Inference Using Impact Numbers for a Multinomial Sampling Design

Tanweer Shapla[1] and Khairul Islam[2]
[1,2]Department of Mathematics, Eastern Michigan University, Ypsilanti, MI 48197 USA

Abstract

Impact numbers reflect the number of people specific to a population among whom one outcome or case is attributable to the exposure of the risk factor. When being exposed to the risk factor is significant for the development of the disease as measured by the standard effect measures such as relative risk, risk difference or attributable risk as appropriate, the impact numbers provide very useful information not possible otherwise. To date, a few studies exist in literature that takes impact numbers into account in making inference. In particular, while confidence interval estimates of impact numbers are investigated for cohort and case-control studies, they are not yet documented adequately for a multinomial sampling design. This paper provides confidence interval estimates of impact numbers for a multinomial sampling design. Real life example and simulation studies are considered to justify performance of these methods.

Key Words: Multinomial Sampling, Impact numbers, Interval Estimate, Principle of Invert and Exchange, Delta Method

1. Introduction

Assessing the risk of a factor to the development of a disease outcome is of great importance to the epidemiological research. Relative risk (RR), odds ratio (OR) or absolute risk increase (ARI) are widely used measures in existing literature for assessing the risk of a factor. Among these measurers, the OR is very popular to epidemiologists because it can be estimated easily for three common types of designs, namely, cohort, case-control and cross-sectional studies. Often, RR, OR and ARI get criticized due to the fact that none of these measures takes into account the actual prevalence of exposure to the population of interest. A simple example that explains the effect of not taking the actual prevalence of exposure into account in measuring risk appears in [1, 2, 3]. It is noted that exposure of industrial workers to various chemicals often entails a high relative risk of carcinoma of the lung, with rates sometimes as much as 40 to 50 times the rate of similar workers not so exposed; smoking, with much lower relative risks, is responsible for many more cases of the disease, simply because the fraction of population exposed to smoking is much larger than that of chemicals. For this kind of reason, when examining diseases with several risk factors varying both in their relative risks and prevalences, it seems inadequate to compare the epidemiological importance of these factors using relative risk alone [3]. Introduced by Levin [2], the attributable risk (AR) on the other hand, takes into account both the prevalence of the risk factor and the strength of association between exposure and disease.

Correspondence: tshapla@emich.edu

Therefore, it is probably the most commonly used epidemiological measure for locating important risk factors of disease in health research and disease prevention program [3-7]. Among several definitions of attributable risk available, the population attributable fraction (PAF), attributable fraction in exposed (AF_e) and population attributable risk (PAR) are widely used in literature [2, 8, 9].

Heller et al. [8] addressed the necessity of measuring impact of a risk factor on a population with some examples: What is the impact of hypertension on the incidence of coronary heart disease (CHD) in the entire community? How many deaths from CHD among smokers are directly attributable to smoking? To answer these sorts of questions they introduced four impact numbers, namely, population impact number (PIN), exposure impact number (EIN), case impact number (CIN) and exposed cases impact number (ECIN) for cohort and case-control studies. The PIN is the reciprocal of the PAR and describes the average number of those in the population among whom one case is attributable to the exposure of the risk factor. The CIN is the reciprocal of the PAF and describes the average number of people with outcome among whom one case is attributable to the exposure of the risk factor. The EIN is the reciprocal of the absolute risk increase (ARI) and describes the average number of exposed person among whom one case is attributable to the exposure of the risk factor. Finally, the ECIN is the reciprocal of the AF_e and describes the average number of exposed cases among whom one case is attributable to the exposure of the risk factor. While Heller et al. [8] considered point estimates and interpretational aspects of impact numbers for cohort and case-control studies, Hildebrandt et al. [10] computed confidence interval estimates for EIN, CIN and ECIN for a cohort study design using the principles of inverting and exchanging the confidence limits of ARI, PAF and AF_e respectively.

To date, however, inference for impact numbers using the confidence interval estimate has not yet been documented adequately for a multinomial sampling scheme. This paper considers constructing confidence interval estimates for impact numbers for a multinomial sampling scheme using delta method and the principle of inverting and exchanging the confidence limits of the standard effect measure they relate to. A real-life example is considered to investigate the impact of drinking on stomach ulcer using various impact numbers and compare performances of confidence interval estimates of these impact numbers using the delta method and the principle of inverting and exchanging limits. A Monte Carlo simulation study is also carried out to compare performance of confidence interval estimates of impact numbers in terms of coverage probability and average length of interval estimates.

2. Set up of a Multinomial Sampling Scheme

Due to the simplicity of the presentation, a multinomial sampling design in the form of a 2x2 contingency table is very popular to epidemiologists. This paper, therefore, considers impact numbers for a multinomial sampling scheme having a form of 2x2 contingency table. Let the two levels of a risk factor E be designated by i=0, 1 and the levels of the disease outcome D by j=0, 1 with 0 (1) meaning the absence (presence) of the exposure and disease outcome. Given a random sample of n individuals, let N_{ij} be the random frequency of n individuals falling into cell at exposure level i (= 0, 1) with disease status $j\,(=0,1)$. Let π_{ij} >0 be the probability of a subject falling into a cell having frequency

$N_{ij} = n_{ij}$. Note that, $\sum_i \sum_j n_{ij} = n$ and $n_{i.} = n_{i0} + n_{i1}$. Also, $\sum_i \sum_j \pi_{ij} = 1$ and $\pi_{i.} = \pi_{i0} + \pi_{i1}$. The table below is a data structure of a multinomial sampling scheme represented by a 2x2 contingency table for a risk factor with dichotomous exposure and a dichotomous disease outcome.

