A Small Sample Correction for Estimating Attributable Risk in Case-Control Studies

Daniel B. Rubin, Food and Drug Administration
A Small Sample Correction for Estimating Attributable Risk in Case-Control Studies

Daniel B. Rubin

Abstract

The attributable risk, often called the population attributable risk, is in many epidemiological contexts a more relevant measure of exposure-disease association than the excess risk, relative risk, or odds ratio. When estimating attributable risk with case-control data and a rare disease, we present a simple bias correction to the standard approach, which also makes it more stable and less variable. As with analogous corrections given by Jewell (1986) for other measures of association, the adjustment often won't make a substantial difference unless the sample size is very small or point estimates are desired within fine strata, but we discuss the possible utility for applications.

KEYWORDS: attributable risk, case-control studies, small sample corrections

Author Notes: This research was conducted while the author was a graduate student at the University of California, Berkeley and is unrelated to his duties at the Food and Drug Administration.
1 Attributable risk

Although an exposure may be strongly associated with a disease, an intervention removing the risk factor can produce a limited public health benefit if the exposure is uncommon. Herein lies the interpretational problem with the measure of exposure-disease association

$$RR = \text{relative risk} = \frac{\text{pr(disease|exposed)}}{\text{pr(disease|unexposed)}}.$$ 

A different measure of association is Levin’s (1953) attributable risk

$$AR = \{\text{pr(disease)} - \text{pr(disease|unexposed)}\} / \text{pr(disease)} = \text{pr(exposed|disease)} (1 - 1/RR),$$

where equality of the two lines can be seen from (5.1) of Jewell (2003). The attributable risk is informally the proportion of cases that could be saved by eliminating exposure, assuming no confounding. For instance, a value of 0.1 suggests a tenth of cases are due to exposure. When contemplating interventions, such knowledge can be more relevant than the relative risk.

While the attributable risk is not directly identifiable with case-control sampling, for a rare disease it is approximately

$$AR^* = \text{pr(exposed|disease)} (1 - 1/OR),$$

for

$$OR = \text{odds ratio} = \frac{\text{pr(disease|exposed)}}{\text{pr(disease free|exposed)}} \times \frac{\text{pr(disease free|unexposed)}}{\text{pr(disease|unexposed)}}.$$ 

The reasoning behind the approximation is that

$$AR^* - AR = \frac{\text{pr(exposed|disease)}}{RR} \left(1 - \frac{\text{pr(disease free|exposed)}}{\text{pr(disease free|unexposed)}}\right),$$

and the factor on the right will be close to zero if the disease is uncommon for both the exposed and unexposed. Case-control studies are most appropriate for rare diseases, so the following discussion assumes $AR^* \approx AR$, and we henceforth take $AR^*$ as our desired measure of association.
In a case-control study, we assess exposure status for \( m \) cases and \( n \) controls. The statistical model is

\[
\begin{align*}
\text{number of cases exposed} &= A \sim \text{Binomial}(m, q), \ 0 < q < 1 \\
\text{number of controls exposed} &= B \sim \text{Binomial}(n, p), \ 0 < p < 1, \ A \perp B \\
\text{number of cases unexposed} &= C = m - A \\
\text{number of controls unexposed} &= D = n - B.
\end{align*}
\]

We can now formally define our parameter of interest as

\[
AR^* = q \left( 1 - \frac{1 - q}{q} \frac{p}{1 - p} \right).
\]

2 Small sample correction

For additional details we refer to Whittemore (1982) or Jewell (2003, Section 7.4), but the usual maximum likelihood estimator is given by

\[
\hat{AR} = \frac{AD - BC}{(A+C)D}.
\]

There is ambiguity when \( D = 0 \), and if the definition is taken literally the estimator has bias of \(-\infty\). Since \( A+C = m \), this part of the two denominators does not cause problems. We propose the small sample correction

\[
\hat{AR}_{SS} = \frac{A}{A+C} - \frac{BC}{(A+C)(D+1)},
\]

where the SS subscript refers to “Small Sample.” The following result is proven in the appendix, and states that bias decreases exponentially with the number of controls.

