International Journal of
Clinical Research & Trials

Research Article Open Access

A Note on Exact Tests and Confidence Intervals for Two-by-Two Contingency Tables in Randomized Trials

Yasutaka Chiba

Clinical Research Center, Kinki University Hospital, 377-2, Ohno-higashi, Osakasayama, Osaka, 589-8511, Japan

Abstract

There are two major exact tests used for hypothesis testing of two-by-two contingency tables: Fisher's exact test and Barnard's exact test. Recently, Chiba (Journal of Biometrics and Biostatistics 2015; 2: 244) developed new exact tests: a conditional exact test, which requires that a marginal total is fixed, and an unconditional exact test, which does not require that a marginal total is fixed and depends rather on the ratio of random assignment. Fisher's exact test can be regarded as a special case of the conditional exact test. For Barnard's and Chiba's exact tests, the confidence intervals linking to them can be constructed in a straightforward manner. In this article, we review these three exact tests, noting the differences in the null hypotheses that they test. Furthermore, using a numerical example, we demonstrate that the confidence interval linking to Barnard's exact test is not in fact an exact confidence interval for the causal effect.

Publication History:

Received: October 02, 2015
Accepted: December 31, 2015
Published: January 03, 2016

Keywords:

Barnard's exact test, Causal null hypothesis, Fisher's exact test, Potential outcome, Principal stratification

Introduction

In a randomized trial to compare two groups where the outcome is binary, the results can be summarized into a two-by-two contingency table and the equality of the response proportions between the two groups compared using a statistical hypothesis test.

Two widely used tests that make such a comparison and do not require any approximations are Fisher's exact test [1,2] and Barnard's exact test [3-5]. The former is more popular than the latter; however, Barnard's exact test has advantages over Fisher's exact test in that it is more powerful for moderate to small samples [6]. Although until recently it was not applied because of the significant computation time needed for the numerical search, it can now be applied easily using a software package such as SAS. Recently, Chiba [7] developed new exact tests; i.e., a conditional exact test, in which one marginal total is fixed, and an unconditional exact test, in which neither marginal total is fixed. Fisher's exact test can be regarded as a special case of Chiba's conditional exact test. The confidence intervals (CIs) linking to Barnard's and Chiba's exact tests can be constructed in a straightforward manner.

In this article, we review these three exact tests by noting the differences in the null hypotheses that they test. Furthermore, using a simple numerical example, we demonstrate that the CI linking to Barnard's exact test is not in fact an exact CI for the causal effect. For this demonstration, we apply the nonparametric bounds [8,9].

Notation and Principal Stratification

We use the following notation through this paper. Let X denote the assigned treatment; $X = 1$ if a subject was assigned to the treatment group, and $X = 0$ if assigned to the control group. Let Y denote the binary outcome; $Y = 1$ if the event occurred, and $Y = 0$ if it did not. Finally, let $Y(x)$ denote the potential outcomes [10] for each subject under $X= x$, which corresponds to the outcomes of the subject had he/she been in the trial group. Then, $\Pr(Y(x) = 1)$ represents a potential response proportion if all subjects are assigned to the group with $X = x$.

Here, we apply the principal stratification approach [11]. This approach considers the following four types of subjects to define the four principal strata:

(i) Individuals for whom the event would occur regardless of the assigned treatment group; i.e., $(Y(1), Y(0)) = (1, 1)$.

(ii) Individuals for whom the event would occur if assigned to the treatment group but would not occur if assigned to the control group; i.e., $(Y(1), Y(0)) = (1, 0)$.

(iii) Individuals for whom the event would not occur if assigned to the treatment group but would occur if assigned to the control group; i.e., $(Y(1), Y(0)) = (0, 1)$.

(iv) Individuals for whom the event would not occur regardless of the assigned treatment group; i.e., $(Y(1), Y(0)) = (0, 0)$.

All subjects belong to one of these four types; however, unfortunately we cannot know the numbers of type (i)–(iv) subjects from the observed data.

