12-31-2019

Topics on high dimensional selective inference

Yan Zhang
New Jersey Institute of Technology

Copyright Warning & Restrictions

The copyright law of the United States (Title 17, United States Code) governs the making of photocopies or other reproductions of copyrighted material.

Under certain conditions specified in the law, libraries and archives are authorized to furnish a photocopy or other reproduction. One of these specified conditions is that the photocopy or reproduction is not to be "used for any purpose other than private study, scholarship, or research." If a, user makes a request for, or later uses, a photocopy or reproduction for purposes in excess of "fair use" that user may be liable for copyright infringement,

This institution reserves the right to refuse to accept a copying order if, in its judgment, fulfillment of the order would involve violation of copyright law.

Please Note: The author retains the copyright while the New Jersey Institute of Technology reserves the right to distribute this thesis or dissertation

Printing note: If you do not wish to print this page, then select "Pages from: first page # to: last page #" on the print dialog screen

ABSTRACT

TOPICS ON HIGH DIMENSIONAL SELECTIVE INFERENCE

by
Yan Zhang

In such applications as identifying differentially expressed genes in micro-array experiments or assessing safety and efficacy of drugs in clinical trials, researchers often report confidence intervals (CIs) and p-values only for the selected parameters, which is called selective inference. While constructing multiple CIs for the selected parameters, it is common practice to ignore issue of selection and multiplicity. Although protection against the effect of selection is sufficient in some cases, simultaneous coverage should be also needed in real applications. For example, in clinical trials, multiple endpoints are considered to assess effects of a drug and the ultimate decision often depends on joint outcome for primary endpoints.

In this dissertation, a new concept of γ-false coverage proportion (γ-FCP) is first presented as a proper measurement for CIs following selection. Such a new measurement has advantages since it takes effect of selection into consideration as well as simultaneous coverage. If a procedure control γ-FCP at a desired level α, then it implies such procedure has high proportion of CIs, which cover the corresponding parameters with high probability. Aiming at keeping γ-FCP at a desired level, two types of procedures are developed. One type is based on unconditional CI; the other type is based on conditional CI, which means CI is conditional on the event of selection. An unconditional CI-based procedure is firstly developed, which is proven to control γ-FCP at a desired level under independence. Theoretically, the result is able to be extended to positive regression dependence. Secondly, a modified unconditional CI-based procedure is presented to control γ-FCP under arbitrary dependence. Thirdly, with approach of conditional CIs, a new conditional CI-based selective inference procedure is developed, which is able to control γ-FCP at

a desired level under independence. Finally a modified conditional CI-based procedure is developed to control γ-FCP under arbitrary dependence.

All of the proposed procedures are evaluated through extensive simulation studies. The effect of nonzero proportion, selection level, and correlation coefficient are evaluated, while we apply the proposed procedures in terms of γ-FCP control and average width of CIs. The simulation studies are then applied to strong dependence such as equal correlation and several weak dependence such as block-wise dependence. The simulation studies show that the proposed procedures are able to either control γ-FCP or have shorter width of CIs than existing methods such as FCR controlling procedures (Benjamini and Yekutieli, 2005). Next, all of the proposed procedures are applied on two sets of micro-array gene expression data. Compared to same existing methods, the proposed conditional CI-based procedure provides (i) shorter width of CI; and (ii) more count of CI not covering zero; and (iii) longer distance of CI away from zero.

TOPICS ON HIGH DIMENSIONAL SELECTIVE INFERENCE

by
Yan Zhang

A Dissertation
Submitted to the Faculty of
New Jersey Institute of Technology and
Rutgers, The State University of New Jersey – Newark
in Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy in Mathematical Sciences

Department of Mathematical Sciences, NJIT
Department of Mathematics and Computer Science, Rutgers-Newark

December 2019

APPROVAL PAGE

TOPICS ON HIGH DIMENSIONAL SELECTIVE INFERENCE

Yan Zhang

Wenge Guo, Dissertation Co-Advisor — Date
Associate Professor, Department of Mathematical Sciences, NJIT

Yixin Fang, Dissertation Co-Advisor — Date
Director, GMA Statistics, AbbVie

Sundarraman Subramanian, Committee Member — Date
Associate Professor, Department of Mathematical Sciences, NJIT

Antai Wang, Committee Member — Date
Associate Professor, Department of Mathematical Sciences, NJIT

Zhi Wei, Committee Member — Date
Professor, Department of Computer Science, NJIT

BIOGRAPHICAL SKETCH

Author: Yan Zhang

Degree: Doctor of Philosophy

Date: September 2019

Undergraduate and Graduate Education:

- Doctor of Philosophy in Mathematical Sciences,
  New Jersey Institute of Technology, Newark, NJ, 2019
- Master of Science in Biostatistics
  New Jersey Institute of Technology, Newark, NJ, 2014
- Bachelor of Science in Biology and Biochemistry
  China Agricultural University, Beijing, China, 2011

Major: Mathematical Sciences

Presentations and Publications:

Introduction to selective inference, *Summer Research Seminar*, DMS, NJIT, June 2017.

Some topics on statistical inference, *Summer Research Seminar*, DMS, NJIT, June 2018.

The dedication of this dissertation is split seven ways:

- To my mother, Mrs. Ling (Lynn) Ye, who supported me to travel overseas and to pursue my dream at U.S.
- To my father, Mr. Jungong Zhang, who made me a person fond of Legos and Teddy bears at the same time.
- 感谢我亲爱的姥姥，我毕业了会努力挣钱给您买好吃的巧克力饼干。
- 感谢全界最好的的姥爷。如果人生就是一场游戏，愿在游戏重启时，我们亦是最好的家人。
- To my aunties and uncles, Mrs. Ying Ye, Dr. Gang Ye, Mr. Rong Qian, Mrs. Xuemin Ye-Han, Mrs. Willa Ye and Dr. Ning Luo, my cousins, Ye, Kento, Kenji and Joy, who gave me a lot of courage to move on.
- To Dr. Silu Sheng, who read this dissertation first and went through the journey with me.
- And to you, if you have stuck with this one until the very end.

ACKNOWLEDGMENT

Rome was not built in a day. I really want to thank all of my advisors, professors, colleagues, families and friends. Without them, it is impossible to complete this dissertation.

First of all, I would like to express my deepest appreciation to my advisors, Dr. Wenge Guo and Dr. Yixin Fang. Dr. Guo, to whom I want to give a special gratitude, offered me the opportunity to work on this wonderful project. During my study at NIJT, Dr. Guo taught me to keep a big picture all the time. This is extremely helpful for both my academic research and future career. The unconditional support that Dr. Guo has given me to balance my research and internship will never be forgotten. I would also like to thank Dr. Fang. His selfless guidance help me a lot during my study at NJIT. For the past three years, he taught me the methodology of critical thinking. Both of Dr. Guo and Dr. Fang have always been the inspiration through enthusiasm and encouragement especially for me to pursue interesting and exciting ideas, and the practicality of their guidance.

Next, I want to thank the rest members of my dissertation committee: Dr. Sundarraman Subramanian, Dr. Antai Wang and Dr. Zhi Wei. They have offered their support and good wills throughout the thesis. Besides my dissertation committee members, I would also like to thank my colleagues, Dr. Yalin Zhu, Dr. Li Yu, Dr. Gan Luan, Mrs. Beibei Li, Mr. Zhongcheng Lin and Mrs. Ziyan Guo, whom I have collaborated with and shared office space with.

The last but not the least, I must say special thanks to my families, Mrs. Ling Ye, Mr. Jungong Zhang, Mr. Dujie Ye, Mrs. Peiyuan Guo, Mrs. Ying Ye, Dr. Gang Ye, Mr. Rong Qian, Mrs. Xuemin Ye-Han, Mrs. Weiqiang (Willa) Ye, Dr. Ning Luo, Ye, Kento, Kenji, Joy, and Dr. Silu Sheng.

TABLE OF CONTENTS

TABLE OF CONTENTS
(Continued)

Chapter **Page**

LIST OF TABLES

LIST OF FIGURES

Figure **Page**

LIST OF FIGURES
(Continued)

Figure **Page**

LIST OF FIGURES
(Continued)

Figure **Page**

LIST OF FIGURES
(Continued)

Figure **Page**

LIST OF FIGURES
(Continued)

Figure **Page**

LIST OF FIGURES
(Continued)

Figure **Page**

CHAPTER 1

INTRODUCTION

1.1 Introduction

In modern scientific investigations, high dimensional data is getting involved, which is difficult to analysis, such as genome-wise association study (GWAS) and micro-array analysis (Lee et al., 2018; Jones et al., 2019). There is a common practice that researchers tend to report only a few confidence intervals (CIs) or p-values for the parameters selected after viewing data (Benjamini et al., 2009). In practice, most statistical analysis involves selective inference, in which CIs and p-values are only reported for the selected variables (Benjamini and Yekutieli, 2005; Efron, 2008; Lee et al., 2013; Peng et al., 2017). Such selected CIs are not able to provide the assumed coverage probability (Benjamini and Yekutieli, 2005). To better understand and investigate about such selective inference in large scale experiments like micro-array or fMRI study, Benjamini and Yekutieli (2005) introduced false coverage rate (FCR) as an appropriate measure to be controlled while constructing a large number of CIs. FCR is widely used in high dimensional inference. Benjamini and Yekutieli (2005) proposed several FCR controlling procedures. However, FCR controlling procedures have limitation because it only takes effects of selection into consideration (Benjamini, 2010). Although protection against effect of selection is sufficient in some cases, simultaneous coverage (Benjamini et al., 2019) is also needed in many real applications, for example, most clinical trials contain multiple endpoint to assess effects of drug and ultimate decision often depends on joint outcome of several selected parameters for primary endpoints. In this dissertation, a new concept of γ-false coverage proportion (γ-FCP) is presented as a proper measurement of simultaneous coverage, in a sense that most of CIs are able to cover the corresponding parameters.

Such new measurement has advantages since it takes simultaneous coverage into consideration as well as effect of selection.

In order to control γ-FCP, we suggest two types of procedures. One type is based on unconditional confidence interval (CI), the other type is based on conditional CI, which is CI conditional on the event of selection (Weinstein et al., 2013). In this dissertation, a general unconditional CI-based selective inference procedure is developed, which is proven to control γ-FCP at a desired level under positive regression dependence (Benjamini and Yekutieli, 2001). An adjusted unconditional CI-based procedure is then developed to control γ-FCP under arbitrary dependence. At the same time, with the approach of conditional CIs (Weinstein et al., 2013), a new conditional CI-based selective inference procedure is developed, which is able to control γ-FCP at a desired level under independence. An adjusted conditional CI-based procedure is then developed to control γ-FCP under arbitrary dependence. All of the proposed procedures are evaluated through extensive simulation studies under independence. We evaluate effect of some factors, such as nonzero proportion, selection level and correlation coefficient, while we apply our proposed unconditional CI-based procedures and conditional CI-based selective inference procedures in terms of γ-FCP control and average width of CIs. The simulation studies are also applied to strong dependence such as equal correlation and several weak dependences such as block-wise dependence. Our simulation studies are able to show that the proposed procedures are able to either control γ-FCP or have shorter width of CIs than existing methods such as FCR controlling procedures (Benjamini and Yekutieli, 2005). Next, all of the proposed procedures are applied on two sets of micro-array gene expression data. Compared to existing methods such as FCR controlling procedures (Benjamini and Yekutieli, 2005), the proposed procedure is demonstrated to provide (i) shorter width of CI; and (ii) more count of CI not covering zero; and (iii) longer distance of CI away from zero.

1.2 Confidence Interval Versus p-value

It is of interest to compare limitations and advantages of two types of widely used inference: CI and p-value. In scientific publications, it is common practice for researcher to misuse and misunderstand p-value. Wasserstein and Lazar (2016) gave a comprehensive discussion about the limitations of p-value. It is worth to mention that a p-value does not measure size of an effect or importance of a result. Moreover, a p-value does not provide a good measure of evidence regarding a model or hypothesis. Besides the disadvantages of p-values have been frequently discussed in literatures, CI can act into scientific publications as a better alternative (Ranstam, 2012). For example, in clinical trails, CI can easily show and evaluate clinical significance, which is practical importance of treatment effect, no matter whether or not it has a real and noticeable effect on daily life.

To better explain the advantages of CI, we offer an illustration example in Figure 1.1 for the comparison between CI and p-value. In this example, there are five CIs and corresponding p-values. Four pairs of CIs are compared to illustrate advantage of CIs against p-values. Each pair is aiming to explain one of advantage of CIs against p-values in details. In this clinical trial example, statistical significance is equivalent to (1) p-value is less than 0.05 or (2) CI not covering 0 (whether CI cross the black dashed vertical line or not in Figure 1.1). In this example, we assume clinical significance as effect is greater than 2 (whether the CI is on the right hand side of dark red dashed vertical line or not in Figure 1.1).

- CI 1 vs CI 2 shows the fact that CI can provide clinically significance, while the corresponding p-value has no information about it. The p-values are both greater than 0.05, only showing no statistical significance for both. Meanwhile, CI 2 is able to show clinical significance, and CI 1 does not.

- CI 3 vs CI 5 presents that CI can provide details about direction of effect. Without point estimation, p-value can not provide details about direction of effect. As the corresponding p-values are less than 0.05 (statistical significance), but CI 3 is able to present negative effect, and CI 5 is able to present positive effect.

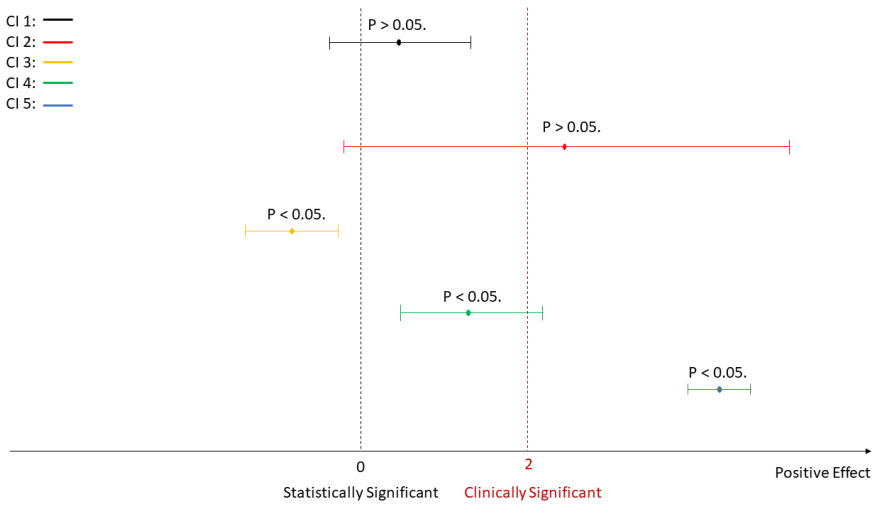

Figure 1.1 Illustration example of five CIs and corresponding p-values. Black vertical dashed line means statistical significant. Red dashed line means clinical significant.

- CI 4 vs CI 5 demonstrates that CI is able to distinguish clearly about the distance between an estimation and a standard level (zero in this example). But p-value can not measure it. The center of CI 2 is much more far away from zero than CI 4, though CI 2's corresponding p-value only display non statistical significance (greater than 0.05), and CI 4's corresponding p-value shows the same.

1.3 Basic Concepts of Multiple Hypothesis Testings

Hypothesis testing is a widely used method of statistical inference. A single hypothesis testing is able to explore one research question, and multiple hypotheses testing are used to investigate multiple study objectives simultaneously. While conducting multiple hypothesis testings, it is critical to define the type I error, which occurs when the null hypothesis is true, but is falsely rejected by testing. With the increasing number of testing hypotheses, multiplicity issue arises in the sense that type I error rate is inflated. Considering the multiple hypothesis testing problem, we first introduce three commonly used type I error rates: familywise error rate (FWER), false discovery rate (FDR) and γ-false discovery proportion (γ-FDP). It is important, in the filed of multiple hypotheses testing, to address the multiplicity issue appropriately and to control overall error rates, meanwhile various multiple testing

procedures (MTPs) have been developed to overcome such issue with the control of proper error rates. Followed by the error measurement, a commonly used multiple hypotheses testing procedures (MTPs), which are then introduced.

1.3.1 Type I Error Rate

To begin with the definitions, we denote R as the number of total rejections and V as the number of false rejections. First of all, we introduce a concept: familywise error rate (FWER).

Definition 1.1 (Familiwise Error Rate (FWER))**.** *Familiwise error rate is defined as probability of making at least one false rejection, that is,*

$$FWER = P\Big[V \geq 1\Big].$$

Remark 1.1. *When dealing with small scale multiple testing problems,* FWER *is commonly used to measure type I error rate. Especially in clinical trials, it is mandatory to strongly control the* FWER *by the Food and Drug Administration (FDA).*

In large-scale multiple testing, it is too conservative for controlling FWER to detect any false null hypothesis. In practice, it may allow a few false discoveries in order to gain power to detect more. Based on this idea, Benjamini and Hochberg (1995) suggest another type I error measure: false discovery rate (FDR).

Definition 1.2 (False Discovery Rate (FDR))**.** *False discovery rate is defined as the expected proportion of false rejections among all rejected hypotheses. That is,*

$$FDR = E\Big[\frac{V}{R \vee 1}\Big],$$

where $R \vee 1 = \max\{R, 1\}$.

To control FDR means that when the experiment is repeated many times, on average we control the false discovery proportion FDP $= \frac{V}{R \vee 1}$. In this sense, it does not consider the variability of FDP. Though we may keep the average of FDP at a desired level, the actual FDP could be quite large. To deal with this issue, Lehmann and Romano (2005) introduced another Type I error measure: γ-false discovery proportion (γ-FDP).

Definition 1.3 (γ-False Discovery Proportion ($\gamma-$FDP))**.** *γ-False discovery proportion is defined as the probability of ratio, which is the false rejections among all rejected hypotheses, beyond a pre-specified value. That is,*

$$\gamma\text{-}FDP = P\Big[\frac{V}{R \vee 1} > \gamma\Big],$$

where γ is pre-specified positive number between 0 and 1.

To sum up, when large-scale of multiple hypothesis testings are involved, FWER is not an appropriate measurement. But FWER can ensure simultaneous correctness of a set of multiple testings. FDR can not guarantee all of the rejecting multiple testings to be true. But FDR is a more useful approach to determining a significance cutoff in high dimensional study. We suggest $\gamma-$FDP among these error as a good measurement, not only for $\gamma-$FDP are able to measure the falsely rejecting multiple testings simultaneously, but also for $\gamma-$FDP can be used to adjust to the large-scale of multiple hypothesis testings.

1.3.2 Multiple Testing Procedures

A commonly used multiple testing procedures (MTPs) is Benjamini-Hochberg procedure (BH procedure). Consider a problem of testing m hypotheses, $H_1, H_2, \ldots, H_m$ with corresponding p-values $p_1, p_2, \ldots, p_m$. Let $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$ denote ordered version of p_i's with corresponding $H_{(1)}, H_{(2)}, \ldots, H_{(m)}$. Reject $H_{(1)}, H_{(2)}, \ldots, H_{(R)}$,

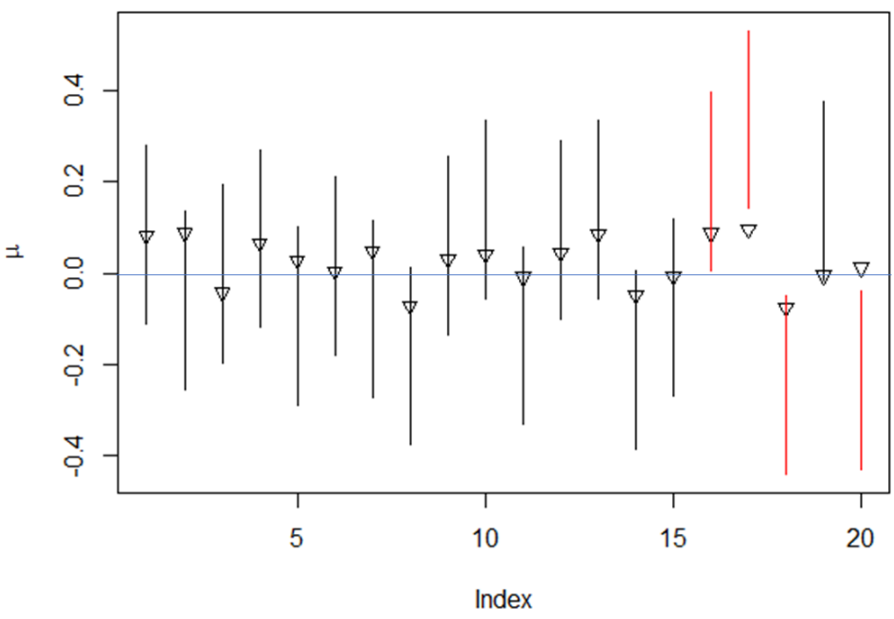

Figure 1.2 Illustration example of twenty CIs (both black segments and red segments) and four selected CIs (only red segments), triangle indicate the corresponding parameters.

where $R = \max\{0 \leq i \leq m : P_{(i)} \leq \alpha_i = \frac{i}{m}\alpha\}$. BH procedure is a FDR controlling procedure under independence or positive regression dependence

1.4 Basic Concepts of Confidence Intervals Based Method

There is a common practice that researchers tend to report only a few CIs for the parameters selected after viewing data. There are two types of serious issues behind such common practice. We name the two types of issues as (1) issue of multiplicity when constructing multiple CIs at same time and (2) issue of selection when reporting a few CIs after viewing the data. And next we illustrate such issues in an illustrative example in Figure 1.2.

We generate data $X_{ij} \sim N(\mu_i, 1), i = 1, 2, ..., 20, j = 1, 2, ..., 100$. And $\bar{X}_i = \frac{1}{100}\sum_{j=1}^{100} X_{ij}$ is calculated as the estimator to select parameter μ_i. Parameter μ_i is selected if $|\bar{X}_i| \geq 0.2$. Once 95% unconditional CI is constructed, then we draw a black segment in Figure 1.2. At the meantime, if the parameter selected and corresponding

95% unconditional CI is constructed, then we draw a red segment in Figure 1.2. μ_i is a uniformly distributed random number between (-0.2, 0.2). It is drawn by triangle in Figure 1.2.

The issue of multiplicity can be shown by all CIs (both black and red segments in Figure 1.2). Among 20 CIs, 2 out of 20 (No.17 and No. 20) are not covering its corresponding true parameters (triangle). In this sense, when constructing multiple CIs without selection, 95% unconditional CIs can not ensure that CIs does not cover the corresponding parameters with probability less than 5% (10% $>$ 5%). When selection are involved (only red segments in Figure 1.2), 2 out of 4 CIs are not covering its corresponding true parameters in the selected ones. This example is designed to see issue of selection. In such sense, when constructing multiple CIs with selection, proportion of a mistake arises for 95% unconditional CIs, since the selected CIs does not cover the corresponding parameters with probability 50% (far away from 5%). Hence it is necessary to address and suggest some new methods for constructing multiple CIs for the selected ones.

1.4.1 False Coverage Rate

Let R_{CI} be the number of constructed CIs and V_{CI} be the number of constructed CIs not covering their respective parameters. False coverage proportion (FCP) is ratio of true parameters, which is not covered by CI. Among the selected CIs, FCP can be denoted as

$$FCP = \frac{V_{CI}}{R_{CI} \vee 1}. \tag{1.1}$$

Benjamini and Yekutieli (2005) suggest a new error rate: false coverage rate (FCR), which is the average rate of FCP.

Definition 1.4 (False Coverage Rate (FCR))**.** *The false coverage rate is defined as the expected proportion of non-covering confidence intervals among all constructed*

confidence intervals, that is,

$$FCR = E\big[FCP\big]. \tag{1.2}$$

For a single parameter ($m = 1$), the FCR equals the probability of constructing a non-covering CI. One single $1 - \alpha$ CI therefore has FCR $< \alpha$. Though, in real world, multiple parameters are more often involved and hence Benjamini and Yekutieli (2005) develop some FCR controlling procedures.

1.4.2 Confidence Interval Based Method: FCR Controlling Procedures

To better understand and investigate about the such selective inference in high dimensional data, such as micro-array or fMRI study, Benjamini and Yekutieli (2005) suggested a procedure for constructing selective multiple CIs (selective CIs), based on a vector of m parameter estimators $\mathbf{T}$. The selection procedure is given by $S(\mathbf{T}) \subseteq \{1, ..., m\}$ and is followed by the construction of some CI for each θ_i, $i \in S(\mathbf{T})$.

Definition 1.5 (Level α FCR-Adjusted Selective CI-Based Procedure)**.**

1. Apply the selection criterion $\hat{S}$ to $\mathbf{T}$, yielding the selected set of parameters as $\hat{S}(\mathbf{T})$.

2. For each selected parameter $\mu_i, i \in \hat{S}(\mathbf{T})$, partition $\mathbf{T}$ into T_i and $\mathbf{T}^{(i)} = \mathbf{T} \setminus \{T_i\}$, and find $R^{(i)}_{min} := \min\{|\hat{S}(\mathbf{T}^{(i)}, T_i = t)| : i \in \hat{S}(\mathbf{T}^{(i)}, T_i = t)\}$.

3. For each selected parameter $\mu_i, i \in \hat{S}(\mathbf{T})$, construct the following confidence interval: $CI_i(\frac{R}{m}\alpha)$.

If components of $\mathbf{T}$ are independent, then the FCR-adjusted selective CI in Definition 1.5 enjoys $FCR \leq \alpha$. To process such method to a complex structure of the selection estimator $\mathbf{T}$, Benjamini and Yekutieli (2005) developed and showed the following results. Before we formally present the concept, a definition of positive regression dependent on a subset (PRDS) was developed by Benjamini and Yekutieli (2001).

Definition 1.6 (Benjamini and Yekutieli, 2001). *The components of* $\mathbf{X}$ *are positive regression dependent on a subset (PRDS) on a give subset* $I_0 \subseteq I = \{1, ..., m\}$*, if for any increasing set* D *(where* $x \in D$ *and* $y \geq x$ *implies that* $y \in D$*) and for each* $i \in I_0, P(\mathbf{X} \in D | X_i = x)$ *is nondecreasing in* x*. Specially, if* $\mathbf{X}$ *is PRDS on any subset of* I*, we can simply denote it as PRDS.*

If the condition change from $P(\mathbf{X} \in D | X_i = x)$ to $P(\mathbf{X} \in D | X_i \geq x)$, for each $i \in I_0$, is nondecreasing in x, we can still denote $\mathbf{X}$ as PRDS. Meanwhile, a definition of concordant of CIs was developed by Benjamini and Yekutieli (2005).

Definition 1.7 (Benjamini and Yekutieli, 2005). *A procedure for selective confidence intervals is concordant if for all values of* μ*, for all* $0 < \alpha < 1$*, and for* $i = 1, ..., m, k = 1, ..., m$*, both* $\{\mathbf{T}^{(i)} : k \leq R_{min}(\mathbf{T}^{(i)})\}$ *and* $\{T_i : \mu_i \notin CI_i(\alpha)\}$ *are either increasing or decreasing sets.*

Theorem 1.1 (Positive Dependence). *If components of* $\mathbf{T}$ *are PRDS and the selection criterion and the confidence intervals are concordant, then the FCR-adjusted selective confidence intervals in Definition 1.5 enjoys* $FCR \leq \alpha$*.*

Moreover, Benjamini and Yekutieli (2005) have proven the Theorem 1.2 under arbitrary dependence.

Theorem 1.2 (General Dependence). *For any monotone unconditional confidence intervals, any selection procedure* $\hat{S}(\mathbf{T})$*, and any dependence structure of the estimators for confidence intervals, the FCR of the FCR-adjusted selective confidence intervals in Definition 1.5 is bounded by* $\alpha \sum_{j=1}^{m} \frac{1}{j}$*.*

All these results allow the researchers to construct multiple selective CIs and still keep FCR at a desired level. In the BH procedure, after sorting the p values $p_{(1)} < ... < p_{(m)}$ and calculating $R = \max\{j : p_{(j)} < j\alpha/m\}$, the R null hypotheses for which $p_{(\cdot)} < R\alpha/m$ are rejected. Our suggested method of adjusting for FCR at level α is defined as following.

Definition 1.8 (Level α FCR-Adjusted BH-Selected CI-Based Procedure)**.**

1. Sort the p-values used for testing the m hypotheses regarding the parameters, $p_{(1)}, p_{(2)}, \ldots, p_{(m)}$.

2. Calculate $R = \max\{j : p_{(j)} \leq j\alpha/m\}$.

3. Select the R parameters for which $p_{(i)} \leq R\alpha/m$*, corresponding to the rejected hypotheses.*

4. Construct a $1 - R\alpha/m$ *confidence interval for each parameter selected.*

1.5 Literature Review

It is a common practice that researchers tend to report only a few CIs or p-values for the parameters selected after viewing data (Benjamini and Yekutieli, 2005; Benjamini et al., 2009; Peng et al., 2017). Benjamini and Yekutieli (2005) demonstrated that CIs which are reported for selected parameters cannot guarantee nominal coverage even on average. Benjamini and Yekutieli (2005) suggested a concept of FCR and several FCR controlling procedures, where CIs are constructed for the selected parameters. Simultaneous selective inference and traditional justification for multiple-comparisons procedures are two distinct goals (Benjamini, 2010). FDR and FCR are viewed as concepts to address directly the dangers that are caused by selective inference, which may alter meaning of reported p-values and CIs, while giving up simultaneous inference. In most large problems, only effect of selection is taken care of (Benjamini and Yekutieli, 2005; Benjamini, 2010).

In recent decades, several progresses have been made about constructing CI after selection. Weinstein et al. (2013) developed methods of constructing conditional CIs and suggested three methods, which can offer FCR control. For these reasons, conditional CIs for the selected parameters are able to be used as an attractive alternative to available general FCR adjusted intervals (Benjamini and Yekutieli, 2005). Based on the previously proposed methods, Weinstein and Yekutieli (2014)

then suggested a procedure, which employs FCR-adjustment to an unconditional CI in order to construct a maximum number of sign-determining CIs. Moreover, Weistein and Ramdas (2019) recently presented a general unconditional CI-based procedure, which can be used to devise online sign classification and control false sign rate (FSR), which is the expected ratio of number of incorrect directional decisions to total number of directional decisions made.

