Journal of Data Science 12(2014), 279-294

Forward selection two sample binomial test

Kam-Fai Wong[1], Weng-Kee Wong[2], Miao-Shan Lin[1]

[1]*Institute of Statistics, National University of Kaohsiung, Kaohsiung, Taiwan*
[2]*Department of Biostatistics, University of California, Los Angeles, CA 90095, U.S.A.*

Abstract: Fisher's exact test (FET) is a conditional method that is frequently used to analyze data in a 2×2 table for small samples. This test is conservative and attempts have been made to modify the test to make it less conservative. For example, Crans and Shuster (2008) proposed adding more points in the rejection region to make the test more powerful. We provide another way to modify the test to make it less conservative by using two independent binomial distributions as the reference distribution for the test statistic. We compare our new test with several methods and show that our test has advantages over existing methods in terms of control of the type 1 and type 2 errors. We reanalyze results from an oncology trial using our proposed method and our software which is freely available to the reader.

Key words: 2×2 contingency table; Binary data; Conditional test; Fisher's exact test; Power function;

1. Introduction

TFisher's exact test (FET) is a popular test for testing whether the two proportions from the two classifications are equal in a 2×2 contingency table when the sample size is small. The test assumes the row and column totals in the table are fixed in advance and so it is a conditional test. The marginal totals are "ancillary statistics " and so do not provide information on the common value of the two proportions when the null hypothesis holds. However, this assumption may not hold in practice and the test has remained somewhat controversial over the years, see for example, Barnard (1947), Tocher (1950), Berkson (1978), Kempthorne (1979), who all argued in different ways that not all 2×2 contingency tables are analyzable by the Fisher's exact test. The debate as to which statistical methodology is most appropriate for analyzing a two-sample comparative binomial trial continues to date.

The FET is commonly used in comparative binomial trials and the test statistic has a hypergeometric distribution that does not depend on the unknown parameter p, the common mean of the binomial proportions under the null hypothesis. Boschloo (1970) and McDonald et al. (1977) noted that the actual probability of type 1 error from the FET is frequently lower than the nominal type 1 error rate Crans and Shuster (2008) reaffirmed similar findings and reported the same phenomenon holds even for sample sizes as large as 125 subjects per group. They provided

an algorithm that included extra points in the rejection region to provide additional power for the test. Our goal in this paper is to propose a new modification of the FET to make it less conservative by using two independent binomial distributions as the reference distribution for the test statistic. We compare our new test with competing methods and show our test has better control over the type 1 and type 2 error rates.

The next section reviews unconditional tests and conditional tests for equality of the proportions in two independent binomial samples. Section 3 describes our proposed test and Section 4 compares its performance with other tests in terms of power and type 1 error rate. We offer a discussion in Section 5 and close in Section 6 with an application of our test to reanalyze an oncology trial for treating colon cancer patients with Cetuximab using our self-developed

2. Tests for two independent binomial

We review two classical unconditional tests and two conditional tests for testing equality of two proportions from two independent binomial samples. The two unconditional tests are the binomial test (BT) and the modified two sample binomial test (MBT) proposed by Suissa and Shuster (1985). The two conditional tests are the Fisher's exact test (FET) and the modified Fisher's exact test (MFET) proposed by Crans and Shuster (2008).

Throughout we have two independent samples of size n1 and n2 and the binary outcomes are coded 0 for failure and 1 for success. Let X and Y be the number of successes from the two samples with binomial distributions having parameters (n1 , p1) and (n2 , p2). The joint distribution of X and Y is

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

To fix ideas, we focus in this paper hypothesis of the form H0: p1 = p2 vs H1: p1 < p2 . Other forms can be similarly dealt with. We denote the observed number of successes from the first and second samples by x∗ and y∗ respectively, and the sample proportions by pˆ1 = $\frac{X}{n_1}$ and pˆ2 = $\frac{Y}{n_2}$

2.1 Two sample binomial test (BT)

An obvious test statistic for comparing the difference of the means of two populations is to use the differ- ence of the two sample means. For testing proportions, the test statistic is

$$\theta_{BT}(x,y) = \hat{p}_1 - \hat{p}_2$$

and the p-value of the test for an observed outcome (x^*, y^*) is

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

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

BT uses the simple statistic for comparing two independent binomial proportions but does not account for the variability in the observed outcome pair (x, y). For example, when we have sample sizes n1 = n2 = 5, the following pair of ordered outcomes are possible: (0, 3), (1, 4) and (2, 5) and any one of them will result in the same significance level for the BT test. However, these possible outcomes occur with different probabilities and this makes it possible that the power of the BT is less than that of FET even though the former uses an exact distribution and the latter uses a conditional distribution.

