Published in final edited form as:

Stat Med. 2017 February 28; 36(5): 876–880. doi:10.1002/sim.7192.

Response to comment on “Randomization inference for treatment effects on a binary outcome”

Joseph Rigdon^a, Wen Wei Loh^b, and Michael G. Hudgens^b

^aQuantitative Sciences Unit, Stanford University, Chapel Hill

^bDepartment of Biostatistics, University of North Carolina, Chapel Hill

Abstract

We thank Professor Yasutaka Chiba [1] for commenting on Rigdon and Hudgens (RH) [2]. Chiba [1] described a certain exact confidence interval reported in RH as “somewhat unnatural.” Chiba also presented an alternative approach to constructing confidence intervals [3]. In this response, we (i) provide a simple explanation why the confidence interval in RH appeared “unnatural,” and (ii) explain the relationship between the RH [2] and Chiba [3] confidence intervals. Essentially the two approaches are equivalent, except RH entails inverting one two-sided test whereas Chiba inverts two one-sided tests. We present a more computationally efficient method (RLH) for computing the RH intervals based on Chiba’s principal stratification formulation of the problem. We also propose a third method based on Blaker [4] which inverts a single two-sided test but forms a confidence interval that is at least as narrow as inverting two one-sided tests. Simulation results show the RLH intervals tend to be as narrow or narrower than the Chiba and Blaker intervals on average.

Keywords

additivity; causal inference; exact confidence interval; permutation tests; randomization inference

1. Introduction

We would like to thank Professor Yasutaka Chiba [1] for his comments on Rigdon and Hudgens (RH) [2]. In this response, we compare Chiba’s approach to constructing confidence intervals (CIs), as described in [3], with the approach in RH. First, we explain that despite the different parameterizations, both approaches test the same null hypotheses for the average causal effect of treatment, using the same randomization distributions to carry out the permutation tests. The difference between the two methods lies in the choice of whether to invert and combine two one-sided tests as Chiba proposed, or to invert a single two-sided test as in RH.

Next, we present a more computationally efficient method (RLH) for computing the RH intervals based on Chiba’s principal stratification formulation of the problem. We also

*Correspondence to: Quantitative Sciences Unit, 1070 Arastradero Road, Palo Alto, California 94304, U.S.A. jrigdon@stanford.edu.

propose a third method based on Blaker [4] which inverts a single two-sided test but forms a CI that is at least as narrow as Chiba's method. Our simulation results show the RLH intervals tend to be narrowest on average among the three methods, but the differences between the methods are reduced either as the magnitude of the average causal effect increases or as the population size increases.

Finally, we provide a simple explanation why a certain CI reported in RH was described by Chiba as "somewhat unnatural" [1]. As the original formulation of the RH approach could be computationally intensive, Monte Carlo sampling was used to approximate p-values when carrying out the hypothesis tests. The interval reported in RH appeared "unnatural" because of the Monte Carlo approximation. Using the RLH approach without Monte Carlo sampling, we obtain the same interval as Chiba [1, 3].

2. Methods

Suppose m of n individuals are randomized to either treatment or control. Assume m is fixed by experimental design. Let $Z_j = 1$ if individual j is assigned treatment and $Z_j = 0$ otherwise. Let the binary outcome of interest be coded as either $Y_j = 1$ or $Y_j = 0$. Assume prior to randomization, each individual has two potential outcomes: $y_j(1)$ if treatment is assigned and $y_j(0)$ if control is assigned. Consequently, the treatment assignment reveals the observed outcome $Y_j = Z_j y_j(1) + (1 - Z_j) y_j(0)$. The parameter of interest is the average causal effect of treatment $\tau = \sum_{j=1}^n \delta_j / n$, where $\delta_j = y_j(1) - y_j(0)$. The data observed from the experiment can be summarized using the notation in Table 1. The observed difference in proportions $T = a/(a + b) - c/(c + d)$ is an unbiased estimator of τ [5].

