Generalized Population Attributable Risk Estimation

Michael J. Kahn[1] , W.M. O'Fallon[2] and JoRean D. Sicks[2]

Technical Report # 54
revised July, 2000

[1]Dept. of Mathematics
St. Olaf College
Northfield, Minnesota 55057

[2]Biostatistics Section
Mayo Clinic
Rochester, Minnesota 55901

Supported in part by Research Grants NS06663 and AR 30582, National Institutes of Health, U.S.P.H.S.

Contents

1. Preface

This technical report is being written to present in one place a unified discussion of the concept of attributable risk(**AR**). This discussion includes not only basic introductory material, but also a unified mathematical construct which allows several natural extensions including adjustment for covariates and multiple risk factors. The most significant contributions are 1) a generalized definition and associated point and interval estimates for **AR** under several study designs and 2) S-PLUS (MathSoft Inc., Becker, Chambers, and Wilks, 1988) software developed for estimation of **AR** and the standard error of the estimate in these settings. A shell archive containing the software and related materials can be accessed via anonymous `ftp` from `ftp.stolaf.edu` in the directory `pub/kahn/atrisk`. Questions can be mailed directly to `kahn@stolaf.edu`.

2. Introduction and Definitions

Levin (1953) seems to have been the first to introduce the idea of attributable risk. In his manuscript, which pertained to the occurrence of lung cancer in males, he reviewed four studies and Table V of his manuscript contains a column entitled, "Indicated percent of all lung cancer 'attributable' to smoking." His derivation of the formula used to calculate this percent was brief but correct. The percents quoted in the table ranged from 56 to 92 and Levin said, "If the latter figure is correct, elimination of smoking would almost eliminate lung cancer (other factors remaining the same) whereas if 56 is nearer the true figure, then elimination of smoking would reduce lung cancer by about one-half, if smoking is a truly causative agent with respect to lung cancer."

This quote is important, not because of any residual lung cancer-smoking controversy, but because it illustrates the potential public health utility of the concept of attributable risk and because it properly warns of the need for causality to be established before attribution can be legitimately converted to action.

Subsequent to Levin's publication, many authors have dealt with the concept of attributable risk, first to discuss its utility and how to estimate it [Cole (1971), Miettinen (1974), Markush (1977)] and then to establish a strong statistical foundation for the estimation of attributable risk [Miettinen (1974), Walter (1975)]. Eventually, [Walter (1976, 1978), Leung (1981), Denman (1983), Kuritz (1987)] expressions for, and estimates of, the standard errors of these various attributable risk estimates were obtained. More recently, there has been much written about how to take into consideration the influence of other factors while estimating the risk attributable to a specific factor of interest [Miettinen (1974), Ejigou (1979), Walter (1980, 1983), Whittemore (1982, 1983), Bruzzi (1985), Kuritz (1988), Benichou (1990, 1991), Coughlin (1991), Drescher (1991), Kooperberg (1991)].

There has also been some recent work concerning a related quantity, prevented fraction, in cross-sectional studies by Garguillo, et al (1995).

2.1. Definitions

Unfortunately, no consensus has arisen as to an "official" terminology or symbolism. In general, the term *attributable risk*, which we will abbreviate **AR**, is used to refer to what is more properly called the *population attributable risk*. Some authors [Miettinen (1974), Kleinbaum (1982)] use the term *etiologic fraction* (EF) to refer to the same concept. Traditionally **AR** will be expressed as a percent, as Levin did, although the usual formulas will not show the multiplier (100) necessary to convert a proportion to a percent.

The concept of attributable risk is deceptively simple. It is the proportion of those diseased members of a population who are diseased because they possess or were exposed to some "risk factor." By way of a simple and completely artificial example, suppose there are 1000 diseased individuals in a population but there would have been only 600 if no one in the population possessed the risk factor. Thus, there are apparently 400 "excess" cases and the attributable risk is 40% (400/1000).

Implicit in this discussion is the idea of a population of individuals, some of whom may possess the risk factor and some of whom may be or may become diseased. Less clearly, there is also some implication regarding time in reference to the population and the disease. At a particular point in time the proportions of people in the population at that time who; i) have the disease, called the *prevalence rate of the disease*, or ii) have the risk factor, called the *prevalence rate of the factor*, are the two main proportions available. A study designed solely to obtain prevalence data is called a *cross-sectional study*. The attributable risk estimates obtained from such data would also refer to the "excess" disease in the population at that point in time.

If members of a population are followed for a period of time to observe newly diagnosed cases of the disease, the rate (cases per year of follow-up) of newly diagnosed disease is called the *incidence rate*. A study designed to estimate incidence of a disease is generically referred to as a *longitudinal study*. Normally, the incidence rate would be estimated separately for those with and without the risk factor and the attributable risk estimates obtained from incidence data would apply to new cases of the disease.

In the expressions which follow, the probabilities may be thought of as prevalence rates or incidence rates or just plain proportions. The precise interpretation of each depends upon the nature of the

information available. As is standard notation, let

$\mathbf{P}(D)$ – probability of disease ,

$\mathbf{P}(F)$ – probability of risk factor ,

$\mathbf{P}(D \mid F)$ – conditional probability of disease among those with the factor ,

$\mathbf{P}(D \mid \overline{F})$ – conditional probability of disease among those without the factor ,

$\mathbf{P}(F \mid D)$ – conditional probability of the factor being present among those who are diseased ,

$\mathbf{P}(F \mid \overline{D})$ – conditional probability of the factor being present among those not diseased ,

$\mathbf{P}(D \mid F)/\mathbf{P}(D \mid \overline{F})$ – relative risk (**RR**).