2.2 Modified two sample binomial test (MBT)

Practitioners typically use unconditional tests based on normal approximation when the sample sizes are large. Unconditional tests are appealing because they are easier to explain and understood by non- statisticians. A large sample test statistic for evaluating equality of two proportions from two independent samples is

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

Similarly, given any observed value (x^*, y^*), the p-value is determined from

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

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

The above test introduced by Suissa and Shuster (1985) can be regarded as a modified version of the

Binomial test and we abbreviate this test as MBT. Both approaches uses two independent binomial distri- butions to calculate their p-values, but the MBT incorporates the variation of pˆ1 − pˆ2 and the variability information from the observed outcome pair (x, y). For instance, when we have sample sizes n1 = n2 = 5, the p-values of the possible outcomes (0, 3), (1, 4) and (2, 5) are all equal to 0.055 for the BT. For MBT, the p−values of the possible outcomes (0, 3), (2, 5) are equal to 0.031 but the p-value of the possible outcome (1, 4) is 0.055. This implies that at the 0.05 nominal significance level, the possible outcomes (0, 3), (1, 4) and (2, 5) are considered not significant for the BT but the possible outcomes (0, 3), (2, 5) are significant for the MBT. This simple example shows that the MBT can be more powerful than the BT test when we have the

same sample size. We note that some outcomes will result in having a zero standard error for the estimated difference in the two proportions. When this happens, modifications will have to be made to the test statistic value. For example, when X = n1 and Y = n2 , we would let θM BT (x, y) = −∞ when pˆ1 = 0 and pˆ2 = 1; let θM BT (x, y) = 0 when pˆ1 = pˆ2 ; and let θM BT (x, y) = ∞ when pˆ1 = 1 and pˆ2 = 0.

2.3 Fisher's exact test (FET)

The FET is widely used in the analysis of 2 × 2 contingency table to test the significance of the association between the two kinds of classification when the sample size is small. FET is an exact conditional test because it assumes that the marginal totals are fixed in advance. This assumption eliminates nuisance parameters in the problem and provides an exact null distribution for the test statistic. Specifically, suppose X and Y are independent random variables each with a binomial distribution. Under the null hypothesis, the conditional distribution of X given X + Y has a hypergeometric distribution, which does not depend on the common value of two binomial proportions:

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

For any observed value (x^*, y^*), the exact p-value is calculated from

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

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

Kempthorne (1979) criticized the method because it did not take into account other possible types of data. For instance, the table could also arise from just fixing only one of the marginal totals or none at all. This sentiment was expressed earlier by Barnard (1947), who also emphasized the need to analyze data depending on how the data was collected, and that not all 2 × 2 tables are analyzable by the FET. Barnard (1947) also pointed that the assumption of having fixed marginal totals can pose interpretation difficulties.

2.4 Modified Fisher's exact test (MFET)

A key assumption of the FET is that the marginal totals in the 2 ×2 table are fixed in advance. Conse- quently, the FET is derived from the conditional sample space rather than the set of all possible outcomes. A long outstanding problem with the FET is that its actual probability of type 1 error can be seriously smaller than the pre-specified type 1 error rate α. Crans and Shuster (2008) proposed an adjustment to FET, that increases the power by adding possible outcomes to the rejection region while maintaining the pre-specified size of the test. The modified FET, which we abbreviate as MFET, defines a new significance level α∗ = α + ε, where α is the pre-specified

nominal level and ε is a small positive number. The critical region is determined by using α∗ instead of α. Specifically, for any given sample sizes (n1 , n2), α∗ is the largest value such that

$$\max_{0\le p_1=p_2\le 1} \sum_{(x,y)\in C_{FET,n_1,n_2,\alpha^*}} f_{X,Y}(x,y \mid p_1,p_2) \le \alpha,$$

where CF ET ,n1 ,n2 ,α∗ = {(x, y)|KF ET (x, y) ≤ α∗}. Crans and Shuster (2008) tabulated adjusted significance levels that link various sample sizes and different significance levels. The cross-reference table enables the researcher to reject the test or not based on the adjusted critical value of the FET.

