On Test of Association using Attributable Risk for a 2 × 2 Contingency Table

Tanweer Shapla[1], Khairul Islam[2]
[1]Eastern Michigan University, 515 Pray-Harrold, Ypsilanti, MI 48197
[2]Texas A&M University-Kingsville, 700 University Blvd., Kingsville, TX 78363

Abstract

Attributable risk is a widely used measure for assessing risk of a factor in public health and biostatistics. It provides the proportion of disease reduction due to the elimination of the risk factor from the population of interest. It is to be noted that attributable risk is rarely used for test of association or independence. The chi-square test investigates if a certain factor is independent of any outcome. However, if the null hypothesis of independence or no association is rejected, this test cannot provide any insight on whether the factor is associated positively or negatively. This paper considers test of hypothesis of independence or no association using attributable risk and discusses sensitivity of power analysis, both theoretically and by Monte Carlo simulation.

Key Words: Attributable risk, independence, power of the test, Monte Carlo simulation

1. Introduction

Attributable risk (AR) is a widely used measure of risk of a factor in public health and epidemiological research. Introduced by Levin (1953), attributable risk is defined as the proportion of disease in the population that could be avoided if the effect associated with the risk factor were totally eliminated from the population of interest. For a dichotomous risk factor with the exposure status $E = i, i = 0,1$ and disease status $D = j, j = 0,1$, AR can be expressed by the expression $AR = \frac{P(D=1)-P(D=1|E=0)}{P(D=1)}$, where $P(D = 1)$ is the overall disease rate in the entire population and $P(D = 1|E = 0)$ is the disease rate in the unexposed population. Using the total law of probability, $P(D = 1) = P(D = 1|E = 0)P(E = 0) + P(D = 1|E = 1)P(E = 1)$, it follows that $AR = \frac{[P(D=1|E=1)-P(D=1|E=0)]P(E=1)}{P(D=1)}$. Other useful equivalent expressions also appear in the literature: $AR = P(E = 1|D = 1)\left(\frac{RR-1}{RR}\right)$, Benichou (1991); Coughlin et al. (1994), or $AR = \frac{P(E=1)(RR-1)}{P(E=1)(RR-1)+1}$,Benichou (1991); Coughlin et al. (1994); Walter (1975, 1976); Fleiss (1979) where $RR = \frac{P(D=1|E=1)}{P(D=1|E=0)}$ is the relative risk of disease in the exposed group. Etiologic fraction (Miettinen, 1974), attributable fraction (Ouellet et al. 1979) and population attributable risk per cent (Cole and MacMahon, 1971) are some other terminologies available in literature to refer to attributable risk.

[1]Corresponding author: tshapla@emich.edu

Walter (1975, 1976) provided distribution and rationale for using attributable risk in health research. Basu and Landis (1995) provided a model-based estimation of AR for a cross-sectional sampling along with the derivation of the asymptotic variance via Taylor series expansion. Benichou (1991) reviewed methods of adjustment for estimation of AR in case-control studies. Whittemore (1982) provided an asymptotic variance of the maximum likelihood estimate of the attributable risk for case-control data in the presence of confounding factors.

Other papers discussed the methods for estimating AR while controlling for confounder factors for various study designs, Benichou (1991); Coughlin et al. (1994); Basu and Landis (1995); Bruzzi et al. (1985); Benichou and Gail (1990); Benichou (2001); Drescher and Schill (1991); Greenland and Drescher (1993); Eide and Gefeller (1995); Graubard and Fears (2005); Cox (2006); Shapla et al. (2009); Islam and Shapla (2013).

Due to simplicity of application and interpretation, attributable risk for a dichotomous risk factor with a dichotomous outcome is very popular in biomedical and health sciences for measuring risk of the factor for the development of the outcome. In this paper, we will explore test of hypothesis of independence or no association of the risk factor and disease outcome for a cross-sectional study using attributable risk. Real life examples and simulation studies will be utilized to justify the usefulness of the test of hypothesis of independence or no association using attributable risk and discuss sensitivity of power analysis.

2. Methods

Let us consider a risk factor with dichotomous exposure status $E = i, i = 0,1$ and a disease outcome with dichotomous status $D = j, j = 0,1$, where 0 (1) means absence (presence) of the exposure and disease outcome. An attributable risk of a disease outcome D due to the factor E is defined by

$$AR = \frac{P(D=1) - P(D=1|E=0)}{P(D=1)} = 1 - \frac{P(D=1|E=0)}{P(D=1)}$$

where $P(D=1)$ is the overall disease rate in the entire population and $P(D=1|E=0)$ is the disease rate in the unexposed population. We wish to test independence of exposure and disease outcome using an attributable risk for a cross-sectional study.

Under a cross-sectional study, a random sample of n individuals is cross-classified by the status of exposure and the disease outcome. Let n_{ij} be the random frequency of individuals falling into a cell at exposure level i (= 0, 1) and disease status $j\,(=0,1)$ with an unknown probability π_{ij}, Of course, $0 < \pi_{ij} < 1$, $\sum_i \sum_j \pi_{ij} = 1$ and $\pi_{i.} = \pi_{i0} + \pi_{i1}$. Given the sample, it follows that $\sum_i \sum_j n_{ij} = n$, $n_{i.} = n_{i0} + n_{i1}$. The table below summarizes the distribution of subjects in a 2×2 table cross-classified by the status of exposure and disease with unknown probability of a subject falling in the cell provided in the parenthesis.