2.1. Exact Confidence Interval of Rigdon and Hudgens (2015)

Two exact CIs for τ were proposed in [2]. The first combines two prediction intervals for attributable effects [6], while the second entails inverting a permutation test. Simulations presented in [2] showed the permutation-based CIs tend to be narrower than the attributable effects-based CIs. Only the permutation-based CI will be considered below.

The observed data reveal one of the two potential outcomes for each individual. Since the missing potential outcome is either 0 or 1 for a binary outcome, there are 2^n possible unique values for the vector $\delta = (\delta_1, \dots, \delta_n)$ given the observed data. Furthermore, there are $n + 1$ unique values of τ that are compatible with the observed data and comprise a set of width one, where the width of a set is defined as the difference between the maximum and minimum elements of the set [2, §1]. A confidence set for τ can be constructed by conducting hypothesis tests about δ as follows.

To test the null hypothesis $H_0: \delta = \delta^0 = (\delta_1^0, \dots, \delta_n^0)$, a test statistic can be chosen, its distribution under H_0 computed, and a measure of extremeness of the observed data defined [7, §4.1]. A natural choice for the test statistic is T . The sampling distribution of T under H_0

can be determined exactly by computing T for each of the $C = \binom{n}{m}$ possible randomizations for $\{Z_j: j = 1, \dots, n\}$ because all potential outcomes are known under the

sharp null H_0 . For randomizations $c = 1, \dots, C$, let t^c denote the value of T under H_0 . Each randomization occurs with probability $1/C$, so a two-sided p-value can be defined as

$\sum_{c=1}^C 1\{|t^c - \tau^0| \geq |t^{\text{obs}} - \tau^0|\} / C$ where t^{obs} is the value of T for the observed data and $\tau^0 = \sum_{j=1}^n \delta_j^0 / n$. The subset of δ^0 vectors where the p-value is greater than or equal to α forms a $100(1 - \alpha)\%$ confidence set for δ . The τ^0 values corresponding to the δ^0 vectors in this confidence set for δ then form a $100(1 - \alpha)\%$ confidence set for τ .

However, one need not test all 2^n hypotheses for $H_0 : \delta = \delta^0$ to obtain an exact confidence set for τ . The set of possible δ^0 values can be partitioned into $(a+1)(b+1)(c+1)(d+1)$ subsets such that all δ^0 in the same subset yield the same p-value. It is thus only necessary to test one δ^0 in each of these subsets [2, §3]; this approach was proposed in RH. When the total number of possible treatment assignments C becomes large, RH suggested approximating the p-value for each hypothesis test with a Monte Carlo random sample of n_{perm} assignments instead of evaluating all C possibilities.

2.2. Exact Confidence Interval of Chiba (2015)

Chiba [3] proposed an exact CI for τ using a principal stratification approach. Let n_{st} denote the number of individuals with $\{y_f(1) = s, y_f(0) = t\}$ for $s, t = 0, 1$, such that $\tau = (n_{10} - n_{01})/n$. Chiba considered null hypotheses about the parameter $\mathbf{n} = (n_{11}, n_{10}, n_{01}, n_{00})$ that are compatible with the observed data, i.e., values of $\mathbf{n}$ that satisfy the conditions:

$$\sum_{s=0}^1 \sum_{t=0}^1 n_{st} = n, \quad n_{11} \leq a+c, \quad n_{10} \leq a+d, \quad n_{01} \leq b+c, \quad n_{00} \leq b+d, \\ n_{11} + n_{10} \leq n - b, \quad n_{11} + n_{01} \leq n - d, \quad n_{00} + n_{10} \leq n - c, \quad n_{00} + n_{01} \leq n - a. \quad (1)$$

Let $n_{st,z}$ denote the number of individuals with $\{y_f(1) = s, y_f(0) = t\}$ and $Z_j = z$ for $s, t = 0, 1$. For m fixed by design, the conditional probability of $(n_{11,1}, n_{10,1}, n_{01,1}, n_{00,1})$ given $\mathbf{n}$ is:

$$p_{\mathbf{n}}^*(n_{11,1}, n_{10,1}, n_{01,1}, n_{00,1}) = \prod_{s=0}^1 \prod_{t=0}^1 \binom{n_{st}}{n_{st,1}} / \binom{n}{m}. \quad (2)$$

