gr,names(sorted_stats[m]))
        }else{
                break
        }
        m<-m-1
}
# We reject the null hypothesis for the following groups
map_names[map_names$Group %in% rej_gr,"Age"]
```

APPENDIX C SELECTED R PROGRAMS FROM CHAPTER 4

- This code implements the proposed algorithm assuming normality in Chapter 4

```
set.seed(123456)
B<-2000
N<-10000
alpha<-0.05

Bt_stats<-function(dat_vals){
        bt<-sample(x=dat_vals,size=B,replace=TRUE)
        mean(bt)
}

tail_genes<-c()
dat<-gene_df
sorted_df<-gene_df[order(z_vals,decreasing = FALSE),]
tail_sorted_df<-tail(sorted_df)
head_sorted_df<-head(sorted_df)

## Computing p-value based on the bootstrapped distribution of means

iteration=1
while(iteration<=200){
        reps<-replicate(N,Bt_stats(dat$z_vals))
        pval<-2*pnorm(abs(mean(reps)),mean=0,sd=1/(sqrt(N*B)),
        lower.tail = FALSE)
        if(pval<alpha){
                idx<-sorted_df$gene_index[c(1:5,(nrow(dat)-4):nrow(dat))]
                ## Why 5? We have largest 5 vals as inf
```

```
                dat<-dat[!(dat$gene_index %in% idx), ]
                sorted_df<-sorted_df[!(sorted_df$gene_index %in% idx),]
                tail_genes<-c(tail_genes,idx)
        }else{
                break
        }
        iteration<-iteration+1
}

tail_df<-gene_df[gene_df$gene_index %in% tail_genes,]
#number of genes rejected by proposed algorithm
nrow(tail_df)
# tail_df$gene_index #significant genes
round((nrow(tail_df)/nrow(gene_df))*100,3)#percentage
```

- This code implements the newly proposed algorithm using Bootstrapped quantiles in Chapter 4

```
two_sample_t<-function(x){
        t_test<-t.test(x[1:47],x[48:72],alternative = "two.sided",
        var.equal = FALSE)
        t_test$statistic
}

## Observed 2-sample t-statistics with unequal variance
obs_t_stats<-apply(gene_dat,1,two_sample_t)
df<-data.frame(cbind(gene_index,obs_t_stats))

B<-10000
```

```
Boot_quantile<-function(x){
        qs<-function(x){
                rm_sample<-sample(df$obs_t_stats,size=n,replace=TRUE)
                q<-quantile(rm_sample,probs=c(alpha,1-alpha))
                q
        }
        boot<-replicate(B,qs(x))
        mn<-apply(boot,1,mean)
        se<-apply(boot,1,sd)
        c(mn=mn,se=se)
}

prop_algo_threshold<-Boot_quantile(df$obs_t_stats)
tail_df<-df[which(df$obs_t_stats<prop_algo_threshold[1]|df$
obs_t_stats>prop_algo_threshold[2]),]

95% confidence interval
prop_algo_threshold[c(1,2)]

The corresponding standard Errors
prop_algo_threshold[c(3,4)]