Table 1: Distribution of individuals by the presence or absence of risk factor and disease outcomes

		Disease Status, D		
		Present (1)	Absent (0)	Total
Risk Factor, E	Present (1)	n_{11} (π_{11})	n_{10} (π_{10})	$n_{1.}(\pi_{1.})$
	Absent (0)	n_{01} (π_{01})	n_{00} (π_{00})	$n_{0.}$ $(\pi_{0.})$
Total		$n_{.1}$ $(\pi_{.1})$	$n_{.0}(\pi_{.0})$	$n(1)$

Given n, the random vector $\mathbf{N} = (N_{00} = n_{00}, N_{01} = n_{01}, N_{10} = n_{10}, N_{11} = n_{11})$ 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 is given by

$L = \log K + n_{00} \log \pi_{00} + n_{01} \log \pi_{01} + n_{10} \log \pi_{10} + n_{11} \log \pi_{11}$.

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

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

Let $\mathbf{p} = (p_{11}, p_{10}, p_{01}, p_{00})$. Then, when n is large, by the Central Limit Theorem (CLT), the random vector $\sqrt{n}\ (\mathbf{p} - \boldsymbol{\pi})$ 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 $\mathbf{p}$ and $\mathrm{diag}(\boldsymbol{\pi})$ is a 4×4 diagonal matrix with diagonal elements π_{ij}.

Let $g_k(\mathbf{p})$ be an estimator of $g_k(\boldsymbol{\pi})$ having a non-zero differential $\frac{\partial g_k(\mathbf{p})}{\partial p_{ij}}$ at $\mathbf{p} = \boldsymbol{\pi}$.

Then, by the use of the delta method [11], $\sqrt{n}(g_k(\mathbf{p}) - g_k(\boldsymbol{\pi}))$ is asymptotically distributed as normal with mean 0 and the variance $\partial_k \boldsymbol{\Sigma} \partial_k'$, where

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

3. Standard Effect Measures for Assessing Risk of a Factor

Some standard effect measures such as relative risk, absolute risk increase and various measures of attributable risk are considered in this section for the completeness of the study as they are widely used in epidemiologic study, and are related to the definitions of impact numbers appeared in Heller et al. [8].

3.1 Relative Risk

The relative risk is the ratio of the risk of disease in the exposed group to that of in the unexposed group given by $\mathrm{RR} = \dfrac{P(D=1 \mid E=1)}{P(D=1 \mid E=0)}$, where $P(\cdot \mid \cdot)$ is the risk of disease measured by the conditional probability. In terms of parameters π_{ij}, it follows that $\mathrm{RR} = \dfrac{\pi_{11}\pi_{0.}}{\pi_{01}\pi_{1.}}$. Let $g_1(\boldsymbol{\pi}) = \dfrac{\pi_{11}\pi_{0.}}{\pi_{01}\pi_{1.}}$. Then, an MLE of RR is given by $g_1(\mathbf{p}) = \dfrac{p_{11}p_{0.}}{p_{01}p_{1.}}$. By the principle of Delta method, an estimator of the variance of $g_1(\mathbf{p})$ is given by $v(g_1(\mathbf{p})) = \left\{ \dfrac{p_{11}p_{0.}}{p_{01}^{\,3}p_{1.}^{\,3}} \left(p_{10}p_{01}p_{0.} + p_{11}p_{00}p_{1.} \right) \right\} / n$. (See Appendix for more detail). A $100(1-\alpha)$ percent confidence interval for *RR* is given by

$\mathrm{CI(RR)} = [\max\{ g_1(\mathbf{p}) - z_{\alpha/2}\sqrt{v(g_1(\mathbf{p}))}, 0\}, g_1(\mathbf{p}) + z_{\alpha/2}\sqrt{v(g_1(\mathbf{p}))}]$. The lower limit is used as the maximum of $\{g_1(\mathbf{p}) - z_{\alpha/2}\sqrt{v(g_1(\mathbf{p}))}\}$ and 0 to ensure an acceptable range for *RR* which lies between 0 and ∞.

3.2 Absolute Risk Increase

The absolute risk increase is commonly used when the risk of disease in an exposed group is higher than that of the unexposed group, and is given by $\mathrm{ARI} = P(D=1 \mid E=1) - P(D=1 \mid E=0)$. When the exposure has a protective effect, the absolute risk reduction (ARR) is used instead of ARI and is given by $\mathrm{ARR} = P(D=1 \mid E=0) - P(D=1 \mid E=1)$. In this paper, it is assumed that the factor is harmful for the development of the disease, that is, the risk in exposed group is higher than that of the unexposed group and hence ARI is being used for the assessment of risk. In terms of parameters π_{ij}, it turns out that $\mathrm{ARI} = \dfrac{\pi_{11}}{\pi_{1.}} - \dfrac{\pi_{01}}{\pi_{0.}}$. Let $g_2(\boldsymbol{\pi}) = \dfrac{\pi_{11}}{\pi_{1.}} - \dfrac{\pi_{01}}{\pi_{0.}}$. Then an MLE of ARI is given by $g_2(\mathbf{p}) = \dfrac{p_{11}}{p_{1.}} - \dfrac{p_{01}}{p_{0.}}$. An estimate of the asymptotic variance of $g_2(\mathbf{p})$ is given by $v(g_2(\mathbf{p})) = \left\{ \dfrac{p_{11}p_{10}}{p_{1.}^{\,3}} + \dfrac{p_{01}p_{00}}{p_{0.}^{\,3}} \right\} / n$ (See Appendix for detail). Then, a $100(1-\alpha)$ percent confidence interval for *ARI* is given by

$\mathrm{CI(ARI)} = [\max\{ g_2(\mathbf{p}) - z_{\alpha/2}\sqrt{v(g_2(\mathbf{p}))}, 0\}, \min\{ g_2(\mathbf{p}) + z_{\alpha/2}\sqrt{v(g_2(\mathbf{p}))}, 1\}]$.