More generally, for any observed pair of outcome (x∗, y∗), the exact p-value of MFET can be determined from

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

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

In the next section, we propose a test that provides a type 1 error rate closer to the nominal alpha level than any of the tests reviewed here or available in the literature.

3. Forward selection two sample Binomial test (FSBT)

Under the assumed set up, the exact distribution of the set of observations (X, Y) is the product of two in- dependent binomial distributions with the BT, MBT and MFET all using the same distribution to calculate their p-values. The only difference is that the order of possible outcomes is defined in different ways. By comparing results from the two tests MBT and MFET, we found that MBT tends to give higher power than the MFET when we have equal sample sizes. However, MFET tends to outperform the MBT in terms of power when we have unequal sample sizes.

We note that FET is more broadly used in practice than the MFET even though FET frequently is a conservative test. MEFT was developed in part to mitigate this issue by calculating the true value of the significance level using the two-independent binomial distribution. The test ranks the possible outcomes derived from the FET p-value and then uses the two-independent binomial distribution to recalibrate the p-value using the observed outcomes. We propose another way to do the ordering where we directly use the two independent binomial distributions as the reference distribution under the null hypothesis. We call our proposed method the forward selection two sample Binomial test because the procedure of selecting possible outcomes into the rejection region in FSBT is similar to the concept of the forward selection method in a multiple linear regression.

Step 1. Set $i = 0$, $A^{(0)} = \Omega_{X,Y}$, the sample space of (X, Y), and $B^{(0)} = \{\}$.

Step 2. Calculate $z_i = \min_{(x,y)\in A^{(i)}} \{ \max_{0\le p_1=p_2\le 1} h(x, y \mid p_1, p_2)\}$,

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

Step 3. Set $C^{(i)} = \{(x, y) \mid (x, y) \in A^{(i)}, \max_{0\le p_1=p_2\le 1} h(x, y \mid p_1, p_2) = z_i\}$.

Step 4. Set $B^{(i+1)} = B^{(i)} \cup C^{(i)}$ and $A^{(i+1)} = A^{(i)} \setminus C^{(i)}$.

Step 5. If $(x^*, y^*) \notin B^{(i+1)}$, increase i by 1 and go to step 2. Otherwise, set $C_{FSBT} = B^{(i+1)}$ and stop.

For an observed outcome (x^*, y^*), the p-value of the test is

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

As an illustrative case, we now apply the above algorithm to the case when we have n1 = 5 and n2 = 5, the observed outcome pair is (0, 4). We want to test the hypothesis

$$H_0 : p_1 = p_2 \text{ vs } H_1 : p_1 < p_2$$

at the $\alpha = 0.025$ significance level. In Step 1, we first determine that

$$A^{(0)} = \{ (0,1), (0,2), (0,3), ..., (5,0) \}$$

and $B^{(0)} = \{\}$. To calculate the value of z_0 in Step 2, we first calculate the approximated value of $\max_{0\le p_1=p_2\le 1} h(x, y \mid p_1, p_2)$ using a grid method for all possible outcomes (x, y) that belong to $A^{(0)}$ and assign the smallest value to z_0. This gives $z_0 = 0.001$, $C^{(0)} = \{(0, 5)\}$, $B^{(1)} = \{(0, 5)\}$ and $A^{(1)} = A^{(0)}\backslash\{(0, 5)\}$ completing Steps 3 and 4.

The next iteration gives $z_1 = 0.007$, $C^{(1)} = \{(0, 4), (1, 5)\}$. This then gives $B^{(2)} = \{(0, 5), (0, 4), (1, 5)\}$ and $A^{(2)} = A^{(1)}\backslash\{(0, 4), (1, 5)\}$. The observed pair $(0, 4)$ belong to $B^{(2)}$ and the algorithm terminates and returns the p-value as 0.011.

In practice, we may not want to reject H_0 if $X/n_1 \ge Y/n_2$ and α is not too large. To reduce the time complexity, we may exclude possible values (x, y) that satisfy $X/n_1 \ge Y/n_2$ from $A^{(0)}$.

