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 a different 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 is possible. Examples are given for binomial and Poisson parameters and the 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 for 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 is 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 2.10.1 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 differences. We return to the motivating example in the 'Poisson two-sample' section below.

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, $p_c = \min\left\{1, 2 * \min\left(F_{\theta_0}(t), \bar{F}_{\theta_0}(t)\right)\right\}$. The name `central` is motivated by the associated inversion confidence intervals which are central intervals, i.e., they guarantee that the true parameter has less than $\alpha/2$ probability of being less (more) than the lower (upper) tail of the $100(1-\alpha)$% confidence interval. 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

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

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

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

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).

Note that $p_c \ge p_b$ for all cases, so that p_b gives more powerful tests than p_c. On the other hand, although generally $p_m < p_c$ it is possible for $p_m > p_c$.

To calculate the associated 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 intervals 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 at 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, 2009).

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; however, it is not rare for p_m or p_b to be inconsistent with the `central` confidence interval (Fay, 2009). We show some examples of such inconsistencies in the examples below.

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 $\theta_0 = .5$ we have $p_c = p_m = p_b$, and there is not a chance of a test-CI inconsistency even when the confidence intervals are not inversions of the test as is done in `binom.test` . When $\theta_0 \ne 0.5$ there may be problems. We explore these cases in the two-sample Poisson case below, since the associated tests reduce through conditioning to one-sample binomial tests.

Note that there are 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, meaning that one could have parameter values that are included in the 90% confidence intervals but not in the 95% confidence intervals (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 are more rare.

The **exactci** package contains the `poisson.exact` function, which has options for each of the three methods and gives p-values with matching confidence intervals. The code `poisson.exact(5,r=1.8,tsmethod="central")` gives confidence intervals the same as above, but a p-value of $p_c = 0.073$; while `poisson.exact(5,r=1.8,tsmethod="minlike")` gives the p-value the same as p_m above, but 95% confidence intervals 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 95% confidence intervals 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,

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

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 for differences between the 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 which as we have seen give a test-CI inconsistency. The `poisson.exact` function avoids this 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 the designs where there are two groups of observations with binary observations. 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 (2009) studies the 2×2 table case with independent observations, so we only briefly give his motivating example here. The usual two-sided application of Fisher's exact test given by `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 (2009) 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 tested 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 of the number 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 number 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 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 cases/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 $\theta = .5$. We can tranform the parameter θ into an odds ratio by

$$\text{Odds Ratio} \equiv \phi = \frac{\theta}{1-\theta} \tag{2}$$

(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 by 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, we can easily get one-sided exact McNemar-type Tests. For two-sided tests we can get three different versions of the two-sided exact McNemar's test using the three `tsmethod` options, but all three are equivalent to the exact version of McNemar's test (see the Appendix in `vignette("exactMcNemar")` in **exact2x2**). Thus, there is only one exact McNemar's test. The difference between the `tsmethod` options is in the calculation of the confidence intervals. The default is to use `central` confidence intervals so that the probability that the true parameter is less than the lower $100(1-\alpha)\%$ confidence interval is guaranteed to be less than or equal to $\alpha/2$, and similarly for the upper confidence interval. 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.

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.

Note that although these 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, 2009). These rare inconsistencies are an unavoidable problem due to the nature of the problem and not to any deficit in the packages. Additionally, those options can have other anomalies (see Vos and Hudson (2008) for the single sample binomial case, and Fay (2009) 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.

Bibliography

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*, 2009. doi: 10.1093/biostatistics/kxp050.

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