Table 1: Distribution of individuals by exposure and disease status

Exposure Status	Disease Status		Total
	$D = 0$	$D = 1$	
$E = 0$ (absent) $E = 1$ (present)	$n_{00}(\pi_{00})$ $n_{10}(\pi_{10})$	$n_{01}(\pi_{01})$ $n_{11}\ (\pi_{11})$	$n_{0.}(\pi_{0.})$ $n_{1.}(\pi_{1.})$
Total	$n_{.0}(\pi_{.0})$	$n_{.1}(\pi_{.1})$	$n(1)$

Using the unknown cell probabilities π_{ij}, we can re-write AR as

$$AR = 1 - \frac{\pi_{01}/\pi_{0.}}{\pi_{.1}} = 1 - \frac{\pi_{01}}{\pi_{0.}\pi_{.1}}$$

or equivalently, as

$$AR = \frac{\pi_{00}\pi_{11} + \pi_{01}\pi_{10}}{\pi_{00}\pi_{11} + \pi_{01}\pi_{10} - \pi_{01}}$$

Note that, under a cross-sectional study, the random vector

$$\boldsymbol{n} = (n_{00}, n_{01}, n_{10}, n_{11})$$

of cell frequencies follows a multinomial distribution with parameters n and

$$\boldsymbol{\pi} = (\pi_{00}, \pi_{01}, \pi_{10}, \pi_{11})$$

for which the log-likelihood function is given by

$$L = logK + n_{00}log(\pi_{00}) + n_{01}log(\pi_{01}) + n_{10}log(\pi_{10}) + n_{11}log(\pi_{11})$$

It follows that the maximum likelihood estimates (MLEs) of π_{ij} are given by

$$p_{ij} = \frac{n_{ij}}{n}; i = 0,1; j = 0,1$$

Then, by the invariance property of the maximum likelihood method, an estimate of AR is given by

$$ar = 1 - \frac{p_{01}}{p_{0.}p_{.1}}$$

Or,

$$ar = \frac{p_{00}p_{11} + p_{01}p_{10}}{p_{00}p_{11} + p_{01}p_{10} - p_{01}}$$

For the purpose of the test of hypothesis of independence or no association, we require distribution of ar and its asymptotic variance. Below we consider asymptotic variance of ar.

2.1 Asymptotic Variance of ar

Let $\boldsymbol{p} = (p_{00}, p_{01}, p_{10}, p_{11})$. Then, when n is large, by the Central Limit Theorem (CLT), the random vector $\sqrt{n}(\boldsymbol{p} - \boldsymbol{\pi})$ is asymptotically distributed as normal $N(\mathbf{0}, \boldsymbol{\Sigma})$, where $\mathbf{0} = (0,\ 0,\ 0,\ 0)$ is a 1×4 vector, $\boldsymbol{\Sigma} = \mathrm{diag}(\boldsymbol{\pi}) - \boldsymbol{\pi}'\boldsymbol{\pi}$ is a 4×4 covariance matrix of $\boldsymbol{p}$ and $\mathrm{diag}(\boldsymbol{\pi})$ is a 4×4 diagonal matrix with diagonal elements π_{ij}.

In order to find an asymptotic distribution of ar, let us note that ar is a function g of $\boldsymbol{p}$:

$$ar = g(\boldsymbol{p}) = \frac{p_{00}p_{11} + p_{01}p_{10}}{p_{00}p_{11} + p_{01}p_{10} - p_{01}}$$

It is easy to see that $g(\boldsymbol{p})$ is an estimator of $g(\boldsymbol{\pi}) = AR$ having a non-zero differential $\frac{\partial g(\boldsymbol{p})}{\partial p_{ij}}$ at $\boldsymbol{p} = \boldsymbol{\pi}$. Then, by the the Delta method, $\sqrt{n}(g(\boldsymbol{p}) - g(\boldsymbol{\pi}))$ is asymptotically distributed as normal with mean 0 and the variance $\boldsymbol{\Sigma}\ \partial'$, where

$$\partial = \left(\frac{\partial g(\boldsymbol{\pi})}{\partial \pi_{00}}, \frac{\partial g(\boldsymbol{\pi})}{\partial \pi_{01}}, \frac{\partial g(\boldsymbol{\pi})}{\partial \pi_{10}}, \frac{\partial g(\boldsymbol{\pi})}{\partial \pi_{11}}\right)$$

It follows that

$$\frac{\partial g(\boldsymbol{\pi})}{\partial \pi_{00}} = \frac{\pi_{01}\pi_{11}}{(\pi_{01} + \pi_{00}\pi_{11} - \pi_{01}\pi_{10})^2}$$

$$\frac{\partial g(\boldsymbol{\pi})}{\partial \pi_{01}} = \frac{-\pi_{00}\pi_{11}}{(\pi_{01} + \pi_{00}\pi_{11} - \pi_{01}\pi_{10})^2}$$

$$\frac{\partial g(\boldsymbol{\pi})}{\partial \pi_{10}} = \frac{-\pi_{01}^2}{(\pi_{01} + \pi_{00}\pi_{11} - \pi_{01}\pi_{10})^2}$$

$$\frac{\partial g(\boldsymbol{\pi})}{\partial \pi_{11}} = \frac{\pi_{00}\pi_{01}}{(\pi_{01} + \pi_{00}\pi_{11} - \pi_{01}\pi_{10})^2}$$

Thus $V(ar)$ is found to be

$$V(ar) = \frac{(1 - AR)^4\{\pi_{00}\pi_{11}(\pi_{00}\pi_{11} + \pi_{00}\pi_{01} + \pi_{01}\pi_{11}) - \pi_{01}(\pi_{00}\pi_{11} - \pi_{01}\pi_{10})^2 + \pi_{10}\pi_{01}^3\}}{n\pi_{01}^3}$$

See Walter (1976), Shapla et al. (2009) for detail derivation.

