The Chernoff Modification of the Fisher Exact Test for the Difference of Two Binomial Probabilities

by

Hajime Takahashi and Kazuki Uematsu

Graduate School of Economics, Hitotsubashi University

Knitachi, Tokyo 186-8601

March 2006

Abstract

To test the difference of two binomial probabilities, we will apply the Fisher exact test modified by Chernoff (2002), which will be called *CMF test*. The key idea is to use the P-value of Fisher exact test as a test statistic and this makes it possible to apply the Fisher exact test for more general problems. The numerical comparison with the likelihood ratio test and the Bayes test are also conducted and we find

Key Words and Phrases: Fisher Exact Test, Binomial Probabilities, Chernoff Modification, Test of Difference of Probabilities

that the performance of *CMF test* is close to that of the likelihood ratio test and it is satisfactory in various situations. We also give a Kullback-Leibler information number of the margins to for the testing problems which justify the stability of P-values of *CMF test*.

1 Introduction

One of the most simple, but the most profound statistical problem is to compare the two binomial probabilities. The problem has been discussed extensively by many authors in the last 100 years, and some of its history and interesting stories may be found in Yates (1984).

Suppose we are to compare the success rate p and q of treatments A and B, where $p = Pr\{Success|A\}$ and $q = Pr\{Success|B\}$. And, following Yates (1984), we summarize the data in the 2x2 table below:

[TABLE 1]

	Treatment A	Treatment B	Margin
Success	a	b	$m_1 = a + b$
Failure	c	d	$m_2 = c + d$
Margin	n_1	n_2	$n = m_1 + m_2 = n_1 + n_2$

In order to save the space, numerical values for particular tables are given in the form $[a, b, c, d]$ for the entries of the table and $\{m_1, m_2, n_1, n_2 : n\}$ is used for margins. We also write (a, m_1, n_1) to denote the whole table if the total sample size n is being fixed, and we assume that the total sample size n is constant throughout the end of this paper.

The most common problem is to test the hypothesis of equal probability against one sided or two sided alternatives. Whereas we sometimes are interested in the difference of these probabilities, see for example Skipka et.al.(2004) where they consider testing the effect of new treatment under the FDA Draft Guidance (1998), they define p and q as failure probabilities though. In this case, we consider, for some Δ, the one sided hypothesis testing problem;

$$H_0 : q = p + \Delta \qquad vs \qquad H_1 : q < p + \Delta \tag{1}$$

When $\Delta = 0$, this is a famous classical problem and the three standard tests (for the equality of these probabilities) are the Pearson method, the Yates continuity correction to the Pearson test (Yates correction for short), and the Fisher exact test, see for example Pearson (1947), Yates (1984), and Fisher (1941). Also we note the recent paper by Chernoff (2002) where the P-value of the Fisher exact test is used as a test statistic, the method of which will be discussed briefly in Section 2. When the sample sizes n_1 and n_2 are large, all three tests are proved to be almost equivalent. The small sample performance of these three test together with others have been discussed extensively in the last half century and it is found that the discrepancies among the performance of these tests in various situations are highly significant (Yates (1984), Little (1989), D'Agostino et.al. (1988) among others). This comes not only from the asymptotic nature of Pearson and Yates correction, since the exact P-values obtained from the binomial distribution also indicates the same tendencies (Chernoff (2002)). There are no clear cut conclusion as to which test is the best, the Fisher exact test is more conservative though (cf. Little (1989)).

When we consider the Fisher exact test, Chernoff (2002) points out that the composite nature of the null hypothesis gives us some complications. Since the margines m_1 and m_2 are just almost ancillary in our problem (cf. Little (1989)), the P-values of the Fisher exact test depend on the unknown common probability p. The dependance to the unknown p of the P-Value of the Fisher exact test may be measured by the K-L (Kullback-Leibler) information number of the margines and the extensive numerical analysis have been conducted by Chernoff (2004). He finds that these K-L information numbers are very small and stable accross the various sample sizes and this may explane the reason why the P-value of the Fisher exact test is stable (Chernoff (2002)). Although the dependence is weak, the P-value of Fisher exact test is still a function of the unknown p. To obtain the overall P-Value, we need it for each p and then we either obtain the maximum of P-values over the range of unknown p of interests or calculate it at the some point, say at MLE. But, unlike Pearson and Yates correction, the Fisher exact test has no test statistic and it is not clear how to analyze the effect of the nuisance parameter.

Now, to solve the problem, Chernoff (2002) introduces a test statistic based on the P-value of the Fisher exact test, which is a sort of unconditional version of the Fisher exact test. Although the Chernoff's method is not completely justified theoretically, it looks reasonable from the numerical examples he considers. We will discuss his method briefly in the next section. The unconditional version of the Fisher test is also considered by Boschloo (1970), who obtains the maximum of the modified P-values over the entire parameter space. But the way he obtains the unconditional P-value looks rather strange to us.

Besides these tests, the likelihood ratio test, the Bayes test and the Barnard's CSM test (Barnard (1947)) are also applied by many authors, especially when the hypothesis is more complicated. These test procedures are considered to be more general and they are proved to be highly efficient in some applications like our problem. Since, in our problem with $\Delta \neq 0$, the Fisher exact test may not be used in its original form, Skipa et. al. (2004) proposes to apply the CSM test, π_{local}-test (Rohmel and Mansman (1999)), Chan's test (Chan (1998)), and the likelihood ratio test. But they rule out the use of the Fisher exact test as they consider it is good only for the case $\Delta = 0$.

Compare to the classical Pearson and Yate tests, the Chernoff modification of the Fisher exact test gives us numerically more stable results for $\Delta = 0$ (Chernoff (2002), we think that it is worth trying to apply the Fisher's method to our problem. Therefore, in this paper, we will apply the Chernoff's modification of the Fisher exact test to this problem. And we find that the Chernoff's modification is as good as the likelihood ratio test and better than some Bayes procedure.

The paper is orrganized as follows; in the section 2 we briefly go over the classical testing procedures for the simpler problems for exposition; they are the Peason test, the Yates correction and the Fisher exact test. We then introduce the Chernoff Mdification of the Fisher test (*CMF test*). We will consider three test procedures for the problem (1) in Section 3. The numerical comparison of the performances of the tests are presented in Section 4.

2 Review of the Fisher exact test and the Chernoff Modification

We will introduce the Chernoff Modification of the Fisher exact test (CMF) in this section, and for this purpose we will start with reviewing the Pearson and the Yates correction for ease of exposition. To facilitate the exposition, we will consider the following classical testing problem throughout the end of this section.,

$$H_0 : p = q \qquad vs \qquad H_1 : p > q. \tag{2}$$

Suppose we have a data of the form [Table 1] and we will obtain the P-value of the observed data by the several methods. We will abuse the notation that variables a and b may be considered as both random and fixed in an obvious manner according to the context. Now, to obtain the P-values of the Pearson test, we consider

$$p_P^{(B)}(z_P, p) = Pr_{p,q}\{Z_P \geq z_P | p = q\}, \tag{3}$$

Here, p is the true parameter value and we have set,

$$Z_P = \frac{\frac{a}{n_1} - \frac{b}{n_2}}{\sqrt{\hat{p}(1-\hat{p})(n_1^{-1} + n_2^{-1})}} = \sqrt{\frac{(ad-bc)^2 n}{m_1 m_2 n_1 n_2}}, \tag{4}$$

$$where, \quad \hat{p} = \frac{m}{n}$$

and z_P denotes the observed Z_P. The superfix (B) indicates that the probability is calculated under the binomial probability: $a \sim Bi(p, n_1)$ and $b \sim Bi(p, n_2)$. Since the null hypothesis is composite, the P-value against the null hypothesis (2) may be obtained by maximizing $p_P^{(B)}(z_P, p)$ with respect to p over the other set W (say) of interest.

$$p_P^{(B)}(z_P : W) = \max\{p_P^{(B)}(z_P, p) : p \in W\}$$

and we simply denote $p_P^{(B)}(z_P) = p_P^{(B)}(z_P : (0,1))$, for the case $W = (0,1)$. The choice of W may be of interest and we will discuss this problem later in this section.

Now, the usual P-value of the Pearson test is given by the normal approximation to the exact probability (3) and is denoted by

$$p_P^{(N)}(z_P) = 1 - \Phi(z_P) \simeq Pr_{p,q}\{Z_P = z_P | p = q\}, \tag{5}$$

where Φ is the cumulative standard normal distribution function. It is easily seen that the usual P-value does not depend on p, which helps us to use the simpler notation. It follows that the P-value against the null hypothesis is given by (5), and the nature of the composite hypothesis becomes implicit.

In the same way, the exact P-value and it's normal approximation of the Yates' correction are given by

$$p_Y^{(B)}(z_Y : W) = \max\{p_Y^{(B)}(z_Y, p) : p \in W\}$$

where,

$$p_Y^{(B)}(z_Y, p) = Pr_p, q\{Z_Y = z_Y | p = q\}, \tag{6}$$

and

$$p_Y^{(N)}(z_Y) = 1 - \Phi(z_Y)$$

respectively, where Z_Y is given by

$$\begin{aligned} Z_Y &= \frac{\frac{a}{n_1} - \frac{b}{n_2}}{\sqrt{\hat{p}(1-\hat{p})(n_1^{-1} + n_2^{-1})}} - \frac{1}{2}\sqrt{\frac{n}{\hat{p}(1-\hat{p})n_1 n_2}} \\ &= \sqrt{\frac{(|ad - bc| - \frac{1}{2}n)^2 n}{m_1 m_2 n_1 n_2}}, \end{aligned}$$

and z_Y is the observed Z_Y.

Note that when P-values of both the Peason and the Yates correction were introduced in the early 20th century, $p_P^{(N)}(z_P)$ and$p_Y^{(N)}(z_Y)$ were inevitably used because of the computational reasons. But now a days, by using computer, it is not difficult to calculate the exact P-values $p_P^{(B)}(z : W)$ and $p_Y^{(B)}(z : W)$. And these exact values are also compared with the P-values from the Fisher exact test, which gives us more direct comparison of these test procedures (Chernoff (2002)). Of course another merit using the normal approximations are that they are independent of the unknown parameter values as we have mentioned above.