In the era of post model selection inference, a valid "post-selection inference" (Berk et al., 2013) was proposed by reducing problem to one of simultaneous inference and therefore suitably widening conventional CIs. Lee et al. (2013) developed a general approach to characterize distribution of a post-selection estimator conditioned on the selection event. A method was developed by Lee and Taylor (2014) to construct valid CIs and hypothesis tests for regression coefficients that account for the selection procedure, which has no required assumptions on design matrix. Fithian et al. (2017), based on classical theory of Lehmann and Scheff'e (1955), derived some powerful unbiased selective tests and CIs for inference in exponential family models after arbitrary selection procedures for linear regression.

Efron (2008) has discussed some issues, as well as associated difficulties, from the empirical Bayes approach. Benjamini (2010) has discussed it from both the Bayesian and the empirical Bayes approach and addressed formally the effects of selection. Benjamini and Gavrilov (2009) have drawn attention to this problem in replicability studies of genomewise scans for association with a disease. Woody Scott (2018) proposed nonparametric empirical-Bayes approach for constructing optimal selection-adjusted CIs.

In this dissertation, we are interested in the two challenges about selective inference. One challenge is about simultaneous selective inference; the other challenge is about high dimensional selective inference (Tian and Taylor, 2018). For the first challenge, Katsevich and Ramdas (2018) addressed simultaneous selective inference

in testing. Benjamini et al. (2019) formally define "simultaneous over the selected" (SoS) error rate, which is the probability that one or more intervals for selected parameters do not cover. Benjamini et al. (2019) suggest a method of constructing SoS controlling CIs for parameters which are selected. For the other challenge of high dimensional selective inference (Tian and Taylor, 2018), as high dimensional inference (Bühlmann and Geer, 2011) is a very important in modern science, Taylor addressed the conditional approach. Even though some results have already been published (Markovic and Leeb 2019; Wasserman and Roeder, 2009), there are still much work need to be done. It remains unknown whether the method can be applied to high dimensional data or not.

1.6 Research Motivation and Outline

In order to measure falsely constructed CIs, Benjamini and Yekutieli (2005) proposed a concept of FCR as well as some FCR controlling procedures. Although such general and special FCR adjusted procedure in Definitions 1.5 and 1.8 can control FCR when we construct multiple CIs for selected parameters in many applications, it has limitations. As control of FCR does not prohibit FCP from varying, even if its average value is bounded. FCR controlling procedure cannot offer simultaneous coverage, which mean multiple CIs can cover most of the corresponding parameters. In many modern applications, simultaneous coverage is important and necessary. For example, several selected parameters for primary endpoints need to be joint so that ultimate decision can be given out in the area of clinical trials. And connection between FDR and the foregoing CIs inspire us to think how we can solve the issue of simultaneous.

In this dissertation, two types of CI are taken into consideration: unconditional CI and conditional CI, which is constructed once the corresponding variable is selected. First of all, a general unconditional CI based selective inference procedure is developed, which can be proven to control γ-FCP at a desired level under

independence. Theoretically, the result is able to be extended to positive regression dependency condition of Benjamini and Yekutieli. Then, an adjusted unconditional CI based procedure is presented to control γ-FCP under arbitrary dependence. With the approach of conditional CIs, a new conditional CI based selective inference procedure is then developed, which is able to control γ-FCP at a desired level under independence. An adjusted conditional CI-based procedure is then developed to control γ-FCP under arbitrary dependence. Finally, the proposed general procedures and conditional CI-based selective inference procedures are evaluated through extensive simulation studies under independence structure. The simulation studies are also extended to strong dependence structures such as equal correlation and several weak dependence structures such as blockwise dependence. The simulation studies are able to show that the new proposed procedures can be more reliable than alternative methods such as Benjamini and Yekutieli (2005) selective inference procedures. Also, the proposed general procedures and conditional CI-based selective inference procedures are applied on two sets of micro-array gene expression data. Compared to alternative methods such as Benjamini and Yekutieli (2005) selective inference procedures, the proposed procedure is demonstrated to be less conservative.

This dissertation is outlined as follows: Chapter 1 provides some basic concepts on multiple testing and background of selective inference. In Chapter 2, we suggest a new simultaneous coverage error measurement: γ-FCP and unconditional CI-based procedures which can control γ-FCP at a desired level. In Chapter 3, we develop a conditional CI approach. We also introduce γ-FCP controlling procedures based on conditional CIs. In Chapters 4 and 5, extensive simulation studies of new proposed methods and real data analysis are included, respectively. In Chapter 6, we summarize all of the results and findings.

CHAPTER 2

UNCONDITIONAL CI-BASED γ-FCP CONTROLLING PROCEDURES

2.1 Introduction

In this chapter, we propose a new error measurement for CI, γ-false coverage proportion (γ-FCP), and then we develop two powerful unconditional CI-based procedures. Often in applied research, unconditional CIs are constructed and reported only for parameters selected after viewing the data (Benjamini and Yekutieli, 2005; Efron, 2008; Lee et al., 2013). Benjamin and Yekutieli (2005) first point out that CIs were often only for parameters selected when constructed and reported, and such selected intervals failed to provide the assumed coverage probability. FCR was suggested as a measurement of interval coverage following selection. Benjamini and Yekutieli (2005) suggested a general procedure (BY2005a), where unconditional $1 - \alpha|S|/m$ CIs are constructed for the $|S|$ selected parameters, in which S is the selected set. Under the positive regression dependency of Benjamini and Yekutieli (2001), FCR is controlled at level α for BY2005a procedure. Meanwhile another FCR controlling procedure under general dependency is proposed (BY2005b). In most of FCR controlling procedures, only effect of selection is taken into consideration (Benjamini and Yekutieli, 2005; Benjamini, 2010). Although protection against effect of selection is sufficient in some cases, simultaneous coverage is also needed in many real applications, for instance, in clinical trials, the ultimate decision often depends on the joint outcome of several selected parameters for primary endpoints (Katsevich and Ramdas, 2018; Benjamini et al., 2019). Benjamini et al. (2019) proposed a Sidak and Bonferroni based procedure. However, the dimension is relatively small (the maximum total size of parameters is 100). So, it still need to be studied whether or not the method can be applied to high dimensional data.

Though FCR (Benjamin and Yekutieli, 2005) is widely used in both theory and practice as a error measurement for CIs. The control of FCR does not prohibit false coverage proportion from varying, even if its average value is bounded. FCR controlling procedures cannot offer simultaneous coverage, which is the probability that one or more intervals do not cover corresponding parameters at the same time. Therefore, we consider a new measurement of control, in a sense that false coverage proportion is bounded, at least with prescribed probability. We are aiming to construct multiple CIs, especially for the selected parameters, while we can offer simultaneous coverage, which implies most of CIs cover the corresponding parameters. By generalizing dual approach between multiple testings and multiple CIs (Lehmann and Romano, 2005; Benjamin and Yekutieli, 2005), we suggest a γ-FCP as an simultaneous error measurement of interval coverage following selection. Two powerful unconditional CI-based procedures are developed, which can control γ-FCP at a desired level under corresponding conditions. The corresponding theoretical results are demonstrated.

The rest of the chapter is organized as follows. Section 2.2 introduces a new error measurement, γ-FCP. The properties and behavior of γ-FCP are derived in detail. In Section 2.3, new selective inference procedures are proposed. In Section 2.4, some desired statistical properties of this procedure are demonstrated. It is proven that selective inference procedures are able to control γ-FCP at a desired level. In Section 2.5, we summarize and discuss current procedures and results together.

2.2 γ-False Coverage Proportion (γ-FCP)

We consider a new error measurement of CIs, in a sense that false coverage proportion is bounded, at least with prescribed probability. Such new error measurement, γ-false coverage proportion (γ-FCP) are hence presented. If a method control γ-FCP, then it implies most of CIs in such methods cover the corresponding parameters.

Definition 2.1 (γ-False Coverage Proportion (γ-FCP)). *γ-False coverage proportion is defined as the probability of FCP in Equation 1.1 beyond a pre-specified value, that is:*

$$\gamma\text{-}FCP = P(FCP > \gamma),$$

where γ is a pre-specified value between 0 and 1.

If a method control γ-FCP, then Definition 2.1 implies that the FCP is not greater than a pre-specified value of γ with high probability. γ-FCP can measure simultaneous coverage, which means most of CIs cover the corresponding parameters. In addition to the definition, it is important to derive the properties of γ-FCP.

2.2.1 Properties of γ-FCP

Before we move on developing new proposed procedures, some properties of the new error measurement γ-FCP are first derived. Dual to the definition of FWER, we present a simultaneous error measurement, familywise coverage rate (FWCR). Let V_{CI} be the number of constructed CIs not covering their respective parameters.

Definition 2.2 (Familywise Coverage Rate (FWCR)). *Familywise coverage rate is the probability of at least one non-covering CI is constructed, which can be denoted as*

$$FWCR = P(V_{CI} \geq 1).$$

Followed by the Definition 2.2, it is easy to find the relationship between γ-FCP and FWCR. When $\gamma = 0$, γ-FCP reduces to FWCR. It is also of interest to compare γ-FCP with FCR. We derive the properties in Lemma 2.1.

Lemma 2.1.

$$\frac{FCR - \gamma}{1 - \gamma} \leq \gamma\text{-}FCP \leq \frac{FCR}{\gamma}. \tag{2.1}$$

Proof.

$$\begin{aligned} \text{FCR} &= E(\text{FCP}|\text{FCP} > \gamma)P(\text{FCP} > \gamma) \\ &+ E(\text{FCP}|\text{FCP} \leq \gamma)P(\text{FCP} \leq \gamma) \\ &\leq \gamma\text{-FCP} + \gamma(1 - \gamma\text{-FCP}), \end{aligned} \tag{2.2}$$

which leads to

$$\frac{\text{FCR} - \gamma}{1 - \gamma} \leq \gamma\text{-FCP} \leq \frac{\text{FCR}}{\gamma},$$

with the last inequality follows from Markov's inequality. □

If a method keeps FCR at a desired level α, then Lemma 2.1 implies that such method controls γ-FCP at level $\frac{\alpha}{\gamma}$. Ratio $\frac{\alpha}{\gamma}$ might be quite large, for example, α and γ are both small. If a method can keep γ-FCP at a desired level α, then Lemma 2.1 implies that it controls FCR in the sense $\text{FCR} \leq \alpha + \gamma(1-\alpha)$. Therefore, in principle, a method that controls γ-FCP in the sense of Equation (2.1) can be used to control FCR and vice versa.

2.2.2 Discussion of γ-FCP for Some Widely Used CIs

Case 1: Constructing Unconditional $1-\alpha$ CIs for All Parameters. For a single parameter, γ-FCP equals to the probability of constructing a non-covering CI. Hence a $1-\alpha$ unconditional CI can guarantee γ-FCP $\leq \alpha$. Next, we move on to the multiple CIs. Without selection, $R_{CI} = m$, and $E(V_{CI}) \leq m\alpha$. Thus, multiple $1-\alpha$ unconditional CIs can guarantee

$$P(\frac{V_{CI}}{R_{CI} \vee 1} > \gamma) \leq \frac{E(V_{CI})}{\lfloor m\gamma \rfloor + 1} \leq \frac{\alpha}{\gamma}. \tag{2.3}$$

It is clear that constructing unconditional $1-\alpha$ CIs for all parameters in Equation 2.3 can keep γ-FCP at level α/γ. If we are able to construct $1-\gamma\alpha$ CIs for all parameters, then the corresponding γ-FCP can be controlled at level α.

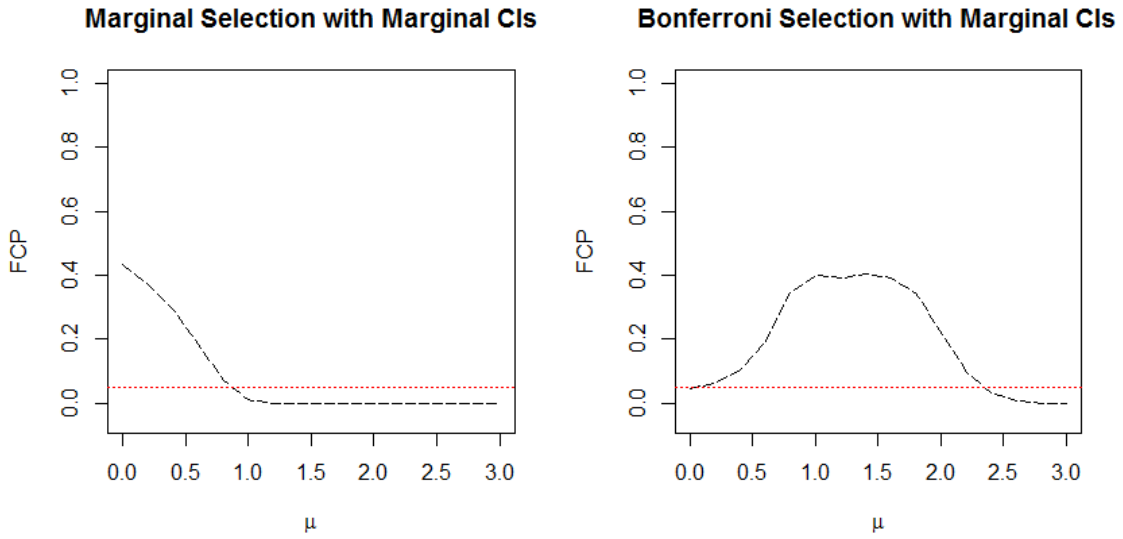

Figure 2.1 Simulation based γ-FCP(dashed line) ($\gamma = 0.10$) of unconditional 0.95 CIs for the unconditional level .05 selection schemes (left panel) and the Bonferroni level .05 selection schemes (right panel).

Case 2: Constructing $1-\alpha$ Unconditional CIs for Independently Selected Parameters. The meaning of the independent selection is that the selection criterion is independent to the data from which we use to estimate CIs. One of the examples is that we construct CIs for the parameters which are determined right before the data are set up, which can control γ-FCP at level $\frac{\alpha}{\gamma}$ as same as Equation (2.3). Another example is that we use two sets, denote $\mathbb{T}_1$ as training set for selection and $\mathbb{T}_2$ as testing set for inference, to construct multiple CIs. Then γ-FCP equals to,

$$\begin{aligned} P_{\mathbb{T}_1,\mathbb{T}_2}(FCP > \gamma) &\leq \frac{1}{\gamma} E_{\mathbb{T}_1,\mathbb{T}_2}(FCP) \\ &= \frac{1}{\gamma} E_{\mathbb{T}_1}(I(R_{CI} > 1) \frac{1}{R_{CI}} E_{\mathbb{T}_2}(V_{CI})) \\ &= \frac{1}{\gamma} E_{\mathbb{T}_1}(I(R_{CI} > 1) \frac{R_{CI}\alpha}{R_{CI}}) \leq \frac{\alpha}{\gamma}, \end{aligned}$$

with the first inequality follows from Markov's inequality. It is apparent to see that constructing unconditional $1-\alpha$ CIs for independently selected parameters can keep γ-FCP at level α/γ. If we construct $1-\gamma\alpha$ CI for independently selected parameters, then the corresponding γ-FCP can be controlled at level α.

Case 3: Constructing a $1-\gamma\alpha$ CIs for Selected Parameters. In this case, $0 \leq R_{CI} \leq m$. The selection criterion is not independent to the estimators for constructing CIs. We process to construct $1-\gamma\alpha$ CIs for selected parameters. We will illustrate the behavior of γ-FCP by a simulation example. Let $T_i \overset{i.i.d.}{\sim} N(\mu_i, 1), i = 1, 2, ..., 200$, be estimators of μ_i. For each simulation, $\mu_i = \mu$ remain fixed. This is done for five values of $\mu = 0, 0.5, 1, 2, 3$. Parameters μ_i are selected for those $|T_i| > Z_{1-0.05/2}$ (marginal selection) and $|T_i| > Z_{1-0.05/2/200}$ (Bonferroni selection). Next, for each selected parameter, $1-\gamma\alpha$ CIs are constructed.

As we can see from Figure 2.1 that $1-\gamma\alpha$ CIs for the selected ones are not able to control γ-FCP when the true parameter μ tend to be small. In other words, our existing methods of constructing CIs for the selected parameters has disadvantage regarding to the poor simultaneous coverage. It is necessary to study and develop a new procedure such that γ-FCP can be controlled at a desired level.

2.3 Unconditional CI-Based γ-FCP Controlling Procedure

Two unconditional CI-based procedures are proposed in this section. We develop unconditional CI-based γ-FCP controlling procedures under independence and dependency, respectively. We denote $\mathbf{Y} = (Y_1, ..., Y_m)$ as selection estimators and $\mathbf{T} = (T_1, ..., T_m)$ as estimator for CI construction, both for parameter $\mu = (\mu_1, ..., \mu_m)$. And the selection procedure is given by $\hat{S}(\mathbf{Y})$, and the size of the selection is $|\hat{S}(\mathbf{Y})|$.

Procedure 2.1 (Unconditional CI-Based Procedure)**.**

1. Apply the selection criterion $\hat{S}$ to $\mathbf{Y} = (Y_1, ...Y_m)$, yielding the selected set of parameters $\hat{S}(\mathbf{Y})$.

2. For each selected parameter $\mu_i, i \in \hat{S}(\mathbf{Y})$, partition $\mathbf{Y}$ into $\mathbf{Y}^{(i)}$ and Y_i, where $\mathbf{Y}^{(i)} = \mathbf{Y} \setminus \{Y_i\}$, and find

$$R_{min}(\mathbf{Y}^{(i)}) := \min_y \{|\hat{S}(\mathbf{Y}^{(i)}, Y_i = y)| : i \in \hat{S}(\mathbf{Y}^{(i)}, y)\}.$$

3. For each selected parameter $\mu_i, i \in \hat{S}(\mathbf{Y})$, construct the following unconditional CI:

$$CI_i\Big(\frac{\lfloor \gamma R_{min}(\mathbf{Y}^{(i)}) \rfloor + 1}{m}\alpha\Big). \tag{2.4}$$

Remark 2.1. *It is worth to mention that $R_{min}(\mathbf{Y}^{(i)})$ can be replaced by R_{CI} for some commonly used plausible selection methods, such as marginal selection and Bonferroni selection. Because $|\hat{S}(\mathbf{Y}^{(i)}, y)|$ assumes a single value, given $\mathbf{Y^{(i)}}$ for values $Y_i = y$ such that parameter μ_i is selected, $i = 1, ..., m$. But there exists some exceptions, such as Benjamini and Hochberg (2000) and Benjamini Krieger and Yekutieli (2006). For these exceptions, $R_{min}(\mathbf{Y}^{(i)}) < R_{CI}$.*

Incorporating R_{CI} into Procedure 2.1, the Equation 2.4 takes on a very simple form. Definition 2.1 immediately implies that the width of such selective CIs decreases as number of selected parameter increases and increases as number of total considered parameter increases. The length is same as Bonferroni adjusted CI if $R_{min}(\mathbf{Y}^{(i)}) < 1/\gamma$. In addition, the average width of CIs is shorter than Bonferroni adjusted CI if $R_{min}(\mathbf{Y}^{(i)}) \geq 1/\gamma$.

2.3.1 Theoretical Results

In this section, we first prove our proposed procedure in Procedure 2.1 ia able to control γ-FCP at level α when $(Y_i, T_i), i = 1, ..., m$ are independent. Next, the result of Procedure 2.1 is extended to control γ-FCP when $(Y_i, T_i), i = 1, ..., m$ are positive regression dependent on subsets.

Theorem 2.1. *If $(Y_i, T_i), i = 1, ..., m$ are independent, then for any selection procedure $\hat{S}(\mathbf{Y})$, the γ-FCP adjusted selective CIs in Procedure 2.1 enjoys γ-FCP $\leq \alpha$.*

Proof. Recall the Definition 2.1 of γ-FCP,

$$\begin{aligned}\gamma\text{-FCP} = P\Big(\frac{V_{CI}}{R_{CI} \vee 1} > \gamma\Big) &= P(V_{CI} \geq \lfloor \gamma R_{CI} \rfloor + 1) \\ &\leq E\Big(\frac{V_{CI}}{\lfloor \gamma R_{CI} \rfloor + 1}\Big).\end{aligned} \tag{2.5}$$

The inequality in Equation (2.5) follows from Markov's inequality. Now we know that

$$\begin{aligned}E\Big(\frac{V_{CI}}{\lfloor \gamma R_{CI} \rfloor + 1}\Big) &= \sum_{i=1}^{m}\sum_{r=1}^{m} \frac{1}{\lfloor \gamma r \rfloor + 1} P\Bigg(i \in \hat{S}, R_{CI} = r, \\ &\qquad \mu_i \notin CI_i\Big(\frac{\lfloor \gamma R_{min}(\mathbf{Y}^{(i)}) \rfloor + 1}{m}\alpha\Big)\Bigg) \\ &= \sum_{i=1}^{m}\sum_{r=1}^{m}\sum_{k=1}^{m} \frac{1}{\lfloor \gamma r \rfloor + 1} P\Bigg(i \in \hat{S}, R_{CI} = r, R_{min}(\mathbf{Y}^{(i)}) = k, \\ &\qquad \mu_i \notin CI_i\Big(\frac{\lfloor \gamma k \rfloor + 1}{m}\alpha\Big)\Bigg) \\ &\leq \sum_{i=1}^{m}\sum_{k=1}^{m} \frac{1}{\lfloor \gamma k \rfloor + 1} P\Bigg(i \in \hat{S}, R_{min}(\mathbf{Y}^{(i)}) = k, \\ &\qquad \mu_i \notin CI_i\Big(\frac{\lfloor \gamma k \rfloor + 1}{m}\alpha\Big)\Bigg) \\ &\leq \sum_{i=1}^{m}\sum_{k=1}^{m} \frac{1}{\lfloor \gamma k \rfloor + 1} P\Bigg(R_{min}(\mathbf{Y}^{(i)}) = k, \mu_i \notin CI_i\Big(\frac{\lfloor \gamma k \rfloor + 1}{m}\alpha\Big)\Bigg).\end{aligned}$$

The first inequality holds since $R_{min}(\mathbf{Y^{(i)}}) \leq R_{CI}$ for each value of $\mathbf{Y^{(i)}}$ and Y_i such that μ_i is selected. The second inequality follows from dropping the condition $i \in \hat{S}$. Then, due to the condition of independence,

$$\begin{aligned}E\Big(\frac{V_{CI}}{\lfloor \gamma R_{CI} \rfloor + 1}\Big) &\leq \sum_{i=1}^{m}\sum_{k=1}^{m} \frac{1}{\lfloor \gamma k \rfloor + 1} P(R_{min}(\mathbf{Y}^{(i)}) = k) P\Bigg(\mu_i \notin CI_i\Big(\frac{\lfloor \gamma k \rfloor + 1}{m}\alpha\Big)\Bigg) \\ &\leq \frac{\alpha}{m}\sum_{i=1}^{m}\sum_{k=1}^{m} P(R_{min}(\mathbf{Y}^{(i)}) = k) = \alpha.\end{aligned}$$

The second inequality is due to the marginal coverage. Hence the desired result follows. □

Note that the condition $(Y_i, T_i), i = 1, ..., m$ are independent, can be generalized as T_i is independent to $\mathbf{Y}^{(i)}$, for all $i = 1, ..., m$. The adjusted level in Procedure 2.1 is sufficient to control γ-FCP at a desired level α. Such increase is important when one can not only characterize the effect of selection but also guarantee the simultaneous coverage is taken into consideration as well. Procedure 2.1 is proven to keep γ-FCP at level α under independence. In real science, there always exists some dependence within the data. Thus we now want to discuss whether or not our new proposed procedure in Procedure 2.1 can control γ-FCP at a desired level under the condition (Y_i, T_i) possessing PRDS. Recall the Definition 1.6 and Definition 1.7, we have the following results.

Theorem 2.2. *If the components of $(\mathbf{Y}^{(i)}, T_i)$ are PRDS, for $i = 1, ..., m$, and the selection criterion $\hat{S}$ and the CIs are concordant, then the γ-FCP adjusted selective CIs in Definition 2.1 enjoys γ-FCP $\leq \alpha$.*

Proof. Recall the proof of Theorem 2.1, we have,

$$\begin{aligned}
\gamma\text{-FCP} &\leq E\Big(\frac{V_{CI}}{\lfloor \gamma R_{CI} \rfloor + 1}\Big) \\
&\leq \sum_{i=1}^{m}\sum_{k=1}^{m} \frac{1}{\lfloor \gamma k \rfloor + 1} P\left(R_{min}(\mathbf{Y^{(i)}}) = k, \mu_i \notin CI_i\Big(\frac{\lfloor \gamma k \rfloor + 1}{m}\alpha\Big)\right) \\
&\leq \frac{\alpha}{m}\sum_{i=1}^{m}\sum_{k=1}^{m} P\left(R_{min}(\mathbf{Y^{(i)}}) = k \middle| \mu_i \notin CI_i\Big(\frac{\lfloor \gamma k \rfloor + 1}{m}\alpha\Big)\right).
\end{aligned}$$

The last inequality is due to the marginal coverage in Definition 2.1. Note that

$$\begin{aligned}
&\sum_{k=1}^{m} P\left(R_{min}(\mathbf{Y^{(i)}}) = k \middle| \mu_i \notin CI_i\Big(\frac{\lfloor \gamma k \rfloor + 1}{m}\alpha\Big)\right) \\
&= \sum_{k=1}^{m}\left[P\left(R_{min}(\mathbf{Y^{(i)}}) \geq k \middle| \mu_i \notin CI_i\Big(\frac{\lfloor \gamma k \rfloor + 1}{m}\alpha\Big)\right)\right. \\
&\left. - \sum_{k=1}^{m} P\left(R_{min}(\mathbf{Y^{(i)}}) \geq k+1 \middle| \mu_i \notin CI_i\Big(\frac{\lfloor \gamma k \rfloor + 1}{m}\alpha\Big)\right)\right],
\end{aligned}$$

which is

$$
\begin{aligned}
&\leq \sum_{k=1}^{m}\left[P\left(R_{min}(\mathbf{Y^{(i)}}) \geq k \middle| \mu_i \notin CI_i\Big(\frac{\lfloor \gamma k \rfloor + 1}{m}\alpha\Big)\right)\right. \\
&\left. - \sum_{k=1}^{m} P\left(R_{min}(\mathbf{Y^{(i)}}) \geq k+1 \middle| \mu_i \notin CI_i\Big(\frac{\lfloor \gamma (k+1) \rfloor + 1}{m}\alpha\Big)\right)\right] \\
&= P\left(R_{min}(\mathbf{Y^{(i)}}) \geq 1 \middle| \mu_i \notin CI_i\Big(\frac{\alpha}{m}\Big)\right) \\
&- P\left(R_{min}(\mathbf{Y^{(i)}}) \geq m+1 \middle| \mu_i \notin CI_i\Big(\frac{\lfloor \gamma (m+1) \rfloor + 1}{m}\alpha\Big)\right) = 1.
\end{aligned}
$$

The inequality holds because of the following argument. Without loss of generality, assume $\{\mathbf{Y}^{(i)} : k \leq R_{min}(\mathbf{Y}^{(i)})\}$ is an increasing set. Then by Definition 1.7, with condition $\hat{S}$ and CIs are concordant, $\{T_i : \mu_i \notin CI_i(\alpha)\}$ is also an increasing set, which in turn can be expressed as an interval $T_i \geq a_i$. By monotone property, which $\alpha \geq \alpha'$ implies that $CI(\alpha) \subseteq CI(\alpha')$. Then we can find that $CI_i(\frac{\lfloor \gamma(k+1) \rfloor + 1}{m}\alpha) \subseteq CI_i(\frac{\lfloor \gamma k \rfloor + 1}{m}\alpha)$, which in turn can be expressed as $T_i \geq b_i'$ and $T_i \geq b_i$. Then $b_i' \geq b_i$. Thus by the PRDS of $(\mathbf{Y}^{(i)}, T_i)$, the inequality follows. Hence

$$\gamma\text{-FCP} \leq \frac{\alpha}{m}\sum_{i=1}^{m} 1 = \alpha.$$

The result in Theorem 2.2 holds for (Y_i, T_i) possessing PRDS property. □

2.4 Modified Unconditional CI-Based Procedure

The results in Theorems 2.1 and 2.2 holds for independence and PRDS, respectively. We now discuss parameter estimators possessing any arbitrary dependence of (Y_i, T_i).

