Unpublished Manuscript , 1–16 DOI: see foootnote
18 September 2009

Confidence Intervals that Match Fisher's Exact or Blaker's Exact Tests[1]

Michael P. Fay

National Institute of Allergy and Infectious Diseases,

Bethesda, MD 20892-7609, U.S.A.

email: mfay@niaid.nih.gov

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.

Key words: Acceptability Function; Conditional Exact Test; Confidence Set; Fisher's exact test; Odds Ratio; Two-by-Two Table.

[1]This is an earlier version of a manuscript that has been published in Biostatistics as a paper with a supplement. There are no substantial differences in content between this draft and the Biostatistics version, but there was rewriting in order to present the paper as a short note with a supplement. The official version of the paper is the Biostatistics version and should be cited as follows: Fay, M.P. (2010). "Confidence Intervals that Match Fisher's Exact or Blaker's Exact Tests" Biostatistics. **11**: 373-374.

1. Introduction

We begin with a motivating data example. Lim, et al (2009) explore whether a certain genetic modification (CCR5 Defficiency) effects the probability of having clinical symptoms given infection with West Nile virus. They first show that there was a highly significant effect on the number of symptoms. Then Lim, et al (2009) present a table with 16 specific symptoms, and test each symptom for significance using a two-sided Fisher's exact test based on a genetic recessive model. They also show odds ratios with the 95% confidence limits based on the asymptotic normality of the log transformed odds ratio (see e.g., Agresti, 1990). We show 3 of the 16 symptoms in Table 1. Notice that Tremors where not significant at the 0.05 level by Fisher's exact test, but the 95% confidence interval does not contain 1 implying a significant effect. Thus, the same 2×2 table for a specific symptom gives conflicting indications of significance.

One might think that using the exact confidence intervals on the odds ratios would solve this problem; however, if the usual exact confidence intervals are used, we obtain conflicting significance for two other symptoms in the table, Vomitting/Diarrehea and Abdominal Pain. Thus, in this case the asymptotic confidence intervals have fewer conflicts with the two-sided Fisher's exact p-values than the usual exact ones. This usual exact confidence interval was derived by Cornfield (1956, see also Agresti and Min, 2001) and will be called the exact conditional tail interval (ECTI). It is the only exact confidence interval for this situation of which we are aware from standard statistical software; in particular, it is the only one given by SAS (version 9.2, Proc Freq), StatXact (StatXact 8 Procs), or in the base implementation of R (version 9.1, see fisher.test). The exact2x2 R package, developed concurrently with this paper, allows other exact confidence intervals, and here we discuss the theoretical implications, algorithms, and bounds on those other intervals.

Theoretically, a simple way to avoid these conflicts is to define a confidence set by inverting

the family of hypothesis tests (see e.g., Casella and Berger, 2002, p. 431). Applying this idea to the example, we consider the family of hypothesis test where each member tests the null hypothesis $H_0 : \beta = \beta_0$ for a different odds ratio β_0. Because Fisher's exact test can be extended for $\beta_0 \neq 1$, the 95% confidence set is easily defined as the set of all β_0 for which the resulting p-value from the two-sided extended Fisher's exact test fails to reject at the 0.05 level. The problem is that the resulting confidence set may be the union of two disjoint intervals. Even if we define the "matching" confidence interval as the smallest one that contains all members of the confidence set of the inversion, the calculation of this interval is not straightforward. That calculation is a main topic of this paper.

[Table 1 about here.]

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 the general acceptability function of Blaker (2000) to create an exact test with confidence intervals for the 2×2 table, which we call Blaker's exact test.

Blaker (2000) gives an algorithm for calculating confidence intervals using the acceptability function applied to a single binomial parameter, but we show here that the bounds on the precision from algorithms of that type are not clear. Baptista and Pike (1977) give an algorithm for calculating the confidence interval that is the inversion of the two-sided Fisher's exact test, although they did not mention the cases when the confidence set is not an interval, and their algorithm will have similar precision ambiguity. In this paper, we give an algorithm for which we may calculate the confidence interval to within some prespecified tolerance level.

