Two-sided Exact Tests and Matching Confidence Intervals for Discrete Data

by Michael P. Fay

Abstract There is an inherent relationship between two-sided hypothesis tests and confidence intervals. A series of two-sided hypothesis tests may be inverted to obtain the *matching* 100(1-α)% confidence interval defined as the smallest interval that contains all point null parameter values that would not be rejected at the α level. Unfortunately, for discrete data there are several different ways of defining two-sided exact tests and the most commonly used two-sided exact tests are defined one way, while the most commonly used exact confidence intervals are inversions of tests defined another way. This can lead to inconsistencies where the exact test rejects but the exact confidence interval contains the null parameter value. The packages **exactci** and **exact2x2** provide several exact tests with the matching confidence intervals avoiding these inconsistencies as much as possible. Examples are given for binomial and Poisson parameters and both paired and unpaired 2×2 tables.

Applied statisticians are increasingly being encouraged to report confidence intervals (CI) and parameter estimates along with p-values from hypothesis tests. The `htest` class of the **stats** package is ideally suited to these kinds of analyses, because all the related statistics may be presented when the results are printed. For exact two-sided tests applied to discrete data, a test-CI inconsistency may occur: the p-value may indicate a significant result at level α while the associated 100(1-α)% confidence interval may cover the null value of the parameter. Ideally, we would like to present a unified report (Hirji, 2006), whereby the p-value and the confidence interval match as much as possible.

A motivating example

I was asked to help design a study to determine if adding a new drug (*albendazole*) to an existing treatment regimen (*ivermectin*) for the treatment of a parasitic disease (*lymphatic filariasis*) would increase the incidence of a rare serious adverse event when given in an area endemic for another parasitic disease (*loa loa*). There are many statistical issues related to that design (Fay et al., 2007), but here consider a simple scenario to highlight the point of this paper. A previous mass treatment using the existing treatment had 2 out of 17877 experiencing the serious adverse event (SAE) giving an observed rate of 11.2 per 100,000. Suppose the new treatment was given to 20,000 new subjects and suppose that 10 subjects experienced the SAE giving an observed rate of 50 per 100,000. Assuming Poisson rates, an exact test using `poisson.test(c(2,10),c(17877,20000))` from the **stats** package (throughout we assume version 2.11.0 for the stats package) gives a p-value of $p = 0.0421$ implying significant differences between the rates at the 0.05 level, but `poisson.test` also gives a 95% confidence interval of $(0.024, 1.050)$ which contains a rate ratio of 1, implying no significant difference. We return to the motivating example in the 'Poisson two-sample' section later.

Overview of two-sided exact tests

We briefly review inferences using the p-value function for discrete data. For details see Hirji (2006) or Blaker (2000). Suppose you have a discrete statistic t with random variable T such that larger values of T imply larger values of a parameter of interest, θ. Let $F_\theta(t) = Pr[T \leq t; \theta]$ and $\bar{F}_\theta(t) = Pr[T \geq t; \theta]$. Suppose we are testing

$$H_0 : \theta \geq \theta_0$$
$$H_1 : \theta < \theta_0$$

where θ_0 is known. Then smaller values of t are more likely to reject and if we observe t, then the probability of observing equal or smaller values is $F_{\theta_0}(t)$ which is the one-sided p-value. Conversely, the one-sided p-value for testing $H_0 : \theta \leq \theta_0$ is $\bar{F}_{\theta_0}(t)$. We reject when the p-value is less than or equal to the significance level, α. The one-sided confidence interval would be all values of θ_0 for which the p-value is greater than α.

We list 3 ways to define the two-sided p-value for testing $H_0 : \theta = \theta_0$, which we denote p_c, p_m and p_b for the `central`, `minlike`, and `blaker` methods, respectively:

central: p_c is 2 times the minimum of the one-sided p-values bounded above by 1, or mathematically,

$$p_c = \min\left\{1, 2 \times \min\left(F_{\theta_0}(t), \bar{F}_{\theta_0}(t)\right)\right\}.$$

The name `central` is motivated by the associated inversion confidence intervals which are central intervals, i.e., they guarantee that the lower (upper) limit of the 100(1-α)% confidence interval has less than $\alpha/2$ probability of being greater (less) than the true parameter. This is called the TST (twice the smaller tail method) by Hirji (2006).

