Power Comparison of Exact Unconditional Tests for Comparing Two Binomial Proportions

Roger L. Berger

Department of Statistics

North Carolina State University

Raleigh, NC 27695-8203

June 29, 1994

Institute of Statistics Mimeo Series No. 2266

Abstract

The powers of six exact, unconditional tests for comparing two binomial proportions are studied. Total sample sizes range from 20 to 100 with balanced and imbalanced designs included. Some previously proposed tests are found to have poor power for imbalanced designs. Two new tests, confidence interval modifications of Boschloo's and Suissa and Shuster's tests, are found to have the best powers. Overall, the modified Boschloo test is recommended.

KEY WORDS: Confidence interval, Contingency table, Homogeneity test, Independence, P value, 2×2 table.

1 INTRODUCTION

In this article, we study the power functions of six exact unconditional tests for comparing two binomial proportions. We find that two new tests are superior to four tests that have been studied previously. Although no one test is uniformly better than all the rest in all situations, we find that Boschloo's (1970) test, with the confidence interval modification of Berger and Boos (1994), generally has the best power properties. The Suissa and Shuster (1985) test, using the pooled variance estimate and the confidence interval modification of Berger and Boos (1994), also has generally good power properties.

This article differs from most previous ones in focusing on the powers of the tests. Many previous articles have studied the sizes of various tests. For example, Upton (1982) compared the sizes of twenty-two tests but did no power comparisons. Storer and Kim (1990) thoroughly compared the sizes of seven tests but only briefly compared their powers at a few alternative parameter points. Haber (1986) presented a small power comparison. Our purpose is to provide a more thorough comparison of the powers of six tests.

The analysis in this article is *unconditional*. That is, the size and power comparisons we make are based on the binomial distributions of the model. There is continuing debate as to whether conditional or unconditional calculations are more relevant for these problems. Little (1989) and Greenland (1991) provide good recent summaries of the issues in this debate. The purpose of this paper is not to continue this debate. Rather, we agree with Greenland that in some (we believe most) situations the unconditional analysis is appropriate. This article is relevant to those situations.

All the tests we compare are *exact* tests. The sizes of the tests are computed using the exact binomial distributions, not normal or chi-squared approximations. We follow the standard Neyman-Pearson paradigm of restricting consideration to level-α tests and then comparing the powers of these tests. For a specified error probability α, all six tests we consider are level-α tests. Tests that are liberal, that sometimes have type-I error probabilities that are greater than α, are not considered. However, the tests do not have sizes exactly equal to the specified α. Because of the discrete nature of this data, equality can (usually) be achieved only with a randomized test. We do not think randomized tests are of any practical interest. So all the tests we consider are level-α; their sizes are not greater than the specified α.

2 MODEL AND TESTS TO BE COMPARED

Let X and Y be independent binomial random variables. The sample size for X is m and the success probability is p_1. The sample size for Y is n and the success probability is p_2. We will use

$$b(x; m, p_1) = C_x^m {p_1}^x (1-p_1)^{m-x}, \quad x = 0, \ldots, m,$$

to denote the binomial probability mass function of X, where $C_x^m = m!/x!/(m-x)!$ is the binomial coefficient. Similarly, $b(y; n, p_2)$ will denote the binomial probability mass function of Y. The sample space of (X, Y) will be denoted by $\mathcal{X} = \{0, \ldots, m\} \times \{0, \ldots, n\}$.

This kind of data is often displayed in a 2×2 contingency table as follows.

	yes	no	
Population 1	X	$m-X$	m
Population 2	Y	$n-Y$	n
	$R = X+Y$	$t-R$	$t = m+n$

In this table, upper case letters denote random variables and lower case letters denote known constants fixed by the sampling scheme. So, t is the total sample size, and R is the observed number of successes. Conditional inference is based on the conditional distribution of X and Y, given the observed marginal $R = r = x + y$.