2.2 Test of Hypothesis of Independence

We wish to test the null hypothesis of independence or no association against the alternative hypothesis of their dependence or association. Symbolically, we might express hypotheses for the test as follows:

H_0: Exposure and disease outcome are independent
H_a: Exposure and disease outcome are not independent

We wish to test this hypothesis using the test statistic involving attribution risk.
Some results stated as propositions and theorems will be of great use in the process of hypothesis testing.

Note that

$$AR = \frac{\pi_{00}\pi_{11} - \pi_{01}\pi_{10}}{\pi_{00}\pi_{11} - \pi_{01}\pi_{10} + \pi_{01}} \qquad (1)$$

Or, equivalently:

$$AR = 1 - \frac{\pi_{01}}{\pi_{0.}\pi_{.1}} \qquad (2)$$

Proposition 1: $AR = 0$ if and only if $\frac{\pi_{11}}{\pi_{1.}} = \frac{\pi_{01}}{\pi_{0.}}$.
Proof: If $AR = 0$, then from (1) we have

$$\pi_{00}\pi_{11} - \pi_{01}\pi_{10} = 0$$

$$\Rightarrow \pi_{00}\pi_{11} = \pi_{01}\pi_{10}$$
$$\Rightarrow \frac{\pi_{11}}{\pi_{10}} = \frac{\pi_{01}}{\pi_{00}}$$
$$\Rightarrow \frac{\pi_{11}}{\pi_{11}+\pi_{10}} = \frac{\pi_{01}}{\pi_{01}+\pi_{00}}$$
$$\Rightarrow \frac{\pi_{11}}{\pi_{1.}} = \frac{\pi_{01}}{\pi_{0.}}$$

For the proof of the reverse direction, let $\frac{\pi_{11}}{\pi_{1.}} = \frac{\pi_{01}}{\pi_{0.}}$.

Then

$$\pi_{11}\pi_{0.} - \pi_{01}\pi_{1.} = 0$$
$$\Rightarrow \pi_{11}(\pi_{01} + \pi_{00}) - \pi_{01}(\pi_{11} + \pi_{10}) = 0$$
$$\Rightarrow \pi_{11}\pi_{01} + \pi_{11}\pi_{00} - \pi_{01}\pi_{11} - \pi_{01}\pi_{10} = 0$$
$$\Rightarrow \pi_{11}\pi_{00} - \pi_{01}\pi_{10} = 0$$

Then, $AR = \frac{\pi_{00}\pi_{11} - \pi_{01}\pi_{10}}{\pi_{00}\pi_{11} - \pi_{01}\pi_{10} + \pi_{01}} = 0$

Proposition 1 states that *the test of no association of the risk factor and the disease outcome in a* 2×2 *cross sectional study is equivalent to test that the rates of disease in exposed and unexposed groups are the same*.

Proposition 2: $AR = 0$ if and only if disease outcome and the exposure factor are independent.

Proof*:* By the definition of conditional probability,

$$P(D = 1|E = 0) = \frac{\pi_{01}}{\pi_{0.}}$$
$$P(D = 1|E = 1) = \frac{\pi_{11}}{\pi_{1.}}$$

By the definition of independence,

$$P(D = 1|E = 0) = P(D = 1) = \pi_{.1}$$
$$P(D = 1|E = 1) = P(D = 1) = \pi_{.1}$$

Thus

$$\frac{\pi_{01}}{\pi_{0.}} = \frac{\pi_{11}}{\pi_{1.}}$$

Then, by proposition 1, $AR = 0$.

Again, if $AR = 0$ then by proposition 1, $\frac{\pi_{11}}{\pi_{1.}} = \frac{\pi_{01}}{\pi_{0.}}$, which is a condition for independence as shown above.

Proposition 2 states that *the test of no association of the risk factor and the disease outcome in a* 2×2 *cross sectional study is equivalent to the test of independence of factor and disease outcome.*

Proposition 3: $AR > 0$ if and only if $\frac{\pi_{11}}{\pi_{1.}} > \frac{\pi_{01}}{\pi_{0.}}$.

Proof: If $AR > 0$, then from (1) it follows that

$$\begin{aligned}
&\pi_{00}\pi_{11} - \pi_{01}\pi_{10} > 0 \\
&\Rightarrow\ \pi_{00}\pi_{11} > \pi_{01}\pi_{10} \\
&\Rightarrow\ \frac{\pi_{11}}{\pi_{01}} > \frac{\pi_{10}}{\pi_{00}} \\
&\Rightarrow\ \frac{\pi_{11}}{\pi_{01}} > \frac{\pi_{11}+\pi_{10}}{\pi_{01}+\pi_{00}} \\
&\Rightarrow\ \frac{\pi_{11}}{\pi_{01}} > \frac{\pi_{1.}}{\pi_{0.}} \\
&\Rightarrow\ \frac{\pi_{11}}{\pi_{1.}} > \frac{\pi_{01}}{\pi_{0.}}
\end{aligned}$$

For the proof of the reverse direction, let $\frac{\pi_{11}}{\pi_{01}} > \frac{\pi_{1.}}{\pi_{0.}}$. Then,

$$\begin{aligned}
&\pi_{11}\pi_{0.} - \pi_{01}\pi_{1.} > 0 \\
&\Rightarrow\ \pi_{11}(\pi_{01} + \pi_{00}) - \pi_{01}(\pi_{11} + \pi_{10}) > 0 \\
&\Rightarrow\ \pi_{11}\pi_{01} + \pi_{11}\pi_{00} - \pi_{01}\pi_{11} - \pi_{01}\pi_{10} > 0 \\
&\Rightarrow\ \pi_{11}\pi_{00} - \pi_{01}\pi_{10} > 0
\end{aligned}$$

Then, $AR = \frac{\pi_{00}\pi_{11} - \pi_{01}\pi_{10}}{\pi_{00}\pi_{11} - \pi_{01}\pi_{10} + \pi_{01}} > 0$

Proposition 3 states that *positive association of the risk factor and the disease outcome in a* 2×2 *cross sectional study is equivalent to the fact that the rate of disease in exposed group is greater than the rate of disease in the unexposed group.*

Proposition 4: $\frac{\pi_{11}}{\pi_{1.}} \geq \frac{\pi_{01}}{\pi_{0.}}$ *if and only if* $\pi_{11} \geq \pi_{1.}\pi_{.1}$.