Chiba proposed the following conditional exact p-value to test a null hypothesis for a given value of $\mathbf{n}$:

$$P_f(t_{\text{obs}}; \mathbf{n}) = \sum_{i=0}^{n_{11}} \sum_{j=0}^{n_{10}} \sum_{k=0}^{n_{01}} \sum_{l=0}^{n_{00}} f(T, t_{\text{obs}}) p_{\mathbf{n}}^*(i, j, k, l) 1\{i+j+k+l=m\}, \quad (3)$$

where $1\{B\} = 1$ if B is true and 0 otherwise, and f is a function of $T = (i+j)/m - \{(n_{11} - i) + (n_{01} - k)\} / (n - m)$ and t_{obs} . Unlike the RH method described above, Chiba's approach

does not explicitly compute a test statistic for all $\binom{n}{m}$ possible randomization assignments, thus potentially obviating the need for Monte Carlo approximation of the p-values $P_f(t_{obs}; \mathbf{n})$. Chiba's $100(1 - \alpha)\%$ CI [3, §1] entails inverting two one-sided tests and equals $[L_1, U_1]$ where:

$$L_1 = \min \left\{ \frac{n_{10} - n_{01}}{n} : P_{f_{L,1}}(t_{obs}; \mathbf{n}) \geq \alpha/2 \right\}, U_1 = \max \left\{ \frac{n_{10} - n_{01}}{n} : P_{f_{U,1}}(t_{obs}; \mathbf{n}) \geq \alpha/2 \right\}, \quad (4)$$

and $f_{L,1} = 1\{(T - t_{obs}) \leq 0\}$ and $f_{U,1} = 1\{(T - t_{obs}) \geq 0\}$.

2.3. Comparison of RH and Chiba

In general, the approach of inverting and combining two separate one-sided tests is termed the *tail method* [8]. An alternative to Chiba's tail method CI is to invert a single two-sided test. A $100(1 - \alpha)\%$ CI analogous to (4) that inverts a single two-sided test is $[L_2, U_2]$, where:

$$L_2 = \min \left\{ \frac{n_{10} - n_{01}}{n} : P_{f_2}(t_{obs}; \mathbf{n}) \geq \alpha \right\}, U_2 = \max \left\{ \frac{n_{10} - n_{01}}{n} : P_{f_2}(t_{obs}; \mathbf{n}) \geq \alpha \right\}, \quad (5)$$

and $f_2 = 1\{|T - (n_{10} - n_{01})/n| \geq |t_{obs} - (n_{10} - n_{01})/n|\}$. The interval $[L_2, U_2]$ in (5) is equivalent to the permutation-based interval RH in [2]. Note in the finite population where n is fixed, one may determine an element n_{st} in $\mathbf{n}$ if given the other three, for example $n_{00} = n - n_{11} - n_{10} - n_{01}$. The set of possible values for the parameter $\mathbf{n}$ that satisfy (1) thus lies in a three-dimensional space [9]. Therefore, under the formulation in (5), one need only test $O(n^3)$ hypotheses for $\mathbf{n}$, as compared to the $O(n^4)$ hypotheses under the RH formulation. Moreover, the requirement for Monte Carlo approximations of the p-values is also potentially avoided. We will henceforth refer to this modified approach as RLH. Li and Deng [10] describe a more computationally efficient approach to construct these intervals which requires only $O(n^2)$ hypothesis tests, but the resulting CIs may potentially be wider.

2.4. Blaker Confidence Interval

An alternative approach that inverts a single two-sided test but forms an interval necessarily contained within the interval using the tail method was proposed by Blaker [4]. Let the minimum one-sided tail probability of T be denoted as $\gamma(T; \mathbf{n}) = \min \{P_{f_{L,1}}(T; \mathbf{n}), P_{f_{U,1}}(T; \mathbf{n})\}$, and let $f_\gamma = 1\{\gamma(T; \mathbf{n}) \leq \gamma(t_{obs}; \mathbf{n})\}$. A $100(1 - \alpha)\%$ CI is $[L_\gamma, U_\gamma]$, henceforth referred to as a *Blaker* interval, where:

$$L_\gamma = \min \left\{ \frac{n_{10} - n_{01}}{n} : P_{f_\gamma}(t_{obs}; \mathbf{n}) \geq \alpha \right\}, U_\gamma = \max \left\{ \frac{n_{10} - n_{01}}{n} : P_{f_\gamma}(t_{obs}; \mathbf{n}) \geq \alpha \right\}. \quad (6)$$

The Chiba, RLH and Blaker intervals can be computed using version 1.3 of the R [11] package `RI2by2` [12].