Relative risk (**RR**) is properly thought of in terms of the ratios of two cumulative incidence rates. However, it is often approximated by the ratio of the odds of being diseased with the factor to the odds of being diseased without the factor. We will assume $\mathbf{RR} > 1$ which implies that the factor is truly a risk factor and not a protective factor.

Kleinbaum, et al. (1982), defined population attributable risk as I^*/I where I is the total number of diseased individuals in the population and I^* the number attributed to the factor. In terms of the probabilities defined above, if there is a homogeneous population of N individuals, then $I = N\,\mathbf{P}(D)$ = expected number of diseased individuals in the population. If the probability of disease among those with the risk factor could be reduced to that of those without the risk factor then $N\,\mathbf{P}(D \mid \overline{F})$ = expected number of diseased individuals. Thus, $I^* = N\,\mathbf{P}(D) - N\mathbf{P}(D \mid \overline{F}) =$ excess number of diseased individuals and

$$\mathbf{AR} = I^*/I = \frac{\mathbf{P}(D) - \mathbf{P}(D \mid \overline{F})}{\mathbf{P}(D)}. \tag{1}$$

This is a commonly seen formula for **AR** whose use requires estimates of the rate (prevalence or incidence) of disease in the total population and of the rate of disease in those without the factor.

Some algebra involving conditional probability arguments and Bayes' theorem lead from equation (1) to the expression (2) below which is actually the form used by Levin,

$$\mathbf{AR} = \frac{\mathbf{P}(F)\,(\mathbf{RR} - 1)}{\mathbf{P}(F)\,(\mathbf{RR} - 1) + 1}\,. \tag{2}$$

This is the most common version of **AR** . Its use in providing an estimate of **AR** requires an estimate of the relative risk, **RR** , and an estimate of $\mathbf{P}(F)$, the prevalence of the risk factor in the population. Such estimates could actually come from different studies.

Apparently Miettinen (1974) was the first to note that more algebra could lead to the following formula for **AR** ,

$$\mathbf{AR} = \mathbf{P}(F \mid D)[\mathbf{RR} - 1]/\mathbf{RR}\,. \tag{3}$$

The use of (3) to estimate **AR** requires an estimate of **RR**, available from all study types, and an estimate of $\mathbf{P}(F \mid D)$, the prevalence of the factor among the diseased individuals. An unbiased estimate of $\mathbf{P}(F \mid D)$ may be obtained from a random sample of cases. Hospitalized patients are unlikely to be a random sample although an estimate based on them might not be seriously biased depending on the disease. A population-based study, using all incidence cases of the disease occurring during some time period, would be the best study from which to base estimates of **AR** using equation (3).

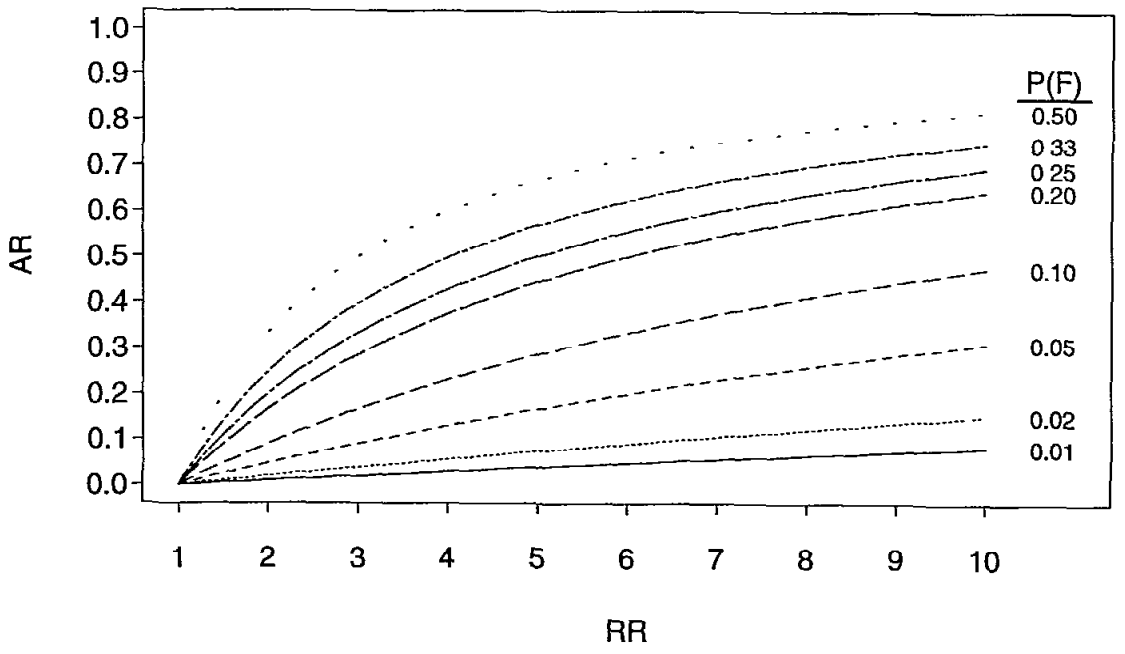
Equations (2) and (3) demonstrate the dual influence of relative risk and the prevalence of the factor on the value of attributable risk. This is further illustrated in figure 1. Thus, a factor which has an extremely high relative risk will have minimal impact on the population if it occurs only rarely. For example, suppose a factor has a relative risk of $\mathbf{RR} = 10$ but is found in only 1% of the population. Then $\mathbf{AR} = .01(10-1)/[.01(10-1)+1] = .09/1.09 = .0826$. That is, such a factor would account for barely 8% of the disease in the population. Conversely, a factor with a relative risk of 2 and a prevalence of 50% would have an **AR** of $.5(2-1)/[.5(2-1)+1] = .5/1.5 = .333$, accounting for 1/3 of the cases of disease. If such a common factor also had a high relative risk (e.g., 10) it would account for over 80% of the disease.

