

CONFIDENCE INTERVALS FOR CAUSAL PARAMETERS

JAMES M. ROBINS

*Occupational Health Program and Department of Biostatistics, Harvard School of Public Health,
665 Huntington Avenue, Boston, MA 02115, U.S.A.*

SUMMARY

Consider an unbiased follow-up study designed to investigate the causal effect of a dichotomous exposure on a dichotomous disease outcome. Under a deterministic outcome model, a standard '95 per cent binomial confidence interval' may fail to cover the causal parameter of interest at the nominal rate when we take the causal parameter to be a parameter associated with the observed study population (regardless of whether the observed study population was sampled from a larger superpopulation). I propose new interval estimators that, in this setting, improve upon the performance of the standard 'binomial confidence interval.'

KEY WORDS Epidemiologic methods Randomized trials Cohort studies Biometry

1. INTRODUCTION

Table I contains data from a follow-up study designed to investigate the causal effect of a dichotomous exposure on a dichotomous disease outcome. The standard analytic approach would report the nominal '95 per cent binomial interval'

$$\begin{aligned} [\hat{m}_1 - \hat{m}_0] \pm 1.96 \left[\frac{\hat{m}_1(1 - \hat{m}_1)}{N_E} + \frac{\hat{m}_0(1 - \hat{m}_0)}{N_{\bar{E}}} \right]^{1/2} \\ = (0.4 - 0.15) \pm 1.96 \left[\frac{(0.4)(0.6)}{100} + \frac{(0.15)(0.85)}{100} \right]^{1/2} \end{aligned} \quad (1)$$

as a 95 per cent confidence interval for the causal parameter of interest (when measuring the magnitude of the exposure effect on a difference scale). Here \hat{m}_1 and \hat{m}_0 are, respectively, the observed rates of disease in the exposed and unexposed study populations, and N_E and $N_{\bar{E}}$ are, respectively, the number of exposed and unexposed subjects in the observed study population. To ascertain whether this '95 per cent binomial interval' is a 95 per cent confidence interval for the causal parameter of interest we need to define formally

1. the parameter of causal interest, and
2. the set of hypothetical repetitions over which we will evaluate the coverage properties of the interval.

Suppose that, for purposes of preserving 'internal validity', we follow the suggestion of Miettinen and Cook¹ and take the causal parameters of interest as parameters associated with the *observed study population* (regardless of whether the observed study population was a sample from a larger superpopulation or target population). Furthermore, suppose we adopt the deterministic causal model independently proposed by Copas,² Rubin,³ and Miettinen and Cook.¹ In this model each subject has a deterministic disease outcome when exposed and another, possibly different,

Table 1. Results of a follow-up study

	E	\bar{E}
D	40	15
\bar{D}	100	100

disease outcome when unexposed. In the actual study we observe only one or the other of these disease outcomes, depending upon the exposure status of the subject. Under this deterministic causal model, two commonly chosen causal parameters of interest are the *causal risk difference* and the *causal risk difference in the exposed*. The causal risk difference is the difference between the rate of disease that we would have observed if the entire study population had been exposed and the rate of disease that we would have observed if the entire study population had been unexposed. The causal risk difference in the exposed is the difference between the observed rate of disease in the exposed subgroup of the study population and the rate of disease that we would have observed in the exposed subgroup had that subgroup been unexposed.

We will show that, even in the absence of bias, the 95 per cent binomial interval can, even in large samples, cover the causal risk difference in the exposed either more or less than 95 per cent of the time under both a randomization and superpopulation sampling model. The actual coverage rate of the nominal 95 per cent binomial interval will depend on the true state of nature. In fact, for certain extreme parameter values, the coverage rate of the 95 per cent binomial interval can approach zero! On the other hand, in large samples, the interval estimator

$$\hat{m}_1 - \hat{m}_0 \pm 1.96 \sqrt{\left\{ \frac{\hat{m}_0(1-\hat{m}_0)}{N_E} + \frac{\hat{m}_0(1-\hat{m}_0)}{N_{\bar{E}}} \right\}} \quad (2)$$

will, in the absence of bias, cover the causal risk difference in the exposed in 95 per cent of hypothetical repetitions under both a superpopulation and a randomization sampling model.

Furthermore, we will show that, when the causal risk difference is non-zero, the 95 per cent binomial interval will cover the causal risk difference more than 95 per cent of the time. In fact, as we will show, the interval estimator

$$[\hat{m}_1 - \hat{m}_0] \pm 1.96 \left[\frac{\hat{m}_1(1-\hat{m}_1)}{N_E} + \frac{\hat{m}_0(1-\hat{m}_0)}{N_{\bar{E}}} + \hat{R} \right]^{1/2}$$

where

$$\hat{R} = \frac{(2\hat{m}_0 - \hat{m}_1)(1 - \hat{m}_1) - \hat{m}_0(1 - \hat{m}_0)}{N_E + N_{\bar{E}}} \quad \text{if } \hat{m}_1 \geq \hat{m}_0 \quad (3)$$

$$\hat{R} = \frac{(2\hat{m}_1 - \hat{m}_0)(1 - \hat{m}_0) - \hat{m}_1(1 - \hat{m}_1)}{N_E + N_{\bar{E}}} \quad \text{if } \hat{m}_0 > \hat{m}_1$$

is strictly narrower than the 95 per cent binomial interval, and yet, in large samples, still covers the causal risk difference at least 95 per cent of the time.