Furthermore, we may have a target alternative for sample size calculation in writing a protocol. In such case, one may include this information to the test procedure. To do it, we can add one more step between step 3 and step 4 in the algorithm as follow

Step 3.5 If there is more than one element in $C^{(i)}$, reset $C^{(i)} = C^*$ where $C^* = \{\arg_{(x,y)\in C^{(i)}} \max_{p_1=p_1^*, p_2=p_2^*} f_{X,Y}(x, y \mid p_1, p_2)\}$,

Table 1 displays p-values for all possible outcomes (x, y) with $x < y$. where the pre-specified target alternatives are $p_1^* = 0.2$ and $p_2^* = 0.45$. For all other possible outcomes (x, y), the p-values are larger than 0.377. With $\alpha = 0.025$, the rejection region of FSBT is $R = \{(0, 5), (0, 4), (1, 5)\}$ and if we set the significance level at $\alpha = 0.05$, the rejection region is $R = \{(0, 5), (0, 4), (1, 5), (2, 5), (0, 3)\}$.

[Table 1 about here.]

4. Performance of the FSBT versus other tests

T We now compare the performance of the various tests discussed in previous sections under different con- figurations and different sample sizes. In subsection 4.1, we compare type 1 error probabilities from the null power function curves when the nominal level $\alpha = 0.025$ and in subsection 4.2, we compare the power function curves for the various tests. We also evaluate other features of the BT, MBT, FET, MFET and FSBT and display their properties graphically, including the area under curve as an indication of the

average performance of the test. For the FSBT, we choose $P_1^* = 0.2$ and $P_2^* = 0.45$ to be the target

Let $R_{n_1,n_2,\alpha}$ be the rejection region of the selected α-sized test. For different values of the proportion p_1 from the first sample and p_2 from the second sample, the power curve of the test is defined by

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

The function $g_{n_1,n_2,\alpha}$ is called the null power function when the null hypothesis $p_1 = p_2$ holds; otherwise, it is called the power function. The size of the test is the supremum of the null power function over the interval $[0,1]$, i.e. $\sup_{0\le p_1=p_2\le 1} g_{n_1,n_2,\alpha}(p_1,p_2)$. In what is to follow, we provide null and power plots for the five tests using different sample sizes. For balanced number of observations from the two samples, we considered cases when $n_1 = n_2 = 10, 25, 50$ and 75, and for unequal number of observations from the two samples, we considered cases when $n_1 = 5$ and $n_2 = 10$, $n_1 = 15$ and $n_2 = 25$, $n_1 = 30$ and $n_2 = 50$, and, $n_1 = 45$ and $n_2 = 75$. These choices are illustrative and for other cases of interest, the plots can be similarly generated from our software described in section 6.

4.1 Probability of type 1 error

G Figures 1a to 1d and Figures 2a to 2d show the null power function curves of BT, FET, MBT, MFET and FSBT when we have balanced and unbalanced sample sizes. Clearly, the curves of FET are far from 0.025 level for all the situations considered. The curves for BT are similarly problematic when we have equal sample sizes and p is close to the boundary. For unequal sample sizes, the maximum of the curve is near 0.025; however, it still has poor performance when p is close to the boundary.

[Figures 1a to 1d about here.]

[Figures 2a to 2d about here.]

We observe that the curves of MBT, MFET and FSBT are uniformly closer to 0.025 than those of FET and BT for balanced sample sizes. When p approaches 0.5, the curves of MBT and MFET overlap each other and decline steeply from the target value of 0.025. In addition, we notice that the curves of MBT and FSBT are closer to the value of 0.025 than the curve of MFET. For unequal sample sizes, the curves of MBT move away from the target as p approaches one

and the curve of MFET no longer drops steeply when p reaches 0.5. In contrast, the FSBT curve is uniformly close to 0.025 and its performance is the best near the boundary among the five methods.

We also use another performance measure of the test by comparing the area under each curve. To do this for each α-sized test, we compute its integral

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

and display its value on the right upper corner of all our figures. For equal sample sizes, the areas under the curves of both the BT and FET are far away from 0.025 as just noted above. However, the areas under the curves of the MBT, MFET and FSBT are closer to 0.025. As the sample sizes increase to 75, we observe that (i) the areas under the curves of the BT and FET are still much below 0.025, (ii) the areas under the MBT and FSBT curves are closer to 0.025 than that of MFET, and (iii) the area under the curve of the FSBT is larger than of those reported for MBT in all cases. For unequal sample sizes, we observe that the area under the curve of MBT is as unsatisfactory as those of BT and FET shown in Figure 6. The areas under both the MFET and FSBT curves are on average closer to 0.025, with the latter being still the closest to the target.

