國立高雄大學統計學研究所

碩士論文

Forward selection two sample binomial test

檢定兩獨立母體比例前進選擇法

研究生：林妙珊 撰

指導教授：黃錦輝 教授

中華民國九十九年七月

致謝辭

本論文的完成，首先感謝我的指導教授 黃錦輝教授，這段時間由於老師在學業上的悉心指導與鼓勵，論文才得以順利完成，其次要感謝 黃文璋教授、 陳瑞彬教授、 俞淑惠教授，和中山大學應用數學系 郭美惠教授，感謝您們在課業上的認真教學與指導，使我對統計學有更深入的認識。還有感謝口試委員 吳雅琪博士及 杜宜軒教授，感謝您們在論文上給予的指導與建議。

另外，感謝蘭屏姐的照顧，謝謝您這兩年來細心且熱心的幫助我們處理大大小小的事情。感謝虹儒學姐和廣杰學長在課業上給予的幫助。還要感謝這兩年一起努力打拼的統計所的同學們，謝謝雨潔、盈慧、周家、慧怡、建中、竣元、建銓和晉煜陪我一起打沒有規則的籃球，尤其感謝和我相同指導教授的珊珊和建銓，謝謝你們這一路上的支持和陪伴。

更要感謝我的父母親從小到大對我的支持和包容，讓我無後顧之憂的專心於學業上，感謝爺爺和奶奶的關心和照顧，也感謝姐姐、弟弟和已逝的愛犬Co Co，謝謝你們一路上的鼓勵、支持和陪伴，你們是我永遠最愛的家人。

最後，在此將這份成果和喜悅與我摯愛的親人、敬愛的師長以及曾經給予我幫助的朋友、學長姐和同學們一起分享。

Forward selection two sample binomial test

by

Miao-Shan Lin

Advisor

Kam-Fai Wong

Institute of Statistics

National University of Kaohsiung

Kaohsiung, Taiwan 811, R.O.C.

July 2010

Contents

List of Figures

List of Tables

檢定兩獨立母體比例前進選擇法

指導教授：黃錦輝 博士
國立高雄大學統計學研究所

學生：林妙珊
國立高雄大學統計學研究所

摘要

來自兩獨立二項分佈的數據，在實務上是廣泛存在的，例如臨床試驗。傳統上，欲比較兩獨立二項分佈母體比例，當遇到小樣本時，一般建議使用費雪精確檢定法。2008年 Gerald G. Crans 和 Jonathan J. Shuster 證實如果將費雪精確檢定法運用在檢定兩獨立二項分佈母體比例是否相同的問題上，則會造成型 I 錯誤的機率遠低於所設定的顯著水準，即使兩組樣本的樣本數都達到 125，依舊遠低於所給定的顯著水準。更進一步的，他們針對費雪精確檢定法提出一個修正方法，他們定義出新的顯著水準，$\alpha^* = \alpha + \varepsilon$，$\alpha$ 代表欲設定之顯著水準，ε 代表一特定的正數。為了方便使用，他們提供在不同樣本數時的新顯著水準和預期之顯著水準的對照表，此修正方法除了提升型 I 錯誤機率問題外，也有效的提昇檢定力；事實上從另一個觀點來看，此修正方法雖然使用兩獨立二項分佈計算真實型 I 錯誤的發生機率，但是實際上可能結果進入拒絕域的順序依然採用超幾何分布做決定，因此，相對於修正費雪精確檢定的方法，在本篇論文中我們給出另一方法來決定可能結果進入拒絕域的順序，在此不使用超幾何分布決定順序，取而代之的是使用兩獨立二項分佈來決定進入拒絕域的順序。最後，我們會將所提出的新方法和一些已經被提出的方法作比較，藉由有限樣本下的數值結果說明這個新的方法的特性和優點，且同時經由檢定力函數圖形看各個方法的曲線變化。

關鍵字：2×2 列聯表、費雪精確檢定法、二元數據、臨床試驗。

Forward selection two sample binomial test

Advisor: Dr. Kam-Fai Wong
Institute of Statistics
National University of Kaohsiung

Student: Miao-Shan Lin
Institute of Statistics
National University of Kaohsiung

Abstract

The data which come from two-sample comparative binomial trial is one of the most widely applied statistical data structure in practice such as in clinical trial. In the comparison of two independent binomial proportions for small sample sizes, one of the commonly used technique is Fisher's exact test. Fisher's exact test is a conditional method which considers the hypergeometric distribution as null distribution. In year 2008, Gerald G. Crans and Jonathan J. Shuster verify that applying Fisher's exact test to this type of comparison will give a lower probability of type I error than that we expected even for sample sizes up to 125 subjects per group. Additionally, they propose an adjusted method that defines new significance levels $\alpha^* = \alpha + \varepsilon$, where α is pre-specified and ε is a small positive number, and gives a cross-reference table which contains all new significance levels links for various sample sizes and target α. The adjustment uniformly improves the test size, rasing the actual probability of type I error and hence more powerful. From another point of view, their proposed method applies two independent binomial distribution to calculate the actual probability of type I error, where the sequence of the possible outcomes for the rejection region is based on hypergeometric distribution. In this paper, we give another sequence to involve the possible outcomes into the rejection region. Instead of

using hypergeometric distribution, we are strictly forward to consider two independent binomial distribution to as the reference. Lastly, we through comparing with some proposed methods, then, the properties and advantages of this method are demonstrated by numerical results. Some figures are presented to show the pattern changes of the power functions for these methods.

Keywords: 2×2 contingency table, Fisher's exact test, binary data, clinical trial.

1 Introduction

The data generated from the comparison of two independent binomial distribution is one of the most widely applied statistical data structure in practice. Then the data can be organized into a 2×2 contingency table, where 2×2 contingency tables are the most elemental data structure leading to ideas of association. Fisher's exact test is a popular test used in the analysis of 2×2 contingency table for small sample sizes. It is used to examine the significance of the association between the two kinds of classification. The most important thing is that Fisher's exact test is a conditional test which assumes that the row and column totals of 2×2 contingency table need to be fixed in advance. That is because Fisher (1935) consider that the marginal totals are "ancillary statistics", and therefore provide no information respecting the configuration of the body of the table. However, in practice, this assumption is not met in many experimental designs and almost all non-experimental ones. Berkson (1978a) advocates that marginal totals do contain information and also indicates that fixed marginal assumption of Fisher's exact test is shown to be incorrect to comparative binomial trial. In fact, Barnard (1947a) is the first to propose that there are at least three different sampling leading to a 2×2 contingency table. Kempthorne (1979) suggests that differentiation among these 2×2 contingency origins is crucial before testing. Kempthorne think that different data structure has its appropriate testing methods. Not all of the 2×2 contingency table data are not applicable for analyzing by Fisher's exact test.

The debate as to which statistical methodology is most appropriate for the analysis of two sample comparative binomial trial has persisted for decades. The main hypothesis testing difficulty in the context of the comparative binomial trial is that it involves the unknown parameter p, the common value of binomial proportion under the null hypothesis. Among several existing methods, Fisher's exact test is most popular and widely used for comparative binomial trial as small sample sizes. Fisher's exact test uses hypergeometric distribution and it does not have any connection with parameter p. However, Boschloo (1970) proposes that the actual probability of type I error is quite lower than

the pre-specified α, and it has been confirmed by several other authors such as McDonald et al (1977). In year 2008, Gerald G. Crans and Jonathan J. Shuster again verify that applying Fisher's exact test to the analysis of comparative binomial trial will give a lower probability of type I error than that we expected even for sample sizes up to 125 subjects per group. Furthermore, they develop a numerical algorithm, which sequence is on the basis of hypergeometric distribution and inflates the rejection region by two independent binomial distribution. They not only propose an adjustment to improve the lower actual probability of type I error for Fisher's exact test but also more powerful.

However, " exact test " idea of Fisher proposing in year 1935, which has been widely considered to be the exemplar of statistical tests for comparative binomial trial. For instance, Barnard's (1947a) the exact unconditional CSM test and Suissa and Shuster's (1985) exact unconditional test are effected by Fisher's " exact test " idea. Therefore, we propose a forward selection binomial test procedure which is similar to the modified Fisher's exact test, but using different way to include possible outcomes into the rejection region. Instead of using hypergeometric distribution, we are strictly forward to consider two independent binomial distribution to as the reference for searching the possible outcomes into the rejection region. Lastly, we compare the curves of exact probability of type I error and power function curve with other existing methods to realize the advantage of forward selection binomial test.

2 Tests for two independent binomial samples

We start by introduce a classical methods, two sample binomial test (BT) and then in Subsection 3.2, we introduce a modified two sample binomial test (MBT) which is proposed by Suissa and Shuster in year 1985. Next in Subsection 3.3, we introduce another classical methods, Fisher's exact test (FET), which is used to test the equality of the proportion of two independent samples. Followed by a modified Fisher's exact test (MFET) which is proposed by Gerald G. Grans and Jonathan J. Shuster in year 2008. In

the end of this section, we propose a forward selection binomial test (FSBT).