2. DESCRIPTIVE EPIDEMIOLOGY

Consider an unmatched follow-up study in which at start of follow-up each of the N study subjects is either exposed (written E or E_1) or unexposed (E_0 or \bar{E}). Measurements are taken at the end of

follow-up on disease status with levels D and \bar{D} . The investigator observes the empirical distribution of E and D . For example, $p(D|E)$ is the proportion of exposed subjects who develop disease. $cRD = p(D|E) - p(D|\bar{E})$ is the crude risk difference. Such observable quantities constitute the parameters of interest for descriptive epidemiology.

3. A DETERMINISTIC MODEL USEFUL IN ETIOLOGIC RESEARCH

In etiologic research, on the other hand, the parameters of interest are intrinsically unobservable (that is, non-identifiable). For example, Miettinen and Cook¹ suggest the expression of causal parameters of interest in etiologic research in terms of comparisons (for example, the difference) between the observed number of cases that occurred in the exposed group (O) and the number of cases that one would have observed in the exposed group had that group been unexposed (that is, the expected number of cases, EX).

Specifically, we define $O/N_E - EX/N_E$ as the causal risk difference in the exposed where N_E is the number of exposed subjects. O/N_E is the observable parameter $p(D|E)$. EX is unobservable, since we cannot observe the outcome of exposed subjects when unexposed.

Note that to have a well-defined number for the expected number of cases in the exposed, we need to entertain a deterministic model in which each study subject is one of four possible types according to response to the presence and absence of exposure. Letting 1 indicate disease occurs and 0 indicate disease does not occur over the study period, we can tabulate the types in the following table.¹⁻⁴

'Common' description of type	Exposed	Unexposed
Type 1 No effect (individual 'doomed')	1	1
Type 2 Exposure causative (individual susceptible)	1	0
Type 3 Exposure preventive (individual susceptible)	0	1
Type 4 No effect (individual immune to disease)	0	0

We define N_j and N_{jE} , respectively, as the total number of subjects and the number of exposed subjects of type j with $j \in \{1, 2, 3, 4\}$. Similarly, $P_{jE} = N_{jE}/N_E$ is the proportion of exposed subjects of type j . Also,

$$O = N_{1E} + N_{2E}, EX = N_{1E} + N_{3E},$$

$$p(D|\bar{E}) = (N_{1\bar{E}} + N_{3\bar{E}}) / N_{\bar{E}} = P_{1\bar{E}} + P_{3\bar{E}}, \text{ and } (O/N_E - EX/N_E) = P_{2E} - P_{3E}.$$

Following Miettinen and Cook¹ and Greenland and Robins⁴ we say *there is no confounding for the causal risk difference in the exposed* if and only if the crude risk difference equals the causal risk difference in the exposed. This condition holds if and only if $p(D|E_0) = EX/N_E$ (that is, the empirical rate of disease in the unexposed equals the rate of disease that we would have observed in the exposed had they been unexposed). Since EX is an unobservable, we can never know whether confounding exists.

Miettinen and Cook¹ restricted their consideration to comparisons between O and EX . In fact, other causal parameters may have interest. For example, we may have interest in comparing the total number of cases that we would have observed if the entire population had been exposed ($N_1 + N_2$) to the total number of cases that we would have observed if the entire population had been unexposed ($N_1 + N_3$).^{2,3} Specifically, define $[(N_1 + N_2) - (N_1 + N_3)] / N = (N_2 - N_3) / N = P_2 - P_3$ as the causal risk difference. This is typically the parameter of interest in the analysis of

randomized trials. We say that there is *no confounding for the (population) causal risk difference if and only if* $cRD = (N_2 - N_3)/N$.

4. THE INTRODUCTION OF SAMPLING VARIABILITY

A superpopulation model

In the study shown in Table I, 40 of 100 exposed subjects developed disease. Then, according to our deterministic model, we know $p(D|E)$ is exactly 0.4. We can have no sampling variability since

1. each exposed subject's outcome is predetermined, and
2. we have not assumed that our study population is a sample drawn from a larger population.

In contrast, the standard approach most epidemiologists take is to report a binomial confidence interval of $0.4 \pm 1.96[(0.4)(0.6)/100]^{1/2}$ for the unknown parameter $p(D|E)$. The model implicit in the standard approach is as follows. The study population constitutes a random sample from a near-infinite superpopulation. The parameter $p(D|E)$ is the proportion of exposed subjects in the superpopulation who become diseased and $\hat{m}_1 = \hat{p}(D|E) = 40/100$ is the proportion of the (sampled) exposed study subjects who become diseased. Under this sampling model, the number of diseased exposed study subjects is a binomial random variable (upon hypothetical resamplings of 100 exposed study subjects from the superpopulation). Therefore, the 95 per cent binomial interval of equation (1) is a valid 95 per cent large-sample confidence interval for the superpopulation cRD .

An alternative model that gives binomial confidence intervals for $p(D|E)$ but does not require the introduction of a hypothetical superpopulation is a model in which

- (a) each exposed study subject's outcome is a Bernoulli random variable, and,
- (b) the probability of developing disease over the study interval is the same for all exposed study subjects.