**Lemma 1.** bias(\(\hat{AR}_{SS}\)) = \(E[\hat{AR}_{SS}] - AR^* = \left( \frac{1-q}{1-p} \right) p^{n+1} = O(e^{-n})\).

For problems in statistics where exact unbiasedness is not possible, exponentially decreasing bias is somewhat unusual. For example, the ratio of two empirical means has bias \( \Omega(n^{-1}) \) in estimating the ratio of population means, where \( n \) is the sample size. The jackknife was originally introduced to reduce such biases to \( O(n^{-2}) \) (Quenouille, 1956), which is a far cry from \( O(e^{-n}) \).
In line with its interpretation of yielding the percentage of cases due to exposure, attributable risk is often reported to two decimal places. Hence, say our estimator is essentially unbiased if bias is less than 0.005. If the exposure is not ubiquitous, so less than half of controls are expected to be exposed, then only nine controls are needed to ensure essential unbiasedness.

Ideally we would only investigate exposures that are associated with disease, or at least neutral. In such a setting, the $(1 - q)/(1 - p)$ factor in the bias is no larger than one.

Table 1 compares the bias of the standard and corrected estimator. As mentioned, the standard estimator does not even have finite bias because $D = 0$ can occur in the denominator, but we level the playing field by conditioning on $\{D \neq 0\}$. The exact bias of the standard estimator can then be calculated by enumerating outcomes for the binomial distribution. Results are shown for a handful of parameter values $(q, p)$, for a small study with ten cases and ten controls. The corrected estimator clearly seems preferable. If we did not condition on $\{D \neq 0\}$ and instead examined unconditional bias through Lemma 1, we would find that for all values in the table the corrected estimator has bias less than 0.0001.

3 Comparison to other adjustments

For estimating different functions of binomial parameters, many adjustments to maximum likelihood have been suggested. For instance, Agresti and Coull (1998) propose forming a confidence interval for a binomial rate by adding two counts to both the number of successes and failures before forming the usual Wald interval. Jewell (1986) proposes estimators for the relative risk and odds ratio that correspond to adding counts to different cells of 2x2 tables. Likewise, Gart (1966) proposes adding 1/2 to each cell of a contingency table before estimating the logarithm of an odds ratio, which can then be used as the basis for a confidence interval.

To compare our bias-corrected $\hat{AR}_{SS}$ to other possible adjustments, we consider estimators $\hat{AR}(A, B, C, D + d)$ and $\hat{AR}(A + d, B + d, C + d, D + d)$ corresponding to adding a constant to the fourth cell (as with $d = 1$ in our method) or to all cells of the 2x2 contingency table.
Table 1: Bias of the standard and corrected estimator, with \( m = 10 \) cases and \( n = 10 \) controls, conditional on \( \{ D \neq 0 \} \). The corrected estimator appears to have smaller bias.

| \( q \) | \( p \) | \( AR^* \) | \( \text{bias}(\overline{AR}|D \neq 0) \) | \( \text{bias}(\overline{AR}_{SS}|D \neq 0) \) |
|---|---|---|---|---|
| 0.1 | 0.1 | 0.0000 | -0.0127 | 0.0000 |
| 0.2 | -0.1250 | -0.0335 | 0.0000 |
| 0.3 | -0.2857 | -0.0694 | 0.0001 |
| 0.4 | -0.5000 | -0.1360 | 0.0009 |
| 0.2 | 0.1 | 0.1111 | -0.0113 | 0.0000 |
| 0.2 | 0.0000 | -0.0297 | 0.0000 |
| 0.3 | -0.1429 | -0.0617 | 0.0000 |
| 0.4 | -0.3333 | -0.1209 | 0.0008 |
| 0.3 | 0.1 | 0.2222 | -0.0099 | 0.0000 |
| 0.2 | 0.1250 | -0.0260 | 0.0000 |
| 0.3 | 0.0000 | -0.0540 | 0.0000 |
| 0.4 | -0.1667 | -0.1058 | 0.0007 |
| 0.4 | 0.1 | 0.3333 | -0.0085 | 0.0000 |
| 0.2 | 0.2500 | -0.0223 | 0.0000 |
| 0.3 | 0.1429 | -0.0463 | 0.0000 |
| 0.4 | 0.0000 | -0.0906 | 0.0006 |