4.2 Power

Figures 3a to 3d and Figures 4a to 4d show the power function curves of FET, BT, MBT, MFET and FSBT when $p2 - p1 = 0.1$ for equal and unequal sample sizes, respectively. The curves of both FET and BT are lower than the curve of BT for equal sample sizes, and lower than both the curves of MFET and FSBT in all the cases. In particular, the curves of BT are obviously lower than the other curves whenever p1 or p2 is close to 0 or 1

[Figures 3a to 3d about here.]

[Figures 4a to 4d about here.]

When the sample sizes are equal, the curves of the MBT, MFET and FSBT almost overlap one another as the sample size increases. However, both the curves of the MBT and FSBT are higher than that of the MFET as p1 or p2 approaches 0 or 1. Moreover, the curves of FSBT are higher than that from the MBT and MFET almost every where. For unequal sample sizes, the curve of the MBT is generally lower than that of FET as p1 strays away from 0.

The area under the curve for each test is shown on the upper corner of the figures. For equal sample sizes, both the MBT and FSBT are close to each other and both are larger than the area under the curve for the MFET. The area under the curve of the FSBT is larger than that of the MBT. For unequal sample sizes, the areas under the curves of both the BT and FET are generally small and we notice that the area of the curve for the MBT is now also small and is as

unsatisfactory as those for the BT and FET. In contrast, the areas under the curves of the FSBT are generally large overall, outperforming even those of the MFET.

5. Discussion

There are several observations from the numerical results in Section 4. The first observation is that both the FET and the BT fail to achieve the target significance level α. The power of the FET is higher than that of the BT when p1 or p2 is close to 0 or 1. The FET utilizes the variability from the hypergeometric distribution and not from the two independent binomial distributions. This explains in part why the perfor- mance of the FET is poor and we do not recommend the FET and the BT for analyzing the comparative binomial trials.

The second observation is that the actual probability of type 1 error from both the BT and FET is smaller than expected and the two tests generally provide low power. The MFET not only has enhanced type 1 error rate, but also has greater power for all sample sizes. The MBT and the MFET have similar properties when we have a balanced design with equal sample sizes in the two groups.

The third observation is that, for equal sample sizes, the curves of the FSBT are unsymmetrical but those of the BT, FET, MBT and MFET are. This is because we use target alternative as the selection principle when we have several candidate points to choose from, in which case we not only consider the information for the common proportion parameter p, but also consider information for the target alternatives.

The fourth observation is that for equal sample sizes, the rejection region of the MBT almost overlaps with that of the FSBT. Even though the MFET uses two independent binomial distributions to calculate the actual probability of type 1 error, the MFET is still based on the hypergeometric distribution. The practical implication of the overlap is that the power of both the MBT and the FSBT are larger than that from the MFET when p1 or p2 approches 0 or 1. This suggests that the MBT and FSBT are suitable tests when we have equal sample sizes.

The fifth observation is that when we have unequal sample sizes, the MBT is not appropriate for ana- lyzing the two-sample comparative binomial trial. The MBT uses the test statistic θM BT to sequence the order in the construction of the rejection region and the denominator in θM BT becomes small when $\hat{p}1$ is close to zero or $\hat{p}2$ is close to unity. Our experience is that the performance of the FSBT is also satisfactory when we have unequal sample sizes.

In summary, the power function curves of the FSBT are almost always higher than the power curves from the other tests considered here. This is especially so when p1 or p2 is close to 0 or 1. The FSBT provides more information for the unknown common parameter p and is generally quite efficient in terms of the number of subjects required in the trial. Another advantage of the test is that it is exact and so does not rely on approximation methods. A drawback of the FSBT is that the order processing required in the test can be time-consuming. However, with improving technology in computing speed, this should not pose a serious problem

Applications

We close with an application of our proposed test to analyze real biomedical study and describe how to use our self-developed software that the reader can freely use to generate the p-value for a one or two-sided test from FSBT and compare results with other tests.

Roock et al. (2008) studied the KRAS mutation status as a candidate marker for predicting survival time in 113 patients with irinotecan refractory metastatic colorectal cancer and treated by cetuximab (CTX) in clinical trials. A predictive model for objective response was constructed using logistic and Cox regres- sion model. Tumor response was classified in one of the following categories according to the response evaluation criteria in solid tumors: complete response (CR), partial response (PR), stable disease (SD) and progressive disease (PD). For purpose of illustrating our proposed analysis using FSBT, we ignore (i) sur- vival outcomes in the study as measured by time of CTX treatment until death (overall survival) or until progression of disease, death from any cause or last radiological assessment (progression free survival) and (ii) patients treated with combitherapy (i.e. CTX with irinotecan). The table below shows results from

28 patients given monotherapy alone, their KRAS status, as measured by wild type or mutant, and binary response status, with SD and PD in one category and PR in the other category:

[Table 2 about here.]