We consider the problem of testing the null hypothesis $H_o : p_1 \geq p_2$ versus the alternative $H_a : p_1 < p_2$. We will compare the powers of the following six tests of this hypothesis. The tests will be denoted by the symbols F, B, B_c, P, P_c and U. The tests will be defined in terms of their p values; $p_A(x, y)$ will denote the p value for the test A. So, for example, $p_B(x, y)$ is the p value for the test B. For each test A, the level-α test is the test that rejects H_o if and only if $p_A(x, y) \leq \alpha$.

1. F – Fisher's (1935) Exact Test.

$$p_F(x,y) = \sum_{a=y}^{\min(n,x+y)} \frac{C^m_{x+y-a} C^n_a}{C^t_{x+y}}$$

This test is the classic *conditional* test of H_o versus H_a, based on the conditional hypergeometric distribution of Y given $X + Y = x + y$. But it often used as an unconditional test, without any regard for concepts like approximate ancillarity that are used to justify conditional inference. Used in this way it is an unconditional level-α test, albeit a rather conservative one. This test has been included in other unconditional test comparisons such as Upton (1982) and Storer and Kim (1990).

2. B – Boschloo's (1970) Test. McDonald, Davis and Milliken (1977) also proposed this test. Assume $p_1 = p_2 = p$.

$$p_B(x,y) = \sup_{0\le p\le 1} P_p(p_F(X,Y) \le p_F(x,y)) = \sup_{0\le p\le 1} \sum_{(a,b)\in R_B(x,y)} b(a;m,p)b(b;n,p),$$

where $R_B(x,y) = \{(a,b) : (a,b) \in \mathcal{X} \text{ and } p_F(a,b) \le p_F(x,y)\}$. Here, Fisher's p value p_F is used as the test statistic, not the p value. The p value is the maximum probability under H_o of observing a value of the test statistic equal to or more extreme than the value observed in the data. This is a standard definition of a p value, such as is found in Bickel and Doksum (1977, Section 5.2.B). B does not seem to be widely used, perhaps because of confusion about the role of p_F as a test statistic, not a p value.

3. B_c – Confidence Interval Modified B Test. Fix β, $0 \le \beta \le 1$.

$$\begin{aligned} p_{B_c}(x,y) &= \sup_{p\in C_\beta} P_p(p_F(X,Y) \le p_F(x,y)) + \beta \\ &= \left(\sup_{p\in C_\beta} \sum_{(a,b)\in R_B(x,y)} b(a;m,p)b(b;n,p) \right) + \beta, \end{aligned}$$

where $R_B(x,y)$ is the same as in the definition of B and C_β is a $100(1-\beta)\%$ confidence interval for p calculated from the data (x,y) and assuming $p_1 = p_2 = p$. B_c differs from B in that the supremum is taken over the confidence interval C_β rather than over the whole range $0 \le p \le 1$, and the error probability β is added to the supremum. If $\beta = 0$, B_c is the same as B. Berger and Boos (1994) showed that this modification of the usual definition of a p value yields a valid p value. That is, the test that rejects H_o if and only if $p_{B_c}(x,y) \leq \alpha$ is an unconditional level-α test. In our comparison, we use $\beta = .001$ (as suggested by Berger and Boos) and the binomial confidence interval in Casella and Berger (1990, Exercise 9.21). The confidence interval is based on $X + Y$, a binomial($m + n, p$) random variable if $p_1 = p_2 = p$.

4. P – Suissa and Shuster's (1985) Z-pooled Test. Haber (1986) also proposed this test. Define the Z-pooled statistic (score statistic) as

$$Z_p(x,y) = \frac{\hat{p}_2 - \hat{p}_1}{\sqrt{\frac{\hat{p}(1-\hat{p})}{m} + \frac{\hat{p}(1-\hat{p})}{n}}},$$

where $\hat{p}_1 = x/m$, $\hat{p}_2 = y/n$ and $\hat{p} = (x+y)/(m+n)$, the pooled estimate of $p_1 = p_2 = p$. Then

$$p_P(x,y) = \sup_{0 \leq p \leq 1} P_p(Z_p(X,Y) \geq Z_p(x,y)) = \sup_{0 \leq p \leq 1} \sum_{(a,b) \in R_P(x,y)} b(a;m,p)b(b;n,p),$$

where $R_P(x,y) = \{(a,b) : (a,b) \in \mathcal{X} \text{ and } Z_p(a,b) \geq Z_p(x,y)\}$. Thus, P is analogous to B, except using Z_p, rather than p_F, as the test statistic.