Equations (2) and (3) are the primary formulas for estimating **AR** . Their derivation provided tools for estimating **AR** but, until it became possible to test hypotheses about **AR** and/or to obtain confidence interval estimates, **AR** was not a particularly useful concept.

The null hypothesis of interest would be that there is no excess disease attributable to the risk factor, $H_0 : \mathbf{AR} = 0$. Obviously, from equations (2) and (3), $\mathbf{RR} = 1$ implies $\mathbf{AR} = 0$ as does lack of the risk factor in the population or among the diseased (i.e. $\mathbf{P}(F) = 0 = \mathbf{P}(F \mid D)$) which would make the whole discussion very uninteresting. Thus, in any realistic situation, hypothesis testing regarding **AR** is equivalent to hypothesis testing regarding **RR**. Since there is an extensive literature on testing hypotheses about **RR** we will not address the issue further.

If $\widehat{\mathbf{AR}}$ is an estimate of **AR**, it is affected by random influences, through the sampling scheme, and so has a probability distribution. This distribution in turn has a variance (V), the square root of which is the standard error of the estimate (SE). Estimating V depends on the sampling scheme (study design) which led to the estimate $\widehat{\mathbf{AR}}$. The first problem is to find an expression for V and then to determine how to estimate it, i.e., to find $\hat{V}$, an estimate of V. Then an estimate of SE is given by $\widehat{SE} = \sqrt{\hat{V}}$.

Miettinen (1974) put it well when he stated that, "The sampling variability of the above estimators poses a rather challenging problem. No results are available." Walter (1975, 1976, 1978) seems to have been the first to seriously attack the problems of determining the distribution of $\widehat{\mathbf{AR}}$ and of



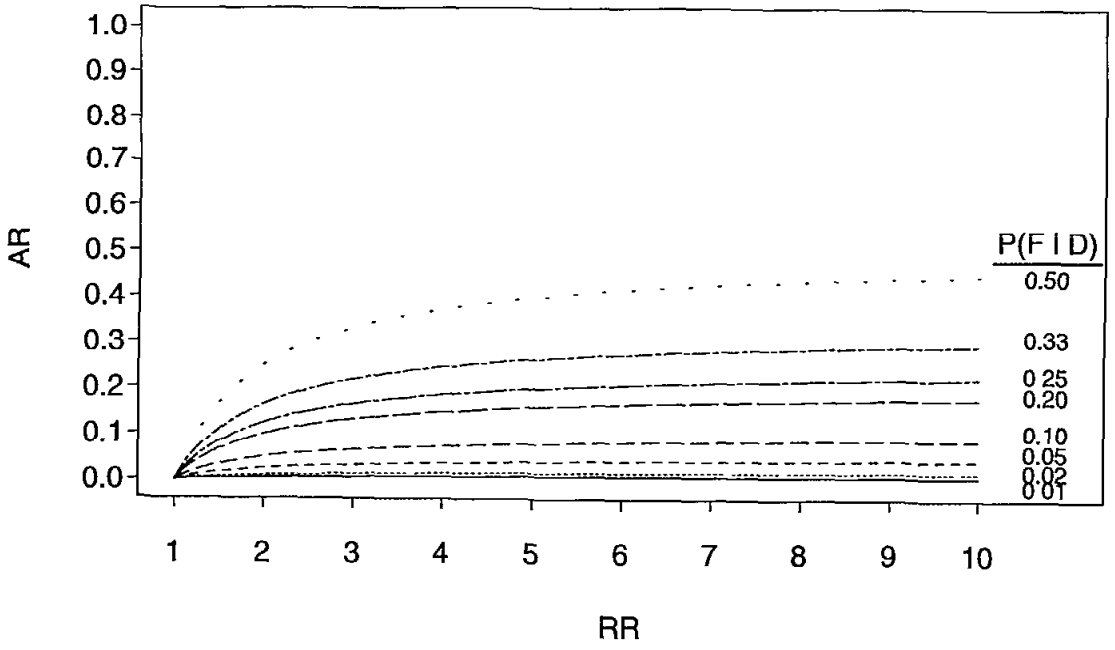


Figure 1 Relationship between **RR** and **AR** as a function of 1) prevalence of risk factor, $\mathbf{P(F)}$ and 2) prevalence of risk factor among diseased $\mathbf{P(F \mid D)}$.

estimating the standard error of $\widehat{\mathbf{AR}}$ in general (i.e., not assuming H_0 to be true). He first established (1975) that $\widehat{\mathbf{AR}}$ has approximately a Gaussian (normal) distribution if the sample sizes are large enough. Then (1976, 1978) he provided some alternative formulas for V and $\hat{V}$ depending on the study design. These formulas are asymptotically correct, i.e., they work well for "large" samples. These two results could then be used to obtain confidence intervals. Thus,

$$\widehat{\mathbf{AR}} \pm z_{\alpha/2}\sqrt{\hat{V}} \tag{4}$$

is approximately a $100(1-\alpha)\%$ confidence interval for $\mathbf{AR}$ where $z_{\alpha/2}$ is the $100(1-\alpha/2)$ percentile of the standard Gaussian distribution.