# %tage of genes rejected by the proposed algorithm
round((nrow(tail_df)/nrow(df))*100,3)
```

APPENDIX D METHODOLOGY FOR COMPUTING EMPIRICAL FDR (Efron (2010)))

The input to the methodology in Efron (2010)) is a collection of correlated standard normal variables. Using a sample of N observations, the method appropriately determines the summary statistic for these correlated normal variates, say, the empirical cumulative distribution function. In addition to the above, the methodology also provides a reasonable estimation of the variance of $\widehat{F}$ by considering the correlation in the data. The methodology splits the correlated random variables into a finite number of classes. The z-values in the same class follow a normal distribution with an identical mean and standard deviation. These classes are characterized by different means and standard deviations. So, a finite collection of (μ_c, σ_c) denoting the mean and standard deviation of the normal distribution for the c^{th} class.

The two major constituents of Efron (2010)) are

1) Finding the distribution ($\widehat{F}$)of correlated normal variables

2) Estimating the correlation parameters in addition to the means, standard deviations of the normal distribution, and the proportion of each predefined class in the sample

The critical assumptions of the methodology described in Efron (2010)) are

i) z_i's follow normal distributions for $i = 1, 2, ..., N$ with different means and variances. $z_i \sim N(\mu_i, \sigma_i^2),\ i = 1, 2, ..., N.$

ii) z_i's might not be independent.

Efron (2010)) used a right-sided cumulative distribution function for convenience.

The author presents competent formulas for computing the mean and covariance of the process $\{\widehat{F}(x), -\infty < x < \infty\}$ in Efron (2010)). However, instead of working with the right-handed cdfs immediately, some valuable results are deduced using a discretized version of the z-values. The mean and covariance of the process $\{\widehat{F}(x), -\infty < x < \infty\}$ are deduced leveraging the results of the discretized version. Using a Poisson Spline Regression, a smooth curve for $\widehat{F}$ is obtained from the histogram of the z-values. Moreover, the methodology uses the

binning strategy to transform the correlated N-random standard normal variables into K-correlated discrete random variables. As per my understanding, $K \leq N$. As mentioned earlier, the transformation is because it is easier to deal with the $K \times K$ covariance matrix of the discrete variables than with the $N \times N$ covariance matrix of the initial correlated standard variables.

The range of the observed z_i values $\mathcal{Z}$ is partitioned into K bins namely $\mathcal{Z}_1, \mathcal{Z}_2, ..., \mathcal{Z}_K$ having an equal bin width of Δ.

$$\mathcal{Z} = \bigcup_{k=1}^{K} \mathcal{Z}_k \tag{.0.1}$$

Additionally, x_k and y_k denote the midpoint and the number of observations in $\mathcal{Z}_K$, respectively.

$$\mathrm{y}_k = \#\ \ z_i \in \mathcal{Z}_k\ \ ,\ k = 1, 2, ..., K \tag{.0.2}$$

Firstly, the mean and covariance of the vector $\mathbf{y} = (\mathrm{y}_1, \mathrm{y}_2, ..., \mathrm{y}_K)'$ are computed. Note that as $\Delta \to 0$ $\mathrm{y}_k \to 0$ or 1 with x_k values for non-void bins denoting the locations of the ordered z_i's under the assumption of no ties. The methodology deduces various properties of **z** through **y**. z_i's are grouped into a finite number of classes.μ_c and σ_c denote the mean and standard deviation of the z_i's in class C_c respectively.

$$z_i \sim N(\mu_c, \sigma_c^2) \text{ for } z_i \in C_c \tag{.0.3}$$

N_c and p_c are defined as the number of elements in class C_c and the proportion of class C_c in the entire sample respectively.

$$N_c = \#\ \ C_c \qquad and \quad p_c = N_c/N \tag{.0.4}$$

Note that $\sum_c N_c = N$ and $\sum_c p_c = 1$.

Let **x** be the vector of bin midpoints having K components. The methodology defines $\mathbf{x}_c$

as below.

$$\mathbf{x}_c = (\mathbf{x} - \mu_c)/\sigma_c = (x_{1c}, ..., x_{kc}, ..., x_{Kc})\prime \tag{.0.5}$$