With these feature of the Pearson and the Yates tests in mind, we next discuss the Fisher exact test. Remember that, given all margins $\{m_1, m_2, n_1, n_2 : n\}$, the [Table 1] is determined by specifying the value of a. Then, given the total sample size n, we may denote the conditional probability of getting the [Table 1] by $Z_F(a, m_1, n_1)$, and it is given by

$$Z_F(a) = Z_F(a, m_1, n_1) = \frac{\binom{n_1}{a}\binom{n_2}{b}}{\binom{n}{m_1}} \tag{7}$$

The idea deriving the formula is easily seen, given all margines, the problem is reduced to the hypergeometric probability. We consider the urn containing n balls, n_1 of which are labeled treatment A and n_2 for treatment B. Then the content of the [Table 1] may be obtained by choosing m_1 balls without replacement from the urn, where a are chosen from the treatment A and the the rest $b = m_1 - a$ are assigned to the treatment B. Hence the conditional probability of getting the Table1 is given by the hypergeometric probability (7).

If we are interested in testing the hypothesis of equal probability against the one sided alternative (2); then the larger value of a is an evidence against the null hypothesis. Hence, the P-value of the observed data $\mathsf{x} = (x_1, m_1, n_1)$ is defined by

$$p_F(\mathsf{x}) = p_F(x_1, m_1, n_1) = \sum_{a=x_1}^{\min\{n_1, m_1\}} Z_F(a, m_1, n_1) \tag{8}$$

Here the composite nature of the hypothesis $H_0 : p = q$ is problematic. If (8) is independent of the unknown parameter p, then it gives us a P-value of the composite hypothesis (2) as in the case of normally approximated Pearson and Yates tests. However, the P-value $p_F(\mathsf{x})$ of the Fisher exact test should depend on the unknown nuisance parameter p, unless the margins are ancillary. Unfortunately it is shown that the margin is not an ancillary statistic, it is only approximately ancillary (See, Little(1989), Chernoff (2002)). Therefore, the P-value of the Fisher exact test should depend on the nuisance parameter, and in order to compare it with the other test procedures, we need to obtain the maximum of P-values over the same range W of interests. In order to obtain the maximum of the type one error probability, we usually need a test statistic T (say), and calculate

$$p_{\bullet}(\cdot;W) = \max_{p \in W} Pr_{p,q}\{T \in CR | p = q\}, \tag{9}$$

where CR is a critical region. Unlike the other tests, the Fisher exact test doesn't have a test statistic, and it is not clear how to define it's P-value for each $p \in W$. To solve this problem, Chernoff proposes a procedure which is stated in the following short sentence: "*For the case where the margines are given, we could use "a" as the test statistic. But if the margines are not specified, a more natural test statistic would be the usual P-value using the Fisher exact test, assuming, without proper justification, that the margines are specified*"; See Chernoff (2002). We will obtain the P-value using Chernoff's suggestion below and we call the method *Chernoff Modification* of the Fisher exact test (**CMF test**) and denote it's P-value as p_{CM} throughout the rest of the paper.

For any given data $\mathsf{X} = (x_1, m_1, n_1)$, we will obtain p_{CM} as follows. To start with, we will define the test statistic $T_F = T_F(X_1, M_1, n_1) = p_F(X_1, M_1, n_1)$ of CMF test,

$$T_F = \sum_{a=X_1}^{\min\{n_1, M_1\}} Z_F(a, M_1, n_1). \tag{10}$$

We will write the observed T_F by $t_F = T_F(x_1, m_1, n_1) = p_F(x_1, m_1, n_1)$. Since small P-value is an evidence against the null hypothesis, it follows that the P-value of (x_1, m_1, n_1) at (p, q) is defined by

$$Pr_{p,q}\{T_F \leq t_F | p = q\} \tag{11}$$

To calculate (11), remember that X_1 and X_2 are independently distributed with $B(n_1, p)$ and $B(n_2, q)$, and $M_1 = X_1 + X_2$, we have

$$\begin{aligned} Pr_{p,q}\{X_1 &= x_1, M_1 = m_1\} \\ &= \binom{n_1}{x_1}\binom{n_2}{x_2} p^{x_1} q^{x_2} (1-p)^{n_1 - x_1} (1-q)^{n_2 - x_2} \qquad (12) \\ where\ x_2 &= m_1 - x_1, n_2 = n - n_1 \end{aligned}$$

Hence, under the null hypothesis $p = q$, the P-value of the observed data $\mathbf{x} = (x_1, m_1, n_1)$ is given by

$$p_{CM}(\mathbf{x}; p)$$

$$= \sum_{(a,m) \in CR} \binom{n_1}{a}\binom{n - n_1}{m - a} p^m (1-p)^{n-m} \tag{13}$$

where,

$$CR = \{(a, m) : T_F(a, m, n_1) \leq T_F(\mathbf{x}) \tag{14}$$

In view of (9), the maximum of $p_{CM}(\mathsf{X}; p)$ over the range W is then denoted by

$$p_{CM}((x_1, m_1, n_1); W) = \max_{p \in W} p_{CM}(\mathsf{X}; p)$$

Similar method is proposed by Boschloo (1970), where the unconditional Fisher exact test is constructed directly on the (x_1, x_2) plane. Whereas Chernoff considers the conditional probability as a test statistic. The numerical comparison of the Chernoff's method and the Pearson test and the Yates correction are given in Chernoff (2002). Although his numerical analysis is very restrictive, the Chernoff modification gives us the most stable P-values over the range of $p \in W$ of interest. And this motivate us to use the CMF test to the more general problems.

3 Hypothesis on the difference of probabilities

We go back to the original testing problem (1) with non-zero and fixed Δ

$$H_0 : q = p + \Delta \qquad vs \qquad H_1 : q < p + \Delta$$

In order to obtain the P-value of the test, we will apply the likelihood ratio test, the Bayes test and the *CMF test* and make some numerical comparisons of their performances.

Let us start with the *CMF test*. For any data $\mathbf{x} = (x_1, m_1, n_1)$, the observed *CMF test* statistic t_F is given by $t_F(x_1, m_1, n_1) = \sum_{a=x_1}^{\min\{n_1,m_1\}} Z_F(a, m_1, n_1)$ (see (10)), and we will define the P-value at p by

$$p_{CM}(\mathbf{x}; p, \Delta) = \max_{q \in Q_p} Pr_{p,q}\{T_F(X_1, M_1, n_1) \leq t_F(\mathbf{x})\} \tag{15}$$

where,

$$Q_p = \{q : \min[1, \max\{p + \Delta, 0\}] \leq q \leq 1\}.$$

It follows from (12) that,

$$\begin{aligned} p_{CM}(\mathbf{x}, p) &= p_{CM}((x_1, m_1, n_1); p, \Delta) \\ &= \max_{q \in Q_p} \sum_{(a,m) \in CR} \binom{n_1}{a} \binom{n_2}{b} p^a q^b (1-p)^c (1-q)^d, \end{aligned} \tag{16}$$

where $b, c,$and d are defined as in [Table 1] for each (a, m), and $CR = \{(a, m) : T_F(a, m, n_1) \leq T_F(\mathbf{x})\}$ denotes the critical region of the observed data (cf. (14)).

Here we may consider several P-values against the null hypothesis: The first one (and may be the most natural one) is obtained by maximizing $p_{CM}(\mathsf{X},p)$ over all possible $p \in AP = \{p : \max(0, -\Delta) \leq p \leq \min(1, 1-\Delta)\}$,

$$p_{CM}^{(AP)} = p_{CM}(\mathsf{X},AP) = \max\{p_{CM}((x_1, m_1, n_1); p, \Delta) : p \in AP\}. \quad (17)$$

The second one is to replace p by it's maximum likelihood estimator under the null hypothesis, $\hat{p}^{(0)}$, and it is denoted by

$$p_{CM}^{(MLE)} = p_{CM}(\mathsf{X},\hat{p}^{(0)}) = p_{CM}((x_1, m_1, n_1); \hat{p}^{(0)}, \Delta)$$

The other variation may be to consider the P-value maximum over the range of p such as the 95% confidence interval of p (see Chernoff (2002)), which we do not consider in this paper though. The derivation of $\hat{p}^{(0)}$ is non-standard and it will be presented below (see (21)).

The calculation of $p_{CM}((x_1, m_1, n_1); p, \Delta)$ may require a grid search in q and it may not only be time consuming, but also it may cause the rounding errors. Hence, we will consider some simplified version, which is motivated by the following heuristic argument. In the simple one sided testing problem (2), the test statistic (10) may tend to take a smaller value (more significant) as the value $\theta = p - q$ gets larger; namely, as we go deep into the region of alternative hypothesis. Therefore, the distribution of $T_F(X_1, M_1, n_1)$ may be stochastically decreasing in θ. This suggests that the maximum in (15) and (16) over $q \in Q_p$ is attained at the boundary point

$(p,q)=(p,p^*)$, where $p^*=\min\{1,\max\{p+\Delta,0\}\}$. Indeed, this can be justified by the lemma below.