The bias of any method depends on the distribution. To evaluate adjustments we consider absolute bias \( \int_R |E_{q,p}[\overline{AR}(A, B, C, D + d)] - AR^*| dq dp \) averaged over region \( R = \{(q, p) : (q, p) \in [0.1, 0.5]^2, q \geq p\} \), and likewise for the adjustment adding a constant to all cells. Exposure rates in this region are higher for cases than controls, as attributable risk is most meaningful when exposure is associated with disease or neutral. In this region the exposure is not so common that more than half of cases or controls will be exposed, but is common enough so that at least one tenth of the population is exposed, as otherwise the average absolute bias is dominated by outliers.

We considered bias in a small study with \( m = 25 \) cases and \( n = 25 \) controls. To evaluate the integral by Monte Carlo, exposure rates \((q,p)\) were uniformly drawn 5,000 times from region \( R \). After each \((q,p)\) draw, biases for adjusted estimators were calculated from the binomial distribution, and absolute biases were then averaged. Results are shown in Figure 1 for adjustment constant \( d \) between 0 and 3. As suggested by the statistical theory, our \( \overline{AR}_{SS} \) gives the smallest bias among all adjustments corresponding to adding a constant to the fourth cell of the contingency table or adding a constant to all cells.
Figure 1: For a study with $m = 25$ cases and $n = 25$ controls, 1 and 2 show absolute bias for small sample adjustments $\hat{AR}(A, B, C, D + d)$ and $\hat{AR}(A + d, B + d, C + d, D + d)$ when averaged over region $R$ of exposure rates. Our adjustment $\hat{AR}_{SS} = \hat{AR}(A, B, C, D + 1)$ gives the smallest bias.

4 Variance

The bias reduction would be a Pyrrhic victory if it induced a large spike in variance. Fortunately, this does not occur because the correction tends to stabilize things. The standard method has infinite variance since $D = 0$ can appear in the denominator, but even after conditioning on the estimator being properly defined, the small sample correction still improves variance. The following result makes this precise, and our proof is in the appendix.
Lemma 2. $\text{var}(\hat{AR}_{SS}|D \neq 0) < \text{var}(\hat{AR}|D \neq 0)$.

The correction is thus meant to reduce both bias and variance, the components of mean squared error, or the traditional measure of risk in statistical decision theory. Our problem is a counterexample to the common situation where bias correction is ill-advised since the variance increase is prohibitive (Doss and Sethuraman, 1989). For instance, Efron and Tibshirani (1993) recommend against correcting an estimator through a bootstrap estimate of bias, due to increased variability. However, there are other examples such as the “German tank problem” of estimating the parameter $\theta$ of a uniform $U(0, \theta)$ distribution in which a bias correction to the MLE improves the overall mean squared error (see Le Cam, 1991, Example 5).

5 Relaying uncertainty

Because the sampling distribution of the usual maximum likelihood estimator is skewed (Jewell, 2003), it is not necessarily informative to attach a standard error to this estimator, and a preferred approach is to form a confidence interval for the attributable risk. The same considerations apply when using small sample corrections. Our recommendation for practice is that our bias correction be used for point estimation, but that standard methods be used for interval estimation. Typically we would work with a normal approximation on the $\log(1 - AR^*)$ scale, and then convert the confidence interval back to one for the desired $AR^*$. Such methods are discussed in Jewell (2003), and in small samples a typical correction for interval estimation would be to add $1/2$ to each of $(A, B, C, D)$.

6 Does the correction make a difference?

Our method has several theoretical advantages that we have discussed, it is simple enough to be implemented with a single line of code, and we avoid definitional problems when $D = 0$. The procedure could prove useful with a small sample size, or when confounding necessitates cross-tabulation and point estimates are desired within relatively fine strata. However, replacing $D$ by $D + 1$ in the estimator will often result in a very minor perturbation, and in this sense our proposal brings to mind the longstanding but frequently meaningless debate over whether to multiply a sample variance by $n/(n-1)$. 