and $x_{kc} = (x_k - \mu_c)/\sigma_c$.Additionally,any vector valued function h of $\mathbf{x}_c$ will be defined as below.

$$\mathbf{h}_c = (h(x_{1c}), ..., h(x_{kc}), ..., h(x_{Kc}))\prime \tag{.0.6}$$

.0.1 Expected value of the count vector **y**

Suppose, P($z_i \in \mathcal{Z}_k | z_i \in C_c$) $= \pi_{kc}$. Here, Prob$_c\{z_i \in \mathcal{Z}_k\}$ = P($z_i \in \mathcal{Z}_k | z_i \in C_c$) Thus π_{kc} is denoted by Prob$_c\{z_i \in \mathcal{Z}_k\}$ and is defined as below.

$$\pi_{kc} = \text{Prob}_c\{z_i \in \mathcal{Z}_k\} = \Delta\varphi(x_{kc})/\sigma_c \tag{.0.7}$$

Where, $\varphi(x) = exp(-x^2/2)/\sqrt{2\pi}$. Note that $\pi_c = (\pi_{1c}, ..., \pi_{kc}, ..., \pi_{Kc})$. So,
P($z_i \in \mathcal{Z}_k$) $= \sum_c p_c \pi_{kc}$ as P($z_i \in C_c$) $= p_c$ and P($z_i \in \mathcal{Z}_k | z_i \in C_c$) $= \pi_{kc}$.

Notice that when Δ is sufficiently small, for
$k = 1, ..., K$ $y_k \sim Binomial(N, p_k)$ where, p_k = P($z_i \in \mathcal{Z}_k$) $= \sum_c p_c \pi_{kc}$. So, $E\{y_k\} = Np_k$
Therefore, for sufficiently small Δ, $E\{\mathbf{y}\} = N \sum_c p_c \pi_c$.

Thus the following holds using .0.7.

$$E\{\mathbf{y}\} = N \sum_c p_c \pi_c = N\Delta \sum_c p_c \varphi(\mathbf{x}_c)/\sigma_c = N\Delta \sum_c p_c \boldsymbol{\varphi}_c/\sigma_c \tag{.0.8}$$

.0.2 Covariance matrix of the count vector **y**

The covariance matrix ($K \times K$)of the count vector **y** is expressed in terms of the $N \times N$ correlation matrix of **z**.The correlation between two z_i's is defined as follows $corr(z_i, z_{i\prime}) = \rho_{ii\prime}$ Note that when $i = i\prime$, $\rho_{ii\prime} = 1$. The total number of possible pairwise correlations is N(N-1)/2 as **z** is an N-dimensional vector. Let M=N(N-1)/2. $g(\rho)$ denotes the uniform correlation density on defined these M possible values. Therefore, $g(\rho) = \frac{1}{M}$ where, $M = \frac{N(N-1)}{2}$.

Suppose $\varphi_\rho(u,v)$ denotes the bivariate density between two correlated standard normal variables, namely u and v with correlation ρ. $\lambda_\rho(u,v)$ (a quantity dependent on ρ)is defined as below.

$$\begin{aligned}&\lambda_\rho(u,v)\\&=\frac{\varphi_\rho(u,v)}{\varphi(u)\varphi(v)}-1\\&=(1-\rho^2)^{-1/2}exp\left\{\frac{2\rho uv-\rho^2(u^2+v^2)}{c(1-\rho^2)}\right\}-1\end{aligned}\tag{.0.9}$$

and $\lambda(u,v)$ (a quantity free of ρ) is defined as the following.

$$\lambda(u,v)=\int_{-1}^{1}\lambda_\rho(u,v)g(\rho)d\rho\tag{.0.10}$$

The integral above is used as a generalization of the mathematical expression below.

$$\sum_{\rho\in\{corr(z_i,z_{i\prime})\}}\lambda_\rho(u,v)g(\rho)\quad\text{where, }\ g(\rho)=\frac{1}{M}\text{ and }\#\{corr(z_i,z_{i\prime})\}=M.$$

The covariance of the count vector **y** is below.

$$\mathbf{cov}(\mathbf{y})=\mathbf{cov}_0+\mathbf{cov}_1\tag{.0.11}$$

Where,

$$\mathbf{cov}_0=N\sum_c p_c\{\text{diag}(\boldsymbol{\pi}_c)-\boldsymbol{\pi}_c\boldsymbol{\pi}_c\prime\}\tag{.0.12}$$

and

$$\mathbf{cov}_1=N^2\sum_c\sum_d p_cp_d\text{diag}(\boldsymbol{\pi}_c)\boldsymbol{\lambda}_{cd}\text{diag}(\boldsymbol{\pi}_d)-N\sum_c p_c\text{diag}(\boldsymbol{\pi}_c)\boldsymbol{\lambda}_{cc}\text{diag}(\boldsymbol{\pi}_c)\tag{.0.13}$$

In the expressions above the summations are over all classes; $diag(\boldsymbol{\pi}_c)$ and $diag(\boldsymbol{\pi}_d)$ denote the $K\times K$ diagonal matrices with diagonal elements π_{kc} and π_{kd} respectively; $\boldsymbol{\lambda}_{cd}$ denotes the $K\times K$ matrix whose klth element is

$\lambda(x_{kc}, x_{ld})$ using the definitions in $.0.5$ and.0.10.Essentially, $\mathbf{cov}_0$ is the covariance when z_i's are independently distributed, and $\mathbf{cov}_1$ is the penalty due to the correlation between z_i's. An increase in N (the number of observations) severely impacts the covariance of the count vector $\mathbf{y}$ through $\mathbf{cov}_1$.$\lambda_p(u, v)$.0.9 is simplified as below using Mehler's Identity.

$$\lambda_\rho(u, v) = \sum_{j \geq 1} \frac{\rho^j}{j!} h_j(u) h_j(v) \tag{.0.14}$$

In the equation above, h_j denotes the jth Hermite Polynomial. The jth moment of the correlation distribution $g(\rho)$ is represented by α_j and defined below.