Walter (1976) proposed a log transformation involving $\ln\left(1-\widehat{\mathbf{AR}}\right)$ whose variance could be estimated more precisely so that corresponding confidence intervals would perhaps be more correct and/or narrower. Leung and Kupper (1981) proposed a logit transformation using $\text{logit}\left(\widehat{\mathbf{AR}}\right) \equiv \ln\left(\widehat{\mathbf{AR}}/\left(1-\widehat{\mathbf{AR}}\right)\right)$ to achieve the same purpose. They established that, for situations in which $.21 < \widehat{\mathbf{AR}} < .79$, the confidence intervals for $\mathbf{AR}$ derived using the logit transformation are shorter than the confidence intervals resulting from using the variances from Walter in equation (4). This result was confirmed by Whittemore (1982) who concluded that there was no advantage to the use of the log transformation and that the logit transformation should be used in the interval specified and otherwise confidence intervals should be obtained using equation (4).

2.2. Special Designs

We will touch briefly on three basic study designs and give examples of their application in the stroke literature.

Design I : A single random (cross-sectional) sample of the population with the sampled subjects being followed for a period of time. This might be described as a cross-sectional sample with a longitudinal component.

Design II : Two random samples (stratified cross-sectional), one of exposed and the other of unexposed members of the population with the sampled subjects being followed for a period of time.

Design III : Two random samples, one of cases of the disease and the other of members of the population without the disease. This is called a *case/control study.* There are two versions of this design depending on whether the controls are selected randomly or matched individually to the cases.

A generic display of the data from each such design involving a single dichotomous risk factor may be presented in the familiar tabular form as below:

		Disease		
		Yes	No	
Risk Factor	Yes	a	b	m_1
	No	c	d	m_2
		n_1	n_2	N

Design I

N - size of random sample from the population

$m_1 = a + b$ - number with the risk factor (exposed) at beginning of follow-up

a - number of exposed individuals who develop disease during follow-up

$m_2 = c + d$ - number without the risk factor (unexposed) at beginning of follow-up

c - number of unexposed who develop disease during follow-up

Design II

m_1 - size of sample of exposed individuals

a - number of exposed individuals who develop disease during follow-up

m_2 - size of sample of unexposed individuals

c - number of unexposed individuals who develop disease during follow-up

Design III-a

n_1 - number of cases

n_2 - number of controls

a - number of cases with the risk factor

b - number of controls with the risk factor

Design III-b (Note that the generic table does not apply here.)

		Controls		
		Exposed	Not Exposed	
Case	Exposed	a	b	
	Not Exposed	c	d	
				N

N - total number of cases (also total number of controls since they are 1 : 1 matched)

a - number of case/control pairs in which both have the risk factor (both are exposed)

b - number of case/control pairs in which case is exposed and control is not

c - number of case/control pairs in which control is exposed and case is not

d - number of case/control pairs in which neither is exposed

We now give forms for estimating **AR** and the associated standard error of the estimator in each of the designs.

2.2.1. Design I

$$\widehat{\mathbf{RR}} = \frac{a}{m_1} / \frac{c}{m_2} = \frac{am_2}{cm_1} . \tag{5}$$

$\mathbf{P}(F)$ is estimated by m_1/N and, using equation (2),

$$\widehat{\mathbf{AR}} = \frac{\frac{m_1}{N}\left[\frac{am_2}{cm_1} - 1\right]}{\frac{m_1}{N}\left[\frac{am_2}{cm_1} - 1\right] + 1} = \frac{(ad - bc)}{n_1 m_2} . \tag{6}$$

From Walter (1978) an estimate of the asymptotic variance of $\widehat{\mathbf{AR}}$ for this design is given by

$$\hat{V} = \frac{cN\left[ad\,(N - c) + bc^2\right]}{n_1^3 m_2^3} . \tag{7}$$

The logit transformation yields confidence intervals as follows:

$$\left(\frac{ad - bc}{(ad - bc) + Nce^{u}}, \frac{ad - bc}{(ad - bc) + Nce^{-u}}\right) \tag{8}$$

where

$$u = z_{\alpha/2}\sqrt{\frac{(a + c)\,(c + d)\,(ad\,(N - c) + bc^2)}{Nc\,(ad - bc)^2}} .$$

2.2.2. Design II

$$\widehat{\mathbf{RR}} = \frac{am_2}{cm_1} . \tag{9}$$

Since $\mathbf{P}(F)$ cannot be estimated directly from this design it must be obtained from some other source. Assume θ is the value of $\mathbf{P}(F)$ so obtained. Then,

$$\widehat{\mathbf{AR}} = \frac{\theta\left(\frac{am_2}{cm_1} - 1\right)}{\theta\left(\frac{am_2}{cm_1} - 1\right) + 1} = \frac{\theta\,(ad - bc)}{\theta\,(ad - bc) + cm_1} , \tag{10}$$

and from Walter (1976), assuming θ is known without error,

$$\hat{V} = \left(\frac{am_2}{ad - bc}\right)^2 \left(\frac{b}{am_1} + \frac{d}{cm_2}\right) . \tag{11}$$

The logit transformation yields confidence intervals as follows:

$$\left(\frac{\theta\left(am_2 - cm_1\right)}{\theta\left(am_2 - cm_1\right) + cm_1 e^{u}}, \frac{\theta\left(am_2 - cm_1\right)}{\theta\left(am_2 - cm_1\right) + cm_1 e^{-u}}\right)$$

where

$$u = z_{\alpha/2}\left(\frac{am_2}{ad - bc}\right)\sqrt{\frac{b}{am_1} + \frac{d}{cm_2}}\,.$$