The maximum and minimum functions in the interval are used to ensure acceptable range for ARI which lies between 0 and 1.

3.3 Population Attributable Risk

The population attributable risk is the excess risk attributable to the exposure of the risk factor in the population and is given by $\text{PAR} = [P(D=1 \mid E=1) - P(D=1 \mid E=0)]\,P(E=1)$. Using the cell probabilities, it is easy to note that $\text{PAR} = \left[\frac{\pi_{11}}{\pi_{1.}} - \frac{\pi_{01}}{\pi_{0.}}\right]\pi_{1.} = \pi_{11} - \frac{\pi_{01}\pi_{1.}}{\pi_{0.}}$. Let $g_3(\boldsymbol{\pi}) = \pi_{11} - \frac{\pi_{01}\pi_{1.}}{\pi_{0.}}$. An MLE of PAR is then given by $g_3(\mathbf{p}) = p_{11} - \frac{p_{01}p_{1.}}{p_{0.}}$. By the principle of delta method, an estimate of the variance of $g_3(\mathbf{p})$ is given by

$$v(g_3(\mathbf{p})) = \left[\frac{1}{p_{0.}^{\,3}}\left\{p_{0.}\left(p_{11}p_{00}^{\,2} + p_{10}p_{01}^{\,2}\right) + p_{1.}^{\,2}p_{01}p_{00}\right\} - (g_3(\mathbf{p}))^2\right]/n$$

(See Appendix for detail). Therefore, a $100(1-\alpha)$ percent confidence interval for *PAR* is given by

$$\text{CI(PAR)} = \max\{v(g_3(\mathbf{p})) - z_{\alpha/2}\sqrt{v(g_3(\mathbf{p}))}, -1\}, \min\{v(g_3(\mathbf{p})) + z_{\alpha/2}\sqrt{v(g_3(\mathbf{p}))}, 1\}]$$

The maximum and the minimum functions are used to ensure acceptable range for PAR which ranges between -1 and 1.

3.4 Population Attributable Fraction

The population attributable fraction is the proportion of disease in the population that could be avoided by completely eliminating the risk factor from the population and is given by $\text{PAF} = \frac{P(D=1) - P(D=1 \mid E=0)}{P(D=1)}$. In terms of parameters π_{ij}, it follows that $\text{PAF} = 1 - \frac{\pi_{01}}{\pi_{.1}\pi_{0.}}$. Let $g_4(\boldsymbol{\pi}) = 1 - \frac{\pi_{01}}{\pi_{0.}\pi_{.1}}$. Therefore, an MLE of PAF is given by $g_4(\mathbf{p}) = 1 - \frac{p_{01}}{p_{.1}p_{0.}}$. Using delta method, it follows that an asymptotic variance of $g_4(\mathbf{p})$ is given by $V(g_4(\mathbf{p})) = \phi^2 \cdot V(\log\hat{\phi})$, where $\phi = \frac{\pi_{01}}{\pi_{0.}\pi_{.1}}$ and $V(\log\hat{\phi})$ is given in [4, 12] by

$V(\log\hat{\phi}) = \left[\frac{1-\pi_{01}}{\pi_{01}} - \frac{\pi_{0.} + \pi_{.1} - 2\pi_{01}}{\pi_{0.}\pi_{.1}}\right]/n$. Therefore, an estimate of the variance of $g_4(\mathbf{p})$ is given by $v(g_4(\mathbf{p})) = \hat{\phi}^2 \cdot \left[\frac{1-p_{01}}{np_{01}} - \frac{p_{0.} + p_{.1} - 2p_{01}}{np_{0.}p_{.1}}\right]$. An alternative, but equivalent, form of an estimate of the variance of $g_4(\mathbf{p})$ appears in [3] and is given by

$v(g_4(\mathbf{p})) = (1-g_4(\mathbf{p}))^4 (p_{0.}p_{.1}) \cdot \{p_{01}(p_{10}p_{01} - p_{11}p_{00}) + p_{11}p_{00}\}$. Once the value of $v(g_4(\mathbf{p}))$ is obtained, an asymptotic $100(1-\alpha)$ percent confidence interval for PAF by using Wald's statistic is given by

$$\mathrm{CI(PAF)} = [v(g_4(\mathbf{p})) - z_{\alpha/2}\sqrt{v(g_4(\mathbf{p}))}, \min\{v(g_4(\mathbf{p})) + z_{\alpha/2}\sqrt{v(g_4(\mathbf{p}))}, 1\}]$$

The upper limit is considered as the minimum of $v(g_4(\mathbf{p})) + z_{\alpha/2}\sqrt{v(g_4(\mathbf{p}))}$ and 1 to ensure an acceptable range for PAF which lies between $-\infty$ and 1.

3.5 Attributable Fraction in Exposed

The attributable fraction in exposed is given by

$$\mathrm{AF_e} = \frac{P(D=1|\,\mathrm{E}=1) - P(D=1|\,E=0)}{P(D=1|\,\mathrm{E}=1)} = 1 - \frac{1}{\mathrm{RR}}.$$

An MLE of $\mathrm{AF_e}$ is then given by $\hat{\mathrm{A}}\mathrm{F}_e = 1 - 1/\hat{\mathrm{R}}\mathrm{R}$.