5. P_c – Confidence Interval Modified P Test. This test is a modification of P in the same way as B_c is a modification of B. The supremum is taken over a confidence interval C_β, and β is added to the error probability. In our comparison, we use the same $\beta = .001$ and confidence interval, as we used for B_c.

6. U – Suissa and Shuster's (1985) Z-unpooled Test. This test is defined in the same way as the P test, but using the Z-unpooled statistic (Wald statistic)

$$Z_u(x,y) = \frac{\hat{p}_2 - \hat{p}_1}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{m} + \frac{\hat{p}_2(1-\hat{p}_2)}{n}}}.$$

Suissa and Shuster showed that, if $m = n$, U is the same test as P. But these tests typically differ for unequal sample sizes.

7. U_c – Confidence Interval Modified U Test. A modification of U using a confidence interval, like B_c and P_c, could be defined. But because U has such poor power properties, as we shall see, U_c is not included in our comparison of tests.

3 COMPARISON OF TESTS

We will compare the power functions of the six tests defined in Section 2. For each test A, the rejection region of the level-α test is

$$R_A = \{(a, b) : (a, b) \in \mathcal{X} \text{ and } p_A(a, b) \le \alpha\}, \tag{1}$$

where p_A is the p value defined in Section 2. In this comparison, we use $\alpha = .10$ throughout. We have done similar comparisons using $\alpha = .05$ and $\alpha = .01$, and the results are qualitatively the same. (We are not advocating the blind use of fixed α levels. But to meaningfully compare the powers of two tests, we need to make the Type-I error probabilities approximately equal. In the Neyman-Pearson paradigm, this is done by first restricting to level-α tests, as we have done here.) The power function for the test A is

$$\beta_A(p_1, p_2) = P_{p_1,p_2}((X, Y) \in R_A) = \sum_{(a,b)\in R_A} b(a; m, p_1)b(b; n, p_2). \tag{2}$$

Because all six tests are level-α tests, $\beta_A(p_1, p_2) \le \alpha$ for all $p_1 \ge p_2$. We want to compare the power functions for $p_1 < p_2$ values.

There is only one pairwise comparison that is definitive for all sample sizes m and n and all α. B is always uniformly more powerful than F. F rejects H_o when $p_F(x, y) \le \alpha$. B rejects H_o when $p_F(x, y) \le \alpha^*$, where α^* is the largest value α' such that

$$P_{p_1,p_2}(p_F(X, Y) \le \alpha') \le \alpha, \text{ for all } p_1 \ge p_2. \tag{3}$$

Because $\alpha' = \alpha$ satisfies (3), $\alpha^* \ge \alpha$ and $R_F \subset R_B$. In all but trivial examples, R_F is a proper subset of R_B, and hence, $\beta_B(p_1, p_2)$ is strictly greater than $\beta_F(p_1, p_2)$ for all $0 < p_1 < p_2 < 1$. (In terms of p values, rather than fixed level-α testing, $p_F(x, y) > p_B(x, y)$ for every (x, y), unless $1 = p_F(x, y) = p_B(x, y)$. That is, F is more conservative.) Given this uniform dominance, which was pointed out by Boschloo in 1970, it is surprising to this author that F is still so frequently used to compare two binomial proportions.