DOI: 10.2202/1557-4679.1252
To see the utility of our method, we should note that it is an analog of relative risk and odds ratio corrections of Jewell (1986). Adjustments follow from the fact that if \( X \sim \text{Bin}(n, p) \), then \((n + 1)/(X + 1)\) and \(X/(n - X + 1)\) are corrected estimators of \(1/p\) and \(p/(1-p)\). Jewell (2003, Section 7.1) states that the “principal value of a small sample adjustment is in its alerting us to situations where the sample size is small enough to have a noticeable impact on an estimator, thereby suggesting that large sample approximations may be suspicious.” This seems like good advice for attributable risk estimation, for which our method should have a modest but positive impact.

7 Example

Jewell (2003) illustrates small sample odds ratio corrections with case-control data from MacMahon et al. (1981) concerning risk factors for pancreatic cancer. Considering coffee consumption as a risk factor, define exposed individuals as those who have one or more cups per day, and unexposed individuals as those who drink no cups per day, and take case or control status to be based on the presence or absence of pancreatic cancer. The values of \((A, B, C, D)\) for men in the study are \((207, 275, 9, 32)\). The rare disease assumption is not in question for this uncommon but often fatal condition. The maximum likelihood estimate of attributable risk is \(\hat{AR} = 0.600\). A standard 95\% confidence interval based on asymptotic normality of \(\log(1 - \hat{AR})\) is \([0.180, 0.805]\), and the small sample correction alluded to in Section 5 yields a similar interval of \([0.163, 0.794]\). This is evidence of a strong association between coffee drinking and pancreatic cancer among men. However, as the counts for \(C\) and \(D\) are relatively small, it is not immediately clear that consistency and asymptotic normality properties can be relied upon when interpreting the point estimate and confidence interval. The corrected point estimate is \(\hat{AR}_{SS} = 0.611\), essentially indistinguishable from the uncorrected estimate of 0.600. This suggests that small sample bias should be much less of a concern than sampling variability, and gives us no reason to distrust asymptotic approximations.
Appendix

Proof of Lemma 1. We begin by noting that since $B \sim \text{Bin}(n, p)$, we have

$$E \left[ \frac{B}{D+1} \right] = E \left[ \frac{B}{n-B+1} \right] = \sum_{k=0}^{n} \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k} \frac{k}{n-k+1}$$

$$= \sum_{k=1}^{n} \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k} \frac{k}{n-k+1}$$

$$= \frac{p}{1-p} \sum_{k=1}^{n} \frac{n!}{(k-1)!(n-(k-1))!} p^{k-1} (1-p)^{n-(k-1)}$$

$$= \frac{p}{1-p} \sum_{i=0}^{n-1} \frac{n!}{i!(n-i)!} p^i (1-p)^{n-i} \text{ for } i = k - 1$$

$$= \frac{p}{1-p} (1 - \text{pr}(B = n)) = \frac{p}{1-p} - \frac{p^n}{1-p} = \frac{p}{1-p} - \frac{p^{n+1}}{1-p}.$$

Further, $E[A/(A + C)] = E[A/m] = q$ since $A \sim \text{Bin}(m, q)$. Likewise, $E[C/(A + B)] = 1 - q$. By the independence of $(A, C)$ and $(B, D)$, we conclude that

$$E[\hat{A}R_{SS}] = E \left[ \frac{A}{A+C} \right] - E \left[ \frac{C}{A+C} \frac{B}{D+1} \right]$$

$$= E \left[ \frac{A}{A+C} \right] - E \left[ \frac{C}{A+C} \right] E \left[ \frac{B}{D+1} \right]$$

$$= q - (1-q) \frac{p}{1-p} + \left( \frac{1-q}{1-p} \right) p^{n+1}.$$

The desired result now follows after expressing

$$AR^* = q \left( 1 - \frac{1-q}{q} \frac{p}{1-p} \right) = q - (1-q) \frac{p}{1-p}. \square$$
Proof of Lemma 2. We first write
\[ \hat{AR} = \hat{AR}_{SS} + \frac{A - m}{m} \left( \frac{n - D}{D} - \frac{n - D}{D + 1} \right), \]
yielding
\[ \text{var}(\hat{AR}|D \neq 0) = \text{var}(\hat{AR}_{SS}|D \neq 0) \]
\[ + \text{var} \left[ \frac{A - m}{m} \left( \frac{n - D}{D} - \frac{n - D}{D + 1} \right) |D \neq 0 \right] \]
\[ + 2 \text{cov} \left[ \hat{AR}_{SS}, \frac{A - m}{m} \left( \frac{n - D}{D} - \frac{n - D}{D + 1} \right) |D \neq 0 \right]. \]

We will show that the covariance term is nonnegative, implying the desired inequality. We can write the covariance as \( \text{cov}(f(A, D), g(A, D)|D \neq 0) \), where both \( f \) and \( g \) are increasing in \( A \) and \( D \).