Proof: Let $\frac{\pi_{11}}{\pi_{1.}} \geq \frac{\pi_{01}}{\pi_{0.}}$. Then, $\frac{\pi_{11}}{\pi_{1.}} \geq \frac{\pi_{11}+\pi_{01}}{\pi_{1.}+\pi_{0.}} = \frac{\pi_{.1}}{1}$, which implies that $\pi_{11} \geq \pi_{1.}\pi_{.1}$.
For the proof of the reverse direction, let $\pi_{11} \geq \pi_{1.}\pi_{.1}$. Then,

$$\frac{\pi_{11}}{\pi_{1.}} \geq \pi_{.1}$$

Also, $\pi_{01} = \pi_{.1} - \pi_{11} \leq \pi_{.1} - \pi_{1.}\pi_{.1} = \pi_{.1}(1 - \pi_{1.}) = \pi_{.1}\pi_{0.}$
Then,

$$\frac{\pi_{01}}{\pi_{0.}} \leq \pi_{.1}$$

which implies that $\frac{\pi_{01}}{\pi_{0.}} \leq \pi_{.1} \leq \frac{\pi_{11}}{\pi_{1.}}$. Then, $\frac{\pi_{11}}{\pi_{1.}} \geq \frac{\pi_{01}}{\pi_{0.}}$.

The facts in propositions 1-4 can be combined together for testing independence or no association regarding the risk factor and the disease.

Therefore, to test H_0: Exposure and disease outcome are independent *versus* H_a: Exposure and disease outcome are not independent, we simply can test

$$H_0: AR = 0\ versus\ H_a: AR \neq 0$$

Given the fact that ar is distributed asymptotically as $N(AR, V(ar))$, we can implement the test using a Z test statistic given by

$$Z = \frac{ar-AR}{\sqrt{V(ar)}} \sim Z(0,1) \text{ asymptotically under } H_0.$$

2.3 Example

This example appears in Fleiss (2003). A total of 2784 subjects in a community has been cross-classified by the presence or absence of the respiratory disease and the locomotor disease. Locomotor disease is the disease of bones and organs of movement. The following table summarizes the distribution of the subjects with respect to respiratory disease and locomotor disease status.

Table 2: Cross-classification of 2784 subjects by the status of the respiratory disease and locomotor disease

Respiratory disease	Locomotor disease $D = 0$	 $D = 1$	Total	Proportion with Locomotor disease
$E = 0$	2376	184	2560	0.07
$E = 1$	207	17	224	0.08
Total	2583	201	2784	0.15

Since the rates of locomotor disease in people with and without respiratory disease (0.08 and 0.07, respectively) are virtually the same, we would like to test whether there is an association between respiratory disease and locomotor disease. Therefore, we are interested to test H_0 : independence between the risk factor and the disease outcome, which is equivalent to test the hypothesis $H_0: AR = 0 \; versus \; H_a: AR \neq 0$. Under null hypothesis, the observed value of the test statistic,

$$Z = \frac{ar - AR}{\sqrt{V(ar)}}$$

is found to be 0.218. Therefore, at 5% significance level the data does not provide sufficient evidence to conclude that there is an effect of respiratory disease in developing locomotor disease. Thus, the two characteristics, respiratory disease and the locomotor disease are independent of each other.

In a similar way, the above test can be used to test for any hypothesized value of AR.

3. Power Analysis for Variation of AR in the set of 2×2 tables

Let us consider the set of all 2×2 tables given by

$$S = \{(\pi_{11}, \pi_{01}, \pi_{10}) : \pi_{11}, \pi_{01}, \pi_{10} > 0; \pi_{11} + \pi_{01} + \pi_{10} < 1\}$$

Note that S can be written as the union of two subsets S^+ and S^-, where

$$S^+ = \left\{(\pi_{11}, \pi_{01}, \pi_{10}) : \pi_{11}, \pi_{01}, \pi_{10} > 0; \pi_{11} + \pi_{01} + \pi_{10} < 1; \frac{\pi_{11}}{\pi_{1.}} \geq \frac{\pi_{01}}{\pi_{0.}}\right\}$$

is the set of all 2×2 tables of positive association and

$$S^- = \left\{(\pi_{11}, \pi_{01}, \pi_{10}) : \pi_{11}, \pi_{01}, \pi_{10} > 0; \pi_{11} + \pi_{01} + \pi_{10} < 1; \frac{\pi_{11}}{\pi_{1.}} < \frac{\pi_{01}}{\pi_{0.}}\right\}$$

is the set of all 2×2 tables of negative association. In other words, on S^+, $AR \geq 0$ and on S^-, $AR < 0$.