Any other power comparison of two tests will depend on the α and sample sizes considered. As mentioned earlier, in this comparison we use $\alpha = .10$. We consider nine sample sizes, namely, (m, n) = (10,10), (13,7), (16,4), (25,25), (33,17), (40,10), (50,50), (65,35) and (80,10). The total sample sizes represented are t = 20, 50 and 100. These represent small to moderately large sample sizes. For each t, the degree of imbalance in m and n is approximately $n{:}m$ = 1:1, 2:1 and 4:1. One might question whether examining only nine potential sample sizes will give enough useful information. Storer and Kim (1990), for example, considered over 1,000 sample sizes. But they concentrated only on the sizes, not the powers, of the tests. Because a fairly clear picture of which tests are superior emerges from our study, we are confident that our conclusions are relevant to other sample sizes as well.

We compare the power functions of the tests in three different ways. First, we compare the sizes of the tests. That is, we compare $\beta_A(p, p)$, for $0 \leq p \leq 1$, for different tests A. Then we compare $\beta_A(p_1, p_2)$, for $0 < p_1 < p_2 < 1$ (on the alternative H_a), for different tests A, in two different ways.

3.1 Comparison of Sizes

For each of the nine values of (m, n), we determined the level-.10 rejection region for each of the six tests using (1). The graphs of $\beta_A(p, p)$ for the nine sample sizes appear in Figure 1. The legend showing which line corresponds to which test appears in the top left plot. These graphs are analogous to Upton's (1982) Figures 4 and 5 and Storer and Kim's (1990) Figure 2. Neither of these papers considered B_c or P_c. But some graphs of $\beta_A(p, p)$ for the other four tests can be found in these papers. A notation like "$P = U$" in the top left corner of a plot means that, for this sample size, the two tests are exactly the same, that is, they have the same rejection region.

For all six tests, the maximum probability of a type-I error occurs on the boundary, $p_1 = p_2 = p$. So the maximum of $\beta_A(p, p)$ over $0 \leq p \leq 1$ is the true size of the test. Ideally, the true size should be close to the nominal value of α = .10. F does not achieve this ideal. The true size of F ranges from .04 to .07 for the nine cases. This is a manifestation of what others have labeled the "extreme conservatism" of F, when used as an unconditional test. The other tests for which the true size is less than .09 are P and U for all three $m = n$ cases and P_c for $(m, n) = (10, 10)$. For the reasonably large values of $m = n = 50$, the poor size of P and U, namely .082, is surprising. But if the two sample points with the next smaller value of the test statistic are included in the rejection region, the size increases to .104. Even for these large sample sizes, there are still large gaps in the values of the attainable sizes for P and U.

If a test is unbiased, then it is similar, that is, $\beta_A(p, p) = \alpha$ for all $0 \leq p \leq 1$. In this discrete problem, the only similar tests are randomized. As Cox (1977) argued, in such a situation the statistically meaningful concept is approximate similarity. Thus we would like $\beta_A(p, p)$ to rise quickly to a value near $\alpha = .10$ and stay close to α for most values of p before dropping back to zero near $p = 1$. We see in Figure 1 that B_c, B and P_c generally exhibit this behavior for all nine sample sizes, But the other three tests do not. For F, $\beta_F(p, p)$ is fairly flat, but it does not reach a value close to α. For P, $\beta_P(p, p)$ is not flat for the $(40, 10)$ and $(80, 20)$ cases. It rises rapidly to a value near .09, but then falls quickly and is near .06 for most of the range of $.25 \leq p \leq 1$. Both Suissa and Shuster (1985) and Storer and Kim (1990) claim that $\beta_P(p, p)$ is relatively flat over the range $.05 \leq p \leq .95$. But we see that in the imbalanced $(40, 10)$ and $(80, 20)$ cases, this is not true. Berger and Boos (1994, Figure 1) show a graph for a two-sided testing problem where $\beta_P(p, p)$ is even more distinctly spiked. Also, as mentioned in the previous paragraph, $\beta_P(p, p)$ does not attain a value very near to α for the $m = n$ cases. U exhibits the worst behavior of all in terms of $\beta_U(p, p)$. In all six $m \neq n$ cases, $\beta_U(p, p) < .05$ for $0 \leq p \leq .5$ and rises to a peak only near $p = 1$. This behavior seems to get worse as t increases.