Here is an outline of the paper. In Section 2 we introduce notation and outline the general

problem. In Section 3 we review three different exact tests for this situation. Section 4 we describe the difficulty in getting precision on the matching confidence intervals for two of those three tests, and propose an algorithm. In Section 5 we explore the extent of the conflicting inferences problem described above for general 2×2 tables, and show that it is not a rare problem. In Section 6 we return to the application of Table 1 and compare the three tests.

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:

$$H_0 : \beta = \beta_0$$

$$H_1 : \beta \neq \beta_0.$$

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 $p_{\beta_0}(\mathbf{x}) \leqslant \alpha$. 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):

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

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 the 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).

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, 2007).

3.2 *Central Fisher's Exact Test*

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

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

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

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

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 subsection but,

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

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, because of the p-value function of equation 5 may not be unimodal in β (see Section 4 below).

3.4 *Blaker's Exact Test*

An alternative method for creating a two-sided p-value is to add to the smaller of the one-sided p-values "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 define the *Blaker p-value* (see Blaker, 2000) as the two-sided p-value which adds to the smaller one-sided one the largest tail probability in the opposite tail which is less than or equal to the observed tail (see equations 6 or 7). 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 with 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, 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 reasons 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 $F_\beta(x) = Pr[X_1 \leqslant x \mid \beta]$, $\bar{F}_\beta(x) = Pr[X_1 \geqslant x \mid \beta]$, 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

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

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.

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

4.1 *Motivation*

To show the non-unimodality in β of the p-value function for the two-sided Fisher's exact test and Blaker's exact test, consider an invented example of a 2×2 table where one group has 7/262=2.67% (i.e., 7 events out of 262 at risk) while the other group has 30/494=6.07%. We plot the p-values for the three tests in Figure 1. The confidence set created by inverting the family of two-sided Fisher's exact tests is not a confidence interval: the resulting 95% confidence set is $\{\beta : \beta \in (0.177, 0.993)$ or $\beta \in (1.006, 1.014)\}$. Similar observations have

been made previously (see Blaker, 2000, Vos and Hudson, 2008). 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, $p_1(\mathbf{x}) = 0.04996$, but for slightly larger or smaller β_0 the p-value is not significant, $p_{1.01}(\mathbf{x}) = 0.05006$ and $p_{0.99}(\mathbf{x}) = 0.05005$. Blaker's exact test can also have this problem, although in this case $p_{1.01}(\mathbf{x}) = 0.0354$ for Blaker's test. The problem is the non-unimodality of the p-value function. Note that this non-unimodality is not a problem for the central Fisher's exact test, and that is why the ECTI are much easier to calculate and have been the default exact intervals in standard software.

[Figure 1 about here.]

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 the invented data mentioned above, and here we choose $\alpha = 0.0501$ to demonstrate a potential problem with the algorithm. The algorithm is to calculate $p_\beta(\mathbf{x})$ for different values of β at equal intervals, starting from an extreme value, moving towards $\beta = 1$, and stopping when $p_\beta(\mathbf{x}) > \alpha$. The points in Figure 2 show the two-sided Fisher's exact p-value calculated 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 1.0138 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 $1.0138 - 0.9860 = 0.0278$.

[Figure 2 about here.]

4.2 *Algorithm*

First consider the Blaker confidence interval. Recall that $p_\beta(\mathbf{x})$ for Blaker's exact test is the sum of two tails of Fisher's non-central hypergeometric distribution, the observed tail and the opposite tail that is closest to but not greater than the observed tail. Note that $F_\beta(x)$ is decreasing in β for all fixed x in $\{x_{min}, \ldots, x_{max} - 1\}$ (Mehta, Patel, and Gray, 1985), and it is increasing in x for $0 < \beta < \infty$ and fixed. Since $\bar{F}_\beta(x) = 1 - F_\beta(x-1)$, we have similar relationships but reversed directions for $\bar{F}_\beta(x)$. Thus, we can write $p_\beta(\mathbf{x})$ for Blaker's exact test as a series of segments, each of which is the sum of an increasing function of β plus a decreasing function of β. This allows us to calculate bounds on $p_\beta(\mathbf{x})$.