$p(D|E)$ is that common probability. We reject this model as biologically implausible since the model implies no between-individual variation in any risk factor for disease. (When the probability of developing disease varies among exposed study subjects, the number of exposed study subjects who become diseased is not a binomial random variable, since the sum of the Bernoulli random variables is not a binomial random variable when the individual Bernoulli parameters are not constant.)

In summary, an investigator who (1) reports a binomial confidence interval for $p(D|E)$ and cRD and (2) who acknowledges that there exists between-individual variation in risk has implicitly assumed that (1) he/she has sampled the study subjects from a near-infinite superpopulation and (2) all inferences concern the parameters of that superpopulation.

Since, in most epidemiologic studies, study subjects do *not* constitute a random sample from any near-infinite superpopulation, the superpopulation model is a fiction. Why then does the model appear so frequently? I offer a possible reason.

An investigator often wishes to generalize his or her findings from the observed study population to some larger population. For example, an investigator, who considers a recommendation of a public health intervention, would hope to have studied a population that represents the population of potential recipients of that intervention. The simplest possible model is to consider the study population as a random sample of a larger population of potential recipients of the intervention. We can justify the use of a hypothetical superpopulation model (but not the use of confidence

intervals) on subjective Bayesian grounds by use of DeFinetti's theorem.⁵ DeFinetti's theorem implies that an investigator should not entertain a superpopulation model if

1. he/she believes the population of potential recipients differs from the study population on disease risk factors to an extent unaccountable by sampling variability, or
2. the size of the pool of potential recipients is not much larger than the size of the study population.

Suppose, in the superpopulation, we have no confounding for the causal risk difference and the causal risk difference in the exposed, that is $cRD = P_2 - P_3 = P_{2E} - P_{3E}$ (as would be the case if the distribution of the four types is the same among the exposed as among the unexposed superpopulation members, that is, $P_{jE} = P_{j\bar{E}}, j \in (1, 2, 3, 4)$). Then, the 95 per cent binomial interval is a valid 95 per cent confidence interval for the superpopulation causal risk difference and the superpopulation causal risk difference in the exposed.

However, some epidemiologists have less concern with the causal parameters of a possibly hypothetical superpopulation than with causal parameters associated with the observed study population.¹ That is, their interest centres on internal rather than external validity. In the next two sections we consider the construction of confidence intervals for causal parameters associated with the observed study population.

Randomization: An alternative sampling model

A number of epidemiologists (for example, Miettinen and Cook¹) make it clear that in a follow-up study they wish to treat

- (a) disease outcomes as deterministic, and
- (b) the causal parameters associated with the observed study population as the causal parameters of interest (regardless of whether that study population was a sample from a larger superpopulation or target population).

It is unlikely that any investigator would willingly assume that, in the observed study population, the distribution of risk factors balanced so well across exposure groups that there was no confounding for the causal risk difference or the causal risk difference in the exposed. None the less, an investigator might willingly make subjective statements such as 'although there may be some small association of risk factors with exposure, I do not believe such associations are systematic'.

In an observational study, we might approximate such a subjective statement by assuming that nature assigned exposure to N_E randomly chosen study subjects and left the remaining $N_{\bar{E}}$ subjects unexposed. That is, the observed study was a randomized trial performed by nature. We consider N_E and $N_{\bar{E}}$ fixed by design. Then the value of the empirical risk difference, $\hat{c}RD$, depends on the particular N_E subjects who received exposure. Therefore, in hypothetical rerandomizations, $\hat{c}RD$ has a well defined distribution. (The \sim appears over $\hat{c}RD$ to stress that we regard $\hat{c}RD$ as a random variable with a distribution defined by hypothetical rerandomizations. The value of $\hat{c}RD$ that we observe reflects the random exposure assignment that actually occurred.) In particular if we define (using the notation in Copas²) $m_1 = (N_1 - N_2)/N$, $m_0 = (N_1 + N_3)/N$, and $\beta = (N_2 + N_3)/N$, then $E(\hat{c}RD) = m_1 - m_0$ (Copas²). $m_1 - m_0$ is the causal risk difference. Thus, although in a randomized trial, in general, $\hat{c}RD \neq m_1 - m_0$, none the less $\hat{c}RD$ is unbiased. [Under this randomization model, one can characterize the population by the parameters (N_1, N_2, N_3, N_4) plus the constraint $N_1 + N_2 + N_3 + N_4 = N = N_E + N_{\bar{E}}$, or equivalently by the parameters (m_1, m_0, β, N) .]

Based on the following theorem, we can construct a conservative 95 per cent confidence interval for the causal risk difference that is narrower than the 95 per cent binomial interval.

Theorem 1

Under hypothetical rerandomizations, $\hat{c}RD$ is asymptotically normal with asymptotic variance.

$$N \text{ var}^A(\hat{c}RD) = \frac{m_1(1-m_1)}{P_E} + \frac{m_0(1-m_0)}{P_{\bar{E}}} + R \quad (4)$$

where $P_E = N_E/N$ and

$$R = m_0 + m_1 - 2m_0m_1 - m_1(1-m_1) - m_0(1-m_0) - \beta \quad (5)$$

where, by its definition, the range of possible β is

$$|m_0 - m_1| \leq \beta \leq \min(m_0 + m_1, 2 - m_0 - m_1).$$

For Proof see Appendix I.