minlike: p_m is the sum of probabilities of outcomes with likelihoods less than or equal to the observed likelihood, or

$$p_m = \sum_{T:f(T)\leq f(t)} f(T)$$

where $f(t) = Pr[T = t; \theta_0]$. This is called the PB (probability based) method by Hirji (2006).

blaker: p_b combines the probability of the smaller observed tail with the smallest probability of the opposite tail that does not exceed that observed tail probability. Blaker (2000) showed that this p-value may be expressed as

$$p_b = Pr[\gamma(T) \leq \gamma(t)]$$

where $\gamma(T) = \min\{F_{\theta_0}(T), \bar{F}_{\theta_0}(T)\}$. The name `blaker` is motivated by Blaker (2000) which comprehensively studies the associated method for confidence intervals, although the method had been mentioned in the literature earlier, see e.g., Cox and Hinkley (1974), p. 79. This is called the CT (combined tail) method by Hirji (2006).

There are other ways to define two-sided p-values, such as defining extreme values according to the score statistic (see e.g., Hirji (2006, Chapter 3), or Agresti and Min (2001)). Note that $p_c \geq p_b$ for all cases, so that p_b gives more powerful tests than p_c. On the other hand, although generally $p_c > p_m$ it is possible for $p_c < p_m$.

If $p(\theta_0)$ is a two-sided p-value testing $H_0 : \theta = \theta_0$, then its $100(1-\alpha)\%$ matching confidence interval is the smallest interval that contains all θ_0 such that $p(\theta_0) > \alpha$. To calculate the matching confidence intervals, we consider only regular cases where $F_\theta(t)$ and $\bar{F}_\theta(t)$ are monotonic functions of θ (except perhaps the degenerate cases where $F_\theta(t) = 1$ or $\bar{F}_\theta(t) = 0$ for all θ when t is the maximum or minimum). In this case the matching confidence limits to the `central` test are (θ_L, θ_U) which are solutions to:

$$\alpha/2 = \bar{F}_{\theta_L}(t)$$

and

$$\alpha/2 = F_{\theta_U}(t)$$

except when t is the minimum or maximum, in which case the limit is set at the appropriate extreme of the parameter space. The matching confidence intervals for p_m and p_b require a more complicated algorithm to ensure precision of the confidence limits (Fay, 2010).

If matching confidence intervals are used then test-CI inconsistencies will not happen for the `central` method, and will happen very rarely for the `minlike` and `blaker` methods. We discuss those rare test-CI inconsistencies in the 'Unavoidable inconsistencies' section later, but the main point of this article is that it is not rare for p_m to be inconsistent with the `central` confidence interval (Fay, 2010) and that particular test-CI combination is the default for many exact tests in the **stats** package. We show some examples of such inconsistencies in the following sections.

Binomial: one-sample

If X is binomial with parameters n and θ, then the `central` exact interval is the Clopper-Pearson confidence interval. These are the intervals given by `binom.test`. The p-value given by `binom.test` is p_m. The matching interval to the p_m was proposed by Stern (1954) (see Blaker (2000)).

When $\theta_0 = 0.5$ we have $p_c = p_m = p_b$, and there is no chance of a test-CI inconsistency even when the confidence intervals are not inversions of the test as is the case in `binom.test`. When $\theta_0 \neq 0.5$ there may be problems. We explore these cases in the 'Poisson: two-sample' section later, since the associated tests reduce through conditioning to one-sample binomial tests.

Note that there is a theoretically proven set of shortest confidence intervals for this problem. These are called the Blyth-Still-Casella intervals in StatXact (StatXact Procs Version 8). The problem with these shortest intervals is that they are not nested, so that one could have a parameter value that is included in the 90% confidence interval but not in the 95% confidence interval (see Theorem 2 of Blaker (2000)). In contrast, the matching intervals of the `binom.exact` function of the **exactci** will always give nested intervals.

Poisson: one-sample

If X is Poisson with mean θ, then `poisson.test` from **stats** gives the exact `central` confidence intervals (Garwood, 1936), while the p-value is p_m. Thus, we can easily find a test-CI inconsistency: `poisson.test(5, r=1.8)` gives a p-value of $p_m = 0.036$ but the 95% central confidence interval of $(1.6, 11.7)$ contains the null rate of 1.8. As θ gets large the Poisson distribution may be approximated by the normal distribution and these test-CI inconsistencies become more rare.