By the use of the delta method, an estimate of the asymptotic variance of $\hat{AF}_e$ is $v(\hat{\mathrm{A}}\mathrm{F}_e) = (1/\hat{\mathrm{R}}\mathrm{R})^4 v(\hat{\mathrm{R}}\mathrm{R})$. Then an asymptotic $100(1-\alpha)$ percent confidence interval for AF_e by using Wald's statistic is given by

$$\mathrm{CI(AF_e)} = \left[\hat{\mathrm{A}}\mathrm{F_e} - z_{\alpha/2}\sqrt{v(\hat{\mathrm{A}}\mathrm{F_e})}, \min\left\{\hat{\mathrm{A}}\mathrm{F_e} + z_{\alpha/2}\sqrt{v(\hat{\mathrm{A}}\mathrm{F_e})}, 1\right\}\right]$$

4. Definition of Impact Numbers and Their Characteristics

Introduced by Heller et al. [8], four impact numbers, namely, population impact number (PIN), exposure impact number (EIN), case impact number (CIN) and exposed case impact number (ECIN) are defined as follows:

$$\mathrm{PIN} = \frac{1}{\mathrm{PAR}}, \mathrm{PAR} \neq 0$$

$$\mathrm{EIN} = \frac{1}{\mathrm{ARI}}, \mathrm{ARI} \neq 0$$

$$\mathrm{CIN} = \frac{1}{\mathrm{PAF}}, \mathrm{PAF} \neq 0$$

$$\mathrm{ECIN} = \frac{1}{\mathrm{AF_e}}, \mathrm{AF_e} \neq 0$$

The necessity of these impact numbers are well documented in [8]. It is clear from above definitions that four impact numbers are defined when the effect measures they relate to are non-zeros, that is, there is an association between the exposure to the risk factor and the disease outcome. Also, with simple algebraic manipulation or by the definition of total law of probability, it easily follows that

$$\mathrm{ARI} = P(D=1|\,E=0)(\mathrm{RR}-1)$$

$$\mathrm{PAR} = P(D=1|\,E=0)P(E=1)(\mathrm{RR}-1)$$

$$\text{PAF} = \frac{P(D=1 \mid E=0)P(E=1)(\text{RR}-1)}{P(D=1)}$$

$$\text{AF}_e = 1 - \frac{1}{\text{RR}}$$

In other words, it is evident that each of these measures can be expressed in terms of RR. As the value of RR increases, the values of ARI, PAR, PAF and AF_e also increase and of course their higher values are indication of stronger association between exposure to the risk factor and disease. But, with the increasing values of ARI, PAR, PAF and AF_e, the impact numbers EIN, CIN, PIN and ECIN decrease and their lower values are indication of lower average numbers for which one disease case corresponds to exposure to the risk factor. When the related effect measure approaches 0^- (0 from the left) or 0^+ (0 from the right), the corresponding impact measure approaches $-\infty$ or $+\infty$. The higher is the value of an impact number, the lower is the impact of the corresponding exposure to the risk of the disease. When impact number approaches $-\infty$ or $+\infty$, the exposure to the risk factor tends to be insignificant for the development of the disease. Hildebrandt et al. [10] recommend using the impact numbers for the presentation of study results in public health research only in the case of studies showing significant exposure effects. In the situation of statistically non-significant study results, they recommended just absolute and relative frequencies complemented by point and interval estimates of a relation effect measure, which can be interpreted easily in all situations, e.g. the risk ratio.

5. Estimation of Impact Numbers

The point estimates of these impact numbers follow directly from the corresponding effect measures by the invariance property of the MLEs reported in section 3. That is,

$$\hat{\text{PIN}} = \frac{1}{g_3(\mathbf{p})},\ g_3(\mathbf{p}) \neq 0$$

$$\hat{\text{EIN}} = \frac{1}{g_2(\mathbf{p})},\ g_2(\mathbf{p}) \neq 0$$

$$\hat{\text{CIN}} = \frac{1}{g_4(\mathbf{p})},\ g_4(\mathbf{p}) \neq 0$$

$$\hat{\text{ECIN}} = \frac{1}{\hat{\text{AF}}_e},\ \hat{\text{AF}}_e \neq 0$$

All point estimates should be accompanied by an estimate of their precision. This is commonly done by quoting confidence intervals (CIs), which take the estimation uncertainty into account. Below, we review the principle of inverting and exchanging the limits of the standard effect measures for constructing confidence interval estimates for impact numbers and also consider their estimates by using the delta method.

5.1 CIs by Inverting and Exchanging the Limits of Effect Measures

Since impact numbers are defined as the reciprocals of some standard effect measures, Hildebrandt et al. [10] apply the principle of inverting and exchanging the confidence limits of the corresponding standard effect measures to construct confidence limits of

associated impact numbers. Following the principle, the $100(1-\alpha)$ per cent confidence intervals for impact numbers are given as follows:

$$\text{CI(EIN)} = \left[\frac{1}{\text{UCL(ARI)}}, \frac{1}{\text{LCL(ARI)}}\right] \quad (1)$$

$$\text{CI(PIN)} = \left[\frac{1}{\text{UCL(PAR)}}, \frac{1}{\text{LCL(PAR)}}\right] \quad (2)$$

$$\text{CI(CIN)} = \left[\frac{1}{\text{UCL(PAF)}}, \frac{1}{\text{LCL(PAF)}}\right] \quad (3)$$

$$\text{CI(ECIN)} = \left[\frac{1}{\text{UCL}(\text{AF}_e)}, \frac{1}{\text{LCL}(\text{AF}_e)}\right] \quad (4)$$

where $\text{UCL}(\cdot)$ and $\text{LCL}(\cdot)$ are respectively the upper and lower confidence limits of the corresponding effect measures.