Suppose X and Y are the number of successes of the treatment group and control group, respectively, which are two independent random variables having binomial distributions with parameters (n_1, p_1) and (n_2, p_2). Thus, the joint distribution of X and Y can be written as

$$f_{X,Y}(x, y \mid p_1, p_2) = \binom{n_1}{x} p_1^x (1-p_1)^{n_1-x} \binom{n_2}{y} p_2^y (1-p_2)^{n_2-y}. \tag{1}$$

Without loss of generality, we focus on one-sided test with the hypothesis statement

$$H_0 : p_1 = p_2 \quad vs \quad H_1 : p_1 < p_2.$$

2.1 Two sample binomial test (BT)

A trivial test statistic for comparing the difference of the means of two populations is the difference of the two sample means, where the test statistic is written as

$$\theta_{BT}(x, y) = \frac{X}{n_1} - \frac{Y}{n_2}.$$

Therefore, the p-value of any observed value (x^*, y^*) can be calculated through equation (1)

$$K_{BT}(x^*, y^*) = \max_{0 \le p_1 = p_2 \le 1} \sum_{(x,y) \in C_{BT}(x^*, y^*)} f_{X,Y}(x, y \mid p_1, p_2),$$

where the ordered set

$$C_{BT}(x^*, y^*) = \{(x, y) \mid \theta_{BT}(x, y) \le \theta_{BT}(x^*, y^*)\}.$$

In addition, the rejection region with nominal level α becomes

$$R_{BT,n_1,n_2,\alpha} = \{(x, y) \mid K_{BT}(x, y) \le \alpha\}.$$

Although BT considers the trivial statistic for comparing two independent binomial proportions, however, the test statistic of BT ignores the information of variability given by the observation (x, y). For example, as sample sizes $n_1 = n_2 = 5$, the possible outcomes

$(0,3)$, $(1,4)$ and $(2,5)$ are regarded as same significance value for BT. But the fact that those possible outcomes have different probability to happen. This such that the power of BT is less than that of FET, although BT applies exact distribution but FET applies conditional distribution. We will give a more detail discussion later.

2.2 Modified two sample binomial test (MBT)

Generally, in contrast to FET, practitioners prefer to use normal approximation test as large sample sizes. However, since Fisher (1935) proposes the " exact test " idea, the use of the " exact test " with comparative binomial trial has been widely consider to be the exemplar of statistical tests of significance. Furthermore, in contrast to the conditional tests, the unconditional tests are easily explained and understood by non-statisticians. Therefore, in year 1985, Suissa and Shuster propose an adjustment for normal approximation test. They use the test statistic given by normal approximation test (Z-test), but the calculation of p-value is based on two independent binomial distribution rather than standard normal distribution which comes from asymptotic theory.

In other words, they use the Z-test statistic as the sequence of including possible outcomes into rejection region. Instead of using asymptotic theory, they apply two independent binomial distribution to calculate the actual probability of type I error. The test statistic is written as

$$\theta_{MBT}(x,y) = \frac{(\hat{p}_1 - \hat{p}_2)}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}},$$

where $\hat{p}_1 = \frac{X}{n_1}$ and $\hat{p}_2 = \frac{Y}{n_2}$, respectively. Therefore, for any observed value (x^*, y^*), the corresponding p-value is also calculated through equation (1)

$$K_{MBT}(x^*, y^*) = \max_{0 \le p_1 = p_2 \le 1} \sum_{(x,y) \in C_{MBT}} f_{X,Y}(x, y \mid p_1, p_2),$$

where the ordered set is

$$C_{MBT}(x^*, y^*) = \{(x,y) \mid \theta_{MBT}(x,y) \le \theta_{MBT}(x^*, y^*)\}.$$

The rejection region with nominal level α becomes

$$R_{MBT,n_1,n_2,\alpha} = \{(x,y) \mid K_{MBT}(x,y) \leq \alpha\}.$$

As a matter of fact, comparing BT with Z-test, we can regard it as an modified test to BT (MBT). They are using two independent binomial distribution to calculate the corresponding p-value of possible outcomes. However, the test statistic of MBT not only consider the variation of $\hat{p}_1 - \hat{p}_2$ but also considers the information of variability given by the observation (x, y), which is ignored by BT. For instance, as sample sizes $n_1 = n_2 = 5$, the p-values of the possible outcomes $(0, 3)$, $(1, 4)$ and $(2, 5)$ are equal to 0.054688 for BT. For MBT, the $p-$values of the possible outcomes $(0, 3)$, $(2, 5)$ are equal to 0.030924, but the p-value of the possible outcome $(1, 4)$ is 0.054689. Here, when we use 0.05 as nominal significant level, for BT, the possible outcomes $(0, 3)$, $(1, 4)$ and $(2, 5)$ are not considered as significant. In contrast to BT, the possible outcomes $(0, 3)$, $(2, 5)$ are considered as significant for MBT. Therefore, MBT can effectively raise the power and not influenced by sample sizes. In the next Section, we will have detailed discussions.

Note that the following cases for each estimator of standard error are equal to zero need to be given additional discussion in advance. Since the alternative hypothesis is $H_1 : p_1 < p_2$, therefore, if we get the possible outcome $(0, n_2)$, we will intuitively think that p_1 must be smaller than p_2. Hence, the possible outcome $(0, n_2)$ is ranked in the first place. In contrast, the possible outcome $(n_1, 0)$ will be intuitively thought that p_1 must be larger than p_2. Moreover, the cases $(0, 0)$ and (n_1, n_2) are regarded as $p_1 = p_2$. Therefore, we directly rank those possible outcomes $(n_1, 0)$, $(0, 0)$ and (n_1, n_2) in last place.

2.3 Fisher's exact test (FET)

R. A. Fisher (1935) in the Design of Experiments describes the following experiment. Muriel Bristol, a colleague of Fisher's at Rothamsted Experiment Station, claims that she is able to distinguish whether the milk or the tea infusion is added to the cup first. To test her claim, Fisher designs an experiment in which Fisher asks her to taste eight cups

of tea. In mixing eight cups of tea, four cups of tea have milk added first, and the other four have tea added first, where the order of presenting the cups to her is randomized. Then she should try to select the four that have milk fist. Before the experiment started, she has been told that there are four cups which have the milk added first. Therefore, the experimental design has been fixed both marginal totals before trial. Whether Ms Bristol really can distinguish whether milk or tea is poured in first. For solving the small sample sizes problem, Fisher proposes an exact test which uses the hypergeometric distribution as null distribution.

Therefore, Fisher's exact test (FET) is proposed and named after its inventor, R. A. Fisher. FET is widely used in the analysis of 2×2 contingency table for small sample sizes. It is used to test the significance of the relationship between two kinds of classification. In fact, FET is an exact conditional test, because it assumes that the marginal totals should be fixed in advance. That is because Fisher regards that if these marginal frequencies which can be admitted without supplying any information, we may recognize the information which only come from guess result. That is, we regard the information of marginal totals supplying as wholly ancillary. Then conditionality of given the marginal totals can provide a simple way to eliminate nuisance parameters. Furthermore, the assumption of FET can get the exact distribution for the test statistic, which is hypergeometric distribution, rather than relying on an approximation.

In the cases that X and Y are independent random variables having binomial distributions. The conditional distribution of X given $X+Y$ has hypergeometric distribution under null hypothesis, which does not have any connection with parameter p, the common value of binomial proportion under the null hypothesis. Specifically,

$$f_{X|X+Y}(x \mid x+y) = \frac{\binom{n_1}{x}\binom{n_2}{y}}{\binom{N}{x+y}} \ , \ max\{0, x+y-N+n_1\} \le x \le min\{x+y, n_1\}.$$

For any observed value (x^*, y^*), the corresponding p-value can be calculated by

$$K_{FET}(x^*, y^*) = \sum_{(x,y)\in C_{FET}(x^*,y^*)} f_{X|X+Y}(x \mid x+y),$$

where the ordered set based on hypergeometric distribution

$$C_{FET}(x^*, y^*) = \{(x, y) \mid x \leq x^*, x + y = x^* + y^*\}.$$

This implies that the rejection region that generates by FET with nominal level α is

$$R_{FET,n_1,n_2,\alpha} = \{(x, y) \mid K_{FET}(x, y) \leq \alpha\}.$$

Although, FET is widely applied to test the significance of the association between the two kinds of classification for small sample sizes. Kempthorne (1979) think it is a pity, that Fisher does not take account of the data which come from different type of structures. In fact, categorical data that result from classifying objects in two different ways which can be broken down into three cases. In addition to marginal totals are fixed, which also contains that only one side is fixed and all not fixed. As early as 1947, Barnard has been thought that the data structure for the categorical data that result from classifying objects in two different ways need to be distinguished. They think that not all of the data structures are appropriate and use FET to examine the significance of the association between the two kinds of classification. Finally, they especially point out that FET is not suitable for testing as the data coming from two-sample comparative binomial trial. In practical, the assumption of fixing marginal totals can pose interpretation difficulties. Therefore, they suggested that a good analysis must be done in accordance with data structure to find the appropriate method. In Section 3, we will have more detailed discussion about using FET in the analysis of the two-sample comparative binomial trial.