Here are the details. Let

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

with $b(x_1, x_{max}+1) \equiv \infty$ and $b(x_{min}-1, x_1) \equiv 0$, and let $F_\beta(x) = 0$ when $x < x_{min}$ and $\bar{F}_\beta(x) = 0$ when $x > x_{max}$. Then we can write Blaker's p-value function as

$$p_b(\mathbf{x}) = \begin{cases} F_b(x) + \bar{F}_b(x_1) & \text{for } b(x-1, x_1) < b \leqslant b(x, x_1); x = x_{min}, \ldots, x_1 - 1 \\ 1 & \text{for } b(x_1 - 1, x_1) \leqslant b \leqslant b(x_1, x_1 + 1) \\ F_b(x_1) + \bar{F}_b(x) & \text{for } b(x_1, x) \leqslant b < b(x_1, x+1); x = x_1 + 1, \ldots, x_{max} \end{cases} \tag{7}$$

Figure 3 helps to explain the Blaker p-value.

[Figure 3 about here.]

For calculating bounds on the error in estimating $p_b(\mathbf{x})$, we first assume that the error in calculating $F_b(x)$ and $\bar{F}_b(x)$ 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 $b(x_1, x_1 + j) < a_1 < a_2 \leqslant b(x_1, x_1 + j + 1)$, we have

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

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 $100(1-\alpha)$ percent Blaker confidence limit, say U:

(1) Set $i = 1$, $j = x_{max}$, $N = N_{div}$, where N_{div} is a positive integer greater than 1.

(2) If $x_1 = x_{max}$ then set $U = \infty$ and stop. Otherwise, calculate $b_{low} = b(x_1, j)$. If $F_{b_{low}}(x_1) > \alpha$ set U equal to the root, b, where $F_b(x_1) = \alpha$, which can be found using a numeric root function (e.g., uniroot in R).

(3) Let $b_{low} = b(x_1, j)$ and calculate $b_{hi} = b(x_1, j-1)$ using a numeric root function.

(4) If $b_{hi} - b_{low} < \delta$, set $U = b_{hi}/2 + b_{low}/2$ and stop. Otherwise continue.

(5) If $\bar{P}\{b_{low}, b_{hi}\} \leqslant \alpha$, decrease j by 1 and go to step (3). If $\underline{\mathrm{P}}\{b_{low}, b_{hi}\} > \alpha$, set $U = b_{hi}$ and stop. Otherwise continue.

(6) 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 (3). If all the $\underline{\mathrm{P}}$ values are greater than α, set $U = b_{hi}$ and stop. Otherwise continue.

(7) If any $\bar{P}(a_{\ell-1}, a_\ell) > \alpha$, set b_{hi} equal to the maximum of any a_ℓ such that $\bar{P}(a_{\ell-1}, a_\ell) > \alpha$. If any $\underline{\mathrm{P}}(a_{\ell-1}, a_\ell) > \alpha$ set b_{low} equal to the maximum a_ℓ of any a_ℓ such that $\underline{\mathrm{P}}(a_{\ell-1}, a_\ell) > \alpha$, otherwise b_{low} remains unchanged. Increase N to $2N$, and increase i by 1. If $i < I_{max}$ go to step (6), if not set $U = b_{low}/2 + b_{hi}/2$ 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.

A similar algorithm could be used for the lower confidence limit.

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).

5. The Extent of the Non-Matching Problem

In the data example from of Lim, et al (2009, see Table 1), there were 2 out of 16 cases where the two-sided Fisher's exact p-value implied different inferences at the 0.05 level than the ECTI. To see if this is a rare occurrence, we systematically check the frequency of this problem in this section.

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 $I_p \neq I_C$. Let the set of possible 2×2 tables for a given n_0 and n_1 be called an n-set. Within each n-set, we check and see if there are any mismatches, if so we say that the n-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 n-sets where n_0 and n_1 are each in $\{5, 6, \ldots, 20\}$. Of these n-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 n-sets of tables mentioned above through exhaustive search and find no mismatches. Thus,

although mismatches between p-values and confidence intervals are possible when using the matching confidence intervals (see Figure 1b), they are not common.

6. Application and Comparison of Methods

In Table 2 we give the odds ratios, two-sided p-values and matching confidence intervals to the data presented in Table 1. The odds ratios are the conditional maximum likelihood ones, rather than the sample odds ratios. Note that for $\alpha = 0.05$ there are no mismatches of inferences between the p-values and matching confidence intervals. Our primary recommendation for this paper is that when presenting both p-values and confidence intervals, you should use the matching confidence intervals.