The confidence interval for impact numbers using the principle of inverting and exchanging the limits of the standard effect measures they relate to often may lead to interpretational difficulties. For example, if a 95% confidence interval for ARI is found to be (-0.25, 0.25), then by the principle of inverting and exchanging the limits, the confidence interval for EIN is (4.0, -4.0), which is a mathematical nonsense [13]. Lesaffre and Pledger [13] pointed out that it is only sensible to invert end points over a region where the reciprocal transformation is continuous. Also, the method of inverting and exchanging numbers does not necessarily give a 95% confidence interval since $P(a < Y < b) = 1-\alpha \equiv P\{g(a) < g(Y) < g(b)\} = 1-\alpha$ is true only if the function $g(\cdot)$ is one to one and direction of inequality is checked [14]. Indeed, an interval of the form (4,-4) is expressed as the union of two disjoint interval: $(-\infty,-4) \cup (4,\infty)$, which is outside the connected interval (4,-4), [14-16]. Even though various interpretation of such interval appears in literature, the width of such interval is infinity which causes problem in evaluation of the simulation results to be compared. Some of these difficulties could be avoided if impact numbers are used when being exposed to the risk factor is found to be significant for the development of the disease as measured by standard effect measures such as RR, OR, ARI or AR for which risk factor could be categorized as either protective or harmful.

In section below, the confidence interval estimates of impact numbers using delta method are considered which are expected to perform reasonably well than those obtained by the principle of inverting and exchanging the limits of standard effect measures. Later, with real life examples we justify that the confidence interval estimates of impact numbers by delta method are better as compared with those by the principle of inverting and exchanging the limits of the standard effect measures in terms of the lengths of intervals.

5.2 CIs by Delta Method

It is easy to find the estimates of the asymptotic variances of the impact numbers if the estimates of the asymptotic variances of the standard effect measures they relate to are known by means of the delta method. Using the delta method, the estimates of the asymptotic variances of $\hat{\mathrm{EIN}}$, $\hat{\mathrm{PIN}}$, $\hat{\mathrm{CIN}}$ and $\hat{\mathrm{ECIN}}$ are given by

$$v(\hat{\mathrm{EIN}}) = \hat{\mathrm{EIN}}^4 \, v(\hat{\mathrm{ARI}})$$

$$v(\hat{\mathrm{PIN}}) = \hat{\mathrm{PIN}}^4 \, v(\hat{\mathrm{PAR}})$$

$$v(\hat{\mathrm{CIN}}) = \hat{\mathrm{CIN}}^4 \, v(\hat{\mathrm{PAF}})$$

$$v(\hat{\mathrm{ECIN}}) = \hat{\mathrm{ECIN}}^4 \, v(\hat{\mathrm{AF}}_e)$$

See Appendix for more detail.

Once the estimates of the asymptotic variance of the impact numbers are obtained, the corresponding asymptotic $100(1-\alpha)$ per cent confidence intervals using Wald's statistics are given by

$$\mathrm{CI}^*(\mathrm{EIN}) = \hat{\mathrm{EIN}} \pm z_{\alpha/2}\sqrt{v(\hat{\mathrm{EIN}})} \qquad (1^*)$$

$$\mathrm{CI}^*(\mathrm{PIN}) = \left[\hat{\mathrm{PIN}} - z_{\alpha/2}\sqrt{v(\hat{\mathrm{PIN}})}, \hat{\mathrm{PIN}} + z_{\alpha/2}\sqrt{v(\hat{\mathrm{PIN}})}\right] \qquad (2^*)$$

$$\mathrm{CI}^*(\mathrm{CIN}) = \hat{\mathrm{CIN}} \pm z_{\alpha/2}\sqrt{v(\hat{\mathrm{CIN}})} \qquad (3^*)$$

$$\mathrm{CI}^*(\mathrm{ECIN}) = \hat{\mathrm{ECIN}} \pm z_{\alpha/2}\sqrt{v(\hat{\mathrm{ECIN}})} \qquad (4^*)$$

These confidence intervals are expected to have lower confidence lengths than those obtained by the principle of inverting and exchanging the confidence limits of effect measures. The performance of these methods is justified with a real life example and simulations in the next sections.

6. Application

An example is considered in this section as an application to compare the performance of confidence interval estimates for impact numbers by means of the principle of inverting and exchanging the limits of the corresponding effect measures and delta method. For the completeness of the discussion, confidence intervals of the related effect measures are included as well.

The data for this example consists of 1329 subjects and appears before in [4, 17, 18]. As reported in Table 2, the exposure (serum cholesterol) has two levels 0 (=SC<200mg%) and 1 (=SC 200+mg%) and disease (coronary heart disease, CHD) has two levels 1(=developed CHD after 6 years) and 0 (=no CHD after 6 years). As noted in [4, 18], since the data was collected by the use of a perspective study design with unstratified

sampling, the formula of a cross-sectional study applies to the data set. We wish to see the impact of a higher serum cholesterol level on CHD.

Table 2: Serum cholesterol (SC) and CHD

Exposure Status	CHD Yes (1)	CHD No (0)	Total
SC 200+mg% (1)	72	684	756
SC<200 mg% (0)	20	553	573
Total	92	1237	1329

Table 3 includes point and 95% confidence interval estimates of all undertaken measures. It follows from Table 3 that the effect of high serum cholesterol level on CHD is statistically significant at 5% level of significance as measured by RR, ARI, PAR, PAF and AF_e.