$$\alpha_j = \int_{-1}^{1} \rho^j g(\rho) d\rho \tag{.0.15}$$

The covariance structure $\widehat{F}$ is expressed in the covariance structure of $\mathbf{y}$. Now,

$$\begin{aligned}
\lambda(u, v) &= \int_{-1}^{1} \lambda_p(u, v) g(\rho) d\rho \quad .0.10 \\
&= \int_{-1}^{1} \left(\sum_{j \geq 1} \frac{\rho^j}{j!} h_j(u) h_j(v) \right) g(\rho) d\rho \quad .0.14 \\
&= \sum_{j \geq 1} \left(\int_{-1}^{1} \rho^j g(\rho) d\rho \right) \frac{h_j(u) h_j(v)}{j!} \quad \text{[as } \rho^j \text{ is uniformly convergent on [-1,1] for j=1,2,...]} \\
&= \sum_{j \geq 1} \frac{\alpha_j}{j!} h_j(u) h_j(v) \quad .0.15
\end{aligned}$$

Thus,

$$\lambda(u, v) = \sum_{j \geq 1} \frac{\alpha_j}{j!} h_j(u) h_j(v) \tag{.0.16}$$

Using the identity above $\boldsymbol{\lambda}_{cd}$ in $.0.13$ is expressed as the following.

$$\boldsymbol{\lambda}_{cd} = \sum_{j \geq 1} \frac{\alpha_j}{j!} h_j(\mathbf{x}_c) h_j(\mathbf{x}_d)\prime \tag{.0.17}$$

We can obtain the following using .0.7.

$$\begin{aligned} \text{diag}(\boldsymbol{\pi}_c)h_j(\mathbf{x}_c) &= \Delta\text{diag}(\varphi(\mathbf{x}_c)h_j(\mathbf{x}_c)/\sigma_c \\ &= (-1)^j\Delta.\boldsymbol{\varphi}_c^{(j)}/\sigma_c \end{aligned} \tag{.0.18}$$

where $\boldsymbol{\varphi}_c^{(j)}$ indicates the jth derivative of $\varphi(u)$ evaluated at each component of $\mathbf{x}_c$ using $\varphi^{(j)}(\mathbf{u}) = (-1)^j\varphi(\mathbf{u})h_j(\mathbf{u})$. Next, the methodology defines a new quantity, namely $\boldsymbol{\phi}^{-(j)}$ as below.

$$\boldsymbol{\phi}^{-(j)} \equiv \sum_c p_c\boldsymbol{\varphi}_c^{(j)}/\sigma_c \tag{.0.19}$$

The above .0.13 can be simplified as follows.

$$\begin{aligned} \mathbf{cov}_1 &= N^2\sum_c\sum_d p_cp_d\text{diag}(\boldsymbol{\pi}_c)\boldsymbol{\lambda}_{cd}\text{diag}(\boldsymbol{\pi}_d) - N\sum_c p_c\text{diag}(\boldsymbol{\pi}_c)\boldsymbol{\lambda}_{cc}\text{diag}(\boldsymbol{\pi}_c) \\ &= N^2\sum_c\sum_d p_cp_d\text{diag}(\boldsymbol{\pi}_c)\bigg(\sum_{j\geq 1}\frac{\alpha_j}{j!}h_j(\mathbf{x}_c)h_j(\mathbf{x}_d)\prime\bigg)\text{diag}(\boldsymbol{\pi}_d) \\ &\quad - N\sum_c p_c\text{diag}(\boldsymbol{\pi}_c)\bigg(\sum_{j\geq 1}\frac{\alpha_j}{j!}h_j(\mathbf{x}_c)h_j(\mathbf{x}_c)\prime\bigg)\text{diag}(\boldsymbol{\pi}_c) \quad \text{using .0.17} \\ &= N^2\sum_{j\geq 1}\frac{\alpha_j}{j!}\bigg(\sum_c p_c\text{diag}(\boldsymbol{\pi}_c)h_j(\mathbf{x}_c)\sum_d p_dh_j(\mathbf{x}_d)\prime\text{diag}(\boldsymbol{\pi}_d)\bigg) \\ &\quad - N\sum_{j\geq 1}\frac{\alpha_j}{j!}\bigg(\sum_c p_c\text{diag}(\boldsymbol{\pi}_c)h_j(\mathbf{x}_c)h_j(\mathbf{x}_c)\prime\text{diag}(\boldsymbol{\pi}_c)\bigg) \\ &= N^2\sum_{j\geq 1}\frac{\alpha_j}{j!}\bigg(\sum_c p_c\Delta\boldsymbol{\varphi}_c^{(j)}/\sigma_c\sum_d p_d\Delta\boldsymbol{\varphi}_d^{(j)\prime}/\sigma_d\bigg) \\ &\quad - N\sum_{j\geq 1}\frac{\alpha_j}{j!}\bigg(\sum_c p_c\Delta\boldsymbol{\varphi}_c^{(j)}/\sigma_c\Delta\boldsymbol{\varphi}_c^{(j)\prime}/\sigma_c\bigg) \quad \text{using .0.18} \\ &= N^2\Delta^2\sum_{j\geq 1}\frac{\alpha_j}{j!}\boldsymbol{\phi}^{-(j)}\boldsymbol{\phi}^{-(j)\prime}/\sigma_c\sigma_d - N\Delta^2\sum_{j\geq 1}\frac{\alpha_j}{j!}\bigg(\sum_c p_c\boldsymbol{\varphi}_c^{-(j)}\boldsymbol{\varphi}_c^{-(j)\prime}/\sigma_c^2\bigg) \quad \text{using .0.19} \end{aligned}$$

Thus .0.13 can be re-written as

$$\mathbf{cov}_1 = N^2\Delta^2\left\{\sum_{j\geq 1}\frac{\alpha_j}{j!}\boldsymbol{\phi}^{-(j)}\boldsymbol{\phi}^{-(j)\prime}/\sigma_c\sigma_d - \frac{1}{N}\sum_{j\geq 1}\frac{\alpha_j}{j!}\Big(\sum_c p_c\boldsymbol{\varphi}_c^{-(j)}\boldsymbol{\varphi}_c^{-(j)\prime}/\sigma_c^2\Big)\right\} \tag{.0.20}$$

α the root mean square (rms) correlation is defined using the second moment of the distribution of the correlation parameter ρ.

$$\alpha = \alpha_2^{1/2} = \left[\int_{-1}^{1}\rho^2 g(\rho)d\rho\right]^{1/2} \tag{.0.21}$$

Using multiple reduction techniques, the methodology provides a simple formula for rms approximation of $\mathbf{cov}_1$ as specified below.

$$\mathbf{cov}_1 \doteq (N\Delta\alpha)^2\boldsymbol{\phi}^{-(2)}\boldsymbol{\phi}^{-(2)\prime}/2 \tag{.0.22}$$