2.4 Modified Fisher's exact test (MFET)

Recently, FET is still commonly used and popular in comparative binomial trials even the sample size is large. However, FET assumes that the total number of successes should be fixed in advance, that can cause interpretation difficulties in application for data coming from comparative binomial trial. That is because the result of FET are derived using conditional sample space rather than the set of all possible outcomes. Furthermore,

more importantly, using Fisher's exact test for the analysis of the two-sample comparative binomial trial, which can pose that the actual probability of type I error is serious less than pre-specified α. For the above-mentioned problems of FET, Crans and Shuster (2008) propose an adjustment to FET, that increases the power by adding possible outcomes to the rejection region while maintaining the test size which we are prespecified. Therefore, the modified FET (MFET) is to define a new significance levels $\alpha^* = \alpha + \varepsilon$, where α is pre-specified and ε is a small positive number, and the critical region will be determined by using α^* instead of α. Because in real-world applications such as in clinical trials, given significance levels in advance are the standard, and statistical tests will be evaluated according to expected significance level.

Another point of view on MFET is that, in fact, they use the hypergeometric distribution to define the ordered set $C_{MFET}(x, y)$. That is, the order of (x, y) is given by $K_{FET}(x, y)$, the p-value of observing (x, y) by using FET. Then, the procedure uses the ordered set $C_{MFET}(x, y)$ as the reference for creating rejection region, where the actual probability of type I error is calculated through equation (1) as BT and MBT. Specifically, for any observed value (x^*, y^*), the corresponding p-value is equal to

$$K_{MFET}(x^*, y^*) = \max_{0 \le p_1 = p_2 \le 1} \sum_{(x,y) \in C_{MFET}} f_{X,Y}(x, y \mid p_1, p_2)$$

where the ordered set is

$$C_{MFET}(x^*, y^*) = \{(x, y) \mid K_{FET}(x, y) \le K_{FET}(x^*, y^*)\}.$$

The rejection region with nominal level α is

$$R_{MFET,n_1,n_2,\alpha} = \{(x, y) \mid K_{MFET}(x, y) \le \alpha\}.$$

According to the equation of K_{MFET}, the value of K_{MFET} is computational complexity than that of K_{FET}. A gridding method is applied to get an approximated value. Specifically, the interval $[0, 1]$ is divided into r equal subintervals, where r is a positive integer. Then, define $p'_i = \frac{i}{r}$, the approximated value of K_{MFET} is calculated by

$$\max_i \sum_{(x,y) \in C_{MFET}} \binom{n_1}{x} \binom{n_2}{y} (p'_i)^{x+y} (1 - p'_i)^{n_1+n_2-x-y},$$

where $i = 0, 1, ..., r$. Moreover, a more accurate approximation for the value of K_{MFET} can be done by chosen a large r. In addition, the p-value of both BT and MBT are also calculated in this way.

Through the approximation, Crans and Shuster demonstrate that for equal sample sizes up to 125, the actual probability of type I error is more less than the prespecified nominal level for the nominal two-sided significance level 0.05. Figure 2.1 shows the numerical results by different dashed lines for sample sizes of 10, 25, 50, and 100, and also the simulation results with 10,000 repetitions when $p = 0.1, 0.2, ..., 0.9$. The simulation results agree that the gridding method is accuracy and efficiency for approximating K_{MFET}.

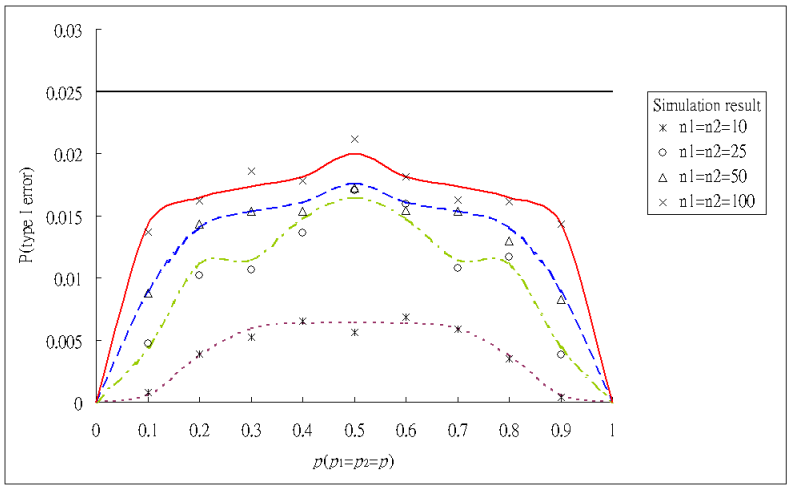


Figure 2.1: The curves of the probabilities of type I error which calculated by gridding method with the sample sizes $n_1 = n_2 = 10$ (dotted line), $n_1 = n_2 = 25$ (dotted-dashed line), $n_1 = n_2 = 50$ (dashed line) and $n_1 = n_2 = 100$ (solid line) where $p_1 = p_2 = p$.

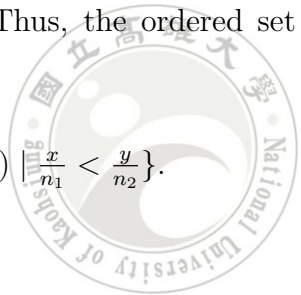
It is noteworthy that as early as 1970, this well known property of FET using in comparative binomial trial has been confirmed by Boschloo. In order to be easily implemented in analysis, Crans and Shuster (2008) give a cross-reference table which contains all new significance levels links for various sample sizes and target significance level. Through the cross-reference table, the decision of whether the test is significant or not is taken by

determining the p-value of FET and comparing it with the new significance level α^* of the table.

2.5 Forward selection two sample binomial test (FSBT)

Rather than MFET method which considers hypergeometric distribution as the order of including the possible outcomes for creating the rejection region, here, we provide another ordered set C_{FSBT} which considers two independent binomial distribution directly to as the reference. In practice, we may not agree to reject H_0 if $\frac{X}{n_1} \geq \frac{Y}{n_2}$. This is true when α is not too large. Thus, we exclude the observed values (x, y) in which $\frac{X}{n_1} \geq \frac{Y}{n_2}$ from the list of candidates.

It is noteworthy that we should have target alternative, p_1^* and p_2^*, when doing sample size calculation. To order the possible outcomes without any tie cases, we need these two values being prespecified. Thus, the ordered set C_{FSBT} is given by the following algorithm.



Step 1. Set $B^{(0)} = \{\}, A^{(0)} = \{(x, y) \mid \frac{x}{n_1} < \frac{y}{n_2}\}$.

Step 2. Set $i = 0$.

Step 3. Calculate $z_i = \min\limits_{(x,y)\in A^{(i)}} \{\max\limits_{0\leq p_1=p_2\leq 1} h(x, y \mid p_1, p_2)\}$,

where $h(x, y \mid p_1, p_2) = f_{X,Y}(x, y \mid p_1, p_2) + \sum\limits_{(u,v)\in B^{(i)}} f_{X,Y}(u, v \mid p_1, p_2)$.

Step 4. $C^{(i)} = \{(x, y) \mid (x, y) \in A^{(i)}, \max\limits_{0\leq p_1=p_2\leq 1} h(x, y \mid p_1, p_2) = z_i\}$.

Step 5. Let $d_i =$ the number of elements belong to $C^{(i)}$.

If $d_i > 1$, set $C^* = \{\arg\max\limits_{(x,y)\in C^{(i)}} \max\limits_{p_1=p_1^*, p_2=p_2^*} f_{X,Y}(x, y \mid p_1, p_2)\}$, else $C^* = C^{(i)}$.

Step 6. $B^{(i+1)} = B^{(i)} \cup C^*,\ A^{(i+1)} = A^{(i)} \setminus C^*$.

Step 7. If $(x^*, y^*) \notin C^*$, increase i by 1 and go to step3. Otherwise, set $C_{FSBT} = B^{(i+1)}$.