Lemma 1

For all possible β , $R \leq 0$. The inequality is strict except when $\beta = 0$ and $m_1 = m_0$ or $|m_1 - m_0| = 1$. For Proof see Appendix I.

Now, we can estimate m_1 and m_0 unbiasedly and consistently by \hat{m}_1 and \hat{m}_0 , respectively, where again \hat{m}_0 is the proportion of unexposed subjects we observe to have disease.² Unfortunately, since β is non-identifiable,² we cannot estimate $N \text{ var}^A(\hat{c}RD)$ consistently. None the less, we can set conservative large sample confidence intervals (that is, confidence intervals guaranteed to cover at least their nominal rate) for $m_1 - m_0$ by deriving a consistent estimator, say $N \text{ var}^A(\hat{c}RD)$, of a number that is at least as large as $N \text{ var}^A(\hat{c}RD)$. We can accomplish this if, in equation (5), we replace β by a consistent estimator of its minimum possible value, that is, by $|\hat{m}_1 - \hat{m}_0|$, and, in both equation (4) and equation (5), we replace m_1 and m_0 by their empirical estimates \hat{m}_1 and \hat{m}_0 . On simplification, we have that if $\hat{m}_1 - \hat{m}_0 \geq 0$,

$$N \text{ var}^A(\hat{c}RD) = \frac{\hat{m}_1(1-\hat{m}_1)}{P_E} + \frac{\hat{m}_0(1-\hat{m}_0)}{P_{\bar{E}}} + (2\hat{m}_0 - \hat{m}_1)(1-\hat{m}_1) - \hat{m}_0(1-\hat{m}_0). \quad (6)$$

Equation (6) forms the basis for the interval estimator in equation (3).

If $\hat{c}RD$ was, as in the superpopulation model, the difference of two binomial proportions, a consistent estimate of its variance is the first two terms on the right hand side of equation (6). It then follows from Lemma 1 that, when the causal risk difference is non-zero, the usual 'binomial interval' for the causal risk difference is unnecessarily conservative since, taken together, the last two terms in equation (6) are negative when $\hat{m}_1 - \hat{m}_0 \neq 0$. In contrast, as pointed out by Copas, the usual binomial test of the null hypothesis $m_1 = m_0$ need not be conservative (since, in equation (4), $R = 0$ when $m_1 = m_0$ and $\beta = 0$).

Example 1

If, as in Table I, $N_E = 100$, $N_{\bar{E}} = 100$, $\hat{m}_1 = \frac{40}{100}$, $\hat{m}_0 = \frac{15}{100}$, then the 'binomial 95 per cent interval' for the causal risk difference, $m_1 - m_0$, is

$$0.25 \pm 1.96 \left[\frac{(0.4)(0.6) + (0.85)(0.15)}{100} \right]^{1/2} = 0.250 \pm 0.118$$

while a conservative 95 per cent confidence interval based on equation (3) is

$$0.25 \pm 1.96 \left[\frac{(0.4)(0.6) + (0.85)(0.15)}{100} + \frac{[2(0.15) - (0.4)](1 - 0.4) - (0.15)(0.85)}{200} \right]^{1/2} \\ = 0.250 \pm 0.102.$$

Thus, the 'binomial interval' is 16 per cent wider than necessary.

None the less, before one begins the routine report of confidence intervals for the causal risk difference in randomized trials based on equation (3), two caveats are in order. First, in moderate (as opposed to large) samples, it will happen frequently that 'Wald' type intervals based on $\hat{c}RD \pm 1.96 [\text{var}^A(\hat{c}RD)]^{1/2}$ (that is, equation (3)) will exclude the null even when standard χ^2 or Fisher exact tests for $m_1 = m_0$ do not 'reject'. In such cases, one should not 'reject' the null hypothesis.² (The problem with Wald confidence intervals is that they evaluate $\text{var}^A(\hat{c}RD)$ at $\hat{m}_1 - \hat{m}_0$, rather than at the null hypothesis $m_1 = m_0$.)

Second, confidence intervals based on equation (3) are valid only when a deterministic model for disease outcome is correct. If outcomes were, in fact, stochastic, the causal risk difference is defined as

$$\left\{ \sum_{i=1}^N [p(D|E, i) - p(D|\bar{E}, i)] \right\} / N$$

where, for example, $p(D|E, i)$ is the probability that subject i will become diseased if exposed. In this stochastic world, $\hat{c}RD$ is still unbiased for the causal risk difference in hypothetical rerandomizations but equation (4) is no longer valid for $\text{var}^A(\hat{c}RD)$. [For example, if $p(D|E, i)$ and $p(D|\bar{E}, i)$ did not depend on i , $\text{var}^A(\hat{c}RD)$ is the usual binomial variance.] One cannot empirically determine whether outcomes are stochastic or deterministic.

Suppose now, following Miettinen and Cook, that our interest is in the causal risk difference in the exposed (rather than in the causal risk difference $m_1 - m_0$) and we have again assumed that

- (a) disease outcomes are deterministic, and
- (b) nature conducts a randomized trial.