Procedure 2.2 (Modified Unconditional CI-based Procedure under Arbitrary Dependence)**.**

Let $c_{\gamma,m} = \sum_{i=1}^{\lfloor \gamma(m-1) \rfloor + 2} \frac{1}{i}$*, then for each selected parameter* $\mu_i, i \in \hat{S}(\mathbf{Y})$*, construct the following unconditional CI:*

$$CI_i\Big(\frac{\lfloor \gamma R_{min}(\mathbf{Y}^{(i)})\rfloor + 1}{mc_{\gamma,m}}\alpha\Big). \tag{2.6}$$

Theorem 2.3. *For any monotone unconditional CIs, any selection procedure $\hat{S}(\mathbf{Y})$, and any dependence of (Y_i, T_i), for $i = 1, ..., m$, the γ-FCP of Procedure 2.2 is bounded by α.*

Proof. Recall the proof of Theorem 2.1, we have

$$\begin{aligned}\gamma\text{-FCP} &\le E\Big(\frac{V_{CI}}{\lfloor \gamma R_{CI}\rfloor + 1}\Big)\\ &\le \sum_{i=1}^{m}\sum_{k=1}^{m}\frac{1}{\lfloor \gamma k\rfloor + 1}P\bigg(R_{min}(\mathbf{Y^{(i)}}) = k, \mu_i \notin CI_i\Big(\frac{\lfloor \gamma k\rfloor + 1}{mc_{\gamma,m}}\alpha\Big)\bigg).\end{aligned}$$

Note that

$$\begin{aligned}&\sum_{k=1}^{m}\frac{1}{\lfloor \gamma k\rfloor + 1}P\bigg(R_{min}(\mathbf{Y^{(i)}}) = k, \mu_i \notin CI_i\Big(\frac{\lfloor \gamma k\rfloor + 1}{mc_{\gamma,m}}\alpha\Big)\bigg)\\ =&\sum_{k=1}^{m}\frac{1}{\lfloor \gamma k\rfloor + 1}P\bigg(R_{min}(\mathbf{Y^{(i)}}) \ge k, \mu_i \notin CI_i\Big(\frac{\lfloor \gamma k\rfloor + 1}{mc_{\gamma,m}}\alpha\Big)\bigg)\\ &-\sum_{k=1}^{m}\frac{1}{\lfloor \gamma k\rfloor + 1}P\bigg(R_{min}(\mathbf{Y^{(i)}}) \ge k+1, \mu_i \notin CI_i\Big(\frac{\lfloor \gamma k\rfloor + 1}{mc_{\gamma,m}}\alpha\Big)\bigg)\\ \le&\sum_{k=1}^{m}\frac{1}{\lfloor \gamma k\rfloor + 1}P\bigg(R_{min}(\mathbf{Y^{(i)}}) \ge k, \mu_i \notin CI_i\Big(\frac{\lfloor \gamma k\rfloor + 1}{mc_{\gamma,m}}\alpha\Big)\bigg)\\ &-\sum_{k=1}^{m}\frac{1}{\lfloor \gamma (k+1)\rfloor + 1}P\bigg(R_{min}(\mathbf{Y^{(i)}}) \ge k+1, \mu_i \notin CI_i\Big(\frac{\lfloor \gamma k\rfloor + 1}{mc_{\gamma,m}}\alpha\Big)\bigg)\\ \le& P\bigg(\mu_i \notin CI_i\Big(\frac{\alpha}{mc_{\gamma,m}}\Big)\bigg)\\ &+\sum_{k=2}^{m}\frac{1}{\lfloor \gamma k\rfloor + 1}\bigg[P\bigg(\mu_i \notin CI_i\Big(\frac{\lfloor \gamma k\rfloor + 1}{m}\alpha\Big)\bigg) - P\bigg(\mu_i \notin CI_i\Big(\frac{\lfloor \gamma (k-1)\rfloor + 1}{mc_{\gamma,m}}\alpha\Big)\bigg)\bigg]\\ =&\sum_{k=1}^{m-1}\Big(\frac{1}{\lfloor \gamma k\rfloor + 1} - \frac{1}{\lfloor \gamma (k+1)\rfloor + 1}\Big)P\bigg(\mu_i \notin CI_i\Big(\frac{\lfloor \gamma k\rfloor + 1}{mc_{\gamma,m}}\alpha\Big)\bigg)\\ &+\frac{1}{\lfloor \gamma m\rfloor + 1}P\bigg(\mu_i \notin CI_i\Big(\frac{\lfloor \gamma m\rfloor + 1}{mc_{\gamma,m}}\alpha\Big)\bigg), s\end{aligned}$$

which is

$$\leq \frac{\alpha}{mc_{\gamma,m}}\left[1+\sum_{k=1}^{m-1}\Big(1-\frac{\lfloor\gamma k\rfloor+1}{\lfloor\gamma(k+1)\rfloor+1}\Big)\right].$$

The first inequality follows from $\lfloor\gamma k\rfloor \leq \lfloor\gamma(k+1)\rfloor$, for $k = 1,..,m$. The third inequality holds due to marginal coverage of unconditional CI. Thus

$$\gamma\text{-FCP} \leq \sum_{i=1}^{m}\frac{\alpha}{mc_{\gamma,m}}\left[1+\sum_{k=1}^{m-1}\Big(1-\frac{\lfloor\gamma k\rfloor+1}{\lfloor\gamma(k+1)\rfloor+1}\Big)\right] = \frac{\alpha}{c_{\gamma,m}}\left(1+\sum_{k=1}^{m-1}\Big(1-\frac{\lfloor\gamma k\rfloor+1}{\lfloor\gamma(k+1)\rfloor+1}\Big)\right).$$

Define $k_i := max\{1 \leq k \leq m-1 : \lfloor\gamma k\rfloor = i\}$, $k_{-1} = 0$, and $i_{max} := \lfloor\gamma(m-1)\rfloor$. Then,

$$\sum_{k=1}^{m-1}\Big(1-\frac{\lfloor\gamma k\rfloor+1}{\lfloor\gamma(k+1)\rfloor+1}\Big) = \sum_{i=0}^{i_{max}}\sum_{k=k_{i-1}+1}^{k_i}\Big(1-\frac{\lfloor\gamma k\rfloor+1}{\lfloor\gamma(k+1)\rfloor+1}\Big)$$

For any i, we have

$$\begin{aligned}\sum_{k=k_{i-1}+1}^{k_i}\Big(1-\frac{\lfloor\gamma k\rfloor+1}{\lfloor\gamma(k+1)\rfloor+1}\Big) &= \sum_{k=k_{i-1}+1}^{k_i-1}\Big(1-\frac{\lfloor\gamma k\rfloor+1}{\lfloor\gamma(k+1)\rfloor+1}\Big)+\Big(1-\frac{\lfloor\gamma k_i\rfloor+1}{\lfloor\gamma(k_i+1)\rfloor+1}\Big)\\ &= 0+\frac{1}{i+2} = \frac{1}{i+2}\end{aligned}$$

And therefore

$$\sum_{k=1}^{m-1}\Big(1-\frac{\lfloor\gamma k\rfloor+1}{\lfloor\gamma(k+1)\rfloor+1}\Big) = \sum_{i=0}^{i_{max}}\Big(\frac{1}{i+2}\Big) = \sum_{i=2}^{i_{max}+2}\Big(\frac{1}{i}\Big)$$

Hence

$$\gamma\text{-FCP} \leq \frac{\alpha}{c_{\gamma,m}}\left(\sum_{i=1}^{\lfloor\gamma(m-1)\rfloor+2}\frac{1}{i}\right) = \alpha$$

□

The immediate corollary is that for any monotone unconditional CIs, any selection procedure $\hat{S}(\mathbf{Y})$, and any dependence of (Y_i, T_i), for $i = 1,...,m$, the γ-FCP of Procedure 2.1 is bounded by $\alpha c_{\gamma,m}$, where $c_{\gamma,m} = \sum_{i=1}^{\lfloor\gamma(m-1)\rfloor+2}\frac{1}{i}$.

2.5 Conclusion

We have proposed a new error measurement of CIs, γ-FCP in Definition 2.1. Two unconditional CI-based γ-FCP controlling procedures are present in Procedure 2.1 and Procedure 2.2. It is sufficient to prove that such Procedure 2.1 can keep γ-FCP at a desired level under independence/PRDS and Procedure 2.2 can keep γ-FCP at a desired level under arbitrary dependence. Besides unconditional CI-based procedures, in the next chapter, we develop procedures, based on the conditional CIs (Weinstein et al., 2013; Benjamini et al., 2019), in which conditional CIs provide shorter width of CIs.

CHAPTER 3

CONDITIONAL CI-BASED γ-FCP CONTROLLING PROCEDURES

3.1 Introduction

Weinstein et al (2013) developed conditional CIs and suggested three methods to offer FCR control. Benjamini et al. (2019) suggested a method of constructing CIs to control simultaneous coverage over selected parameters. In this chapter, we develop conditional CI-based γ-FCP controlling procedures, which can take the effect of selection into consideration and offer simultaneous coverage over selected parameters as well. Such new conditional CI approach is based on conditional CI, which is constructed for a parameter if it is selected. A 95% conditional CIs offers *P(conditonal CI covers its parameter | the parameter is selected)* $\geq$ 0.95. However, conditional CI is quite challenging to obtain. We first illustrate the conditional CIs with an example. Let $T \sim N(\mu, 1)$ be the estimator for μ. We are interested in the value of parameter only if T is large enough, i.e., $T \geq 1.96$. Since only $T|T \geq 1.96$ is observed, such conditional T no longer follows a normal distribution $N(\mu, 1)$. Let $f_\mu(t)$ and $F_\mu(t)$ be the probability density function and cumulative distribution function of T. Then the conditional probability density function of $T|T \geq 1.96$ is different from unconditional probability density function of T. Hence, we are not able to directly use unconditional probability density function to estimate such conditional CIs. In fact, it is appropriate to use the conditional probability density function to estimate conditional CIs.

The rest of the chapter is organized as follows. In Section 3.2, we address the preliminaries of conditional CI. Then, in Section 3.3, new conditional CI-based γ-FCP controlling procedure is developed. In Section 3.4, a modified conditional CI-based procedure is developed, which is proven to keep γ-FCP at a desired level under arbitrary dependence. We also derive the properties of the new conditional

CI-based procedures in Section 3.5, as well as the discussion of conditional CI. Finally, we summarize the all of the procedures and results in Section 3.6.

3.2 Preliminaries

We first introduce a conditional error measurement, which complement to the unconditional error measurement we introduce in Section 2.2. Let $\hat{S}$ be the index set of selected parameters. For any subset $S \subseteq \{1, ..., m\}$, where selection criterion is $\hat{S} = S$.

Definition 3.1 (Conditional γ-False Coverage Proportion (γ-cFCP$_S$))**.** *Conditional γ-false coverage proportion is defined as, condition on a selection rule $\hat{S} = S$, the probability of FCP beyond a pre-specified value, where FCP is the ratio of non-covering CIs among all constructed cCIs, that is,*

$$\gamma\text{-}cFCP_S = P\Big[\frac{V_{CI}}{R_{CI} \vee 1} > \gamma | \hat{S} = S\Big].$$

γ-cFCP is an application tool for us to develop conditional CI-based γ-FCP controlling procedures. If a procedure control γ-$cFCP_S$ at a desired level, then it implies control of γ-FCP.

3.3 Conditional CI-Based γ-FCP Controlling Procedure

Denote $\mathbf{Y} = (Y_1, ..., Y_m)$ as selection statistics and $\mathbf{T} = (T_1, ..., T_m)$ as the estimator for CIs construction, both for parameter $\mu = (\mu_1, ..., \mu_m)$. And the selection procedure is given by $\hat{S}(\mathbf{Y})$, and the total size of the selection is $|\hat{S}(\mathbf{Y})|$.

Procedure 3.1 (Conditional CI-Based Procedure)**.**

1. Apply the selection criterion $\hat{S}$ to $\mathbf{Y}$, yielding the selected set of parameters $\hat{S}(\mathbf{Y})$. Note that $|S|$ is the number of selection conditional on $\hat{S}(\mathbf{Y}) = S$.

2. For each selected parameter $\mu_i, i \in \hat{S}(\mathbf{Y})$, construct the following conditional CIs: $cCI_i(u)$, where u satisfies the function,

$$\sum_{j=0}^{\lfloor \gamma|S| \rfloor} \binom{|S|}{j} u^j (1-u)^{|S|-j} = 1 - \alpha. \tag{3.1}$$

3.3.1 Theoretical Results

We prove our proposed procedure in Procedure 3.1 can control γ-FCP when $(Y_i, T_i), i = 1, ..., m$ are independent.

Theorem 3.1. *If $(Y_i, T_i), i = 1, ..., m$ are independent, conditional on the selection $S(\hat{\mathbf{Y}}) = S$, we construct exact cCI(u) for μ_i in Definition 3.2, then the γ-cFCP$_S$ is bounded by α.*

Proof. Let $V_S(u)$ be the number of non-covering constructed cCI(u) for the selected parameters, conditional on $\hat{S}(\mathbf{Y}) = S$. Under condition (Y_i, T_i) are independent for $i = 1, ..., m,$, $V_S(u)$ follows a binomial distribution. By Definition 3.1, to control γ-cFCP$_S$ at level α as

$$\begin{aligned} \gamma\text{-cFCP} &= Pr(V_{|S|} \geq \lfloor \gamma|S| \rfloor + 1 | S = S) \\ &= \sum_{j=\lfloor \gamma|S| \rfloor + 1}^{|S|} \binom{|S|}{j} u^j (1-u)^{|S|-j} = 1 - \sum_{j=0}^{\lfloor \gamma|S| \rfloor} \binom{|S|}{j} u^j (1-u)^{|S|-j} = \alpha. \end{aligned}$$

By using double expectation, γ-FCP can be kept at level α. □

3.4 Modified Conditional CI-Based Procedure

The results in Theorem 3.1 holds for independence. We now discuss parameter estimators possessing any arbitrary dependence of (Y_i, T_i). Though the independence case is relatively complicated, the dependence case is quite intuitive.

Procedure 3.2 (Modified Procedure 3.1 for Dependence)**.**

For each selected parameter $\mu_i, i \in \hat{S}(\mathbf{Y})$, construct the following conditional

CIs:

$$cCI_i\Big(\frac{\lfloor\gamma|S|\rfloor+1}{|S|}\alpha\Big).$$

Theorem 3.2. *For any selection procedure $\hat{S}(\mathbf{Y})$, and any dependence structure of (Y_i, T_i), for $i = 1, ..., m$, the γ-cFCP$_S$ adjusted selective conditional CIs in Definition 3.2 enjoys γ-cFCP$_S \leq \alpha$.*

Proof. By the Procedure 3.1, we have

$$\begin{aligned}
\gamma\text{-}cFCP_S &= P\Big(\frac{V_S}{R_S \vee 1}|\hat{S} = S\Big) \\
&= P(V_S \geq \lfloor\gamma|S|\rfloor + 1|\hat{S} = S) \\
&\leq \frac{1}{\lfloor\gamma|S|\rfloor+1}E(V_S|\hat{S} = S) \\
&= \frac{1}{\lfloor\gamma|S|\rfloor+1}\sum_{i\in S}P(\mu_i \notin cCI_i|\hat{S} = S) \\
&\leq \frac{\lfloor\gamma|S|\rfloor+1}{\lfloor\gamma|S|\rfloor+1}\alpha \leq \alpha
\end{aligned}$$

□

If the number of $|S|$ of the selected parameters is less than $1/\gamma$, then it is enough to construct conditional CIs at level $1 - \alpha/|S|$.

3.5 Discussion

The following the questions are discussed: (1) What is the meaning that (Y_i, T_i), for $i = 1, ..., m$, are independent? (2) How can we obtain u? How can we understand u? (3) Which procedure can be more powerful, Procedure 3.1 or Procedure 3.2, in terms of shorter CI width?

Explanation of the condition: (Y_i, T_i), for $i = 1, ..., m$, are independent Since T_i is the estimator of CIs construction for parameter μ_i, such condition can be replaced by $T_i|_{\hat{S}=S}$ that are independent to each other for all $i \in S$. Such condition may be not

easy to guarantee. For example, when $Y_i \neq T_i$ as the selection statistics for parameter μ_i, such condition can be simplified as T_i is independent to $\mathbf{Y}$, and T_i is independent for $i \in S$. When $Y_i = T_i$, with simple selection rule $T_i > c$, where c is a constant, we can find

$$T_i|_{\hat{S}=S} = \begin{cases} T_i|T_i > c & \text{if } i \in S \\ T_i|T_i \leq c & \text{if } i \notin S \end{cases}$$

If we can find conditional $T_i|T_i > c$, then condition can be updated as conditional $T_i|T_i > c$ is independent for $i \in S$.

Discussion about u We move on to discuss about u. From Equation (3.1), we can tell u is a value which is affected by the size of selection $|S|$, α, and γ, which can be denoted as $u = u(\alpha, \gamma, |S|)$. We use stochastic ordering as application tools to derive (i) monotonicity of $u(\alpha, \gamma, |S|)$ in α and γ, respectively; and (ii) limitation of $u(\alpha, \gamma, |S|)$ when $|S|$ is very large.

Lemma 3.1. *$u(\alpha, \gamma, |S|)$ is an increasing function in α.*

The details of proofs are in Appendix.

Lemma 3.2. *$u(\alpha, \gamma, |S|)$ is a nondecreasing function in γ.*

The details of proofs are in Appendix. The foregoing conditional CIs procedure is monotone in α.:

$$\alpha \geq \alpha' \text{ implies that } cCI_i(\alpha) \subseteq cCI_i(\alpha'). \tag{3.2}$$

Lemma 3.3. *If* $V_n \sim Bin(n, u_n)$, *the corresponding* γ*-cFCP*$= P(V_n \geq \gamma n) = \alpha$, *where* $0 < \alpha < 1$, *then* $\lim_{n\to\infty} u_n = \gamma$.

The details of proofs are in the Appendix. From Lemma 3.3, we derive an fact that u tends to be a constant when the size of selection $|S|$ is large, and such constant is γ. With this strong lemma, it implies that we can construct cCI(γ) for a large scale of selected parameters, which keeps γ-FCP at a desired level.

Discussion about u versus $\frac{\lfloor\gamma|S|\rfloor+1}{|S|}\alpha$ We discuss about whether or not $u(\alpha, \gamma, |S|) > \frac{\lfloor\gamma|S|\rfloor+1}{|S|}\alpha$, in a sense that conditional CI-based procedure under independence has shorter CIs width than under dependence. Numerical studies are performed to compare between $u(\alpha, \gamma, |S|)$ and $\frac{\lfloor\gamma|S|\rfloor+1}{|S|}\alpha$. Our numerical studies shows the ratio $r = \frac{u(\alpha,\gamma,|S|)}{\frac{\lfloor\gamma|S|\rfloor+1}{|S|}\alpha}$ versus α, γ or $|S|$, respectively. The setting and figures of numerical study are in Appendix. The numerical studies shows the fact that $u(\alpha, \gamma, |S|)$ is not smaller than $\frac{\lfloor\gamma|S|\rfloor+1}{|S|}\alpha$, which further implies Procedure 3.1 has equal or shorter CIs width than Procedure 3.2.

Conditional CI for Selected Parameters To implement the procedure in Sections 3.3 and 3.4, we discuss the methods about constructing conditional CIs. Weinstein, Fithian and Benjamini (2013) introduced a powerful method to construct conditional CIs for one-sample problem. We develop a method to construct conditional CIs for two-sample problem. In applied research, two-sample design is commonly used to determine whether two population means are equal or not (Fithian et al, 2015; Tian et al, 2018). The construction of such design can be shown with a classic example, $\{X_{1j} : 1 \leq j \leq n_1\}$, and $\{X_{2j} : 1 \leq j \leq n_2\}$ be two independent random samples, such as,

$$X_{1j} \overset{\text{i.i.d}}{\sim} N(\mu_1, \sigma_1^2),$$
$$X_{2j} \overset{\text{i.i.d}}{\sim} N(\mu_2, \sigma_2^2).$$

We want to construct conditional CIs for $\mu_1 - \mu_2$, which guarantee the conditional coverage, that is,

$$P(\mu_1 - \mu_2 \in CI(\alpha) | i \in S) \geq 1 - \alpha,$$

where S is the selected set of parameters. We choose Y as selection estimator and T as the estimator for CIs, which are defined as

$$Y = \bar{X}_1 + \bar{X}_2,$$
$$T = \bar{X}_1 - \bar{X}_2.$$

The selection rule is based on Y. For the convenience of discussion, let $\sigma_1^2 = \sigma_1^2 = 1$. (Y, T) follows a bivariate normal distribution,

$$\begin{pmatrix} Y \\ T \end{pmatrix} \sim \mathcal{N} \left[\begin{pmatrix} \mu_1 + \mu_2 \\ \mu_1 - \mu_2 \end{pmatrix}, \begin{pmatrix} \frac{1}{n_1} + \frac{1}{n_2} & 0 \\ 0 & \frac{1}{n_1} + \frac{1}{n_2} \end{pmatrix} \right].$$

Since $Cov(Y, T) = 0$, we can conclude that Y is independent to T. And conditional CIs for $\mu_1 - \mu_2 | Y$ is same to the unconditional CIs for $\mu_2 - \mu_2$.

$$P(\mu_1 - \mu_2 \in CI(\alpha) | Y) = P(\mu_1 - \mu_2 \in CI(\alpha)).$$

In Appendix, we construct conditional CI when Y is dependent to T.

3.6 Conclusion

We have proposed two new conditional CI-based γ-FCP controlling procedures. So far, there are four γ-FCP controlling procedures are proposed, in two types: unconditional and conditional CIs. We summarize the procedures in Table 3.1. In Chapter 4, we conduct extensive simulation study about the proposed procedures. Such studies are intend to dig about the performance of the new procedures. In

Table 3.1 Summary of Unconditional CI-Based Procedures and Conditional CI-Based Procedures

Procedure	Type of CI	Required Condition	Adjusted Level
1	unconditional CI	independence/PRDS	$\frac{\lfloor \gamma \lvert S \rvert \rfloor + 1}{m}\alpha$
2	unconditional CI	Dependence	$\frac{\lfloor \gamma \lvert S \rvert \rfloor + 1}{m c_{\gamma,m}}\alpha$
3	conditional CI	Independence	$u(\alpha, \gamma, \lvert S \rvert)$
4	conditional CI	Dependence	$\frac{\lfloor \gamma \lvert S \rvert \rfloor + 1}{\lvert S \rvert}\alpha$

Note that $c_{\gamma,m} = \sum_{i=1}^{\lfloor \gamma(m-1) \rfloor + 2} \frac{1}{i}$.

Chapter 5, we study two real data sets while we apply our newly proposed procedures. The analysis of real data provides us an intuitive view of pros and cons about our newly proposed procedures. A possible future work is to develop weighted methods for constructing conditional CIs. There are three reasons that we are interested in weighted methods: (1) to incorporate previous knowledge about the parameters; (2) to construct powerful CI; and (3) to represent difference in importance of estimators. More details will be discussed in the Chapter 6.

CHAPTER 4

SIMULATION STUDIES

4.1 Introduction

In this chapter, we perform extensive simulation studies for our proposed procedures. Simulation studies are designed to (1) evaluate effect of nonzero proportion, selection level and correlation coefficient under independence, while we apply our proposed procedures in terms of γ-FCP control and average width of CIs; and (2) apply to strong dependence such as equal correlation and several weak dependence such as block-wise dependence. We explore the performance of the proposed procedures in four different situations: (i) under independence for one sample case; (ii) under independence for two sample case; (iii) under dependence for one sample case; and (iv) under dependence for two sample case. In Section 4.2, we clearly introduce the methods of constructing CI for unconditional CI-based γ-FCP controlling procedures in Chapter 2 (Procedures 1 and 3 refers to independence and dependence) and conditional CI-based γ-FCP controlling procedures in Chapter 3 (Procedures 2 and 4 refers to independence and dependence) and the existing FCR controlling procedures (Procedures 5 and 6 refers to independence and dependence, which were suggested by Benjamini and Yekutieli, 2005). From Section 4.3 to Section 4.6, we show the performance of the proposed procedures in four different situations: (i) - (iv). In Section 4.7, a concluding remark is given.

4.2 Preliminary

We introduce the methods of constructing CIs for one sample case. X_i follows $N(\mu_i, \sigma)$ with unknown parameter μ_i and known parameter σ, $\bar{X}_i$ are used to estimate μ_i. To simplify the case, we assume σ^2 as the sample size of X_i. Therefore $\bar{X}_i \sim N(\mu_i, 1)$.

Table 4.1 Summary of Various Methods of Constructing CIs for One Sample Case

Procedure	CIs for One Sample Case
1	$(\bar{x}_i - Z_{\frac{\lfloor \gamma R \rfloor + 1}{2m}\alpha}, \bar{x}_i + Z_{\frac{\lfloor \gamma R \rfloor + 1}{2m}\alpha})$
2	$(f_l(\bar{x}_i, u_{\alpha,\gamma,R}), f_u(\bar{x}_i, u_{\alpha,\gamma,R}))$
3	$(\bar{x}_i - Z_{\frac{\lfloor \gamma R \rfloor + 1}{2mc_{\gamma,m}}\alpha}, \bar{x}_i + Z_{\frac{\lfloor \gamma R \rfloor + 1}{2mc_{\gamma,m}}\alpha})$
4	$(f_l(\bar{x}_i, \frac{\lfloor \gamma R \rfloor + 1}{R}\alpha), f_u(\bar{x}_i, \frac{\lfloor \gamma R \rfloor + 1}{R}\alpha))$
5	$(\bar{x}_i - Z_{\frac{R}{2m}\alpha}, \bar{x}_i + Z_{\frac{R}{2m}\alpha})$
6	$(\bar{x}_i - Z_{\frac{R}{2mc^*_{\gamma,m}}\alpha}, \bar{x}_i + Z_{\frac{R}{2mc^*_{\gamma,m}}\alpha})$

Given the proposed procedures and the existing selective inference procedures, there are six corresponding methods of constructing CIs for μ_i for one sample case. In Table 4.1, we list six methods of CIs for one sample case corresponding to the selective inference procedures. We denote that (1) x_i as the observed value of X_i; (2) R is the number of selected parameters μ_i, i.e. $R = 1, 2, ..., m$, where m is total parameter size; (3) γ is a pre-specified value between 0 and 1; (4) $u_{\alpha,R,\gamma} := \text{argmin}_{0 \le u \le 0.5}\{F_u(\lfloor \gamma R \rfloor) = 1 - \alpha\}$, $F_u(\cdot)$ is the cumulative distribution function of a binomial distribution $Bin(R, u)$, which is introduced in Chapter 3; (5) $c_{\gamma,m} = \sum_{i=1}^{\lfloor \gamma(m-1) \rfloor + 2} \frac{1}{i}$, which is introduced in Chapter 2; (6) $c^*_{\gamma,m} = \sum_{i=1}^{m} 1/i$, which was proposed by Benjamini and Yekutieli (2005); and (7) $f_l(\bar{x}_i, \alpha^*)$ and $f_u(\bar{x}_i, \alpha^*)$ are lower and upper conditional CI bond of μ_i, which we modify from Weinstein, Fithian and Benjamini (2013), α^* is the adjusted α level in conditional CI-based procedures.

Next, we introduce the methods of constructing CIs for two sample case. Given that X_{1i} follows $N(\mu_{1i}, \sigma)$ and X_{2i} follows $N(\mu_{2i}, \sigma)$ with unknown parameters μ_{1i}, μ_{2i} and known parameter σ, $\bar{X}_{1i} - \bar{X}_{2i}$ are used to estimate $\mu_{1i} - \mu_{2i}$. To simplify the case, we assume $2\sigma^2$ as the sample size of X_{1i} (or X_{2i}). Therefore $\bar{X_{1i}} - \bar{X_{2i}} \sim N(\mu_{1i} - \mu_{2i}, 1)$.

Table 4.2 Summary of Various Methods of Constructing CIs for Two Sample Case

Procedure	CIs for Two Sample Case
1	$(\bar{x}_{1i} - \bar{x}_{2i} - Z_{\frac{\lfloor \gamma R \rfloor + 1}{2m}\alpha}, \bar{x}_{1i} - \bar{x}_{2i} + Z_{\frac{\lfloor \gamma R \rfloor + 1}{2m}\alpha})$
2	$(\bar{x}_{1i} - \bar{x}_{2i} - Z_{u_{\alpha,R,\gamma}}, \bar{x}_{1i} - \bar{x}_{2i} + Z_{u_{\alpha,R,\gamma}})$
3	$(\bar{x}_{1i} - \bar{x}_{2i} - Z_{\frac{\lfloor \gamma R \rfloor + 1}{2mc_{\gamma,m}}\alpha}, \bar{x}_{1i} - \bar{x}_{2i} + Z_{\frac{\lfloor \gamma R \rfloor + 1}{2mc_{\gamma,m}}\alpha})$
4	$(\bar{x}_{1i} - \bar{x}_{2i} - Z_{\frac{\lfloor \gamma R \rfloor + 1}{2R}\alpha}, \bar{x}_{1i} - \bar{x}_{2i} + Z_{\frac{\lfloor \gamma R \rfloor + 1}{2R}\alpha})$
5	$(\bar{x}_{1i} - \bar{x}_{2i} - Z_{\frac{R}{2m}\alpha}, \bar{x}_{1i} - \bar{x}_{2i} + Z_{\frac{R}{2m}\alpha})$
6	$(\bar{x}_{1i} - \bar{x}_{2i} - Z_{\frac{R}{2mc^*_{\gamma,m}}\alpha}, \bar{x}_{1i} - \bar{x}_{2i} + Z_{\frac{R}{2mc^*_{\gamma,m}}\alpha})$

In Table 4.2, given the Procedures 1 - 6, there are six corresponding methods of constructing CIs for $\mu_{1i} - \mu_{2i}$ under two sample case.