Thus, for any observed value (x^*, y^*), the corresponding $p - value$ is equal to

$$K_{FSBT}(x^*, y^*) = \max_{0 \le p_1 = p_2 \le 1} \sum_{(x,y) \in C_{FSBT}} f_{X,Y}(x, y \mid p_1, p_2).$$

The rejection region with nominal level α is

$$R_{FSBT,n_1,n_2,\alpha} = \{(x, y) \mid K_{FSBT}(x, y) \le \alpha\}.$$

In general, it gives target alternative in protocol before studying clinical trials, so we use the condition of target alternative as $d_i > 1$ for some i. If the condition of target alternative does not consider during the course of ordering as d_i is more than 1, in turn, we lay down that all such points are to give the same rank. This mimics to the idea given by Barnard in year 1947, yet the computing power is the limitation in that moment. Therefore, Barnard proposes another procedure called C.S.M. test rather than using the above algorithm to form the p-value. In fact, many researchers support the idea, but they think that the calculation of C.S.M is still limited by computational difficulties such as Upton (1982). Barnard later refutes C.S.M. test in favor of FET in year 1949. Despite his disclaimer and other practitioners in support of FET, exact unconditional inference have gained tremendous popularity over the last few decades. Here, the algorithm of FSBT extends the idea of the Barnard's primitive idea.

2.5.1 Example

The following example provides a detailed and step-by-step explanation for the order process of FSBT using sample sizes $n_1 = 5$ and $n_2 = 5$. The setup of the problem is as follows. The independent random variables X and Y have binomial distributions with parameters (n_1, p_1) and (n_2, p_2), respectively. We are interested in testing the following hypothesis

$$H_0 : p_1 = p_2 \quad vs \quad H_1 : p_1 < p_2$$

at significance level $\alpha = 0.025$. Additionally, suppose we observe the pair $(0, 4)$ and the prespecified target alternative are $p_1^* = 0.2$ and $p_2^* = 0.45$.

Initially, we have

$$A^{(0)} = \left\{ \begin{array}{l} (0,1),(0,2),(0,3),(0,4),(0,5),(1,2),(1,3),(1,4), \\ (1,5),(2,3),(2,4),(2,5),(3,4),(3,5),(4,5) \end{array} \right\}$$

and $B^{(0)} = \{\}$. To calculate the value of z_0, we first calculate the approximated value of $\max_{0 \le p_1 = p_2 \le 1} h(x, y \mid p_1, p_2)$ through gridding method for all possible outcomes (x, y) belong to $A^{(0)}$, then assign the smallest value to z_0. This tends to give $z_0 = 0.0010$, $C^{(0)} = \{(0,5)\}$, $d_0 = 1$, $C^* = \{(0,5)\}$, $B^{(1)} = \{(0,5)\}$ and $A^{(1)} = A^{(0)} \backslash \{(0,5)\}$.

Next, follow the above procedure, we have $z_1 = 0.0068$, $C^{(1)} = \{(0,4),(1,5)\}$ and $d_1 = 2$. Since $d_1 > 1$, we use target alternative p_1^* and p_2^*, which are equal to 0.2 and 0.45, respectively, to get C^*. We then have $C^* = \{(0,4)\}$ since $f_{X,Y}(0,4 \mid 0.2, 0.45) > f_{X,Y}(1,5 \mid 0.2, 0.45)$. This implies that $B^{(2)} = \{(0,5),(0,4)\}$ and $A^{(2)} = A^{(1)} \backslash \{(0,4)\}$. Moreover, since the observed pair $(0,4)$ belong to $B^{(2)}$, the algorithm is end and the corresponding p-value is equal to $z_1 = 0.0068$.

Table 2.1 gives the p-values for all possible outcomes in $A^{(0)}$. For all the other possible outcomes (x, y), which p-values are greater than 0.3770. Moreover, for $\alpha = 0.025$, the rejection region of FSBT is

$$R_{5,5,0.025,FSBT} = \{(0,5),(0,4),(1,5)\}.$$

However, if we set the significance level $\alpha = 0.05$, then the rejection region is

$$R_{5,5,0.05,FSBT} = \{(0,5),(0,4),(1,5),(2,5),(0,3)\}.$$

Table 2.1: P-values of each possible outcome (x, y) when $n_1 = n_2 = 5$.

	y					
x	0	1	2	3	4	5
0	-	0.2454	0.0911	0.0283	0.0068	0.0010
1	-	-	0.3605	0.1397	0.0547	0.0107
2	-	-	-	0.2793	0.1719	0.0309
3	-	-	-	-	0.3770	0.0914
4	-	-	-	-	-	0.2454
5	-	-	-	-	-	-

3 Comparisons of the testing procedures

The main work of this section is to compare the tests which are mentioned in Section 2 under certain chosen sample sizes. Suppose $R_{n_1,n_2,\alpha,*}$ be the corresponding rejection region of a chosen test, then the actual probability of rejecting null hypothesis is equal to

$$g_{n_1,n_2,\alpha,*}(p_1, p_2) = \sum_{(x,y)\in R_{n_1,n_2,\alpha,*}} f_{X,Y}(x, y \mid p_1, p_2).$$

The function $g_{n_1,n_2,\alpha,*}$ is called null power function as $p_1 = p_2 = p$, otherwise, it is called power function. Then the size of the test is obtained from the supremum of null power function, that is

$$\sup_{0\le p_1=p_2\le 1} g_{n_1,n_2,\alpha,*}(p_1, p_2).$$

In Section 3.1, we compare the possible probability of type I error by null power function curves with nominal level $\alpha = 0.025$, and then compare the power by power function curves in Section 3.2. For a further understanding of these methods, the comparisons of BT, MBT, FET, MFET and FSBT are through numerical computation and demonstrated by the figures. Furthermore, we also calculate the area under curve. In addition, $p_1^* = 0.2$ and $p_2^* = 0.45$ are chosen to be the target alternative for FSBT.

3.1 Probability of type I error

Figure 3.1 shows the null power function curves of FET, BT, MBT, MFET and FSBT under certain chosen sample sizes. We can see that the curves of FET are obviously far from 0.025 in all cases. On the other hand, BT has the similar problem as FET for equal sample sizes, especially when p is close to boundary. Although, for unequal sample sizes, the maximum of curve is close to 0.025, it still has poor performance when p is close to the boundary.

Next, for equal sample sizes, the curves of MBT, MFET and FSBT are uniformly close to 0.025. And the shapes of the curves of BT, FET, MBT and MFET are symmetrical, but the shape of FSBT is not symmetrical. Moreover, with the sample sizes increasing, the curves of both MBT and FSBT are almost overlapped to each other and having higher values than that of MFET as p close to 0 or 1. However, for unequal sample sizes, the shapes of curves of MBT gradually decrease with p being close to 1.

We give also the area under curve, the maximum probability of null power function curve and the corresponding point p of maximum value in Tables 3.1 and 3.2. It is noteworthy that the area under curve is equivalent to calculate

$$\int_0^1 g_{n_1,n_2,\alpha,*}(p,p)dp.$$

As the results in Table 3.1, the area under curve of both BT and FET are significantly less than 0.025 even the sample size is up to 100. On the other hand, MFET shows the limitation of using the hypergeometric distribution as the reference of building order set, so that the area under curve does not converge to 0.025 as sample size increase. In the case of equal sample sizes, both MBT and FSBT mimic to each other, however, the areas under curve of FSBT are slightly higher than that of MBT in all the cases.

For unequal sample sizes (see Table 3.2), BT, FET, MFET and FSBT have similar conclusion as equal sample sizes. However, the areas under curve of MBT are poor in these cases.

3.2 Power

In this subsection, we give further comparison of the power of the five testing methods through power function curves under certain chosen sample size. Figure 3.2 and 3.3 show the power function curves of FET, BT, MBT, MFET and FSBT when $p2 - p1 = 0.1$ and $p2 - p1 = 0.25$, respectively.

Figure 3.2 shows that the curves of both FET and BT are lower than the other curves. Especially, the curves of BT are obviously lower than the other curves when p_1 or p_2 close to 0 or 1 in all cases. Even if the distance between p_1 and p_2 is increased to 0.025 (Figure 3.2), the curves of BT is still have poor performance when p_1 or p_2 close to 0 or 1.

The results of power function curves have the same pattern as that of null power function. For the results of area under curves, please refer to Tables 3.3 to 3.6.

It is noteworthy that the area under curves when $p_2 - p_1 = 0.1$ and $p_2 - p_1 = 0.25$ are calculated by

$$\frac{1}{0.9}\int_0^{0.9} g_{n_1,n_2,\alpha,*}(p_1, p_1 + 0.1)dp_1 \text{ and } \frac{1}{0.75}\int_0^{0.75} g_{n_1,n_2,\alpha,*}(p_1, p_1 + 0.25)dp_1,$$

respectively.