Before giving secondary recommendations we review some 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.

As one might expect from Figure 1, small changes in the *data* can have large changes in the two-sided Fisher's exact p-value (see Dupont, 1986) or Blaker's exact p-value. Vos and Hudson (2008) emphasized a different point for other discrete tests, which we would like to emphasize for these two tests. It is possible that small changes in the data in the direction *away* from the null can lead to *less* significant tests. Let us modify the invented data so that 2 more individuals were observed with no events in the first group, giving proportions: 7/264=2.65% and 30/494=6.07%. This is clearly further away from the null than the original example, since the first group, 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 ($p_1(\mathbf{x}) = 0.04996$) to non-significance with the 2 added individuals ($p_1(\mathbf{x}) = 0.05005$). Although the same phenomena can occur with Blaker's exact test and often the Blaker p-values equal those of the two-sided Fisher's exact test, in this modified example the Blaker p-value is different, ($p_1(\mathbf{x}) = 0.0356$). Unlike the two-sided Fisher's exact p-values, the p-values from the central Fisher's exact test properly show the ordering, giving a larger p-value for the original data set ($p_1(\mathbf{x}) = 0.0518$) than the modified one ($p_1(\mathbf{x}) = 0.0493$).

[Table 2 about here.]

7. Discussion

We recommend that whenever 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 $p = 0.0499$ and $p = 0.0501$ are much closer in terms of significance than the pair with $p = 0.0499$ and $p = 0.0001$.

Acknowledgements

The author thanks Dean Follmann for providing the data example and Mike Proschan for discussions on the non-central hypergeometric distribution.

References

Agresti, A. (1990). *Categorical Data Analysis* Wiley: New York.

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.

Cornfield, J. (1956). "A statistical problem arising from retrospective studies." *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability* **4:** 135-148.

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.

Lim, J.K., Mcdermott, D.H., Lisco, A., Foster, G.A., Krysztof, D., Follmann,D., Stramer, S.L., Murphy, P.M. (2009). "CCR5 Deficiency is a Strong Risk Factor for Clinical Illness in the Early Stages of West Nile Virus Infection." (unpublished manuscript).

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.