4.3 Numerical Comparison under Independence for One-Sample Case

We generated $m = 2000$ normal random variables $\{\bar{X}_i, ..., \bar{X}_m\}$ with covariance matrix $\widetilde{\Sigma}$ and mean vector $\widetilde{\mu} = (\mu_1, ..., \mu_m)$. CIs are constructed only for the selected μ_i, $i \in \{1, 2, ..., m\}$. The mean vector $\widetilde{\mu}$ has $100\pi\%$ of nonzero mean μ_i. That is, $100\pi\%$ proportion of $\mu_i = 0.5$, meanwhile the remaining is zero. Covariance matrix $\widetilde{\Sigma}$ is defined as

$$\widetilde{\Sigma} = \begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{pmatrix} . \tag{4.1}$$

Simulation studies are first performed as nonzero proportion π varies, that is π from 0 to 1. This is done for four values of $\rho = 0, 0.2, 0.5, 0.8$, and two values of selection

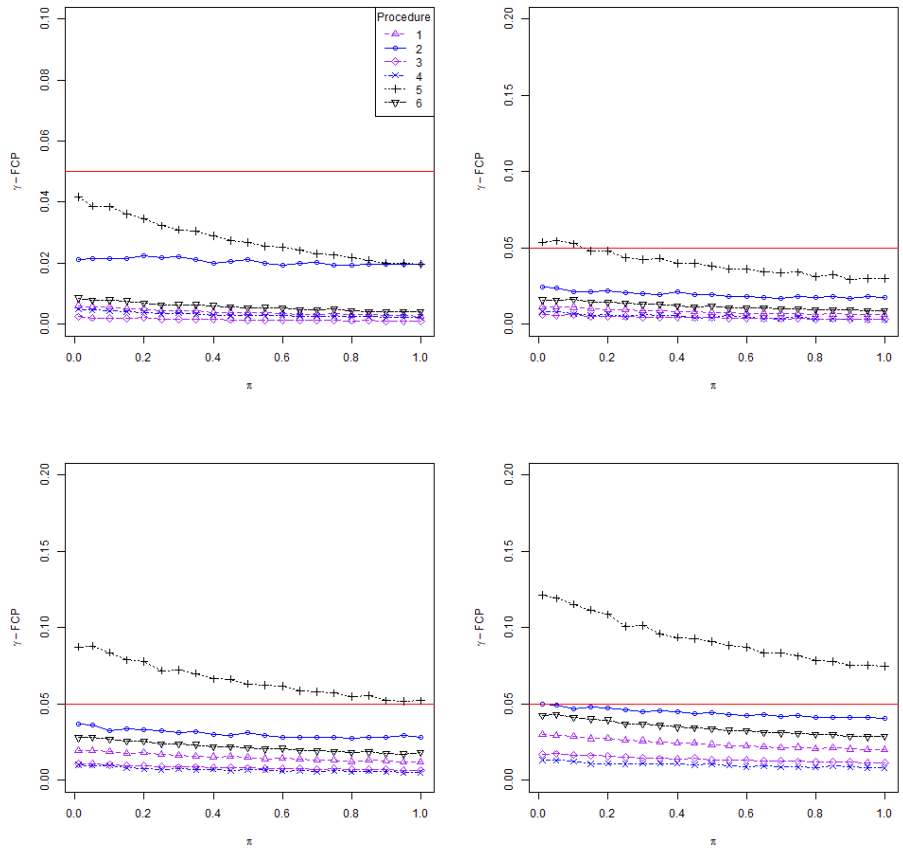

Figure 4.1 Estimated γ-FCP of our proposed unconditional CI-based γ-FCP controlling procedures and conditional CI-based γ-FCP controlling procedures along with Benjamini and Yekutieli procedures for one sample case, with correlation coefficient $\rho = 0$ (first row left panel), $\rho = 0.2$ (first row right panel), $\rho = 0.5$ (second row left panel), $\rho = 0.8$ (second row right panel) and selection level $s = 0.20$. Here, the value of nonzero proportion π is from 0 to 1. $m = 2000, \alpha = 0.05$.

level $s = 0.20$ and $s = 0.40$, corresponding to Figures 4.1 - 4.4. Next, simulation studies are performed as selection level s varies, that is s from 0 to 1. The covariance matrix is same as Equation 4.1, this is done for four values of $\rho = 0, 0.2, 0.5, 0.8$, and two values of selection level $\pi = 0.10$ and $\pi = 0.20$, corresponding to Figures 4.5 - 4.8.

In Figures 4.1 and 4.3, we compare the estimated γ-FCP with respect to the nonzero proportion π with value from 0 to 1 for one sample case. Figures 4.1 and 4.3 show that the estimated γ-FCP is bounded by 0.05 as π varies under independence ($\rho = 0$) for all six procedures. And the estimated γ-FCP is controlled by 0.05 as π

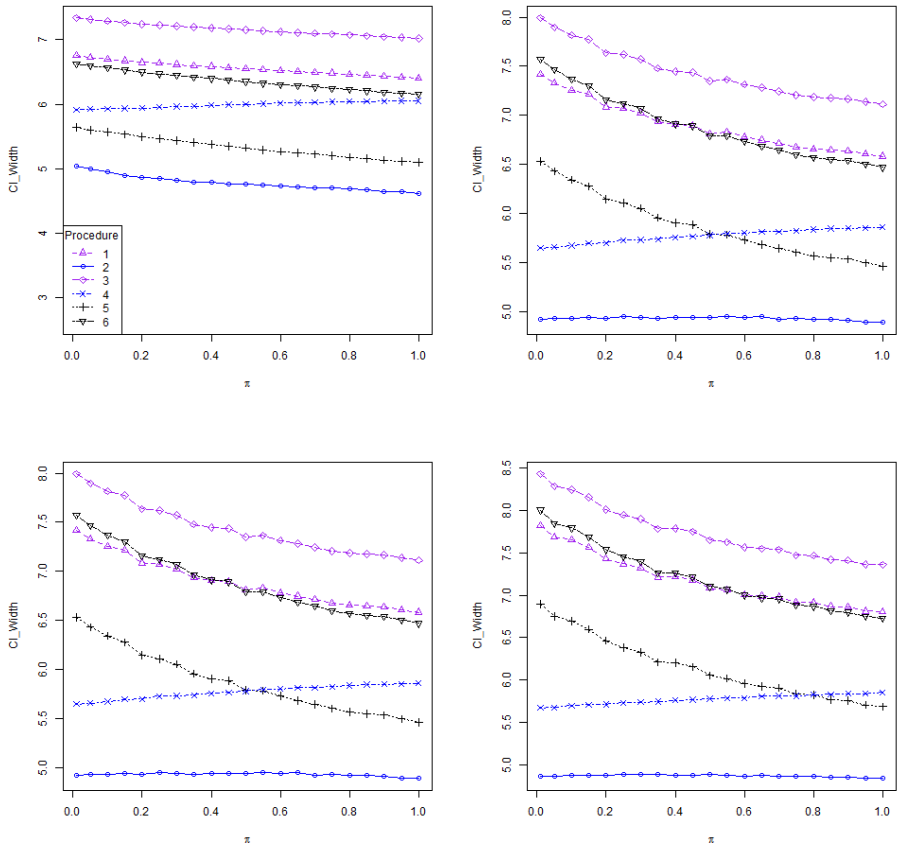

Figure 4.2 Estimated average width of CI of our proposed unconditional CI-based γ-FCP controlling procedures and conditional CI-based γ-FCP controlling procedures along with Benjamini and Yekutieli procedures for one sample case, with correlation coefficient $\rho = 0$ (first row left panel), $\rho = 0.2$ (first row right panel), $\rho = 0.5$ (second row left panel), $\rho = 0.8$ (second row right panel) and selection level $s = 0.20$. Here, the value of nonzero proportion π is from 0 to 1. $m = 2000, \alpha = 0.05$.

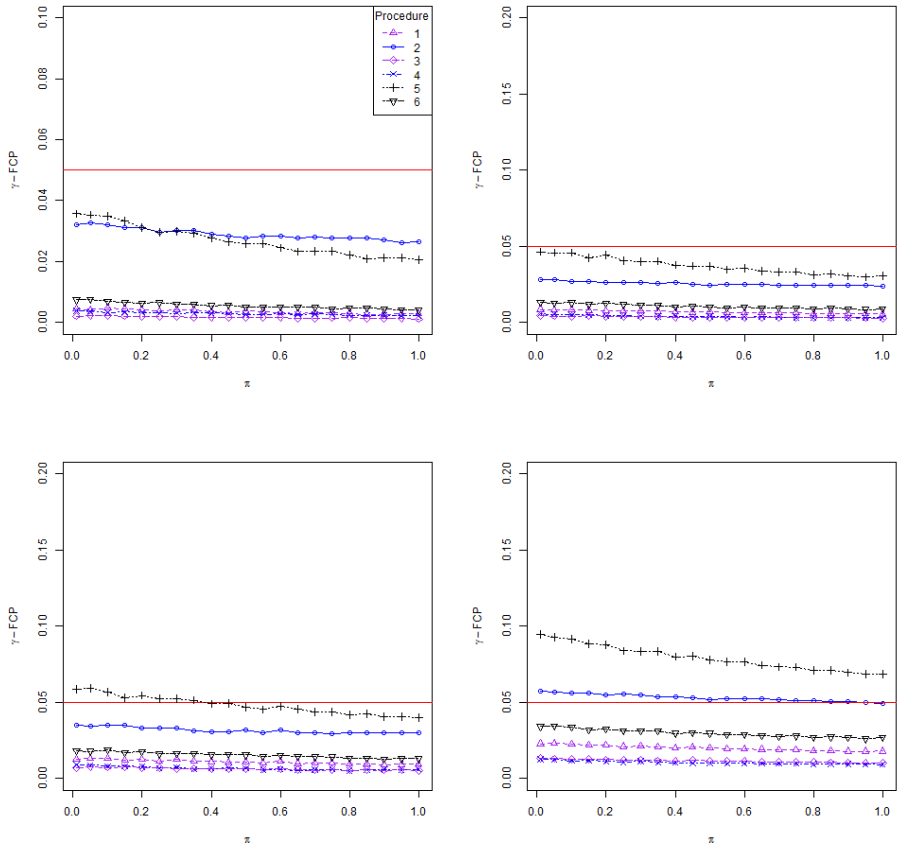

Figure 4.3 Estimated γ-FCP of our proposed unconditional CI-based γ-FCP controlling procedures and conditional CI-based γ-FCP controlling procedures along with Benjamini and Yekutieli procedures for one sample case, with correlation coefficient $\rho = 0$ (first row left panel), $\rho = 0.2$ (first row right panel), $\rho = 0.5$ (second row left panel), $\rho = 0.8$ (second row right panel) and selection level $s = 0.40$. Here, the value of nonzero proportion π is from 0 to 1. $m = 2000, \alpha = 0.05$.

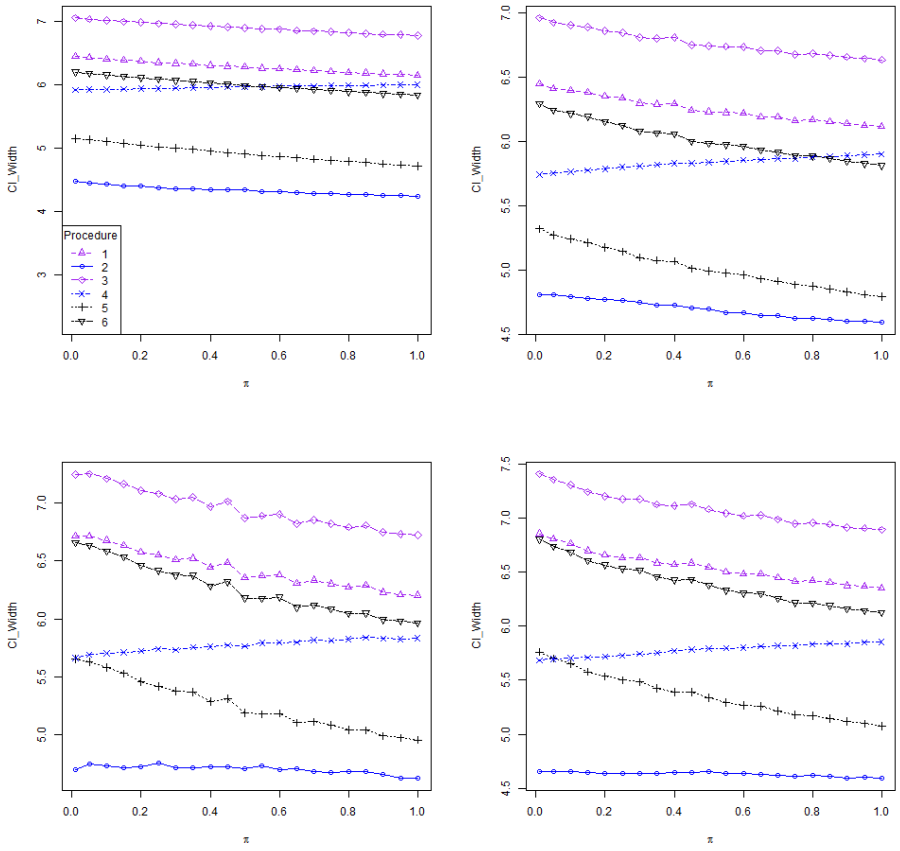

Figure 4.4 Estimated average width of CI of our proposed unconditional CI-based γ-FCP controlling procedures and conditional CI-based γ-FCP controlling procedures along with Benjamini and Yekutieli procedures for one sample case, with correlation coefficient $\rho = 0$ (first row left panel), $\rho = 0.2$ (first row right panel), $\rho = 0.5$ (second row left panel), $\rho = 0.8$ (second row right panel) and selection level $s = 0.40$. Here, the value of nonzero proportion π is from 0 to 1. $m = 2000, \alpha = 0.05$.

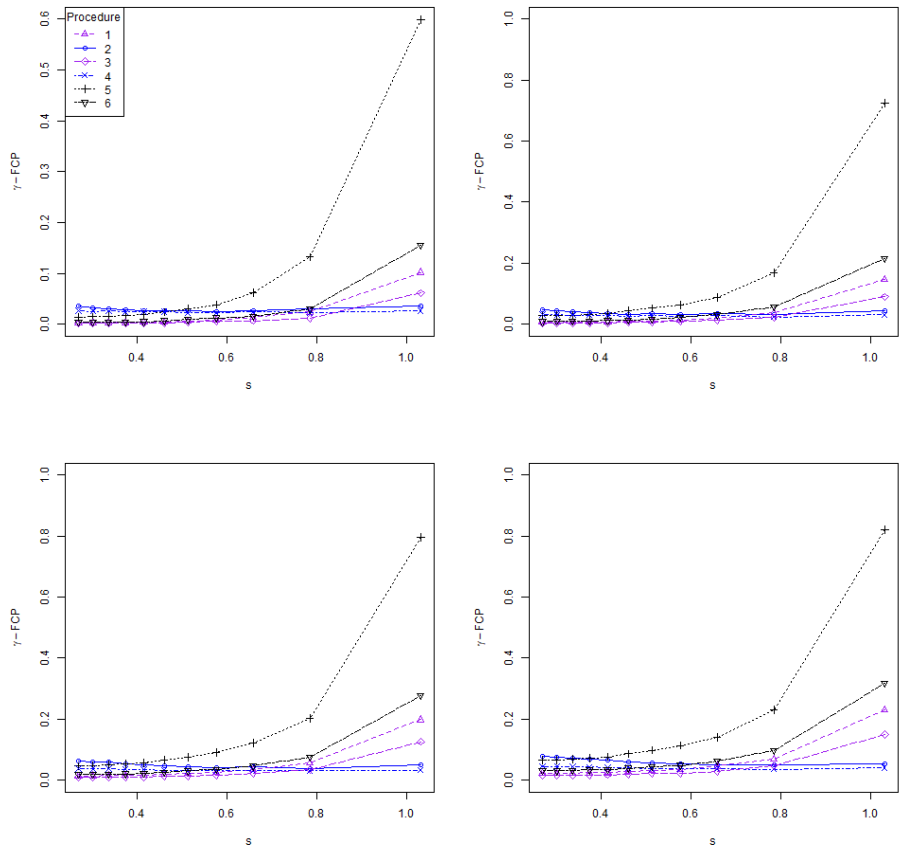

Figure 4.5 Estimated γ-FCP of our proposed unconditional CI-based γ-FCP controlling procedures and conditional CI-based γ-FCP controlling procedures along with Benjamini and Yekutieli procedures for one sample case, with correlation coefficient $\rho = 0$ (first row left panel), $\rho = 0.2$ (first row right panel), $\rho = 0.5$ (second row left panel), $\rho = 0.8$ (second row right panel) and nonzero proportion $\pi = 0.10$. Here, the value of selection level s is from 0 to 1. $m = 2000, \alpha = 0.05$.

varies when $\rho = 0.2$ and 0.5 for Procedures 1,2,3,4 and 6. The estimated γ-FCP is not controlled by 0.05 as π varies with extreme large $\rho = 0.5$ or 0.8 for Procedure 5. In Figures 4.2 and 4.4, we compare the estimated average width of CIs with respect to the nonzero proportion π with value from 0 to 1. Figures 4.2 and 4.4 show that when s and ρ are fixed, the average width of CI of Procedures 1, 3, 5 and 6 decrease as nonzero proportion π increase. The average width of CI of Procedures 2 and 4 does not vary no matter which value of π is chosen. Procedure 3 has the widest CI; whereas Procedure 2 has the least wide CI when s and ρ are fixed.

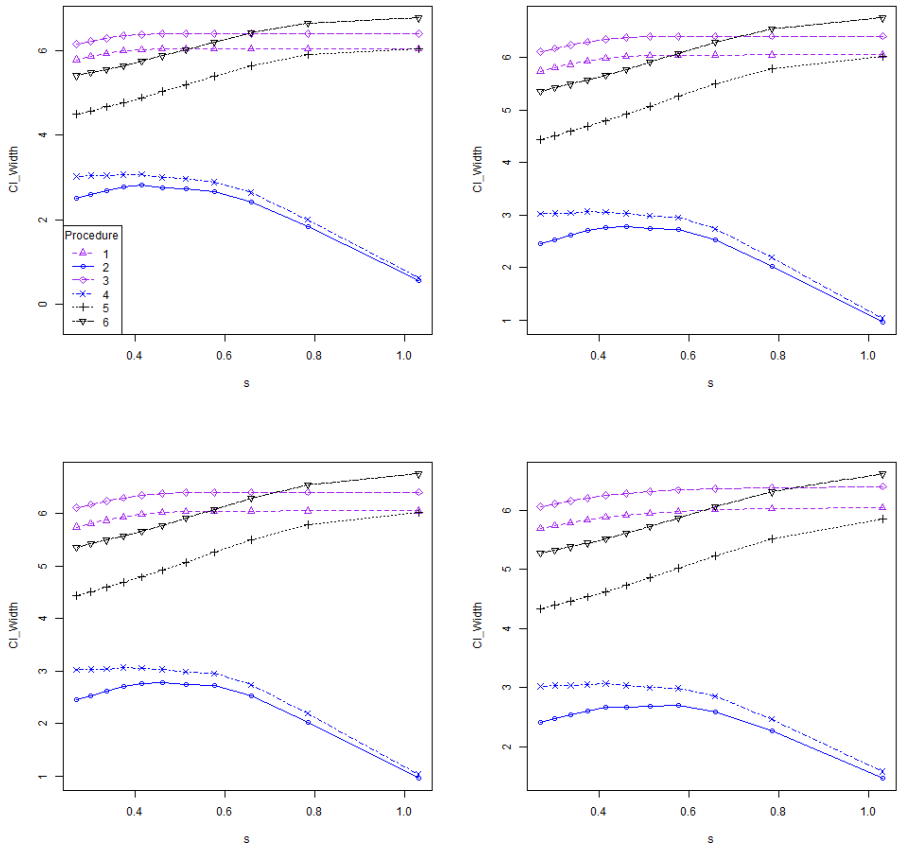

Figure 4.6 Estimated average width of CI of our proposed unconditional CI-based γ-FCP controlling procedures and conditional CI-based γ-FCP controlling procedures along with Benjamini and Yekutieli procedures for one sample case, with correlation coefficient $\rho = 0$ (first row left panel), $\rho = 0.2$ (first row right panel), $\rho = 0.5$ (second row left panel), $\rho = 0.8$ (second row right panel) and nonzero proportion $\pi = 0.10$. Here, the value of selection level s is from 0 to 1. $m = 2000, \alpha = 0.05$.

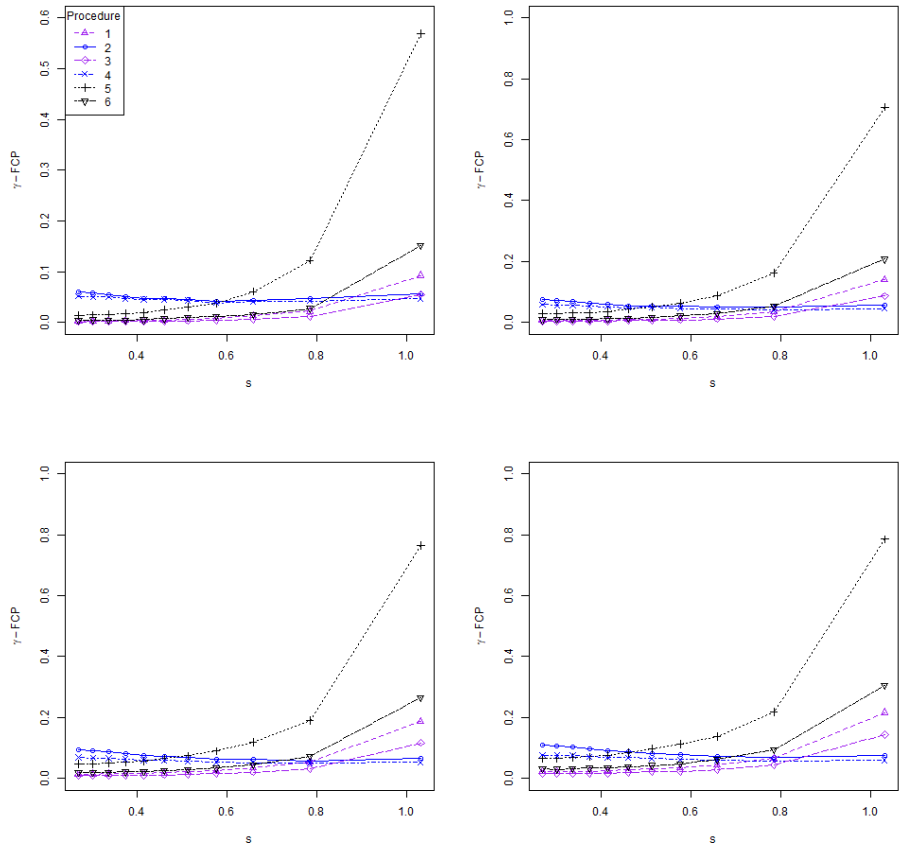

Figure 4.7 Estimated γ-FCP of our proposed unconditional CI-based γ-FCP controlling procedures and conditional CI-based γ-FCP controlling procedures along with Benjamini and Yekutieli procedures for one sample case, with correlation coefficient $\rho = 0$ (first row left panel), $\rho = 0.2$ (first row right panel), $\rho = 0.5$ (second row left panel), $\rho = 0.8$ (second row right panel) and nonzero proportion $\pi = 0.20$. Here, the value of selection level s is from 0 to 1. $m = 2000, \alpha = 0.05$.

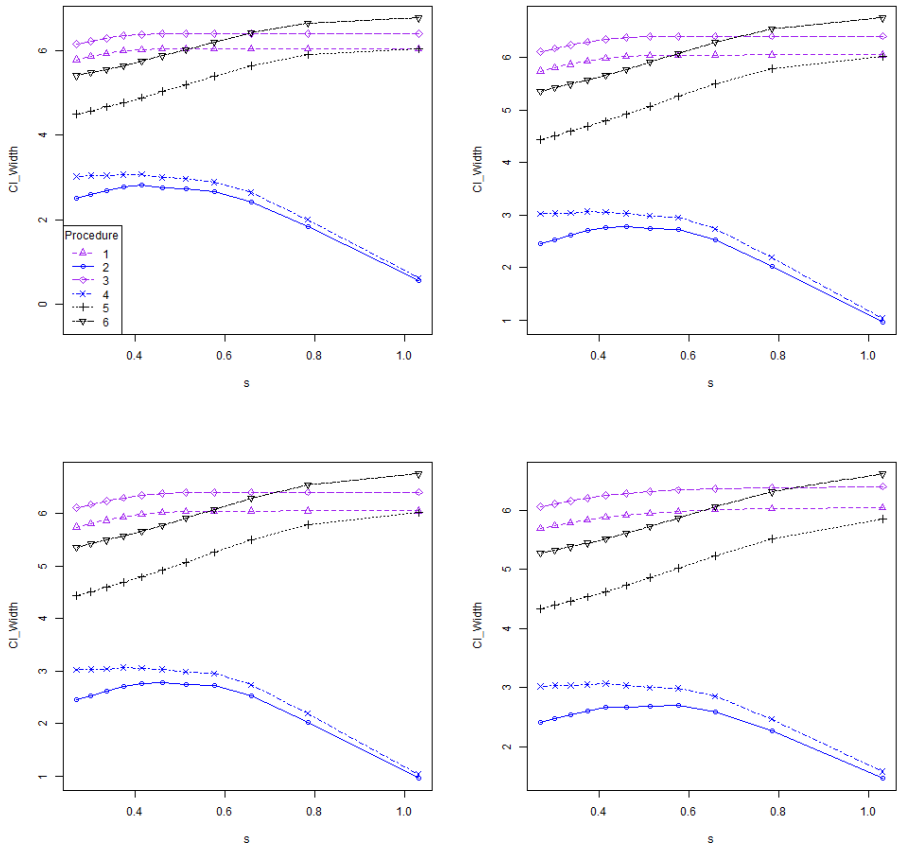

Figure 4.8 Estimated average width of CI of our proposed unconditional CI-based γ-FCP controlling procedures and conditional CI-based γ-FCP controlling procedures along with Benjamini and Yekutieli procedures for one sample case, with correlation coefficient $\rho = 0$ (first row left panel), $\rho = 0.2$ (first row right panel), $\rho = 0.5$ (second row left panel), $\rho = 0.8$ (second row right panel) and nonzero proportion $\pi = 0.10$. Here, the value of selection level s is from 0 to 1. $m = 2000, \alpha = 0.05$.

In Figures 4.5 and 4.7, we compare the estimated γ-FCP with respect to selection parameter s with value from 0 to 1. Figures 4.5 and 4.7 show the estimated γ-FCP of Procedures 2 and 4 are bounded by 0.05 as selection parameter s varies when $\rho = 0$. Procedures 1, 3, 5 and 6 only control γ-FCP when s is small. Figures 4.5 and 4.7 also show that, under the case when $\rho = 0.2, 0.5$ and 0.8, the estimated γ-FCP of Procedures 2 and 4 is stilled controlled. The estimated γ-FCP of Procedures 1, 3, 5 and 6 is out of control when the value of s is large. In Figures 4.6 and 4.8, we compare the estimated average width of CI with respect to selection parameter s with value from 0 to 1. Figures 4.6 and 4.8 show the CI width of Procedures 2 and 4 decreases as selection parameter s increases. And Figures 4.6 and 4.8 also show that Procedure 2 can always provide the shortest CI, and Procedure 1 provides the widest CI width. Therefore, we suggest using Procedure 2 (conditional CI-based γ-FCP controlling procedure) under independence for one sample case.

4.4 Numerical Comparison under Independence for Two-Sample Case

We generated $m = 2000$ two-samples normal random, which are $\{\bar{X}_{11}, \bar{X}_{12}, ..., \bar{X}_{1m}\}$ with mean vector $\widetilde{\mu_1} = (\mu_{11}, ..., \mu_{1m})$ and covariance matrix $\widetilde{\Sigma}$ for treatment group and $\{\bar{X}_{21}, \bar{X}_{22}, ..., \bar{X}_{2m}\}$ with mean vector $\widetilde{\mu_2} = (\mu_{21}, ..., \mu_{2m})$ and same covariance matrix $\widetilde{\Sigma}$ for control group. $\widetilde{\Sigma}$ is defined as Matrix equation 4.1. The mean vector $\widetilde{\mu_1}$ has $100\pi\%$ of nonzero mean μ_{1i}. That is, $100\pi\%$ proportion of $\mu_{1i} = 0.5$, meanwhile the remaining is zero; and mean vector $\widetilde{\mu_2} = \widetilde{0}$. Simulation studies are first designed as nonzero proportion π varies, that is π from 0 to 1. The covariance matrix is Matrix equation 4.1, this is done for four values of $\rho = 0, 0.2, 0.5, 0.8$, and two values of selection level $s = 0.20$ and $s = 0.40$, corresponding to Figures 4.9 - 4.12. Next, simulation studies are performed as selection level s varies, that is s from 0 to 1. The covariance matrix is Matrix 4.1, this is done for four values of $\rho = 0, 0.2, 0.5, 0.8$, and

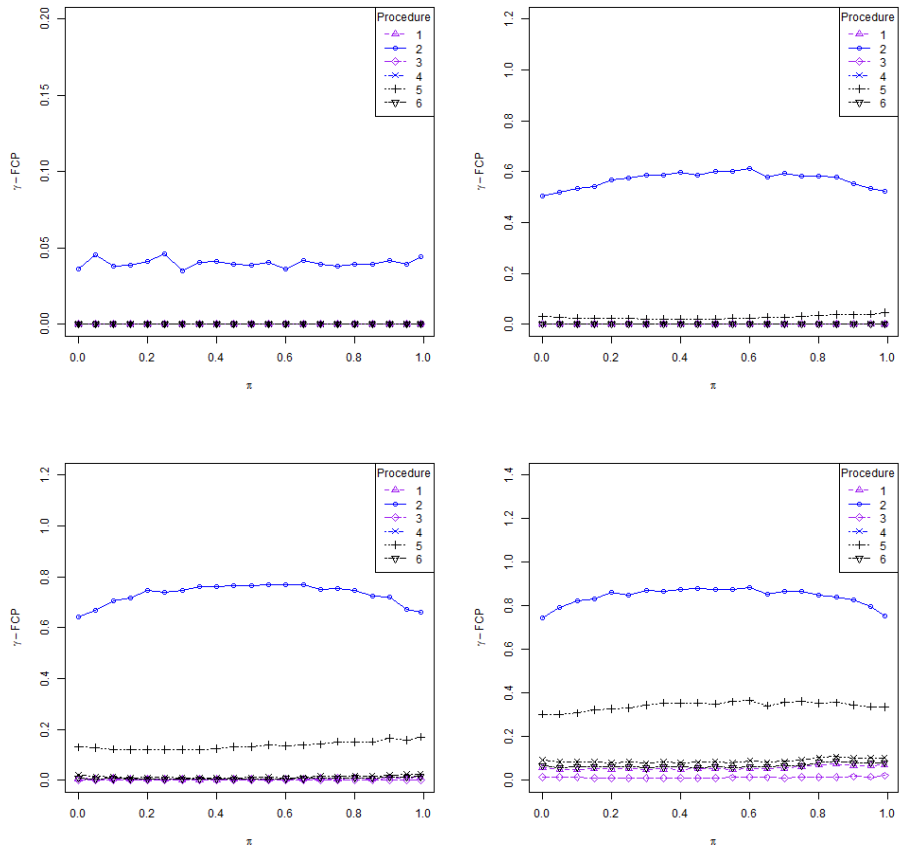

Figure 4.9 Estimated γ-FCP of our proposed unconditional CI-based γ-FCP controlling procedures and conditional CI-based γ-FCP controlling procedures along with Benjamini and Yekutieli procedures for two sample case, with correlation coefficient $\rho = 0$ (first row left panel), $\rho = 0.2$ (first row right panel), $\rho = 0.5$ (second row left panel), $\rho = 0.8$ (second row right panel) and selection level $s = 0.20$. Here, the value of nonzero proportion π is from 0 to 1. $m = 2000, \alpha = 0.05$.

two values of selection level $\pi = 0.10$ and $\pi = 0.20$, corresponding to Figures 4.13 - 4.16.