For equal sample sizes as $p_2 - p_1 = 0.1$, in Table 3.3, the areas under curve of both BT and FET are lower than the other even for sample sizes up to 100. Moreover, we can obviously observe that the areas under curve of both MBT and FSBT are close to each other and it is higher than of MFET. However, the areas under curve of FSBT are higher than of MBT. For unequal sample sizes in Table 3.4, the areas under curve of both BT and FET are still low, but of MBT become poor as BT and FET. Moreover, the areas under curve of FSBT are still higher than of MFET. Furthermore, even though all of the areas under curve are raised when $p_2 - p_1 = 0.025$, there are the same results as above described.

4 Discussion and conclusion

According to the numerical results in Section 3, several conclusions can be summarized. The first is that both FET and BT have the problem of achieving target level α. Therefore, FET and BT are not suitable for analyzing the comparative binomial trials. The second is that, both MFET and MBT have effective improvement for FET and BT, respectively. However, for unequal sample sizes, the modification of MBT is not appropriate to BT. Furthermore, although MFET uniformly improves FET, but the fact that MFET use the hypergeometric distribution to as the sequence of order. MFET still has the limitation of the conditional distribution. Therefore, generally speaking, for analyzing the two-sample comparative binomial trial, our method FSBT performs well and is easy to understand and interpret.

The shape of the null power function curve of FSBT is not symmetrical due to the using of target alternative. If we do not consider the condition of target alternative during the course of ordering as $d_i > 1$, in turn, we lay down that all such possible outcomes are to give the same rank. Then, the null power function curve of FSBT becomes symmetrical. This will tend to have smaller power than of considering target alternative.

As the example in Subsection 2.5.1, when $z_1 = 0.0068$, the possible outcomes $(0, 4)$ and $(1, 5)$ are the next candidates for sequence. If we do not consider the condition of target alternative, they are given the same rank. Then, the p-values of the possible outcomes $(0, 4)$ and $(1, 5)$ are equal to 0.0107. If we set target $\alpha = 0.01$, the rejection region only contains the possible outcome $(0, 5)$. However, if we consider the condition of target alternative, the rejection region will contain two possible outcomes $(0, 4)$ and $(1, 5)$. Moreover, the null power function curve of FSBT is closer to target $\alpha = 0.01$ than of which is not consider the target alternative condition.

Furthermore, we want to further understand the rationality of the sequence of the possible outcomes for the rejection region. Therefore, here we give two examples to discussion. For equal sample sizes, as $n_1 = n_2 = 25$, the rejection region of MFET does not contain the possible outcomes $(0, 4)$ and $(21, 25)$, which are included in the rejection

region of both MBT and FSBT. The rejection region of FSBT includes the possible outcomes $(10, 17)$ and $(7, 14)$, which are not included in of both MBT and MFET. The rejection region of FSBT does not contain the possible outcome $(6, 13)$, but it is included in of both MBT and MFET.

In the first place, we discuss the possible outcomes $(10, 17)$, $(7, 14)$, and $(6, 13)$. For MBT, the numerator values of these possible outcomes $(10, 17)$, $(7, 14)$ and $(6, 13)$ are same to each other, so the sequence of possible outcomes is affected by denominator. The denominator of the possible outcomes $(10, 17)$, $(7, 14)$ and $(6, 13)$ are 0.1353, 0.1339 and 0.1312, respectively. The denominator of $(6, 13)$ is smaller than the other, so it is included in rejection region. Then for MFET, we can observe that the probabilities of the possible outcomes $(10, 17)$, $(7, 14)$ and $(6, 13)$ are 0.0289, 0.0257 and 0.0232, respectively, which are calculated by hypergeometric distribution. The probability of the possible outcome $(6, 13)$ is smaller than the other, so it is significance than the possible outcome $(7, 14)$ and $(6, 13)$. And then FSBT may be effected by parameter p.

Next, we discuss the possible outcomes $(0, 4)$ and $(21, 25)$. In generally, we tend not to believe that $p_1 = p_2$ when the distance between $\hat{p}_1$ and $\hat{p}_2$ is large. The rejection region of both FSBT and MBT contain the possible outcomes $(0, 4)$ and $(21, 25)$ which distance between X and Y are equal to 4. However, the rejection region of FSBT does not contain the possible outcome $(6, 13)$ and of MBT does not contain the possible outcomes $(10, 17)$ and $(7, 14)$, which distance are equal to 7. In this case, it is reasonable for MBT. Because when $\hat{p}_1$ or $\hat{p}_2$ close to 0 or 1, the denominator will be small. The denominator of the possible outcome $(6, 13)$ is 0.1312 and of both $(0, 4)$ and $(21, 25)$ are equal to 0.0733. Moreover, when the distance of numerator is in a reasonable range, the main effect of sequence of MBT is still dependent on denominator. In addition, we also can think that if the placebo group has no success cases, and treatment group have four success cases. At this time, we will tend to believe that the success rates of the two groups may be different when the distance between the success rates of two groups is reasonable. Therefore, for FSBT, the possible outcomes $(0, 4)$ and $(21, 25)$ are significance than $(10, 17)$ and $(7, 14)$,

it is reasonable.

And then, for unequal sample sizes, as $n_1 = 15$ and $n_2 = 25$, the possible outcomes

$$\left\{ \begin{array}{l} (12,25),(11,25),(10,24),(9,23),(9,22), \\ (8,22),(7,20),(6,19),(5,17),(3,13) \end{array} \right\}$$

are not included in the rejection region of MBT, but which is included in the rejection region of FSBT. And the rejection region of MBT does not contain these possible outcomes

$$\left\{ \begin{array}{l} (2,11),(4,15),(5,17),(6,19),(7,20), \\ (8,21),(8,22),(9,23),(10,24),(11,25) \end{array} \right\},$$

but which is included in the rejection region of MFET. Moreover, the rejection region of MBT contains the possible outcome $(0,6)$, which is not contained in the rejection region of both MFET and FSBT.

From these case, we can see that MBT trend to easily accept some possible outcomes into rejection region, which is only little contribution to power. That is because MBT uses the test statistic θ_{MBT} to sequence the order for the rejection region, where θ_{MBT} is affected by standard error estimator of $\hat{p_1} - \hat{p_2}$. According to the standard error estimator of $\hat{p_1} - \hat{p_2}$, we can find that the value of denominator of θ_{MBT} would easily tend to smaller than the other while $\hat{p_1}$ or $\hat{p_2}$ close to 0 or 1. Moreover, here we focus on discussing the case of $n_1 < n_2$, so we can find that the possible outcomes, which $\hat{p}_1 = 0$, can be easier ordered in the front than the possible outcomes which $\hat{p}_2 = 1$.

References

[1] Agresti, A. (2002). *Categorical Data Analysis.* John Wiley & Sons, Inc.

[2] Barnard, G. A. (1945). A new test for 2×2 tables. *Nature,* **156**, 177.

[3] Barnard, G. A. (1947a). Significance tests for 2×2 tables. *Biometrika,* **34**, 123-138.

[4] Barnard, G. A. (1947b). The meaning of a significance level. *Biometrika,* **34**, 179-182.

[5] Berkson, J. (1978a). In dispraise of the exact test. *J. Statist. Planning Inference,* **2**, 27-42.

[6] Berkson, J. (1978b). Do the marginal totals of the 2×2 table contain relevant information respecting the table proportions? *J. Statist. Planning Inference,* **2**, 43-44.

[7] Boschloo, R. D. (1970). Raised conditional level of significance for the 2×2 table when testing the equality of two probabilities. *Statistica Neerlandica,* **24(1)**, 1-35.

[8] Crans, G. G. and Shuster, J. J. (2008). How conservative is Fisher's exact test? A quantitative evaluation of the two-sample comparative binomial trial. *Statistics in Medicine,* **27**, 3598-3611.

[9] Fisher, R. A. (1935). The logic of inductive inference. *Journal of the Royal Statistical Society,* **98**, 39-82.

[10] Fisher, R. A. (1971). *The Design of Experiments.* Hafner Publishing Company, New York.

[11] Kempthorne, O. (1979). In dispraise of the exact test: reactions. *J. Statist. Planning Inference,* **3**, 199-213.

[12] Martín Andrés, A. and Herranz Tejedor, I. (2009). Comments on 'How conservative is Fisher's exact test? A quantitative evaluation of the two-sample comparative binomial trial'. *Statistics in Medicine,* **28**, 173-174.

[13] McDonald, L. L., Davis, B. M. and Milliken, G. A. (1977). A non-randomized unconditional test for comparing two proportions in a 2×2 contingency table. *Technometrics,* **19**, 145-150.

[14] Suissa, S., and Shuster, J. J. (1985). Exact unconditional sample sizes for the 2×2 binomial trial. *Journal of the Royal Statistical Society, Series A,* **148**, 317-327.

[15] Upton, G. J. G. (1982). A comparison of Alternative Tests for the 2×2 comparative trial. *Journal of the Royal Statistical Society, Series A,* **145**, 86-105.