To review the three exact tests, let us assume that the generic two-by-two contingency table in Table 1 is obtained from a randomized trial, where a, b, c, d, and n are the numbers of subjects. The risk difference can be calculated as follows:

$$\mathrm{RD} \quad := \frac{a}{a+b} - \frac{c}{c+d}$$

	Event		
Group	Yes ($Y = 1$)	No ($Y = 0$)	Total
Treatment ($X = 1$)	a	b	$a + b$
Control ($X = 0$)	c	d	$c + d$
Total	$a + c$	$b + d$	n

Table 1: Generic two-by-two contingency table, where a, b, c, d, and n indicate the numbers of subjects.

Here, we consider the case of $\mathrm{RD}_O \geq 0$, but a similar discussion holds for the case of $\mathrm{RD}_O \leq 0$. Table 2 lists simple example data for a hypothetical randomized trial with the assignment ratio of 1:1, $\mathrm{RD}_O = 3/5 - 1/5 = 0.4$.

***Corresponding Author:** Dr. Yasutaka Chiba, Clinical Research Center, Kinki University Hospital, 377-2, Ohno-higashi, Osakasayama, Osaka 589-8511, Japan; E-mail: chibay@med.kindai.ac.jp

Citation: Chiba Y (2016) A Note on Exact Tests and Confidence Intervals for Two-by-Two Contingency Tables in Randomized Trials. Int J Clin Res Trials 1: 102. doi: https://doi.org/10.15344/2456-8007/2016/102

	Event		
Group	Yes ($Y = 1$)	No ($Y = 0$)	Total
Treatment ($X = 1$)	1	4	5
Control ($X = 1$)	3	2	5

Table 2: Results from a hypothetical randomized trial with small sample.

Review of Exact Tests

Fisher's exact test

Let us denote a generic two-by-two contingency table under the null hypothesis using random variables W_1 and W_0 such as in Table 3. Here, the null hypothesis for Fisher's exact test is as follows:

	Event		
Group	Yes ($Y = 1$)	No ($Y = 0$)	Total
Treatment ($X = 1$)	w_1	$a + b – w_1$	$a + b$
Control ($X = 0$)	w_0	$c + d – w_0$	$c + d$
Total	$w_0 + w_1$	$n – w_0 – w_1$	n

Table 3: Generic two-by-two contingency table under the null hypothesis with realized values of random variables W_0 and W_1.

H_0: $Y(1) = Y(0)$ for all subjects,

which is referred to as the sharp causal null hypothesis [12].

Under this null hypothesis, subjects are limited to those with ($Y(1)$, $Y(0)$) = (0, 0) or (1, 1), which implies that an outcome for a subject is constant regardless of the assigned group. Then, subjects with $Y = 1$ are those with($Y(1)$, $Y(0)$) = (1, 1), and similarly subjects with $Y = 0$ are those with($Y(1)$, $Y(0)$) = (0, 0). Therefore, under the sharp causal null hypothesis, $w_0 + w_1 = a + c$ and $n – w_0 – w_1 = b + d$ from Tables 1 and 3.

The probability of $W_1 = w_1$ under the sharp causal null hypothesis is given by the hypergeometric distribution as follows:

$$p_{w_1} = \binom{a+c}{w_1}\binom{b+d}{a+b-w_1}\Big/\binom{n}{a+b}$$

where $\binom{j}{k} = {}_jC_k = \dfrac{j!}{k!(j-k)!}$ and max $\{0, a–d\} \le w_1 \le$ min $\{a + b, a + c\}$. This is the probability that w_1 subjects of $(a + c)$ subjects who experienced the event, and $(a + b – w_1)$ subjects of $(b + d)$ subjects who did not experience the event are selected in the treatment group, when $(a + b)$ subjects of the total n subjects are selected in the treatment group.