In Figures 4.9 and 4.11, we compare the estimated γ-FCP with respect to the nonzero proportion π with value from 0 to 1. Figures 4.9 and 4.11 show that the estimated γ-FCP is bounded by 0.05 as π varies when $\rho = 0$ for all six procedures. The estimated γ-FCP of Procedure 2 is bounded by and most close to 0.05 as π varies when $\rho = 0$. And the estimated γ-FCP is still controlled by 0.05 as π varies when $\rho = 0.2, 0.5$ or 0.8 for Procedures 1,3,4 and 6. The estimated γ-FCP is not controlled by 0.05 as π varies when $\rho = 0.2, 0.5$ or 0.8 for Procedure 2. The estimated

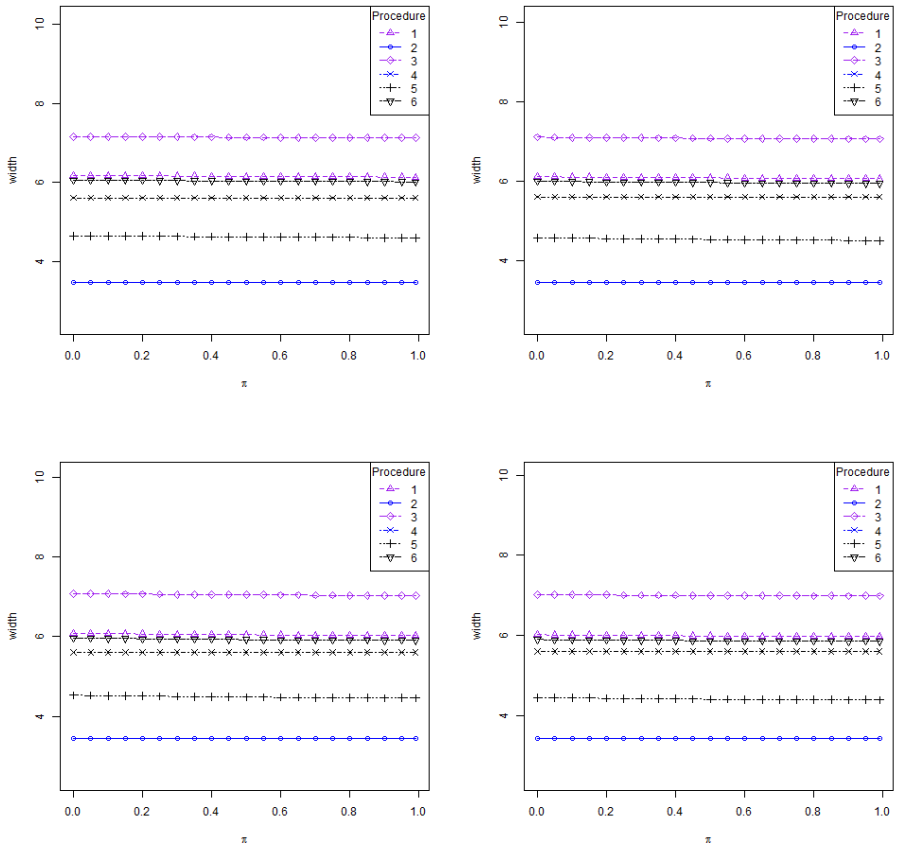

Figure 4.10 Estimated average width of CI of our proposed unconditional CI-based γ-FCP controlling procedures and conditional CI-based γ-FCP controlling procedures along with Benjamini and Yekutieli procedures for two sample case, with correlation coefficient $\rho = 0$ (first row left panel), $\rho = 0.2$ (first row right panel), $\rho = 0.5$ (second row left panel), $\rho = 0.8$ (second row right panel) and selection level $s = 0.20$. Here, the value of nonzero proportion π is from 0 to 1. $m = 2000, \alpha = 0.05$.

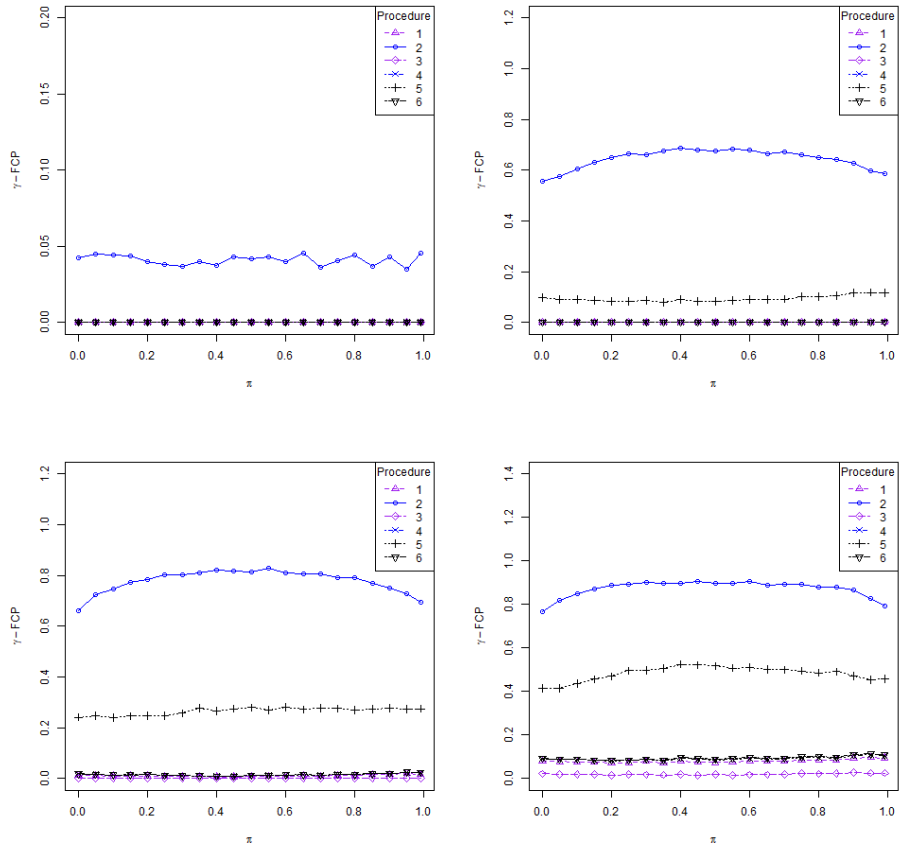

Figure 4.11 Estimated γ-FCP of our proposed unconditional CI-based γ-FCP controlling procedures and conditional CI-based γ-FCP controlling procedures along with Benjamini and Yekutieli procedures for two sample case, with correlation coefficient $\rho = 0$ (first row left panel), $\rho = 0.2$ (first row right panel), $\rho = 0.5$ (second row left panel), $\rho = 0.8$ (second row right panel) and selection level $s = 0.40$. Here, the value of nonzero proportion π is from 0 to 1. $m = 2000, \alpha = 0.05$.

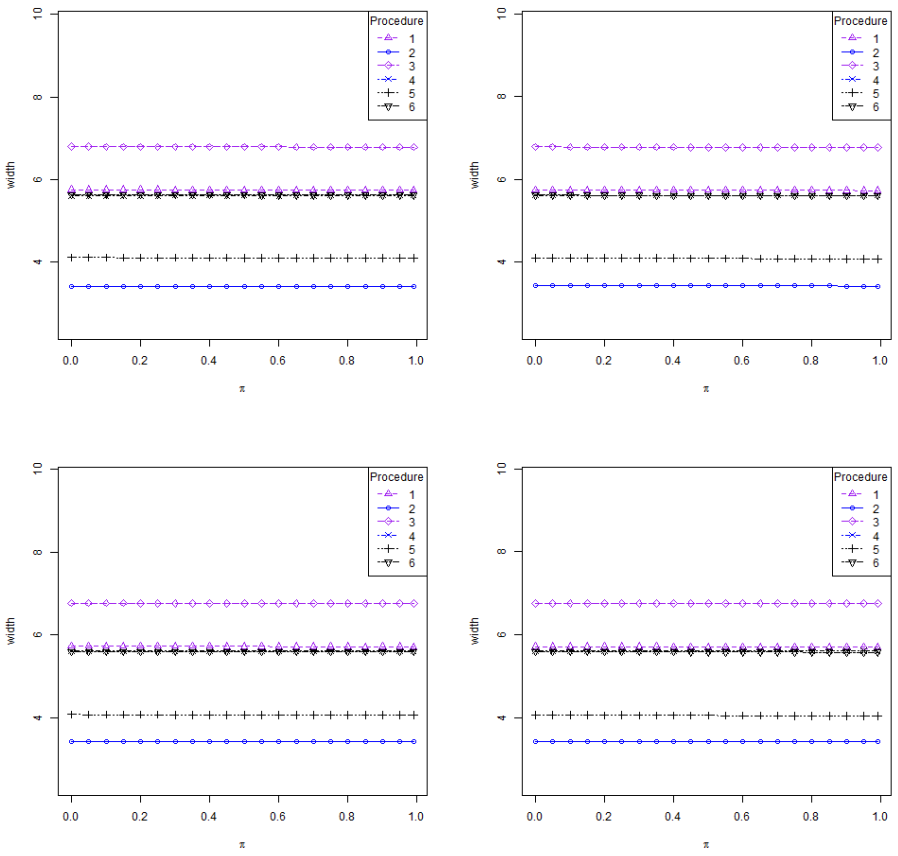

Figure 4.12 Estimated average width of CI of of our proposed unconditional CI-based γ-FCP controlling procedures and conditional CI-based γ-FCP controlling procedures along with Benjamini and Yekutieli procedures for two sample case, with correlation coefficient $\rho = 0$ (first row left panel), $\rho = 0.2$ (first row right panel), $\rho = 0.5$ (second row left panel), $\rho = 0.8$ (second row right panel) and selection level $s = 0.40$. Here, the value of nonzero proportion π is from 0 to 1. $m = 2000, \alpha = 0.05$.

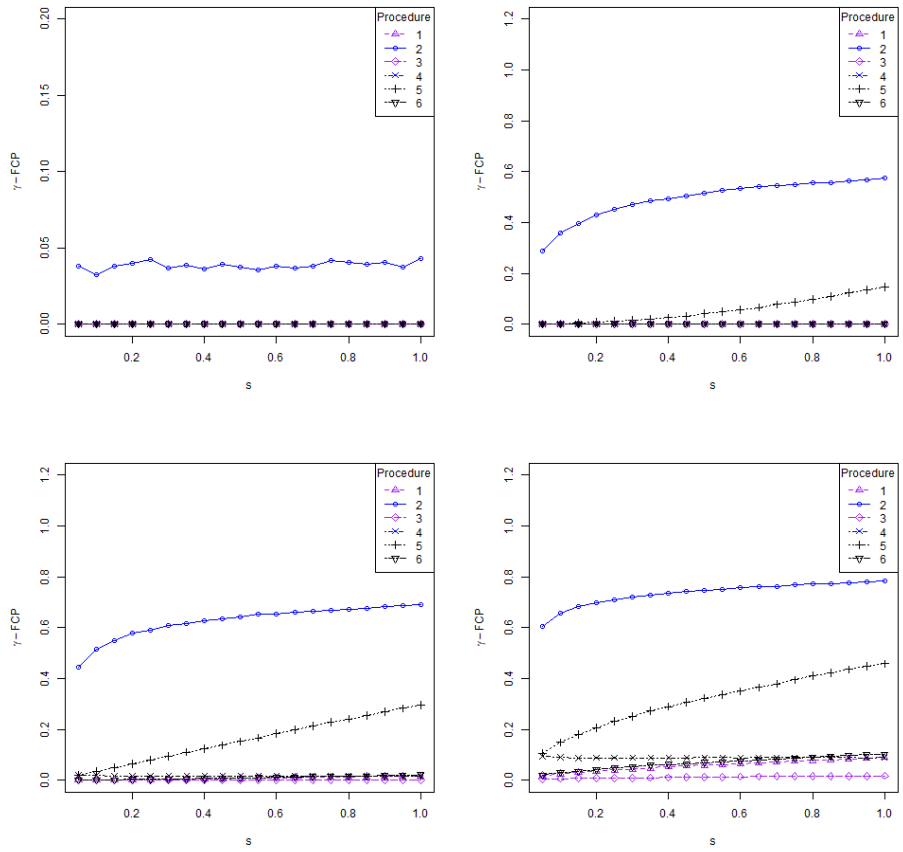

Figure 4.13 Estimated γ-FCP of our proposed unconditional CI-based γ-FCP controlling procedures and conditional CI-based γ-FCP controlling procedures along with Benjamini and Yekutieli procedures for two sample case, with correlation coefficient $\rho = 0$ (first row left panel), $\rho = 0.2$ (first row right panel), $\rho = 0.5$ (second row left panel), $\rho = 0.8$ (second row right panel) and nonzero proportion $\pi = 0.10$. Here, the value of selection level s is from 0 to 1. $m = 2000, \alpha = 0.05$.

γ-FCP is not controlled by 0.05 as π varies with extreme large ρ for Procedure 5. In Figures 4.10 and 4.12, we compare the estimated CI width with respect to the nonzero proportion π with value from 0 to 1. Figures 4.10 and 4.12 show that when s and ρ are fixed, the CI width of six procedures does not vary no matter which value of π is chosen. This is the result of the adjusted level, which decides the CI width, is not a function of/not affected by π. Procedure 3 has the widest CI; whereas Procedure 2 has the least wide CI when s and ρ are fixed.

In Figures 4.13 and 4.15, we compare the estimated γ-FCP with respect to selection parameter s with value from 0 to 1. Figures 4.13 and 4.15 show the estimated

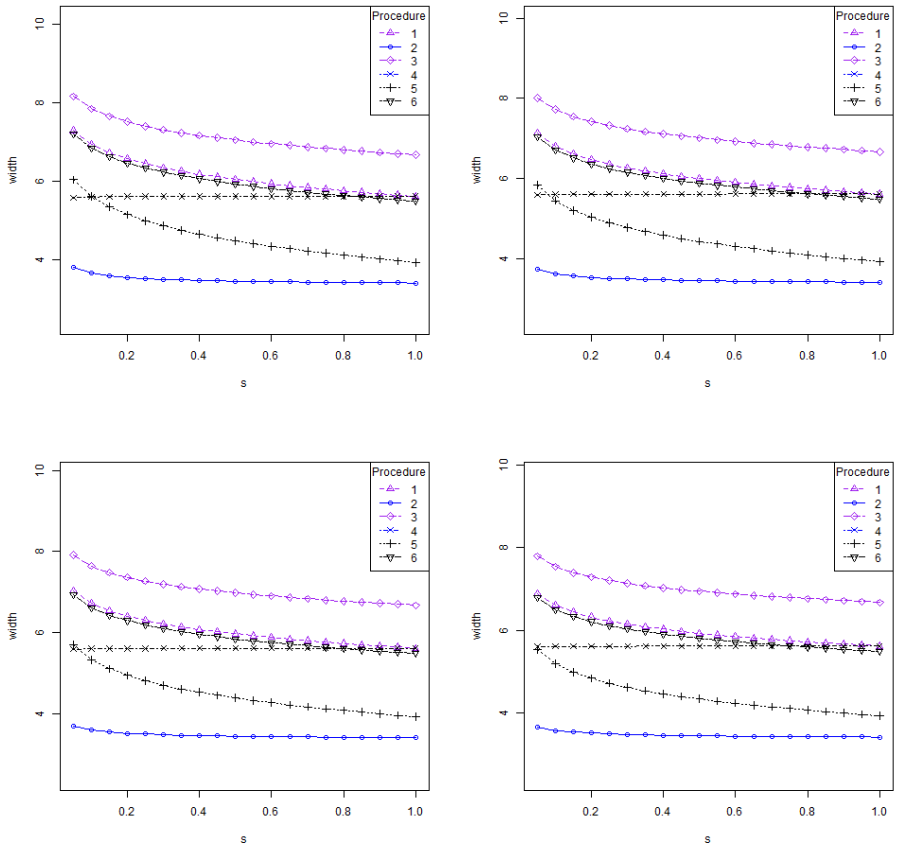

Figure 4.14 Estimated average width of CI of our proposed unconditional CI-based γ-FCP controlling procedures and conditional CI-based γ-FCP controlling procedures along with Benjamini and Yekutieli procedures for two sample case, with correlation coefficient $\rho = 0$ (first row left panel), $\rho = 0.2$ (first row right panel), $\rho = 0.5$ (second row left panel), $\rho = 0.8$ (second row right panel) and nonzero proportion $\pi = 0.10$. Here, the value of selection level s is from 0 to 1. $m = 2000, \alpha = 0.05$.

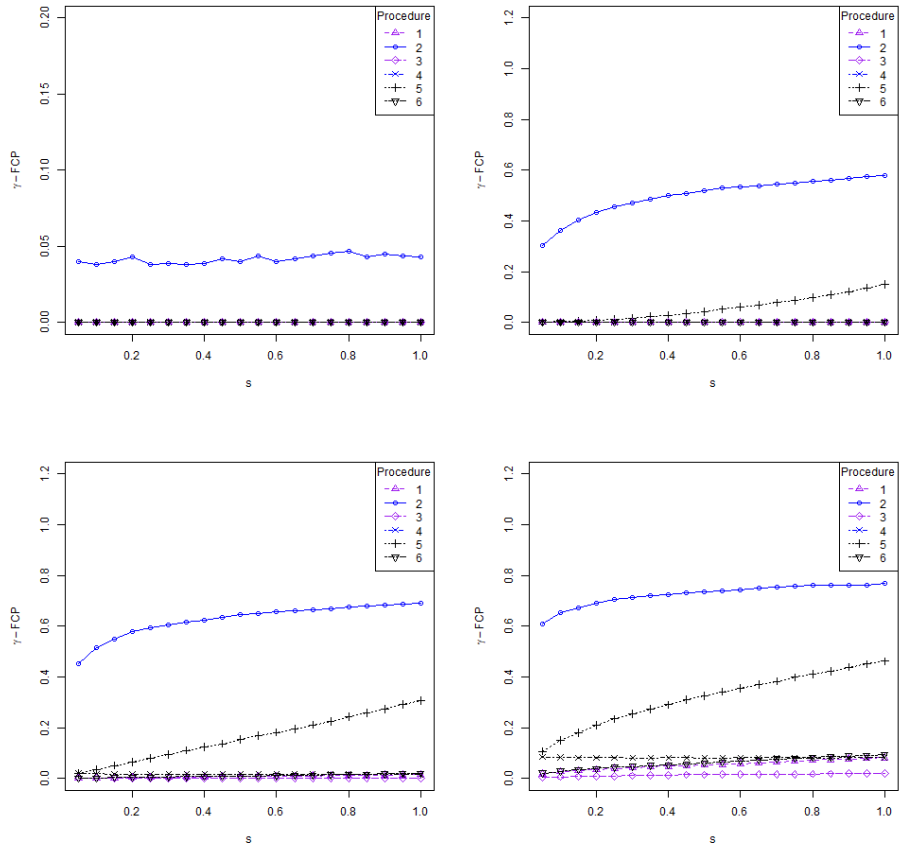

Figure 4.15 Estimated γ-FCP of our proposed unconditional CI-based γ-FCP controlling procedures and conditional CI-based γ-FCP controlling procedures along with Benjamini and Yekutieli procedures for two sample case, with correlation coefficient $\rho = 0$ (first row left panel), $\rho = 0.2$ (first row right panel), $\rho = 0.5$ (second row left panel), $\rho = 0.8$ (second row right panel) and nonzero proportion $\pi = 0.20$. Here, the value of selection level s is from 0 to 1. $m = 2000, \alpha = 0.05$.

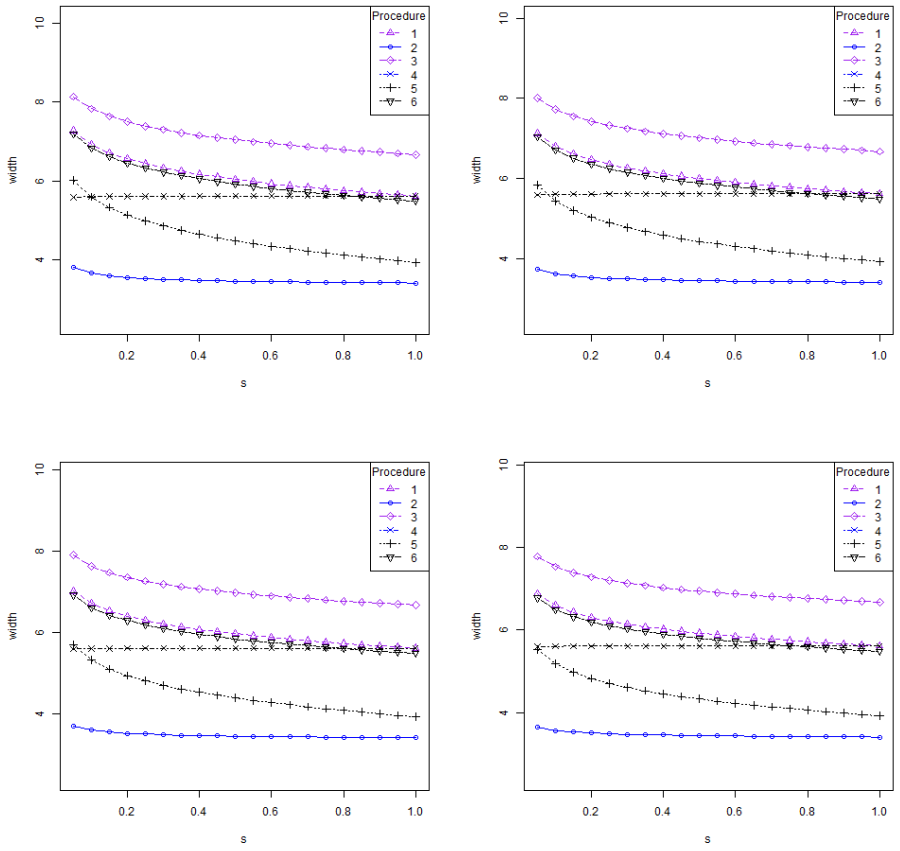

Figure 4.16 Estimated average width of CI of our proposed unconditional CI-based γ-FCP controlling procedures and conditional CI-based γ-FCP controlling procedures along with Benjamini and Yekutieli procedures for two sample case, with correlation coefficient $\rho = 0$ (first row left panel), $\rho = 0.2$ (first row right panel), $\rho = 0.5$ (second row left panel), $\rho = 0.8$ (second row right panel) and nonzero proportion $\pi = 0.10$. Here, the value of selection level s is from 0 to 1. $m = 2000, \alpha = 0.05$.

γ-FCP of all procedures are bounded by 0.05 as selection parameter s varies when $\rho = 0$. Figures 4.13 and 4.15 also show that, under the case when $\rho = 0.2, 0.5$ and 0.8, the estimated γ-FCP of Procedures 1,3,4,6 is stilled controlled. The estimated γ-FCP of both Procedure 2 is out of control, no matter what value of s is. The estimated γ-FCP of Procedure 6 are out of control for most value of s, except small value.

In Figures 4.14 and 4.16, we compare the estimated CI width with respect to selection parameter s with value from 0 to 1. Figures 4.14 and 4.16 show the CI width of six procedures decreases as selection parameter s increases. Since the more parameters are selected, the CI width is shorter. And Figures 4.14 and 4.16 also show that Procedure 2 can always provide the shortest CI, and Procedure 1 provides the most wide CI width. Therefore, we suggest using Procedure 2 (conditional CI-based γ-FCP controlling procedure) under independence for one sample case.

4.5 Numerical Comparison under Dependence for One-Sample Case

We generated $m = 2000$ normal random variables $\{\bar{X}_i, ..., \bar{X}_m\}$ with covariance matrix $\widetilde{\Sigma}$ and mean vector $\widetilde{\mu} = (\mu_1, ..., \mu_m)$. CIs are constructed only for the selected μ_i, $i \in \{1, 2, ..., m\}$. The mean vector $\widetilde{\mu}$ has $100\pi\%$ of nonzero mean μ_i. That is, $100\pi\%$ proportion of $\mu_i = 0.5$, meanwhile the remaining is zero. Given that ρ is the correlation relationship between X_{ij}, $j = 1, 2, ..m$, which is a value between 0 and 1. We considered (1) a equal correlation structure, where covariate matrix is defined as Equation 4.1. And (2) a mixed correlation structure is considered as well, where covariate matrix is defined as

$$\widetilde{\Sigma} = \begin{pmatrix} 1 & -\rho & \rho & -\rho & \cdots & \rho & -\rho \\ -\rho & 1 & -\rho & \rho & \cdots & -\rho & \rho \\ \rho & -\rho & 1 & -\rho & \cdots & \rho & -\rho \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ -\rho & \rho & -\rho & \rho & \cdots & -\rho & 1 \end{pmatrix}, \tag{4.2}$$

where the element in $\widetilde{\Sigma}$ is defined as ρ_{ij}, where i and j are row and column of the element, where $\rho_{ij} = (-1)^{i-j}\rho$. The covariance matrices are supposed to be positive semi-definite. We need to ensure the covariance matrix in Matrix Equation 4.2 is positive semi-definite. See the proof in Appendix. And (3) a block-wise structure is also considered, where within each block (where size is $m_b = 200$),

$$\widetilde{\Sigma}_{m_b} = \begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{pmatrix},$$

and between blocks,

$$\widetilde{\Sigma} = \begin{pmatrix} \widetilde{\Sigma}_{m_b} & \widetilde{0} & \cdots & \widetilde{0} \\ \widetilde{0} & \widetilde{\Sigma}_{m_b} & \cdots & \widetilde{0} \\ \vdots & \vdots & \ddots & \vdots \\ \widetilde{0} & \widetilde{0} & \cdots & \widetilde{\Sigma}_{m_b} \end{pmatrix}. \tag{4.3}$$

And last (4) we consider AR structure, where

$$\widetilde{\Sigma} = \begin{pmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{m-1} \\ \rho & 1 & \rho & \cdots & \rho^{m-2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{m-1} & \rho^{m-2} & \rho^{m-3} & \cdots & 1 \end{pmatrix}. \tag{4.4}$$

There are four types of covariance matrix, Matrix Equations 4.1 - 4.4. The nonzero proportion $\pi = 0.1$ and $\pi = 0.2$, and two values of selection level $s = 0.20$ and $s = 0.40$. Correlation coefficient ρ is from 0 to 1, corresponding to Figures 4.17 - 4.20. Figure 4.17 displays the γ-FCP and average CI width under equal correlation structure as ρ varies from 0 to 1. Procedures 1, 3 and 4 control γ-FCP at level 0.05 as correlation coefficient ρ varies (γ-FCP of Procedures 1 is close to 0.05 in this

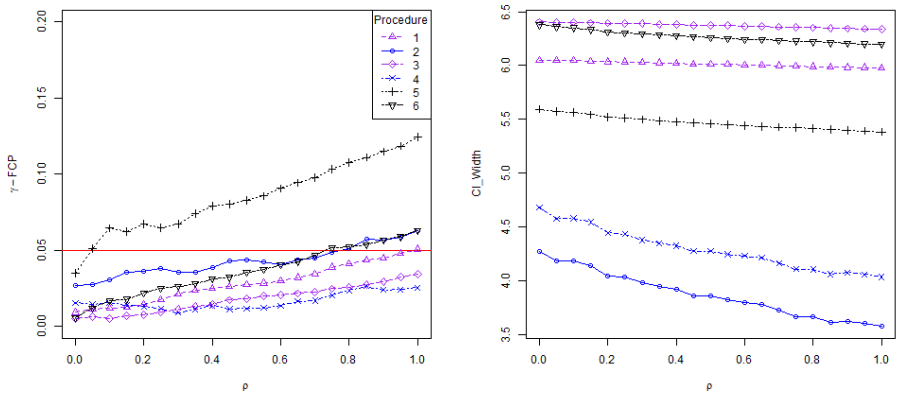

Figure 4.17 Estimated γ-FCP (left panel) and average width of CI (right panel) of our proposed unconditional CI-based γ-FCP controlling procedures and conditional CI-based γ-FCP controlling procedures along with Benjamini and Yekutieli procedures for one sample case under equal correlation dependence, with $\pi = 0.10$, $s = 0.2$. Here correlation coefficient ρ is from 0 to 1. $m = 2000, \alpha = 0.05$.

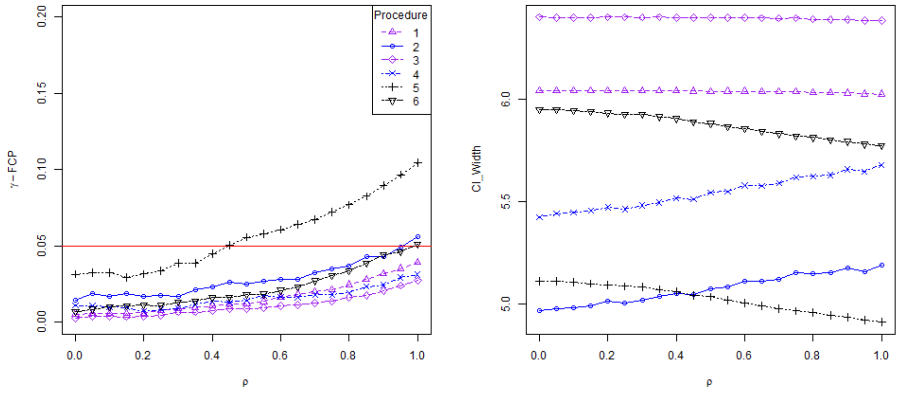

Figure 4.18 Estimated γ-FCP (left panel) and average width of CI (right panel) of our proposed unconditional CI-based γ-FCP controlling procedures and conditional CI-based γ-FCP controlling procedures along with Benjamini and Yekutieli procedures for one sample case under mixed correlation dependence, with $\pi = 0.10$, $s = 0.2$. Here correlation coefficient ρ is from 0 to 1. $m = 2000, \alpha = 0.05$.

