More Powerful Tests from Confidence Interval P Values

Roger L. Berger
Department of Statistics
North Carolina State University
Raleigh, NC 27695-8203

December 20, 1995
Institute of Statistics Mimeo Series No. 2281

Abstract

In this article, the problem of comparing two independent binomial populations is considered. It is shown that the test based on the confidence interval p value of Berger and Boos (1994) often is uniformly more powerful than the standard unconditional test. This test also requires less computational time.

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

1 INTRODUCTION

The problem of comparing two binomial proportions has been considered for many years. The most commonly used test is Fisher's Exact Test (Fisher, 1935), a conditional test. Barnard(1945, 1947) proposed an unconditional test for this problem. Although unconditional tests are usually more powerful than conditional tests, they are computationally much more complex. But recent advances in computing have made unconditional tests practical, and they are beginning to appear in statistical software packages such as *StatXact 3 for Windows*. In this article it is shown that unconditional tests based on the confidence interval p value of Berger and Boos (1994) are often uniformly more powerful than the standard unconditional tests.

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. The binomial probability mass function of X will be denoted by

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

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 $\mathcal{X} = \{0, \ldots, m\} \times \{0, \ldots, n\}$. $\mathcal{X}$ contains $(m+1)(n+1)$ points.

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 $R = r = x + y$.

Consider the problem of testing

$$H_o : p_1 = p_2 \quad \text{versus} \quad H_a : p_1 < p_2. \tag{1}$$

Exact tests for this problem will be considered. The sizes of the tests are computed using the exact binomial distributions, not normal or chi-squared approximations. The standard Neyman-Pearson paradigm of restricting consideration to level-α tests and then comparing the powers of these tests will be followed. For a specified error probability α, all tests considered 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. Because randomized tests are not of any practical interest, this article considers only nonrandomized tests.

The analysis in this article is *unconditional*. That is, the size and power comparisons 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 article is not to continue this debate. Rather, suffice it to say that this article is relevant to those situations in which the unconditional analysis is appropriate.

2 USUAL UNCONDITIONAL TEST

Barnard (1945, 1947) first proposed an unconditional test for this problem. Because of the computational difficulty of unconditional tests, they were not widely used until recently. NOw, computing technology makes the use of unconditional tests feasible.

A commonly used unconditional test is the the Z test proposed by Suissa and Shuster (1985) and Haber (1986). Define the Z-pooled statistic (score statistic) as

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

where $\hat{p}_1 = x/m$, $\hat{p}_2 = y/n$, and $\hat{p} = (x+y)/(m+n)$, the pooled estimate of $p_1 = p_2 = p$ under H_o. Then, the p value for testing (1), using the test statistic Z, is

(2) $$p_Z(x,y) = \sup_{0\le p\le 1} P_p\left(Z(X,Y) \ge Z(x,y)\right) = \sup_{0\le p\le 1} \sum_{(a,b)\in R_Z(x,y)} b(a;m,p)b(b;n,p),$$

where $R_Z(x,y) = \{(a,b) : (a,b) \in \mathcal{X} \text{ and } Z(a,b) \ge Z(x,y)\}$. 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). Rejection of H_o if and only if $p_Z \le \alpha$ defines a level-α test of (1). The calculation of the supremum in (2) must be done numerically. Typically there is no simple formula for this value. This numeric maximization has been the cause of the computational difficulty of unconditional tests.

3 CONFIDENCE INTERVAL P VALUE

Berger and Boos (1994) proposed a new method of computing a p value. In the problem of comparing two binomial proportions, if H_o is true, $p = p_1 = p_2$ is a nuisance parameter. Let $C_\beta(x,y)$ denote a $100(1-\beta)\%$ confidence interval for p calculated from the data (x,y) and assuming $p_1 = p_2 = p$. The confidence interval used in this article is the Clopper and Pearson (1934) interval based on $X+Y$, a binomial($m+n, p$) random variable if $p_1 = p_2 = p$. This interval is easily computed from the formula

$$\frac{a}{a+(b-a+1)F_{2(b-a+1),2a,\beta/2}} \leq p \leq \frac{(a+1)F_{2(a+1),2(b-a),\beta/2}}{b-a+(a+1)F_{2(a+1),2(b-a),\beta/2}} \tag{3}$$

where $a = x+y$, $b = m+n$, and $F_{\nu,\eta,\beta/2}$ is the upper $100(\beta/2)$ percentile of an F distribution with ν and η degrees of freedom.

The confidence interval p value, based on the statistic Z is defined by

$$\begin{aligned} p_C(x,y) &= \sup_{p \in C_\beta(x,y)} P_p\left(Z(X,Y) \geq Z(x,y)\right) + \beta \\ &= \left(\sup_{p \in C_\beta(x,y)} \sum_{(a,b) \in R_Z(x,y)} b(a;m,p)b(b;n,p) \right) + \beta, \end{aligned}$$

where $R_Z(x,y)$ is the same as in the definition of p_Z. p_C differs from p_Z in that the supremum is taken over the confidence interval $C_\beta(x,y)$ rather than over the whole range $0 \leq p \leq 1$, and the error probability β is added to the supremum. If $\beta = 0$, p_C is the same as p_Z. 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 $p_C(X,Y) \leq \alpha$ is an unconditional level-α test. The error probability is specified by the experimenter. Different values of β yield different p values and tests. In this article, $\beta = .001$ is used (as suggested by Berger and Boos).

Berger and Boos proposed the confidence interval based p value for two reasons. The first is computational. In both p_Z and p_C, the function to be maximized is the same. The maximization over the smaller set, C_β, can be much simpler. The second is statistical. Having observed the data, we should be able to estimate p and should not need to consider values of p that are completely unsupported by the data. In p_C, only those "plausible" values that are in the confidence interval are considered.

This article points out that the confidence interval p value can have a third advantage. It can produce tests with higher power than the usual p value. And, remember, this is achieved with less computational effort.

4 EXAMPLE

To see the improvement that can be obtained by using p_C rather than p_Z, consider constructing a level-α test with $\alpha = .10$ for sample sizes $m = 33$ and $n = 17$.

An enumeration of $p_Z(x, y)$ for all the sample points in $\mathcal{X}$ shows that the point $(x, y) = (23, 15)$ has the largest value of $p_Z(x, y)$ that is less than or equal to $\alpha = .10$. $Z(23, 15) = 1.454$ and $p_Z(23, 15) = .0823$. The sample point with the next smaller value of Z is $(x, y) = (0, 1)$ with $Z(0, 1) = 1.407$ and $p_Z(0, 1) = .1548$. So $(0, 1)$, or any other sample point with a smaller value of Z, is not in the level $\alpha = .10$ rejection region based on p_Z.

Let the *p value function* for a sample point (x, y) be defined by

(4) $$\alpha^Z(p; x, y) = \sum_{(a,b) \in R_Z(x,y)} b(a; m, p) b(b; n, p).$$

The p value function is the function that is maximized in calculating p_Z and p_C. In Figure 1, $\alpha^Z(p; 23, 15)$ (solid line) and $\alpha^Z(p; 0, 1)$ (long dashed line) are shown. $\alpha^Z(p; 23, 15) < .10 = \alpha$ for all values of p. Its maximum value is $p_Z(23, 15) = .0823$ which occurs at $p = .524$. Because $p_Z(23, 15)$ is the largest p value not exceeding $\alpha = .10$, .0823 is also the actual size of the level $\alpha = .10$ test constructed using p_Z. The addition of the single sample point $(0, 1)$ to the rejection region causes a large spike to appear in $\alpha^Z(p; 0, 1)$. The maximum of $\alpha^Z(p; 0, 1)$ is $p_Z(0, 1) = .1548$ which occurs at $p = .026$. It is not unusual for the addition of a single sample point with $x + y$ close to 0 or $m + n$ to create a spike like this.

In Figure 2, the sample points in $\mathcal{X}$ with $p_Z(x, y) \leq .10$ are marked by $\square$'s. These are the elements of the level $\alpha = .10$ rejection region defined by p_Z. Because the actual size of this test is only .0823, not too close to $\alpha = .10$, it seems possible that some more points, some of the points marked by $\times$'s and $+$'s, for example, might be added to this rejection region and the resulting test could still be level $\alpha = .10$. The points marked by $+$'s and $\times$'s are the points that satisfy the "convexity" property of Barnard (1947). It would take a great deal of computation to try each point individually, then try pairs or triples of points, to determine points that could be added. But the use of the confidence interval p value easily identifies some points that can be added.

Consider $(x, y) = (21, 14)$ with $Z(21, 14) = 1.368$. There are four sample points with $Z(21, 14) < Z(x, y) < Z(23, 15)$, namely, $Z(0, 1) = 1.407$, $Z(26, 16) = 1.401$, $Z(9, 8) = 1.399$, and $Z(3, 4) = 1.394$. The p value function $\alpha^Z(p; 21, 14)$ is shown in Figure 3. The maximum of this function, .1549, is greater than .10 because $Z(0, 1) > Z(21, 14)$. But $p_C(21, 14)$ is calculated by maximizing over the 99.9% confidence interval for p, calcu-

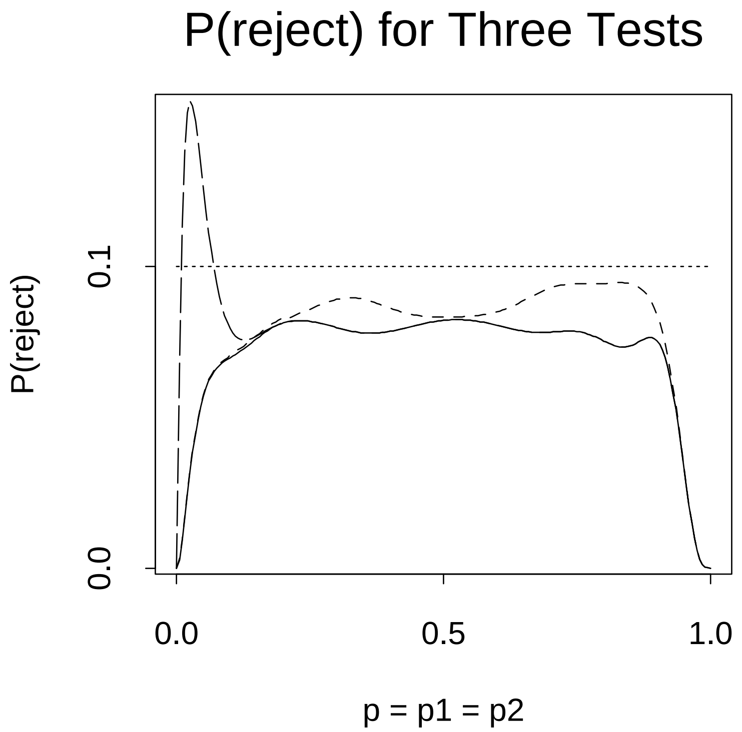

Figure 1: $P(\text{reject})$ for three rejection regions defined by $\{Z \geq Z(23, 15)\}$ (solid line), $\{Z \geq Z(0, 1)\}$ (long dashes), and $\{p_C \leq .10\}$ (short dashes).

lated from the data $(x, y) = (21, 14)$. Using (3), this confidence interval is $[.459, .881]$. These confidence limits are shown in Figure 3. The maximum of $\alpha^Z(p; 21, 14)$ over this interval is .0946 and this maximum occurs at $p = .830$. Therefore, $p_C(21, 14) = .0946 + \beta = .0946 + .0010 = .0956$. Because, $p_C(21, 14) \leq .10$, $(21, 14)$ is in the level $\alpha = .10$ rejection region defined by p_C. In addition, two other points are in the level $\alpha = .10$ rejection region defined by p_C. These are $(26, 16)$ with $p_C(26, 16) = .0949$ and $(9, 8)$ with $p_C(9, 8) = .0906$. These three added points are the points marked with $\times$'s in Figure 2.

Because the rejection region of the level $\alpha = .10$ test defined by p_Z is a proper subset of the rejection region of the level $\alpha = .10$ test defined by p_C, the test defined by p_C is uniformly more powerful.

The probability of the level $\alpha = .10$ rejection region defined by p_C, that is, the rejection region consisting of all the $\square$'s and $\times$'s in Figure 2, as a function of $p = p_1 = p_2$

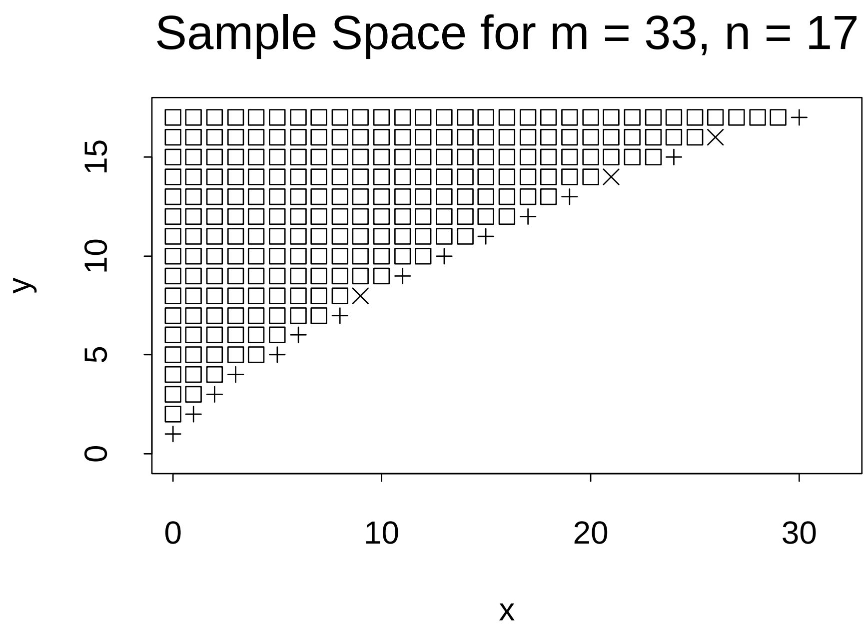

Figure 2: □'s are sample points in level $\alpha = .10$ test defined by p_Z. +'s and $\times$'s are sample points that might be added. $\times$'s are three points added by test defined by p_C.

is graphed in Figure 1 with a short dashed line. This probability is less than .10 for all values of p, because this is a level $\alpha = .10$ test. But this probability is much closer to .10 than the probability of the rejection region for the Suissa and Shuster test defined by p_Z. The actual size of the p_C test is .0946, the maximum of this function.

5 CONSISTENCY OF IMPROVEMENT

In the previous section it was shown that, for $\alpha = .10$ and $(m, n) = (33, 17)$, the confidence interval p value defines a uniformly more powerful, level-α test than the usual unconditional p value. The question arises as to the generality of this phenomenon. To investigate this we enumerated the level-α rejection regions defined by p_Z and p_C for $\alpha = .10$, .05 and .01 for each of nine different sample sizes, $(m, n) = (10, 10)$, $(13, 7)$, $(16, 4)$, $(25, 25)$, $(33, 17)$, $(40, 10)$, $(50, 50)$, $(65, 35)$, and $(80, 20)$. These sample sizes were chosen to represent small to moderately large total sample sizes and balanced to 4:1 unbalanced designs.

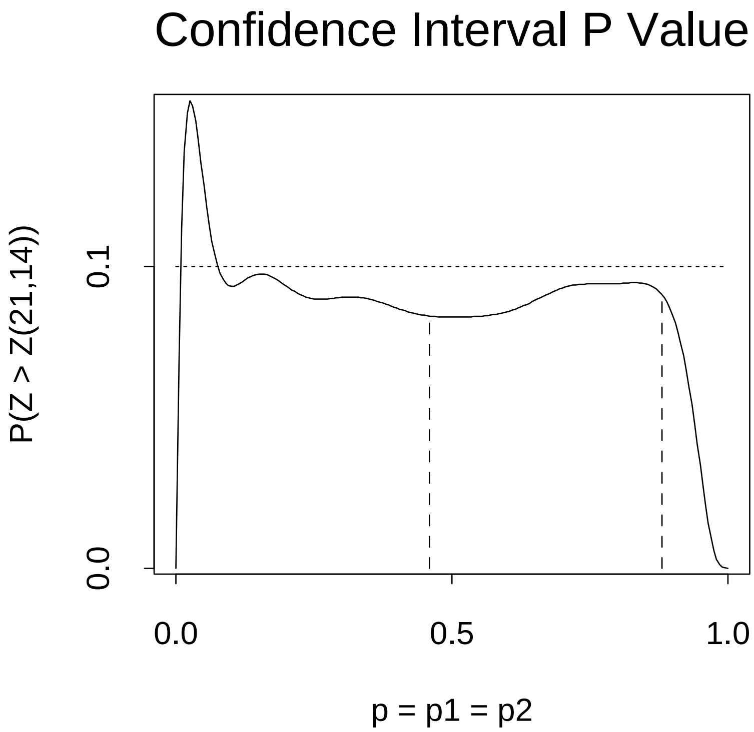

Figure 3: Confidence interval p value $p_C(21, 14) = .0956$ is the maximum of $\alpha^Z(p; 21, 14) + .001$ over the 99.9% confidence interval $[.459, .881]$. $\alpha^Z(p; 21, 14)$ is the function shown and the confidence interval is marked by vertical lines.

In 15 out of the 27 cases, the rejection region defined by p_Z is a proper subset of the rejection region defined by p_C. So the confidence interval p value defines a uniformly more powerful, level-α test. In another 9 out of the 27 cases, the rejection regions defined by the two p values are exactly the same. In one case, $\alpha = .01$ and $(m, n) = (50, 50)$, neither rejection region contained the other and the power functions of the two tests crossed. In the remaining two cases. $\alpha = .01$ and $(m, n) = (13, 7)$ and $(25, 25)$, the rejection region defined by p_C is a proper subset of the rejection region defined by p_Z, and p_Z defines a uniformly more powerful test. The power functions for the nine tests with $\alpha = .10$ are described more fully by Berger (1994).

Thus, in most cases, the confidence interval p value defines a test that is the same or uniformly more powerful than the test defined by the usual unconditional p value. Only infrequently will the test defined by p_C be inferior to the test defined by p_Z. And in all cases, the computation required for p_C is less than that required for p_Z.

The reason that the rejection region defined by p_C usually contains the rejection region defined by p_Z is the following fact. If there is no sample point in $\mathcal{X}$ such that $\alpha - \beta < p_Z(x, y) \leq \alpha$, then every sample point with $p_Z(x, y) \leq \alpha$ also satisfies $p_C(x, y) \leq \alpha$. That is, every sample point in the level-α rejection region defined by p_Z is also in the level-α rejection region defined by p_C. This fact is true because, if $p_Z(x, y) \leq \alpha$, then $p_Z(x, y) \leq \alpha - \beta$, and, hence,

$$\begin{aligned} p_C(x, y) &= \sup_{p \in C_\beta(x,y)} P_p\left(Z(X, Y) \geq Z(x, y)\right) + \beta \\ &\leq \sup_{0 \leq p \leq 1} P_p\left(Z(X, Y) \geq Z(x, y)\right) + \beta \\ &= p_Z(x, y) + \beta \\ &\leq (\alpha - \beta) + \beta = \alpha. \end{aligned}$$

When β is small compared to α, as with the $\alpha = .001$ recommended by Berger and Boos (1994) that is used in this article, it often happens that there is no sample point with $\alpha - \beta < p_Z(x, y) \leq \alpha$. In such cases the test defined by p_C will be at least as powerful as the test defined by p_Z. Note, this property applies in general to confidence interval p values, not just this problem and this test statistic Z.

6 OTHER TEST STATISTICS

Other statistics besides $Z(x, y)$, such as the likelihood ratio test statistic and $\hat{p}_2 - \hat{p}_1$, can be used to test (1). Santner and Duffy (1989, Exercises 5.11 and 5.12), Haber (1987), and Martín and Silva (1994) list several possible statistics. The experience with Z suggests that if another statistic is used, the confidence interval p value might provide improved power over the usual unconditional p value.

The power comparisons of Haber (1987) and Martín and Silva (1994) suggest that the two statistics $Z(x, y)$ and

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

produce tests with the highest power. The statistic $B(x, y)$ was first proposed by Boschloo (1970) and McDonald, Davis, and Milliken (1977). $B(x, y)$ is the conditional p value of

Fisher's Exact Test (Fisher, 1935). Here, $B(x,y)$ is not used as a p value, but, rather, as a statistic to order the sample points. Small values of $B(x,y)$ give evidence for H_a so the unconditional p value based on $B(x,y)$ is

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

where $R_B(x,y) = \{(a,b) : (a,b) \in \mathcal{X} \text{ and } B(a,b) \le B(x,y)\}$. The confidence interval p value based on B is defined as in (4), namely,

$$p_{CB}(x,y) = \left(\sup_{p\in C_\beta(x,y)} \sum_{(a,b)\in R_B(x,y)} b(a;m,p)b(b;n,p)\right) + \beta.$$

Berger (1994) found that the p value function $\alpha^B(p;x,y)$ tends to be flatter than $\alpha^Z(p;x,y)$, especially for unequal sample sizes. This agrees with Martín and Silva's (1994) finding that the unconditional test based on B usually has higher power than the test based on Z, especially when $m \neq n$. So there is less room for improvement of Boschloo's test. But, Berger (1994) found that, as with p_Z and p_C, the confidence interval p value, p_{CB}, usually defined a test that was the same or uniformly more powerful than the test defined by p_B.

In comparing the tests based on the two confidence interval p values, Berger (1994) did not find a clear preference. Usually the power functions of these two tests crossed with one test having higher power for some parameter values and the other having higher power for other parameter values. Usually, the power function defined by p_{CB} was higher on a majority of the parameter space.

In their power comparison of tests for (1), Martín and Silva (1994) considered two computationally intensive tests they called M and M'. M is the test proposed by Barnard (1945, 1947), and M' is a simplified version of M. Both methods involve construction of a rejection region by adding one sample point at a time, with a good deal of computation required to determine which point is added next. Martín and Silva report that M' and M require about 10 and 85 times the computation time required by p_Z or p_B, respectively. But, M and M' do provide some improvement in power. In this article it has been shown that confidence interval p values provide an improvement in power over p_Z or p_B, but with *less* computation. It remains to be determined if the improvement in power provided by p_C or p_{CB} is comparable to the improvement provided by M' or M.

7 CONCLUSIONS

Confidence interval p values can improve the power of standard unconditional tests for comparing two binomial populations. They also require less computational effort. Thus, they offer a promising new method for the analysis of 2×2 tables.

Similar, but less extensive, comparisons have have been made for two-sided tests. The results are qualitatively the same. The confidence interval p value often defines a more powerful test than the standard p value.

XUN2X2 is a FORTRAN program that will compute the standard and confidence interval p values discussed in this article. The program will also perform unconditional tests for multinomial, rather than two independent binomials, 2×2 tables. XUN2X2 may be obtained by sending the one line message "get exact from general" to statlib@lib.stat.cmu.edu.

References

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

Barnard, G. A. (1947). Significance tests for 2×2 tables. *Biometrika*, 34:123–138.

Berger, R. L. (1994). Power comparison of exact unconditional tests for comparing two binomial proportions. Technical Report 2266, North Carolina State University Statistics Department.

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*, 89:1012–1016.

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

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.

Clopper, C. J. and Pearson, E. S. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26:404–413.

Fisher, R. A. (1935). The logic of inductive inferences. *Journal of the Royal Statistical Society, Series A*, 98:39–54.

Greenland, S. (1991). On the logical justification of conditional tests for two-by-two contingency tables. *American Statistician*, 45:248–251.

Haber, M. (1986). An exact unconditional test for the 2×2 comparative trial. *Psychological Bulletin*, 99:129–132.

Haber, M. (1987). A comparison of some conditional and unconditional exact tests for 2×2 contingency tables. *Communications in Statistics–Simulation and Computation*, 16:999–1013.

Little, R. J. A. (1989). Testing the equality of two independent binomial proportions. *American Statistician*, 43:283–288.

Martín Andrés, A. and Silva Mato, A. (1994). Choosing the optimal unconditioned test for comparing two independent proportions. *Computational Statistics and Data Analysis*, 17:555–574.

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.

Mehta, C. and Patel, N. (1995). *StatXact 3 for Windows: User Guide*. Cytel Software, Cambridge, MA.

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

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.