Lemma 1 *Let $F_\theta(t)$ be the distribution function of T_F. Then it is stochastically decreasing in the sence of Lehmann (1957)*

$$\theta<\theta^{'} \Rightarrow F_\theta(t)\geq F_{\theta'}(t) \quad for\ all\ t \tag{18}$$

Proof. It is easily seen that,

$$T_F(X_1+1,M_1,n_1)\leq T_F(X_1,M_1,n_1).$$

And, by the straightforward algebra, we have

$$\begin{aligned}
& T_F(X_1,M_1+1,n_1)-T_F(X_1,M_1,n_1)\\
= & \sum_{a=X_1}^{\min(n_1,M_1+1)} Z_F(a,M_1+1,n_1)-\sum_{a=X_1}^{\min(n_1,M_1)} Z_F(a,M_1,n_1)\\
\geq & \sum_{a=X_1}^{\min(n_1,M_1)} Z_F(a,M_1+1,n_1)-\sum_{a=X_1}^{\min(n_1,M_1)} Z_F(a,M_1,n_1)\\
= & \binom{n}{n_1}^{-1}\sum_{a=X_1}^{\min(n_1,M_1)}\left[\binom{M_1+1}{a}\binom{n-M_1-1}{n_1-a} - \binom{M_1}{a}\binom{n-M_1}{n_1-a}\right]\\
= & \binom{n}{n_1}^{-1}\sum_{a=X_1}^{\min(n_1,M_1)}\left[\binom{M_1+1}{a}\binom{n-M_1-1}{n_1-a-1} + \binom{M_1}{a-1}\binom{n-M_1}{n_1-a}\right]>0
\end{aligned}$$

It follows that the test statistic T_F is a decreasing function of X_1 and increasing in M_1. It follows that there exists a function $B(X_1, t)$ depending only on X_1 and t, for which

$$T_F \leq t \Leftrightarrow 0 \leq Y_1 \leq B(X_1, t)$$

where, $Y_1 = M_1 - X_1 \sim Bi(n_2, p - \theta)$. Moreover,

$$\begin{aligned} \frac{d}{d\theta} P(0 &\leq Y_1 \leq B(X_1, t) | X_1 = x_1) \\ &= (B(x_1, t) + 1)(p - \theta)^{B(x_1,t)+1}(1 - p + \theta)^{n_2 - B(x_1,t)} \binom{n_2}{B(x_1, t) + 1} > 0, B(x_1, t) \neq n_2. \end{aligned}$$

By, interchanging the order of integration and the differentiation, the lemma follows readily. ■

To obtain the P-values with this version, we first calculate $p_{CM}((x_1, m_1, n_1); p, \Delta) = Pr_{p,p^*}\{p_F(X_1, M_1, n_1) \leq p_F(x_1, m_1, n_1)\}$. It follows from (12) that,

$$\begin{aligned} & p_{CM}((x_1, m_1, n_1); p, \Delta) \\ = & \sum_{(a,m) \in CR} \binom{n_1}{a} \binom{n_2}{m - a} p^a p^{*m-a}(1 - p)^{n_1 - a}(1 - p^*)^{n - m - n_1 + a} \end{aligned}$$

and then we finally obtain the P-values $p_{CM}^{(H_0)}$ and $p_{CM}^{(MLE)}$ by

$$p_{CM}^{(H_0)} = \max\{p_{CM}((x_1, m_1, n_1); p, \Delta) : p \in AP\} \quad (19)$$

$$p_{CM}^{(MLE)} = p_{CM}((x_1, m_1, n_1); \hat{p}^{(0)}, \Delta)$$

respectively. In the next section, we will use these P-values to compare with the P-values from the other tests.

We will next consider the likelihood ratio test and the Bayes test. The log-likelihood for the model is

$$\begin{aligned} l(x_1, x_2, p, q) &= K(x_1, x_2) + x_1 \log p \\ &\quad + x_2 \log q + (n_1 - x_1)\log(1-p) + (n_2 - x_2)\log(1-q) \end{aligned} \tag{20}$$

Since Δ is known, the maximum of $l(x_1, x_2, p, q)$ under H_0 is attained at the boundary $(p, p+\Delta)$ (Rohmel and Mansmann (1999)), the maximum likelihood estimator $\hat{p}^{(0)}$ of p under H_0 is given by the solution of the nonlinear equation

$$\frac{x_1}{p} - \frac{n_1 - x_1}{1-p} + \frac{x_2}{p+\Delta} - \frac{n_2 - x_2}{1-p-\Delta} = 0 \tag{21}$$

We consider the log of the likelihood ratio Z_L for the hypothesis H_0 and the alternative H_1 at the maximum likelihood estimates under the two hypotheses. The maximum under the alternative is easily obtained by

$$\hat{p} = \frac{x_1}{n_1}, \ \hat{q} = \frac{x_2}{n_2}$$

It follows that the log-likelihood ratio is

$$\begin{aligned} Z_L &= x_1 \log \hat{p}^{(0)} + x_2 \log(\hat{p}^{(0)} + \Delta) + (n_1 - x_1)\log(1 - \hat{p}^{(0)}) \\ &\quad + (n_2 - x_2)\log(1 - \hat{p}^{(0)} - \Delta) - n_1 Ent(\hat{p}) - n_2 Ent(\hat{q})) \end{aligned}$$

where $Ent(\cdot)$ is the entropy given by

$$Ent(p) = -[p\log(p) + (1-p)\log(1-p)]$$

Then, the P-value at p is given by

$$\begin{aligned} p_L(z_L; p, \Delta) &= \Pr\{Z_L = z_L | p, p^*\} \\ &= \sum_{(a,b)\in CRA} \binom{n_1}{a}\binom{n_2}{b} p^a p^{*b}(1-p)^{n_1-a}(1-p^*)^{n_2-b} \end{aligned}$$

where $CRA = \{(a,b) : Z_L = z_L\}$ and $p^* = \min\{1, \max\{p+\Delta, 0\}\}$ are defined as before. We also write $p_L = p_L(z_L; AP) = \max\{p_L(z_L; p, \Delta) : p \in AP\}$, $p_{L_1} = p_{L_1}(z_L; \hat{p}^{(0)}) = p_L(z_L; \hat{p}^{(0)}, \Delta)$ and z_L is an observed Z_L.

Finally, we will consider the Bayes method. We assume that the prior distribution of (p, q) is a mixture of two distributions. It takes p and q being independent Beta distributions $Be(\alpha_1, \beta_1)$ and $Be(\alpha_2, \beta_2)$. We will denote the distribution of (p, q) by,

$$\mathsf{L}(p, q) = Be(\alpha_1, \beta_1) * Be(\alpha_2, \beta_2),$$

which means that p has the Beta distribution with parameters $\boldsymbol{\alpha}_1 = (\alpha_1, \beta_1)$ and q has the Beta distribution with parameters $\boldsymbol{\alpha}_2 = (\alpha_2, \beta_2)$ and p and q are independent. Then the posterior distribution, given data (x_1, x_2) (in terms of the notation in the section 1 this is $(x_1, x_1 + x_2, n_1)$) has the same form and is given by

$$\mathsf{L}(p, q | (x_1, x_2)) = Be(\alpha_1^*, \beta_1^*) * Be(\alpha_2^*, \beta_2^*),$$

where $\alpha_1^* = a_1 + x_1, \beta_1^* = \beta_1 + n_1 - x_1, \alpha_2^* = \alpha_2 + x_2, \beta_2^* = \beta_2 + n_2 - x_2$. Then the posterior probability of the hypothesis $H_0 : q = p + \Delta$ being true is given by

$$p_B((\alpha_1, \beta_1), (\alpha_2, \beta_2); (x_1, x_2)) = \frac{1}{Be(\alpha_1^*, \beta_1^*)Be(\alpha_2^*, \beta_2^*)} \int_{t-s=\Delta} s^{\alpha_1^*}(1-s)^{\beta_1^*} t^{\alpha_2^*}(1-t)^{\beta_2^*} ds dt. \tag{22}$$

We will use $p_B(\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2) = p_B((\alpha_1, \beta_1), (\alpha_2, \beta_2); (x_1, x_2))$ as a P-value of the data from the Bayes method.

There are numerous tests which have been developed in the last half century besides these tests we have discussed in this paper (see for example Skipka et. al. (2004)). However, no one have ever tried to apply the Fisher exact test other than the test of equality of the two success probabilities. Although it is not comprehensive, we will perform a numerical comparison of our P-value p_{CM} with the likelihood ratio test p_L and the Bayes test p_B in the next section.

4 Numerical Comparison of tests

We will make a numerical comparison of the tests discussed in the previous section and for this purpose we consider the probability of rejecting the null hypothesis. As have been mentioned, we use $p_B(\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2) = p_B((\alpha_1, \beta_1), (\alpha_2, \beta_2); (x_1, x_2))$ as a proxy for the P-value of the Bayes method. We include the Bayes prcedure with the reservation because $p_B(\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2)$ is not a rejection probability. As to the CMF tests, we first order the sample points in the (x_1, x_2) plane according to the value of the test statistics. We then calculate

$$\sum_{all\ Data} Pr_{p,p+\theta}\{Rejects\ H_o|Data\} Pr_{p,p+\theta}\{Data\}. \tag{23}$$

When $\theta = \Delta$ this gives us a unconditional size of the test, and for $\theta < \Delta$ this gives us a power. To obtain the P-values for the likelihood ratio test, we will utilized the asymptotic distribution of the likelihood ratio obtained by Munk (2004), the accuracy of which is not fully investigated yet. We will compare the following tests:

(1) *the likelihood ratio test* $p_L : (LR)$.

As to the Chernoff modification of the Fisher test, we consider;

(2) *the Chernoff modification with the unknown parameter p estimated by the maximum likelihood* $p_{CM}^{(MLE)} : (Fisher1)$, and

(3) *the Chernoff modification with the maximum over all possible p values* $p_{CM}^{(H_0)}$: $(Fisher2)$.

We choose the following priors for the Bayes tests;

(4) *Bayesian approach with Uniform prior* $(\alpha_i = \beta_i = 1, i = 1, 2) : (Bayes1)$ and

(5) *Bayes test with prior* $(\alpha_i = \beta_i = \frac{1}{2}, i = 1, 2) : (Bayes2)$.

We present the P-values together with the power of the tests against the nuisance parameter $p \in (0, 1)$ of the probability of the success for selected sample sizes. The results are presented in Figure 1 to 12 below. We present the sizes of the tests on the Figure 1 through 6, whereas the powers are shown in the Figure 7 to 12.