2.2.3. Design III-a

In case-control designs the relative risk estimate is based on the odds ratio. Thus, in this unmatched design,

$$\widehat{\mathbf{RR}} = \frac{ad}{bc} \tag{12}$$

and, using equation (3) and assuming the cases in the study represent all cases so $\mathbf{P}(F \mid D)$ can be estimated by a/n_1,

$$\widehat{\mathbf{AR}} = \left(\frac{a}{n_1}\right)\left(\frac{\frac{ad}{bc} - 1}{\frac{ad}{bc}}\right) = \frac{ad - bc}{dn_1}\,, \tag{13}$$

and from Walter (1978),

$$\hat{\mathbf{V}} = \left(\frac{cn_2}{dn_1}\right)^2\left(\frac{a}{cn_1} + \frac{b}{dn_2}\right)\,. \tag{14}$$

The logit transformation yields the following confidence interval for $\mathbf{AR}$,

$$\left(\frac{ad - bc}{(ad - bc) + c\left(b + d\right)e^{u}}, \frac{ad - bc}{(ad - bc) + c\left(b + d\right)e^{-u}}\right)$$

where

$$u = z_{\alpha/2}\frac{d\left(b + d\right)}{ad - bc}\sqrt{\frac{a}{c\left(a + c\right)} + \frac{b}{d\left(b + d\right)}}\,.$$

2.2.4. Design III-b

With one control "matched" to each case

$$\widehat{\mathbf{RR}} = \frac{b}{c}\,,$$

and, using equation (3) and assuming the cases in the study represent all cases so $\mathbf{P}(F \mid D)$ can be estimated by m_1/N,

$$\widehat{\mathbf{AR}} = \frac{a + b}{N}\left(\frac{\frac{b}{c} - 1}{\frac{b}{c}}\right) = \frac{\left(a + b\right)\left(b - c\right)}{bN}\,. \tag{15}$$

Then, from Kuritz and Landis (1987)

$$\hat{V} = \left(\frac{1}{bN}\right)^2\left(a\left(b - c\right)^2 + \frac{\left(b^2 + ac\right)^2}{b} + c\left(a + b\right)^2 - \frac{\left(a + b\right)^2\left(b - c\right)^2}{N}\right)\,. \tag{16}$$

2.3. *Examples*

2.3.1. *Example (Design I)*

Davis, et al (1987), reported on a study of all residents of Rochester, MN, who had a general examination in 1960. Those over age 50 with no history of stroke or TIA at the time of this examination were followed to determine the occurrence of such events. The results of the first 5 years of follow-up are summarized in reference to a diagnosis of hypertension (on antihypertensive therapy or $BP \geq 160/95$ mmHg in the medical record) in the following table.

		Stroke/TIA		
		Yes	No	
Hypertension	Yes	70	892	962
	No	9	805	814
		79	1697	1776

Using equations (5), (6), and (7) we obtain

$$\widehat{\mathbf{RR}} = \frac{(70)(814)}{(9)(962)} = 6.58\,,$$

$$\hat{\mathbf{P}}(F) = \frac{962}{1776} = .542\,,$$

$$\widehat{\mathbf{AR}} = \frac{(70)(805) - (9)(892)}{(79)(814)} = .751\,,$$

$$\hat{V} = 9(1776)\frac{(70)(805)(1767) + (892)(9^2)}{(79^3)(814^3)} = .00599\,,$$

$$\widehat{SE} = .0774\,.$$

The 95% confidence interval for **AR** is $.751 \pm 1.96(.0774)$ or (60.0% , 90.3%).

The logit-transformed 95% confidence interval for **AR** is (57.3% , 87.2%).

From this example it would appear that between 57% and 90% of cases of stroke or TIA occurring during the 5-year follow-up could be attributed to the diagnosis of hypertension.

2.3.2. *Example (Design II)*

Wiebers, et al., (1990) reported on the five-year follow-up of 566 patients with asymptomatic carotid bruit and 428 without bruit. Among the 566, sixty-three (63) experienced unilateral or bilateral carotid system cerebral ischemic symptoms (TIA or cerebral infarction) during the follow-up whereas among the 428, only thirteen (13) experienced such events. This may be summarized as:

	Carotid System Cerebral Ischemic Symptoms		
	Yes	No	
With Bruit	63	503	566
Without Bruit	13	415	428

From equations (9), (10), and (11), respectively

$$\widehat{\mathbf{RR}} = \frac{(63)(428)}{(13)(566)} = 3.66\,,$$

$$\widehat{\mathbf{AR}} = \frac{\theta\left((63)(415)-(503)(13)\right)}{\theta\left((63)(415)-(503)(13)\right)+(13)(566)} = \frac{\theta(19606)}{\theta(19606)+7358}\,,$$

$$\begin{aligned}\hat{V} &= \theta^2\left(\frac{(63)(13)(566)(428)}{(\theta((63)(415)-(503)(13))+(13)(566))^2}\right)^2\left(\frac{503}{(63)(566)}+\frac{415}{(13)(423)}\right)\\ &= \theta^2\left(\frac{198401112}{(\theta(19606)+7358)^2}\right)^2(.08869286)\,.\end{aligned}$$

In an earlier study Sandok, et al (1982), reported the prevalence of asymptomatic carotid bruit to be 12.6% while in this study the process of selecting a sample free of bruit led to a prevalence estimate of 10.3%. Using $\theta_1 = .126$ and $\theta_2 = .103$ in the prior equations yields respectively, two estimates, $\widehat{\mathbf{AR}}_1$ and $\widehat{\mathbf{AR}}_2$, for $\mathbf{AR}$ as follows:

$$\widehat{\mathbf{AR}}_1 = \frac{(19606)(.126)}{(19606)(.126)+7358} = .251 \quad (25.1\%)\,,$$

$$\widehat{\mathbf{AR}}_2 = \frac{(19606)(.103)}{(19606)(.103)+7358} = .215 \quad (21.5\%)\,.$$

