A MORE POWERFUL EXACT TEST FOR A PRACTICAL DIFFERENCE BETWEEN BINOMIAL PROPORTIONS

Justin S. Kopit, Roger L. Berger, North Carolina State University
Justin S. Kopit, Department of Statistics, N.C. State University, Raleigh, NC 27695

Key Words: Superiority, Exact test, Confidence interval p value.

Abstract:

In this article, we construct a new exact test for whether the difference between two binomial proportions exceeds a specified quantity $\gamma, \gamma \geq 0$. Our test, which is based on confidence region p values, is often uniformly more powerful than the commonly used exact test, which is based on a Wald statistic.

1. Introduction

Often, researchers want to know whether the difference between two binomial proportions exceeds a specified quantity γ, $\gamma \geq 0$. For instance, a health insurer might not cover an expensive new treatment unless its success rate exceeds that of the current one by γ units or more. (Alternatively, researchers might want to find out if the ratio of proportions, i.e., the relative risk, exceeds a specified quantity.) When researchers are interested in whether one proportion exceeds another by at least γ units, $\gamma \geq 0$, they are said to be testing for superiority or for a practical improvement.

Let $X \sim bin(p_1, n_1)$ and $Y \sim bin(p_2, n_2)$ where p_1 is the success rate for treatment 1, the new or experimental treatment, and p_2 is the success rate for treatment 2, the reference treatment. Denote the sample space of (X, Y) by $\chi = \{0 \ldots n_1\} \times \{0 \ldots n_2\}$. Typically, if we want to know whether $p_1 - p_2 > \gamma$, we test

$$H_0 : p_1 - p_2 \leq \gamma \quad \text{versus} \quad H_a : p_1 - p_2 > \gamma. \tag{1}$$

(We want overwhelming evidence before concluding that $p_1 - p_2 > \gamma$.)

Testing for a difference between proportions is straightforward for large samples because all the commonly used test statistics have either a normal or chi-squared distribution asymptotically. It's not straightforward for small samples, however, and there has been a spate of new research in recent years on "exact unconditional" tests for a difference between binomial proportions, especially for a zero difference (Suissa and Shuster, 1985, Berger, 1996, Chan, 1998). (Note that by "exact" test, we mean one in which probabilities are computed based on the true underlying distribution of (X, Y) rather than on an asymptotic approximation of some test statistic.) This research has been motivated, at least in part, by better computing: while exact unconditional tests are more powerful than exact conditional tests, such as Fisher's exact test (Suissa and Shuster, 1985, Liddell, 1978, Berger, 1996), they are also a lot more computationally intensive.

Thanks to Anastasios A. Tsiatis for his valuable support on this project.

In this paper, we show how to construct a confidence region p value test for superiority. Specifically, we start with a standard exact unconditional test based on a Wald statistic,

$$Z(x, y) = \frac{\widehat{p_1} - \widehat{p_2} - \gamma}{\sqrt{\frac{\widehat{p_1}(1-\widehat{p_1})}{n_1} + \frac{\widehat{p_2}(1-\widehat{p_2})}{n_2}}}, \tag{2}$$

where $\widehat{p_1} = x/n_1$ and $\widehat{p_2} = y/n_2$, and then apply the confidence region p value method of Berger and Boos to get a new test. (Note that all confidence region p value tests are derivative, i.e., they are built off of other tests.)

2. The Wald Test for Superiority

Exact unconditional tests use p values as their ultimate test statistics. Recall that the standard definition of a p value is the supremum, over all points in the null hypothesis, of the probability of observing a value of the test statistic equal to or more extreme than the value of the observed data. (Bickel and Doksum, 1977, Section 5.2B). Thus the p value at (x, y) in our problem, which we denote by $p_Z(x, y)$, is

$$
\begin{aligned}
&p_Z(x,y)\\
&= \sup_{p_1-p_2\le\gamma} P_{(p_1,p_2)}\{Z(X,Y)\ge Z(x,y)\}\\
&= \sup_{p_1-p_2\le\gamma} \sum_{(a,b)\in R_Z(x,y)} bin(a;n_1,p_1)bin(b;n_2,p_2),
\end{aligned}
$$

where $R_Z(x,y) = \{(c,d) : Z(c,d) \ge Z(x,y)\}$ and $bin(w;n,p) = \binom{n}{w}p^w(1-p)^w$. Rejecting $H_0 : p_1 - p_2 \le \gamma$ in favor of $H_a : p_1 - p_2 > \gamma$ for all (x,y) such that $Z(x,y) \ge Z_\alpha$, where Z_α is any level α critical value, is equivalent to rejecting for all (x,y) such that $p_Z(x,y) \le \alpha$. Thus, we use $p_Z(x,y)$ as our test statistic, rejecting whenever it is less than α.