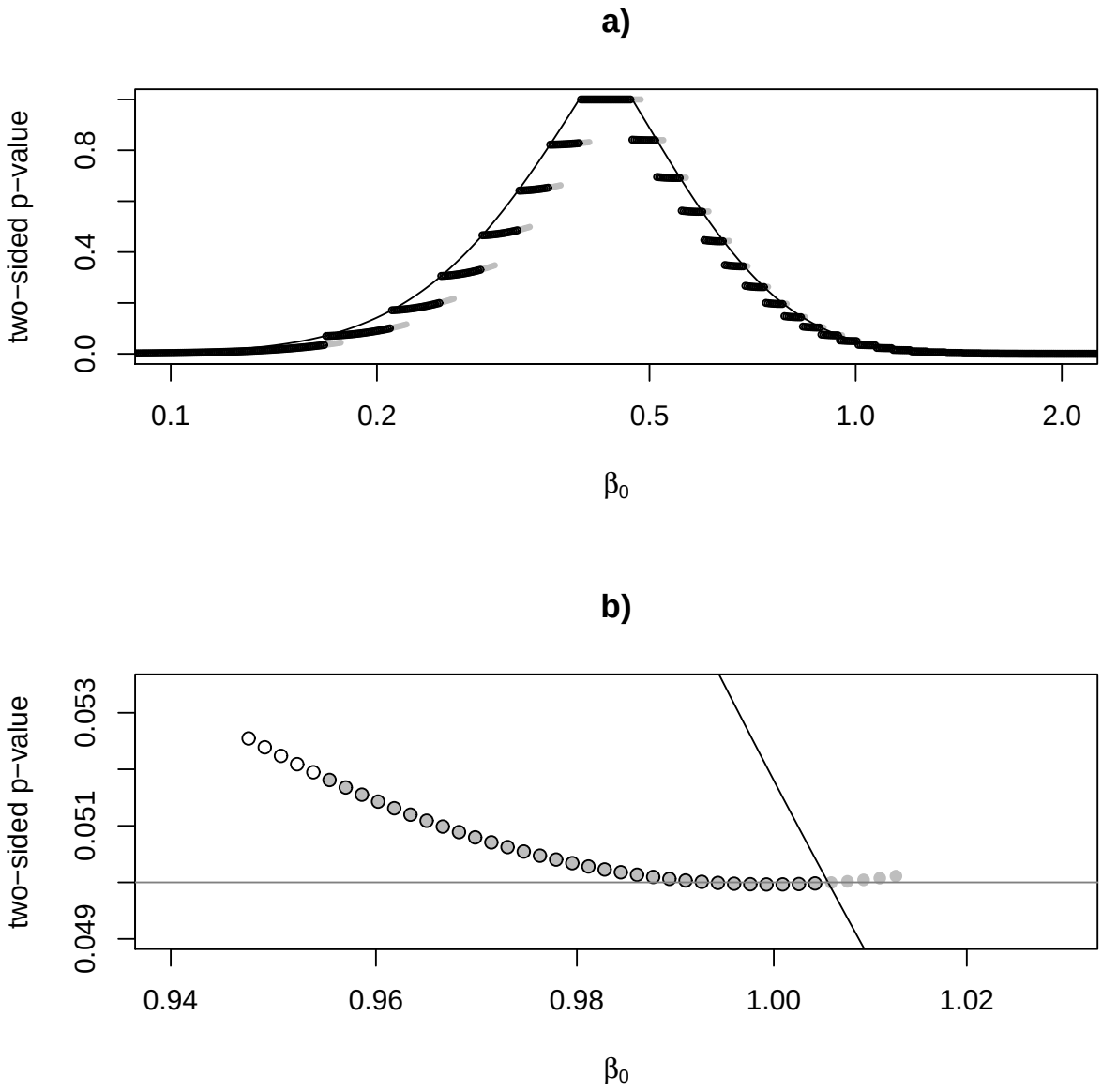

Figure 1. P-values from the three two-sided exact tests for testing $\beta = \beta_0$ for different values of β_0. Solid gray dots (appearing as thick gray lines in Figure a) are two-sided Fisher's exact test, black open dots (appearing as thick black lines in Figure a) are Blaker's exact test, and gray dots outlined in black (appearing as thick black lines in Figure a) are where those two p-values are equal. The thin black line is the central Fisher's exact, the horizontal gray line is the reference line at 0.05. Figure b is a blow-up of a portion of Figure a.