A direct calculation using STATA 13 shows an improper chi-square analysis test produced a value of 3.3816 for the Pearson chi-squared test statistic and a p-value of 0.066. Fisher's exact tests produced a p-value of0.087for a one-sided test and a p-value of0.128 for a two-sided test.

We created a software that produces p-values for our proposed test upon input from the user. The software first prompts for the sample sizes n1 and n2 and number of cases x1 and x2 . The two sample sizes do not have to be equal. The software then prompts for target alternatives and if there is none, input values should be p1 = 0 and p2 = 0. The 2 × 2 table is then displayed along with the alternatives if they were specified. The software automatically computes the p-value for a one-sided test first followed by the p-value for a two-sided test. For the above problem without specifying the target alternative, the p-value for a one-sided test is 0.032 the p-value for a two-sided test is 0.042.

Acknowledgments

The research of Wong reported in this paper was partially supported by the National Institute Of General Medical Sciences of the National Institutes of Health under Award Number R01GM107639. The contents in this paper are solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

[1] Barnard G. A. (1947). Significance tests for 2 × 2 tables. Biometrika 34, 123–138.

[2] Berkson J. (1978). In dispraise of the exact test: Do the marginal totals of the 2 × 2 table contain relevant information respecting the table proportions? Journal of Statistical Planning and Inference 2, 27–42.

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

[4] 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.

[5] De Roock W., Piessevaux H., De Schutter J., Janssens M., De Hertogh G., Personneni N., Biesmans B., Van Laethem J. L., Peeters M., Humblet Y., Van Cutsem E. and Tejpar S. (2008). KRAS wild-type state predicts survival and is associated to early radiological response in metastatic colorectal cancer treated with cetuximab. Annals of Oncology 19, 508–515.

[6] Kempthorne O. (1979). In dispraise of the exact test: reactions. Journal of Statistical Planning and Inference 3, 199–213

[7] 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.

[8] 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.

[9] Tocher K. D. (1950). Extension of hte Neyman-Pearson theory of tests to discontinuous variats. Biometrika 37, 130–144

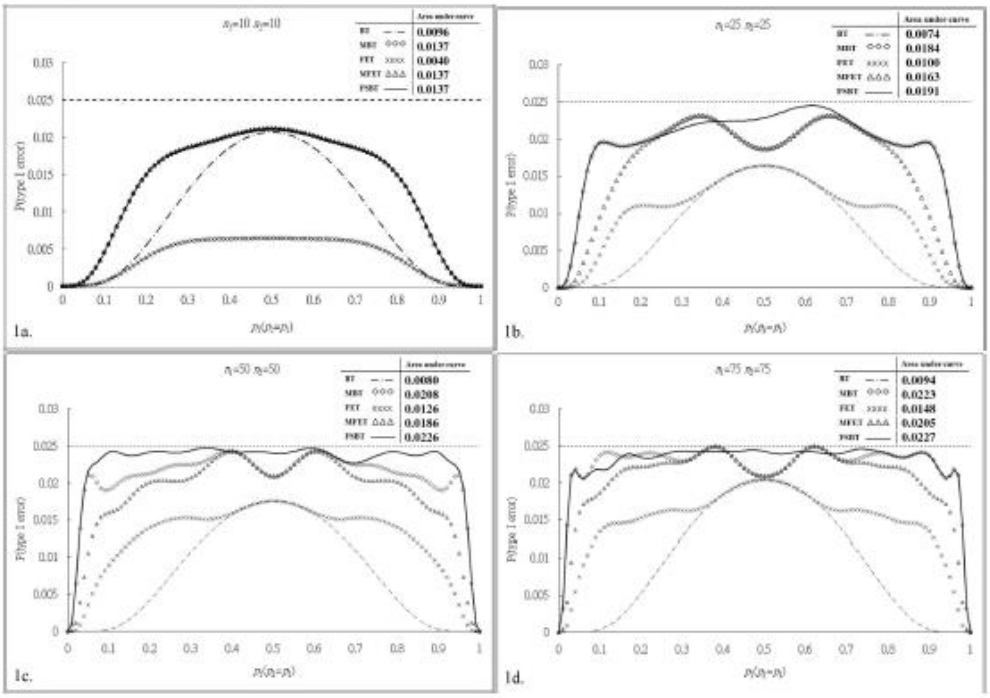
Received November 14, 2014; accepted September 10, 2014.