An explanation can be given for the relative flatness of the graphs for B and B_c. Let $p_{\max}$ be the maximum value of $p_F(a, b)$ over all $(a, b) \in R_B$. Then we can write

$$\begin{aligned} \beta_B(p, p) &= \sum_{r=0}^{m+n} P((X, Y) \in R_B \mid X + Y = r) P_p(X + Y = r) \\ &= \sum_{r=0}^{m+n} P(p_F(X, Y) \leq p_{\max} \mid X + Y = r) P_p(X + Y = r). \end{aligned} \tag{4}$$

For each r, $P((X, Y) \in R_B \mid X + Y = r) = P(p_F(X, Y) \leq p_{\max} \mid X + Y = r)$ is a hypergeometric probability that does not depend on p. In fact,

$$P(p_F(X, Y) \leq p_{\max} \mid X + Y = r) = p_F(r - b, b)$$

where b is the smallest value such that $p_F(r - b, b) \leq p_{\max}$. Thus, in (4), each $P(p_F(X, Y) \leq$

$p_{\max} \mid X+Y = r)$ is as close as possible to $p_{\max}$, without exceeding it. Equation (4) shows that $\beta_B(p,p)$ is a weighted average of $P(p_F(X,Y) \le p_{\max} \mid X+Y = r)$ values with weights $P_p(X+Y = r)$. It does not vary much with p because it is a weighted average of values that are nearly all close to $p_{\max}$. However, if $p_{\max} < 1$, then $P(p_F(X,Y) \le p_{\max} \mid X+Y = r)$ drops to zero for r values near 0 and $m+n$. So as $p \to 0$ or 1, most of the weight in (4) shifts to these values, and $\beta_B(p,p)$ decreases to zero. Because B_c is closely related to B, the values of $P((X,Y) \in R_{B_c} \mid X+Y = r)$ are also fairly constant in r. But for U, P and P_c, the values of $P((X,Y) \in R_A \mid X+Y = r)$ sometimes vary a good deal. Then the function $\beta_B(p,p)$ varies as the weights on these values shift with p.

3.2 Power Comparison Over H_a

The power function $\beta_A(p_1,p_2)$ is a function of two variables. The comparison of two such functions, to compare two tests, may be done in several ways. In this and the next subsection, we present two different comparisons.

Table 1 presents pairwise comparisons for each pair of the six tests. The information in Table 1 summarizes an overall comparison for all alternative values $0 < p_1 < p_2 < 1$. The row and column headings indicate which pair of tests is being compared. Within each pairwise comparison are nine entries, corresponding to the nine sample sizes considered. The nine positions in each 3×3 block correspond to the nine sample sizes in this pattern,

$$\begin{array}{ccc} (10,10) & (13,7) & (16,4) \\ (25,25) & (33,17) & (40,10) \\ (50,50) & (65,35) & (80,10) \end{array}.$$

The three rows correspond to the three total sample sizes, $t = 20, 50, 100$. The left column corresponds to equal sample sizes, and the sample sizes become more imbalanced as you move to the right. An entry of "=" indicates that the power functions of the two tests are exactly equal because the rejection regions for the two tests are exactly equal. An entry of "<" means that the column test is uniformly more powerful than the row test because the rejection region of the row test is a proper subset of the rejection region of the column test. An entry of ">" means that the row test is uniformly more powerful than the column test because the column test's rejection region is a proper subset of the row test's rejection region. In situations where none of the preceding uniform comparisons apply, the power functions for the two tests were computed (using (2)) on a grid of 5050 points, $p_2 = .005(.010).995$, $p_1 = .005(.010)p_2$. The proportion of these points at which the column test's power exceeds the row test's power is recorded in Table 1, as a percent. This is an approximation of the proportion of the total area of H_a at which the column test's power exceeds the row test's power.