Fixing the row sum $\pi_{1.}$ (and also $\pi_{0.}$, thereby), let

$$S_{\pi_{1.}} = \{(\pi_{11}, \pi_{01}, \pi_{10}) : \pi_{11}, \pi_{01}, \pi_{10} > 0; \pi_{11} + \pi_{01} + \pi_{10} < 1;\ \pi_{1.} \text{is fixed}\}$$

A 2×2 table in $S_{\pi_{1.}}$ is completely determined when π_{11} and π_{01} are determined. The set $S_{\pi_{1.}}$ can be written as the union of two subsets $S^{+}_{\pi_{1.}}$ and $S^{-}_{\pi_{1.}}$ where

$$S^{+}_{\pi_{1.}} = \left\{(\pi_{11}, \pi_{01}, \pi_{10}) : \pi_{11}, \pi_{01}, \pi_{10} > 0; \pi_{11} + \pi_{01} + \pi_{10} < 1, \pi_{1.} \text{is fixed}, \frac{\pi_{11}}{\pi_{1.}} \geq \frac{\pi_{01}}{\pi_{0.}}\right\}$$

and

$$S^{-}_{\pi_{1.}} = \left\{(\pi_{11}, \pi_{01}, \pi_{10}) : \pi_{11}, \pi_{01}, \pi_{10} > 0; \pi_{11} + \pi_{01} + \pi_{10} < 1, \pi_{1.} \text{is fixed}, \frac{\pi_{11}}{\pi_{1.}} < \frac{\pi_{01}}{\pi_{0.}}\right\}.$$

Lemma 1: In the set $S_{\pi_{1.}}$, $\pi_{1.}$ can be fixed in the range $\pi_{11} < \pi_{1.} < 1 - \pi_{01}$.

Proof: Note that

$$\pi_{11} < \pi_{1.}$$

Also,

$$\begin{aligned} &\pi_{11} + \pi_{01} + \pi_{10} < 1 \\ &\Rightarrow \pi_{1.} + \pi_{01} < 1 \\ &\Rightarrow \pi_{1.} < 1 - \pi_{01} \end{aligned}$$

Thus,

$$\pi_{11} < \pi_{1.} < 1 - \pi_{01}$$

3.1 Useful Results for Variation of AR in the Sets of 2×2 Table

In this section, we state some results as theorems for the variations of AR in the set $S_{\pi_{1.}}$

Theorem 1

In the set $S_{\pi_{1.}}$, if we fix π_{01}, then $AR = 1 - \frac{\pi_{01}}{\pi_{0.}\pi_{.1}}$ is increasing in π_{11}.

Proof: In the set $S_{\pi_{1.}}$, if we fix π_{01}, then we can write $AR = 1 - \frac{\pi_{01}}{\pi_{0.}\pi_{.1}} = 1 - \frac{\pi_{01}}{\pi_{0.}(\pi_{01}+\pi_{11})}$, which is only a function of π_{11}, since $\pi_{0.}$ is fixed in $S_{\pi_{1.}}$.

Taking the derivative of AR with respect to π_{11}, we get

$$\frac{\partial AR}{\partial \pi_{11}} = -\frac{\pi_{01}}{\pi_{0.}}\left[-\frac{1}{(\pi_{01}+\pi_{11})^2}\right] = \frac{\pi_{01}}{\pi_{0.}(\pi_{01}+\pi_{11})^2} > 0$$

Thus, AR is increasing in π_{11} in the set $S_{\pi_{1.}}$.

Theorem 2

For any given point (π_{01}, π_{11}) in S, AR is decreasing in $\pi_{1.}$ in the range $\pi_{11} < \pi_{1.} < 1 - \pi_{01}$.

Proof: For a fixed point (π_{01}, π_{11}) of $S_{\pi_{1.}}$, we can write $AR = 1 - \frac{\pi_{01}}{\pi_{0.}(\pi_{01}+\pi_{11})} = 1 - \frac{\pi_{01}}{(1-\pi_{1.})(\pi_{01}+\pi_{11})}$, which is only a function of $\pi_{1.}$.

Taking the derivative of AR with respect to $\pi_{1.}$, we get

$$\frac{\partial AR}{\partial \pi_{1.}} = -\frac{\pi_{01}}{(\pi_{01}+\pi_{11})}\left[-\frac{1}{(1-\pi_{1.})^2}(-1)\right] = -\frac{\pi_{01}}{(\pi_{01}+\pi_{11})(1-\pi_{1.})^2} < 0\ .$$

Therefore, AR with a fixed point (π_{01}, π_{11}) is decreasing for $\pi_{11} < \pi_{1.} < 1 - \pi_{01}$.

In view of these two theorems, we expect that the test of positive association has a larger power at a point with larger value of AR than at a point with a smaller value of AR.

3.2 Simulation of Power of the Test using ar

In this section, a Monte Carlo study has been carried out to assess the power of the test

$$H_0: AR = 0 \; versus \; H_a: AR \neq 0$$

based on ar by fixing the row sum marginals, the entry π_{01}, and gradually increasing the value of π_{11} for varying large sample cases.

We consider three different combinations for the row sum marginals described below.

Case 1: The proportion of exposed group is $\pi_{1.} = 10\%$, unexposed group is $\pi_{0.} = 90\%$. This kind of situation may arise in real life when a small proportion of people are exposed to a particular risk factor; for example, exposed to a chemical, coal dust, etc. in the community.

Case 2: The proportion of exposed group is $\pi_{1.} = 90\%$, unexposed group is $\pi_{0.} = 10\%$. For example, in most of the third world countries, a big portion of the population is usually exposed to contaminated water, polluted air, unhealthy environment, etc.

Case 3: The proportions of exposed and unexposed groups are equal, that is, $\pi_{0.} = \pi_{1.} = 50\%$. This situation may arise in real life when there is a possibility of the outbreak of a particular disease, for example, flu, and the community has been urged to take the preventive medication.