Table 3: Estimates with 95% CIs for various measures using data in Table 2

Measures	Estimator	CI		CI^*		Length	Length CI^*
RR	2.729	1.410	4.048				
ARI	0.060	0.034	0.086				
PAR	0.034	0.019	0.049				
PAF	0.491	0.478	0.504				
AF_e	0.634	0.457	0.811				
PIN	29.4	20.4	52.6	16.7	42.1	32.2	25.4
EIN	16.7	11.63	29.41	9.52	23.88	17.8	14.4
CIN	2.04	1.984	2.092	1.985	2.095	0.11	0.11
ECIN	1.58	1.233	2.188	1.14	2.02	0.96	0.88

How are the impacts of high serum cholesterol level on CHD in terms of the impact numbers? The estimate of CIN is 2.04, which implies that for every 2 persons who had a CHD, on average one case is attributable to higher level of serum cholesterol. The estimate for ECIN is 1.58, which implies that for every 2 persons with higher serum cholesterol level who had a CHD, on average one case is attributable to higher level of serum cholesterol. The PIN estimate of 29.4 implies that for every 29 persons in the population, on average one CHD is attributable to higher level of serum cholesterol. The EIN estimate of 16.7 implies that for every 17 persons with a higher level of serum cholesterol level, on average one CHD is attributable to the higher level of serum cholesterol. It also follows from Table 3 that the 95% confidence interval estimates of impact numbers, CI^*, provided by delta method has lower confidence length than those of CI using the principle of inverting and exchanging the confidence limits of the corresponding effect measures.

7. Simulation Studies

In order to evaluate the finite sample performance of the confidence interval estimates of impact numbers, a Monte Carlo simulation study is carried out in this section. The simulation performance is measured in terms of the coverage probability and confidence length. We dropped the confidence interval estimate CI for comparison with that of CI* because any estimate CI of the form [4,-4] as mentioned in section 5.1 by the principle of inverting and exchanging the limits of the standard effect measures they relate to results in confidence length of ∞. This fact would make the comparison of CI and CI* unreasonable and unjustifiable and hence the coverage probability and confidence lengths of CI* based on the delta method appears in Table 4 of simulation results.

For simulation purpose, the known multinomial distributions are considered with values of parameter n chosen arbitrarily as 150, 250, 500 and 1000 for each of three choices of $\boldsymbol{\pi} = (\pi_{00}, \pi_{01}, \pi_{10}, \pi_{11})$ by (0.3, 0.1, 0.3, 0.3), (0.3, 0.1, 0.15, 0.45) and (0.34, 0.06, 0.15, 0.45) respectively which results in true values of RR respectively as 2, 3 and 5. The corresponding three combinations of (PIN, EIN, CIN, ECIN) in the population are: (6.67, 4.0, 2.67, 2.00), (3.33, 2.00, 1.83, 1.50) and (2.78, 1.67, 1.42, 1.25). The coverage probability is calculated over 10,000 repeated samples generated from a multinomial distributions with parameters *n* and $\boldsymbol{\pi}$. The average confidence length is calculated from those interval estimates contain the true parameter values of the impact numbers. The result of simulation appears in Table 4 for a Monte Carlo simulation of size 10,000.

Table 4: Estimates of coverage probabilities with corresponding confidence length (in parenthesis) for 95% confidence interval estimates, CI*, of various impact numbers (INs) for arbitrarily selected values of parameters *n* and RRs.

INs	*n*	RR=2	RR=3	RR=5
PIN	150	0.9167 (36.4274)	0.9420 (2.1772)	0.9450 (1.4069)
EIN		0.9133 (21.9311)	0.9339 (1.1864)	0.9390 (0.7192)
CIN		0.9077 (14.1296)	0.9303 (1.2034)	0.9360 (0.6531)
ECIN		0.9068 (8.5554)	0.9309 (0.7093)	0.9378 (0.3862)
PIN	250	0.9282 (8.3121)	0.9493 (1.6298)	0.9474 (1.0551)
EIN		0.9239 (4.8746)	0.9463 (0.8891)	0.9436 (0.5387)
CIN		0.9230 (3.2744)	0.9440 (0.9038)	0.9322 (0.4904)
ECIN		0.9224 (1.9438)	0.9443 (0.5323)	0.9328 (0.2897)
PIN	500	0.9430 (4.8679)	0.9516 (1.1212)	0.9472 (0.7307)
EIN		0.9390 (2.8542)	0.9515 (0.6134)	0.9439 (0.3729)
CIN		0.9391 (1.9234)	0.9501 (0.6245)	0.9403 (0.3418)
ECIN		0.9381 (1.1439)	0.9516 (0.3687)	0.9415 (0.2022)
PIN	1000	0.9478 (3.2112)	0.9503 (0.7768)	0.9499 (0.5139)
EIN		0.9469 (1.8830)	0.9524 (0.4249)	0.9484 (0.2626)
CIN		0.9451 (1.2710)	0.9492 (0.4327)	0.9478 (0.2411)
ECIN		0.9458 (0.7565)	0.9494 (0.2549)	0.9467 (0.1424)

As reported in Table 4, the coverage probability of 95% confidence interval estimates of various impact numbers gets closer to the nominal level of 0.95 as the value of the parameter *n* increases. One explanation to this situation is that the estimate $\mathbf{p} = (p_{11}, p_{10}, p_{01}, p_{00})$ of $\boldsymbol{\pi} = (\pi_{00}, \pi_{01}, \pi_{10}, \pi_{11})$ is a consistent estimator for large

values of the parameter *n*. Therefore, a sufficiently large value of *n* is required when a simulation is performed from the given multinomial distribution with parameters *n* and $\boldsymbol{\pi}$. It also follows from the result that the length of confidence interval estimates gets smaller as the sample size *n* gets larger.

8. Conclusion

The impact numbers each reflects the impact of a risk factor on different populations - PIN to the entire population, CIN to those with outcome, EIN to those exposed and ECIN to those exposed and have outcomes. Therefore, the choice any of these measures is determined by the purpose of the study and the interest of the researchers. Due to the interpretational simplicity of the impact numbers, these measures may be preferred by people having difficulties in interpreting standard effect measures such as RR, OR, PAR, PAF and $AR_{e,}$. Also, when the standard effect measures are significant, the inclusion of impact numbers may help communicate the study result better to policy makers, health administrators, and consumers and thus may add a new dimension to epidemiological research by allowing researchers to describe the risk of a factor from different perspective on different populations.