with $\boldsymbol{\phi}^{-(2)}$ in .0.19 depending on the second derivative of the normal density,$\varphi^{(2)}(u) = \varphi(u).(u^2-1)$.

.0.3 Derivation of the covariance of the right-sided empirical cumulative distribution function

To derive the expectation and covariance matrix of $\widehat{F}$, a $K \times K$ matrix, namely B, is defined below.

$$\mathbf{B}_{kk\prime} = \begin{cases} 1 & if\ k \leq k\prime \\ 0 & if\ k > k\prime, \end{cases} \tag{.0.23}$$

A K-vector $\widehat{\mathbf{F}}$ is defined as next.

$$\widehat{\mathbf{F}} = \frac{1}{N}\mathbf{By} \tag{.0.24}$$

Following the definition .0.24 the kth component of $\widehat{\mathbf{F}}$ is actually the proportion of z_i's in bins $\mathcal{Z}_{k\prime}$, where $k\prime \geq k$. As Δ and x_k denote the bin width and mid-point of the bin respectively for all $k \in 1, 2, ..., K$.

Therefore mathematically,

$$\widehat{F_k} = \#\big\{z_i \geq x_k - \Delta/2\ \ /N \qquad k = 1, 2, ..., K \tag{.0.25}$$

Now, we know that $E\widehat{\{F\}} = \mathbf{B}E\{\mathbf{y}\}$ and the covariance matrix of $E\{\widehat{\mathbf{F}}\}$ can be given by $\mathbf{Bcov}(\mathbf{y})\mathbf{B}'/N^2$. Thus,

$$\begin{aligned} E\{\widehat{F_k}\} &= \sum_c p_c \Big[\sum_{k\prime \geq k} \Delta\varphi\Big(\frac{x_{k\prime} - \mu_c}{\sigma_c} \Big)/\sigma_c \Big] \\ &\doteq \sum_c p_c \int_{x_{kc}}^{\infty} \varphi(u)du \\ &= \sum_c p_c \Phi^+(x_{kc}) \end{aligned} \tag{.0.26}$$

where $\Phi^+(u) = 1 - \Phi(u)$. Letting $\Delta \to 0$ make .0.26 exact. $\widehat{\mathbf{F}}$ has covariance matrix $\mathbf{Bcov}(\mathbf{y})\mathbf{B}\prime/N^2$. Using the equations .0.11 .0.12 and .0.13 and the fact that $\widehat{\mathbf{F}}$ has covariance matrix $\mathbf{Bcov}(\mathbf{y})\mathbf{B}\prime/N^2$ one can express $\widehat{\mathbf{F}}$ as below.

$$\mathbf{Cov}(\widehat{\mathbf{F}}) = \mathbf{Cov}_0 + \mathbf{Cov}_1 \tag{.0.27}$$

where $\mathbf{Cov}_0$ has the kl^{th} entry is

$$\frac{1}{N} \sum_c p_c \{ \Phi^+(max(x_{kc}, x_{lc})) - \Phi^+(x_{kc})\Phi^+(x_{lc}) \} \tag{.0.28}$$

and

$$\mathbf{Cov}_1 = \sum_{j\geq 1} \frac{\alpha_j}{j!} \boldsymbol{\varphi}^{-(j-1)} \boldsymbol{\varphi}^{-(j-1)\prime} - \frac{1}{N} \sum_{j\geq 1} \frac{\alpha_j}{j!} \Big\{ \sum_c p_c \boldsymbol{\varphi}_c^{-(j-1)} \boldsymbol{\varphi}_c^{-(j-1)\prime} \Big\} \tag{.0.29}$$

Where, p_c is defined in .0.4, x_{kc} and x_{lc} are from .0.5, α_j is defined in .0.15 and

$$\boldsymbol{\varphi}^{-(j-1)} = \sum_c p_c \boldsymbol{\varphi}_c^{-(j-1)} = \sum_c p_c \boldsymbol{\varphi}^{-(j-1)}(\mathbf{x}_c) \tag{.0.30}$$

Applying reduction steps similar to that used to find $\mathbf{cov}_1$, the rms approximation of $\mathbf{Cov}_1$ is computed below.

$$\mathbf{Cov}_1 \doteq \alpha^2 \boldsymbol{\varphi}^{-(1)} \boldsymbol{\varphi}^{-(1)\prime}/2 \tag{.0.31}$$

.Where, $\boldsymbol{\varphi}^{-(1)}$ depends on the first derivative of the normal density, $\varphi^{(1)}(u) = -\varphi(u)u$.

.0.4 Estimation of the correlation parameter α (Efron (2010)))

Two effective estimates of α as highlighted in Efron (2010)) are used for computing the root mean square correlation as a function of aforesaid m and ν. The formulas for computing the effective rms correlation are presented below.

$$\widehat{\alpha}^2 = \frac{n_0}{n_0 - 1}\left(\nu - \frac{1}{n_0 - 1}\right) \tag{.0.32}$$

$$\tilde{\alpha}^2 = \tilde{\nu} - \frac{3}{n_0 - 5}\tilde{\nu}^2 \quad \left[\tilde{\nu} = \frac{(n_0 - 3)\nu - 1}{n_0 - 5}\right] \tag{.0.33}$$

As seen above, computing $\widehat{\alpha}$ is simpler than computing $\tilde{\alpha}$ Efron (2010)) prefers $\widehat{\alpha}$ as the rms correlation value.

Based on the discussion so far, it may seem like class components like $p_c,\ \mu_c$ and σ_c are also required to be estimated. However, the computations above can be reduced to simpler ones under certain assumptions. Below is how the computational redundancies are avoided, as presented in Efron (2010)).

Using the equations $.0.3$ and $.0.4$ the marginal density $f(z)$ can be expressed as below.

$$f(z) = \sum_c p_c \varphi\left(\frac{z - \mu_c}{\sigma_c}\right)\frac{1}{\sigma_c} \tag{.0.34}$$