The **exactci** package contains the `poisson.exact` function, which has options for each of the three methods of defining p-values and gives matching confidence intervals. The code `poisson.exact(5, r=1.8, tsmethod="central")` gives the same confidence interval as above, but a p-value of $p_c = 0.073$; while `poisson.exact(5, r=1.8, tsmethod="minlike")` returns a p-value equal to p_m, but a 95% confidence interval of $(2.0, 11.8)$. Finally, using `tsmethod="blaker"` we get $p_b = 0.036$ (it is not uncommon for p_b to equal p_m) and a 95% confidence interval of $(2.0, 11.5)$. We see that there is no test-CI inconsistency when using the matching confidence intervals.

Poisson: two-sample

For the control group, let the random variable of the counts be Y_0, the rate be λ_0 and the population at risk be m_0. Let the corresponding values for the test group be Y_1, λ_1 and m_1. If we condition on $Y_0 + Y_1 = N$ then the distribution of Y_1 is binomial with parameters N and

$$\theta = \frac{m_1 \lambda_1}{m_0 \lambda_0 + m_1 \lambda_1}$$

This parameter may be written in terms of the ratio of rates, $\rho = \lambda_1 / \lambda_2$ as

$$\theta = \frac{m_1 \rho}{m_0 + m_1 \rho}$$

or equivalently,

$$\rho = \frac{m_0 \theta}{m_1 (1 - \theta)}. \tag{1}$$

Thus, the null hypothesis that $\lambda_1 = \lambda_0$ is equivalent to $\rho = 1$ or $\theta = m_1 / (m_0 + m_1)$, and confidence intervals for θ may be transformed into confidence intervals for ρ by Equation 1. So the inner workings of the `poisson.exact` function when dealing with two-sample tests simply use the `binom.exact` function and transform the results using Equation 1.

Let us return to our motivating example (i.e., testing the difference between observed rates 2/17877 and 10/20000). As in the other sections, the results from `poisson.test` output p_m but the 95% central confidence intervals, as we have seen, give a test-CI inconsistency. The `poisson.exact` function avoids such test-CI inconsistency in this case by giving the matching confidence interval; here are the results of the three `tsmethod` options:

tsmethod	p-value	95% confidence interval
central	0.061	(0.024, 1.050)
minlike	0.042	(0.035, 0.942)
blaker	0.042	(0.035, 0.936)

Analysis of 2×2 tables, unpaired

The 2×2 table may be created from many different designs, consider first designs where there are two groups of observations with binary outcomes. If all the observations are independent, even if the number in each group is not fixed in advance, proper inferences may still be obtained by conditioning on those totals (Lehmann and Romano, 2005). Fay (2010) considers the 2×2 table with independent observations, so we only briefly present his motivating example here. The usual two-sided application of Fisher's exact test: `fisher.test(matrix(c(4,11,50,569), 2, 2))` gives $p_m = 0.032$ using the `minlike` method, but 95% confidence interval on the odds ratio of $(0.92, 14.58)$ using the `central` method. As with the other examples, the test-CI inconsistency disappears when we use either the `exact2x2` or `fisher.exact` function from the **exact2x2** package.

Analysis of 2×2 tables, paired

The case not studied in Fay (2010) is when the data are paired, the case which motivates McNemar's test. For example, suppose you have twins randomized to two treatment groups (Test and Control) then test on a binary outcome (pass or fail). There are 4 possible outcomes for each pair: (a) both twins fail, (b) the twin in the control group fails and the one in the test group passes, (c) the twin on the test group fails and the one in the control group passes, or (d) both twins pass. Here is a table where the numbers of sets of twins falling in each of the four categories are denoted a, b, c and d:

	Test	
Control	Fail	Pass
Fail	a	b
Pass	c	d

In order to test if the treatment is helpful, we use only the numbers of discordant pairs of twins, b and c, since the other pairs of twins tell us nothing about whether the treatment is helpful or not. McNemar's test statistic is

$$Q \equiv Q(b,c) = \frac{(b-c)^2}{b+c}$$

which for large samples is distributed like a chi-squared distribution with 1 degree of freedom. A closer approximation to the chi-squared distribution uses a continuity correction:

$$Q_C \equiv Q_C(b,c) = \frac{(|b-c|-1)^2}{b+c}$$

In R this test is given by the function `mcnemar.test`.