3. Results

3.1. Simulation Study

We compared the Chiba interval (4), the RLH interval (5), and the Blaker interval (6) in a simulation study. The simulations were carried out as described in scenarios (i) and (ii) of [2], and the results are summarized in Tables 2 and 3 respectively. The CIs using all three methods had coverage greater than the nominal level. For the scenarios where exactly 50% were assigned to treatment, all three methods returned identical intervals. For the scenarios where either 30% or 70% were assigned to treatment, the RLH intervals were as narrow or narrower than both the Blaker and Chiba intervals on average for smaller values of n . The Blaker intervals tended to be slightly narrower than the Chiba intervals. The differences between the methods were reduced when either n or the magnitude of the true average causal effect τ increased.

3.2. Counterexamples

While the RLH intervals tend to be narrower than the Chiba intervals, it is possible that the Chiba intervals are narrower for particular data sets. For example, suppose $a = 11$, $b = 1$, $c = 7$, $d = 21$. Then both the Chiba and Blaker 95% CIs equal $[0.400, 0.775]$ whereas the RLH interval equals $[0.375, 0.775]$.

The Blaker CI may also be narrower than both the Chiba and RLH CIs for particular data sets. For example, suppose $a = 7$, $b = 5$, $c = 1$, $d = 27$. For these data the Blaker 95% CI equals $[0.275, 0.750]$, whereas the RLH interval equals $[0.250, 0.750]$ and the Chiba interval equals $[0.250, 0.775]$.

3.3. Application

In the vaccine adherence example presented in §4.3 of [2], $n = 96$ injection drug users were randomized to a monetary incentive group or an outreach group. Of the $m = 48$ individuals in the monetary incentive group, $a = 33$ were adherent, and of the $n - m = 48$ in the outreach group, $c = 11$ were adherent. The estimated average causal effect was $t^{obs} = 33/48 - 11/48 \approx 0.46$. The RH approach entails testing $(33 + 1)(15 + 1)(11 + 1)(37 + 1) = 248064$ hypotheses to obtain an exact CI. Given the large number of possible treatment assignments

$$C = \binom{96}{48} \approx 6 \times 10^{27}$$

, RH approximated the p-value for each hypothesis test with $n_{perm}=100$ re-randomizations. The resulting 95% CI in [2] was $[0.28, 0.64]$, which Chiba [1] described as “somewhat unnatural.” However, because only 100 re-randomizations were used, the resulting CI in RH was heavily dependent on the random seed, with different random seeds producing different intervals. Using RLH without Monte Carlo sampling, the interval equals $[0.28125, 0.59375]$, which is the same as Chiba’s interval [1, 3]. The Blaker interval also equals $[0.28125, 0.59375]$.

Acknowledgments

The authors thank the Editor for the invitation to respond to Professor Chiba's comments. This research was supported in part by the National Institutes of Health and by a Gillings Innovation Laboratory award from the UNC Gillings School of Global Public Health. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