Now if **f** is denoted by $f(\mathbf{x})$ which is the density evaluated at the K-vector of bin midpoints, then assuming N_c's are fixed $\Delta.\mathbf{f} = \sum_c p_c \boldsymbol{\pi}_c$ .0.7 and .0.12 can be expressed as below.

$$\mathbf{cov}_0 = N\left\{\text{diag}(\Delta\mathbf{f}) - \sum_c p_c \boldsymbol{\pi}_c \boldsymbol{\pi}'_c\right\} \tag{.0.35}$$

Instead of assuming fixed N_c's, a more pragmatic approach would be assuming that $N_1, N_2, ..., N_C$ are a multinomial sample of size N with probabilities $p_1, p_2, ..., p_C$. Under the assumption mentioned earlier, the above equation .0.35 can be written as next.

$$\mathbf{cov}_0 = N\left\{\text{diag}(\Delta\mathbf{f}) - \Delta^2\mathbf{f}\mathbf{f}'\right\} \tag{.0.36}$$

Now, a smooth estimate of $\mathbf{f}$, namely $\widehat{\mathbf{f}}$, is obtained using a Poisson spline regression. Thus under the assumption that all σ_c values are the same, without knowing the class-specific structure, the $\mathbf{Cov}(\widehat{\mathbf{F}})$ can be obtained from the equations mentioned below. $\mathbf{Cov}_0$ for $\widehat{\mathbf{F}}$ is given by

$$(\widehat{\mathbf{Cov}}_0)_{kl} = \frac{1}{N}\{\widehat{F}_{max(k,l)} - \widehat{F}_k\widehat{F}_l \tag{.0.37}$$

and acknowledging the fact that a smooth estimate of $\widehat{f}(z)$ of $f(z)$ can be differentiated the penalty terms are given by

$$\begin{aligned}\mathbf{Cov}_1 &= \frac{(\sigma_0^2\alpha)^2}{2}\mathbf{f}^{(1)}\mathbf{f}^{(1)\prime} \\ \mathbf{cov}_1 &= \frac{(N\Delta\sigma_0^2\alpha)^2}{2}\mathbf{f}^{(2)}\mathbf{f}^{(2)\prime}\end{aligned} \tag{.0.38}$$

Therefore the correlation penalty standard deviation of $\widehat{F}(x_k)$ can be obtained below.

$$\text{sd}_1\{\widehat{F_k}\} = (\widehat{\mathbf{Cov}}_1)_{kk}^{1/2} = \frac{\widehat{\sigma}_0^2\alpha}{\sqrt{2}}|\widehat{f}^{(1)}(x_k)| \tag{.0.39}$$

APPENDIX E PROOF OF THEOREM 1.1 (Chen (2016)))

Proof. $\Theta^c_{(k+1)} = \Theta$ and $C^k_0\big(\boldsymbol{y}\big) = \Theta$ make sense as in this case $\boldsymbol{\theta} \in \Theta^c_{(k)} \cap \Theta^c_{(k-1)} \cap ... \cap \Theta^c_{(1)}$ and $\Theta = \mathbb{R}^k$.

Thus,

$\Theta^c_{(k+1)} \cap \Theta^c_{(k)} \cap ... \cap \Theta^c_{(1)} \cap C^{k+1}_0\big(\boldsymbol{y}\big)$

$=\Theta \cap \Theta^c_{(k)} \cap ... \cap \Theta^c_{(1)} \cap \Theta$

$=\Theta^c_{(k)} \cap ... \cap \Theta^c_{(1)}$ [As $\Theta = \mathbb{R}^k$]

where, $\Theta^c_{(i)} = \mathbb{R} \times ... \times \mathbb{R} \times \left(\Theta^*_{(i)}\right)^c \times \mathbb{R} \times ... \times \mathbb{R}$

Thus, $\cap^k_{i=1}\Theta^c_{(i)} = \left(\Theta^*_{(k)}\right)^c \times \left(\Theta^*_{(k-1)}\right)^c \times ... \times \left(\Theta^*_{(1)}\right)^c$

Notations: $A \cap B = AB$ and $\Theta_0 = \Theta = \mathbb{R}^k$

Let $\Delta_1 = \Theta_{(k)} \quad [\Theta_{(k)} = \mathbb{R}^{k-1} \times \Theta^*_{(k)}]$,

$\Delta_2 = \Theta^c_{(k)}\Theta_{(k-1)} \quad [\Theta_{(k)} = \mathbb{R}^{k-2} \times \Theta^*_{(k-1)} \times \left(\Theta^*_{(k)}\right)^c]$

$\vdots$

$\Delta_i = \Theta^c_{(k)}\Theta^c_{(k-1)}...\Theta^c_{(k-i+2)}\Theta_{(k-i+1)}$

$[\Theta^c_{(k)}\Theta^c_{(k-1)}...\Theta^c_{(k)}\Theta^c_{(k-i+2)}\Theta_{(k-i+1)} = \mathbb{R}^{k-i} \times \Theta^*_{(k-i+1)} \times (\Theta^*_{(k-i+2)})^c \times ... \times (\Theta^*_{(k-1)})^c \times (\Theta^*_{(k)})^c]$

$\Delta_{i+1} = \Theta^c_{(k)}\Theta^c_{(k-1)}...\Theta^c_{(k-i+2)}\Theta^c_{(k-i+1)}\Theta_{(k-i)}$

$[\Theta^c_{(k)}\Theta^c_{(k-1)}...\Theta^c_{(k)}\Theta^c_{(k-i+2)}\Theta^c_{(k)}\Theta_{(k-i+1)} = \mathbb{R}^{k-i-1} \times \Theta^*_{(k-i)} \times (\Theta^*_{(k-i+1)})^c \times ... \times (\Theta^*_{(k-1)})^c \times (\Theta^*_{(k)})^c]$

$\vdots$

$\Delta_{k-1} = \Theta^c_{(k)}\Theta^c_{(k-1)}...\Theta^c_{(3)}\Theta_{(2)}$

$[\Theta^c_{(k)}\Theta^c_{(k-1)}...\Theta^c_{(3)}\Theta_{(2)} = \mathbb{R} \times \Theta^*_{(2)} \times (\Theta^*_{(3)})^c \times ... \times (\Theta^*_{(k-1)})^c \times (\Theta^*_{(k)})^c]$

$\Delta_k = \Theta^c_{(k)}\Theta^c_{(k-1)}...\Theta^c_{(3)}\Theta^c_{(2)}\Theta_{(1)}$