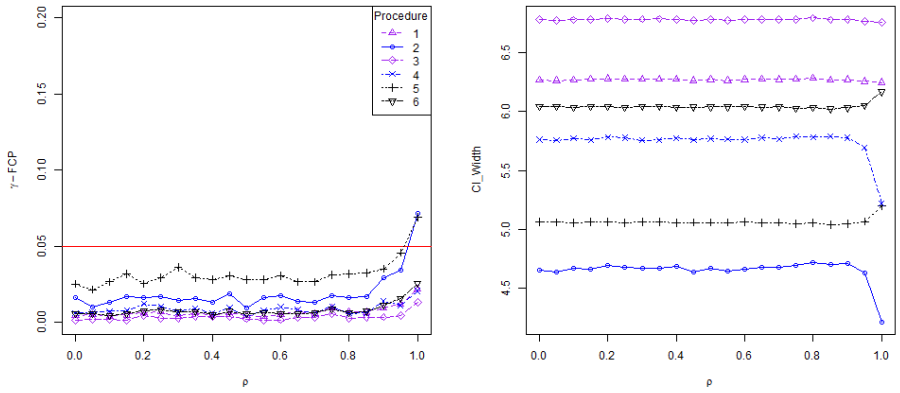

Figure 4.19 Estimated γ-FCP (left panel) and average width of CI (right panel) of our proposed unconditional CI-based γ-FCP controlling procedures and conditional CI-based γ-FCP controlling procedures along with Benjamini and Yekutieli procedures for one sample case under block-wise dependence, block number is 40, with $\pi = 0.10$, $s = 0.2$. Here correlation coefficient ρ is from 0 to 1. $m = 2000, \alpha = 0.05$.

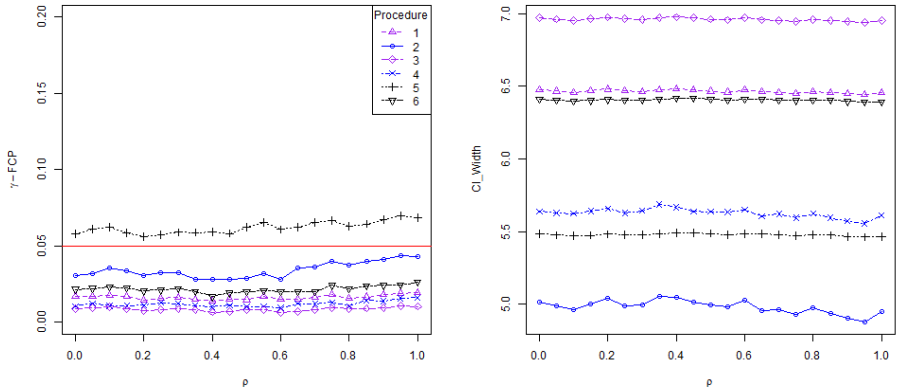

Figure 4.20 Estimated γ-FCP (left panel) and average width of CI (right panel) of our proposed unconditional CI-based γ-FCP controlling procedures and conditional CI-based γ-FCP controlling procedures along with Benjamini and Yekutieli procedures for one sample case under AR structure, with $\pi = 0.10$, $s = 0.2$. Here correlation coefficient ρ is from 0 to 1. $m = 2000, \alpha = 0.05$.

case). Procedures 2 and 6 control γ-FCP at level 0.05 when correlation coefficient ρ is moderate. Procedure 5 does not control γ-FCP at level 0.05 except $\rho = 0$. As ρ increases, the average CI width of Procedures 2 and 4 decrease. Meanwhile the other procedure slightly decreases. Procedure 4 outperforms than the others. Since it has the shortest CI width among the procedures that controls γ-FCP under dependence at level 0.05.

Figure 4.18 shows the γ-FCP and average CI width under mixed correlation structure as ρ varies from 0 to 1. Procedures 1, 3 and 4 control γ-FCP at level 0.05 as correlation coefficient ρ varies. Procedures 2 and 6 control γ-FCP at level 0.05 for most values of correlation coefficient ρ (except $\rho = 1$ in this case). Procedure 5 does not control γ-FCP at level 0.05 when correlation coefficient ρ are not large (smaller than 0.5 in this case). As ρ increases, average CI width of Procedures 2 and 4 increase. Meanwhile the other procedure slightly decreases. Procedure 4 outperforms than the others. Since it has the shortest CI width among the procedures that controls γ-FCP under dependence at level 0.05. If we exclude the case $\rho = 1$, Procedure 2 is better than Procedure 4, since it has the shortest CI width. Both Figures 4.17 and 4.18 indicate that Procedure 3 performs the worst in terms of the CI width.

Figure 4.19 displays the γ-FCP and average CI width under block-wise structure as ρ varies from 0 to 1. The block size is 5 and the block number is 40. Procedure 1, 3, 4 and 6 control γ-FCP at level 0.05 as correlation coefficient ρ varies. Procedure 2 and 5 control γ-FCP at level 0.05 when correlation coefficient ρ is not large (smaller than and equal to 0.9 in this case). As ρ increases, average CI width of Procedure 1 - 6 slightly change except $\rho = 1$.

Figure 4.20 shows the γ-FCP and average CI width under AR structure as ρ varies from 0 to 1. Procedures 1, 2, 3, 4 and 6 control γ-FCP at level 0.05 as correlation coefficient ρ varies. Procedure 5 does not control γ-FCP for all value of correlation coefficient ρ. As ρ increases, average CI width of Procedures 1 - 6 slightly

change. Figures 4.17 to 4.20 indicate that Procedure 3 performs the worst in terms of the CI width, though it controlled γ-FCP at level 0.05. Procedure 4 outperforms against the others. Since it has the shortest CI width among the procedures that controls γ-FCP under dependence at level 0.05. If we exclude some large value ρ (i.e.: $\rho = 0.75, 0.8, ..., 1$), Procedure 2 performs better than Procedure 4. Therefore, under the various dependence structure, there is no such procedure which is always more powerful than the others. In different practical situation, one needs to choose a most suitable procedure.

4.6 Numerical Comparison under Dependence for Two-Sample Case

We generated $m = 2000$ two-samples random variable, which are $\{X_{11}, X_{12}, ..., X_{1m}\}$ with mean vector $\widetilde{\mu_1} = (\mu_{11}, ..., \mu_{1m})$ and covariance matrix $\widetilde{\Sigma}$ for treatment group and $\{X_{21}, X_{22}, ..., X_{2m}\}$ with mean vector $\widetilde{\mu_2} = (\mu_{21}, ..., \mu_{2m})$ and same covariance matrix $\widetilde{\Sigma}$ for control group. $\widetilde{\Sigma}$ is similarly defined as Matrix Equation 4.1. The mean vector $\widetilde{\mu_1}$ has $100\pi\%$ of nonzero mean μ_{1i}. That is, $100\pi\%$ proportion of $\mu_{1i} = 0.5$, meanwhile the remaining is zero; and mean vector $\widetilde{\mu_2} = \widetilde{0}$. Given that ρ is the correlation relationship between X_{ij}, $j = 1, 2, ..m$, and ρ is between 0 and 1. We considered (1) a equal correlation structure, where covariate matrix is similarly defined as Matrix Equation 4.2; (2) a mixed correlation structure, where covariate matrix is similarly defined as Matrix Equation 4.3; a block-wise structure, where covariate matrix is similarly defined as Matrix Equations 4.4 - 4.5; and (4) AR structure, where covariate matrix is similarly defined as Matrix Equation 4.6. The nonzero proportion $\pi = 0.1$ and $\pi = 0.2$, and two values of Selection level $s = 0.20$ and $s = 0.40$. Correlation coefficient ρ is from 0 to 1, corresponding to Figures 4.21 - 4.24.

Figure 4.21 displays the γ-FCP and average CI width under equal correlation structure as ρ varies from 0 to 1. As seen from Figure 4.21, only when $\rho = 0$, all the procedures control γ-FCP at level 0.05. Procedures 1, 3 and 4 control γ-FCP at

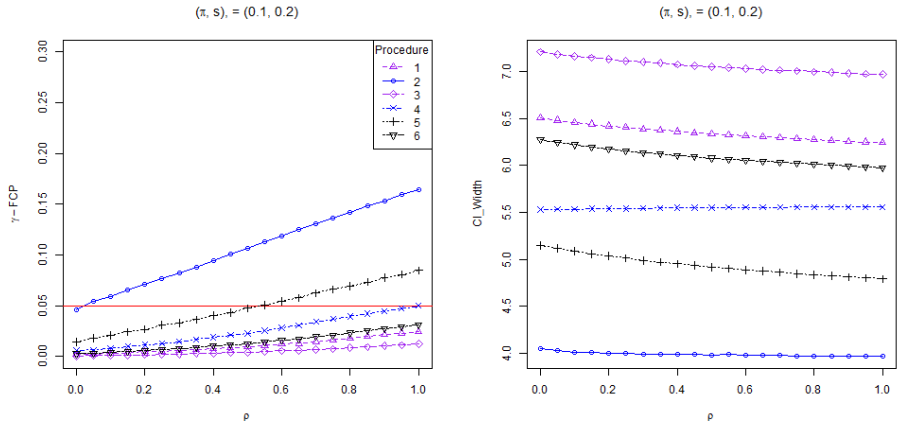

Figure 4.21 Estimated γ-FCP (left panel) and average width of CI (right panel) of our proposed unconditional CI-based γ-FCP controlling procedures and conditional CI-based γ-FCP controlling procedures along with Benjamini and Yekutieli procedures for two sample case under equal correlation structure, with $\pi = 0.10$, $s = 0.2$. Here correlation coefficient ρ is from 0 to 1. $m = 2000, \alpha = 0.05$.

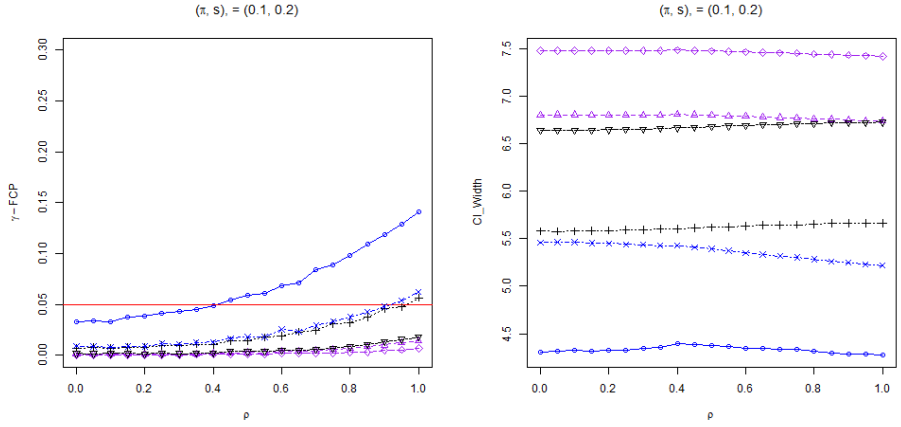

Figure 4.22 Estimated γ-FCP (left panel) and average width of CI (right panel) of our proposed unconditional CI-based γ-FCP controlling procedures and conditional CI-based γ-FCP controlling procedures along with Benjamini and Yekutieli procedures for two sample case under mixed correlation structure, with $\pi = 0.10$, $s = 0.2$. Here correlation coefficient ρ is from 0 to 1. $m = 2000, \alpha = 0.05$.

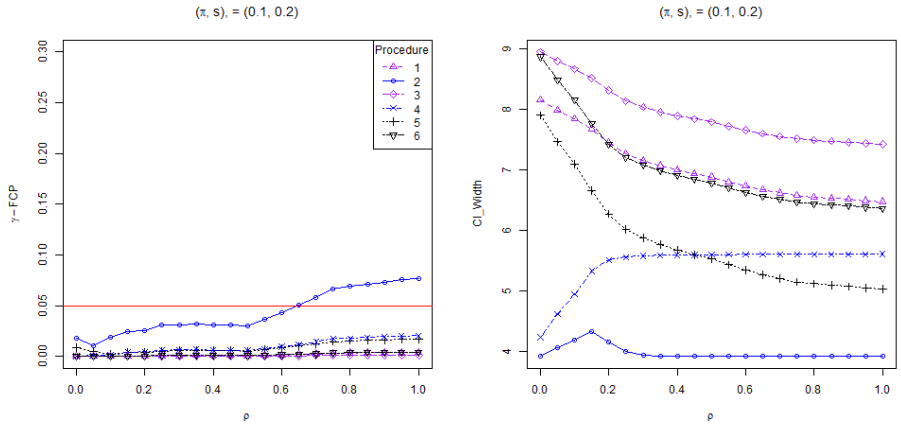

Figure 4.23 Estimated γ-FCP (left panel) and average width of CI (right panel) of our proposed unconditional CI-based γ-FCP controlling procedures and conditional CI-based γ-FCP controlling procedures along with Benjamini and Yekutieli procedures for two sample case under block-wise structure, block size is 200, with $\pi = 0.10$, $s = 0.2$. Here correlation coefficient ρ is from 0 to 1. $m = 2000, \alpha = 0.05$.

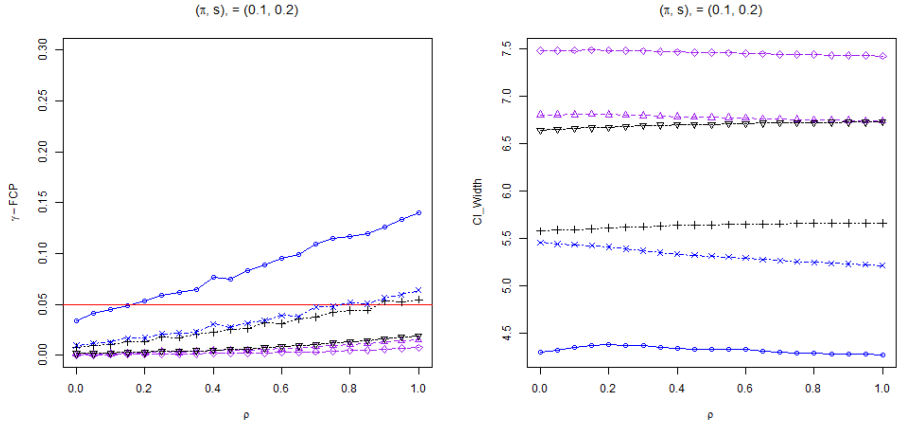

Figure 4.24 Estimated γ-FCP (left panel) and average width of CI (right panel) of our proposed unconditional CI-based γ-FCP controlling procedures and conditional CI-based γ-FCP controlling procedures along with Benjamini and Yekutieli procedures for two sample case under AR structure, with $\pi = 0.10$, $s = 0.2$. Here correlation coefficient ρ is from 0 to 1. $m = 2000, \alpha = 0.05$.

level 0.05 under equal dependence. Procedures 4 has the shortest CI width; at the same time it can control γ-FCP under equal correlation structure. Figure 4.22 shows the γ-FCP and average CI width under mixed correlation structure as ρ varies from 0 to 1. As seen from Figure 4.22, Procedures 1, 3 and Benjamini & Yekutieli can control γ-FCP at level 0.05 under mixed dependence. The other three procedures, including Procedure 2 can control γ-FCP at level 0.05 when ρ is moderate. Figure 4.23 displays the γ-FCP and average CI width under block-wise structure as ρ varies from 0 to 1. Procedures 1,3,4 and Benjamini & Yekutieli can control γ-FCP at level 0.05 under block-wise dependence. Procedure 2 can control γ-FCP at level 0.05 when ρ is moderate. Procedures 4 has the shortest CI width; at the same time it can control γ-FCP under block-wise structure. Figure 4.24 shows the γ-FCP and average CI width under AR structure as ρ varies from 0 to 1. Procedures 1,3,4 and Benjamini & Yekutieli can control γ-FCP at level 0.05 under AR dependence. Procedure 2 can control γ-FCP only when ρ is small. It worth to mention that Procedure 3 can control γ-FCP under all four types of dependence configurations, since it has the widest CI width. Therefore, under the various dependence structure, there is no such procedure which is always more powerful than the others. In different practical situations, one needs to choose the most suitable procedure.

4.7 Conclusion

Through the extensive simulation studies, we evaluate our proposed procedures with existing FCR-controlling procedures (BY2005). The simulation studies are performed, regarding to the estimated γ-FCP and average width of CI. Under the independence for both one sample case and two sample case, Procedure 2 is more powerful in the sense that Procedure 2 has shorter CI width and it is able to control γ-FCP at level α as well. Under the various dependence structure, there is no such procedure which is always more powerful than the others with proper control of

γ-FCP under dependence. In different practical situations, one needs to choose a most suitable procedure, for example, our proposed procedures (Procedures 1, 3 and 4) perform well under weak dependence such as block-wise dependence in terms of estimated γ-FCP, while they may lose γ-FCP control under strong dependence such as equal correlation in some scenario.

CHAPTER 5

REAL DATA ANALYSES

5.1 Introduction

In this chapter, we focus on real data analyses, while applying all of the proposed procedures and existing procedures (Benjamini and Yekutieli FCR controlling procedures) to two real data sets. One is microarray data of prostate cancer, the other one is microarray data of HIV. The two real data sets are available at the following website: http://statweb.stanford.edu/∼ckirby/brad/LSI/datasets-and-programs/datasets.html, which is a useful website to provide high dimensional data.

We are interested in the microarray data of prostate cancer and HIV, since both types of the disease are recognized as high risk in death (Dhanasekaran et al, 2001; Giri et al, 2006; Gallerano et al, 2015). Prostate cancer is one of the most common types of cancer in men, among the most heterogeneous of cancers. Acquired immune deficiency syndrome (AIDS) is caused by infection of HIV, which is commonly known as high risk in death as well. In the recent years, these two diseases have caused more than an estimated 35 million deaths worldwide. It is necessary to conduct genetic background check to detect the risk of a disease, as prostate cancer is not caused by one single gene. In fact, many different genes have been implicated to the disease. Microarray expression analysis has been used to identify genes that might anticipate the clinical behavior of the disease. Meanwhile, Wout et al.(2003) conduct a study on several classes of genes inhibited by HIV infection. Microarray analysis contribute to study of HIV host cell interactions and permitted identification of specific cellular pathways previously implicated in HIV infection.

For Sections 5.2 and 5.3, we construct CIs for our proposed procedures as well as BY FCR controlling procedures. Procedures 1 and 2 perform our proposed unconditional CI-based γ-FCP controlling procedures (independence, and dependence,

respectively), Procedures 3 and 4 perform our proposed conditional CI-based γ-FCP controlling procedures (independence, and dependence, respectively), Procedures 5 and 6 perform BY procedures (independence, and dependence, respectively). We conduct the analysis in four aspects: (a) the average width of CIs, (b) the count number of CIs not covering zero, (c) the average distance between CIs and zero, and (d) the average distance between CIs and zero only if the CIs are constructed by all 6 procedures. In Section 5.4, we summarize all the results.

5.2 Analysis of Microarray Data for Prostate Cancer

The principal goal of the study was to discover a small number of genes of interest (GOI), that is genes whose expression levels differs between the prostate and normal subjects. Once the genes are identified, the interest is to conduct CIs of GOI. The data containing genetic expression levels for $N = 6,033$ genes were obtained for $n = 102$ male objects, in which there are $n_1 = 50$ normal control subjects and $n_2 = 52$ prostate cancer patients. Let X_{ij} = expression level for gene i on patient j. Note that we calculate three descriptive statistics: $\bar{X}_{1,i}$, $\bar{X}_{2,i}$ and variance s_i^2, which are defined as:

$$\begin{aligned} \bar{X}_{1,i} &= \sum_{j=1}^{n_1} X_{ij}/n_1, \text{ and } \bar{X}_{2,i} = \sum_{j=n_1+1}^{n_1+n_2} X_{ij}/n_2, \\ s_i^2 &= \frac{\sum_{j=1}^{n_1} (X_{ij} - \bar{X}_{1,i})^2 + \sum_{j=n_1+1}^{n_1+n_2} (X_{ij} - \bar{X}_{2,i})^2}{n_1 + n_2 - 2}. \end{aligned} \tag{5.1}$$

Based on Equation 5.1, we generate two types of estimator:

$$\begin{aligned} S_i &= \bar{X}_{1,i} + \bar{X}_{2,i} \text{ for gene selection,} \\ Y_i &= \frac{\bar{X}_{2,i} - \bar{X}_{1,i}}{s_i \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \text{ for CI construction.} \end{aligned} \tag{5.2}$$

S_i is independent to Y_i, which implies distribution for $Y_i|S_i$ is same as unconditional Y_i, which is student t distribution. The GOI is determined in the following methods:

Table 5.1 Average Width of CIs by Procedure 1 to Procedure 6 for GOI of Prostate Cancer Microarray Data

Parameter		**Procedure**					
α	γ	1	3	5	2	4	6
0.010	0.05	7.776	**4.173**	6.180	8.627	**6.956**	7.400
0.025	0.05	7.319	**4.132**	5.613	8.213	**6.449**	6.921
0.050	0.05	6.956	**4.096**	5.151	7.887	**6.041**	6.538
0.100	0.05	6.575	**4.056**	4.652	7.548	**5.608**	6.135
0.010	0.10	7.437	**3.490**	6.180	8.367	**6.580**	7.400
0.025	0.10	6.960	**3.458**	5.613	7.940	**6.046**	6.921
0.050	0.10	6.579	**3.431**	5.151	7.604	**5.613**	6.538
0.100	0.10	6.179	**3.400**	4.652	7.253	**5.151**	6.135
0.010	0.20	7.079	**2.718**	6.180	8.090	**6.179**	7.400
0.025	0.20	6.569	**2.693**	5.613	7.650	**5.613**	6.921
0.050	0.20	6.170	**2.672**	5.151	7.302	**5.151**	6.538
0.100	0.20	5.755	**2.648**	4.652	6.938	**4.651**	6.135

select gene i if $S_i \notin (S_{i,0.10}, S_{i,0.90})$, where $S_{i,p}$ is the $100 * p\%$ quantile among all S_i. CIs of Procedure 1 - 6 are constructed for GOI. The number of GOI for the data is 1208 (out of 6033). We calculate CIs, based on four different levels of $\alpha = 0.01, 0.025, 0.05, 0.10$, and three different levels of $\gamma = 0.05, 0.10, 0.20$. The results for prostate are displayed in Tables 4.3 - 4.10.

In Table 5.1, Procedure 3 always provides the shortest width of CIs within all three independence Procedures 1, 3 and 5. At the same time, Procedure 4 provides the shortest width of CIs within all three dependence Procedures 2, 4 and 6. To sum up all Procedures 1 - 6, Procedure 3 always provides the shortest width of CIs among

Table 5.2 Number of CIs not Covering Zero by Procedure 1 to Procedure 6 for GOI of Prostate Cancer Microarray Data

Parameter		**Procedure**					
α	γ	1	3	5	2	4	6
0.010	0.05	2	**75**	10	1	**4**	2
0.025	0.05	2	**77**	15	1	**6**	4
0.050	0.05	4	**78**	19	2	**13**	6
0.100	0.05	6	**80**	38	2	**15**	10
0.010	0.10	2	**138**	10	1	**6**	2
0.025	0.10	4	**143**	15	2	**13**	4
0.050	0.10	6	**151**	19	2	**15**	6
0.100	0.10	10	**155**	38	2	**19**	10
0.010	0.20	3	**275**	10	2	**10**	2
0.025	0.20	6	**278**	15	2	**15**	4
0.050	0.20	10	**280**	19	2	**19**	6
0.100	0.20	14	**287**	38	4	**38**	10

all procedures. For the genetics study, estimated CIs of GOI are expected to exclude from zero, which means scientific effectiveness.

In Table 5.2, it is obviously to see that Procedure 3 can provide the most large count number of CIs not covering zero among all three independence Procedures 1, 3 and 5. Procedure 4 can offer the most large number of CIs not covering zero for all dependence Procedures 2, 4 and 6. Combined the results in Table 5.2, Procedure 3 is over-perform against the others in the sense for the amount of nonzero CIs.

Given $\alpha = 0.010, 0.025$, Procedure 3 in Table 5.3 has the most far-away distance to zero among all three independence Procedures 1, 3 and 5. Given $\alpha = 0.050, 0.10$, Procedure 5 has bigger far-away distance to zero than Procedure 3. Meanwhile, for

Table 5.3 Distance between CIs and Zero by Procedure 1 to Procedure 6 for GOI of Prostate Cancer Microarray Data

Parameter		**Procedure**					
α	γ	1	3	5	2	4	6
0.010	0.05	0.4997	**0.6319**	0.5863	0.3116	**0.4899**	0.4455
0.025	0.05	0.4729	**0.6358**	0.6238	0.4545	**0.5527**	0.4983
0.050	0.05	0.4902	0.6299	**0.6540**	0.4444	**0.5870**	0.5294
0.100	0.05	0.5338	0.6387	**0.6460**	0.4729	**0.6267**	0.6018
0.010	0.10	0.4533	**0.6487**	0.5863	0.4106	**0.5390**	0.4455
0.025	0.10	0.4880	**0.6441**	0.6238	0.4174	**0.5845**	0.4983
0.050	0.10	0.5393	0.6346	**0.6540**	0.4796	**0.6240**	0.5294
0.100	0.10	0.5868	0.6356	**0.6459**	0.4373	**0.6542**	0.6018
0.010	0.20	0.4380	**0.6306**	0.5863	0.4081	**0.5865**	0.4455
0.025	0.20	0.5393	**0.6328**	0.6238	0.4951	**0.6240**	0.4983
0.050	0.20	0.5868	0.6315	**0.6540**	0.4338	**0.6542**	0.5294
0.100	0.20	0.6107	0.6302	**0.6459**	0.4899	**0.6461**	0.6018

all three dependence Procedures 2, 4 and 6, Procedure 4 is outstanding than the other two as shown in Table 5.3. We also compared all the six procedure together. 6/12 combination of α and γ has Procedure 3 as the most far-away distance to zero among all the procedures, and 3/12 combination of α and γ has Procedure 4 as the most far-away distance to zero among all the procedures, and 3/12 combination of α and γ has Procedure 5 as the most far-away distance to zero among all the procedures.

There is no absolute winner regarding the average distance between all the constructed CIs and zero. Hence, we further study the average distance between the constructed CIs and zero if the CIs are commonly not covering zero. As shown in Table 5.4, Procedure 3 provides the longest distance between the constructed CIs

Table 5.4 Distance between CIs and Zero by Procedure 1 to Procedure 6 for Commonly Selected and Nonzero CIs of GOI for Prostate Cancer Microarray Data

Parameter		**Procedure**					
α	γ	1	3	5	2	4	6
0.010	0.05	0.7370	**2.5384**	1.5351	0.3116	**1.1468**	0.9251
0.025	0.05	0.9014	**2.4950**	1.7542	0.4546	**1.3365**	1.1004
0.050	0.05	0.9098	**2.3395**	1.8122	0.4444	**1.3673**	1.1184
0.100	0.05	0.9593	**2.2187**	1.9208	0.4729	**1.4428**	1.1790
0.010	0.10	0.8757	**2.8491**	1.5041	0.4106	**1.3039**	0.8941
0.025	0.10	0.9076	**2.6586**	1.5810	0.4174	**1.3648**	0.9272
0.050	0.10	0.9917	**2.5661**	1.7060	0.4796	**1.4750**	1.0123
0.100	0.10	0.9744	**2.3640**	1.7379	0.4373	**1.4886**	0.9961
0.010	0.20	0.9140	**3.0944**	1.3635	0.4081	**1.3636**	0.7535
0.025	0.20	1.0305	**2.9736**	1.5136	0.4951	**1.5137**	0.8598
0.050	0.20	0.9953	**2.7486**	1.5093	0.4338	**1.5094**	0.8155
0.100	0.20	1.0815	**2.6349**	1.6329	0.4899	**1.6330**	0.8911

and zero within the common selected ones for all three independence Procedures 1, 3 and 5. And Procedure 4 have the longest distance between the constructed CIs and zero within the common selected ones for all three dependence Procedures 2, 4 and 6. Moreover, for all the combination in our analysis, Procedure 3 is the most far-away distance to zero among all the procedures.

5.3 Analysis of Microarray Data for HIV

Now we consider the microarray data for HIV, which contains genetic expression levels for $N = 7,688$ genes were obtained for $n_1 = 4$ normal control subjects and $n_2 = 4$ HIV patients. The principal goal of the study is to discover a small number of genes

of interest (GOI), that is, genes whose expression levels differs between the prostate and normal subjects. Once the genes are identified, the interest is to conduct the CIs of GOI. Let X_{ij} = expression level for gene i on patient j. Note that we calculate three descriptive statistics: $\bar{X}_{1,i}$, $\bar{X}_{2,i}$ and pooled variance s_i^2, which are same defined as Equation 5.1. But if we keep using Equation 5.2 in the previous prostate study, the central histogram is less dispersed than a standard normal distribution as shown in Figure 5.1. Hence, we use a different CI estimator: logarithm of Y_i in Equation 5.2. Specially, we generate two types of estimator:

$$\begin{aligned} S_i &= \bar{X}_{1,i} + \bar{X}_{2,i} \text{ for gene selection,} \\ Y_i &= \log(\frac{\bar{X}_{2,i} - \bar{X}_{1,i}}{s_i}) \text{ for CI construction.} \end{aligned} \tag{5.3}$$

The GOI is determined in the following methods: select gene i if $S_i \nsubseteq (S_{i,0.10}, S_{i,0.90})$, where $S_{i,p}$ is the $100{*}p\%$ quantile among all S_i. CIs of Procedures 1 - 6 are constructed for GOI. The selection size is 1544 (out of 7680). We construct CIs for Procedures 1-6. And we construct CIs, based on four different $\alpha = 0.01, 0.025, 0.05, 0.10$, and three different $\gamma = 0.05, 0.10, 0.20$. The results for HIV data are displayed in Tables 5.5 - 5.8.

In Table 5.5, Procedure 3 always provides the shortest width of CIs among all three independence procedures, Procedures 1, 3 and 5. Meanwhile Table 5.5 shows that Procedure 4 has the shortest width of CIs among all three dependence procedures, Procedures 2, 4 and 6. Considering all six procedures, Procedure 3 shows the shortest width of CIs.

From Table 5.6, Procedure 3 has the most count number of CIs not covering zero among all three independence procedures, Procedures 1, 3 and 5. And Table 5.6 points out Procedure 4 provides the most count number of CIs not covering zero among all three dependence procedures, Procedure 2, 4 and 6. Considering all six procedures, Procedure 3 has the most number of CIs not covering zero.