Then, the causal risk difference in the exposed is a random variable (and not a population parameter) since O and EX both depend on the particular set of exposed subjects. For a random variable, we define the analogue of a confidence interval as a *prediction interval*. Specifically, a prediction interval is a rule that gives, for any data realization, an interval. This interval may or may not contain the unobservable random variable $(O - EX) / N_E$ associated with that realization. If, in hypothetical rerandomizations, 95 per cent of realizations produce an interval that includes (the realization-specific) $(O - EX) / N_E$, we have, by definition, a 95 per cent prediction interval for $(O - EX) / N_E$.

Appendix I shows that a valid 95 per cent large sample prediction interval for $(O - EX) / N_E$ is

$$\hat{c}RD \pm 1.96 \left[\frac{\hat{m}_0(1 - \hat{m}_0)N}{N_{\bar{E}}N_E} \right]^{1/2}. \tag{7}$$

This prediction interval is not conservative, that is, it will almost cover $(O - EX) / N_E$ exactly 95 per cent of the time in large samples.

Example 2

Given the data in Table I, we compute a prediction interval for the causal risk difference in the

exposed as

$$0.250 \pm 1.96 [(0.15)(0.85)200 / (100)^2]^{1/2} = 0.250 \pm 0.099$$

under the randomization distribution. Thus, the usual 'binomial interval' (computed in example 1) is 18 per cent too wide. None the less, for certain states of nature, binomial intervals for the causal risk difference in the exposed (in contrast to binomial intervals for the causal risk difference) may be anti-conservative, that is they may fail to obtain their nominal coverage rate. To see this, note that we can rewrite equation (7) as equation (2). The prediction interval given in equations (7) and (2) differs from the usual binomial interval only in that the second term under the square root sign in equation (2) is $[\hat{m}_0(1 - \hat{m}_0)]/N_E$ rather than $[\hat{m}_1(1 - \hat{m}_1)]/N_E$. Thus, if $[\hat{m}_1(1 - \hat{m}_1)]/N_E$ is greater than $[\hat{m}_0(1 - \hat{m}_0)]/N_E$ (as in our example) binomial confidence intervals will be conservative. If $[\hat{m}_1(1 - \hat{m}_1)]/N_E$ is less than $[\hat{m}_0(1 - \hat{m}_0)]/N_E$, binomial confidence intervals will be anti-conservative. (This follows from the fact that, in large samples, the prediction interval given by equation (2) covers the causal risk difference in the exposed in precisely 95 per cent of hypothetical rerandomizations.) In fact, in the limit as $m_1 \rightarrow 1$ and $N_E/N \rightarrow 0$, the rate at which the usual binomial interval covers the random variable $(O-EX)/N_E$ approaches 0!

We could extend this randomization model to include measured covariates as follows. If we have obtained data on a dichotomous covariate C , and we believe that within levels of C there is no systematic association of exposure with unmeasured risk factors, then we could assume that, for $i \in (0, 1)$, nature exposed at random N_{EC} subjects and left $N_{\bar{E}C}$ subjects unexposed, where, for example, N_{EC} is the observed number of exposed subjects in stratum C_i .

Relationships between the superpopulation and randomization models

One might find it surprising that the variance of the cRD in the deterministic randomization model given by equation (4) is less than the binomial variance, even asymptotically. Here we try to make it intuitively clear why this is so, by, in a sense, embedding our randomization model within the superpopulation model.

Consider a superpopulation model with deterministic outcomes. We can formally approximate subjective statements such as 'exposure is not systematically associated with other risk factors in the observed study population' in our superpopulation model by assuming that $P_{jE} = P_{j\bar{E}}$, $j \in \{1, 2, 3, 4\}$. Under this assumption, $cRD = P_2 - P_3 = P_{2E} - P_{3E}$, that is, there exists no confounding in the superpopulation. Even so, we will, in general, have confounding for both the 'causal risk difference' and the 'causal risk difference in the exposed' in the observed (that is, sampled) study population, since $\hat{P}_{jE} = \hat{P}_{j\bar{E}}$, $j \in \{1, 2, 3, 4\}$ will be false due to chance associations of exposure with risk factors in the sample. (We have put a \sim over the 'parameters' of the observed study population to emphasize that these 'parameters' are random variables under hypothetical resamplings of the superpopulation.) Suppose an investigator accepts the superpopulation model as a sampling model, but his interest lies in the causal risk difference of the observed study population (that is, $\hat{P}_2 - \hat{P}_3$) rather than the causal risk difference of the superpopulation $P_2 - P_3$. $\hat{P}_2 - \hat{P}_3$ and $P_2 - P_3$ will, in general, differ from one another. This reflects the fact that $\hat{P}_2 - \hat{P}_3$ is a random variable with non-zero variance since its value depends on the particular N_E exposed and $N_{\bar{E}}$ unexposed subjects sampled from the superpopulation.

The large-sample results described below hold under a limiting model in which, as N_E and $N_{\bar{E}} \rightarrow \infty$, both N_E/M and $N_{\bar{E}}/M \rightarrow 0$ where M is the size of the superpopulation. That is, the sampled study population constitutes a negligible fraction of the superpopulation.