Because the risk difference under the null hypothesis can be expressed as

$$\mathrm{RD_N} := \frac{w_1}{a+b} - \frac{w_0}{c+d}$$

the one-sided p-value can be calculated by

$$p = \sum_{w_1=\max\{0,a-d\}}^{\min(a+b,a+c)} I(z)\binom{a+c}{w_1}\binom{b+d}{a+b-w_1}\Big/\binom{n}{a+b}$$
$$= \sum_{w_1=a}^{\min(a+b,a+c)} \binom{a+c}{w_1}\binom{b+d}{a+b-w_1}\Big/\binom{n}{a+b},$$

where $I(z) := \mathrm{RD_N} - \mathrm{RD_O} = 1$ if $z \ge 0$ and $I(z) = 0$ if $z < 0$ with $z = \mathrm{RD_N} - \mathrm{RD_O}$. The second equation is derived because $w_0 = a + c – w_1$ and thus

$$\mathrm{RD_N} - \mathrm{RD_O} = \left(\frac{w_1}{a+b} - \frac{w_0}{c+d}\right) - \left(\frac{a}{a+b} - \frac{c}{c+d}\right)$$
$$= \left(\frac{1}{a+b} + \frac{1}{c+d}\right)(w_1 - a).$$

This is the p-value for Fisher's exact test. For the hypothetical data listed in Table 2, we have

$$p = \sum_{w_1=3}^{4}\binom{4}{w_1}\binom{6}{5-w_1}\Big/\binom{10}{5} = 0.2619$$

Barnard's exact test

Let the response probability for the group with $X = x$ be $\Pr(Y = 1 \mid X = x) = \pi_x$. Then, the probability of observing w_1 and w_0 is given by the following product of two binomial probabilities:

$$p_{\pi_1,\pi_0} = \binom{a+b}{w_1}\pi_1^{w_1}(1-\pi_1)^{a+b-w_1}\binom{c+d}{w_0}\pi_0^{w_0}(1-\pi_0)^{c+d-w_0}$$

The null hypothesis for Barnard's exact test is as follows:

H_0: $\pi_1 = \pi_0$.

Therefore, for $\pi_1 = \pi_0 = \pi$, the one-sided p-value can be calculated by

$$p_\pi = \sum_{w_1=0}^{a+b}\sum_{w_0=0}^{c+d} I(z)\binom{a+b}{w_1}\binom{c+d}{w_0}\pi^{w_0+w_1}(1-\pi)^{n-w_0-w_1}$$

Unfortunately, because this calculation of the p-value includes a nuisance parameter π, we cannot yield the p-value immediately. Thus, we yield the p-value by calculating the p-values for all possible π and choosing the maximum value; i.e., $p = \sup\{p_\pi\}$. This is the p-value for Barnard's exact test. For the hypothetical data in Table 2, we have

$$p = p_{0.4950}$$
$$= \sum_{w_1=0}^{5}\sum_{w_0=0}^{5} I(z)\binom{5}{w_1}\binom{5}{w_0}0.4950^{w_0+w_1}(1-0.4950)^{10-w_0-w_1}$$
$$= 0.1719.$$

An SAS program to yield the p-values for Fisher's and Barnard's exact tests is given in the appendix.

Chiba's exact test

Let n_{st} denote the number of subjects with ($Y(1)$, $Y(0)$) = (s, t), where s, t = 0, 1. If all subjects are assigned to the treatment group ($X = 1$), then $\Pr(Y(1) = 1) = (n_{11} + n_{10}) / n$, because only subjects with type (i) or (ii) would experience the event. Likewise, if all subjects are assigned to the control group ($X = 0$), then $\Pr(Y(0) = 1) = (n_{11} + n_{01}) /n$, because only subjects with type (i) or (iii) would experience the event.

The null hypothesis for Chiba's exact test is as follows:

H_0: $n_{10} = n_{01}$.

This null hypothesis corresponds to $\Pr(Y(1) = 1) = \Pr(Y(0) = 1)$, which is referred to as the weak causal null hypothesis [12].

Let us assume that of the n_{st} subjects, $n_{st,1}$ subjects were assigned to the treatment group ($X = 1$), and $n_{st,0}$ subjects were assigned to the control group ($X = 0$) by random assignment at a 1:r ratio. This leads to the two-by-two contingency table shown in Table 4. Applying the binomial probabilities, the one-sided p-value can be calculated by

$$p_{n_{11},n_{10},n_{01},n_{00}} = \sum_{n_{11,1}=0}^{n_{11}} \sum_{n_{10,1}=0}^{n_{10}} \sum_{n_{01,1}=0}^{n_{01}} \sum_{n_{00,1}=0}^{n_{00}} I'(z) \prod_{s=0}^{1} \prod_{t=0}^{1} \binom{n_{st}}{n_{st,1}} \left(\frac{1}{1+r}\right)^{n_{st,1}} \left(\frac{r}{1+r}\right)^{n_{st,0}} \qquad (1)$$