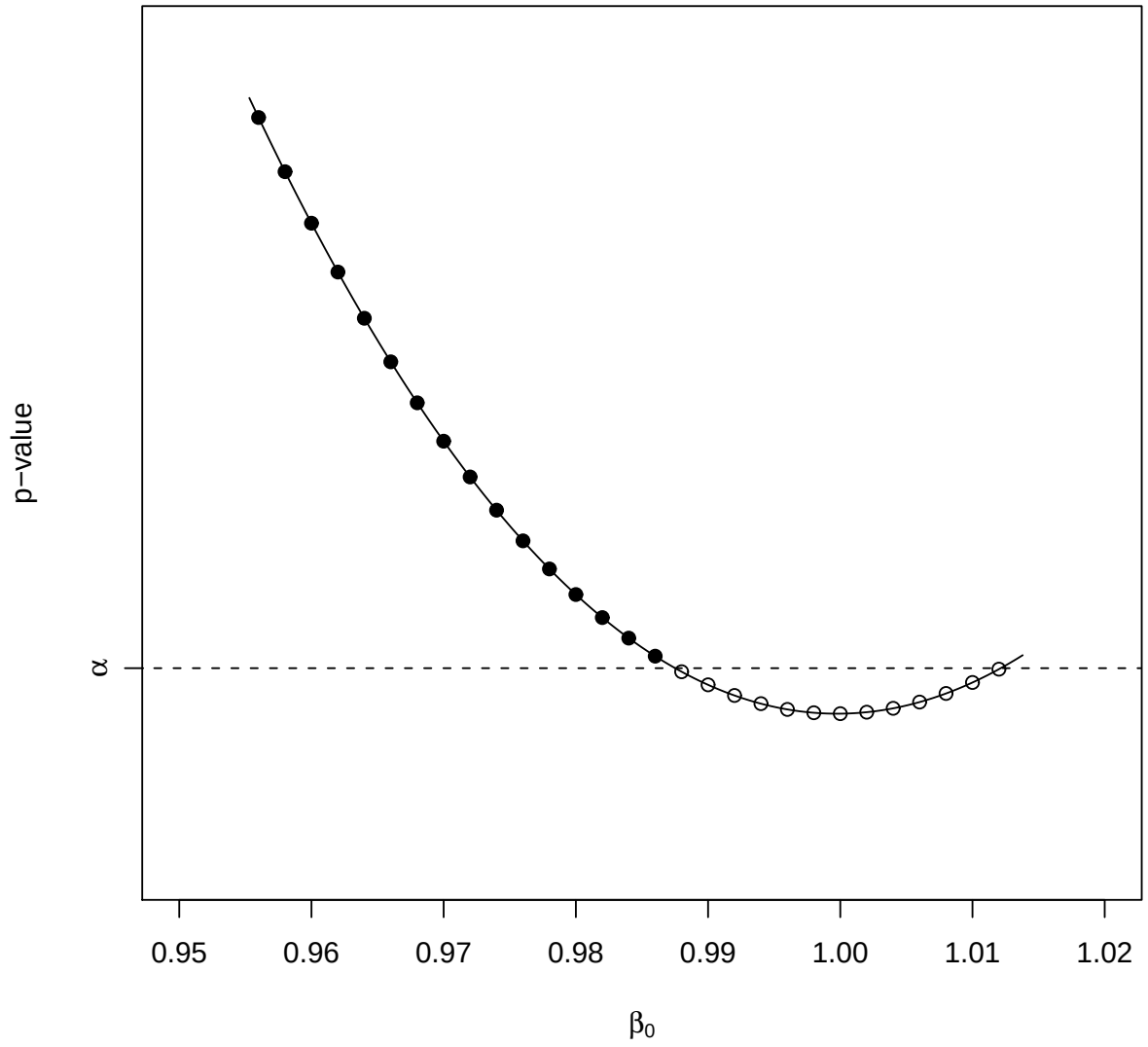

Figure 2. Figure to show difficulty with Blaker's algorithm. P-values evaluated at the points, $1 \pm j * 0.002, j = 0, 1, \ldots$.