The Monte Carlo simulation has been performed following the scheme given below.
1. Fix significance level α, and the Monte Carlo sample size M.
2. For each of the cases 1-3, form a 2×2 table satisfying$\frac{\pi_{11}}{\pi_{1.}} > \frac{\pi_{01}}{\pi_{0.}}$.
3. Generate a random sample from the given multinomial distribution for different values of n and find the value of the test statistic $Z = \frac{ar - AR}{\sqrt{V(ar)}}$ under the alternative hypothesis.
4. Compare the observed value of the test statistic with the critical value z_α and reject the null hypothesis if the observed value is greater than or equal to the critical value.
5. Repeat steps 3-4 M times and count the number of rejections. The proportion of rejection over all the simulations gives the estimated power.
6. Keeping the same row sum marginals and fixing the entry π_{01}, we consider a new 2×2 table such that $\acute{\pi}_{11} > \pi_{11}$ and $\frac{\acute{\pi}_{11}}{\pi_{1.}} \geq \frac{\pi_{01}}{\pi_{0.}}$. We repeat steps 3-5 M times to find the power with the new configuration given by $(\pi_{00}, \pi_{01}, \acute{\pi}_{10}, \acute{\pi}_{11})$.

The simulation results for M=10000 have been summarized in Tables 3, 4, and 5 for cases 1, 2 and 3, respectively for different values of n. From the simulation results, it is evident that the power of the test statistic is increasing with the increase in the value of AR for all three cases considered in the study, which is consistent with the theoretical development above.

Table 3: Estimated power using ar for the case $\pi_{1.}$= 0.1, $\pi_{0.}$= 0.9

Matrix of probabilities		Estimated power						
	ar	$n = 30$	50	70	90	120	150	200
0.540 0.360 0.055 0.045	0.0123	0.0960	0.0999	0.1086	0.1164	0.1171	0.1208	0.1338
0.540 0.360 0.050 0.050	0.0244	0.1170	0.1349	0.1429	0.1525	0.1642	0.1857	0.2096
0.540 0.360 0.045 0.055	0.0361	0.1479	0.1538	0.1872	0.2177	0.2369	0.2682	0.3289
0.540 0.360 0.040 0.060	0.0476	0.1554	0.1937	0.2433	0.2815	0.3404	0.3996	0.4761
0.540 0.360 0.035 0.065	0.0588	0.1970	0.2537	0.3078	0.3656	0.4463	0.5223	0.6276
0.540 0.360 0.030 0.070	0.0698	0.2323	0.3042	0.3800	0.4582	0.5585	0.6404	0.7648
0.540 0.360 0.025 0.075	0.0805	0.2670	0.3614	0.4648	0.5570	0.6648	0.7612	0.8655
0.540 0.360 0.020 0.080	0.0909	0.2964	0.4207	0.5425	0.6420	0.7652	0.8512	0.9254
0.540 0.360 0.015 0.085	0.1011	0.3511	0.4917	0.6335	0.7363	0.8500	0.9147	0.9718
0.540 0.360 0.010 0.090	0.1111	0.3931	0.5539	0.7049	0.8122	0.9051	0.9572	0.9901

Table 4: Estimated power using ar for the case $\pi_{1.}$= 0.9, $\pi_{0.}$= 0.1

Matrix of probabilities		Estimated power						
	ar	$n = 30$	50	70	90	120	150	200
0.060 0.040 0.495 0.405	0.1011	0.0257	0.0639	0.1286	0.1373	0.1377	0.1394	0.1386
0.060 0.040 0.450 0.450	0.1837	0.0127	0.0993	0.1637	0.1849	0.1917	0.2044	0.2423
0.060 0.040 0.405 0.495	0.2523	0.0115	0.1596	0.2151	0.2394	0.2766	0.3041	0.3705
0.060 0.040 0.360 0.540	0.3103	0.0496	0.2395	0.2800	0.3413	0.3936	0.4573	0.5449
0.060 0.040 0.315 0.585	0.3600	0.0990	0.2981	0.3721	0.4288	0.5218	0.5933	0.7006
0.060 0.040 0.270 0.630	0.4030	0.1828	0.3712	0.4621	0.5295	0.6441	0.7153	0.8299
0.060 0.040 0.225 0.675	0.4406	0.2570	0.4511	0.5469	0.6446	0.7460	0.8251	0.9108
0.060 0.040 0.180 0.720	0.4737	0.3232	0.5317	0.6499	0.7396	0.8383	0.9011	0.9577
0.060 0.040 0.135 0.765	0.5031	0.4385	0.6033	0.7425	0.8166	0.9053	0.9513	0.9842
0.060 0.040 0.090 0.810	0.5294	0.4826	0.6929	0.8031	0.8820	0.9487	0.9763	0.9932

Table 5: Estimated power using ar for the case $\pi_{1.} = 0.5$, $\pi_{0.} = 0.5$