Note that for functions \( h_1 \) and \( h_2 \) and a random variable \( X \) such that \( h_1(X) \) and \( h_2(X) \) are bounded, that \( \text{cov}(h_1(X), h_2(X)) \geq 0 \) if both \( h_1 \) and \( h_2 \) are monotone increasing or monotone decreasing. See, for instance, Egozcue et al. (2009) or the references therein for this inequality that they attribute to Chebyshev.

To compress notation, suppose we have already conditioned on \( \{D \neq 0\} \) in the formulas below. The law of total covariance tells us
\[ \text{cov}(f(A, D), g(A, D)) = E[\text{cov}(f(A, D), g(A, D)|D)] \]
\[ + \text{cov}(E[f(A, D)|D], E[g(A, D)|D]). \]

The aforementioned Chebyshev inequality tells us both terms in the sum are nonnegative, implying the desired result. For the first term, after conditioning on \( D \) we see that \( f(A, D) \) and \( g(A, D) \) are increasing functions of \( A \), so the covariance inside the expectation should be nonnegative. For the second term, both \( E[f(A, D)|D] \) and \( E[g(A, D)|D] \) are increasing functions of \( D \), as both \( f \) and \( g \) are increasing in \( A \) and \( D \), so the covariance is again nonnegative. □
References

[1] Agresti, A. and Coull, B.A. (1998). Approximate is better than “exact” for interval estimation of binomial proportions. The American Statistician, 52, 199-126.

[2] Doss, H. and Sethuraman, J. (1989). The price of bias reduction when there is no unbiased estimate. Annals of Statistics, 17, 440-442.

[3] Benichou, J. (1991). Methods of adjustment for estimating the attributable risk in case-control studies: a review. Statistics in Medicine, 10, 1753-1773.

[4] Egozcue, M., Fuentes Garcia, L, and Wong, W. (2009). On some covariance inequalities for monotonic and non-monotonic functions. Journal of Inequalities in Pure and Applied Mathematics, volume 10, issue 3, article 75.

[5] Efron and Tibshirani (1993). An Introduction to the Bootstrap. London, UK: Chapman & Hall/CRC.

[6] Gart, J.J. (1966). Alternative analyses of contingency tables. Journal of the Royal Statistical Society, Series B, 28, 164-179.

[7] Jewell, N.P. (1986). On the bias of commonly used measures of association for 2 × 2 tables. Biometrics, 42, 351-358.

[8] Jewell, N.P. (2003). Statistics for Epidemiology. Boca Raton, FL: Chapman & Hall/CRC.

[9] Le Cam, L. (1991). Maximum likelihood, an introduction. University of California, Berkeley Department of Statistics technical report 168. Available at www.stat.berkeley.edu/users/rice/LeCam/papers/tech168.pdf.

[10] Levin, M.L. (1953). The occurrence of lung cancer in man. Acta Unio Int Contra Cancrum, 9, 531-541.

[11] Lui, K.J. (2001). Interval estimation of the attributable risk in case-control studies with matched pairs. Journal of Epidemiology and Community Health, 55, 885-890.
[12] MacMahon, B., Yen, S., Trichopoulos, D., Warren, K., and Nardi, G. (1981). Coffee and cancer of the pancreas. *New England Journal of Medicine*, 304, 630-633 and 1604-1606.

[13] Quenouille, M.H. (1956). Notes on bias in estimation. *Biometrika*, 43, 353-360.

[14] Whittemore, A.S. (1982). Statistical methods for estimating attributable risk from retrospective data. *Statistics in Medicine*, 1, 229-243.