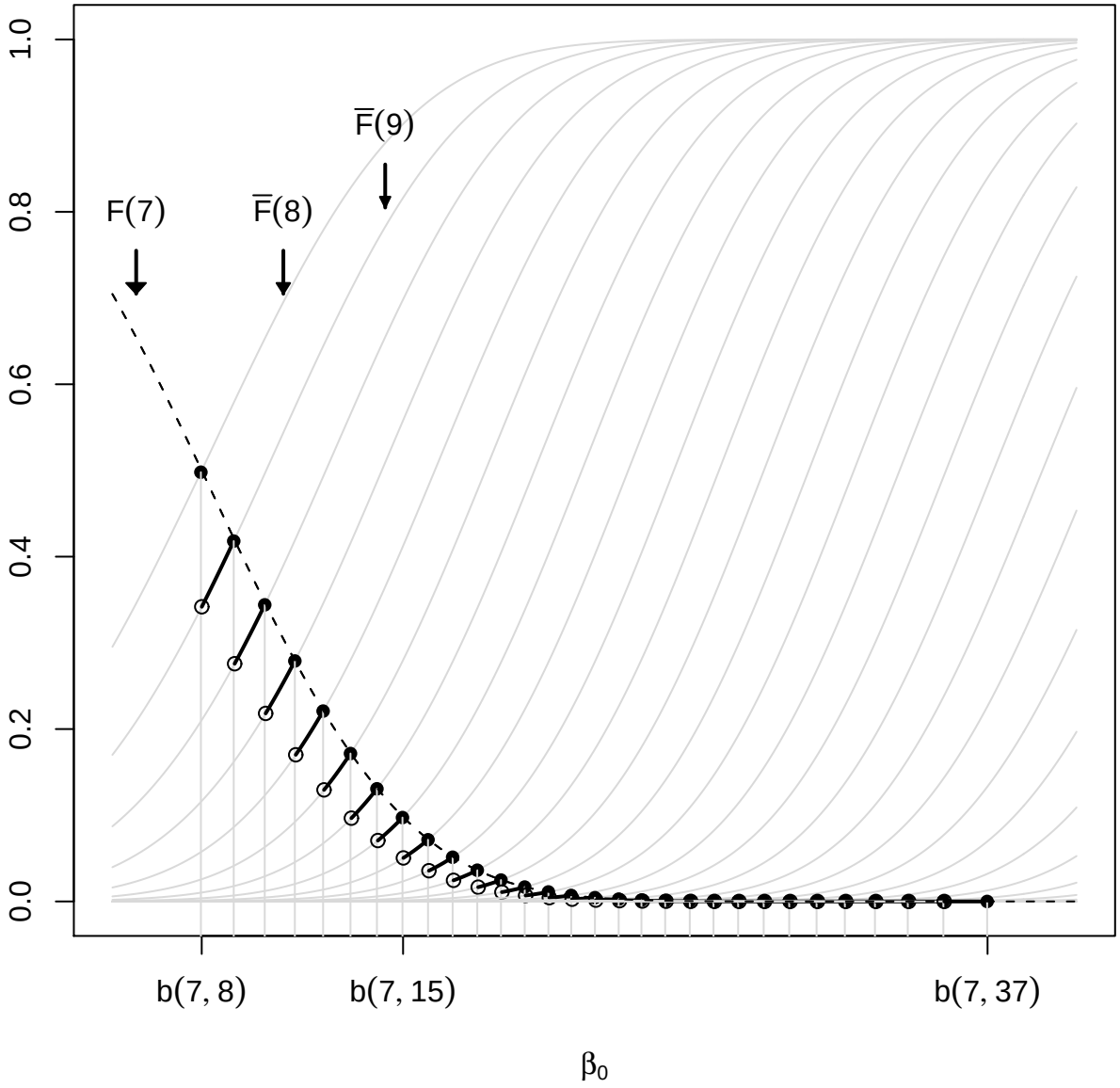

Figure 3. Diagram of Blaker p-values for invented table. For example, in the interval $b(7,8) < b \leqslant b(7,9)$ the p-value is the sum of the $F_b(7)$ (the dotted line in that interval) and $\bar{F}_b(9)$ (the solid black line segment in that interval).

Table 1
Standard Analyses on Data from Lim, et al (2009)

	Table Counts				2-sided	Odds	95% Asy		95% Exact	
Symptom	HY*	WY	HN	WN	p-value	Ratio	C.I.		C.I.	
Tremors	1	4	14	615	0.113	10.98	1.15	104.66	0.21	119.89
Vomitting/Diarrhea	5	78	10	541	0.035	3.47	1.16	10.41	0.90	11.46
Abdominal Pain	4	50	11	569	0.032	4.14	1.27	13.47	0.92	14.58

* HY=Homozygous for CCR5 deficiency, with symptom; WY= Wild type and Heterozygous for CCR5 deficiency, with symptom; HN and WN are subjects with genetics similarly defined but without symptoms. Odds ratio is the sample odds ratio. The p-value is from the usual two-sided Fisher's exact test. The 95% asymptotic confidence interval uses the log transformation and the delta method and the 95% exact confidence interval is the ECTI.

Table 2
Exact Two-sided Tests with Matching Confidence Intervals on Data from Lim, et al (2009)

	Odds	Fisher's Two-Sided			Central Fisher			Blaker's Exact		
symptom	Ratio*	p	95% C.I.		p	95% C.I.		p	95% C.I.	
Tremors	10.85	0.113	0.42	89.89	0.226	0.21	119.89	0.113	0.42	89.89
Vomitting/Diarrhea	3.46	0.035	1.11	11.14	0.071	0.90	11.46	0.035	1.11	11.27
Abdominal Pain	4.12	0.032	1.17	14.17	0.063	0.92	14.58	0.032	1.17	14.22

* Odds ratio calculated by conditional maximum likelihood.