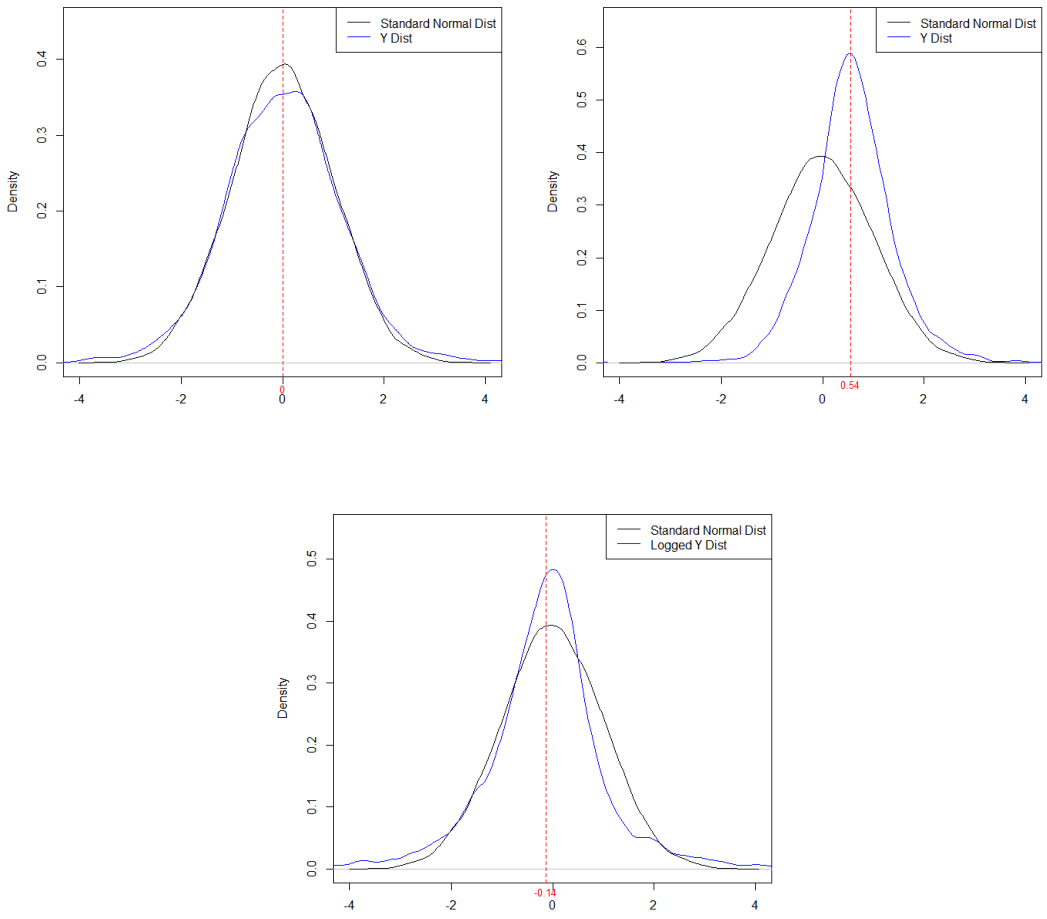

Figure 5.1 Distribution of standard normal and distribution of the estimator Y for prostate microarray data (first row left panel), distribution of standard normal and the estimator Y for HIV microarray data (first row right panel), distribution of standard normal and the logarithm of estimator Y for HIV microarray data (second row).

Table 5.5 Average Width of CIs by Procedure 1 to Procedure 6 for GOI of HIV Microarray Data

Parameter		**Procedure**					
α	γ	1	3	5	2	4	6
0.010	0.05	7.780	**4.149**	6.180	8.647	**6.960**	7.413
0.025	0.05	7.323	**4.112**	5.614	8.234	**6.453**	6.935
0.050	0.05	6.960	**4.081**	5.152	7.909	**6.045**	6.554
0.100	0.05	6.580	**4.045**	4.653	7.571	**5.612**	6.151
0.010	0.10	7.437	**3.465**	6.180	8.382	**6.580**	7.413
0.025	0.10	6.960	**3.437**	5.614	7.957	**6.045**	6.935
0.050	0.10	6.580	**3.413**	5.152	7.621	**5.612**	6.554
0.100	0.10	6.179	**3.386**	4.653	7.271	**5.151**	6.151
0.010	0.20	7.079	**2.699**	6.180	8.105	**6.179**	7.413
0.025	0.20	6.580	**2.677**	5.614	7.666	**5.612**	6.935
0.050	0.20	6.179	**2.658**	5.152	7.318	**5.150**	6.554
0.100	0.20	5.755	**2.637**	4.653	6.954	**4.651**	6.151

Table 5.6 Number of CIs not Covering Zero by Procedure 1 to Procedure 6 for GOI of HIV Microarray Data

Parameter		**Procedure**					
α	γ	1	3	5	2	4	6
0.010	0.05	20	**120**	45	13	**28**	21
0.025	0.05	21	**123**	59	19	**40**	28
0.050	0.05	28	**126**	72	19	**46**	37
0.100	0.05	37	**128**	92	20	**59**	45
0.010	0.10	21	**171**	45	17	**37**	21
0.025	0.10	28	**174**	59	19	**46**	28
0.050	0.10	37	**177**	72	20	**59**	37
0.100	0.10	45	**183**	92	22	**72**	45
0.010	0.20	27	**262**	45	19	**45**	21
0.025	0.20	37	**264**	59	20	**59**	28
0.050	0.20	45	**267**	72	21	**72**	37
0.100	0.20	53	**270**	92	28	**92**	45

In Table 5.7, 11/12 combination of α and γ has Procedure 3 as the most far-away distance to zero among all independence procedures, Procedures 1, 3 and 5. Except for the combination $\alpha = 0.10$ and $\gamma = 0.05$, Procedure 5 has slightly advantage against Procedure 3. Table 5.7 shows that 8/12 combination of α and γ has Procedure 4 as the most far-away distance to zero among all dependence procedures, Procedures 2, 4 and 6. Combined all six procedures, 11/12 combination of α and γ has Procedure 3 as the most far-away distance to zero.

Table 5.8 shows the average distance between constructed CIs and zero if the CIs are commonly not covering zero for all procedures. Table 5.8 demonstrate that Procedure 3 is the most far-away distance to zero among there independence

Table 5.7 Distance between CIs and Zero by Procedure 1 to Procedure 6 for GOI of HIV Microarray Data

Parameter		**Procedure**					
α	γ	1	3	5	2	4	6
0.010	0.05	0.6189	**0.8103**	0.6903	0.4917	0.6918	**0.7267**
0.025	0.05	0.7589	**0.8065**	0.7312	0.4715	0.6565	**0.7015**
0.050	0.05	0.6918	**0.8010**	0.7866	0.6000	0**7307**	0.6662
0.100	0.05	0.6552	0.8043	**0.8059**	0.7025	0.**7320**	0.7022
0.010	0.10	0.7184	**0.8320**	0.6903	0.4675	0.6552	**0.7267**
0.025	0.10	0.6918	**0.8303**	0.7312	0.5811	0.**7307**	0.7015
0.050	0.10	0.6552	**0.8270**	0.7866	0.6826	0.**7320**	0.6662
0.100	0.10	0.6909	**0.8120**	0.8059	0.7428	0.**7874**	0.7022
0.010	0.20	0.6697	**0.8179**	0.6903	0.5225	0.6909	**0.7267**
0.025	0.20	0.6552	**0.8206**	0.7312	0.6645	0.**7320**	0.7015
0.050	0.20	0.6909	**0.8191**	0.7866	0.7609	0.**7874**	0.6662
0.100	0.20	0.7501	**0.8187**	0.8059	0.6940	0.**8068**	0.7022

procedures, Procedures 1, 3 and 5. And Procedure 4 has the most far-away distance to zero among there dependence procedures, Procedures 2, 4 and 6, as shown in Table 5.8.

5.4 Conclusion

All the results and analysis about real data are summarized in Table 5.9, in which we show that either Procedure 3 or Procedure 4 have advantage among the procedures under independence or arbitrary dependence, receptively. We can figure out some findings: (1) Procedure 3 has shortest average width of CIs , and (2) Procedure 3 has the most count number of nonzero covering CIs, and (3) there is no absolute winner

Table 5.8 Distance between CIs and Zero by Procedure 1 to Procedure 6 for Commonly Selected and Nonzero CIs of GOI for HIV Microarray Data

Parameter		**Procedure**					
α	γ	1	3	5	2	4	6
0.010	0.05	1.2805	**3.0959**	2.0802	0.8468	**1.6904**	1.4638
0.025	0.05	1.2291	**2.8345**	2.0837	0.7737	**1.6642**	1.4231
0.050	0.05	1.2888	**2.7282**	2.1930	0.8145	**1.7463**	1.4920
0.100	0.05	1.2898	**2.5569**	2.2532	0.7942	**1.7734**	1.5039
0.010	0.10	1.2573	**3.2430**	1.8854	0.7845	**1.6858**	1.2690
0.025	0.10	1.2989	**3.0604**	1.9720	0.8007	**1.7564**	1.3113
0.050	0.10	1.3058	**2.8890**	2.0197	0.7853	**1.7894**	1.3188
0.100	0.10	1.3297	**2.7264**	2.0928	0.7839	**1.8443**	1.3435
0.010	0.20	1.3281	**3.5182**	1.7773	0.8150	**1.7780**	1.1608
0.025	0.20	1.3386	**3.2900**	1.8214	0.7956	**1.8223**	1.1608
0.050	0.20	1.3840	**3.1443**	1.8976	0.8146	**1.8984**	1.1966
0.100	0.20	1.3956	**2.9546**	1.9466	0.7958	**1.9476**	1.1973

Table 5.9 Summary for the Independence Procedures and Dependence Procedures, Which Has More Far away Distance from Zero than the Other Procedures

Table	Independence procedures	Dependence procedures
1	Procedure 3	Procedure 4
2	Procedure 3	Procedure 4
3	Procedure 3*	Procedure 4
4	Procedure 3	Procedure 4
5	Procedure 3	Procedure 4
6	Procedure 3	Procedure 4
7	Procedure 3**	Procedure 4***
8	Procedure 3	Procedure 4

Note*: Among all 12 combinations of (α, γ), 6 combination shows that Procedures 3 has more far away distance from zero than other independence procedures.

Note**: Among all 12 combinations of (α, γ), 11 combination shows that Procedure 3 has more far away distance from zero than other independence procedures.

Note***: Among all 12 combinations of (α, γ), 8s combination shows that Procedure 4 has more far away distance from zero than other dependence procedures.

when comparing the distance between CIs and zero for all selected ones as shown in Table 5.9, and (4) Procedure 3 is more likely to provide a CIs away from zero for commonly selected ones among all six procedures. Considering the theoretical results, if one can control γ-FCP, then a method which can provide a shorter nonzero covering CIs, as well as longer distance between CIs and zero, is preferred.

CHAPTER 6

CONCLUSION AND FUTURE WORK

In this dissertation, a new concept, γ-FCP, is presented as a measure of simultaneous coverage for the multiple CIs following the selection. We suggest four new and powerful procedures: (i) an unconditional CI-based procedure, (ii) a modified unconditional CI-based procedure, (iii) a conditional CI-based procedure, and (iv) a modified conditional CI-based procedure. Among all of the new procedures, (i) is developed to control γ-FCP under PRDS/independence; (ii) and (iv) are developed to control γ-FCP under arbitrary dependence; and (iii) is developed to control γ-FCP under independence.

We evaluate the performance of our proposed procedures via the extensive simulation studies in terms of estimated γ-FCP and average width of CIs. The effect of nonzero proportion, selection level, and correlation coefficient are evaluated, while we apply the proposed procedures in terms of γ-FCP control and average width of CIs. The simulation studies are applied to strong dependence such as equal correlation and several weak dependences such as block-wise dependence. Our simulation studies are able to show that the proposed procedures are able to either control γ-FCP or have shorter width of CIs than existing methods such as FCR controlling procedures (Benjamini and Yekutieli, 2005). Next, all of the proposed procedures are applied on two sets of micro-array gene expression data. Compared to same existing methods, the proposed conditional CI-based procedure is demonstrated to provide (i) shorter width of CI; and (ii) more count of CI not covering zero; and (iii) longer distance of CI away from zero.

A potential work is to present a weighted CI-based procedure, which can keep γ-FCP at a desired level. With the Poisson binomial distribution, we have some theoretical results for weighted conditional CI γ-FCP controlling procedure under

independence condition. But it is not computationally efficient. In the future, (1) we plan to extent our current result from independence to arbitrary dependence; and (2) we are interested in further studying and developing a more explicit and computational procedure; and (3) besides the conditional CI based procedures, we want to develop a weighted unconditional CI based procedure to control γ-FCP at a desired level under (i) independence and (ii) arbitrary dependence structure.

APPENDIX A

PROOFS FOR SELECTIVE INFERENCE

Proof of Lemma 3.1. To begin with the discussion, we will first introduce a idea: stochastic ordering and its related properties to help us learn about $V_S(u)$.

Definition A.1 (Stochastic Ordering)**.** *A random variable X is less than a random variable Y in the stochastic order if*

$$P(X > x) \leq P(Y > x) \text{ for all } x \in (-\infty, \infty),$$

This is denoted as $X \leq_{st} Y$.

With Definition A.1, we can learn about the comparison between two random variables. Hence we can figure out more information about the random variable $V_S(u)$.

Lemma A.1 (Klenke and Mattner, 2010)**.** *Let X, Y be two random variable, where $X \sim Bin(n, p_1)$ and $Y \sim Bin(n, p_2)$. If $p_1 \leq p_2$ then $X \leq_{st} Y$.*

Lemma A.2. *If $X \leq_{st} Y$ and $f(\cdot)$ is a non-decreasing function, then $E(f(X)) \leq E(f(Y))$.*

With these two lemma, we can further compare any random variable which follows a binomial distribution.

Proof. We will prove this Lemma by contradiction. Assume there exists $\alpha' > \alpha$ such that $u(\alpha', \gamma, |S|) \leq u(\alpha, \gamma, |S|)$, which can be denoted as $u' \leq u$. Based on the condition $\hat{S} = S$, Then

$$V_S(u') \sim Bin(|S|, u'), \text{ and } V_S(u) \sim Bin(|S|, u).$$

According to Lemma A.1, $V_S(u') \leq_{st} V_S(u)$. Then,

$$\begin{aligned}\alpha &= P(V_S(u) \geq \lfloor \gamma|S| \rfloor + 1 | \hat{S} = S) \\ &= E(I(V_S(u) \geq \lfloor \gamma|S| \rfloor + 1 | \hat{S} = S)) \\ &\geq E(I(V_S(u') \geq \lfloor \gamma|S| \rfloor + 1 | \hat{S} = S)) \\ &= P(V_S(u') \geq \lfloor \gamma|S| \rfloor + 1 | \hat{S} = S) = \alpha',\end{aligned}$$

which leads to a contradiction. Here, the inequality holds because of the Lemma A.2. □

We can assume the foregoing conditional CIs procedure is monotone in the conditional confidence level:

$$\alpha \geq \alpha' \text{ implies that } cCI_i(\alpha) \subseteq cCI_i(\alpha'). \tag{A.1}$$

Proof of Lemma 3.2.

Proof. We will prove this Lemma by contradiction. Assume there exist $\gamma' > \gamma$ such that $u(\alpha, \gamma', |S|) < u(\alpha, \gamma, |S|)$, for convince, we use the short term $u' < u$. Based on the condition $\hat{S} = S$, Then

$$V_S(u') \sim Bin(|S|, u'), \text{ and } V_S(u) \sim Bin(|S|, u).$$

According to Lemma 3.1, $V_S(u') \leq_{st} V_S(u)$. Then,

$$\begin{aligned}\alpha &= P(V_S(u) \geq \lfloor \gamma|S| \rfloor + 1 | \hat{S} = S) \\ &= E(I(V_S(u) \geq \lfloor \gamma|S| \rfloor + 1 | \hat{S} = S)) \\ &> E(I(V_S(u') \geq \lfloor \gamma|S| \rfloor + 1 | \hat{S} = S)) \\ &= P(V_S(u') \geq \lfloor \gamma|S| \rfloor + 1 | \hat{S} = S) \\ &\geq P(V_S(u') \geq \lfloor \gamma'|S| \rfloor + 1 | \hat{S} = S) = \alpha,\end{aligned}$$

which leads to a contradiction. Here, the first inequality holds because of the Lemma 3.1. □

Remark A.1. *$u(\alpha, \gamma, |S|)$ is not a monotone function in $|S|$. But if we fix $\lfloor \gamma|S| \rfloor$, then $u(\alpha, \gamma, |S|)$ is a monotone function in $|S|$. Since $\lfloor \gamma|S| \rfloor = i - 1$ is equivalent to $|S| \in (\frac{i}{\gamma} - \frac{1}{\gamma}, \frac{i}{\gamma} - 1)$, where $i = 1, 2, ...,$ in which case γ decrease as $|S|$ increase. Consider that $u(\alpha, \gamma, |S|)$ is a nondecreasing function in γ. And thus $u(\alpha, \gamma, |S|)$ is a non-increasing function in $|S|$ when $\lfloor \gamma|S| \rfloor$ is fixed.*

Lemma A.3. *If u satisfy Equation (3.1) and we can construct cCI(u) for μ conditional on $\hat{S} = S$ such that*

$$P(\mu \notin cCI(u)|\hat{S} = S) \leq u.$$

Then γ-cFCP$_S$ in Definition 3.1 can be controlled.

Before introducing conditional CIs, we assume to construct exact adjusted level u, namely $P(\mu_i \notin cCI_i(u)|\hat{S} = S) = u$. To solve Equation (3.1) numerically may result in the fact u is a approximation value, namely $P(\mu_i \notin cCI_i(u)|\hat{S} = S) \leq u$ instead of $P(\mu_i \notin cCI_i(u)|\hat{S} = S) = u$. Hence we wonder whether γ-cFCP$_S$ can be controlled for the case that conditional CIs with adjusted level u such that $P(\mu_i \notin cCI_i(u)|\hat{S} = S) \leq u$.

Proof. Assume we construct cCI(u) and cCI(u') for μ_i conditional on $\hat{S} = S$ such that

$$\begin{aligned} P(\mu_i \notin cCI_i(u)|\hat{S} = S) &= u, \\ P(\mu_i \notin cCI_i(u')|\hat{S} = S) &= u' \leq u, \end{aligned} \tag{A.2}$$

where γ-cFCP$= P(V_S(u) \geq \lfloor \gamma|S| \rfloor + 1) = \alpha$. Applying Lemma A.1, according to Equation 3.6, the corresponding $V_S(u) \geq_{st} V_S(u')$. Then by applying Lemma A.2, we have

$$P(V_S(u') \geq \lfloor \gamma|S| \rfloor + 1) \leq P(V_S(u) \geq \lfloor \gamma|S| \rfloor + 1) = \alpha.$$

Thus, the desired result follows. □

Proof of Lemma 3.3 We want to introduce an interesting fact that u tends to be a constant when the size of selection is large, and such constant is γ.

Proof. We will prove this Lemma by contradiction. Assume there exists a subsequence $\{u_{n_j}\}$ of $\{u_n\}$ such that $\lim_{n_j\to\infty} u = c \neq \gamma$. First, let $c > \gamma$, for $0 < \epsilon < \frac{c-\gamma}{2}$:

$$\exists \text{ a large number } N \text{ such that for } n_j \geq N, u_{n_j} > c - \epsilon > \gamma + \epsilon.$$

Let $W_n \sim Bin(n, \gamma + \epsilon)$. When $n_j \geq N, V_{n_j} \geq_{st} W_{n_j}$. And hence

$$\alpha = P(V_{n_j} \geq \gamma n_j) \geq P(W_{n_j} \geq \gamma n_j) = P(\frac{W_{n_j}}{n_j} \geq \gamma) \to 1, \, as \; n_j \to \infty,$$

which leads to a contradiction. The first inequality follows from Lemma A.2. The last equality holds due to the strong law of large number, that is $P(\lim_{n_j\to\infty} \frac{W_{n_j}}{n_j} = \gamma + \epsilon) = 1$. Second, let $c < \gamma$, for $0 < \epsilon < \frac{\gamma-c}{2}$:

$$\exists \text{ a large number } N \text{ such that for } n_j \geq N, u_{n_j} < c + \epsilon < \gamma - \epsilon.$$

Let $W_n \sim Bin(n, \gamma - \epsilon)$. When $n_j \geq N, V_{n_j} \leq_{st} W_{n_j}$.

$$\alpha = P(V_{n_j} \geq \gamma n_j) \leq P(W_{n_j} \geq \gamma n_j) = P(\frac{W_{n_j}}{n_j} \geq \gamma) \to 0, \, as \; n_j \to \infty,$$

which leads to a contradiction. The first inequality follows from Lemma A.2. The last equality holds due to the strong law of large number, that is $P(\lim_{n_j\to\infty} \frac{W_{n_j}}{n_j} = \gamma - \epsilon) = 1$. To sum up, $\lim_{|S|\to\infty} u = \gamma$. □

Discussion about u versus $\frac{\lfloor \gamma|S| \rfloor + 1}{|S|}\alpha$

Figures A.1 and A.2 draw the ratio $r = \frac{u}{\frac{\lfloor \gamma|S| \rfloor + 1}{|S|}\alpha}$ versus α. Figures A.3 and A.4 draw the ratio r versus γ. Figures A.5 and A.6 draw the ratio r versus $|S|$.

The settings of Figures A.1 to A.6 are defined as fellowing. Figure A.1: $\gamma = 0.1$. This is done for four values of $|S|$: $5, 20, 80$ and 500. Figure A.2: $|S| = 20$. This is done for four values of γ : $0.05, 0.1, 0.15, 0.2$. Figure A.3: $\alpha = 0.05$. This is done

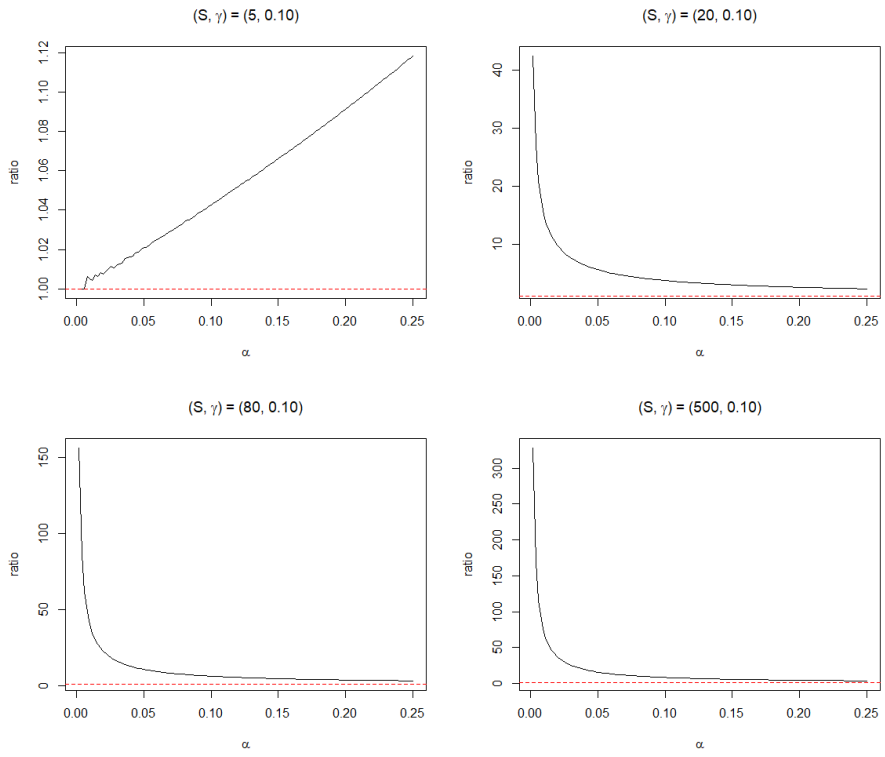

Figure A.1 Ratio r (black line) versus α with $\gamma = 0.1$. The red line shows $r = 1$. This is done for four values of $|S|$: 5 (first row left panel), 20 (first row right panel), 80 (second row left panel), 500 (second row right panel).

for four values of $|S|$: $5, 20, 80, 500$. Figure A.4: $|S| = 20$. This is done for four values of α : 0.05,0.1,0.15,0.2. Figure A.5: $\alpha = 0.05$. This is done for four values of γ : 0.05,0.10,0.15,0.2. Figure A.6: $\gamma = 0.10$. This is done for four values of α : 0.05,0.1,0.15,0.20.

The numerical studies have the following conclusion: Figures A.1 to A.2 shows that (1) given the $\gamma, |S|$ are fixed, smaller α results in larger ratio r, which in turns means $u(\alpha, \gamma, |S|)$ is greater than $\frac{\lfloor \gamma|S| \rfloor + 1}{|S|}\alpha$; (2) given the $\gamma, |S|$ are fixed as value 0.10 and 5 respectively, the ratio r not only increase as α increase, but also is greater than 1; (3) given the $\gamma, |S|$ are fixed as value 0.10 (or 0.05, or 0.15, or 0.20) and 20 (or 80, or 500) respectively, the ratio r not only decrease as α increase, but also is greater than 1.

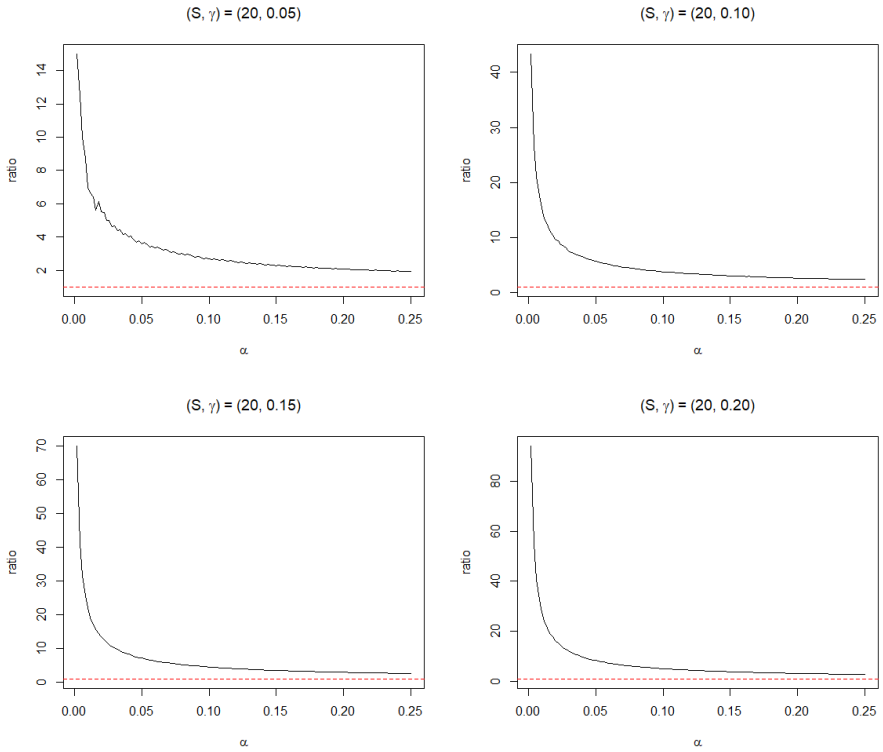

Figure A.2 Ratio r (black line) versus α with $|S| = 20$. The red line shows $r = 1$. This is done for four values of γ: 0.05 (first row left panel), 0.1 (first row right panel), 0.15 (second row left panel), 0.2 (second row right panel).

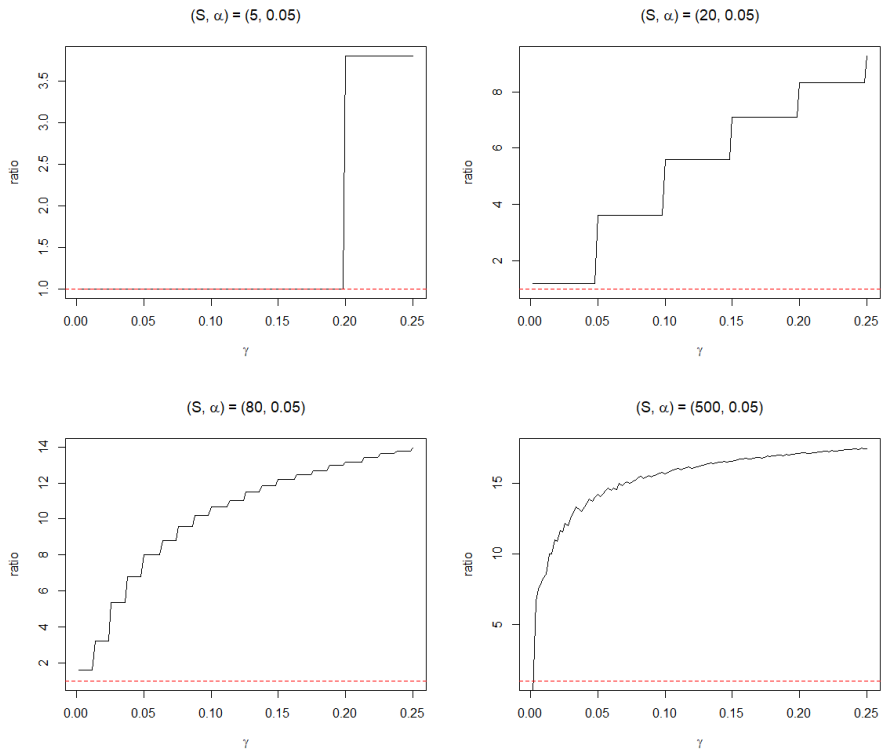

Figure A.3 Ratio r (black line) versus γ with $\alpha = 0.05$. The red line shows $r = 1$. This is done for four values of $|S|$: 5 (first row left panel), 20 (first row right panel), 80 (second row left panel), 500 (second row right panel).