	Event		
Group	Yes ($Y = 1$)	No ($Y = 0$)	Total
Treatment ($X = 1$)	$n_{11,1} + n_{10,1}$	$n_{00,1} + n_{01,1}$	$n_{11,1} + n_{10,1} +$ $n_{01,1} + n_{00,1}$
Control ($X = 0$)	$n_{11,0} + n_{01,0}$	$n_{00,0} + n_{10,0}$	$n_{11,0} + n_{10,0} +$ $n_{01,0} + n_{00,0}$
Total	$n_{11} + n_{10,1} + n_{01,0}$	$n_{00} + n_{01,1} + n_{10,0}$	n

Table 4: Generic two-by-two contingency table with the numbers for the four types of subjects defining the four principal strata.

where $I'(Z)$=1 if z≥ 0 and $I'(Z)$=0 if z< 0 with , z=RD'$_N$- RD$_O$ and

$$\mathrm{RD}'_N := \frac{n_{11,1} + n_{10,1}}{n_{11,1} + n_{10,1} + n_{01,1} + n_{00,1}} - \frac{n_{11,0} + n_{01,0}}{n_{11,0} + n_{10,0} + n_{01,0} + n_{00,0}}$$

where n_{st} (s, t = 0, 1) must satisfy the following conditions:

$$n_{10} = n_{01}, \sum_{s=0}^{1}\sum_{t=0}^{1} n_{st} = n, \begin{cases} n_{11} \le a + c \\ n_{10} \le a + d \\ n_{01} \le b + c \\ n_{00} \le b + d \end{cases} \text{and} \begin{cases} a \le n_{11} + n_{10} \le n - b \\ c \le n_{11} + n_{01} \le n - d \\ d \le n_{00} + n_{10} \le n - c \\ b \le n_{00} + n_{01} \le n - a \end{cases} \qquad (2)$$

Similar to Barnard's exact test, because this calculation of the p-value includes the nuisance parameters n_{11}, n_{10}, n_{01} and n_{00}, we cannot yield the p-value directly. Thus, we yield the p-value by calculating the p-values for all possible combinations of (n_{11}, n_{10}, n_{01} and n_{00}) and choosing the maximum value; i.e., p = sup{ $p_{n_{11},n_{10},n_{01},n_{00}}$}. This is the p-value for Chiba's unconditional exact test. For the hypothetical data in Table 2, we have

$$p = p_{4,0,0,6} = \sum_{n_{11,1}=3}^{4} \binom{4}{n_{11,1}} \binom{6}{5 - n_{11,1}} \left(\frac{1}{2}\right)^{10} = 0.1592$$

For Chiba's conditional exact test,

$$p_{n_{11},n_{10},n_{01},n_{00}} = \sum_{n_{11,1}=0}^{n_{11}} \sum_{n_{10,1}=0}^{n_{10}} \sum_{n_{01,1}=0}^{n_{01}} \sum_{n_{00,1}=0}^{n_{00}} I(z) \prod_{s=0}^{1} \prod_{t=0}^{1} \binom{n_{st}}{n_{st,1}} \Big/ \binom{n}{a+b}$$

is applied rather than (1), and $\Sigma_s \Sigma_t n_{st,1} = a + b$ is added to the conditions in (2). As with Fisher's exact test, the calculation of this p-value is based on the probability that $n_{st,1}$ subjects of n_{st} subjects are selected in the treatment group when (a + b) subjects of the total n subjects are selected in the treatment group. Note that the special case of n_{10} = n_{01} = 0, for which subjects are limited to those with (Y(1), Y(0)) = (0, 0) or (1, 1), corresponds to Fisher's exact test. For the hypothetical data listed in Table 2, we have

$$p = p_{4,0,0,6} = \sum_{n_{11,1}=3}^{4} \binom{4}{n_{11,1}} \binom{6}{5 - n_{11,1}} \Big/ \binom{10}{5} = 0.2619$$