The corresponding variance estimates would be:

$$\hat{V}_1 = (.126)^2\left(\frac{198401112}{((19606)(.126)+7358)^2}\right)^2(.08869286) = (.126)^2(2.053914)^2(.08869286) = .005940109\,,$$

$$\widehat{SE}_1 = \sqrt{\hat{V}_1} = .077\,,$$

$$\hat{V}_2 = (.103)^2\left(\frac{198401112}{((19606)(.103)+7358)^2}\right)^2(.08869286) = (.103)^2(2.2562)^2(.08869286) = .004789809\,,$$

$$\widehat{SE}_2 = \sqrt{\hat{V}_2} = .069\,.$$

The corresponding 95% confidence intervals for **AR** would be

$$.251 \pm 1.96(.0771) \text{ or } (10.0\%, 40.2\%)$$

using θ_1 and

$$.215 \pm 1.96(.069) \text{ or } (8.0\%, 35.1\%)$$

using θ_2.

The 95% confidence intervals based on the logit transformation are (13.1%, 42.8%) using θ_1 and (11.0%, 38.0%) using θ_2.

Thus, it would appear that somewhere between 8% and 40% of episodes of "cerebral ischemic symptoms" could be attributed to the presence of asymptomatic carotid bruits. This example shows that even relatively minor differences in the value of θ can have rather major effects on $\hat{V}$ and, consequently, on the corresponding interval estimates.

2.3.3. Example (Design III-a)

In a recent study, all cases of cerebral infarcts occurring in residents over the age of 50 of Rochester, MN, during the period from 1960-1984 (inclusive) were identified, as were a corresponding number of controls. All medical records were reviewed for evidence of hypertension (on therapy or with recorded $BP \geq 160/95$ mmHg). The data are summarized in the following table:

		Cases	Controls
Hypertension	Yes	938	763
	No	384	559
		1322	1322

From equations (12), (13), and (14) respectively,

$$\widehat{\mathbf{RR}} = \frac{(938)(559)}{(384)(763)} = 1.79\,,$$

$$\hat{\mathbf{P}}(F \mid D) = \frac{938}{1322} = .71\,,$$

$$\widehat{\mathbf{AR}} = \frac{(938)(559) - (763)(384)}{(559)(1322)} = .313\,,$$

$$\hat{V} = \left(\frac{(384)(1322)}{(559)(1322)}\right)^2 \frac{938}{(384)(1322)} = \frac{763}{(559)(1322)} = .0014\,,$$

$$\widehat{SE} = .0369 .$$

This gives a 95% confidence interval of $.313 \pm 1.96(.0369)$ or $(24.1\%, 38.5\%)$. The 95% CI based on the logit transformation is $(24.6\%, 38.9\%)$.

2.3.4. Example (Design III-b)

The case-control study of ischemic stroke described in the previous example was actually designed as a matched-pairs study with a single control being matched to each case on the basis of gender, age at time of stroke (±5 years) and calendar year at time of stroke (±5 years). The data are summarized in the following matched-pair tables and associated analysis.

		Controls		
		HBP	No HBP	
Case	HBP	563	375	
	No HBP	200	184	
				1322

$$\widehat{\mathbf{RR}} = \frac{375}{200} = 1.875 ,$$

$$\hat{\mathbf{P}}(F \mid D) = 938/1322 = .71 ,$$

$$\widehat{\mathbf{AR}} = \frac{(938)(375-200)}{(375)(1322)} = .331 ,$$

$$\hat{V} = \left(\frac{1}{(375)(1322)}\right)^2 \left((562)(175)^2 + \frac{\left(375^2 + (563)(200)\right)^2}{375} + (200)(938)^2 - \frac{(938)^2(175)^2}{1322}\right) = .00014 ,$$

$$\widehat{SE} = .0374 .$$

Thus, a 95% confidence interval is $.331 \pm 1.96(.0374)$ or $(25.8\%, 40.4\%)$.

The unmatched analysis of the same data resulted in an **AR** estimate of 31.3% not appreciably different from the 33.1% obtained by this, more proper, analysis.

2.3.5. Example, matched set case-control

Finally, there is another design, a matched-set case-control design, in which the notion of attributable risk is of interest. This design does not readily admit to a simple tabular summary or analysis as has been discussed for the previous designs. It is similar to a matched case-control design except that now there may be multiple controls matched to each case.

The example used in this manuscript comes from a population based study of temporal arteritis (Machado, et al, 1989). In this study, a case was defined to be a patient suffering from temporal arteritis. Four controls were matched to each case by age (within a year) and gender.

One question of interest was the extent, if any, to which smoking contributed to the risk of temporal arterits. After adjusting for whether a patient had a history of angina, the methods to be described in section 3.3.3 of this manuscript yield an estimated attributable risk of 0.235 with a standard error of approximately 0.056.

3. The mathematics of attributable risk

In this section we review some of the mathematical background concerning population attributable risk from section 2 and generalize estimation of **AR** in case-control designs.

Recall that for a dichotomous risk factor, F, the population attributable risk (**AR**) is defined by

$$\mathbf{AR} = \frac{\mathbf{P}(D) - \mathbf{P}(D \mid \overline{F})}{\mathbf{P}(D)} = 1 - \frac{\mathbf{P}(D \mid \overline{F})}{\mathbf{P}(D)},$$

which, as we saw in (3), can be rewritten as

$$\mathbf{AR} = \mathbf{P}(F \mid D)\left(1 - \frac{1}{\mathbf{RR}}\right). \tag{17}$$

From this we can see how both $\mathbf{P}(F \mid D)$ and **RR** affect **AR**, namely
$0 \leq \mathbf{AR} \leq \min\left(\mathbf{P}(F \mid D), 1 - \frac{1}{\mathbf{RR}}\right)$.