Maximizing a function over a two-dimensional grid of points is much more computationally intensive than maximizing a function over an interval of points. If we can show that the supremum of $P_{(p_1,p_2)}\{Z(X,Y) \ge Z(x,y)\}$ over $(p_1,p_2) \in H_0$ occurs on the boundary of H_0, then we can maximize $P_{(p_1,p_1-\gamma)}\{Z(X,Y) \ge Z(x,y)\}$, which is a function of p_1 only. Sidik (1997) gave sufficient conditions for maximizing on the boundary.

Sidik's Theorem: Let the families of density functions of the independent random variables X and Y be $f(x,\theta_1)$ and $g(y,\theta_2)$, θ_1 and θ_2 being real-valued parameters. Let $\mathbf{R}$ be a set such that if (x,y) is in $\mathbf{R}$, so too are $(x+1,y)$ and $(x,y-1)$. (Such a set is "Barnard Convex" (Barnard, 1947).) If the family of distributions of X is stochastically increasing in θ_1 and that of Y in θ_2, then

$$P_{(\theta_1,\theta_2)}[(X,Y)\in\mathbf{R}] \le P_{(\theta_1',\theta_2')}[(X,Y)\in\mathbf{R}]$$

for all (θ_1',θ_2') such that $\theta_1 \le \theta_1'$ and $\theta_2' \le \theta_2$.

Some of the conditions in Sidik's theorem are obviously true. First, X and Y are independent by assumption. Second, since X and Y are both binomial random variables, they are stochastically increasing in their proportion parameters. Is $R_Z(x,y)$ Barnard convex, however? One can show, by looking at the partial derivatives of $Z(x,y)$, that, while $R_Z(x,y)$ isn't Barnard convex for all points in the sample space, it is for points in the set A, where $A = \{(x,y) : \widehat{p_1} - \widehat{p_2} \ge \gamma\}$. But, this is the region in which we are most interested since 0.01, 0.05, and 0.1 level rejection regions will all be contained in A. Thus, by Sidik's theorem, we know that for $(x,y) \in A$,

$$
\begin{aligned}
&p_Z(x,y)\\
&= \sup_{p_1-p_2=\gamma} \sum_{(a,b)\epsilon R_z(x,y)} bin(a;n_1,p_1)bin(b;n_2,p_1-\gamma)\\
&= \sup_{\gamma\le p_1\le 1} \sum_{(a,b)\epsilon R_z(x,y)} bin(a;n_1,p_1)bin(b;n_2,p_1-\gamma).
\end{aligned}
$$

3. The Drawback of the Standard p Value

As stated previously, the standard definition of a p value overcomes any dependence on a parameter by maximizing over the entire range of that parameter. This definition simply neglects the information contained in the data about the parameter and the supremum often occurs at a value that is incongruous with any reasonable parameter estimate. Thus, by the standard definition, a p value is often larger than it would have been if the maximization had been restricted to a set of values that were "consistent" with the observed value. This results in fewer sample points being included in a level α rejection region, i.e., in less powerful tests.

Berger and Boos (1994) introduced a way to incorporate information available on the nuisance parameter into the calculation of a p value.

The Berger and Boos Confidence Region p Value. Let θ be a parameter that can take on a range of values under H_0, i.e., a parameter that isn't fixed, and let $C_\beta(\mathbf{x})$ be a $1-\beta$ confidence region (or interval) for θ. Suppose that $T(\mathbf{x})$ is a test statistic and that the larger the value of T, the stronger the evidence against H_0. Let $p(\theta|\mathbf{x}) = P_\theta(T(\mathbf{X}) \ge T(\mathbf{x}))$ and note that $p(\theta|\mathbf{x})$, the "p value function," is the quantity we maximize to compute a standard p value at $\mathbf{x}$. Then the confidence region p value is

$$p_C(\mathbf{x}) = \sup_{\theta\epsilon C_\beta(\mathbf{x})} p(\theta|\mathbf{x}) + \beta. \tag{3}$$

β in Equation 3 is added as a penalty for maximizing over a reduced set. In practice, β should be a lot smaller than α, since a β close to or greater than α would guarantee that $p_C > \alpha$ for most or all $\mathbf{x}$. Berger and Boos suggest using values of β in the neighborhood of 0.001 to 0.005. They also proved that $p_C(\mathbf{x})$ is a valid p value, i.e., $P_\theta(p_C(\mathbf{x}) \le \alpha) \le \alpha$ for all θ in H_0. Thus, $p_C(\mathbf{x}) \le \alpha$ defines a valid level α test, i.e., $p_C(\mathbf{x})$ is not liberal.

4. A Confidence Region p Value Test

The main step in deriving a confidence region p value is to construct a $(1-\beta)$ confidence region. Since the supremum of the p value function, i.e. the function we maximize to compute the standard p value, occurs on the boundary of H_0, at least for points in A, it's natural to construct a confidence region on the boundary of H_0. This amounts to constructing an interval for $p_1 \in [\gamma,1]$ since the boundary H_0 is the set of ordered pairs $(p_1,p_1-\gamma)$, $p_1 \in [\gamma,1]$. We constructed an interval for p_1 based on $X \sim$

$bin(p_1, n_1)$ and $Y \sim bin(p_1 - \gamma, n_2)$ using what we'll call the"monotone" method. We denote the p value associated with this interval as p_{C_M}.

The Monotone Method (Theorem 9.2.3 in Casella and Berger). Let T be a discrete or continuous statistic with cdf $F_T(t|\theta)$ a decreasing function of θ. For each t, define $\theta_L(t)$ and $\theta_U(t)$ as

$$P(T \leq t|\theta_L(t)) = \frac{\beta}{2} \quad \text{and} \quad P(T \geq t|\theta_U(t)) = \frac{\beta}{2}.$$

Then $[\theta_L(t),\, \theta_U(t)]$ is a $1-\beta$ confidence interval for θ.

We set $T = X+Y$ and computed the lower and upper bounds of the interval numerically. Denote this interval by C_M^*. To use Equation 3 we actually need a two-dimensional confidence set for $(p_1, p_2) \in H_0$. (For starters, we don't even know that the supremum of the p value function occurs on the boundary of H_0 for all sample points.) The following lemma, the proof of which is beyond the scope of this paper, shows how to define this two-dimensional set, which we denote by C_M, in terms of the our one dimensional interval, C_M^*.

Lemma 1. Suppose H_0 is true, i.e. $p_1 - p_2 \leq \gamma$. Let $C_M^*(x,y) = [L_M(x,y), U_M(x,y)] = C_M^*(t) = [L_M(t),\, U_M(t)]$ denote the $(1-\beta)$ "monotone" confidence interval for $p_1 \in [\gamma, 1]$ based on $X \sim bin(p_1, n_1)$ and $Y \sim bin(p_1 - \gamma, n_2)$, X and Y independent. Recall that $t = x + y$. Then $C_M(x,y) = C_M(t) = \{(p_1,\, p_2) : (p_1,\, p_2) \in H_0,\; p_2 \geq L_M(t) - \gamma,\; p_1 \leq U_M(t)\}$ is a $(1-\beta)$ confidence region for $(p_1,\, p_2) \in H_0$. (See Figure 1.)

Thus, the confidence region p value is

$$p_{C_M}(x,y) = \sup_{C_M(x,y)} \sum_{(a,b)\in R_z(x,y)} \{bin(a; n_1, p_1) bin(b; n_2, p_2)\} + \beta.$$

If $(x,y) \in A$,

$$\sup_{(p_1,p_2)\in C_M(x,y)} P_{(p_1,p_2)}\{Z(X,Y) \geq Z(x,y)\} = \sup_{p_1 \in C_M^*(x,y)} P_{p_1}\{Z(X,Y) \geq Z(x,y)\},$$

by Sidik's theorem, and

$$p_{C_M}(x,y) = \sup_{p_1 \in C_M^*(x,y)} \sum_{R_z(x,y)} \{bin(a; n_1, p_1) bin(b; n_2, p_1 - \gamma)\} + \beta.$$

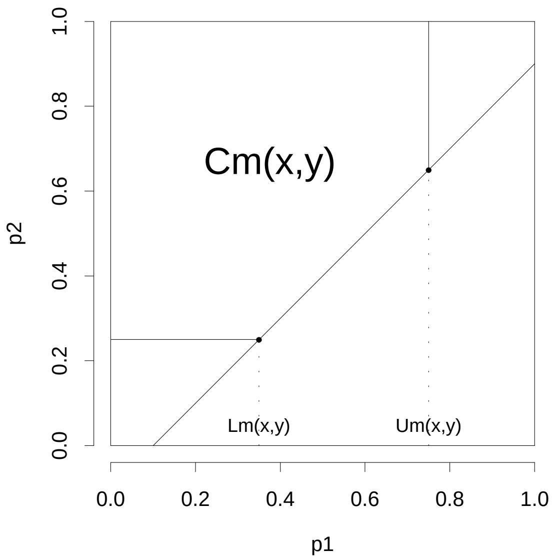

Figure 1: $C_M(x,y)$ is the $(1-\beta)$ confidence region for $(p_1, p_2) \in H_0$ defined by Lemma 1. The diagonal line is $p_1 - p_2 = \gamma$.

5. Two Illustrative Comparisons

How well our confidence region p value performs relative to p_Z depends on the relative sizes of n_1 and n_2. Generally, the more unbalanced the samples, the better the performance of our p values relative to p_Z. For highly unbalanced samples, $R_{C_M}(\alpha)$, the rejection region determined by $p_{C_M} \leq \alpha$, typically contains many more points than $R_Z(\alpha)$, the rejection region determined by $p_Z \leq \alpha$. In addition, $R_Z(\alpha)$ is often a proper subset of $R_{C_M}(\alpha)$, in which case p_{C_M} is uniformly more powerful than p_Z. When $n_1 \approx n_2$, however, p_Z is relatively "efficient," and $R_{C_M}(\alpha)$ and $R_Z(\alpha)$ tend to be very similar or even identical.

In this section, we give two examples, the first with $n_1 \gg n_2$ and the second with $n_1 = n_2$. In both, we test $H_0 : p_1 - p_2 \leq 0.1$ vs. $H_a : p_1 - p_2 > 0.1$ and use 99.9% confidence intervals for p_1.

Example 1: $n_1 = 45$, $n_2 = 15$, $\alpha = 0.05$. $R_Z(0.05) = \{(x,y) : p_Z(x,y) \leq 0.05\}$, which has 150 points, is a proper subset of $R_{C_M}(0.05) = \{(x,y) : p_{C_M}(x,y) \leq 0.05\}$, which has 167! The cause of p_Z's poor performance relative to p_{C_M} in this example, and for $n_1 \gg n_2$ in general, is rooted in the behavior of its p value functions. Generally, the p value function depends on p_1 and p_2. But, we're most concerned with the p values of points in $A = \{(x,y) : \widehat{p_1} - \widehat{p_2} \geq \gamma\}$. For these points, the maximum value of p value function is guaranteed to occur at a parameter point on the boundary of H_0. Thus, for these points, the relevant part of the p value function depends on p_1 only. Denote the p value function at (x,y) by $p_Z(x,y|p_1)$. When $n_1 \gg n_2$, $p_Z(x,y|p_1)$ tends to reach its maximum at a small value of p_1 and to be right skewed. (In contrast, for $n_1 \ll n_2$, $p_Z(x,y|p_1)$ tends to reach its maximum at large values of p_1 and to be left-skewed.) Thus, if the confidence interval for p_1 is in the right tail, the confidence region p value will be a lot smaller than the standard p value.

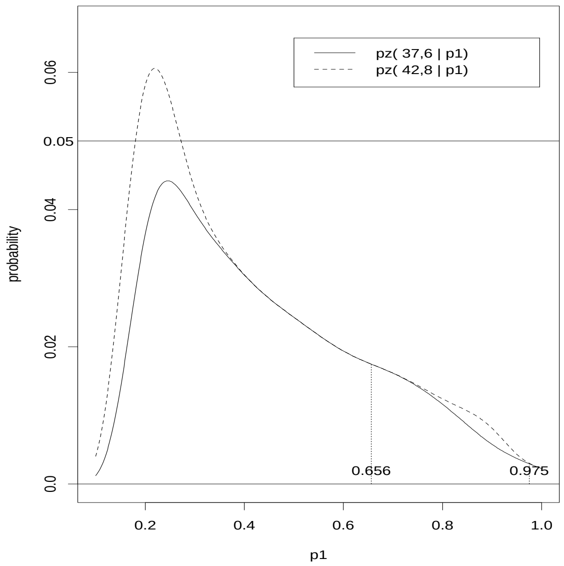

Figure 2: $(37,6)$ is the point with the largest p value less than or equal to 0.05. $(42,8)$ isn't in the original rejection region, but is in $R_{C_M}(0.05)$. The sup of $p_Z(42,8|p_1)$ over $C^*_M(42,8) = \{0.6561, 0.9754\}$ is 0.0175 and thus $(42,8) \in R_{C_M}(0.05)$.

Figure 2 depicts the p value functions for the points $(37,6)$ and $(42,8)$. The point $(37,6)$, which is contained in $R_Z(0.05)$, is the point with the largest p value less than or equal to 0.05, with $p_Z(37,6) = 0.04419$. Since $p_Z(37,6|p_1) = P(R_Z(0.05)|p_1)$ drops off sharply in its right tail, we would expect to be able to add some points to $R_Z(0.05)$ that would raise the curve in the right tail without having much of an effect on it for small values of p_1. $(42,8)$ is one such point. It is not an element of $R_Z(0.05)$ because $p_Z(42,8) = 0.06056$, but is an element of $R_{C_M}(0.05)$. $C^*_M(42,8) = \{0.6561, 0.9754\}$ is "located" in the right tail and the supremum of $p_Z(42,8|p_1)$ over $C^*_M(42,8)$ is much smaller than the supremum over $[0.1,1]$, with $p_{C_M}(42,8) = 0.0185$! Sixteen other sample points were added to $R_Z(0.05)$ in this way to form $R_{C_M}(0.05)$.

Example 2: $n_1 = n_2 = 30$, $\alpha = 0.01$. In this case, $R_{C_M}(0.01) = \{(x,y) : p_{C_M}(x,y) \leq 0.01\}$ has 193 points. Although $R_{C_M}(0.01)$ is a proper subset of $R_Z(0.10) = \{(x,y) : p_Z(x,y) \leq 0.01\}$, which has 195 points, the two sets aren't very different. Thus, restricting the maximization to a 99.9% confidence interval for p_1 didn't, in general, result in smaller p values. This happens because, for balanced samples, p value functions are symmetric about $\frac{(1+\gamma)}{2}$ (0.55 in this example) and are either bimodal or unimodal and closer, uniformly, to their maxima, leaving less room for additional points. Typically, a confidence interval will contain the value of p_1 at which $p_Z(x,y|p_1)$ reaches its maximum (or maxima if the curve is bimodal). Thus, because β is added as a penalty, p_{C_M} tends to be a bit larger than p_Z.

6. Broad Comparisons of p_{C_M} to p_Z

We compared tests based on p_{C_M} to those based on p_Z at α levels of 0.1, 0.05, and 0.01 for nine different samples – $(10,10)$, $(13,7)$, $(15,5)$, $(30,30)$, $(40,20)$, $(45,15)$, $(50,50)$, $(66,34)$, $(75,25)$ – using 99.9% confidence intervals. These samples represent small ($N = n_1 + n_2 = 20$), medium ($N = 60$) and moderately large ($N = 100$) experiments with ratios of n_1 to n_2 of $1:1$, $2:1$, and $3:1$. These comparisons are summarized in Table 1.

p_{C_M}'s performance relative to p_Z is closely tied to the ratio of n_1 to n_2. N, the total size of the experiment, also plays a role. The ratio of n_1 to n_2 doesn't seem to affect how many points, if any, are dropped from $R_Z(\alpha)$ by p_{C_M}. Regardless of the ratio, p_{C_M} never cuts more than a few points from R_Z. The ratio of n_1 to n_2 does affect how many points are added to R_Z, however. As stated previously, the more unbalanced the samples, the more skewed the p value functions and more room for new points.

p_{C_M} was as or more powerful uniformly than p_Z in 16 of the 27 cases, i.e., the rejection region determined by p_Z was a subset of the rejection region determined by p_{C_M} in 16 of 27 cases. In another 7 cases, neither test was uniformly more powerful than the other. In the four remaining cases, p_Z was uniformly more powerful than p_{C_M}.

It's instructive to break down these results by the ratio of n_1 to n_2.

		1:1			2:1			3:1		
α		$(10,10)$	$(30,30)$	$(50,50)$	$(13,7)$	$(40,20)$	$(66,34)$	$(15,5)$	$(45,15)$	$(75,25)$
0.01	add	0	0	0	0	2	16	0	12	37
	drop	0	2	6	2	4	1	0	3	2
	$R_Z(\alpha)$	12	195	635	10	157	524	5	105	378
	$R_{C_M}(\alpha)$	12	193	629	8	155	539	5	114	413
0.05	add	0	0	0	0	7	20	0	17	40
	drop	0	4	0	0	1	0	0	0	2
	$R_Z(\alpha)$	20	249	747	17	204	640	10	150	485
	$R_{C_M}(\alpha)$	20	245	747	17	210	660	10	167	523
0.10	add	0	2	0	0	3	10	1	12	23
	drop	0	0	0	0	0	0	0	0	1
	$R_Z(\alpha)$	25	275	812	21	237	710	13	183	561
	$R_{C_M}(\alpha)$	25	277	812	21	240	720	14	195	583

Table 1: Comparisons of tests based on p_{C_M} to those based on p_Z. Note: "add" refers to how many points are in $R_{C_M}(\alpha)$ but not in $R_Z(\alpha)$; "drop" to how many points are in $R_Z(\alpha)$ but not in $R_{C_M}(\alpha)$. "$R_Z(\alpha)$" and "$R_{C_M}(\alpha)$" give the number of points in $R_Z(\alpha)$ and $R_{C_M}(\alpha)$.

$1:1$. The rejection regions, R_Z and R_{C_M}, are equivalent in 5 of 9 cases. R_Z is a proper subset of R_{C_M} in 1 case and R_{C_M} is a proper subset of R_Z in 3. In none of the nine cases, however, do R_Z and R_{C_M} differ by more than 6 points!

$2:1$. p_Z is uniformly more powerful than p_{C_M} in just one case. The tests are equivalent in two cases. In three cases, R_Z is a proper subset of R_{C_M}, and thus p_{C_M} is uniformly more powerful than p_Z. In the remaining three cases, neither rejection region is a subset of the other and the power functions cross. However, in two of these cases, many times more points are added to R_Z than are kicked out. (The 0.01 level test based on p_{C_M} for sample sizes $(66,34)$ adds 16 points to $R_Z(0.01)$ while dropping just one. The 0.05 level test based on samples of size $(40,20)$ adds seven points while dropping one, and the 0.10 level test based on the same experiment adds two points and drops one.)

$3:1$. In two cases, the tests are equivalent. In another three, R_Z is a proper subset of R_{C_M}. (One, 12, and 17 points are added to R_Z.) In four instances, neither rejection region is a subset of the other. However, in these cases, many times more points are added to R_Z than are dropped from it. (On average, about sixteen points are added to R_Z for each that is dropped.) p_Z is never uniformly more powerful than p_{C_M}.

We should also say a few words about the effect of N, the total number of patients or trials in the experiment, on p_{C_M}'s performance. We have stressed that p_{C_M} performs best relative to p_Z for unbalanced samples. We need to offer a caveat. For very small samples, such as $N = 20$, p_{C_M} and p_Z were roughly the same, regardless of the ratio of n_1 to n_2. For $N = 20$, 99.9% confidence intervals are very wide and there is little difference between maximizing over a confidence interval and maximizing over the entire range of p_1. (On the other hand, p values are much further apart from one another, i.e., much more "discrete," for $N = 20$ than for $N = 60$ or $N = 100$. Thus, adding $\beta = 0.001$ to each p_Z didn't wind up forcing many points out, either.)

7. Conclusion

Rejection regions based on $p_{C_M} \leq \alpha$ ranged from being roughly equivalent to those determined by $p_Z \leq \alpha$ to being significantly larger. And, we could improve the performance of p_{C_M} if we tailored our choice of β to size of the experiment.

LITERATURE CITED

BARNARD, G.A., (1947), "Significance Tests for 2×2 Tables," *Biometrika*, 34, 123-138.

BERGER, R.L., and BOOS, D.D., (1994), "*P* Values Maximized Over a Confidence Set for the Nuisance Parameters," *Journal of the American Statisical Association*, 89, 1012-1016.

BERGER, R.L., (1996), "More Powerful Tests From Confidence Interval *p* Values," *The American Statistician*, 50, 314-317.

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

CASELLA, G., and BERGER, R. L., (1990), *Statistical Inference*, Pacific Grove: Wadsworth and Brooks/Cole.

CHAN, I.S.F., (1998), "Exact Tests of Equivalence and Efficacy with a Non-zero Lower Bound for Comparative Studies" *Statistics in Medicine*, 17, 1403-1413.

FISHER, R.A., (1934,1970), *Statistical Methods for Research Workers*, Edinburgh: Oliver and Boyd.

LIDDELL, D., (1978), "Practical Tests for 2×2 Contingency Tables" *The Statistician*, 25, 295-304.

SIDIK, K., (1997), *Exact Unconditional Tests for Discrete Data*, Doctoral Dissertation, Raleigh: North Carolina State University.

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