As applied to real-life example to construct confidence interval estimates of impact numbers, the length of confidence interval using CI^{*} is much lower than that of CI used by Hildebrandt et al. [10]. However, the comparison of the two methods are not considered in simulation because the confidence length by the principle of inverting and exchanging the limits of standard effect measures often provides confidence length of ∞ due to the fact that some interval estimate takes the form of (4,-4) mentioned in section 5.1 even if the effect of the factor is significant in true population. Also, from the example appears in section 6, it is evident that confidence interval estimates of impact numbers provided by delta method has shorter confidence length as compared with those by the principle of inverting and exchanging the limits of the standard effect measures. Therefore, it suffices to justify the performance of confidence interval estimates of impact numbers given by the delta method from the Monte Carlo simulation in terms of coverage probability and confidence length. The result of the simulation is satisfactory since the coverage probability of the interval estimates of impact numbers gets closer to the nominal level of 0.95 for sufficiently large values of *n* and the length gets smaller for larger values of *n*. It also appears that the coverage probability is sensitive to the value of the relative risk of the associated multinomial population in that larger relative risk corresponds to relatively higher coverage probability and lower confidence length.

References

1. Doll, R. (1959). Occupational lung cancer: a review. *British Journal of Industrial Medicine*, **16**: 181-190.
2. Levin, M. L. (1953). The occurrence of lung cancer in man. *Acta Unio Internationalis Contra Cancrum*, **9**: 531-541.
3. Walter S. D. (1976). The estimation and interpretation of attributable risk in health research. *Biometrics*, **32**: 829-849.
4. Lui, K. J. (2001). Notes on interval estimation of the attributable risk in cross-sectional sampling. *Statistics in Medicine*, **20**: 1797-1809.
5. Benichou, J. (1991). Methods of adjustment for estimating the attributable risk in case-control studies: a review. *Statistics in Medicine*, **10**: 1753-1773.

6. Whittemore, A. S. (1982). Statistical methods for estimating attributable risk from retrospective data. *Statistics in Medicine*, **1**: 229-243.
7. Lilienfeld, A. M. (1973). Epidemiology of infectious and non-infectious disease: some comparisons. *American Journal of Epidemiology*, **97**: 135-147.
8. Heller, R. F., Dobson, A. J., Attia, J. and Page, J. (2002). Impact numbers: measures of risk factor impact on the whole population from case-control and cohort studies. *J Epidemiol Community Health*, **56**: 606-610.
9. Rothman, K. J. and Greenland, S. (1998). *Modern Epidemiology*, 2nd edn. Lippincott-Raven: Philadelphia.
10. Hiderbrandt, M., Bender, R., Ulrich, G. and Blettner, M. (2006). Calculating confidence intervals for impact numbers. *BMC Medical Research Methodology*, **6**:32.
11. Agresti, A. (2002). *Categorical Data Analysis*, 2nd edn. Wiley: New York.
12. Lui, K. J. (2004). *Statistical Estimation of Epidemiological Risk*. Wiley.
13. Lesaffre, E. and Pledger, G. W. (1999). A note on the number needed to treat. *Contr. Clin, Trials*, **20**: 439-447.
14. Hutton, J.L. (2000). Number needed to treat: properties and problems. *J. R. Statist. Soc. A*, **163**: 403-419.
15. Altman, D. G. (1998). Confidence intervals for the number needed to treat. *Br. Med. J.*, **317**: 1309-1312.
16. Bender, R., Kuss, O., Hildebrandt, M., and Gehrmann, U. (2007). Estimating adjusted NNT measures in logistic regression analysis. *Statistics in Medicine*. **26**: 5586-5595.
17. Walter, S. D. (1976). Calculation of attributable risks from epidemiologic data. *International Journal of Epidemiology*, **7**: 175-182.
18. Leung, H. M. and Kupper, L. L. (1981). Comparisons of confidence intervals for attributable risk. *Biometrics*, **37**: 293-302.

Appendix

Using notations of section 2 it follows that

$$\partial_k \boldsymbol{\Sigma} \partial_k' = \sum_i \sum_j \pi_{ij} \left(\frac{\partial g_k(\boldsymbol{\pi})}{\partial \pi_{ij}} \right)^2 - \left[\sum_i \sum_j \sum_k \pi_{ij} \left(\frac{\partial g_k(\boldsymbol{\pi})}{\partial \pi_{ij}} \right) \right]^2$$

The above expression is a very useful for simplifying the expression for the variance of $g_k(\mathbf{p})$. An estimator of $V(g_k(\mathbf{p}))$, $v(g_k(\mathbf{p}))$, can be computed by substituting the MLEs p_{ij} for π_{ij}.

Let $g_1(\boldsymbol{\pi}) = RR = \frac{\pi_{11}\pi_{0.}}{\pi_{01}\pi_{1.}}$. It is easy to see that

$$\frac{\partial g_1(\boldsymbol{\pi})}{\partial \pi_{11}} = \frac{\pi_{0.}\pi_{10}}{\pi_{01}\pi_{1.}^2}, \frac{\partial g_1(\boldsymbol{\pi})}{\partial \pi_{10}} = -\frac{\pi_{0.}\pi_{11}}{\pi_{01}\pi_{1.}^2}, \frac{\partial g_1(\boldsymbol{\pi})}{\partial \pi_{01}} = -\frac{\pi_{11}\pi_{00}}{\pi_{1.}\pi_{01}^2} \text{ and } \frac{\partial g_1(\boldsymbol{\pi})}{\partial \pi_{00}} = \frac{\pi_{11}}{\pi_{01}\pi_{1.}}$$

Let $\partial_1 = \left(\frac{\partial g_1(\boldsymbol{\pi})}{\partial \pi_{11}}, \frac{\partial g_1(\boldsymbol{\pi})}{\partial \pi_{10}}, \frac{\partial g_1(\boldsymbol{\pi})}{\partial \pi_{01}}, \frac{\partial g_1(\boldsymbol{\pi})}{\partial \pi_{00}} \right)$.