1. Chiba Y. A note on exact confidence interval for causal effects on a binary outcome in randomized trials. *Statistics in Medicine*. 2016; 35(10):1739–1741. [PubMed: 27059605]
2. Rigdon J, Hudgens MG. Randomization inference for treatment effects on a binary outcome. *Statistics in Medicine*. 2015; 34(6):924–935. [PubMed: 25471299]
3. Chiba Y. Exact tests for the weak causal null hypothesis on a binary outcome in randomized trials. *Journal of Biometrics & Biostatistics*. 2015; 6(244)
4. Blaker H. Confidence curves and improved exact confidence intervals for discrete distributions. *Canadian Journal of Statistics*. 2000; 28(4):783–798.
5. Neyman, J. On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Annals of Agricultural Science* 1923. In: Dabrowska, DM., Speed, TP., translators. *Statistical Science*. Vol. 5. 1990. p. 465-472.
6. Rosenbaum PR. Effects attributable to treatment: Inference in experiments and observational studies with a discrete pivot. *Biometrika*. 2001; 88(1):219–231.
7. Rubin DB. Practical implications of modes of statistical inference for causal effects and the critical role of the assignment mechanism. *Biometrics*. 1991; 47(4):1213–1234. [PubMed: 1786315]
8. Agresti A. Dealing with discreteness: Making 'exact' confidence intervals for proportions, differences of proportions, and odds ratios more exact. *Statistical Methods in Medical Research*. 2003; 12(1):3–21. [PubMed: 12617505]
9. Copas J. Randomization models for the matched and unmatched 2×2 tables. *Biometrika*. 1973; 60(3):467–476.
10. Li X, Ding P. Exact confidence intervals for the average causal effect on a binary outcome. *Statistics in Medicine*. 2016; 35(6):957–960. [PubMed: 26833798]
11. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2016. URL <https://www.R-project.org/>
12. Rigdon J, Loh WW, Hudgens MG. RI2by2: Randomization Inference for Treatment Effects on a Binary Outcome. R package version 1.3.

Table 1

Observed data from experiment

Randomization Assignment	Outcome		Total
	$Y = 1$	$Y = 0$	
$Z = 1$	a	b	$a + b = m$
$Z = 0$	c	d	$c + d = n - m$
Total	$a + c$	$b + d$	n

Table 2

Simulation results under scenario (4) in [2]. Table entries give the average empirical width [coverage] of 95% CIs, where τ is the true average treatment effect, % treatment is the percent of n total individuals assigned to treatment in each experiment, Chiba CIs are given by (4), RLH CIs by (5), and Blaker CIs by (6).

n	Method	30% treatment			50% treatment			70% treatment		
		$\tau = 0.2$	$\tau = 0.5$	$\tau = 0.95$	$\tau = 0.2$	$\tau = 0.5$	$\tau = 0.95$	$\tau = 0.2$	$\tau = 0.5$	$\tau = 0.95$
20	Chiba	0.76[1.00]	0.72[1.00]	0.39[1.00]	0.72[1.00]	0.69[1.00]	0.30[1.00]	0.76[1.00]	0.72[1.00]	0.39[1.00]
	Blaker	0.73[1.00]	0.68[1.00]	0.39[1.00]	0.72[1.00]	0.69[1.00]	0.30[1.00]	0.73[1.00]	0.67[1.00]	0.39[1.00]
	RLH	0.71[1.00]	0.66[1.00]	0.39[1.00]	0.72[1.00]	0.69[1.00]	0.30[1.00]	0.71[1.00]	0.66[1.00]	0.39[1.00]
40	Chiba	0.58[0.99]	0.51[0.99]	0.25[1.00]	0.54[1.00]	0.47[1.00]	0.17[1.00]	0.58[0.99]	0.51[0.99]	0.25[1.00]
	Blaker	0.57[0.99]	0.49[0.99]	0.23[1.00]	0.54[1.00]	0.47[1.00]	0.17[1.00]	0.57[0.99]	0.49[0.99]	0.23[1.00]
	RLH	0.56[0.99]	0.49[0.99]	0.24[1.00]	0.54[1.00]	0.47[1.00]	0.17[1.00]	0.56[0.99]	0.49[0.99]	0.24[1.00]
60	Chiba	0.48[0.99]	0.41[0.99]	0.18[1.00]	0.45[1.00]	0.38[1.00]	0.13[1.00]	0.48[0.99]	0.41[0.99]	0.18[1.00]
	Blaker	0.48[0.99]	0.40[0.99]	0.17[1.00]	0.45[1.00]	0.38[1.00]	0.13[1.00]	0.48[0.99]	0.41[0.99]	0.17[1.00]
	RLH	0.47[0.99]	0.40[0.99]	0.18[1.00]	0.45[1.00]	0.38[1.00]	0.13[1.00]	0.47[0.99]	0.40[0.99]	0.18[1.00]
100	Chiba	0.39[0.99]	0.32[1.00]	0.13[1.00]	0.36[1.00]	0.29[1.00]	0.10[1.00]	0.39[0.99]	0.32[1.00]	0.13[1.00]
	Blaker	0.38[0.99]	0.31[1.00]	0.12[1.00]	0.36[1.00]	0.29[1.00]	0.10[1.00]	0.38[0.99]	0.32[1.00]	0.12[0.99]
	RLH	0.38[0.99]	0.31[1.00]	0.13[1.00]	0.36[1.00]	0.29[1.00]	0.10[1.00]	0.38[0.99]	0.31[1.00]	0.13[0.99]