As an example, consider B_c and P_c (first column, fourth row). For $(m, n) = (25, 25)$, $(50, 50)$ and $(16, 4)$, the two tests are exactly the same, with the same rejection regions and power functions. For $(m, n) = (10, 10)$, B_c is uniformly more powerful than P_c because R_{P_c} is a proper subset of R_{B_c}. But P_c is uniformly more powerful than B_c for $(m, n) = (65, 35)$. For the other four sample sizes, $\beta_{B_c}(p_1, p_2)$ and $\beta_{P_c}(p_1, p_2)$ cross with $\beta_{B_c}(p_1, p_2)$ exceeding $\beta_{P_c}(p_1, p_2)$ between 67% and 75% of the time on H_a.

Table 1 indicates that F and U have very poor power properties. B_c, P_c and B are all uniformly more powerful than F for all nine sample sizes. P is uniformly more powerful than F for eight of the nine cases and $\beta_P(p_1, p_2)$ exceeds $\beta_F(p_1, p_2)$ 68% of the time in the ninth case. For $m = n$, Suissa and Shuster (1985) showed that P and U are the same test. For all of the $m = n = 10(1)150$ designs with $\alpha = .05$, .025 and .01, Suissa and Shuster (1985) found that U and P are uniformly more powerful than F. But we see that this dominance does not continue for unequal sample sizes. F is more powerful than P on 32% of H_a for $(m, n) = (80, 20)$. And the power of U is often poor when compared to the power of F. $\beta_F(p_1, p_2)$ exceeds $\beta_U(p_1, p_2)$ a majority of the time for all the $m \neq n$ cases. The four tests B_c, B, P_c and P all are superior to U also. Generally, their power functions dominate $\beta_U(p_1, p_2)$ over all, or almost all, of H_a. Storer and Kim (1990), based only on size comparisons, also conclude that P is superior to U for unequal sample size designs.

The comparisons of B versus B_c and P versus P_c, that is the comparison of the "confidence interval modified" versus unmodified versions, is interesting. In both cases we see that for every sample size the confidence interval method is the same as or uniformly more powerful than the unmodified version. Thus the intuitive justification given in Berger and Boos (1994) seems to translate into genuine power improvements.

There will certainly be combinations of m, n and α for which $\beta_{P_c}(p_1, p_2)$ and $\beta_P(p_1, p_2)$ or $\beta_{B_c}(p_1, p_2)$ and $\beta_B(p_1, p_2)$ cross. But the evidence in Table 1 and Figure 1 is that, in general, the confidence interval modification offers an improvement in size and power.

Overall, from the first column of Table 1, we see that the B_c has good power properties when compared with the other five tests. For many sample sizes B_c is uniformly more powerful than the other tests. Only the test P_c for the single case $(m, n) = (65, 35)$ is uniformly more powerful than B_c. And when the power functions of B_c and another test cross, $\beta_{B_c}(p_1, p_2)$ is always larger on at least 67% of H_a; it is often larger on over 90% of H_a.

3.3 Magnitude of Power Differences

In the last section, comparisons regarding which power functions are largest were given. But there is no indication in Table 1 as to the magnitude of any power differences. To investigate this we present the graphs in Figures 2a, b and c. In Figure 2, we graph $\beta_A(p_1, p_2) - \beta_F(p_1, p_2)$ for each of the five tests $A = B_c$, B, P_c, P and U. All the power functions increase as $p_2 - p_1$ increases. This makes it difficult to judge differences in the graphs of the power functions. Differences in power are easier to see if we graph differences in power functions, as we have done. Note also that to compare two tests, for example B and P, $\beta_B(p_1, p_2) - \beta_P(p_1, p_2) = (\beta_B(p_1, p_2) - \beta_F(p_1, p_2)) - (\beta_P(p_1, p_2) - \beta_F(p_1, p_2))$. So a difference in two graphs in Figure 2 is equal to a difference in the power functions for the corresponding two tests.

We graph these functions along three "slices" in H_a defined by $p_1 + p_2 = 0.5$, 1.0 and 1.5. For $p_1 + p_2 = 0.5$ and 1.5, $p_2 - p_1$ ranges from 0.0 to 0.5. For $p_1 + p_2 = 1.0$, $p_2 - p_1$ ranges from 0.0 to 1.0. But for $p_2 - p_1 > 0.5$, almost all the power functions are near 1.0 and the differences are near 0.0. So for $p_1 + p_2 = 1.0$, we also only show the graph for $0 \leq p_2 - p_1 \leq 0.5$. Figures 2a, b and c correspond to $t = 20$, 50 and 100, respectively. Each column contains the three plots for the three slices for a particular (m, n) value. As in Figure 1, a notation like "$U = P$" in the top left of a plot indicates that two tests are exactly equal for this sample size. Note that the vertical scales differ in Figures 2a, b and c. In many cases Figure 2a is showing increases in power of about

.2 over F; Figure 2c is showing increases of about .1.

Figure 2 reiterates the poor power of U. In many cases, the function for U is negative, indicating that U has worse power than F. In many imbalanced cases, β_U is so small that the graph decreases off the bottom of the plot.

The importance in looking at both Figure 2 and Table 1 can be seen in a comparison of B_c and P_c. For $(m, n) = (65, 35)$, Table 1 shows that P_c is uniformly more powerful than B_c. But in Figure 2, the P_c and B_c curves are almost indistinguishable. The maximum difference is only about .006, occurring at $p_1+p_2 = 0.5$ and $p_2-p_1 = .18$. So P_c's power advantage appears minimal. But for $(m, n) = (10, 10)$, Table 1 indicates that B_c is uniformly more powerful than P_c. And in Figure 2 the maximum difference is seen to be .054 at $p_1 + p_2 = 1.0$ and $p_2 - p_1 = .30$, a much more substantial advantage.

In 26 out of the 27 plots in Figure 2, the function for B_c (solid line) is highest or very near to the highest. The one exception is the $(m, n) = (13, 7)$ and $p_1 + p_2 = 1.5$ slice where the P and P_c power exceeds the B_c power by as much as .065. For each of the three (m, n) values for which B_c is uniformly more powerful than B, there is at least one slice on which the power of B_c is noticeably larger than the power of B. The power of P_c is often approximately equal to the power of B_c. But there are several slices; $(m, n) = (40, 10)$ and $p_1+p_2 = 0.5$, $(m, n) = (33, 17)$ and $p_1+p_2 = 1.0$, $(m, n) = (13, 7)$ and $p_1+p_2 = 1.0$ and all of the $(m, n) = (10, 10)$ slices; where the power of B_c is substantially larger than the power of P_c. The overall impression from Figure 2 is that B_c has the consistently best power.

4 Conclusions and Generalizations

Based on these power comparisons, we conclude that B_c has the best power properties among these six tests. A user will not be far from optimal in using B_c for the entire range of sample sizes considered in this paper. Often, B_c is uniformly more powerful than other tests. When the power functions of B_c and another test cross, B_c always has the higher power over at least 67% of H_a. Usually the dominance is on over 90% of H_a. In only one case is one test (P_c) uniformly more powerful than B_c. And, in that case, Figure 2 indicates that the increase in power is negligible. Figure 1 shows that B_c is also most successful in being approximately similar on $p_1 = p_2$.

The two tests P_c and B are reasonable alternatives to B_c. B and B_c are exactly the same in six out of nine cases. But for larger sample sizes with imbalanced data, B_c is uniformly more powerful. P_c and B_c are exactly the same in three cases, and once P_c dominates B_c. But when their power functions cross, B_c's power is larger on a majority of H_a.

U has very poor power. For imbalanced sample sizes, its power is often much worse than F's. U should not be used for comparing two binomial proportions.

P's power is usually inferior to B_c's and P_c's. Sometimes, the power of P is little better than the power of F. If one wants to use the test statistic Z_p, because of ease of interpretation, then one should use P_c to obtain an increase in power. Santner and Duffy (1989) recommend P for this problem. But our results show that B_c and P_c are better tests.

Fisher's Exact Test F has poor unconditional power. We believe that in a comparison of binomial proportions, the user rarely is concerned with abstract notions such as approximate ancillarity. Rather the user wants high power. In such a situation, F is a poor test to use. Of course, if the true sampling distribution is hypergeometric (not independent binomials), as in Greenland's (1991) Example 1 or Kempthorne's (1979) "Mayo Clinic" example, then F is an appropriate test.

One reason for the continued use of F, despite its poor power, might be the impression that unconditional tests are computationally difficult. The maximization in computing the "sup" in the various p values must be done numerically. Even recent articles, such as Storer and Kim (1990) and Agresti (1992), label these computations as "computationally difficult" and "computationally intensive." But we disagree with these characterizations. A FORTRAN program, XUN2X2, is available that will calculate the p value for any of the tests discussed in this article in less than a second on a 486DX2/50 computer, if $t \leq 100$. Larger sample sizes might take a few seconds. The program will also perform unconditional tests for multinomial, rather than two independent binomials, 2×2 tables. These calculations might take a minute or two, but are still very doable. The program XUN2X2 may be obtained by sending the one line message "get exact from general" to statlib@lib.stat.cmu.edu.

Other test statistics, such as the likelihood ratio statistic and $\hat{p}_2 - \hat{p}_1$, might be used to construct an unconditional test. Santner and Duffy (1989, Exercises 5.11 and 5.12) list several choices. Our experience with Z_p and Z_u suggests that, if this is done, the confidence interval modified version of the test might well provide improved power over the unmodified version.

In this article, we have presented comparisons for only one-sided tests and $\alpha = .10$. We have performed similar, but less extensive, calculations for other α values and two-sided tests. The results we obtained are qualitatively the same. We make the same recommendations for these problems, too.

REFERENCES

Agresti, A. (1992), "A Survey of Exact Inference for Contingency Tables" (with discussion), *Statistical Science*, 7, 131-177.

Berger, R. L., and Boos, D. D. (1994), "P Values Maximized Over a Confidence Set for the Nuisance Parameter," *Journal of the American Statistical Association*, 89. To appear.

Bickel, P. J., and Doksum, K. A. (1977), *Mathematical Statistics: Basic Ideas and Selected Topics*, San Francisco: Holden-Day.

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.

Casella, G., and Berger, R. L. (1990), *Statistical Inference*, Belmont, CA: Wadsworth.

Cox, D. R. (1977), "The Role of Significance Tests," *Scandinavian Journal of Statistics*, 4, 49-70.

Fisher, R. A. (1935), *The Design of Experiments*, Edinburgh: Oliver & Boyd.

Greenland, S. (1991), "On the Logical Justification of Conditional Tests for Two-by-Two

Contingency Tables," *The American Statistician*, 45, 248-251.

Haber, M. (1986), "An Exact Unconditional Test for the 2×2 Comparative Trial," *Psychological Bulletin*, 99, 129-132.

Kempthorne, O. (1979), "In Dispraise of the Exact Test: Reactions," *Journal of Statistical Planning and Inference*, 3, 199-213.

Little, R. J. A. (1989), "On Testing the Equality of Two Independent Binomial Proportions," *The American Statistician*, 43, 283-288.

McDonald, L. L., Davis, B. M., and Milliken, G. A. (1977), "A Nonrandomized Unconditional Test for Comparing Two Proportions in 2×2 Contingency Tables," *Technometrics*, 19, 145-157.

Santner, T. J., and Duffy, D. E. (1989), *The Statistical Analysis of Discrete Data*, New York: Springer.

Suissa, S. and Shuster, J. J. (1985), "Exact Unconditional Sample Sizes for the 2×2 Binomial Trial, *Journal of the Royal Statistical Society*, Ser. A, 148, 317-327.

Storer, B. E. and Kim, C. (1990), "Exact Properties of Some Exact Test Statistics for Comparing Two Binomial Proportions," *Journal of the American Statistical Association*, 85, 146-155.

Upton, G. J. G. (1982), "A Comparison of Alternative Tests for the 2×2 Comparative Trial," *Journal of the Royal Statistical Society*, Ser. A, 145, 86-105.