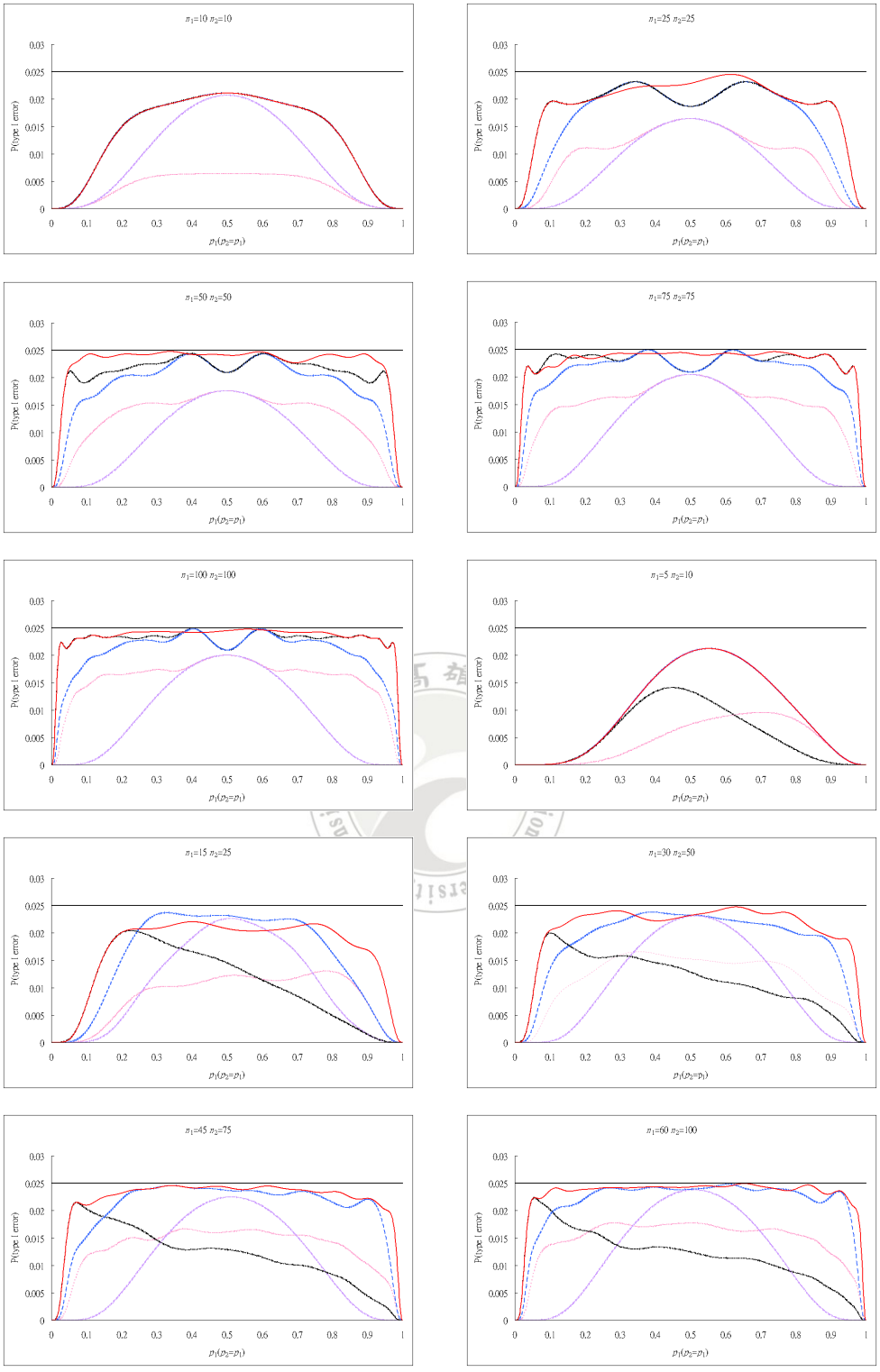


Figure 3.1: The curves of the power functions of FET (dotted line), BT (dotted-dashed line), MBT (double-dotted dashed line), MFET(dashed line) and FSBT (solid line) when $p_1 = p_2$.

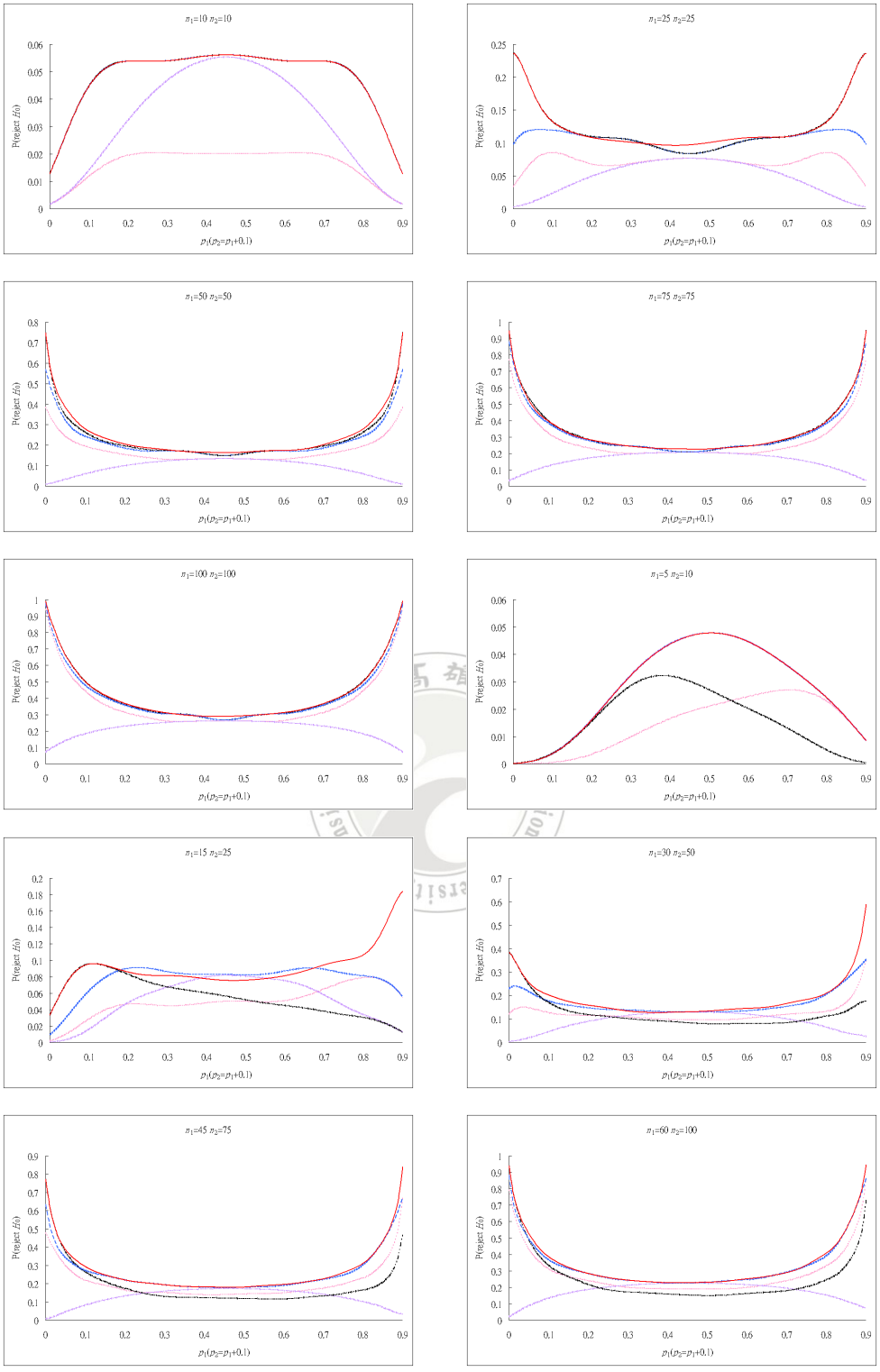


Figure 3.2: The curves of the power functions of FET (dotted line), BT (dotted-dashed line), MBT (double-dotted dashed line), MFET (dashed line) and FSBT (solid line) when $p_2 = p_1 + 0.1$.