Consider the set of hypothetical resamplings of the superpopulation in which the vector $(\hat{N}_1, \hat{N}_2, \hat{N}_3, \hat{N}_4)$ is fixed at its observed value, but $(\hat{N}_{1E}, \hat{N}_{2E}, \hat{N}_{3E}, \hat{N}_{4E})$ is not fixed. Then, the conditional

distribution of this latter vector, given the former vector, is the same as the distribution of the vector $(\hat{N}_{1E}, \hat{N}_{2E}, \hat{N}_{3E}, \hat{N}_{4E})$ in a randomized trial with parameters equal to $(\hat{N}_1, \hat{N}_2, \hat{N}_3, \hat{N}_4)$. That is, we have generated the randomization distribution as a conditional distribution within the superpopulation model. It follows that the conditional expectation of the empirical crude risk difference $E[\hat{cRD} | (\hat{N}_1, \hat{N}_2, \hat{N}_3, \hat{N}_4)]$ is $(\hat{P}_2 - \hat{P}_3)$ and the conditional asymptotic variance is given by equation (4) where, $m_1, m_0,$ and $\beta,$ now will depend on the particular superpopulation subjects sampled. Now, it is a well-known identity that

$$\text{var}^A(\hat{cRD}) = \text{var}^A[E^A(\hat{cRD} | (\hat{N}_1, \hat{N}_2, \hat{N}_3, \hat{N}_4))] + E^A[\text{var}^A(cRD | (\hat{N}_1, \hat{N}_2, \hat{N}_3, \hat{N}_4))]$$

We have previously noted that in the superpopulation model, unconditionally, $\text{var}^A(\hat{cRD})$ is the usual binomial variance for the risk difference cRD . Furthermore, as discussed above, $\text{var}^A[E^A(\hat{cRD} | (\hat{N}_1, \hat{N}_2, \hat{N}_3, \hat{N}_4))] = \text{var}^A[\hat{P}_2 - \hat{P}_3]$ is non-zero and of the same order as $\text{var}^A(\hat{cRD})$. It follows that, in expectation, $\text{var}^A[\hat{cRD} | (\hat{N}_1, \hat{N}_2, \hat{N}_3, \hat{N}_4)]$ (given by equation (4)) must be less than the binomial variance.

Thus, equation (3) gives a valid conditional confidence interval for $\hat{P}_2 - \hat{P}_3$. It follows that this interval is also an unconditional 95 per cent prediction interval for the random variable $\hat{P}_2 - \hat{P}_3$. On the other hand, the unconditional confidence interval for $P_2 - P_3$ (which is the usual binomial interval) is wider because of the remaining uncertainty in $P_2 - P_3$ given knowledge of $\hat{P}_2 - \hat{P}_3$. Thus, an investigator who (1) accepts a deterministic superpopulation model but (2) has interest in the causal risk difference and/or the causal risk difference in the exposed of the (sampled) study population should report (unconditional) prediction intervals for the causal risk difference as given in equation (3) and for the causal risk difference in the exposed as given in equation (2).

Now suppose that disease outcomes were stochastic, and, for any population of size N , we define the causal risk difference as

$$\left\{ \sum_{i=1}^N [p(D|E, i) - p(D|\bar{E}, i)] \right\} / N.$$

If $p(D|E, i) - p(D|\bar{E}, i)$ depends on i , we can again use the identity that the unconditional variance is the expectation of the conditional variance plus the variance of the conditional expectation to show that a standard 95 per cent binomial interval for the causal risk difference of the sampled population is conservative.

5. CONFIDENCE INTERVALS FOR THE CAUSAL RISK RATIO

In this section, we provide 95 per cent confidence and prediction intervals for the causal risk ratio $(N_1 + N_2)/(N_1 + N_3)$ and the causal risk ratio in the exposed $(N_{1E} + N_{2E})/(N_{1E} + N_{3E})$ under the deterministic randomization model of Section 4.

A 95 per cent large sample prediction interval for the natural logarithm of the causal risk ratio in the exposed is

$$\ln(\hat{m}_1, \hat{m}_{1E}) \pm 1.96 \sqrt{\left\{ \frac{(1 - \hat{m}_0)}{\hat{m}_0 N_E} + \frac{(1 - \hat{m}_{0E})}{\hat{m}_{0E} N_E} \right\}} \tag{8}$$

For Proof see Corollary to Theorem A1.

A conservative 95 per cent confidence interval for the causal risk ratio that is strictly narrower than the standard 95 per cent binomial interval is

$$\ln(\hat{m}_1/\hat{m}_0) + 1.96 \sqrt{\left\{ \frac{(1 - \hat{m}_1)}{\hat{m}_1 N_E} + \frac{(1 - \hat{m}_0)}{(\hat{m}_0) N_{\bar{E}}} - R \right\}} \tag{9}$$

where

$$R = (\hat{m}_1 - \hat{m}_0) (\hat{m}_1 \hat{m}_0) \quad \text{if } \hat{m}_1 \geq \hat{m}_0,$$

$$R = (\hat{m}_0 - \hat{m}_1) (\hat{m}_1 \hat{m}_0) \quad \text{if } \hat{m}_0 > \hat{m}_1.$$

For Proof see Corollary A1 of the Appendix.