If the sample sizes are over 25 in both group, the performances of the tests are similar and the size of these tests are stable over all value of p. When sample sizes are moderate (say, one of the sample is less than 20), we found the Bayesian tests give us rather instable sizes. On a whole, Fisher 1 and Fisher 2, and LR are very stable in p, especially in the mid-range, say at $p \in (.2, .8)$. Whereas, p in the neighborhood of 0 and 1, they are quite conservative, which also reflects the power of these tests. Bayes 1 and 2 are fine for larger sample sizes, but when one of the sample size dropes less than 10, we observe some size distortions. As to the power of the tests, we have examined only cases where $q^* - p^* = 0$ when $\Delta = 0.1$ and $q^* - p^* = -0.2$ when $\Delta = -0.1$. We cannot find dominating test procedure. The performances of these tests are quite "random" and we cannot conclude which test is to be recommended for smaller sample sizes. We, however, can conclude that if the sample sizes of both sample exceed 25, any of these test may be applied. But, when the sample sizes are not too big, we recommend the either Fisher 1 or LR, as the performance of LR and Fisher1 are quite similar and both are fairly stable.

To support the stable nature of Fisher 1, we will also present the Kullback-Leibler information numbers of the margins for $\frac{n_1}{n_2} = c$ with $c = 1, 2$, and 5 in Figures 13 and 14. We find that, in most cases, the information carried by the margines are very small and stable with respect to the sample sizes. We also found the same phenomena as Chernoff (2004) that the K-L information number is decreasing in sample size. These results explain the stability of the Fisher type test for the problem of this paper. We observe the asymmetry of the performances of Chernoff Modification in Δ, the reason for this phenomena may be explained by the results of Munk (2004), which however awaits further studies.

5 Figures and Tables

5.1 Unconditional Size of the tests

Figuire 1 to 6 are the sizes of the tests for $\Delta = \theta = 0.1(Fig.1-3)$ and $\Delta = \theta = -0.1(Fig.4-6)$ for several sample sizes.

(Put Fig 1 to 6 after this line):

Figure 1 :Sample Size $(10, 5)$

Figure 2 Sample size $(25, 5)$

Figure 3 Sample size $(50, 50)$

Figure 4 Sample size $(10, 5)$

Figure 5 Sample size $(25, 5)$

Figure 6 Sample size $(50, 50)$

5.2 Unconditional Power of the tests

Figures 7 to 12 present the power of the test. Figures 7 to 9 are power of the test for $\Delta = 0.1$ and at $\theta = q - p = 0 < \Delta$, and Figures 10 to 12 are power of the test for $\Delta = -0.1$ and at $\theta = q - p = -0.2 < \Delta$.

(Put Fig 7 to 12 around here)

Figure 7 :Sample Size $(10, 5)$

Figure 8 Sample size $(25, 5)$

Figure 9 Sample size $(50, 50)$

Figure 10 Sample size $(10, 5)$

Figure 11 Sample size $(25, 5)$

Figure 12 Sample size $(50, 50)$

5.3 Tables

Figure 13 (d=0.1) and 14 (d=-0.1) show the Kullback-Leiber information numbers of the margines. Each numgber is 100 times the actual number. p^* and q^* are the parameter values in the alternative hypothesis. The entories are the average of $KL_2(H, K) = \min_p E\{L_2|p.p + d\}$ and $KL_2(K, H) = \min_p E\{-L_2|p^*.q^*\}$, where $L_2 = \log\{f(m|p, p + d)/f(m|p^*.q^*)\}$

(Put Fig 13 and 14 around here)

6 References

Barnard, G.A., (1947) "Significance tests for 2x2 tables" Biometrika, 34, 123-138.

Boschloo, R.D., (1970) "Raised conditional level of significance for the 2x2 table when testing the equality of two probabilities", Statist. Neerlandica 24, 1-35.

Chan, I.S.F., (1998) "Exact tests of equivalence and efficacy with a non-zerolower bound for comparative studies", Statist. Med., 17, 1403-1413.

Chernoff, H. (2002) "Another view of the classical problem of comparing two probabilities" J. Iranian Statist. Soc. 1, 35-54.

Chernoff, H. (2004) "Information, for testing the equality of two probabilities, from the margins of the 2x2 table" J. Ststistic. Plan. Inf.,121, 209-214.

D'Agostino, R.B., and Case, W., and Belanger, A.,(1988) "The appropriateness of some common procedures for testing equality of two independent Binomial proportions" The American Statistician, 42,198-202.

FDA (1998) "Guidance for industry. Complicated urinary track infections and pyelonephritis-developing antimicrobial drugs. Drafts guidance. Rockvill, MD: FDA, Center for Drug Evaluation and Research.

Fisher, R.A. (1041) "Statistical Methods or Research Workers, 8th ed.", Edinburg: Oliver and Boyd.

Little, R.J.A., (1989) "Testing the equality of two independent binomial proportions", The American Statistician 43, 283-28.

Pearson, E.S., (1947) "The choice of statistical tests illustrated on the interpretation of tests classed in a 2x2 Table", Biometrika, 34, 139-167.

Rohmel, J. and Mansman, U., (1999) "Unconditional non-asymptotic one-sided tests for independent binomial proportions when the interest lies in showing non-inferiority and/or superiority", Biometrical J. 42, 149-170.

Skipka, G., Munk, A. and Freitag, G., (2004) "Unconditional exact tests for the difference of binomial probailities-contrasted and compared", Compt. Statist. Data Anal. 47, 757-773.

Uematsu, K.,(2005) "On the tests of significance for the 2x2 Table", MA thesis submitted to Hitotsubashi University.

Yates, F., (1984) "Tests of significance for 2x2 contingency tables (with discussion)", J. Roy. Statist. Soc. Ser. A 147, 426-463.

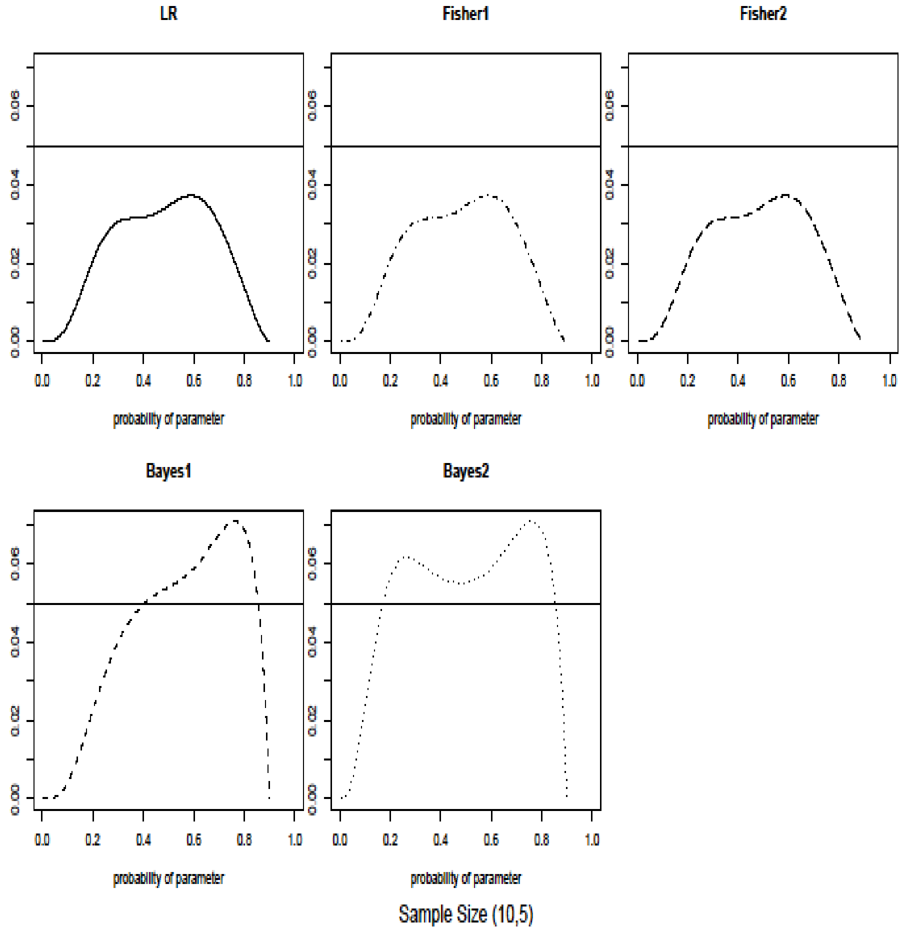

LR
Fisher1
Fisher2
Bayes1
Bayes2
0.00
0.02
0.04
0.06
0.0
0.2
0.4
0.6
0.8
1.0
probability of parameter
Sample Size (10,5)

Figure 1:

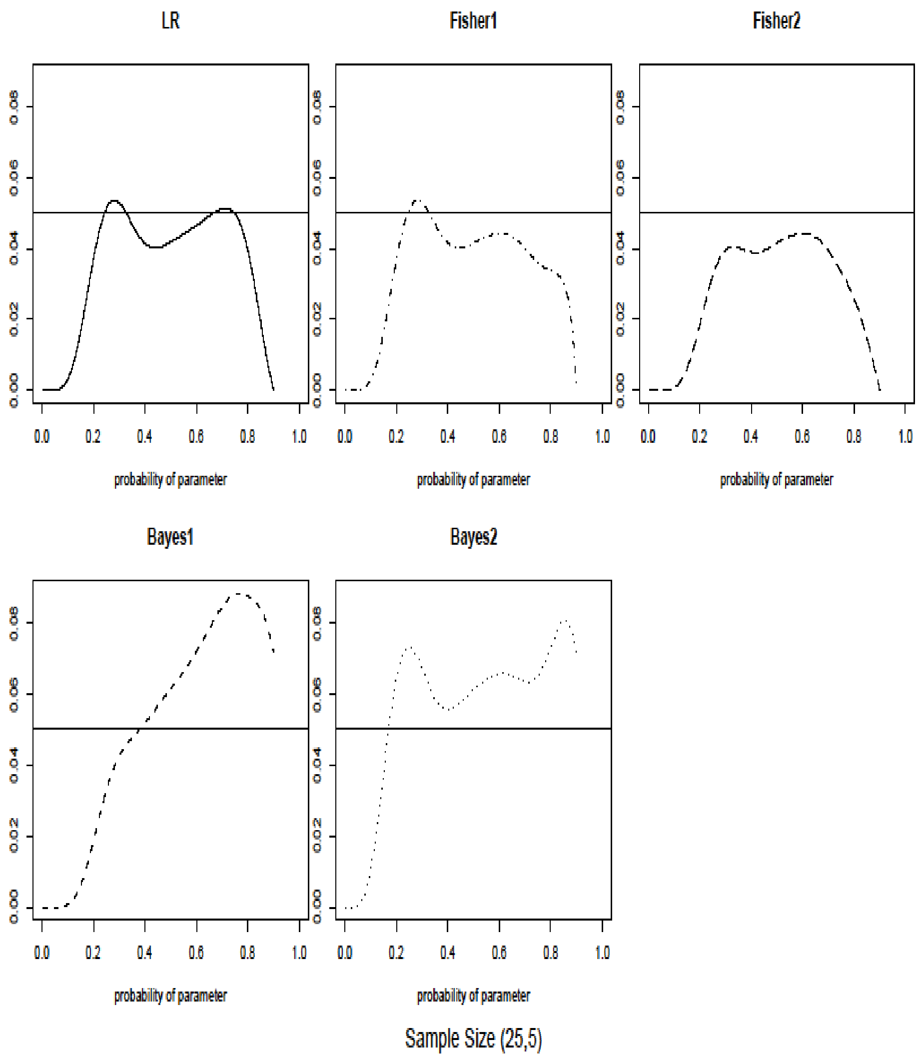

LR
Fisher1
Fisher2
Bayes1
Bayes2
0.00
0.02
0.04
0.06
0.08
0.0
0.2
0.4
0.6
0.8
1.0
probability of parameter
Sample Size (25,5)

Figure 2:

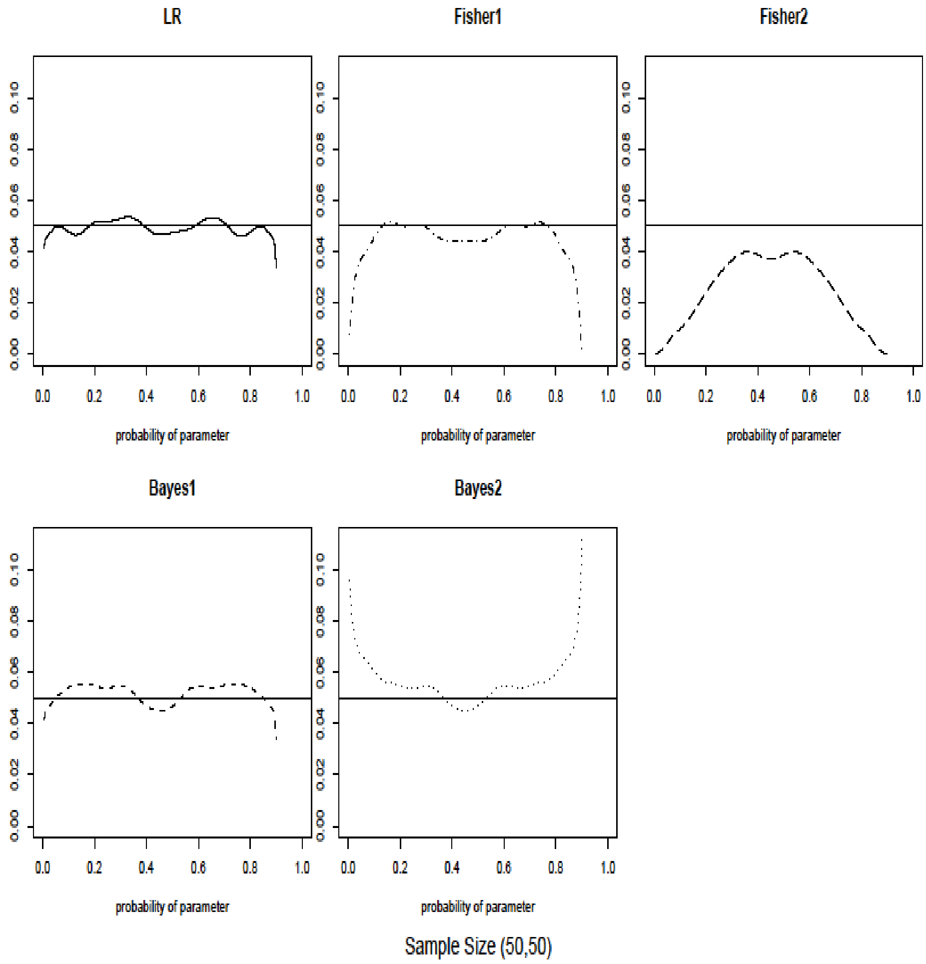

LR
Fisher1
Fisher2
Bayes1
Bayes2
probability of parameter
Sample Size (50,50)

Figure 3:

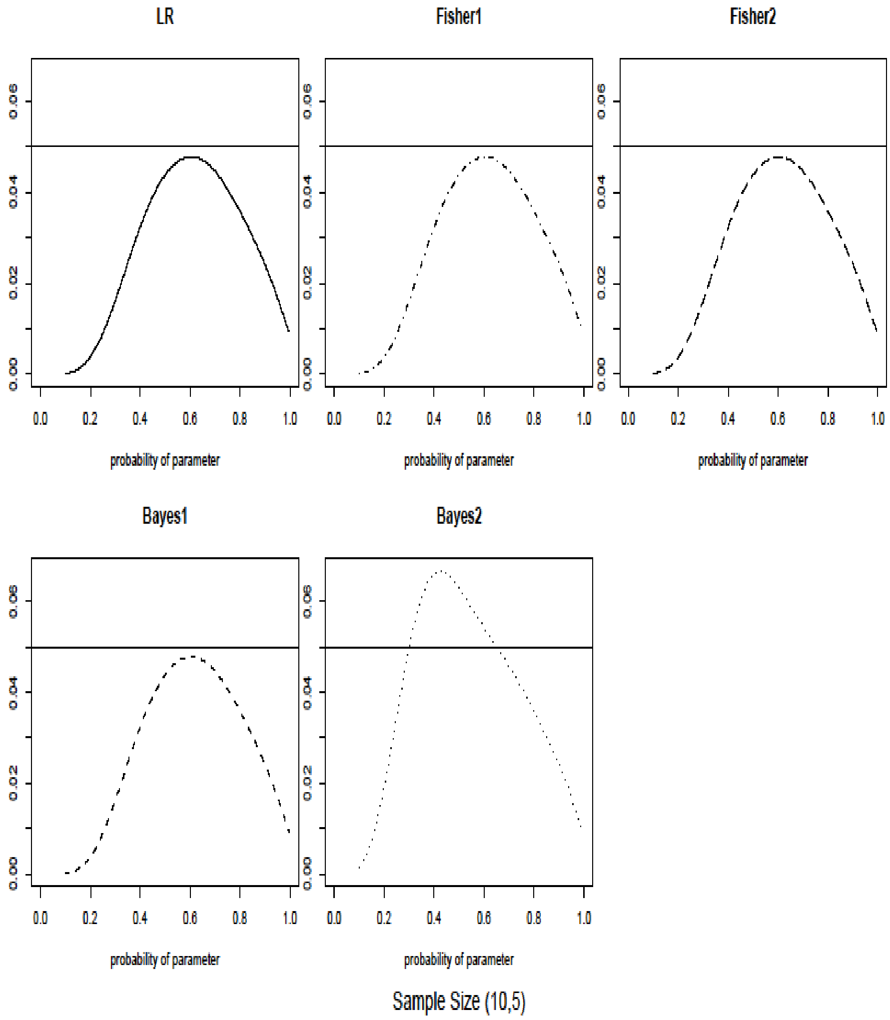

LR
Fisher1
Fisher2
Bayes1
Bayes2
0.00
0.02
0.04
0.06
0.0
0.2
0.4
0.6
0.8
1.0
probability of parameter
Sample Size (10,5)

Figure 4:

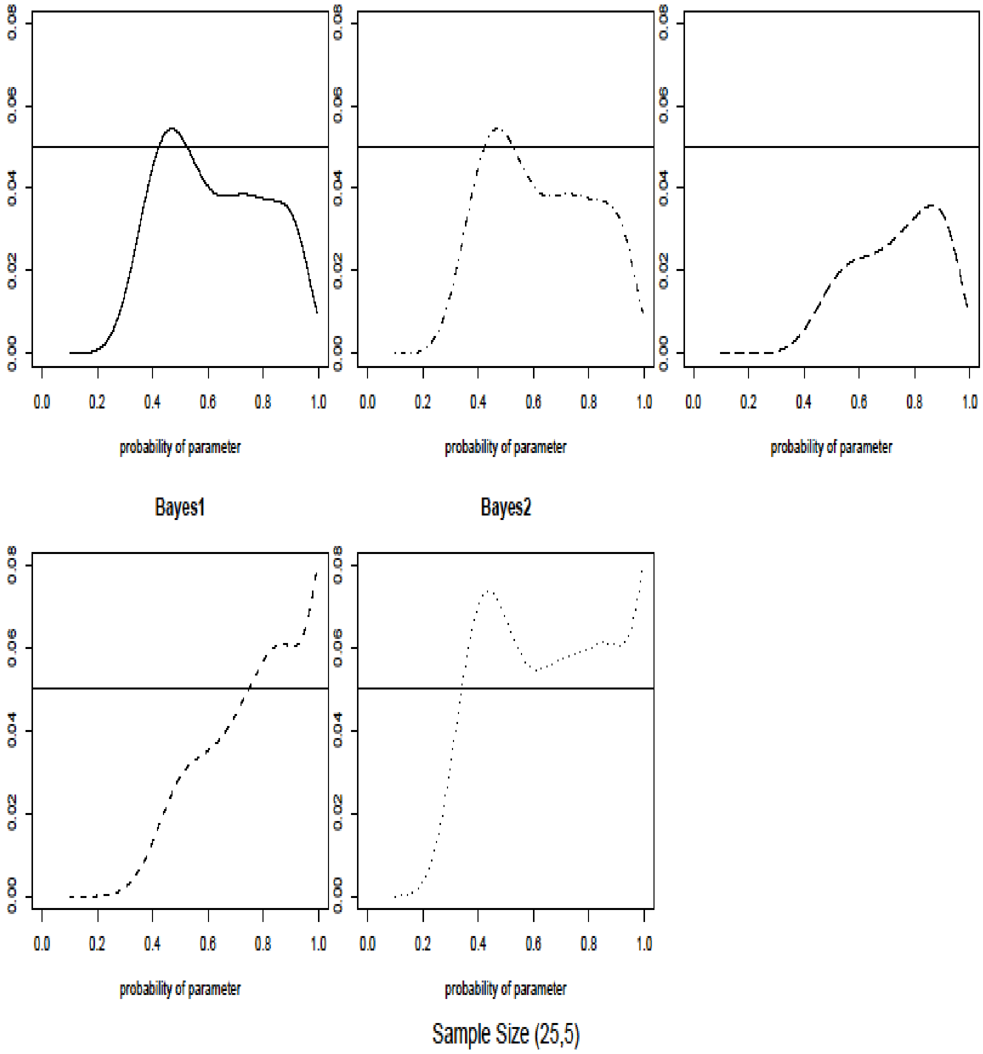
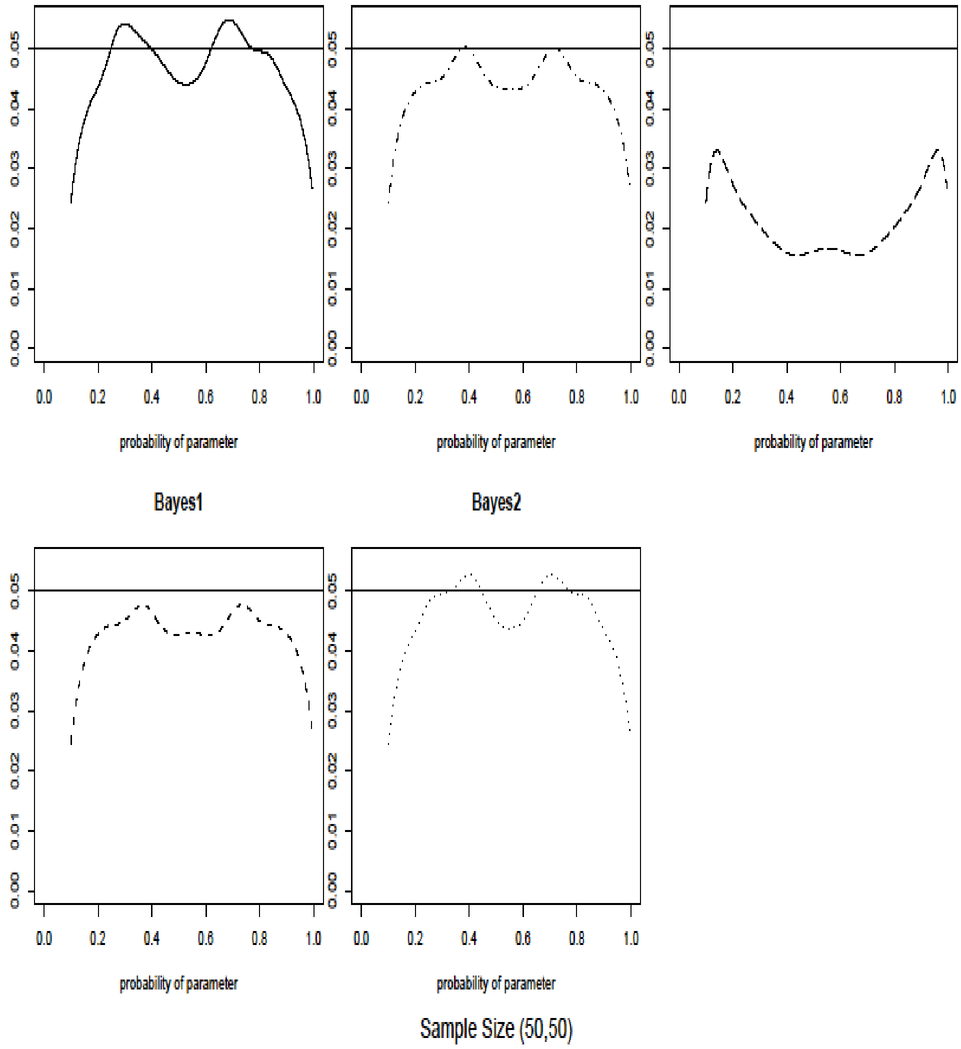

0.00
0.02
0.04
0.06
0.08
0.0
0.2
0.4
0.6
0.8
1.0
probability of parameter
Bayes1
Bayes2
Sample Size (25,5)

Figure 5:



0.00
0.01
0.02
0.03
0.04
0.05
0.0
0.2
0.4
0.6
0.8
1.0
probability of parameter
Bayes1
Bayes2
Sample Size (50,50)

Figure 6:

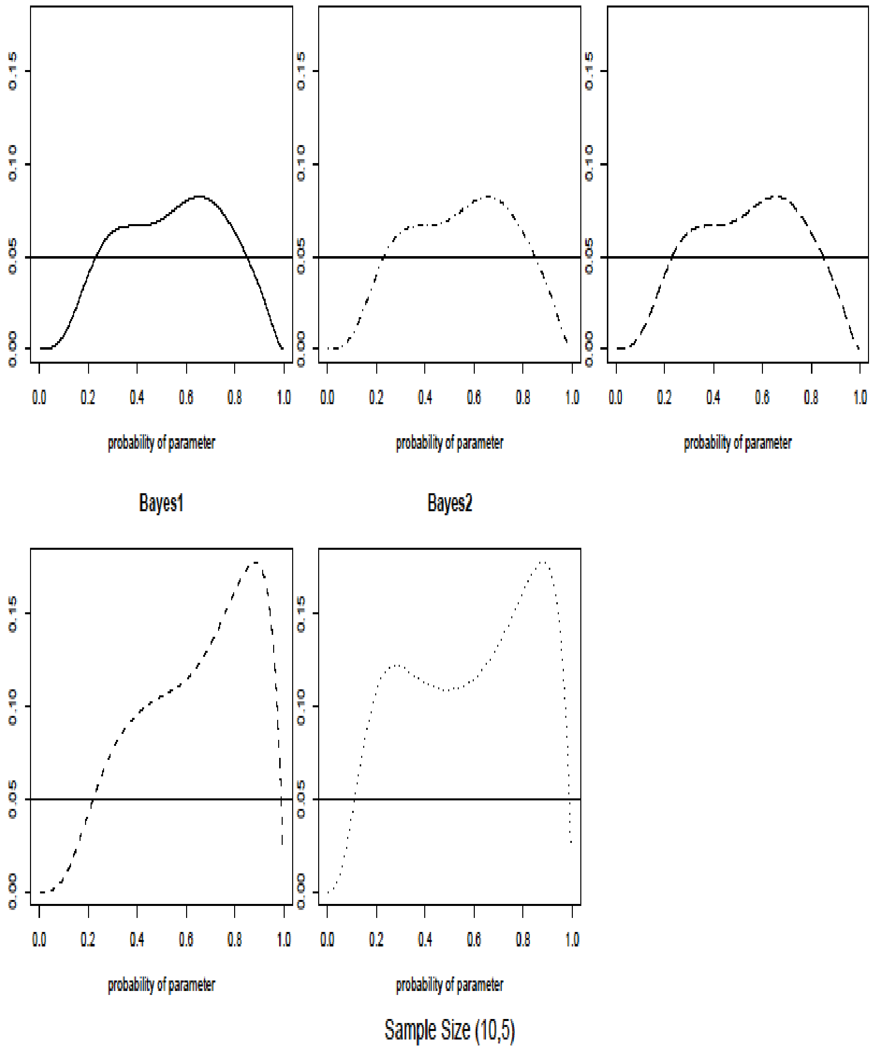
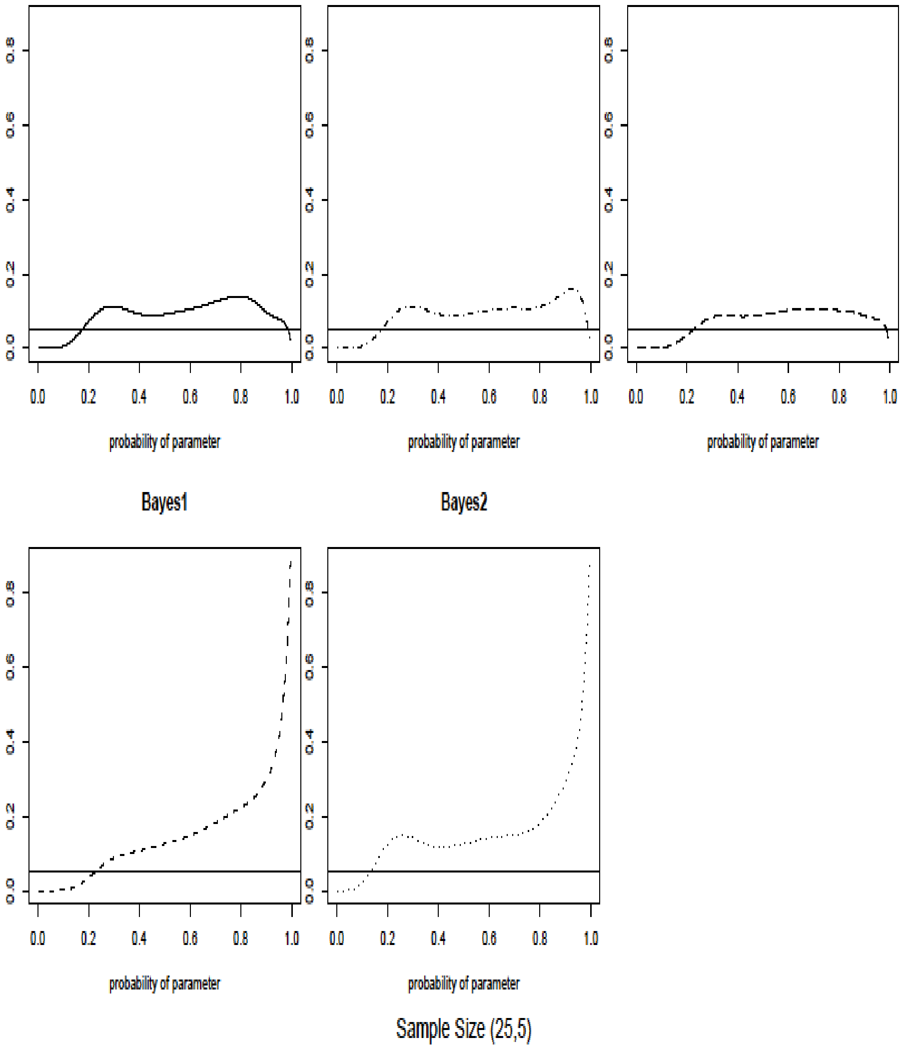
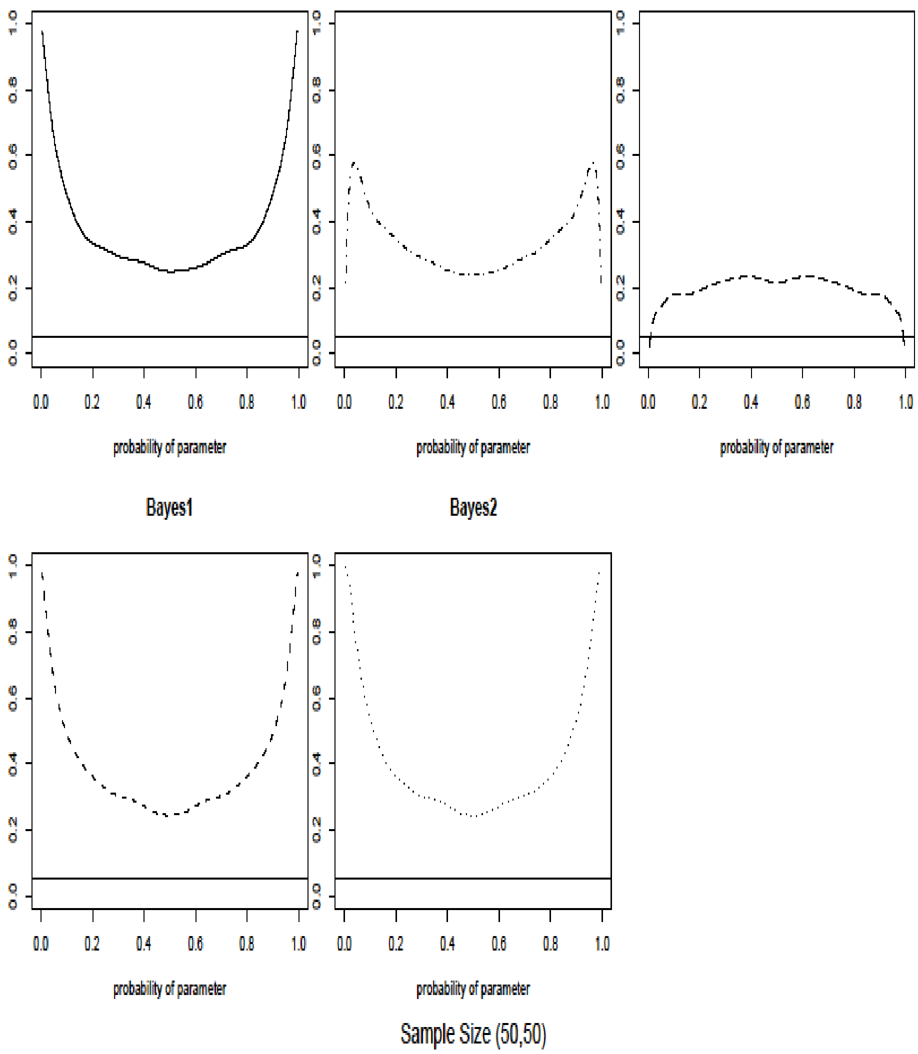