Case-control data may be analyzed this way as well. Suppose you have a set of people with some rare disease (e.g., a certain type of cancer); these are called the cases. For this design you match each case with a control who is as similar as feasible on all important covariates except the exposure of interest. Here is a table:

	Exposed	
Not Exposed	Control	Case
Control	a	b
Case	c	d

For this case as well we can use Q or Q_C to test for no association between case/control status and exposure status.

For either design, we can estimate the odds ratio by b/c, which is the maximum likelihood estimate (see Breslow and Day (1980), p. 165). Consider some hypothetical data (chosen to highlight some points):

	Test	
Control	Fail	Pass
Fail	21	9
Pass	2	12

When we perform McNemar's test with the continuity correction we get $p = 0.070$ while without the correction we get $p = 0.035$. Since the inferences are on either side of the traditional 0.05 cutoff of significance, it would be nice to have an exact version of the test to be clearer about significance at the 0.05 level. From the **exact2x2** package using `mcnemar.exact` we get the exact McNemar's test p-value of $p = .065$. We now give the motivation for the exact version of the test.

After conditioning on the total number of discordant pairs, $b + c$, we can treat the problem as $B \sim Binomial(b + c, \theta)$, where B is the random variable associated with b. Under the null hypothesis $\theta = 0.5$. We can transform the parameter θ into an odds ratio by

$$\text{Odds Ratio} \equiv \phi = \frac{\theta}{1-\theta} \qquad (2)$$

(Breslow and Day (1980), p. 166). Since it is easy to perform exact tests on a binomial parameter, we can perform exact versions of McNemar's test internally using the `binom.exact` function of the package **exactci** then transform the results into odds ratios via Equation 2. This is how the calculations are done in the `exact2x2` function when `paired=TRUE`. The `alternative` and the `tsmethod` options work in the way one would expect. So although McNemar's test was developed as a two-sided test testing the null that $\theta = 0.5$ (or equivalently $\phi = 1$), we can easily extend this to get one-sided exact McNemar-type Tests. For two-sided tests we can get three different versions of the two-sided exact McNemar's p-value function using the three `tsmethod` options, but all three are equivalent to the exact version of McNemar's test when testing the usual null that $\theta = 0.5$ (see the Appendix in `vignette("exactMcNemar")` in **exact2x2**). If we narrowly define McNemar's test as only testing the null that $\theta = 0.5$ as was done in the original formulation, there is only one exact McNemar's test; it is only when we generalize the test to test null hypotheses of $\theta = \theta_0 \neq 0.5$ that there are differences between the three methods. Those differences between the `tsmethod` options become apparent in the calculation of the confidence intervals. The default is to use `central` confidence intervals so that they guarantee that the lower (upper) limit of the 100(1-α)% confidence interval has less than $\alpha/2$ probability of being greater (less) than the true parameter. These guarantees on each tail are not true for the `minlike` and `blaker` two-sided confidence intervals; however, the latter give generally tighter confidence intervals.

Graphing P-values

In order to gain insight as to why test-CI inconsistencies occur, we can plot the p-value function. This type of plot explores one data realization and its many associated p-values on the vertical axis representing a series of tests modified by changing the point null hypothesis parameter (θ_0) on the horizontal axis. There is a default plot command for `binom.exact`, `poisson.exact`, and `exact2x2` that plots the p-value as a function of the point null hypotheses, draws vertical lines at the confidence limits, draws a line at 1 minus the confidence level, and adds a point at the null hypothesis of interest. Other plot functions (`exactbinomPlot`, `exactpoissonPlot`, and `exact2x2Plot`) can be used to add to that plot for comparing different methods. In Figure 1 we create such a plot for the motivating example. Here is the code to create that figure:

```
x <- c(2, 10)
n <- c(17877, 20000)
poisson.exact(x, n, plot = TRUE)
exactpoissonPlot(x, n, tsmethod = "minlike",
    doci = TRUE, col = "black", cex = 0.25,
    newplot = FALSE)
```