Difference in Null Hypotheses

Table 5 lists the null hypotheses for the three exact tests. In many actual randomized trials, the most interesting null hypothesis will be the weak causal null hypothesis, where the causal risk difference is zero; i.e., Pr(Y(1) = 1) – Pr(Y(0) = 1) = 0. Chiba's exact test is a hypothetical test for the weak causal null hypothesis; however, Fisher's and Barnard's exact tests are not. The weak causal null hypothesis holds whenever the sharp causal null hypothesis holds, but rejection of the sharp causal null hypothesis does not imply rejection of the weak causal null hypothesis; i.e., Pr(Y(1) = 1)–Pr(Y(0) = 1)≠ 0 [7]. Barnard's exact test includes nothing about the causal effect.

Exact test	Null hypothesis
Fisher	Y(1) = Y(0) for all subjects (sharp causal null hypothesis)
Barnard	Pr(Y = 1 | X = 1) = Pr(Y = 1 | X = 0)
Chiba	n_{10} = n_{01}(corresponding to the weak causal null hypothesis Pr(Y(1) = 1) = Pr(Y(0) = 1))

Table 5: Null hypotheses for Fisher's, Barnard's and Chiba's exact tests.

Nevertheless, Fisher's exact test can be a hypothesis test for the weak causal null hypothesis under the monotonicity assumption [13,14], which implies that there are no subjects with (Y(1), Y(0)) = (0, 1). Barnard's exact test can be a hypothesis test for the weak causal null hypothesis under the exchangeability assumption [15], which implies that, in a randomized trial with a 1:1 ratio, the number of type (i)–(iv) subjects in the treatment group is exactly equal to that in the control group. See Chiba [7] for further details.

Exact Confidence Intervals

We can construct the exact CIs linking to Barnard's and Chiba's exact tests, whereas the exact CI linking to Fisher's exact test cannot be constructed in a straightforward manner. Specifically, the exact CI linking to Barnard's exact test can be yielded easily using the FREQ procedure in SAS (version 9.4, SAS Institute, Cary, NC, USA). The program used to achieve this is described in the appendix.

We applied the exact CIs to the data listed in Table 2. The 95% CIs are given in Table 6, where the risk difference was –0.4.

Exact test	p-value	95% CI
Barnard	0.1719	(–0.3049, 0.8665)
Chiba (Unconditional)	0.1592	(–0.2000, 0.7000)
Chiba (Conditional)	0.2619	(–0.2000, 0.7000)

Table 6: 95% confidence intervals yielded from the data in Table 2, where the nonparametric bounds are –0.3≤ Pr(Y(1) = 1)–Pr(Y(0) = 1)≤ 0.7.

To examine whether these CIs are in fact exact CIs for the causal effect, we applied nonparametric bounds [8,9]. Nonparametric bounds are a range within which the causal risk difference must exist and are given by

–{Pr(Y = 1, X = 0) + Pr(Y = 0, X = 1)}
≤ Pr(Y(1) = 1)–Pr(Y(0) = 1)
≤ Pr(Y = 1, X = 1) + Pr(Y = 0, X = 0).

For the hypothetical data in Table 2, we have

–(1/10 + 2/10) = –0.3
≤ Pr(Y(1) = 1)–Pr(Y(0) = 1)
≤ (3/10 + 4/10) = 0.7.

Since the nonparametric bounds describe a range within which the causal risk difference must exist, the upper limit of 95% CI must be smaller than the upper nonparametric bound, and the lower limit must be larger than the lower nonparametric bound. However, for the CI linking to Barnard's exact test, the upper limit (0.8665) was larger than the upper bound of 0.7, and the lower limit (–0.3049) was smaller than the lower bound of –0.3. This demonstrates that the 95% CI linking to Barnard's exact test includes values that the causal risk difference cannot take. As a consequence, the exact CI is not in fact an exact CI for the causal effect. Barnard's exact test is exact for $\Pr(Y = 1 \mid X = 1) - \Pr(Y = 1 \mid X = 0)$ but not for the causal effect $\Pr(Y(1) = 1) - \Pr(Y(0) = 1)$. The limits of the CIs linking to Chiba's exact test cannot be outside the nonparametric bounds, because the inequalities in (2) correspond to the nonparametric bounds.