Table 1 The p-value of the Forward Selection Binomial Test (FSBT) for each possible outcome (x, y) when $n_1 = n_2 = 5$.

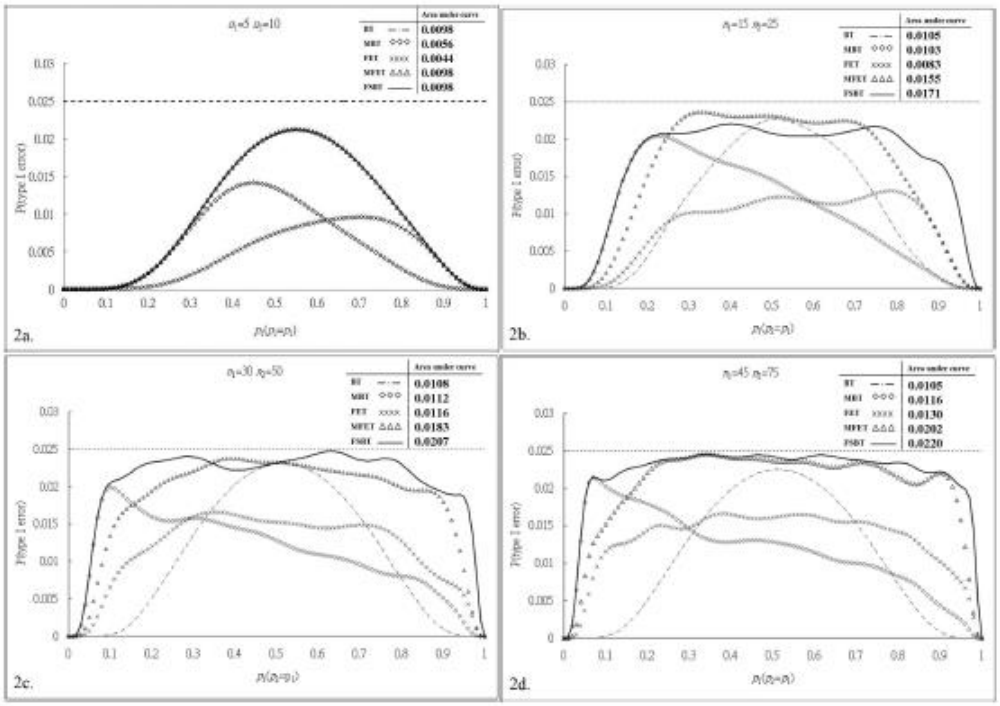
	y				
x	1	2	3	4	5
0	0.245	0.091	0.028	0.007	0.001
1	-	0.360	0.140	0.055	0.011
2	-	-	0.279	0.172	0.031
3	-	-	-	0.377	0.091
4	-	-	-	-	0.245

Table 2 Binary responses (partial response (PR) versus stable disease (SD) and progressive disease (PD) combined, from 28 of the 108 patients given monotherapy only by their KRAS mutation status.