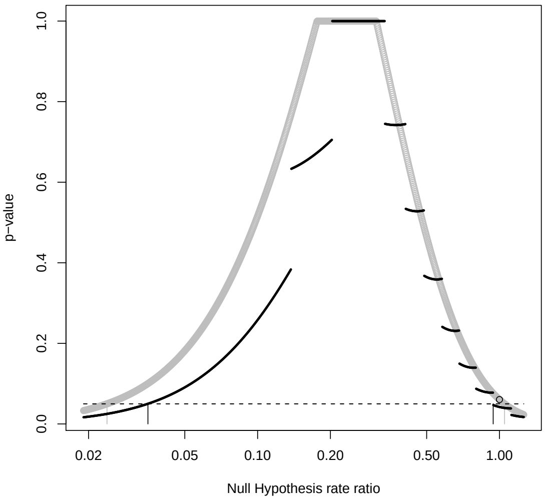

Figure 1: Graph of p-value functions for motivating two-sample Poisson example. Gray is `central` method, black is `minlike` method. Vertical lines are 95% confidence limits, black circle is `central` p-value at null rate ratio of 1.

We see from Figure 1 that the `central` method has smoothly changing p-values, while the `minlike` method has discontinuous ones. The usual confidence interval is the inversion of the `central` method (the limits are the vertical gray lines, where the dotted line at the significance level intersects with the gray p-values), while the usual p-value at the null that the rate ratio is 1 is where the black line is. To see this more clearly we plot the lower right hand corner of Figure 1 in Figure 2.

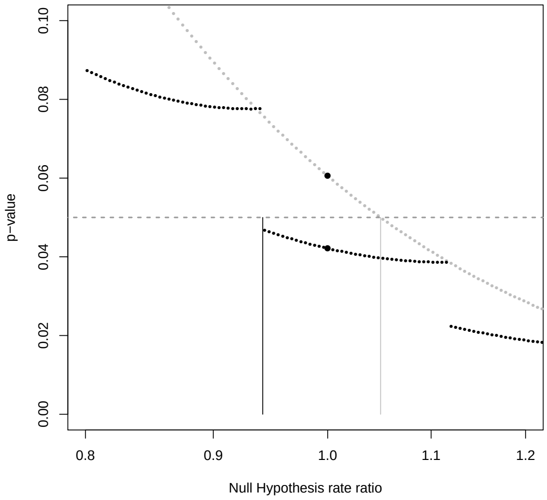

Figure 2: Graph of p-value functions for motivating two-sample Poisson example. Gray is `central` method, black is `minlike` method. Vertical lines are upper 95% confidence limits, solid black circles are the respective p-values at null rate ratio of 1.

From Figure 2 we see why the test-CI inconsistencies occur, the `minlike` method is generally more powerful than the `central` method, so that is why the p-values from the `minlike` method can reject a specific null when the confidence intervals from the `central` method imply failing to reject that same null. We see that in general if you use the matching confidence interval to the p-value, there will not be test-CI inconsistencies.

Unavoidable Inconsistencies

Although the **exactci** and **exact2x2** packages do provide a unified report in the sense described in Hirji (2006), it is still possible in rare instances to obtain test-CI inconsistencies when using the `minlike` or `blaker` two-sided methods (Fay, 2010). These rare inconsistencies are unavoidable due to the nature of the problem rather than any deficit in the packages.

To show the rare inconsistency problem using the motivating example, we consider the unrealistic situation where we are testing the null hypothesis that the rate ratio is 0.93 at the 0.0776 level. The corresponding confidence interval would be a 92.24% $= 100 \times (1 - 0.0776)$ interval. Using `poisson.exact(x, n, r=.93, tsmethod="minlike", conf.level=1-0.0776)` we reject the null (since $p_m = 0.07758 < 0.0776$) but the 92.24% matching confidence interval contains the null rate ratio of 0.93. In this situation, the confidence set that is the inversion of the series of tests is two disjoint intervals ($[0.0454, 0.9257]$ and $[0.9375, 0.9419]$), and the matching confidence interval fills in the hole in the confidence set. This is an unavoidable test-CI inconsistency. Figure 3 plots the situation.