Discussion and Conclusion

We have reviewed three exact tests for two-by-two contingency tables in the context of randomized trials and demonstrated that the exact CI linking to Barnard's exact test is not in fact an exact CI for the causal effect.

Researchers may encounter a situation in which they would like to examine the weak causal null hypothesis but cannot apply a hypothesis test for it. In such a situation, we recommend that they employ Fisher's exact test when the sample size is small, for the following two reasons. First, we have a higher violation possibility of the exchangeability assumption for a relatively small sample size. Second, violation of the monotonicity assumption (i.e., at least one subject with $(Y(1), Y(0)) = (1, 0)$ exists in the trial) will not be guaranteed for a small sample size. Conversely, in the case in which monotonicity cannot be assumed, we recommend that researchers employ Barnard's exact test when the sample size is large, because it will not be claimed that exchangeability holds at least approximately. In general, Barnard's exact test is more powerful than Fisher's exact test for moderate to small samples [6]. Therefore, this recommendation will derive a conservative result. Nevertheless, such a result may be welcomed in randomized trials to avoid a large probability of type I error.

As demonstrated in the case with a small sample size, the CI linking to Barnard's exact test was not in fact an exact CI for the causal effect. Therefore, we recommend that researchers do not apply this exact CI for small sample sizes.

In many randomized trials, the most interesting null hypothesis will be the weak causal null hypothesis. Nevertheless, to the best of our knowledge, exact tests for it have received little investigation. Such exact tests and CIs linking to them should be investigated further and applied to actual randomized trials.

Appendix

The following is an SAS program used to yield the p-values from Fisher's and Barnard's exact tests, as well as the exact CI linking to Barnard's exact test.

```
data dat0;
   input x y count;
   cards;
   1 1 3
   1 0 2
   0 1 1
   0 0 4
   run;
proc freq data=dat0;
   tables x*y/norow nocol nopercent Fisher alpha = 0.05;
   exact Barnard riskdiff (method=score);
   weight count;
run;
```

Competing Interests

The author declares that he has no competing interests.

Funding

This work was supported partially by Grant-in-Aid for Scientific Research (No. 15K00057) from Japan Society for the Promotion of Science.

References

1. Fisher RA (1925) Statistical Methods for Research Workers, Edinburgh: Oliver and Boyd.
2. Fisher RA (1935) The logic of inductive inference. J Roy Stat Soc Ser A 98: 39-54.
3. Barnard GA (1945) A new test for 2×2 tables. Nature 156:177.
4. Barnard GA (1947) Significance tests for 2 X 2 tables. Biometrika 34: 123-138.
5. Barnard GA (1949) Statistical inference. J Roy Stat Soc Ser B 11: 115-139.
6. Mehrotra DV, Chan ISF, Berger RL (2003) A cautionary note on exact unconditional inference for a difference between two independent binomial proportions. Biometrics 59: 441-450.
7. Chiba Y (2015) Exact Tests for the weak causal null hypothesis on a binary outcome in randomized trials. J Biomet Biostat 6: 244.
8. Manski CF (1990) Nonparametric bounds on treatment effects. American Economic Review 80: 319-323.
9. Pearl J (1995) Causal inference from indirect experiments. Artif Intell Med 7: 561-582.
10. Rubin DB (1978) Bayesian inference for causal effects: the role of randomization. Ann Stat 6: 34-58.
11. Frangakis CE, Rubin DB (2002) Principal stratification in causal inference. Biometrics 58: 21-29.
12. Greenland S (1992) On the logical justification of conditional tests for two-by-two contingency tables. Am Stat 45: 248-251.
13. Angrist JD, Imbens GW,Rubin DB (1996)Identification of causal effects using instrumental variables (with discussion). J Am Stat Assoc 91: 444-472.
14. Manski CF (1997) Monotone treatment response. Econometrica 65: 1311-1334.
15. Greenland S, Robins JM (1986) Identifiability, exchangeability, and epidemiological confounding. Int J Epidemiol 15: 413-419.