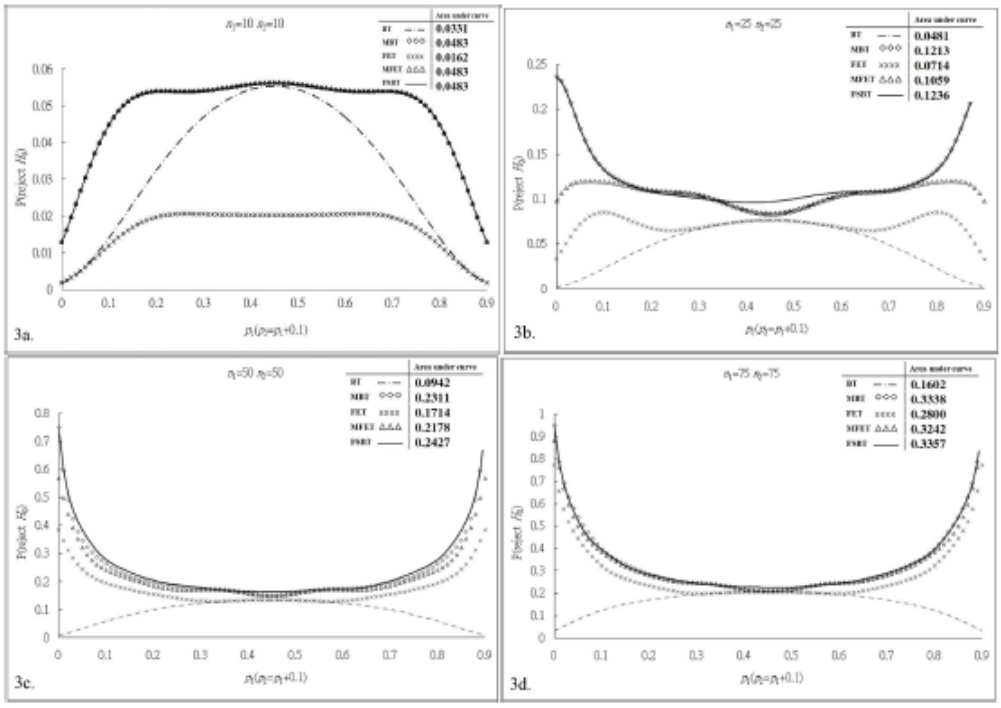
	PR	$SD + PD$	$Total$
wild type	5	13	18
mutant	0	10	10
total	5	23	28



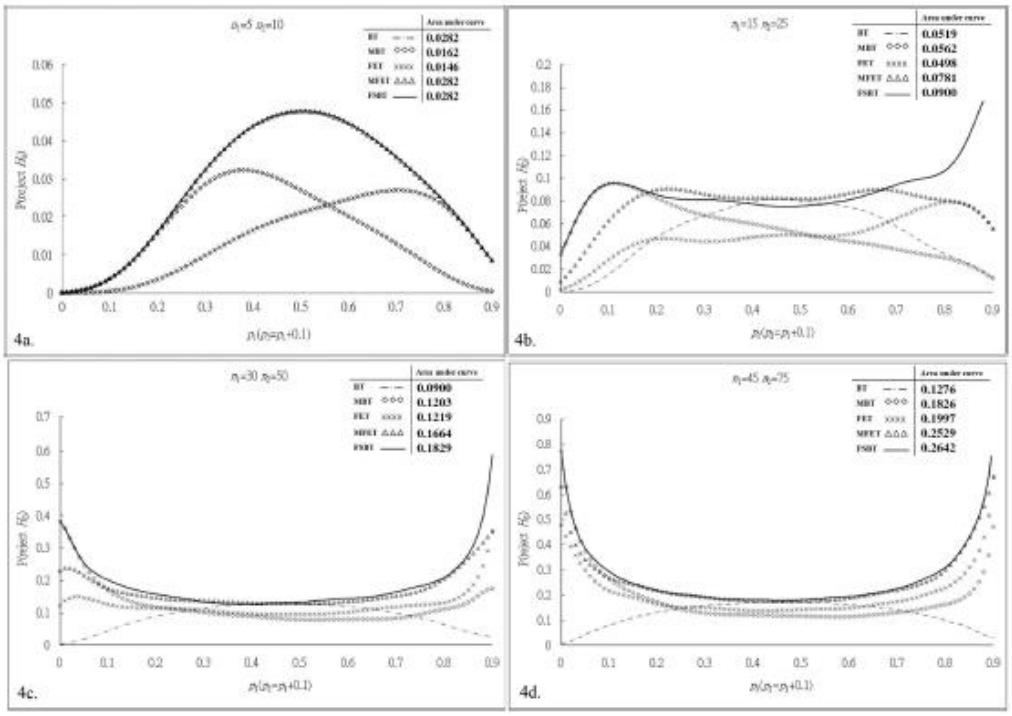
Figures 1a, 1b, 1c and 1d: Null power function curves of BT, FET, MBT, MFET and FSBT for balanced sample sizes.



Figures 2a, 2b, 2c and 2d: Null power function curves of BT, FET, MBT, MFET and FSBT for unbalanced sample sizes.



Figures 3a, 3b, 3c and 3d: Power function curves of BT, FET, MBT, MFET and FSBT for balanced sample sizes.



Figures 4a, 4b, 4c and 4d: Power function curves of BT, FET, MBT, MFET and FSBT for unbalanced sample sizes.