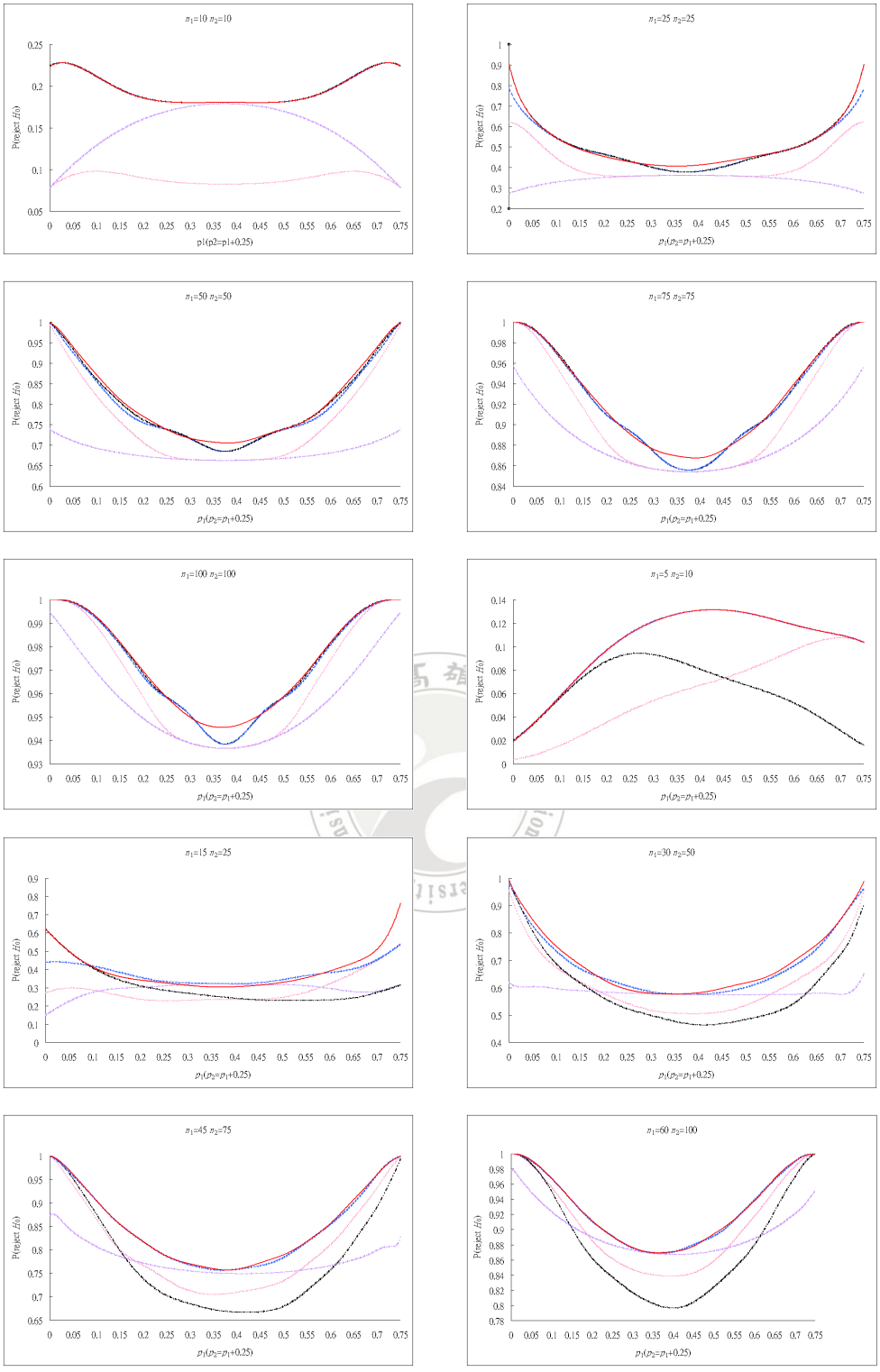


Figure 3.3: The curves of the power functions of FET (dotted line), BT (dotted-dashed line), MBT (double-dotted dashed line), MFET (dashed line) and FSBT (solid line) when $p_2 = p_1 + 0.25$.

Table 3.1: The results of comparisons under null hypothesis for equal sample sizes.

(n_1, n_2)		BT	MBT	FET	MFET	FSBT
$(10, 10)$	max	0.0207	0.0211$*$	0.0064	0.0211$*$	0.0211$*$
		$(p = 0.5)$	$(p = 0.5)$	$(p = 0.5)$	$(p = 0.5)$	$(p = 0.5)$
	area under curve	0.0096	0.0137	0.0040	0.0137	0.0137
$(25, 25)$	max	0.0164	0.0232	0.0164	0.0232	0.0245
		$(p = 0.5)$	$(p = 0.343)$,	$(p = 0.5)$	$(p = 0.344)$,	$(p = 0.614)$
			$(p = 0.657)$		$(p = 0.656)$	
	area under curve	0.0074	0.0184	0.0100	0.0163	0.0191
$(50, 50)$	max	0.0176	0.0244	0.0176	0.0243	0.0248
		$(p = 0.5)$	$(p = 0.394)$,	$(p = 0.5)$	$(p = 0.397)$,	$(p = 0.337)$
			$(p = 0.606)$		$(p = 0.603)$	
	area under curve	0.0080	0.0208	0.0126	0.0186	0.0226
$(75, 75)$	max	0.0204	0.0249	0.0204	0.0249	0.0246
		$(p = 0.5)$	$(p = 0.38)$,	$(p = 0.5)$	$(p = 0.38)$,	$(p = 0.738)$
			$(p = 0.62)$		$(p = 0.62)$	
	area under curve	0.0094	0.0223	0.0148	0.0205	0.0227
$(100, 100)$	max	0.0200	0.0248	0.0200	0.0248	0.0248
		$(p = 0.5)$	$(p = 0.403)$,	$(p = 0.5)$	$(p = 0.404)$,	$(p = 0.559)$
			$(p = 0.597)$		$(p = 0.596)$	
	area under curve	0.0092	0.0226	0.0157	0.0203	0.0232

1. $*$ represents that these three tests have the same rejection region as $\alpha = 0.025$.
2. The row of max is used to record the maximum value of null power function.
3. p is used to record the location of the maximum occurs.
4. The row of area under curve is used to record the area under curve.

Table 3.2: The results of comparisons under null hypothesis for unequal sample sizes.

(n_1, n_2)		BT	MBT	FET	MFET	FSBT
$(5, 10)$	max	$0.0212*$	0.0141	0.0096	$0.0212*$	$0.0212*$
		$(p = 0.552)$	$(p = 0.45)$	$(p = 0.703)$	$(p = 0.552)$	$(p = 0.552)$
	area under curve	0.0098	0.0056	0.0044	0.0098	0.0098
$(15, 25)$	max	0.0227	0.0204	0.0130	0.0237	0.0220
		$(p = 0.509)$	$(p = 0.226)$	$(p = 0.784)$	$(p = 0.328)$	$(p = 0.401)$
	area under curve	0.0105	0.0103	0.0083	0.0155	0.0171
$(30, 50)$	max	0.0231	0.0200	0.0166	0.0238	0.0248
		$(p = 0.515)$	$(p = 0.101)$	$(p = 0.352)$	$(p = 0.389)$	$(p = 0.629)$
	area under curve	0.0108	0.0112	0.0116	0.0183	0.0207
$(45, 75)$	max	0.0225	0.0215	0.0166	0.0246	0.0246
		$(p = 0.512)$	$(p = 0.071)$	$(p = 0.38)$	$(p = 0.342)$	$(p = 0.342)$
	area under curve	0.0105	0.0116	0.0130	0.0202	0.0220
$(60, 100)$	max	0.0238	0.0223	0.0177	0.0248	0.0250
		$(p = 0.51)$	$(p = 0.055)$	$(p = 0.283)$	$(p = 0.609)$	$(p = 0.653)$
	area under curve	0.0112	0.0118	0.0146	0.0213	0.0229

1. $*$ represents that these three tests have the same rejection region as $\alpha = 0.025$.
2. The row of max is used to record the maximum value of null power function.
3. p is used to record the location of the maximum occurs.
4. The row of area under curve is used to record the area under curve.

Table 3.3: The results of comparisons under alternative hypothesis as $p_2 = p_1 + 0.1$ for equal sample sizes.