Matrix of probabilities		Estimated power						
	ar	$n = 30$	50	70	90	120	150	200
0.300 0.200 0.275 0.225	0.0588	0.1407	0.1392	0.1389	0.1439	0.1583	0.1622	0.1887
0.300 0.200 0.250 0.250	0.1111	0.1793	0.1935	0.2256	0.2514	0.2989	0.3331	0.4075
0.300 0.200 0.225 0.275	0.1579	0.2350	0.2805	0.3469	0.4065	0.4969	0.5718	0.6776
0.300 0.200 0.200 0.300	0.2000	0.2957	0.4030	0.5008	0.5925	0.6986	0.7853	0.8791
0.300 0.200 0.175 0.325	0.2381	0.3738	0.5293	0.6554	0.7602	0.8597	0.9145	0.9699
0.300 0.200 0.150 0.350	0.2727	0.4786	0.6704	0.7854	0.8809	0.9449	0.9779	0.9945
0.300 0.200 0.125 0.375	0.3043	0.5842	0.7694	0.8914	0.9453	0.9839	0.9953	0.9994
0.300 0.200 0.100 0.400	0.3333	0.6711	0.8633	0.9527	0.9837	0.9970	0.9998	1.0000
0.300 0.200 0.075 0.425	0.3600	0.7855	0.9336	0.9857	0.9967	0.9996	1.0000	1.0000
0.300 0.200 0.050 0.450	0.3846	0.8458	0.9727	0.9968	0.9995	1.0000	1.0000	1.0000

3.3 Estimating Size of the Test

In this section, we will study the size of the test statistics to test H_0 : independence versus H_a : positive association based on ar**.** In order to do so, we consider two matrices formed by the probabilities of two given multinomial distributions for which both the values of AR are equal to zero. For each of the matrices, the simulation study has been performed in the following way.

1. Generate a random sample from the given multinomial distribution and find the value of the test statistic given by $Z = \frac{ar-AR}{\sqrt{V(ar)}}$.
2. Compare the value of the statistic with the critical value z_α and reject the null hypothesis if $Z \geq z_\alpha$.
3. Repeat steps 1-2 for M times. The proportion of rejection over all the simulations gives the estimated size for each test statistic.

Table 6: Estimated level ($\alpha = 0.05$) for test Z

Probability Matrix	n	$Z = \frac{ar - AR}{\sqrt{V(ar)}}$
0.540 0.360 0.060 0.040	30	0.0483
	50	0.0495
	70	0.0546
	90	0.0534
	120	0.0530
	150	0.0526
	200	0.0553
0.060 0.040 0.540 0.360	30	0.0312
	50	0.0429
	70	0.0418
	90	0.0560
	120	0.0656
	150	0.0595
	200	0.0554

4. Result Discussions

In order to study the power of the test $H_0{:}\, AR = 0\ versus\ H_a{:}\, AR \neq 0$, we perform a Monte Carlo simulation in the set $S_{\pi_{1.}}$ while fixing the entry π_{01} and gradually increasing π_{11}. We consider three different combinations of the rates of exposure and unexposure, namely, Case 1: ($\pi_{1.}$= 0.1, $\pi_{0.}$= 0.9), Case 2: ($\pi_{1.}$= 0.9, $\pi_{0.}$= 0.1), and Case 3: ($\pi_{1.}$= 0.5, $\pi_{0.}$= 0.5). The values of n to be considered in the simulation are 30, 50, 70, 90, 120, 150, and 200. The estimated powers are displayed in Tables 3-5 for a Monte Carlo simulation of size M=10000. From the results, it is evident that the power of the test increases as AR increases for a given sample size n. It also reveals that for a given 2×2 table, the power increases as the sample size increases. As values of AR increases, the power gets closer to one for relatively larger n.

In order to study the size of the test $H_0{:}\, AR = 0\ versus\ H_a{:}\, AR \neq 0$, we form two 2×2 tables from two given multinomial distributions with AR values equal to zero. The estimated sizes from different values of n are presented in Table 6 in section 3.3. It appears that the estimated sizes are close to the nominal level of 0.05 as sample size increases for both contingency tables considered.

The results presented in Tables 3-6 reveal that the test of no association based on AR performs well in estimating the power and size of the test.

5. Conclusions

The AR plays an important role in public health and epidemiology to locate the important risk factors of a disease outcome. While a substantial amount of research has been done in developing the point and asymptotic variance estimations of AR in case-control, cohort and cross-sectional study designs, the use of AR for the test of hypothesis in reference to the association between disease and exposure factors lacks in literature.

In this paper, we wish to employ AR to test the hypotheses H_0: Exposure and disease outcome are independent versus H_a: Exposure and disease outcome are not independent. It appears in literature that the AR for a 2×2 cross sectional study is very popular in biomedical and health science research due to the simplicity of its application and interpretation. Therefore, we restrict our attention for a 2×2 cross sectional study. However, the process can be generalized to higher dimensional table as well.

As a part of the study, we presented four propositions in section 2.2 along with the proof. Proposition 1 states that no association of the risk factor and the disease outcome in a 2×2 cross sectional study is equivalent to the fact that the rates of disease in exposed and unexposed groups are the same. Proposition 2 states that the test of no association of the risk factor and the disease outcome in a 2×2 cross sectional study is equivalent to the test of independence of factor and disease outcome. Proposition 3 states that positive association of the risk factor and the disease outcome in a 2×2 cross sectional study is equivalent to the fact that the rate of disease in exposed group is greater than the rate in the unexposed group. Proposition 4 states that the rate of disease in the exposed group is greater than or equal to the rate of disease in the unexposed group if and only if the rate of diseased, exposed group is greater than or equal to the product of the overall exposure rate and the overall disease rate. The application of results of four propositions leads us to the test of hypothesis of independence or no association between the risk factor and disease outcome equivalently by testing

$$H_0: AR = 0 \; versus \; H_a: AR \neq 0.$$

We consider an example in section 2.3 to study if there is an association between the respiratory disease and locomotor disease using AR. A total of 2784 subjects was cross-classified according to the disease and exposure status. On the basis of the asymptotic distribution of AR, we form the test statistic that follows approximately a standard normal distribution. Based on the data, the observed value of the test statistic is found to be $z = 0.218$. Comparing the observed value with a two-tailed critical value under a standard normal curve, we fail to reject the null hypothesis at 5% level of significance and hence conclude that the data does not provide sufficient evidence to indicate that there is an effect of respiratory disease in developing locomotor disease.