0.00
0.05
0.10
0.15
0.0
0.2
0.4
0.6
0.8
1.0
probability of parameter
probability of parameter
probability of parameter
Bayes1
Bayes2
probability of parameter
probability of parameter
Sample Size (10,5)

Figure 7:



probability of parameter
probability of parameter
probability of parameter
Bayes1
Bayes2
probability of parameter
probability of parameter
Sample Size (25,5)

Figure 8:



1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
probability of parameter
probability of parameter
probability of parameter
Bayes1
Bayes2
probability of parameter
probability of parameter
Sample Size (50,50)

Figure 9:

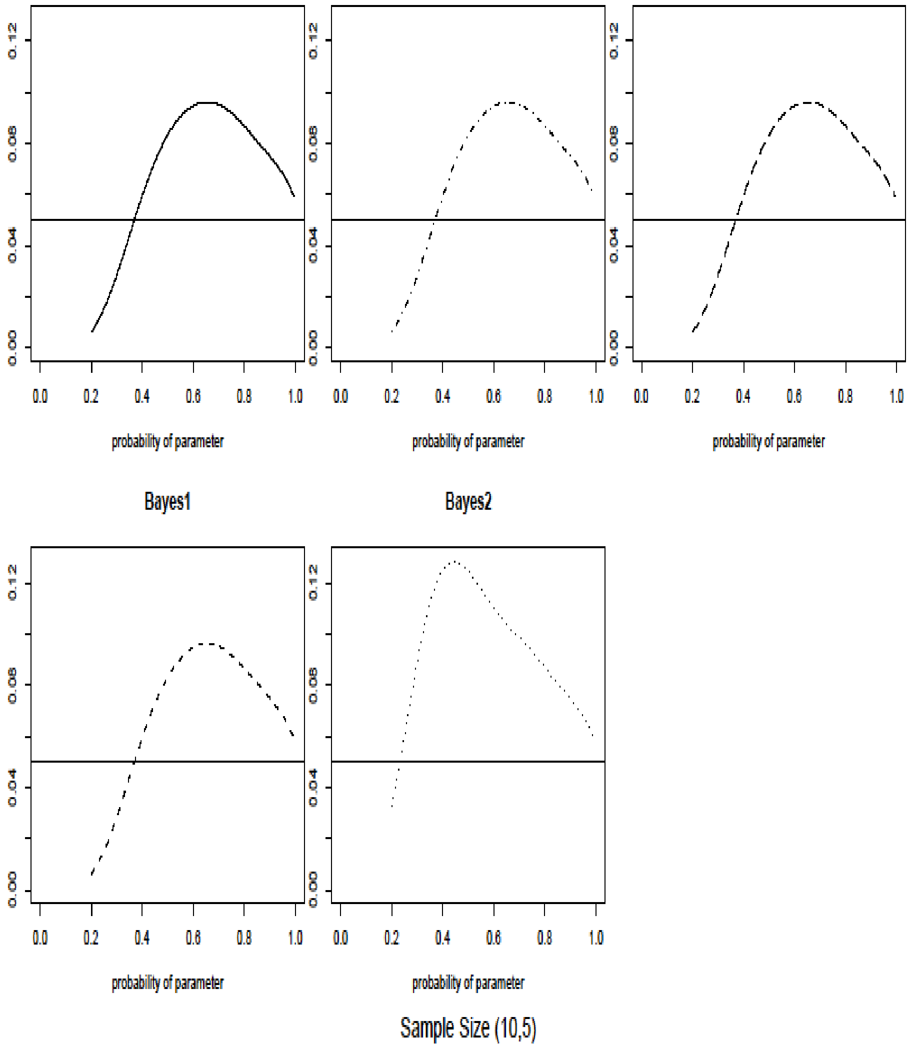

0.00
0.04
0.08
0.12
0.0
0.2
0.4
0.6
0.8
1.0
probability of parameter
Bayes1
Bayes2
Sample Size (10,5)

Figure 10:

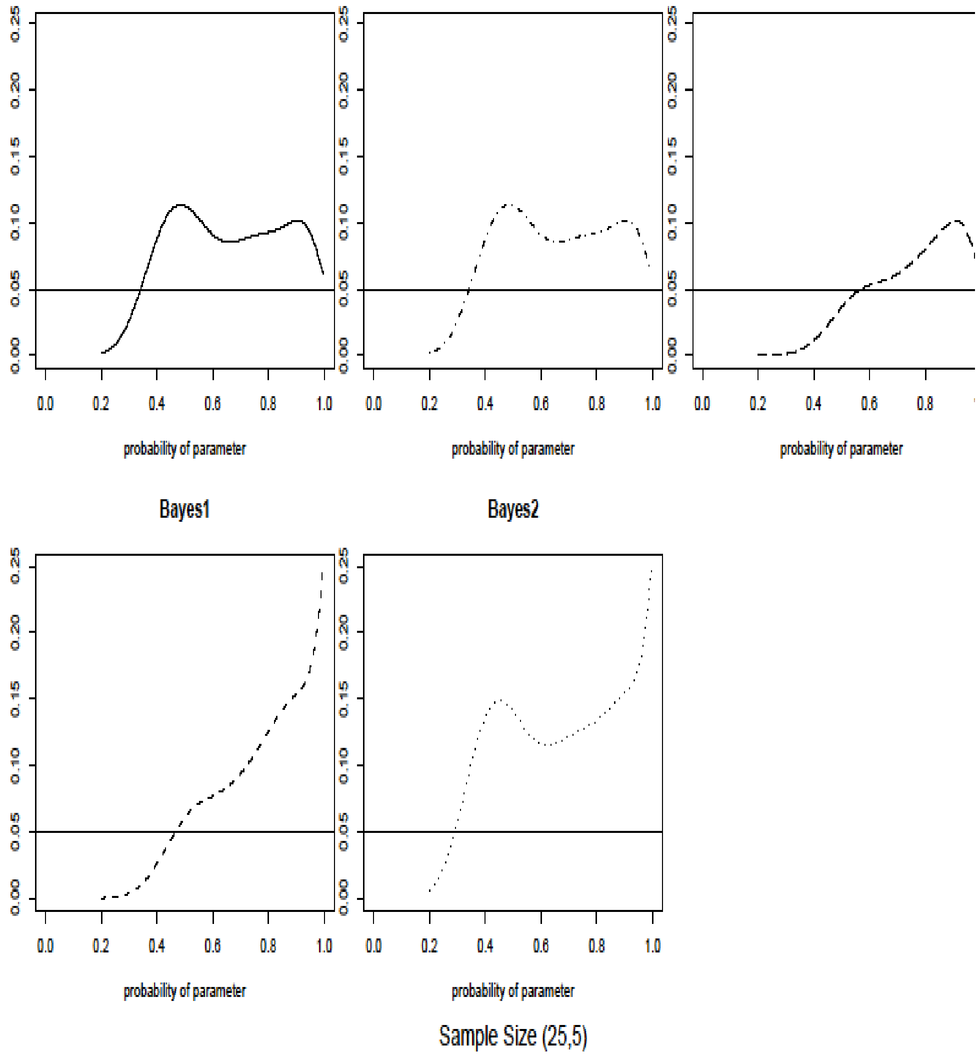

0.00
0.05
0.10
0.15
0.20
0.25
0.0
0.2
0.4
0.6
0.8
1.0
probability of parameter
Bayes1
Bayes2
Sample Size (25,5)

Figure 11:

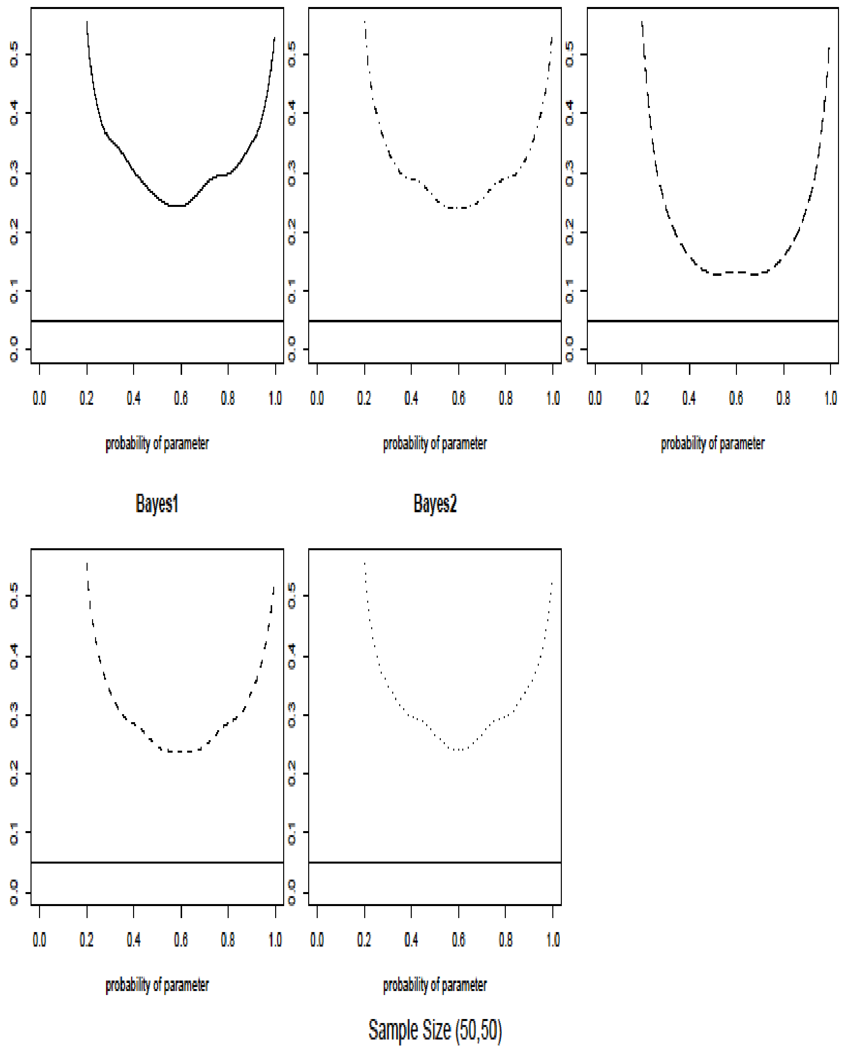

0.0
0.1
0.2
0.3
0.4
0.5
0.0
0.2
0.4
0.6
0.8
1.0
probability of parameter
probability of parameter
probability of parameter
Bayes1
Bayes2
probability of parameter
probability of parameter
Sample Size (50,50)

Figure 12:

$d = 0.1$

Sample Sizes

p*	q*	(10;5)	(20;10)	(40;20)	(80;40)	(160;80)	(320;160)
0.1	0.1	0.0168	0.0162	0.0159	0.0158	0.0157	0.0157
0.2	0.2	0.0052	0.0051	0.0050	0.0049	0.0049	0.0049
0.3	0.3	0.0030	0.0029	0.0029	0.0029	0.0028	0.0028
0.4	0.4	0.0023	0.0022	0.0022	0.0022	0.0022	0.0022
0.5	0.5	0.0021	0.0021	0.0020	0.0020	0.0020	0.0020
0.6	0.6	0.0023	0.0022	0.0022	0.0022	0.0022	0.0022
0.7	0.7	0.0030	0.0029	0.0029	0.0029	0.0028	0.0028
0.8	0.8	0.0052	0.0051	0.0050	0.0049	0.0049	0.0049
0.9	0.9	0.0169	0.0163	0.0159	0.0158	0.0157	0.0157

p*	q*	(10;10)	(25;25)	(50;50)	(100;100)	(200;200)
0.1	0.1	0.0210	0.0203	0.0201	0.0200	0.0197
0.2	0.2	0.0065	0.0063	0.0063	0.0062	0.0062
0.3	0.3	0.0038	0.0037	0.0036	0.0036	0.0036
0.4	0.4	0.0029	0.0028	0.0028	0.0028	0.0027
0.5	0.5	0.0027	0.0026	0.0026	0.0025	0.0025
0.6	0.6	0.0029	0.0028	0.0028	0.0028	0.0027
0.7	0.7	0.0038	0.0037	0.0036	0.0036	0.0036
0.8	0.8	0.0065	0.0063	0.0063	0.0062	0.0062
0.9	0.9	0.0210	0.0203	0.0201	0.0200	0.0197

p*	q*	(25;5)	(50;10)	(100;20)	(200;40)	(400;80)
0.1	0.1	0.0063	0.0062	0.0061	0.0061	0.0061
0.2	0.2	0.0020	0.0019	0.0019	0.0019	0.0019
0.3	0.3	0.0011	0.0011	0.0011	0.0011	0.0011
0.4	0.4	0.0009	0.0009	0.0008	0.0008	0.0008
0.5	0.5	0.0008	0.0008	0.0008	0.0008	0.0008
0.6	0.6	0.0009	0.0009	0.0008	0.0008	0.0008
0.7	0.7	0.0011	0.0011	0.0011	0.0011	0.0011
0.8	0.8	0.0020	0.0019	0.0019	0.0019	0.0019
0.9	0.9	0.0063	0.0062	0.0061	0.0061	0.0061

Figure 13:

d = −0.1

p*	q*	(10;5)	(20;10)	(40;20)	(80;40)	(160;80)	(320;160)
0.3	0.1	0.0401	0.0385	0.0377	0.0374	0.0372	0.0371
0.4	0.2	0.0255	0.0245	0.0241	0.0239	0.0238	0.0237
0.5	0.3	0.0207	0.0200	0.0196	0.0195	0.0194	0.0193
0.6	0.4	0.0201	0.0194	0.0191	0.0189	0.0188	0.0188
0.7	0.5	0.0231	0.0224	0.0220	0.0218	0.0217	0.0217
0.8	0.6	0.0329	0.0318	0.0313	0.0310	0.0309	0.0309
0.9	0.7	0.0671	0.0648	0.0636	0.0631	0.0628	0.0627

p*	q*	(10;10)	(25;25)	(50;50)	(100;100)	(200;200)
0.3	0.1	0.0629	0.0608	0.0602	0.0598	0.0597
0.4	0.2	0.0356	0.0346	0.0342	0.0341	0.0340
0.5	0.3	0.0270	0.0262	0.0260	0.0259	0.0258
0.6	0.4	0.0249	0.0241	0.0239	0.0238	0.0237
0.7	0.5	0.0270	0.0262	0.0260	0.0259	0.0258
0.8	0.6	0.0356	0.0346	0.0342	0.0341	0.0340
0.9	0.7	0.0629	0.0608	0.0602	0.0598	0.0597

p*	q*	(25;5)	(50;10)	(100;20)	(200;40)	(400;80)
0.3	0.1	0.0123	0.0120	0.0119	0.0118	0.0118
0.4	0.2	0.0086	0.0085	0.0084	0.0083	0.0083
0.5	0.3	0.0075	0.0073	0.0073	0.0072	0.0072
0.6	0.4	0.0077	0.0075	0.0075	0.0074	0.0074
0.7	0.5	0.0094	0.0092	0.0092	0.0091	0.0091
0.8	0.6	0.0146	0.0144	0.0142	0.0142	0.0141
0.9	0.7	0.0359	0.0352	0.0349	0.0347	0.0346

Figure 14: