Confidence Intervals that Match Fisher's Exact or Blaker's Exact Tests

Version: July 20, 2009

Michael P. Fay

Summary

The two-sided Fisher's exact test is one of the most common tests for testing independence in a 2 by 2 table, or equivalently, of testing that the odds ratio is different from one. We desire a confidence interval on the odds ratio that contains the null odds ratio if and only if the test fails to reject the null. Unfortunately, the confidence set created by inverting the family of two-sided Fisher's exact tests may consist of more than one interval. Even if we create the smallest interval that contains this confidence set, the resulting "matching" interval is not the usual confidence interval reported for odds ratios conditional on the marginals of the table. This usual interval matches with a different implementation of Fisher's exact test, the typically less powerful but more directionally balanced test that rejects if the minimum of two one-sided Fisher's exact tests reject at one half the nominal significance level. We discuss these two exact two-sided tests and a third one suggested by Blaker (2000, Canadian Journal of Statistics, 783-798), and study the matching confidence intervals for each test. The R package exact2x2 is provided to calculate all three tests and their matching intervals.

1 Introduction

For comparing two groups when the response is binary, one of the most common ways of testing for a difference between the groups is Fisher's exact test. When the two-sided version of that test is applied to the 2×2 table given by Example 1 of Table 1 the p-value is 0.0437 which denotes significance at the conventional 0.05 level. Statistical significance only signifies that the observed data is extreme under the null hypothesis of independence and tells little about the magnitude of the actual effect, so it is recommended (see e.g., the CONSORT statement: Moher, Schultz, Altman, et al, 2001, item 17) to report effect size and confidence intervals along with p-values.

One might think that since Fisher's exact test has such a long history and that it continues to be used frequently, it would be obvious what are the appropriate confidence intervals to be used with the test; however, it

is not at all clear from the literature and available software. In R (version 2.9.1) the fisher.test function applied to the table above gives the conditional maximum likelihood estimate of the odds ratio (OR) 0.219 along with the 95% confidence interval (CI) of $(0.039, 1.06)$. We call this interval the exact conditional tail interval (ECTI) (see equation 4 below). The ECTI is not consistent with the two-sided Fisher's exact test, since it implies non-significance at the 0.05 level. What we want is an interval that is strongly consistent with the two-sided Fisher's exact test, where a strongly consistent $100(1-\alpha)\%$ confidence interval contains the null value of the parameter if and only if the corresponding test fails to reject at the α level. The problem is not with the R software, since the only exact confidence interval for the odds ratio offered by SAS (version 9.1) and StatXact (StatXact 8 Procs) is the ECTI. Although there are confidence intervals that more often match the inferences from Fisher's exact test, we show here that there is not a method that will always produce a strongly consistent confidence interval.

There are two problems. First, the $100(1-\alpha)\%$ ECTI is strongly consistent with a different two-sided Fisher's exact test (see Section 3.2 below). The second problem is that if we invert the usual two-sided Fisher's exact test, it is not guaranteed that the resulting confidence set will be an interval (i.e., there may be a "hole" in the set). This problem with inversion of tests not being intervals has been extensively studied for the single binomial parameter (see e.g., Casella and Berger, 2002, p. 431, or Blaker, 2000).

Blaker (2000) and Agresti and Min (2001) both give excellent discussions of the formation and properties of two-sided confidence intervals for all kinds of discrete data in many more situations than the 2×2 table, but neither paper explicitly examines the confidence set that is an inversion of Fisher's two-sided exact test. We do that in this paper. Additionally, we apply a general method of Blaker (2000) to create an exact test with confidence intervals for the 2×2 table.

2 Outline of the Problem

For the 2×2 table, we use the model with $\mathbf{X} = [X_0, X_1]$, where $X_i \sim Binomial(n_i, \pi_i)$ and are independent of each other and the n_i are fixed and known. There are other models for the 2×2 table but for most it is reasonable to condition on the marginals so that inferences can be calculated from Fisher's noncentral hypergeometric distribution as we do here (see e.g., Lehmann and Romano, 2005, or Yates, 1984). Unconditional tests are not discussed in this paper, and for a comparison of the two types of tests see

Agresti (2001) and the references cited there. For this paper the parameter of interest is the odds ratio, $\beta = \frac{\pi_1(1-\pi_0)}{\pi_0(1-\pi_1)}$, and the nuisance parameter is $\psi = \pi_0 + \pi_1$. The distribution of $\mathbf{X}$ is completely described by the parameter vector $\theta = [\beta, \psi]$.

We are interested in confidence intervals about β, so we consider the family of two-sided hypothesis tests indexed by β_0 where the hypotheses are:

$$H_0 : \beta = \beta_0, 0 < \psi < 1$$
$$H_1 : \beta \neq \beta_0, 0 < \psi < 1.$$

The usual application only considers the case where $\beta_0 = 1$. In Section 3 we discuss three families of tests associated with these hypotheses. For any of these three families, let $p_{\beta_0}(\mathbf{x})$ be the two-sided p-value associated the null $H_0 : \beta = \beta_0$, where we reject when $p_{\beta_0}(\mathbf{x}) \leq \alpha$. A conceptually simple way to create confidence sets from any family is to invert the tests, so that the $100(1-\alpha)\%$ confidence set is (see e.g., Casella and Berger, 2002):

$$C(\mathbf{x}, 1-\alpha) = \{\beta : p_\beta(\mathbf{x}) > \alpha\}. \quad (1)$$

The confidence set given by equation 1 is said to be *strongly consistent* with the family of tests, since the $100(1-\alpha)\%$ confidence interval does not contain β_0 if and only if the α-level test corresponding to $H_0 : \beta = \beta_0$ rejects. We call this confidence set the *inversion* of the family of tests. Since this inversion is not guaranteed to be an interval (see Blaker, 2000 or Section 4.1), following Blaker (2000) we use the smallest interval which contains all of the parameter values of the inversion (i.e., it fills in the holes of the inversion if they exist). We call this interval the *matching confidence interval* to the family of tests (or to one member of that family).

The major point of this paper is that when confidence intervals are used to supplement information from a test, the matching confidence interval should be used since this interval will avoid (as much as is possible) the problem of rejecting the null at the α-level but including the null parameter in the $100(1-\alpha)\%$ confidence interval.

3 Three Two-Sided Exact Conditional Tests for 2×2 Tables

3.1 Preliminaries

Each of the null hypotheses in the family of hypotheses described by equation 1 is a point hypothesis in terms of β. If we condition on $X_0 + X_1$, the

sufficient statistic for ψ, then we obtain a likelihood without ψ terms:

$$Pr[X_1 = x; \beta] = f_\beta(x) = \frac{\binom{n_1}{x}\binom{n_0}{k-x}\beta^x}{\sum_{i=x_{min}}^{x_{max}}\binom{n_1}{i}\binom{n_0}{k-i}\beta^x}, \text{ for } x \in [x_{min}, x_{max}],$$

where $k = x_0 + x_1$, $x_{min} = \max(0, n_0 - k)$ and $x_{max} = \min(k, n_1)$. This distribution is Fisher's non-central hypergeometric distribution (see e.g., Fog, 2008).

Once we condition on the marginals, the table is completely described by x_1, and smaller values of x_1 suggest smaller odds ratios. Since we are only considering non-randomized tests, there is only one commonly used exact one-sided test, the one-sided Fisher's exact test, and it is based on the ordering of x_1. The exact versions of other non-randomized historical one-sided tests are constructed this way and are all equivalent (see Davis, 1986 or StatXact 8 Procs Manual).

3.2 Central Fisher's Exact Test

The one-sided Fisher's exact tests have p-values of either

$$\begin{aligned} p_\beta^{(lte)}(\mathbf{x}) &= \sum_{i:i\le x_1} f_\beta(i) \\ &\textit{or} \\ p_\beta^{(gte)}(\mathbf{x}) &= \sum_{i:i\ge x_1} f_\beta(i), \end{aligned} \tag{2}$$

and we can create a two-sided test with p-value equal to

$$p_\beta(\mathbf{x}) = \min\left\{1, 2 * \min\left(p_\beta^{(lte)}(\mathbf{x}), p_\beta^{(gte)}(\mathbf{x})\right)\right\} \tag{3}$$

This doubling of the one-sided p-value is a common and simple method for defining the two-sided p-value (Gibbons and Pratt, 1975).

The inversion of this test is an interval because the one-sided p-values given in equations 2 are unimodal in β. Unimodality follows from the monotonicity in β of each side (see the Appendix of Mehta, Patel, and Gray, 1985) and equation 3. The matching interval is the exact conditional tail interval (ECTI) mentioned previously. Specifically, let the ECTI be

$C(\mathbf{x}, 1-\alpha) = [L(x_1, 1-\alpha), U(x_1, 1-\alpha)]$ which are the solutions to (see e.g., StatXact 8 Procs Manual):

$$\begin{aligned} L(x_1, 1-\alpha) &= \begin{cases} 0 & \text{if } x_1 \text{ is } x_{min} \\ \{\beta : \sum_{i:i \geq x_1} f_\beta(i) = \alpha/2\} & \text{otherwise} \end{cases} \\ U(x_1, 1-\alpha) &= \begin{cases} \infty & \text{if } x_1 \text{ is } x_{max} \\ \{\beta : \sum_{i:i \leq x_1} f_\beta(i) = \alpha/2\} & \text{otherwise} \end{cases} \end{aligned} \tag{4}$$

This is a central interval meaning that we can bound the probability that the true β is less than the lower interval by $\alpha/2$ and similarly for the upper interval. Because of this property we call the test associated with the $p_\beta(\mathbf{x})$ of equation 3 the *central* Fisher's exact test. The test is also known as twice the one-sided Fisher's exact test.

3.3 Two-sided Fisher's Exact Test

The usual p-value associated with the two-sided Fisher's exact test is not the central one mentioned in the previous section but,

$$p_\beta(\mathbf{x}) = \sum_{i:f_\beta(i) \leq f_\beta(x_1)} f_\beta(i) \tag{5}$$

This p-value uses the "principal of minimum likelihood", which has little formal motivation and can lead to absurd inferences in some situations (Gibbons and Pratt, 1975). However, in the case of the conditional test on the 2×2 table, the principle of minimum likelihood gives reasonable answers because for fixed β the non-central hypergeometric distribution is unimodal in the x_1 values so that the values of x_1 in which we fail to reject will always be a set of consecutive integers (see e.g., Liao and Rosen, 2001).

Based on common current usage (see R help for fisher.test, SAS help for Proc Freq, and StatXact manual), we will call this test *the* two-sided Fisher's exact test, despite the fact that Fisher himself appeared to prefer the central Fisher's exact test (Yates, 1984, p. 444).

The inversion of this test may not be an interval as will be shown in Section 4.1. Consequently, because of the p-value function of equation 5 is not unimodal in β, one cannot simply find two values of β where $p_\beta(\mathbf{x}) = \alpha$, since there may be more than two. Calculation of the matching interval is discussed in Section 4. These matching two-sided Fisher's exact intervals have been suggested by Baptista and Pike (1977), although they did not mention the cases when the confidence set is not an interval (see also Table 2 of Agresti and Min, 2001).

3.4 Blaker's Exact Test

An alternative method for creating a two-sided p-value is to add to the one-sided p-value "an attainable probability in the other tail which is as close as possible to the one tailed P-value obtained" (Gibbons and Pratt, 1975). To maximize power, we follow the recommendation of Blaker (2000) and use the largest tail probability in the opposite tail which is less than or equal to the observed tail. See Appendix A for an explicit representation of that p-value. We call the resulting test, Blaker's exact test. From first principles, this is as reasonable if not more reasonable (see Gibbons and Pratt, 1975) as using the principle of minimum likelihood as is done Fisher's two-sided exact test. In the 2×2 table case the two-sided Fisher exact p-values will for many null hypotheses in the family coincide with the Blaker p-values. When the two p-values do not coincide, and when the principle of minimum likelihood may lead the two-sided Fisher to add more probability in the opposite tail than the observed one, and it is hard to see how this is desired over the smaller p-values of Blaker. We know of no commonly used statistical property for which the two-sided Fisher's exact test performs better than Blaker's exact test, and the greatest reason for using the former test may be tradition and ease of explanation.

Blaker (2000, see also Blaker and Spjøtvoll, 2000) showed that the p-value described above can be written in the following way. Let $F_\beta(x) = Pr[X_1 \leq x \mid \beta]$, $\bar{F}_\beta(x) = Pr[X_1 \geq x \mid \beta]$, and $\gamma(x, \beta) = \min\{F_\beta(x), \bar{F}_\beta(x)\}$, then the p-value (also called the acceptability function) of Blaker (2000) is

$$p_\beta(\mathbf{x}) \quad = \quad Pr\left[\gamma(X_1, \beta) \leq \gamma(x_1, \beta)\right]. \tag{6}$$

As with the two-sided Fisher's exact test the inversion of the test is a confidence set which may not be an interval since $p_\beta(\mathbf{x})$ of equation 6 is not necessarily unimodal in β. For applications we use the matching confidence interval which fills in the holes in the inversion and the calculation has the same problems as for the matching intervals to the two-sided Fisher's exact test which will be described in the next section.

4 Calculation of Intervals for Non-Unimodal P-value Functions

We begin by noting the non-unimodality in β of the p-value function for the two-sided Fisher's exact test and Blaker's exact test, which points out the difficulty of the calculating the matching intervals. Similar observations have

been made previously (see Blaker, 2000, Vos and Hudson, 2008). Then we give a method for calculating the confidence interval, and what we can say about the bound on its accuracy. The examples in this section are chosen not because they are typical, but because they are atypically difficult to calculate.

4.1 The Strongly Consistent Confidence Set is not Guaranteed to be an Interval

Ideally, we would want to use the inversion of the two-sided Fisher's exact test for our confidence intervals; however, the resulting confidence set is not guaranteed to be an interval. In Example 2 of Table 1 (chosen to highlight this point) the confidence set created by inverting the family of Fisher's exact tests is not a confidence interval: the resulting 95% confidence set is $\{\beta : \beta \in (0.178, 0.998) \text{ or } \beta \in (1.010, 1.018)\}$. In Figure 1 we plot $p_{\beta_0}(\mathbf{x})$ from the three tests. We see that for $\beta_0 = 1$ for the two-sided Fisher's exact test the p-value is significant at the 0.05 level, $p_1(\mathbf{x}) = 0.04999$, but for slightly larger or smaller β_0 the p-value is not significant, $p_{1.015}(\mathbf{x}) = 0.05008$ and $p_{0.99}(\mathbf{x}) = 0.05015$. Blaker's exact test can also have this problem, although in this case $p_{1.015}(\mathbf{x}) = 0.0354$ for Blaker's test. The problem is the non-unimodality of the p-value function and this motivates the matching confidence intervals defined previously. The central Fisher's exact test is continuous and unimodal and avoids many of these problems.

4.2 An Algorithm for Calculating Confidence Intervals on Non-monotonic P-value Functions

Blaker (2000) gave a simple algorithm for the calculation of the confidence interval for the single binomial parameter using his acceptability function. We describe a similar algorithm pictorially applied to a two-sided 2×2 table example. Figure 2 shows an example where the p-value is calculated only a certain points (specifically at $1 \pm j * 0.002, j = 0, 1, 2, \ldots$). The solid points are the ones above α. Other p-values measured to the right of the last open circle are below the range of the vertical axis, so that the largest calculated odds ratio that gives p-values greater than α is 0.986. The actual upper value of the matching confidence interval is 1.01375 since that is the largest β_0 such that $p_{\beta_0}(\mathbf{x}) > \alpha$. Thus, although the p-values are measured every 0.002 the error in the upper limit calculated this way is over ten times larger than 0.002 since $1.01375 - 0.98600 = 0.02775$. (The example is purely for illustration. It is Table 2 except subtracting 2 non-events in Group B and

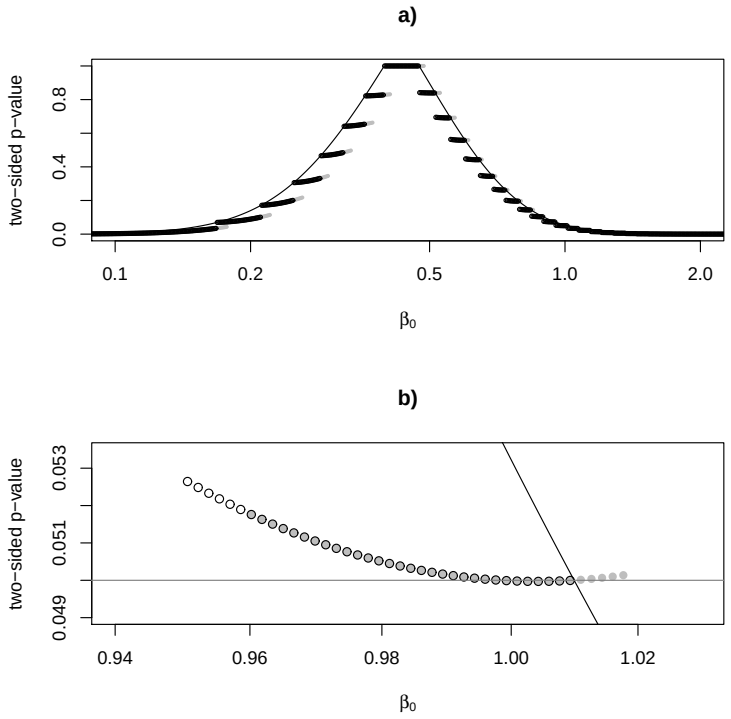

Figure 1: P-values from the three two-sided exact tests for testing $\beta = \beta_0$ for different values of β_0. Solid gray dots are two-sided Fisher's exact test, black open dots are Blaker's exact test, black line is central Fisher's exact, gray line is the reference line at 0.05. Figure b) is a blow-up of a portion of Figure a).

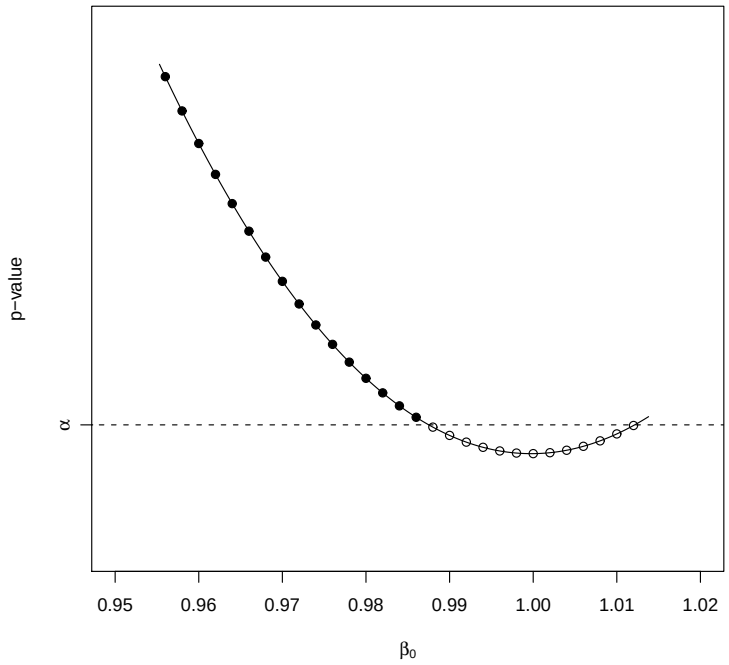

Figure 2: Figure to show difficulty with Blaker's algorithm. P-values evaluated at the points, $1 \pm j * 0.002, j = 0, 1, \ldots$.

using $\alpha = 0.0501$.)

In Appendix A we describe an algorithm for calculating the matching confidence interval for either the two-sided Fisher's exact test or Blaker's exact test. This algorithm allows calculation of the matching confidence limits either within a pre-specified tolerance level, or gives precision on the limits if the p-value function is very flat when it is very close to α. The algorithm is implemented in an R package called exact2x2. The package gives matching confidence intervals for all three tests described in Section 3.

5 Comparison of the Three Tests

First, consider the tables where $n_0 = n_1$. In these cases the noncentral hypergeometric distribution is symmetric, and since we know it is unimodal in the x_1 (see e.g., Liao and Rosen, 2001), the two-sided Fisher's exact test

and Blaker's exact test give equivalent p-values and confidence intervals.

For completeness, we list some important properties that all three tests share. All three tests are exact tests, meaning that the p-values are valid, and the only conservativeness of the tests is due to the discrete nature of the data. All three tests are nested, meaning that if a test fails to reject at the α_1 level then it must also fail to reject for all $\alpha > \alpha_1$. The matching confidence intervals are similarly nested (see Blaker, 2000). Because of the discrete nature of the data, none of the tests are unbiased. Although a randomized version of the one-side exact test is uniformly most powerful unbiased (Tocher, 1950), as is typically done in applications, we only consider non-randomized tests.

Whenever the central Fisher's exact test rejects, then Blaker's exact test also rejects, but not vise versa. Thus, Blaker's exact test is always more powerful than the central Fisher's exact test (see Figure 1). Blaker showed this result except with more generality (see Blaker, 2000, Corollary 1). This property does not hold for the two-sided Fisher's exact test. Although most of the time p-values from the central Fisher's exact test are larger than those of the two-sided Fisher's exact test, this is not always true (see Figure 1b).

For the central tests and matching intervals, besides the interpretational advantage of being central intervals, additionally the p-value function of the central test is continuous and unimodal in β_0. So the calculation of the confidence interval is easier and all matching confidence sets are intervals. Vos and Hudson (2008) emphasized a different point for other discrete test, which we would like to emphasize for these tests. It is possible that small changes in the data in the direction *away* from the null can lead to *less* significant tests. Suppose 2 more individuals were observed with no events in Group A, giving Example 3 of Table 1. Example 3 is clearly further away from the null than Example 2, since group A, which had the lower event rate in the original example, has an even lower one when those two individuals are added. The two-sided Fisher's exact p-value moves from significance for the original example ($p_1(\mathbf{x}) = 0.04999$) to non-significance with the 2 added individuals ($p_1(\mathbf{x}) = 0.05001$). In these two examples as is often the case Blaker's exact p-values exactly equal those of the two-sided Fisher's exact test. In contrast the p-values from the central Fisher's exact test properly show the ordering, giving a larger p-value for Example 2 (p=0.0532) than Example 3 (p=0.0506).

6 The Extent of the Non-Matching Problem

The examples of the previous sections were chosen to highlight specific atypical problems that may arise. Here we return to the motivating problem and explore how often we get differing significance inferences between the p-value and the confidence interval especially when using the p-value from the two-sided Fisher's exact test with the ECTI.

Suppose we are testing $H_0 : \beta_0 = 1$ at the 0.05 level, let I_p be the an indicator of whether the p-value from a test is less than or equal to 0.05, and let I_C be an indicator of whether the confidence interval *does not* contain 1 (i.e., implies rejection of H_0). Define a mismatch as any table which has $I_P \neq I_C$. Consider a set of 2×2 tables where n_0 and n_1 are fixed. Within the possible tables of each of these sets, we check and see if there are any mismatches, if so we say that the set has a mismatch problem.

First, we consider the situation where for I_p the p-value comes from the two-sided Fisher's exact test, and for I_C the confidence interval comes from the ECTI. Although this situation is not recommended, we study it because it appears to be the state of the current readily available software. We consider the 256 sets where n_0 and n_1 are each in $\{5, 6, \ldots, 20\}$. Of these sets, 234/256 or 91.4 percent have a mismatch problem. Thus, this problem is not a rare one.

Now consider the situation where each of the three tests uses its matching confidence interval. For the central Fisher's exact test, there will theoretically be no mismatches because the matching confidence interval is the inversion. For the two-sided Fisher's exact and Blaker's exact tests and the associated matching confidence intervals, we check the 256 sets of tables mentioned above and through exhaustive search. We find no mismatches in any of the 256 sets for either of those two test/matching confidence interval pairs. Thus, although mismatches between p-values and confidence intervals are possible when using the matching confidence intervals (see Table 1, Example 2), they are not necessarily common. We repeat that the examples of Table 1 were chosen to highlight specific points, not because they were typical.

7 Discussion

We recommend that whenever when confidence intervals for odds ratios are given together with p-values from a test, that the matching confidence intervals to the family of tests be presented. Because of the non-unimodality

of both Blaker's exact test and the two-sided Fisher's exact test, we cannot create strongly consistent confidence intervals, and there is a small possibility of rejecting the null that the odds ratio is one but including the value of 1 in the matching confidence interval. To avoid this problem the central Fisher's exact test (i.e., the other two-sided Fisher's exact test that uses twice the one-sided p-value) could be used. Although this central test is not as powerful as Blaker's exact test (nor is it likely to be as powerful as the usual two-sided Fisher's exact test), the resulting confidence intervals are central which allow more natural interpretation than the other two intervals. Finally, although the results of the hypothesis test are formally binary (reject or fail to reject), often it makes sense to examine the p-values which give a more nuanced view, allowing us to see that a pair of tables with p-values of $p = 0.04999$ and $p = 0.0501$ are much closer in terms of significance than the pair with $p = 0.04999$ and $p = 0.00001$.

References

Agresti, A. (2001). "Exact inference for categorical data: recent advances and continuing controversies" *Statistics in Medicine.* **20:** 2709-2722.

Agresti, A. and Min, Y. (2001). "On Small-sample Confidence Intervals for Parameters in Discrete Distributions," *Biometrics*, **57:** 963-971.

Baptista, J. and Pike, M.C. (1977). "Exact two-sided confidence limits for the odds ratio in a 2×2 table." *Journal of the Royal Statistical Society, Series C* **26** 214-220.

Blaker, H. (2000). "Confidence curves and improved exact confidence intervals for discrete distributions" *Canadian Journal of Statistics* **28:** 783-798.

Blaker, H. and Spjøtvoll, E. (2000). "Paradoxes and improvements in interval estimation." *American Statistician* **54:** 242-247.

Casella, G. and Berger, R.L. (2002). *Statistical Inference, second edition.* Duxbury: Pacific Grove, CA.

Davis, L.J. (1986). "Exact Tests for 2×2 Contingency Tables." *American Statistician* **40:** 139-141.

Dupont, W.D. (1986). "Sensitivity of Fisher's exact test to minor pertubations in 2×2 contingency tables." *Statistics in Medicine* **5:** 629-635.

Fog, A. (2008). "Sampling methods for Wallenius' and Fisher's Noncentral Hypergeometric Distributions." *Communications in Statistics-Simulation and Computation* **37:** 241-257.

Gibbons, J.D. and Pratt, J.W. (1975). "P-values: Interpretation and Methodology *American Statistician* **29:** 20-25.

Lehmann, E.L., and Romano, J.P. (2005). *Testing Statistical Hypotheses, third edition* Springer: New York.

Liao, J.G., and Rosen, O. (2001). "Fast and stable algorithms for computing and sampling from the noncentral hypergeometric distribution." *American Statistician* **55:** 366-369.

Mehta, C.R., Patel, N.R., and Gray, R. (1985). "Computing an exact confidence interval for the common odds ratio in several 2×2 contingency tables." *Journal of the American Statistical Association* **80:** 969-973.

Moher, D., Schultz, K.F., and Altman, D.G. (2001). "The CONSORT statment: revised recommendations for improving the quality of reports of parallel group randomized trials." *BMC Medical Research Methodology* **1:** 2.

StatXact 8 Procs User Manual (2007). Cytel Software Corporation: Cambridge MA.

Tocher, K.D. (1950). "Extension of the Neyman-Pearson Theory of Tests to Discontinuous Variates" *Biometrika* **37:** 130-144.

Vos, P.W., and Hudson, S. (2008). "Problems with binomial two-sided tests and the associated confidence intervals" *Australian and New Zealand Journal of Statistics* **50:** 81-89.

Yates, F. (1984). "Test of significance for 2×2 contingency tables. (with discussion)" *Journal of the Royal Statistical Society, Series A* **147:** 426-463.

Appendix A

We first consider the Blaker confidence interval. Note that $F_\beta(x) = \sum_{i=x_{min}}^{x} f_\beta(i)$ and $\bar{F}_\beta(x) = \sum_{i=x}^{x_{max}} f_\beta(i)$. We let

$$b(x_a, x_b) \quad = \quad \{\beta : F_\beta(x_a) = \bar{F}_\beta(x_b)\}$$

Let $b(x_1, x_{max}+1) = \infty$ and $b(x_{min}-1, x_1) = 0$, and let $F_\beta(x) = 0$ when $x < x_{min}$ and $\bar{F}_\beta(x) = 0$ when $x > x_{max}$. Then the another form of the Blaker p-value is

$$p_b(\mathbf{x}) = \begin{cases} F_b(x_1) + \bar{F}_b(x_1 + j + 1) & \text{for } b(x_1, x_1 + j) < b \leq b(x_1, x_1 + j + 1); j = 1, 2, \ldots, x_0 \\ 1 & \text{for } b(x_1 - 1, x_1) \leq b \leq b(x_1, x_1 + 1) \\ \bar{F}_b(x_1) + F_b(x_1 - j - 1) & \text{for } b(x_1 - j - 1) \leq b < b(x_1 - j, x_1); j = 1, 2, \ldots, x_1 \end{cases}$$

Figure 3 helps to explain the Blaker p-value.

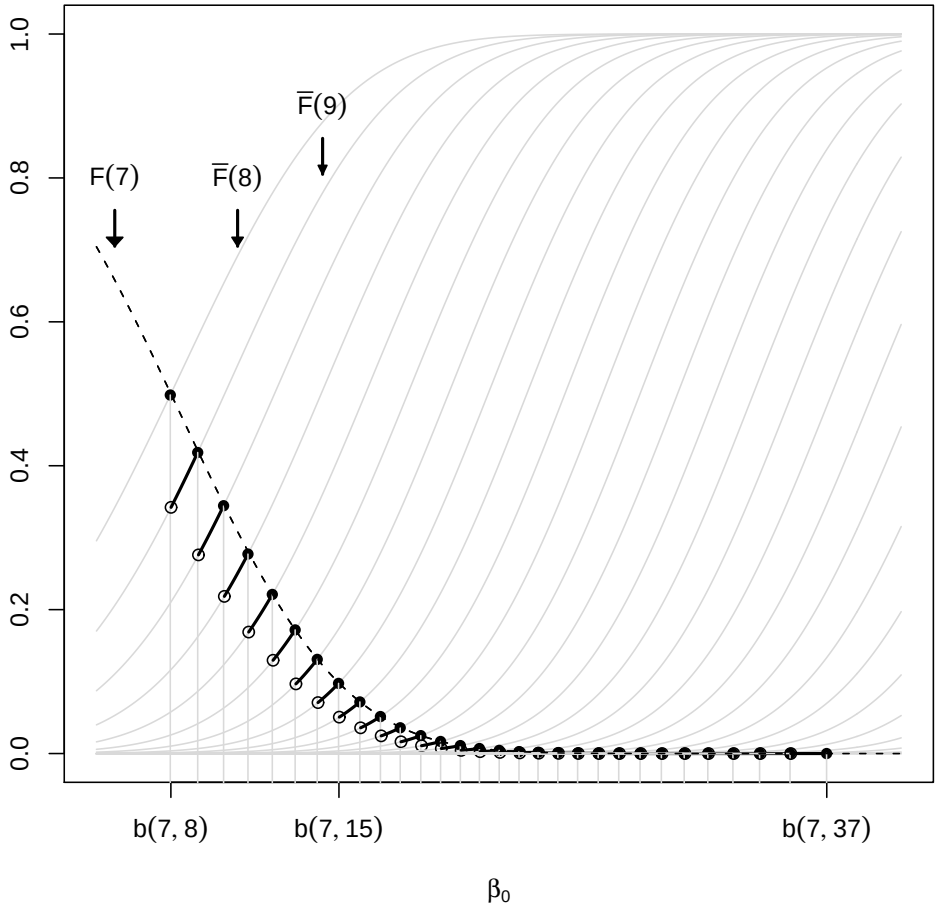

Figure 3: Diagram of Blaker p-values for Table 2. For example, in the interval $b(7,8) < b \leq b(7,9)$ the p-value is the sum of the $F_b(7)$ (the dotted line in that interval) and $\bar{F}_b(9)$ (the solid black line segment in that interval).

Table 1: Hypothetical 2×2 Data Examples

			Example 1		
	Event	No Event	two-sided Fisher	p=0.04371	CI=(0.0435,0.9170)
Group A	6	12	central Fisher	p=0.06059	CI=(0.0389,1.0565)
Group B	12	5	Blaker's exact	p=0.04371	CI=(0.0422,0.9170)
			Example 2		
	Event	No Event	two-sided Fisher	p=0.04999	CI=(0.1780,1.0182)
Group A	7	255	central Fisher	p=0.05322	CI=(0.1560, 1.0099)
Group B	30	466	Blaker's exact	p=0.04999	CI=(0.1683, 0.9983)
			Example 3		
	Event	No Event	two-sided Fisher	p=0.05001	CI=(0.1766, 1.0101)
Group A	7	257	central Fisher	p=0.05062	CI=(0.1548,1.0019)
Group B	30	466	Blaker's exact	p=0.05001	CI=(0.1670,1.0018)

For calculating bounds on the error in estimating $p_b(\mathbf{x})$, we first assume that the error in calculating $F_b(x)$ and $\bar{F}_b$ is small enough that it can be ignored, i.e., it is much smaller than the desired tolerance of the limits denoted δ. Because of the monotonicity in b of both $F_b(x)$ and $\bar{F}_b(x)$, for all $b \in (a_1, a_2)$ where $b(x_1, x_1 + j) < a_1 < a_2 \leq b(x_1, x_1 + j + 1)$, we have

$$\underline{\mathrm{P}}(a_1, a_2) \equiv F_{a_1}(x_1) + \bar{F}_{a_2}(x_1) \leq p_b(\mathbf{x}) \leq F_{a_2}(x_1) + \bar{F}_{a_1}(x_1) \equiv \bar{P}(a_1, a_2)$$

We can use these bounds to create an algorithm that can either find the confidence limits to within some pre-specified tolerance level, δ, or output bounds on those confidence limits. Here is an outline of an algorithm to calculate the upper Blaker confidence limit, a similar algorithm could be used for the lower confidence limit:

1. Set $i = 1$, $j = x_1 + x_0$, $N = N_{div}$, where N_{div} is a positive integer greater than 1.

2. Calculate $b_{low} = b(x_1, j)$ and $b_{hi} = b(x_1, j + 1)$ using a numeric root function (e.g., `uniroot` in R).

3. If $b_{hi} - b_{low} < \delta$, set the upper confidence interval equal to $b_{hi}/2 + b_{low}/2$ and stop. Otherwise continue.

4. If $\bar{P}\{b_{low}, b_{hi}\} \leq \alpha$, decrease j by 1 and go to step 2. If $\underline{\mathrm{P}}\{b_{low}, b_{hi}\} > \alpha$, set the upper confidence limit to b_{hi} and stop. Otherwise continue.

5. Divide up the interval $(b_{low}, b_{hi}]$ into N pieces where the ith piece is $(a_{i-1}, a_i]$ and $a_0 = b_{low}$ and $a_N = b_{hi}$. Calculate $\bar{P}$ and $\underline{\mathrm{P}}$ for each piece. If all the $\bar{P}$ values are less than or equal to α decrease j by 1 and go to step 2. If all the $\underline{\mathrm{P}}$ values are greater than α, set the upper confidence limit to b_{hi} and stop. Otherwise continue.

6. If any $\bar{P}(a_{hi-1}, a_{hi}) \leq \alpha$, set $b_{hi} = a_{hi}$ where a_{hi} is the minimum value such that $\bar{P}(a_{hi-1}, a_{hi}) \leq \alpha$. If any $\underline{\mathrm{P}}(a_{low}, a_{low+1}) > \alpha$ set $b_{low} = a_{low}$ where a_{low} is the minimum value such that $\underline{\mathrm{P}}(a_{low}, a_{low+1}) > \alpha$. Increase N to $2N$, and increase i by 1. If $i < I_{max}$ go to step 5, if not set the upper confidence limit as $b_{low}/2 + b_{hi}/2$ and output $(b_{low}, b_{hi}]$ as bounds on the limit and if $b_{hi} - b_{low} > \delta$ give a warning that the tolerance level was not reached.

For the matching interval to the two-sided Fisher exact test, we follow the same outline, except $b(x_a, x_b)$ is defined as

$$b(x_a, x_b) \quad = \quad \{\beta : f_\beta(x_a) = f_\beta(x_b)\}.$$

This works because the non-parametric hypergeometric distribution is unimodal in x_1 as can be shown by writing the ratio $f_\beta(x)/f_\beta(x+1)$ and showing that it is a monotone function of x (see e.g., Liao and Rosen, 2001).