Figures A.3 to A.4 show that (1) given $\alpha, |S|$ are fixed, larger γ results in larger ratio r, which in turns means $u(\alpha, \gamma, |S|)$ is greater than $\frac{\lfloor \gamma |S| \rfloor + 1}{|S|}\alpha$; (2) given $\alpha, |S|$ are fixed as value 5 and 0.05 respectively, $r = 1$ when $0 \leq \gamma \leq 1$; (3) given $\alpha, |S|$ are fixed, the ratio r does not only increase as α increases, but also is greater than or equal to 1. Figures A.5 to A.6 show that (1) given α, γ are fixed, larger $\|S|$ results in larger ratio r, which in turns means $u(\alpha, \gamma, |S|)$ is greater than $\frac{\lfloor \gamma |S| \rfloor + 1}{|S|}\alpha$; (2) given α, γ are fixed, the ratio r are not only increase as α increase, but also is greater than or equal to 1.

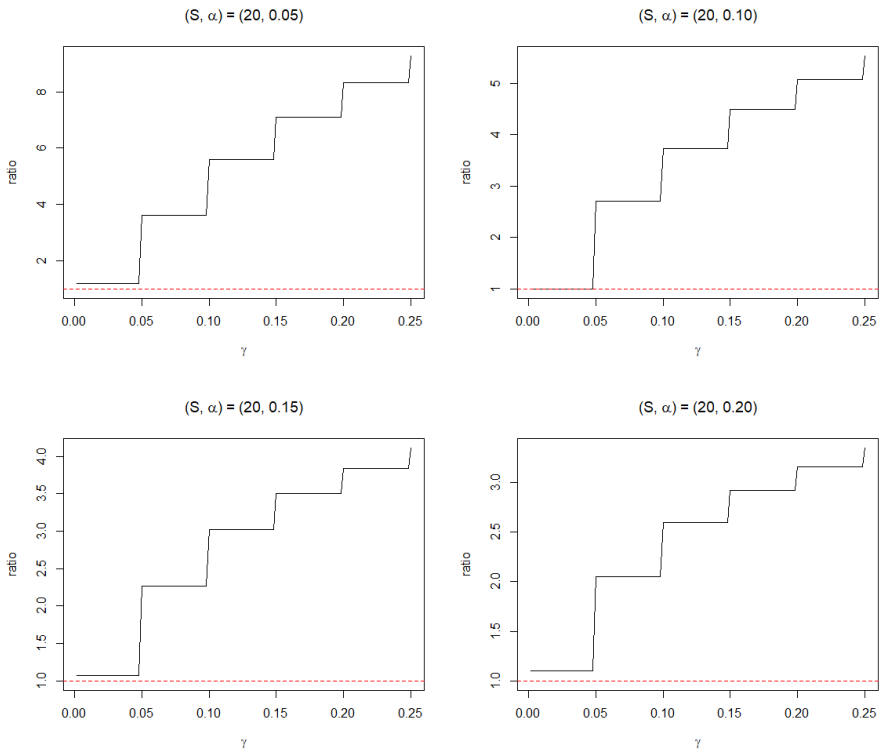

Figure A.4 Ratio r (black line) versus γ, $|S| = 20$. The red line shows $r = 1$. This is done for four values of α: 0.05 (first row left panel), 0.1 (first row right panel), 0.15 (second row left panel), 0.2 (second row right panel).

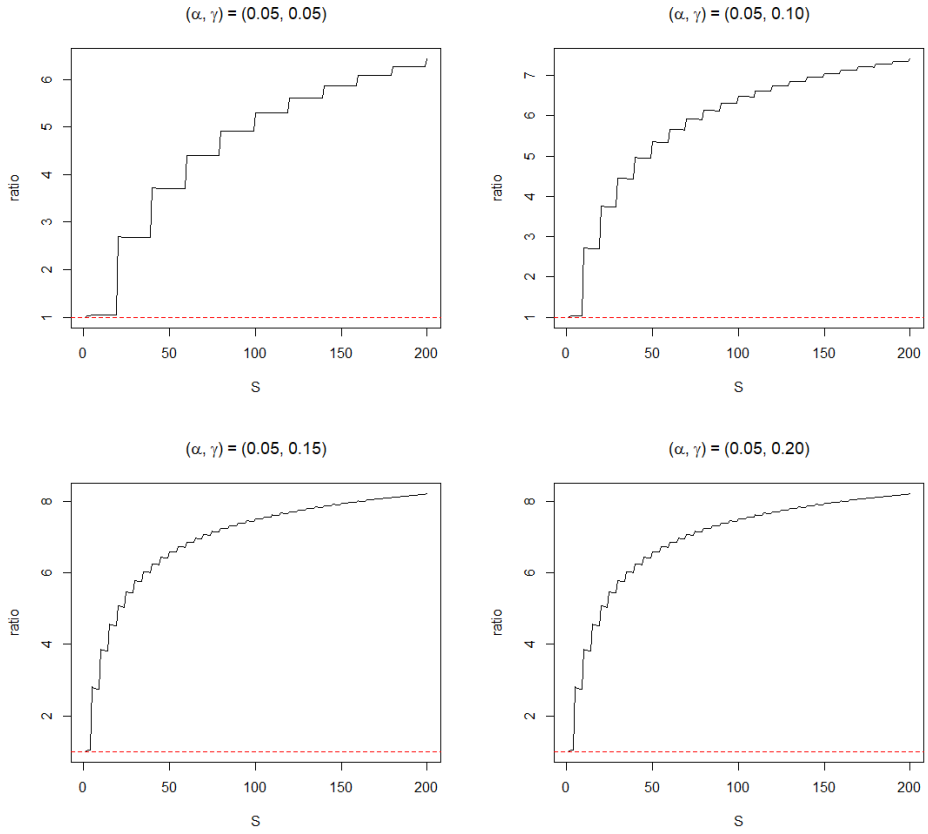

Figure A.5 Ratio r (black line) versus $|S|$ with $\alpha = 0.05$. The red line shows $r = 1$. This is done for four values of γ: 0.05 (first row left panel), 0.1 (first row right panel), 0.15 (second row left panel), 0.2 (second row right panel).

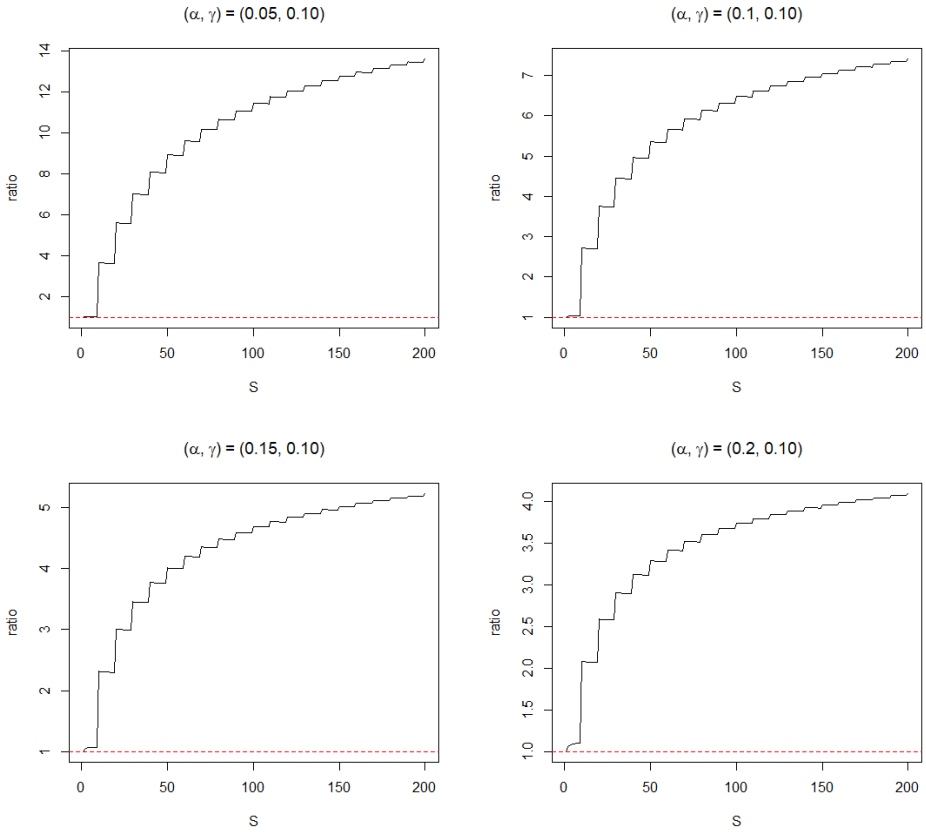

Figure A.6 Ratio r (black line) versus $|S|$ with $\gamma = 0.10$. The red line shows $r = 1$. This is done for four values of α: 0.05 (first row left panel), 0.1 (first row right panel), 0.15 (second row left panel), 0.2 (second row right panel).

Methods of Constructing Conditional Confidence Intervals.

The following will discuss the three cases as above for any dependence in details. Let w_1, w_2 be two weighed values, which we can pre-specify $w_1 \neq 1, w_2 \neq 1$. Informally, the values of w_1 and w_2 show the dependence between Y and T. We will summarize how we can construct conditional CI of $\mu_1 - \mu_2$ for the first case under dependence in details.

Details for Case 1: Known variance $\sigma_1^2 = \sigma_2^2 = \sigma^2$. We define the selection estimator Y and a CI estimator T as

$$Y = w_1\bar{X}_1 + \bar{X}_2,$$
$$T = w_2\bar{X}_1 - \bar{X}_2.$$

Hence (Y, T) follows a bivariate normal distribution,

$$\begin{pmatrix} Y \\ T \end{pmatrix} \sim \mathcal{N}\left[\begin{pmatrix} w_1\mu_1 + \mu_2 \\ w_2\mu_1 - \mu_2 \end{pmatrix}, \begin{pmatrix} \frac{w_1^2}{n_1} + \frac{1}{n_2} & \frac{w_1 w_2}{n_1} - \frac{1}{n_2} \\ \frac{w_1 w_2}{n_1} - \frac{1}{n_2} & \frac{w_2^2}{n_1} + \frac{1}{n_2} \end{pmatrix} \sigma^2 \right].$$

The conditional distribution of $T|Y = s$ follows a normal distribution $\mathcal{N}(\mu_{T|Y=s}, \sigma_{T|Y=s})$. Let $r = n_1/n_2$. Then,

$$\begin{aligned}
\mu_{T|Y=s} &= \mu_T + \rho\frac{\sigma_T}{\sigma_Y}(s - \mu_Y) \\
&= w_2\mu_1 - \mu_2 + \frac{w_1 w_2 - r}{\sqrt{(w_1^2 + r)(w_2^2 + r)}} \frac{\sqrt{\frac{w_2^2}{n_1} + \frac{1}{n_2}}}{\sqrt{\frac{w_1^2}{n_1} + \frac{1}{n_2}}}(s - w_1\mu_1 - \mu_2) \\
&= \frac{w_1 w_2 - r}{w_1^2 + r}s + \frac{w_1 + w_2}{w_1^2 + r}(r\mu_1 - w_1\mu_2), \\
\sigma^2_{T|Y=s} &= (1 - \rho^2)\sigma_T^2 = (1 - \frac{(w_1 w_2 - r)^2}{(w_1^2 + r)(w_2^2 + r)})(\frac{w_2^2}{n_1} + \frac{1}{n_2})\sigma \\
&= \frac{(w_1 + w_2)^2}{n_2(w_1^2 + r)}\sigma.
\end{aligned}$$

where $\rho = \frac{Cov(Y,T)}{\sigma_Y \sigma_T} = \frac{\frac{w_1 w_2}{n_1} - \frac{1}{n_2}}{\sqrt{\frac{w_1^2}{n_1} + \frac{1}{n_2}}\sqrt{\frac{w_2^2}{n_1} + \frac{1}{n_2}}} = \frac{w_1 w_2 - r}{\sqrt{(w_1^2 + r)(w_2^2 + r)}}$. If $w = r$ is fixed, then $\mu_{T|Y=s}$ is a linear function of $\mu_1 - \mu_2$. Then the selection estimator Y and the CI estimator T can be updated as,

$$\begin{aligned} Y &= r\bar{X}_1 + \bar{X}_2, \\ T &= w\bar{X}_1 - \bar{X}_2. \end{aligned} \tag{A.3}$$

Then when applying a fixed selection rule $Y = s$, a simple one-sided $(1-\alpha)$ conditional C.I. for $(\mu_1 - \mu_2)$ can be defined as following.

Result A.1. *A simple one-sided* $(1-\alpha)$ *conditional C.I. for* $(\mu_1 - \mu_2)$ *is*

$$\Big(-\infty, f_1(\bar{X}_1, \bar{X}_2) + \sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} Z_{1-\alpha} \Big), \tag{A.4}$$

where $f_1(\bar{X}_1, \bar{X}_2) = \dfrac{1+r}{w+r} T - \dfrac{w-1}{w+r} s.$

Details for Case 2: Known variance $\sigma_1^2 \neq \sigma_2^2$. We define a selection estimator S and a CI estimator T as

$$\begin{aligned} S &= r_1\bar{X}_1 + r_2\bar{X}_2, \\ T &= w\bar{X}_1 - r_2\bar{X}_2. \end{aligned} \tag{A.5}$$

where $r_1 = \frac{n_1}{n_2}$, $r_2 = \frac{\sigma_1}{\sigma_2}$ and $\sigma_1^2 = r_2^2\sigma_2^2 = \sigma^2$. (S, T) follows a bivariate normal distribution,

$$\begin{pmatrix} S \\ T \end{pmatrix} \sim \mathcal{N}\left[\begin{pmatrix} r_1\mu_1 + r_2\mu_2 \\ w\mu_1 - r_2\mu_2 \end{pmatrix}, \begin{pmatrix} \frac{r_1^2}{n_1} + \frac{1}{n_2} & \frac{r_1 w}{n_1} - \frac{1}{n_2} \\ \frac{r_1 w}{n_1} - \frac{1}{n_2} & \frac{w^2}{n_1} + \frac{1}{n_2} \end{pmatrix} \sigma^2 \right].$$

The conditional distribution of $T|S=s$ follows a normal distribution $\mathcal{N}(\mu_{T|S=s}, \sigma_{T|S=s})$.

$$\mu_{T|S=s} = \frac{w-1}{r_1+1}s + \frac{r_1+w}{r_1+1}(\mu_1-\mu_2),$$
$$\sigma^2_{T|S=s} = \frac{(r_1+w)^2}{n_2(r_1^2+r_1)}\sigma.$$

Then when applying a fixed selection rule $S=s$, a simple one-sided $(1-\alpha)$ conditional C.I. for $(\mu_1-\mu_2)$ can be defined as following.

Result A.2. A simple one-sided $(1-\alpha)$ conditional C.I. for $(\mu_1-\mu_2)$ is,

$$\Big(-\infty, f_2(\bar{X}_1, \bar{X}_2) + \sigma\sqrt{\frac{1}{n_1}+\frac{1}{r_2 n_2}}Z_{1-\alpha}\Big), \tag{A.6}$$

where $f_2(\bar{X}_1, \bar{X}_2) = \dfrac{r_1+1}{w+r_1}T - \dfrac{w-1}{w+r_1}s.$

Details for Case 3: Unknown variance $\sigma_1^2 = \sigma_2^2 = \sigma^2$ We define a selection estimator S and a CI estimator T as

$$\begin{aligned} S &= r\bar{X}_1 + \bar{X}_2, \\ T &= w\bar{X}_1 - \bar{X}_2. \end{aligned} \tag{A.7}$$

Hence, (S,T) follows a bivariate t distribution,

$$\begin{pmatrix} S \\ T \end{pmatrix} \sim \mathbf{t}\left[\begin{pmatrix} r_1\mu_1 + r_2\mu_2 \\ w\mu_1 - r_2\mu_2 \end{pmatrix}, \begin{pmatrix} \dfrac{r_1^2}{n_1}+\dfrac{1}{n_2} & \dfrac{r_1 w}{n_1}-\dfrac{1}{n_2} \\ \dfrac{r_1 w}{n_1}-\dfrac{1}{n_2} & \dfrac{w^2}{n_1}+\dfrac{1}{n_2} \end{pmatrix}\sigma^{\mathbf{2}}\right].$$

The conditional distribution of $T|S=s$ follows a t distribution $\mathbf{t}(\mu_{\mathbf{T|S=s}}, \sigma_{\mathbf{T|S=s}})$.

$$\mu_{T|S=s} = \frac{w-1}{r_1+1}s + \frac{r_1+w}{r_1+1}(\mu_1-\mu_2),$$
$$\sigma^2_{T|S=s} = \frac{(r_1+w)^2}{n_2(r_1^2+r_1)}\sigma.$$

Then when applying a fixed selection rule $S=s$, a simple one-sided $(1-\alpha)$ conditional C.I. for $(\mu_1-\mu_2)$ can be defined as following.

Result A.3. Then a simple one-sided $(1-\alpha)$ cCI for $(\mu_1 - \mu_2)$ is

$$\Big(-\infty, f_3(\bar{X}_1, \bar{X}_2) + s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} t_{(1-\alpha, n_1+n_2-2)} \Big), \tag{A.8}$$

where $f_3(\bar{X}_1, \bar{X}_2) = \dfrac{1+r}{w+r} T - \dfrac{w-1}{w+r} s$, $s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$, $s_1^2 = \frac{\sum_{i=1}^{n_1} (X_{1i} - \bar{X}_1)^2}{n_1-1}$, $s_2^2 = \frac{\sum_{i=1}^{n_2} (X_{2i} - \bar{X}_2)^2}{n_2-1}$.

Since the value of w shows the dependence between selection estimator S and CI estimator T. Now we focus on the w and compare the unconditional CI with conditional CI.

Case 1: $w = 1$**.** All the cCI is same as unconditional CI since CI estimator T is independent to selection estimator Y. There is no need to discuss this part. But what we concern more is the situation $w \neq 1$.

Case 2: $w \neq 1$**.** Now, we use first case as an example. We want to compare Result A.1 with unconditional CI. Without loss of any generality, we assume $r = 1$. Now the selection estimator of cCI and unconditional CI are the same as

$$Y = \bar{X}_1 + \bar{X}_2.$$

But the CI estimator are different. Let $T_1 = \bar{X}_1 - \bar{X}_2$ be the unconditional CI estimator and $T_2 = w\bar{X}_1 - \bar{X}_2$ be the cCI estimator. Given the fixed selection rule $Y = s$, we can show that these two estimators can construct same CI eventually. That is, the upper bound of one-side CI can be showed as

$$\frac{2}{w+1}(w\bar{X}_1 - \bar{X}_2) - \frac{w-1}{w+1}(\bar{X}_1 + \bar{X}_2) + \textit{error term} = \bar{X}_1 - \bar{X}_2 + \textit{error term}.$$

To sum up, no matter whether w is equal to 1 or not the cCI is same as the unconditional CI. In this sense, though we may assume selection estimator Y and CI estimator T have some dependence, we can still easily construct cCI. And this cCI

can deal with the effect of selection. The selection rule we suggest is $Y = s$. And this simple selection rule is not commonly used in real research. Since it is too strict to make any discovery. In the future, we want to extend our results to selection rule $Y \geq s$, which is much more acceptable as a selection rule.

Proof of the Covariance Matrix is Positive Semi-definite

Proof. By the definition of positive semi-definite, a $n \times n$ Hermitian complex matrix $\widetilde{\Sigma}$ is said to be positive semi-definite if $x^T \widetilde{\Sigma} x \geq 0$ for all non-zero x in $\mathbb{R}^n$. Let A be a arbitrary nonzero vector $A = (a_1, a_2, ..., a_{2n-1}, a_{2n})^T$, where $a_i \neq 0$, $i = 1, 2, ..., 2n$ and Hermitian complex matrix $\widetilde{\Sigma}$ is $2n$ by $2n$ matrix as defined in Equation 4.2 (m = 2n). Given a fixed $\rho \in [0, 1]$ for any nonzero vector A,

$$A^T \widetilde{\Sigma} A = \begin{pmatrix} a_1 & a_2 & \cdots & a_{2n-1} & a_{2n} \end{pmatrix} \begin{pmatrix} 1 & -\rho & \rho & -\rho & \cdots & \rho & -\rho \\ -\rho & 1 & -\rho & \rho & \cdots & -\rho & \rho \\ \rho & -\rho & 1 & -\rho & \cdots & \rho & -\rho \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ -\rho & \rho & -\rho & \rho & \cdots & -\rho & 1 \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_{2n-1} \\ a_{2n} \end{pmatrix}$$

$$= (1-\rho) \sum_{i=1}^{2n} a_i^2 + \rho (\sum_{j=1}^{n} a_{2j-1})^2 - 2\rho (\sum_{j=1}^{n} a_{2j-1})(\sum_{k=1}^{n} a_{2k}) + \rho (\sum_{k=1}^{n} a_{2k})^2$$

$$= (1-\rho) \sum_{i=1}^{2n} a_i^2 + \rho (\sum_{j=1}^{n} a_{2j-1} - \sum_{k=1}^{n} a_{2k})^2 \geq 0.$$

Hence the covariance matrix in Equation 4.2 is positive semi-definite. □

REFERENCES

[1] Y. Benjamini. Simultaneous and selective inference: current successes and future challenges. *Biometrical Journal*, 52: 708-721, 2010.

[2] Y. Benjamini and M. Bogomolov. Selective inference on multiple families of hypotheses. *Journal of the Royal Statistical Society: Series B*, 76: 297-318, 2014.

[3] Y. Benjamini, Y. Hechtlinger and P. Stark. Confidence intervals for selected parameters. *arXiv:1906.00505*, 2019.

[4] Y. Benjamini, R. Heller and D. Yekutieli. Selective inference in complex research. *Philosophical Transactions of The Royal Society A Mathematical Physical and Engineering Sciences*, 367: 4255–4271, 2009.

[5] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B*, 57: 289-300, 1995.

[6] Y. Benjamini and Y. Hochberg. On the adaptive control of the false discovery rate in multiple testing with independent statistics. *Journal of Educational and Behavioral Statistics*, 25: 60-83, 2000.

[7] Y. Benjamini, Y. Hochberg and P. Stark. Confidence intervals with more power to determine the sign: two ends constrain the means. *Journal of the American Statistical Association*, 93: 309-331, 1998.

[8] Y. Benjamini, A. Krieger and D. Yekutieli. Adaptive linear step-up procedures that control the false discovery rate. *Biometrika*, 93: 491-507, 2006.

[9] Y. Benjamini and D. Yekutieli. The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 29: 1165-1188, 2001.

[10] Y. Benjamini and D. Yekutieli. False discovery rate: adjusted multiple confidence intervals for selected parameters. *Journal of the American Statistical Association*, 100: 71-81, 2005.

[11] R. Berk, L. Brown, A. Buja, K. Zhang and L. Zhao. Valid post-selection inference. *The Annals of Statistics*, 41: 802-837, 2013.

[12] P. Buhlmann and S. Geer. *Statistics for high-dimensional data. Methods, theory and applications*, 1st edition, Springer, Berlin Heidelberg, 2011.

[13] S. Dhanasekaran, T. Barrette, D. Ghosh, R. Shah, S. Varambally, K. Kurachi, K. Pienta, M. Rubin and A. Chinnaiyan. Delineation of prognostic biomarkers in prostate cancer. *Nature*, 412: 822-826, 2001.

[14] B. Efron. Microarrays, empirical bayes and the two-groups model (with discussion). *Statistical Science*, 23: 1-22, 2008.

[15] B. Efron. Tweedie's formula and selection bias. *Journal of the American Statistical Association*, 105: 1602-1614, 2011.

[16] W. Fithian, J. Elith, T. Hastie and D. Keith. Bias correction in species distribution models: pooling survey and collection data for multiple species. *Methods in Ecology and Evolution*, 6: 424–438, 2015.

[17] W. Fithian, D. Sun and J. Taylor. Optimal inference after model selection. *arXiv:1410.2597v4*, 2017.

[18] D. Gallerano, P. Ndlovu, I. Makupe, M. Focke-Tejkl, K. Fauland, E. Wollmann, E. Puchhammer-Stöckl, W. Keller, E. Sibanda and R. Valenta. Comparison of the specificities of IgG, IgG-subclass, IgA and IgM reactivities in African and European HIV-infected individuals with an HIV-1 clade C proteome-based array. *PLoS One*, 10: e0117204, 2015.

[19] M. Giri, M. Nebozhyn, L. Showe and L. Montaner. Microarray data on gene modulation by HIV-1 in immune cells: 2000-2006. *Journal of Leukocyte Biology*, 80: 1031-1043, 2006.

[20] W. Guo and J. Romano. A generalized Sidak-Holm procedure and control of generalized error rates under independence. *Statistical Applications in Genetics and Molecular Biology*, 6: 1-35, 2007.

[21] W. Guo and J. Romano. Analysis of error control in large scale two-stage multiple hypothesis testing. *arXiv:1703.06336*, 2017.

[22] Y. Hong. On computing the distribution function for the poisson binomial distribution. *Computational Statistics and Data Analysis*, 59: 41-51, 2013.

[23] J. Ioannidis. Why most published research findings are false. *PLoS Medicine*, 2: 696-701, 2005.

[24] S. Jones, J. Lane and M. Weedon. Genome-wide association analyses of chronotype in 697,828 individuals provides insights into circadian rhythms. *Nature Communications*, 10: 343-353, 2019.

[25] E. Katsevich and A. Ramdas. Towards "simultaneous selective inference": post-hoc bounds on the false discovery proportion. *arXiv:1803.06790v3*, 2018.

[26] D. Kivaranovic and H. Leeb. On the length of post-model-selection confidence intervals conditional on polyhedral constraints. *arXiv:1803.01665v2*, 2019.

[27] A. Klenke and L. Mattner. Stochastic ordering of classical discrete distributions. *arXiv:0903.1361*, 2010.

[28] J. Lee, D. Sun, Y. Sun and E. Jonathan. Exact post-selection inference, with application to the lasso. *Annals of Statistics*, 44: 907-927, 2013.

[29] J. Lee and J. Taylor Exact post model selection inference for marginal screening. *Advances in Neural Information Processing Systems 27 (NIPS)*, 2014.

[30] J. Lee, R. Wedow and D. Cesarini. Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. *Nature Genetics*, 50: 1112–1121, 2018.

[31] E. Lehmann and J. Romano. *Testing statistical hypotheses*, 3rd edition, Springer, New York, 2005.

[32] E. Lehmann and H. Scheff'e. Completeness, similar regions, and unbiased estimation: Part ii. *Sankhya: The Indian Journal of Statistics*, 15: 219–236, 1955.

[33] J. Markovic, L. Xia and J. Taylor. Unifying approach to selective inference with applications to cross-validation. *arXiv:1703.06559v3*, 2018.

[34] J. Peng, C. Lee, K. Davis and W. Wang. Stepwise confidence intervals for monotone dose-response studies. *Biometrics*, 64: 877-85, 2008.

[35] J. Peng, W. Liu, F. Bretz and Z. Shkedy. Multiple confidence intervals for selected parameters adjusted for the false coverage rate in monotone dose-response microarray experiments. *Biometrical Journal*, 59: 732-745, 2017.

[36] J. Pratt. Length of confidence intervals. *Journal of the American Statistical Association*, 56: 549-567, 1961.

[37] J. Ranstam. Why the p-value culture is bad and confidence intervals a better alternative. *Osteoarthritis and Cartilage*, 20: 805-808, 2012.

[38] J. Rossouw, G. Anderson, R. Prentice, A. Lacroix, C. Kooperberg, M. Stefanick, R. Jackson, S. Beresford, B. Howard, K. Johnson, J. Kotchen and J. Ockene. Risks and benefits of estrogen plus progestin in healthy postmenopausal women: principal results from the women's health initiative randomized controlled trial. *The Journal of the American Medical Association*, 288: 321-333, 2002.

[39] S. Sarkar. On methods controlling the false discovery rate. *Sankhya: The Indian Journal of Statistics, Series A*, 70: 135-168, 2008.

[40] B. Soric. Statistical 'discoveries' and effect-size estimation. *Journal of the American Statistical Association*, 84: 608-610, 1989.

[41] L. Sun and S. Bull. Reduction of selection bias in genomewide studies by resampling. *Genetic Epidemiology*, 28: 352–67, 2005.

[42] J. Taylor and R. Tibshirani. Statistical learning and selective inference. *Proceedings of the National Academy of Sciences*, 112: 7629-7634, 2015.

[43] L. Tian, H. Fu, S. Ruberg, H. Uno, and L. Wei. Efficiency of two sample tests via the restricted mean survival time for analyzing event time observations. *Biometrics*, 74: 694–702, 2018.

[44] X. Tian and J. Taylor. Selective inference with a randomized response. *The Annals of Statistics*, 46: 679-710, 2018.

[45] A. Wout, G. Lehrman, S. Mikheeva, G. Keeffe, M. Katze, R. Bumgarner, G. Geiss and J. Mullins. Cellular gene expression upon human immunodeficiency virus type 1 infection of CD4(+) T-cell lines. *Journal of Virology*, 77: 1392-1402, 2003.

[46] C. Wang, M. Chen, E. Schifano, J. Wu and J. Yan. Statistical methods and computing for big data. *arXiv:1502.07989*, 2015.

[47] R. Wasserstein and N. Lazar. The ASA's statement on p-values: context, process, and purpose. *The American Statistician*, 70: 129-133, 2016.

[48] L. Wasserman and K. Roeder. High-dimensional variable selection. *The Annals of Statistics*, 37: 2178-2201, 2009.

[49] A. Weinstein, W. Fithian and Y. Benjamini. Selection adjusted confidence intervals with more power to determine the sign. *Journal of the American Statistical Association*, 108: 165-176, 2013.

[50] A. Weinstein and A. Ramdas. Online control of the false coverage rate and false sign rate. *arXiv:1905.01059*, 2019.

[51] A. Weinstein and D. Yekutieli. Selective sign-determining multiple confidence intervals with FCR control. *arXiv:1404.7403v2*, 2014.

[52] S. Woody and J. Scott. Optimal post-selection inference for sparse signals: a nonparametric empirical-Bayes approach. *arXiv:1810.11042v1*, 2018.

[53] D. Yekutieli. Adjusted bayesian inference for selected parameters. *Journal of the Royal Statistical Society: Series B*, 74: 515-541, 2012.

[54] H. Zhong and R. Prentice. Bias-reduced estimators and confidence intervals for odds ratios in genome-wide association studies. *Biostatistics*, 9: 621–34, 2008.