(n_1, n_2)		BT	MBT	FET	MFET	FSBT
$(10, 10)$	max	0.0553	0.0561$*$	0.0204	0.0561$*$	0.0561$*$
		$(p_1 = 0.45)$	$(p_1 = 0.45)$	$(p_1 = 0.268)$,	$(p_1 = 0.45)$	$(p_1 = 0.45)$
				$(p_1 = 0.632)$		
	area under curve	0.0331	0.0484	0.0162	0.0484	0.0484
$(25, 25)$	max	0.0765	0.2364	0.0853	0.1201	0.2364
		$(p_1 = 0.45)$	$(p_1 = 0)$,	$(p_1 = 0.098)$	$(p_1 = 0.832)$	$(p_1 = 0)$,
			$(p_1 = 0.9)$			$(p_1 = 0.9)$
	area under curve	0.0481	0.1213	0.0714	0.1059	0.1236
$(50, 50)$	max	0.1343	0.7497	0.3839	0.5688	0.7497
		$(p_1 = 0.45)$	$(p_1 = 0)$,	$(p_1 = 0)$,	$(p_1 = 0)$,	$(p_1 = 0)$,
			$(p_1 = 0.9)$	$(p_1 = 0.9)$	$(p_1 = 0.9)$	$(p_1 = 0.9)$
	area under curve	0.0942	0.2311	0.1714	0.2178	0.2427
$(75, 75)$	max	0.2062	0.9496	0.7729	0.8811	0.9496
		$(p_1 = 0.45)$	$(p_1 = 0)$,	$(p_1 = 0)$,	$(p_1 = 0)$,	$(p_1 = 0)$,
			$(p_1 = 0.9)$	$(p_1 = 0.9)$	$(p_1 = 0.9)$	$(p_1 = 0.9)$
	area under curve	0.1602	0.3338	0.2800	0.3242	0.3357
$(100, 100)$	max	0.2617	0.9922	0.9424	0.9763	0.9923
		$(p_1 = 0.45)$	$(p_1 = 0)$,	$(p_1 = 0)$,	$(p_1 = 0)$,	$(p_1 = 0)$,
			$(p_1 = 0.9)$	$(p_1 = 0.9)$	$(p_1 = 0.9)$	$(p_1 = 0.9)$
	area under curve	0.2167	0.4170	0.3695	0.4059	0.4212

1. $*$ represents that these three tests have the same rejection region as $\alpha = 0.025$.
2. The row of max is used to record the maximum value of power function.
3. p_1 is used to record the location of the maximum occurs.
4. The row of area under curve is used to record the area under curve.

Table 3.4: The results of comparisons under alternative hypothesis as $p_2 = p_1 + 0.1$ for unequal sample sizes.

(n_1, n_2)		BT	MBT	FET	MFET	FSBT
$(5, 10)$	max	0.0478	0.0323$*$	0.0270	0.0478$*$	0.0478$*$
		$(p_1 = 0.505)$	$(p_1 = 0.381)$	$(p_1 = 0.705)$	$(p_1 = 0.505)$	$(p_1 = 0.505)$
	area under curve	0.0283	0.0162	0.0146	0.0283	0.0283
$(15, 25)$	max	0.0817	0.0958	0.0793	0.0912	0.1841
		$(p_1 = 0.453)$	$(p_1 = 0.111)$	$(p_1 = 0.813)$	$(p_1 = 0.224)$	$(p_1 = 0.9)$
	area under curve	0.0519	0.0562	0.0498	0.0781	0.0900
$(30, 50)$	max	0.1293	0.3839	0.3526	0.3526	0.5886
		$(p_1 = 0.464)$	$(p_1 = 0)$	$(p_1 = 0.9)$	$(p_1 = 0.9)$	$(p_1 = 0.9)$
	area under curve	0.0900	0.1203	0.1219	0.1664	0.1829
$(45, 75)$	max	0.1723	0.7729	0.6711	0.6711	0.8410
		$(p_1 = 0.462)$	$(p_1 = 0)$	$(p_1 = 0.9)$	$(p_1 = 0.9)$	$(p_1 = 0.9)$
	area under curve	0.1276	0.1826	0.1997	0.2529	0.2642
$(60, 100)$	max	0.2247	0.9424	0.8626	0.8828	0.9470
		$(p_1 = 0.459)$	$(p_1 = 0)$	$(p_1 = 0.9)$	$(p_1 = 0)$	$(p_1 = 0.9)$
	area under curve	0.1783	0.2447	0.2777	0.3300	0.3374

1. $*$ represents that these three tests have the same rejection region as $\alpha = 0.025$.
2. The row of max is used to record the maximum value of power function.
3. p_1 is used to record the location of the maximum occurs.
4. The row of area under curve is used to record the area under curve.

Table 3.5: The results of comparisons under alternative hypothesis as $p_2 = p_1 + 0.25$ for equal sample sizes.

(n_1, n_2)		BT	MBT	FET	MFET	FSBT
$(10, 10)$	max	0.1788∗	0.2278	0.0977	0.2278∗	0.2278∗
		$(p_1 = 0.375)$	$(p_1 = 0.724)$,	$(p_1 = 0.099)$,	$(p_1 = 0.724)$,	$(p_1 = 0.724)$,
			$(p_1 = 0.026)$	$(p_1 = 0.651)$	$(p_1 = 0.026)$	$(p_1 = 0.026)$
	area under curve	0.1475	0.1965	0.0891	0.1965	0.1965
$(25, 25)$	max	0.3617	0.9038	0.6217	0.7863	0.9038
		$(p_1 = 0.375)$	$(p_1 = 0)$,	$(p_1 = 0)$,	$(p_1 = 0)$,	$(p_1 = 0)$,
			$(p_1 = 0.75)$	$(p_1 = 0.75)$	$(p_1 = 0.75)$	$(p_1 = 0.75)$
	area under curve	0.3394	0.5034	0.4172	0.4955	0.5093
$(50, 50)$	max	0.7378	0.9995	0.9930	0.9979	0.9995
		$(p_1 = 0)$,	$(p_1 = 0)$,	$(p_1 = 0)$,	$(p_1 = 0)$,	$(p_1 = 0)$,
		$(p_1 = 0.75)$	$(p_1 = 0.75)$	$(p_1 = 0.75)$	$(p_1 = 0.75)$	$(p_1 = 0.75)$
	area under curve	0.6824	0.8001	0.7566	0.7955	0.8066
$(75, 75)$	max	0.9569	0.9999	0.9999	0.9999	0.9999
		$(p_1 = 0)$,	$(p_1 = 0)$,	$(p_1 = 0)$,	$(p_1 = 0)$,	$(p_1 = 0)$,
		$(p_1 = 0.75)$	$(p_1 = 0.75)$	$(p_1 = 0.75)$	$(p_1 = 0.75)$	$(p_1 = 0.75)$
	area under curvee	0.8835	0.9239	0.9078	0.9233	0.9258
$(100, 100)$	max	0.9946	0.9999	0.9999	0.9999	0.9999
		$(p_1 = 0)$,	$(p_1 = 0)$,	$(p_1 = 0)$,	$(p_1 = 0)$,	$(p_1 = 0)$,
		$(p_1 = 0.75)$	$(p_1 = 0.75)$	$(p_1 = 0.75)$	$(p_1 = 0.75)$	$(p_1 = 0.75)$
	area under curve	0.9562	0.9718	0.9651	0.9715	0.9729

1. $*$ represents that these three tests have the same rejection region as $\alpha = 0.025$.
2. The row of max is used to record the maximum value of power function.
3. p_1 is used to record the location of the maximum occurs.
4. The row of area under curve is used to record the area under curve.

Table 3.6: The results of comparisons under alternative hypothesis as $p_2 = p_1 + 0.25$ for unequal sample sizes.

(n_1, n_2)		BT	MBT	FET	MFET	FSBT
$(5, 10)$	max	$0.1314*$	0.0942	0.1076	$0.1314*$	$0.1314*$
		$(p_1 = 0.428)$	$(p_1 = 0.27)$	$(p_1 = 0.701)$	$(p_1 = 0.428)$	$(p_1 = 0.428)$
	area under curve	0.1034	0.0635	0.0610	0.1034	0.1034
$(15, 25)$	max	0.3203	0.6217	0.5387	0.5387	0.7639
		$(p_1 = 0.349)$	$(p_1 = 0)$	$(p_1 = 0.75)$	$(p_1 = 0.75)$	$(p_1 = 0.75)$
	area under curve	0.2927	0.3006	0.2879	0.3772	0.3896
$(30, 50)$	max	0.6519	0.9929	0.9626	0.9806	0.9930
		$(p_1 = 0.75)$	$(p_1 = 0)$	$(p_1 = 0.75)$	$(p_1 = 0)$	$(p_1 = 0)$
	area under curve	0.5823	0.5920	0.6195	0.6824	0.6910
$(45, 75)$	max	0.8760	0.9999	0.9995	0.9998	0.9999
		$(p_1 = 0.006)$	$(p_1 = 0)$	$(p_1 = 0)$	$(p_1 = 0)$	$(p_1 = 0)$
	area under curve	0.7775	0.7734	0.8035	0.8452	0.8475
$(60, 100)$	max	0.9790	0.9999	0.9999	0.9999	0.9999
		$(p_1 = 0.001)$	$(p_1 = 0)$	$(p_1 = 0)$	$(p_1 = 0)$	$(p_1 = 0)$
	area under curve	0.8962	0.8802	0.9049	0.9261	0.9266

1. $*$ represents that these three tests have the same rejection region as $\alpha = 0.025$.
2. The row of max is used to record the maximum value of power function.
3. p_1 is used to record the location of the maximum occurs.
4. The row of area under curve is used to record the area under curve.