# Exact McNemar's Test and Matching Confidence Intervals

Michael P. Fay

January 26, 2010

## McNemar's Original Test

Consider paired binary response data. 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

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

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

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

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 contol 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

```
> x <- matrix(c(21, 9, 2, 12), 2, 2)
> mcnemar.test(x)

        McNemar's Chi-squared test with continuity correction

data:  x
McNemar's chi-squared = 3.2727, df = 1, p-value = 0.07044
```

Without the continuity correction we get

```
> mcnemar.test(x, correct = FALSE)

        McNemar's Chi-squared test

data:  x
McNemar's chi-squared = 4.4545, df = 1, p-value = 0.03481
```

Since the inferences change so much, and 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. We study that in the next section.

# Exact Version of McNemar's 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 $\theta$ into an odds ratio by

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

(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 by using the 'binom.exact' function of the package 'exactci' then transform the results into odds ratios via equation 1. 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. In the appendix we show that all three two-sided methods give the same p-value and they all are equivalent to the exact version of McNemar's test. So there is only one defined 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.

Using $x$ defined earlier, here is the exact McNemar's test with the central confidence intervals:

```
> mcnemar.exact(x)

        Exact McNemar test (with central confidence intervals)

data:  x
b = 2, c = 9, p-value = 0.06543
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.02336464 1.07363844
sample estimates:
odds ratio
 0.2222222
```

# Appendix: Equivalence of two-sided p-values

For the two-sided exact tests, the sample space is $B \in \{0, 1, \ldots, b+c\}$. Let $n = b + c$ and let the binomial mass function under the null hypothesis of $\theta = .5$ (i.e., $\phi = 1$) be

$$f(x) = \left( \begin{array}{c} n \\ x \end{array} \right) \left(\frac{1}{2}\right)^x \left(\frac{1}{2}\right)^{n-x} = 2^{-n} \left( \begin{array}{c} n \\ x \end{array} \right).$$

The exact McNemar p-value is defined as

$$p_e \quad = \quad \sum_{x:Q(x,n-x) \geq Q(b,c)} f(x)$$

Here are the definitions of the exact p-values for the three two-sided methods. For the 'central' method, it is

$$p_c \quad = \quad \min\Big\{1, 2 * \min\Big(F(x), \bar{F}(x)\Big)\Big\}$$

where $F(x) = \sum_{i=0}^{x} f(i)$ and $\bar{F}(x) = 1 - F(x-1)$ and $F(-1) = 0$. For the 'minlike' method the p-value is

$$p_m \quad = \quad \sum_{x:f(x) \leq f(b)} f(x)$$

For the 'blaker' method the p-value is

$$p_b \quad = \quad \sum_{x:\min\left\{F(x),\bar{F}(x)\right\} \leq \min\left\{F(b),\bar{F}(b)\right\}} f(x)$$

To show the equivalence of $p_e, p_c, p_m$, and $p_b$ we first rewrite the summation indices in $p_e$. Note that

$$Q(x, n-x) = \frac{4(x - \frac{n}{2})^2}{n}$$

so the summation indices may be rewritten as:

$$\{x : Q(x, n-x) \geq Q(b, c)\} = \left\{x : \left|x - \frac{n}{2}\right| \geq \left|b - \frac{n}{2}\right|\right\} \tag{2}$$

In other words, $p_e$ is just the sum of $f(x)$ for all $x$ that are as far away or further from the center $(n/2)$ as $b$. Note that $f(x)$ is increasing for all $x < n/2$ and decreasing for all $x > n/2$. Further note that $f(x) = f(n-x)$ for all $x$ so that $F(x) = \bar{F}(n-x)$ for all $x$. Thus, it makes sense that all 4 p-values are equivalent for the case when $\theta = .5$.

Here are the details showing the equivalence. We break up the possibilities into three cases:

**Case 1,** $b = n/2$**:** In this case, from equation 2, the whole sample space is covered so $p_e = 1$. Also $F(x) = \bar{F}(x) > 1/2$ so $p_c = 1$. Because the unique peak of the $f(x)$ function happens at $n/2$ when $n$ is even ($n$ must be even when $b = n/2$), we can see that $p_m = 1$. Also because of that peak, $\min\left\{F(x), \bar{F}(x)\right\}$ is maximized at $b = n/2$ and $p_b = 1$.

**Case 2,** $b < n/2$**:** In this case, the set of all $x$ described by equation 2 is all $x \leq b$ and all $n - x \geq n - b$ so that $p_e = F(b) + \bar{F}(n-b)$. Also, $\min\left\{F(x), \bar{F}(x)\right\}$ is $F(x)$, and

$$F(x) = \sum_{i: f(i) \leq f(x) \text{ and } i < n/2} f(i).$$

Further,

$$F(x) = \bar{F}(n-x) = \sum_{i: f(i) \leq f(x) \text{ and } i > n/2} f(i).$$

So $p_c = 2 * F(b) = F(b) + \bar{F}(n-b) = \sum_{i: f(i) \leq f(b)} f(i)$ which is equivalent to $p_m$. Also since $F(x) = \bar{F}(n-x)$ we can see that all values of $x$ with $\min\left\{F(x), \bar{F}(x)\right\} \leq F(b)$ will also give the same p-value and $p_b$ is equivalent to the other p-values.

**Case 3,** $b > n/2$**:** By symmetry, we can show through similar arguments to Case 2 that all 4 p-values are equivalent.

# References


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

McNemar, Q. (1947). "Note on the sampling error of the difference between correlated proportions or percentages" Psychometrika 12:153-157.