Table 3

Simulation results under scenario (ii) in [2]. Table entries give the average empirical width [coverage] of 95% CIs, where γ is the degree of additivity, % treatment is the percent of n total individuals assigned to treatment in each experiment, Chiba CIs are given by (4), RLH CIs by (5), and Blaker CIs by (6).

n	Method	30% treatment			50% treatment			70% treatment		
		$\gamma = 0.2$	$\gamma = 0.8$	$\gamma = 1$	$\gamma = 0.2$	$\gamma = 0.8$	$\gamma = 1$	$\gamma = 0.2$	$\gamma = 0.8$	$\gamma = 1$
20	Chiba	0.77[1.00]	0.76[0.99]	0.76[0.98]	0.72[1.00]	0.72[1.00]	0.72[0.99]	0.77[1.00]	0.76[0.99]	0.76[0.98]
	Blaker	0.74[1.00]	0.73[0.98]	0.73[0.97]	0.72[1.00]	0.72[1.00]	0.72[0.99]	0.73[1.00]	0.73[0.99]	0.73[0.97]
	RLH	0.71[1.00]	0.71[0.98]	0.71[0.97]	0.72[1.00]	0.72[1.00]	0.72[0.99]	0.71[1.00]	0.71[0.99]	0.71[0.97]
40	Chiba	0.58[1.00]	0.58[1.00]	0.58[0.98]	0.55[1.00]	0.55[0.99]	0.55[0.98]	0.58[1.00]	0.58[0.98]	0.58[0.97]
	Blaker	0.57[1.00]	0.57[0.99]	0.57[0.96]	0.55[1.00]	0.55[0.99]	0.55[0.97]	0.57[1.00]	0.57[0.97]	0.57[0.95]
	RLH	0.57[1.00]	0.56[0.99]	0.56[0.97]	0.55[1.00]	0.55[0.99]	0.55[0.98]	0.57[1.00]	0.56[0.97]	0.56[0.96]
60	Chiba	0.49[1.00]	0.49[0.98]	0.49[0.97]	0.46[1.00]	0.46[0.99]	0.46[0.97]	0.49[1.00]	0.49[0.98]	0.49[0.97]
	Blaker	0.48[1.00]	0.48[0.98]	0.48[0.96]	0.46[1.00]	0.46[0.99]	0.46[0.97]	0.48[1.00]	0.48[0.98]	0.48[0.97]
	RLH	0.48[1.00]	0.48[0.98]	0.48[0.96]	0.46[1.00]	0.46[0.99]	0.46[0.97]	0.48[1.00]	0.48[0.98]	0.48[0.97]
100	Chiba	0.39[1.00]	0.39[0.98]	0.39[0.97]	0.36[1.00]	0.36[0.98]	0.36[0.96]	0.39[1.00]	0.39[0.98]	0.39[0.97]
	Blaker	0.39[1.00]	0.39[0.97]	0.39[0.96]	0.36[1.00]	0.36[0.98]	0.36[0.96]	0.39[1.00]	0.39[0.98]	0.39[0.96]
	RLH	0.39[1.00]	0.39[0.97]	0.39[0.96]	0.36[1.00]	0.36[0.98]	0.36[0.96]	0.39[1.00]	0.39[0.98]	0.39[0.96]