In section 3, we develop some useful results to analyze the power of the test $H_0: AR = 0 \; versus \; H_a: AR \neq 0$ for varying values of AR in the sets of 2×2 table. To this end, we denote the sets of all 2×2 tables by S and define the set $S_{\pi_{1.}}$by fixing the row sum $\pi_{1.}$ and hence $\pi_{0.}$ The result in Theorem 1 states that AR is increasing in π_{11} in the set $S_{\pi_{1.}}$while fixing the value of π_{01}. Theorem 2 claims that for any given point (π_{01}, π_{11}) in S, AR is decreasing in $\pi_{1.}$ in the range $\pi_{11} < \pi_{1.} < 1 - \pi_{01}$. Therefore, it is expected that the test of no association versus positive association has a larger power at a 2×2 table with a larger value of AR than at a 2×2 table with a smaller value of AR. In order to verify this, we carry out a Monte Carlo simulation in section 3.2 in the set $S_{\pi_{1.}}$ while fixing the entry π_{01} and gradually increasing π_{11} for varying values of exposure rate $\pi_{1.}$, and the sample size, n considered to be large. The estimated powers are displayed in Tables 3-5 for a Monte Carlo simulation of size M=10000. From the results, it is evident that the power of the test increases as AR increases for a given sample size n. It also reveals that for a given 2×2 table, the power increases as sample size increases. As values of AR increases, the power gets closer to one for relatively larger n considered in the simulation.

In order to study the size of the test $H_0: AR = 0\ versus\ H_a: AR \neq 0$, we form two 2×2 tables for which AR values are equal to zero. The estimated sizes from different values of n are presented in Table 6 in section 3.3. From the results of the study, it reveals that the estimated sizes are close to the nominal level of 0.05 as sample size increases for both contingency tables considered here. Overall, the performance of the study implies that the test of no association based on AR is satisfactory in terms of estimating power and size of the test, and hence the test of no association or independence of a factor with disease outcome should be undertaken with confidence using test statistic that would involve AR.

References

M. L. Levin, *The occurrence of lung cancer in man*, Acta Unio Internationalis contra Cancrum, 9 (1953), pp. 531-541.

J. Benichou, *Methods of adjustment for estimating the attributable risk in case-control studies,* Statistics in Medicine, 10 (1991), pp. 1753-1773.

S. S. Coughlin, J. Benichou, and D. L. Weed, *Attributable risk estimation in case-control studies*, Epidemiologic Reviews, 16 (1994), pp. 51-64.

S. D. Walter, *The distribution of Levin's measure of attributable risk*, Biometrika, 62 (1975), pp. 371-375.

S. D. Walter, *The estimation and interpretation of attributable risk in health research*, Biometrics, 32 (1976), pp. 829-849.

J. L. Fleiss, *Inference about population attributable risk from cross-sectional studies*, American Journal of Epidemiology, 110 (1979), pp. 103-104.

O. S. Miettinen, *Proportion of disease caused or prevented by a given exposure, trait or intervention*, Am. J. Epidemiol., 99 (1974), pp. 325-332.

B. L. Ouellet, J. M. Romeder, and J. M. Lance, *Premature mortality attributable to smoking hazardous drinking in Canada,* American Journal of Epidemiology, 109 (1979), pp. 451-463.

P. Cole and B. MacMahon, *Attributable risk percent in case-control studies*, Br. J. Prev. Soc. Med., 25 (1971), pp. 242-244.

S. Basu and J. R. Landis, *Model-based estimation of population attributable risk under cross-sectional sampling*, American Journal of Epidemiology, 142 (1995) pp. 1338-1343.

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

P. Bruzzi et al., Estimating the population attributable risk for multiple risk factors using case-control data, *American Journal of Epidemiology,* 122 (1985) pp. 904-914.

J. Benichou, and M. H. Gail, *Variance calculations and confidence intervals for estimates of the attributable risk based on logistic models*, Biometrics, 46 (1990), pp. 991-1003.

J. Benichou, *A review of adjusted estimators of attributable risk*, Statistical Methods in Medical Research, 10 (2001), pp. 195-216.

K. Drescher and W. Schill, *Attributable risk estimation from case-control data via logistic regression*, Biometrics, 47 (1991), pp. 1247-1256.

S. Greenland and K. Drescher, *Maximum likelihood estimation of the attributable fraction from logistic models*, Biometrics, 49 (1993), pp. 865-872.

G. E. Eide and O. Gefeller, *Sequential and average attributable fractions as aids in the selection of preventive strategies*, Journal of Clinical Epidemiology, 48 (1995), pp. 645-655.

B. I. Graubard and T. R. Fears, *Standard errors for attributable risk for simple and complex*

sample designs, Biometrics, 61 (2005), pp. 847-855.

C. Cox, *Model-based estimation of attributable risk in case-control studies*, Statistical Methods in Medical Research, 15 (2006), pp. 611-625.

T. J. Shapla, T. T. Nguyen, and J. T. Chen, *Multilevel attributable risk in cross-sectional studies*, Journal of Statistical Computation and Simulation, 79 (2009), pp. 39-54.

K. Islam and T. J. Shapla, *Inference using attributable risk for a 2×2 case–control study*, Journal of Statistical Computation and Simulation, 83 (2013), pp. 355-369.

J. Fleiss, B. Levin, and M. C. Paik, *Statistical Methods for Rates and Proportions*, Wiley-Interscience, New Jersey, 2003.