6. SUMMARY

Under a deterministic outcome model (in which each subject is one of four types depending on response to the presence and absence of exposure), we constructed a large-sample confidence (or prediction) interval for the causal risk difference that is narrower than the standard 'binomial interval' and yet attains its nominal coverage rate under both a randomization and a superpopulation sampling model. In addition, we show that standard binomial intervals for the causal risk difference in the exposed may be either conservative or anticonservative. To correct this deficiency, we constructed a valid large-sample prediction interval for the causal risk difference in the exposed.

Our confidence and prediction intervals are invalid if

1. individual disease outcomes are stochastic.
2. the causal parameters of interest are those of a hypothetical superpopulation rather than those of the observed study population, or
3. there is bias (for example there is confounding in the superpopulation).

ACKNOWLEDGEMENTS

Research for this paper was funded in part by grants from the following organizations: NIEHS # 5 R23 ES03405, NIEHS # 5 P30 ES00002, American Lung Association # LA 3 22 85, the American Heart Association, and the Massachusetts Chapter of the American Heart Association. The author wishes to thank Sander Greenland, David Freedman and Persi Diaconis for helpful advice.

APPENDIX I

We define $m_1, m_0, \hat{m}_1, \hat{m}_0, \hat{c}RD = \hat{m}_1 - \hat{m}_0, P_E, P_{\bar{E}}, N_E, N_{\bar{E}}, N,$ and β are as in Section 4 and N_1, N_2, N_3 and N_4 are as in Section 3.

Proof of Theorem 1

It follows from the equations on page 471 of Copas² (s, n_i in Copas is \hat{m}_i in our notation) that $E(\hat{c}RD) = m_1 - m_0$ and

$$\text{var}(\hat{c}RD) = \frac{1}{N-1} \left[\frac{N_{\bar{E}}}{N_E} m_1 (1 - m_1) + \frac{N_E}{N_{\bar{E}}} m_0 (1 - m_0) + m_0 + m_1 - 2m_0 m_1 - \beta \right] \quad (10)$$

where $|m_1 - m_0| \leq \beta \leq \min(m_0 + m_1, 2 - m_0 - m_1)$.

We next show that $\hat{c}RD$ is asymptotically normal as $N \rightarrow \infty$ with N_E/N fixed:

$$\hat{c}RD = \frac{N_{1E} + N_{2E}}{N_E} - \frac{(N_1 + N_3) - (N_{1E} + N_{3E})}{N_{\bar{E}}}.$$

but the random vector (N_{1E}, N_{2E}, N_{3E}) is jointly asymptotically normal since the exposed subpopulation constitutes a simple random sample of the total population taken without replacement. Therefore $\hat{c}RD$ is asymptotically normal, since $\hat{c}RD$ is a linear combination of the random variables N_{1E} , N_{2E} , and N_{3E} . Equation (10) then implies that $N \text{ var}^A(\hat{c}RD)$ is given by equation (4), since equation (4) derives from equation (10) by replacing $N - 1$ by N and rearranging terms.

Proof of Lemma 1

It is straightforward to show that $R = 0$ if $\beta = 0$ and $m_1 = m_0$. Since R is a strictly decreasing function of β , we have, as noted by Copas, that $R < 0$ if $m_1 = m_0$ and $\beta \neq 0$.²

Next, without loss of generality, assume $m_1 - m_0 > 0$. Then, from the bounds on the range of β , we have, upon substituting $m_1 - m_0$ for β in equation (5), that $R \leq (2m_0 - m_1)(1 - m_1) - m_0(1 - m_0)$. Write $m_1 = m_0 + x$. Then

$$\frac{\hat{c}R}{\hat{c}x} = -(2m_0 - m_1) - (1 - m_1) = 2(m_1 - m_0) - 1 = 2x - 1.$$

Therefore, for $0 < x < \frac{1}{2}$, R is a strictly decreasing function of x , and thus it follows that $R < 0$ from the results given in the preceding paragraph for the case $m_1 = m_0$.

It only remains to show that for $\frac{1}{2} < x < 1$, $R < 0$. Since for $\frac{1}{2} < x < 1$, R is strictly increasing in x , we need only evaluate R at $\max(x) = -m_0$, that is, at $m_1 = 1$, but at $m_1 = 1$, $R = -m_0(1 - m_0) < 0$, provided $m_0 \neq 0$.

Corollary A1

In a deterministic outcome model, under hypothetical rerandomizations, a conservative large-sample 95 per cent confidence interval for the natural logarithm of the causal risk ratio is given by equation (9).

Proof: Since $E(\hat{m}_1) = (N_1 + N_2) / N$ and $E(\hat{m}_0) = (N_1 + N_3) / N$, it follows that

$$E^A[\ln(\hat{m}_1, \hat{m}_0)] = \ln[(N_1 + N_2) / (N_1 + N_3)].$$

Furthermore,

$$\begin{aligned} N \text{ var}^A[\ln(\hat{m}_1, \hat{m}_0)] &= N \left\{ \frac{1}{m_1^2} \text{var}^A(\hat{m}_1) + \frac{1}{m_0^2} \text{var}^A(\hat{m}_0) - \frac{2}{m_1 m_0} \text{cov}^A(\hat{m}_1, \hat{m}_0) \right\} \\ &= \left[\frac{N_{\bar{E}}(1 - m_1)}{N_{\bar{E}} m_1} + \frac{N_{\bar{E}}(1 - m_0)}{N_{\bar{E}} m_0} + \frac{m_0 + m_1 - 2m_0 m_1 - \beta}{m_1 m_0} \right] \\ &= \frac{(1 - m_1)}{m_1 P_{\bar{E}}} + \frac{(1 - m_0)}{m_0 P_{\bar{E}}} + \frac{m_0 + m_1 - 2m_0 m_1 - B - m_0(1 - m_1) - m_1(1 - m_0)}{m_1 m_0} \end{aligned}$$

where, again, we have used the equations on page 471 of Copas.² Equation (9) follows upon substituting $|\hat{m}_1 - \hat{m}_0|$ for β and \hat{m}_1 and \hat{m}_0 for m_1 and m_0 , respectively.