$[\Theta^c_{(k)}\Theta^c_{(k-1)}...\Theta^c_{(3)}\Theta^c_{(2)}\Theta_{(1)} = \Theta^*_{(1)} \times (\Theta^*_{(2)})^c \times (\Theta^*_{(3)})^c \times ... \times (\Theta^*_{(k-1)})^c \times (\Theta^*_{(k)})^c]$

$\Delta_{k+1} = \Theta^c_{(k)}\Theta^c_{(k-1)}...\Theta^c_{(3)}\Theta^c_{(2)}\Theta^c_{(1)}\Theta_0$

$[\Theta^c_{(k)}\Theta^c_{(k-1)}...\Theta^c_{(3)}\Theta^c_{(2)}\Theta^c_{(1)}\Theta_0 = (\Theta^*_{(1)})^c \times (\Theta^*_{(2)})^c \times (\Theta^*_{(3)})^c \times ... \times (\Theta^*_{(k-1)})^c \times (\Theta^*_{(k)})^c]$ where $\Theta^*_{(i)} \subseteq \mathbb{R}$ denotes the one-dimensional null parameter space corresponding to the hypothesis $H_{(i)}$.

Then,

$$\bigcup_{i=1}^{k+1} \quad \text{as} \quad H^*_{(i)} \cup (H^*_{(i)})^c = \mathbb{R}\ \forall i \quad \text{and} \quad \Theta = \mathbb{R}^k \tag{.0.40}$$

Define $D_{k+1}(\boldsymbol{y}) = \bigcup_{i=1}^{k} C_i^k(\boldsymbol{y})$ where $C_i^k(\boldsymbol{y}) = \{\boldsymbol{\theta} \,|P_i(\boldsymbol{y}|\boldsymbol{\theta}) > \alpha/k\}$.

$C_i^k(\boldsymbol{y})$ can be re-written as $C_i^k(\boldsymbol{y}) = \mathbb{R}^{i-1} \times (C_i^k(\boldsymbol{y}))^* \times \mathbb{R}^{k-i}$ where

$(C_i^k(\boldsymbol{y}))^* = \{\theta_i | P_i(\boldsymbol{y}|\theta_i) > \alpha/k\} \subseteq \mathbb{R}$.

Thus,

$D_{k+1}(\boldsymbol{y}) = C_1^k(\boldsymbol{y}))^* \times \mathbb{R}^{k-1} \bigcup ... \bigcup \mathbb{R}^{i-1} \times (C_i^k(\boldsymbol{y}))^* \times \mathbb{R}^{k-i} \bigcup ... \bigcup \mathbb{R}^{k-1} \times (C_k^k(\boldsymbol{y}))^*$ where

$(C_i^k(\boldsymbol{y}))^* = \{\theta_i | P_i(\boldsymbol{y}|\theta_i) > \alpha/k\} \subseteq \mathbb{R}$.

Define $D_i = C_{k-i+1}^i(\boldsymbol{y}) \quad i = 1, 2, ..., k$.

$C_{k-i+1}^i(\boldsymbol{y}) = \{\boldsymbol{\theta} | P_{(k-i+1)}(\boldsymbol{y}) > \alpha/i\} = \mathbb{R}^{k-i} \times (C_{k-i+1}^i(\boldsymbol{y}))^* \times \mathbb{R}^{i-1} \quad \forall i.$

Where $(C_{k-i+1}^i(\boldsymbol{y}))^* = \{\theta_{(k-i+1)} | P_{k-i+1}(\boldsymbol{y}|\theta_{(k-i+1)}) > \alpha/i\}$.

$(C_{k-i+1}^i(\boldsymbol{y}))^*$ essentially denotes the confidence set corresponding to the $(k-i+1)^{th}$ ordered hypothesis.

The following statement is true from lemma 3.1 in Chen (2016)).

$\mathbb{P}(y : \boldsymbol{\theta} \in C_{(i)}^{k-i+1}(\boldsymbol{y})) \geq 1 - \alpha \quad i = 1, 2, ..., k.$

Now, $D_k = C_{(k)}^k(\boldsymbol{y})$, $D_{k-1} = C_{(2)}^{k-1}(\boldsymbol{y})$, ..., $D_1 = C_{(k)}^1(\boldsymbol{y})$

In other words,

$$D_i = C_{(k-i+1)}^i \quad \text{for} \quad i = 1, 2, ..., k \tag{.0.41}$$

Thus,

$D_k = C_{(i)}^{k-i+1}$ with i=1,

$D_{k-1} = C_{(i)}^{k-i+1}$ with i=2,

$\vdots$

and

$D_1 = C_{(i)}^{k-i+1}$ with i=k.

Therefore using lemma 3.1 in Chen (2016)), we get

$$\mathbb{P}(\theta \in D_i) \geq 1 - \alpha \quad \forall i \tag{.0.42}$$

Also, $D_{k+1}(y) = \bigcup_i C_i^k(\boldsymbol{y})$.

Therefore, $D_{k+1} \supseteq C^k_{(1)}(y)$ and from lemma 3.1 in Chen (2016)),, we know that $\mathbb{P}(\boldsymbol{\theta} \in C^k_{(1)}(y)) \geq 1-\alpha$.

Thus, $\mathbb{P}(\theta \in D_{k+1}(\boldsymbol{y})) \geq \mathbb{P}(\boldsymbol{\theta} \in C^k_{(1)}(\boldsymbol{y})) \geq 1-\alpha$.

To show that $\mathbb{P}(\boldsymbol{\theta} \in \bigcup_{i=1}^{k+1} \Delta_i D_i(\boldsymbol{y})) \geq 1-\alpha$.

Where Δ_i and D_i are defined as above for $i = 1, 2, ..., k+1$.

From equation .0.41, we have the following.

$\mathbb{P}(\theta_{(i_0)} \in C^{i_0}_{(k-i_0+1)}) = \mathbb{P}(\widehat{P}_{(k-i_0+1)} \geq \alpha/i_0) = 1-\alpha/i_0$ as $\widehat{P} \sim Uniform(0,1)$.

Now, we have an index m s.t.

i) $\widehat{P}_{(m)} \geq \frac{\alpha}{k-m+1}$ and

ii) for any index $m < i \leq k$,

$\widehat{P}_{(i)} < \frac{\alpha}{k-i+1}$

To show $\mathbb{P}(\boldsymbol{\theta} \in \Theta^c_{(k+1)} \cap \Theta^c_{(k)} \cap ... \cap \Theta^c_{(m+1)} \cap C^{k-m+1}_{(m)}) \geq 1-\alpha$ we need to prove that

$\bigcup_{i=1}^{k+1} \Delta_i D_i \subseteq \boldsymbol{\theta} \in \Theta^c_{(k+1)} \cap \Theta^c_{(k)} \cap ... \cap \Theta^c_{(m+1)} \cap C^{k-m+1}_{(m)}$.

Now,

$$\bigcup_{i=1}^{k+1} \Delta_i D_i = \left[\bigcup_{i \leq k-m} \Delta_i D_i\right] \cup \left[\bigcup_{i > k-m} \Delta_i D_i\right] \tag{.0.43}$$