It follows that $\sum_i \sum_j \pi_{ij} \left(\frac{\partial g_1(\boldsymbol{\pi})}{\partial \pi_{ij}} \right) = 0$ and

$\sum_i \sum_j \pi_{ij} \left(\frac{\partial g_1(\boldsymbol{\pi})}{\partial \pi_{ij}} \right)^2 = \frac{\pi_{11}\pi_{0.}}{{\pi_{01}}^3 {\pi_{1.}}^3} \left(\pi_{10}\pi_{01}\pi_{0.} + \pi_{11}\pi_{00}\pi_{1.} \right)$. Then by delta method, $nV(g_1(\mathbf{p})) = \partial_k \boldsymbol{\Sigma} \partial_k' = \frac{\pi_{11}\pi_{0.}}{{\pi_{01}}^3 {\pi_{1.}}^3} \left(\pi_{10}\pi_{01}\pi_{0.} + \pi_{11}\pi_{00}\pi_{1.} \right)$. An MLE of $V(g_1(\mathbf{p}))$ is then

$$v(g_1(\mathbf{p})) = \left\{ \frac{p_{11}p_{0.}}{{p_{01}}^3 {p_{1.}}^3} \left(p_{10}p_{01}p_{0.} + p_{11}p_{00}p_{1.} \right) \right\} / n .$$

In a similar manner, letting $g_2(\boldsymbol{\pi}) = \frac{\pi_{11}}{\pi_{1.}} - \frac{\pi_{01}}{\pi_{0.}}$, it follows that

$\frac{\partial g_2(\boldsymbol{\pi})}{\partial \pi_{11}} = \frac{\pi_{10}}{{\pi_{1.}}^2}$, $\frac{\partial g_2(\boldsymbol{\pi})}{\partial \pi_{10}} = -\frac{\pi_{11}}{{\pi_{1.}}^2}$, $\frac{\partial g_2(\boldsymbol{\pi})}{\partial \pi_{01}} = -\frac{\pi_{00}}{{\pi_{0.}}^2}$ and $\frac{\partial g_2(\boldsymbol{\pi})}{\partial \pi_{00}} = \frac{\pi_{01}}{{\pi_{0.}}^2}$. Also it follows that

$\sum_i \sum_j \pi_{ij} \left(\frac{\partial g_2(\boldsymbol{\pi})}{\partial \pi_{ij}} \right) = 0$ and $\sum_i \sum_j \pi_{ij} \left(\frac{\partial g_2(\boldsymbol{\pi})}{\partial \pi_{ij}} \right)^2 = \frac{\pi_{11}\pi_{10}}{{\pi_{1.}}^3} + \frac{\pi_{01}\pi_{00}}{{\pi_{0.}}^3}$. Then by the delta method and the property of MLEs, an MLE of the variance of $g_2(\mathbf{p})$ is given by

$$v(g_2(\mathbf{p})) = \left\{ \frac{p_{11}p_{10}}{{p_{1.}}^3} + \frac{p_{01}p_{00}}{{p_{0.}}^3} \right\} / n .$$

Let $g_3(\boldsymbol{\pi}) = PAR = \pi_{11} - \frac{\pi_{01}\pi_{1.}}{\pi_{0.}}$. It follows that

$\frac{\partial g_3(\boldsymbol{\pi})}{\partial \pi_{11}} = \frac{\pi_{00}}{\pi_{0.}}$, $\frac{\partial g_3(\boldsymbol{\pi})}{\partial \pi_{10}} = -\frac{\pi_{01}}{\pi_{0.}}$, $\frac{\partial g_3(\boldsymbol{\pi})}{\partial \pi_{01}} = -\frac{\pi_{1.}\pi_{00}}{{\pi_{0.}}^2}$ and $\frac{\partial g_3(\boldsymbol{\pi})}{\partial \pi_{00}} = \frac{\pi_{1.}\pi_{01}}{{\pi_{0.}}^2}$.

Then, $\sum_i \sum_j \pi_{ij} \left(\frac{\partial g_3(\boldsymbol{\pi})}{\partial \pi_{ij}} \right) = \frac{\pi_{11}\pi_{00} - \pi_{10}\pi_{01}}{\pi_{0.}}$ and

$$\sum_i \sum_j \pi_{ij} \left(\frac{\partial g_3(\boldsymbol{\pi})}{\partial \pi_{ij}} \right)^2 = \frac{\pi_{0.}(\pi_{11}{\pi_{00}}^2 + \pi_{10}{\pi_{01}}^2) + {\pi_{1.}}^2 \pi_{01}\pi_{00}}{{\pi_{0.}}^3} .$$

Also, $g_3(\boldsymbol{\pi}) = \pi_{11} - \frac{\pi_{01}\pi_{1.}}{\pi_{0.}} = \frac{\pi_{11}(\pi_{00} + \pi_{01}) - \pi_{01}(\pi_{11} + \pi_{10})}{\pi_{0.}} = \frac{\pi_{11}\pi_{00} - \pi_{10}\pi_{01}}{\pi_{0.}}$.

Then, by the principle of the delta method, an estimate of the variance of $g_3(\mathbf{p})$ is given by

$$v(g_3(\mathbf{p})) = \left[\frac{1}{{p_{0.}}^3} \left\{ p_{0.}({p_{11}}{p_{00}}^2 + {p_{10}}{p_{01}}^2) + {p_{1.}}^2 p_{01}p_{00} \right\} - (g_3(\mathbf{p}))^2 \right] / n$$