Theorem A1

Under hypothetical rerandomizations,

$$\hat{c}RD \pm 1.96 \sqrt{\left\{ \frac{\hat{m}_0(1 - \hat{m}_0)N}{(N_{\bar{E}} N_{\bar{E}})} \right\}} \tag{11}$$

is a 95 per cent large-sample prediction interval for

$$(O-EX)/N_E \equiv [(N_{1E} + N_{2E}) - (N_{1E} + N_{3E})]/N.$$

Proof: Since $N_{1E} + N_{3E} = N_1 + N_3 - N_{1\bar{E}} - N_{3\bar{E}}$, we have

$$(O-EX)/N_E = \left[\hat{m}_1 + \hat{m}_0 \frac{N_{\bar{E}}}{N_E} \right] - \frac{m_0 N}{N_E}.$$

Now $\hat{m}_1 + \hat{m}_0 \frac{N_{\bar{E}}}{N_E}$ is an observed random variable and $\frac{m_0 N}{N_E}$ is a population parameter.

Furthermore, since $E(\hat{m}_0) = m_0$ and, as shown by Copas,

$$\text{var}(\hat{m}_0) = \frac{N_E m_0 (1 - m_0)}{(N - 1) N_{\bar{E}}}$$

a 95 per cent large-sample confidence interval for $\frac{N}{N_E} m_0$ is

$$\frac{N}{N_E} \hat{m}_0 \pm 1.96 \frac{N}{N_E} \sqrt{\left\{ \frac{N_E \hat{m}_0 (1 - \hat{m}_0)}{N N_{\bar{E}}} \right\}} = \frac{N}{N_E} \hat{m}_0 \pm 1.96 \sqrt{\left\{ \frac{\hat{m}_0 (1 - \hat{m}_0) N}{(N_E N_{\bar{E}})} \right\}}.$$

This implies that a 95 per cent prediction interval for $(O-EX)/N_E$ is

$$\hat{m}_1 + \hat{m}_0 \frac{N_{\bar{E}}}{N_E} - \frac{N}{N_E} \hat{m}_0 \pm 1.96 \sqrt{\left\{ \frac{\hat{m}_0 (1 - \hat{m}_0) N}{(N_E N_{\bar{E}})} \right\}}$$

which simplifies to equation (11).

Corollary: A 95% large-sample prediction interval for $\ln(O, EX)$ is given by equation (8). Using the fact that

$$\ln(EX/N_E) = \ln\left(\hat{m}_0 \frac{N_{\bar{E}}}{N_E} - \frac{m_0 N}{N_E}\right)$$

and that

$$\hat{m}_0 \frac{N_{\bar{E}}}{N_E} - \frac{m_0 N}{N_E} = -\hat{m}_0,$$

we expand $\ln\left(\hat{m}_0 \frac{N_{\bar{E}}}{N_E} - \frac{m_0 N}{N_E}\right)$ around $\frac{\hat{m}_0 N}{N_E}$ in a Taylor series to obtain

$$\ln(EX/N_E) = \ln(-\hat{m}_0) + \frac{1}{\hat{m}_0} ((m_0 - \hat{m}_0)N/N_E) + O_p(1/N)$$

since $((m_0 - \hat{m}_0)N/N_E)^2$ is $O_p(1/N)$.

Therefore, by the proof of Theorem A1, a 95 per cent large-sample prediction interval for $\ln(O/EX) = \ln(\hat{m}_1) - \ln(EX/N_E)$ is

$$\ln(\hat{m}_1/\hat{m}_0) \pm \frac{1}{m_0} 1.96 \sqrt{\left\{ \frac{\hat{m}_0 (1 - \hat{m}_0) N}{N_E N_{\bar{E}}} \right\}}$$

which can be rewritten as equation (8).

REFERENCES

1. Miettinen, O. S. and Cook, E. F. 'Confounding: essence and detection', *American Journal of Epidemiology*, **114**, 593-603 (1981).
2. Copas, J. B. 'Randomization models for matched and unmatched 2×2 tables', *Biometrika*, **60**, 467-476 (1973).
3. Rubin, D. B. 'Bayesian inference for causal effects: The role of randomization', *Annals of Statistics*, **6**, 34-58 (1978).
4. Greenland, S., and Robins, J. M. 'Identifiability, exchangeability, and epidemiological confounding', *International Journal of Epidemiology*, **15**, 413-419 (1986).
5. DeFinetti, B. *Probability, Induction, and Statistics*, Wiley, New York, 1972.