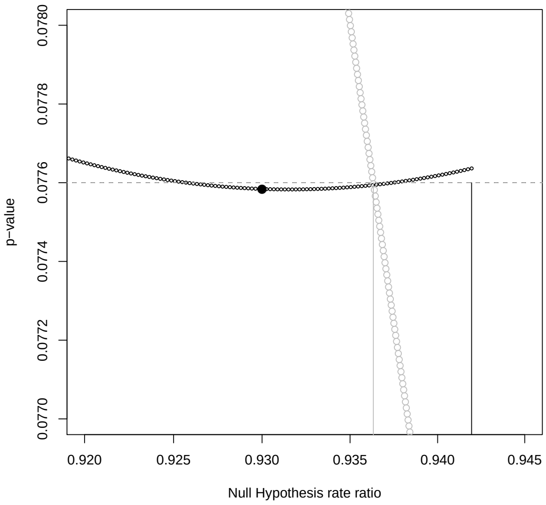

Figure 3: Graph of p-value functions for motivating two-sample Poisson example. Gray is `central` method, black is `minlike` method. Circles are p-values, vertical lines are the upper 92.24% confidence limits, solid black circle is `minlike` p-value at null rate ratio of 0.93. The unrealistic significance level of 0.0776 and null rate ratio of 0.93 are chosen to highlight the rare unavoidable inconsistency problem.

Additionally, the options `tsmethod="minlike"` or `"blaker"` can have other anomalies (see Vos and Hudson (2008) for the single sample binomial case, and Fay (2010) for the two-sample binomial case). For example, the data reject, but fail to reject if an additional observation is added *regardless* of the value of the additional observation. Thus, although the power of the `blaker` (or `minlike`) two-sided method is always (almost always) greater than the `central` two-sided method, the `central` method does avoid all test-CI inconsistencies and the previously mentioned anomalies.

Discussion

We have argued for using a unified report whereby the p-value and the confidence interval are calculated from the same p-value function (also called the evidence function or confidence curve). We have provided several practical examples. Although the theory of these methods have been extensively studied (Hirji, 2006), software has not been readily available. The **exactci** and **exact2x2** packages fill this need. We know of no other software that provides the `minlike` and `blaker` confidence intervals, except the **PropCIs** package which provides the Blaker confidence interval for the single binomial case only.

Finally, we briefly consider closely related software. The **rateratio.test** package does the two-sample Poisson exact test with confidence intervals using the `central` method. The **PropCIs** package does several different asymptotic confidence intervals, as well as the Clopper-Pearson (i.e. `central`) and Blaker exact intervals for a single binomial proportion. The **PropCIs** package also performs the mid-p adjustment to the Clopper-Pearson confidence interval which is not currently available in **exactci**. Other exact confidence intervals are not covered in the current version of **PropCIs** (Version 0.1-6). The **coin** and **perm** packages give very general methods for performing exact permutation tests, although neither perform the exact matching confidence intervals for the cases studied in this paper.

I did not perform a comprehensive search of commercial statistical software; however, SAS (Version 9.2) (perhaps the most comprehensive commercial statistical software) and StatXact (Version 8) (the most comprehensive software for exact tests) both do not implement the `blaker` and `minlike` confidence intervals for binomial, Poisson and 2x2 table cases.

Bibliography

A. Agresti and Y. Min. On small-sample confidence intervals for parameters in discrete distributions. *Biometrics*, 57:963–971, 2001.

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

N. Breslow and N. Day. *Statistical Methods in Cancer Research: Volume 1: Analysis of Case Control Studies*. International Agency for Research in Cancer, Lyon, France, 1980.

D. Cox and D. Hinkley. *Theoretical Statistics*. Chapman and Hall, London, 1974.

M. Fay. Confidence intervals that match Fisher's exact or Blaker's exact tests. *Biostatistics*, 11:373–374, 2010.

M. Fay, C. Huang, and N. Twum-Danso. Monitoring rare serious adverse events from a new treatment and testing for a difference from historical controls. *Controlled Clinical Trials*, 4:598–610, 2007.

F. Garwood. Fiducial limits for the Poisson distribution. *Biometrika*, pages 437–442, 1936.

K. Hirji. *Exact Analysis of Discrete Data*. Chapman and Hall/CRC, New York, 2006.

E. Lehmann and J. Romano. *Testing Statistical Hypotheses, third edition*. Springer, New York, 2005.

T. Stern. Some remarks on confidence and fiducial limits. *Biometrika*, pages 275–278, 1954.

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

Michael P. Fay
National Institute of Allergy and Infectious Diseases
6700-A Rockledge Dr. Room 5133, Bethesda, MD 20817
USA
`mfay@niaid.nih.gov`