3.1. Case-Control Studies

Equation (17) is most useful for estimating **AR** in case-control designs since sampling is stratified by disease status. However, the probabilities, $\mathbf{P}(D \mid F)$ and $\mathbf{P}(D \mid \overline{F})$ are not directly estimable and, hence, $\mathbf{RR} = \mathbf{P}(D \mid F)/\mathbf{P}(D \mid \overline{F})$ is not directly estimable in case-control studies. Since sampling is stratified by disease status, the estimable quantities are $\mathbf{P}(F \mid D)$, $\mathbf{P}(F \mid \overline{D})$. Hence, $\mathbf{odds}(F \mid D) = \mathbf{P}(F \mid D)/\mathbf{P}(\overline{F} \mid D)$, $\mathbf{odds}(F \mid \overline{D}) = \mathbf{P}(F \mid \overline{D})/\mathbf{P}(\overline{F} \mid \overline{D})$ and the *odds ratio* (**OR**),

defined by

$$\mathbf{OR} = \frac{\mathbf{odds}(F \mid D)}{\mathbf{odds}(F \mid \overline{D})} = \frac{\mathbf{P}(F \mid D)\mathbf{P}(\overline{F} \mid \overline{D})}{\mathbf{P}(F \mid \overline{D})\mathbf{P}(\overline{F} \mid D)} .$$

are estimable. Equivalently, the above expression can be written as

$$\mathbf{OR} = \frac{\mathbf{odds}(D \mid F)}{\mathbf{odds}(D \mid \overline{F})} = \frac{\mathbf{P}(D \mid F)\mathbf{P}(\overline{D} \mid \overline{F})}{\mathbf{P}(D \mid \overline{F})\mathbf{P}(\overline{D} \mid F)} .$$

Thus, we see that

$$\mathbf{OR} = \mathbf{RR}\left[\frac{\mathbf{P}(\overline{D} \mid \overline{F})}{\mathbf{P}(\overline{D} \mid F)}\right] ,$$

which can also be written as

$$\mathbf{RR} = \frac{(1 - \mathbf{P}(D))\,\mathbf{OR} + \mathbf{P}(D)\left(\frac{\mathbf{P}(F \mid D)}{\mathbf{P}(F \mid \overline{D})}\right)}{(1 - \mathbf{P}(D)) + \mathbf{P}(D)\left(\frac{\mathbf{P}(F \mid D)}{\mathbf{P}(F \mid \overline{D})}\right)} .$$

So, $\mathbf{RR} \approx \mathbf{OR}$ if

i) $\mathbf{P}(D) \approx 0$ (i.e. disease is rare), or

ii) $\mathbf{P}(F \mid D) << \mathbf{P}(F \mid \overline{D})$.

If either of i) or ii) above are reasonable assumptions then the **OR** estimate is nearly the same as the **RR** estimate. The rare disease assumption, i), is often tenable in case-control studies where odds ratios are estimated as measures of association between risk factors and disease. Under the assumption that $\mathbf{RR} > 1$, ii) is untenable.

3.1.1. Odds Ratio and Logistic Regression

Logistic regression, as a way of modelling probabilities, odds and odds ratios, provides an important tool for expanding analysis of case-control studies from the single dichotomous risk factor to the more realistic circumstance in which there are multiple risk factors and/or confounders and/or effect modifiers. Let

$$F = \begin{cases} 0 & \text{if individual is not exposed} \\ 1 & \text{if individual is exposed} \end{cases}$$

and

$$D = \begin{cases} 0 & \text{if individual is not diseased}(\textit{control}) \\ 1 & \text{if individual is diseased}(\textit{case}) . \end{cases}$$

For a real-valued parameter, θ_1, the logistic model specifies

$$\mathbf{P}(D = 1 \mid F = 1) = \frac{1}{1 + \mathrm{e}^{-\theta_1}} = 1 - \mathbf{P}(D = 0 \mid F = 1) ,$$

which implies that $\mathbf{odds}(D=1\,|\,F=1)=\mathrm{e}^{\theta_1}$. This then yields

$$\mathrm{logit}\,(P\,(D=1\,|\,F=1))\equiv\ln(\mathbf{odds}\,(D=1\,|\,F=1))\;=\;\theta_1\,.$$

Similarly, for $\theta_0\in\mathbb{R}$,

$$\mathrm{logit}\,(P\,(D=1\,|\,F=0))\equiv\ln(\mathbf{odds}\,(D=1\,|\,F=0))\;=\;\theta_0\,.$$

Together these yield

$$\ln(\mathbf{OR})=\theta_1-\theta_0\,.$$

More generally,

$$\mathbf{P}(D=1\,|\,F)=\frac{1}{1+\mathrm{e}^{-(\alpha+\beta F)}}=1-\mathbf{P}(D=0\,|\,F)\,.$$

Thus,

$$\mathbf{odds}(D=1\,|\,F)=\mathrm{e}^{\alpha+\beta F}\Leftrightarrow\mathrm{logit}\,(\mathbf{P}\,(D=1\,|\,F))=\alpha+\beta F\,,$$

which gives

$$\begin{aligned}\ln(\mathbf{OR})&=\ln\left(\frac{\mathbf{odds}\,(D=1\,|\,F=1)}{\mathbf{odds}\,(D=1\,|\,F=0)}\right)\\&=\ln(\mathbf{odds}(D=1\,|\,F=1))-\ln(\mathbf{odds}(D=1\,|\,F=0))\\&=\alpha+\beta\,(1)-(\alpha+\beta\,(0))\\&=\beta\end{aligned}$$

$$\Leftrightarrow\quad\mathbf{OR}\;=\;\mathrm{e}^{\beta}\,.$$

Hence, in a case-control design logistic regression can be used for estimating **OR** (and thus, **RR**) by estimating β.