Observe that $i \leq k-m$ implies that $k-i+1 > m$. Now from the 2^{nd} criterion mentioned in the statement of the theorem we get $\forall i s.t. k \geq i \geq m, \quad \widehat{P}_{(i)} < \frac{\alpha}{k-i+1}$.

Thus by putting $(k-i+1)$ in place of i we get $\widehat{P}_{(k-i+1)} < \frac{\alpha}{i}$ as $\frac{\alpha}{k-(k-i+1)+1} = \frac{\alpha}{k-k+i-1+i} = \alpha/i$. Now, from lemma 3.2 in Chen (2016)),, we know that if $\widehat{P}_{(k-i+1)} < \alpha/i$ then $C^i_{(k-i+1)}(\boldsymbol{y}) \subseteq \Theta^c_{(k-i+1)}$.

Therefore, $\forall i \leq k-m$ (in other words $\forall i$ s.t. $(k-i+1) > m$) we have

$\Delta_i D_i = \Theta^c_{(k)} ... \Theta^c_{(k-i+2)} \Theta_{(k-i+1)} C^i_{(k-i+1)}(\boldsymbol{y}) \subseteq \Theta^c_{(k)} ... \Theta^c_{(k-i+2)} \Theta_{(k-i+1)} \Theta^c_{(k-i+1)} = \phi$

as $C^i_{(k-i+1)}(\boldsymbol{y}) \subseteq \Theta^c_{(k-i+1)}$. Therefore, $\Delta_i D_i = \phi \quad \forall i \leq k-m$ as $\Theta_{(k-i+1)} \cap \Theta^c_{(k-i+1)} = \phi \forall i$. Thus,

$$\bigcup_{i \leq k-m} \Delta_i D_i = \phi \tag{.0.44}$$

Now, substituting .0.44 in .0.43 we get

$$\begin{aligned}
\bigcup_{i=1}^{k+1} \Delta_i D_i &= \bigcup_{i>k-m}^{k+1} \\
&= \bigcup_{i=k-m+1}^{k+1} \Delta_i D_i \\
&= [\Delta_{k-m+1} D_{k-m+1}] \cup ... \cup [\Delta_{k+1} D_{k+1}] \\
&= [\Theta^c_{(k)} ... \Theta^c_{(m+1)} \Theta_{(m)} C^{k-m+1}_{(m)}(\boldsymbol{y})] \cup ... \cup [\Theta^c_{(k)} ... \Theta^c_{(m)} \Theta^c_{(m-1)} ... \Theta^c_{(2)} \Theta^c_{(1)} (\cup_i C^k_{(i)}(\boldsymbol{y}))] \\
&= \Theta^c_{(k)} \Theta^c_{(k-1)} ... \Theta^c_{(m+1)} [\Theta_{(m)} C^{k-m+1}_{(m)}(\boldsymbol{y}) \cup (\Theta^c_{(m)} \Theta_{(m-1)} C^{k-m+2}_{(m-1)}(\boldsymbol{y}) \cup \\
&\quad (\Theta^c_{(m)} \Theta^c_{(m-1)} \Theta_{(m-2)} C^{k-m+3}_{(m-2)}(\boldsymbol{y}) \cup ... \cup (\Theta^c_{(m)} \Theta^c_{(m-1)} ... \Theta^c_{(1)} (\cup_i C^k_{(i)}(\boldsymbol{y})))]
\end{aligned}$$

by using the definition of Δ_i and D_i for $i = 1, 2, ..., k+1$. Thus,

$$\bigcup_{i=1}^{k+1} \Delta_i D_i \subseteq \Theta^c_{(k)} \Theta^c_{(k-1)} ... \Theta^c_{(m+1)} [\Theta_{(m)} C^{k-m+1}_{(m)}(\boldsymbol{y}) \cup \Theta^c_{(m)}] \tag{.0.45}$$

As $\widehat{P}_{(m)} > \frac{\alpha}{k-m+1}$ (criterion i) in the theorem statement), therefore by lemma 3.2 in Chen (2016)), we have $C^{k-m+1}_{(m)} \supseteq \Theta^c_{(m)}$ as $C^t_i(\boldsymbol{y})$ is directed towards Θ^c_i for $i, t = 1, 2, ..., k$.

Therefore,

$\Theta_{(m)} C^{k-m+1}_{(m)}(\boldsymbol{y}) \cup \Theta^c_{(m)}$

$= (\Theta_{(m)} \cup \Theta^c_{(m)}) \cap [C^{k-m+1}_{(m)}(\boldsymbol{y}) \cup \Theta^c_{(m)}]$

$= \Theta \cap C^{k-m+1}_{(m)}(\boldsymbol{y}) \quad [\text{As } \Theta^c_{(m)} \subseteq C^{k-m+1}_{(m)}(\boldsymbol{y}),\ \Theta^c_{(m)} \cup C^{k-m+1}_{(m)}(\boldsymbol{y}) = C^{k-m+1}_{(m)}(\boldsymbol{y})]$

$= C^{k-m+1}_{(m)}(\boldsymbol{y})$

Thus, .0.45 becomes $\bigcup_{i=1}^{k+1} \Delta_i D_i \subseteq \Theta^c_{(k)} \Theta^c_{(k-1)} ... \Theta^c_{(m+1)} C^{k-m+1}_{(m)}(\boldsymbol{y})$.

Therefore, $\mathbb{P}(\Theta^c_{(k)} \Theta^c_{(k-1)} ... \Theta^c_{(m+1)} C^{k-m+1}_{(m)}(\boldsymbol{y})) \geq \mathbb{P}(\bigcup_{i=1}^{k+1} \Delta_i D_i)$. Now, we just need to show that $\mathbb{P}(\bigcup_{i=1}^{k+1} \Delta_i D_i) \geq 1 - \alpha$.

Let $\theta \in \bigcup_{i=1}^{k+1} \Delta_i D_i$ be arbitrary.

Then, $\theta \in \Delta_i D_i$ for some i by construction as $\Delta_i^{'}$s form a partition on $\Theta = \mathbb{R}^k$.

Therefore, $\mathbb{P}(\theta \in \Delta_i) = 1$.

From .0.42 we have $\mathbb{P}(\theta \in D_j) \geq 1 - \alpha \quad \forall j$.

Therefore, it is sufficient to show that $\mathbb{P}(\theta \in \Delta_i D_i) \geq 1 - \alpha$.

Now,

$\mathbb{P}(\theta \in \Delta_i D_i)$

$=1 - \mathbb{P}(\theta \in \Delta_i^c \cup \theta \in D_i^c)$ by using De-Morgan's Law

$\geq 1 - \mathbb{P}(\theta \in \Delta_i) - \mathbb{P}(\theta \in D_i^c)$ [As $\mathbb{P}(\theta \in \Delta_i^c \cup \theta \in D_i^c) \leq \mathbb{P}(\theta \in \Delta_i) + \mathbb{P}(\theta \in D_i^c)$]

$\geq 1 - 0 - \alpha$ [As $\mathbb{P}(\theta \in D_i) \geq 1 - \alpha$]

$=1 - \alpha$ □