3.1.2. Polychotomous Exposures with No Confounding Factors

As before, define

$$D=\begin{cases}0 & \text{if individual is not diseased}\\1 & \text{if individual is diseased}\end{cases}$$

Now we will assume a polychotomous exposure variable. That is, suppose that there exist $K+1$ exposure categories, $E_0,\ldots,E_K$ and for $i=1,\ldots,K$, define

$$F_i=\begin{cases}0 & \text{if individual is not exposed at level } E_i\\1 & \text{if individual is exposed at level } E_i\end{cases}$$

Define $\mathbf{F}_i$ to be the vector in $\mathbb{R}^K$ with i^{th} component equal to 1 and all other components equal to 0, and $\mathbf{F}_0 = (0,\ldots,0)' \in \mathbb{R}^K$. Define the risk of exposure level i relative to exposure level 0 (i.e. $\mathbf{RR}_{i0}$) by

$$\mathbf{RR}_{i0} = \frac{P(D\,|\mathbf{F}_i)}{P(D\,|\mathbf{F}_0)}\,.$$

$\mathbf{OR}_{i0}$ is defined similarly, namely

$$\mathbf{OR}_{i0} \equiv \frac{P(D=1\,|\mathbf{F}_i)}{P(D=0\,|\mathbf{F}_i)}\frac{P(D=0\,|\mathbf{F}_0)}{P(D=1\,|\mathbf{F}_0)}$$

The logistic model then specifies for $i = 0,\ldots,K$

$$\mathbf{odds}(D=1\,|\mathbf{F}_i) = \mathrm{e}^{\alpha+\beta'\mathbf{F}_i} \;\Leftrightarrow\; \mathrm{logit}\,(\mathbf{P}\,(D=1\,|\,\mathbf{F}_i)) = \alpha+\beta'\mathbf{F}_i\,,$$

for parameters α and $\beta' = (\beta_1,\ldots,\beta_K)$. In this case,

$$\mathrm{logit}\,(\mathbf{P}\,(D=1\,|\,\mathbf{F}_i)) = \alpha+\beta_i\;,\;\; \mathrm{logit}\,(\mathbf{P}\,(D=1\,|\,\mathbf{F}_0)) = \alpha\;,\;\; \ln(\mathbf{OR}_{i0}) = \beta_i \;\Leftrightarrow\; \mathbf{OR}_{i0} = \mathrm{e}^{\beta_i}\,.$$

As per our earlier discussion, $\mathbf{OR}_{i0} \approx \mathbf{RR}_{i0}$ provided that $\mathbf{P}(D=1)$ is small. Thus, for case control studies we will use $\mathbf{OR}_{i0}$ as an approximation to $\mathbf{RR}_{i0}$ and we estimate $\mathbf{OR}_{i0}$ by $\widehat{\mathbf{OR}}_{i0} = \mathrm{e}^{\hat{\beta}_i}$.

3.1.3. Continuous Risk Factor

If there are "risk factors" such as blood pressure, body mass index or level of tobacco use, the logistic regression model provides a useful tool for assessing the relationship between that factor and the disease. Thus, if $\mathbf{F}$ represents a (possibly vector-valued) risk factor, then

$$\ln(\mathbf{OR}_{\mathbf{F}\,\mathbf{F}_0}) = \beta'(\mathbf{F}-\mathbf{F}_0)$$

is a measure of the risk to an individual with level $\mathbf{F}$ relative to an individual with level $\mathbf{F}_0$. This is equivalent to $\mathbf{OR}_{i0}$ as defined in the previous section using $\mathbf{F} = \mathbf{F}_i$.

3.1.4. Multiple Risk Factors and Confounding Factors

Within the above settings it is possible, indeed likely, that multiple risk factors and confounding factors must be accounted for. If the vector $\mathbf{X} = (X_1,\ldots,X_m)$ represents such confounding factors and $\mathbf{F} = (F_1,\ldots,F_K)$, then the logistic model defines

$$\mathrm{logit}(P(D=1|\,\mathbf{F},\mathbf{X})) =$$

$$\alpha+\gamma'\mathbf{X}+\beta'\mathbf{F}\,. \tag{18}$$

Then with the vector of baseline exposure levels denoted as $\mathbf{F}_0$, we have

$$\ln(\mathbf{OR}_{\mathbf{F}\,\mathbf{F}_0}\,|\,\mathbf{X}) = \beta'(\mathbf{F}-\mathbf{F}_0)$$

as before.

3.2. *General Target Distribution*

Finally, we describe how to generalize the target distribution. Instead, of referring each observed exposure level $\mathbf{F}$ to a single exposure value (vector) $\mathbf{F}_0$ as above, suppose the target exposure level is itself determined by the observed exposure level. For example, instead of targeting all HBP to $\overline{HBP}$, we target levels of diastolic blood pressure (DBP) to be reduced by 10%.

In general, we would like to choose a target distribution based on $\mathbf{F}_0 \equiv \mathbf{F}_0(\mathbf{F})$. In the above blood pressure example, $DBP_0 = .9\,DBP$. Using $\mathbf{F}_0$ we can, for a given vector of exposure levels, create the corresponding vector of target exposure levels.

OR is still of interest and is estimable by

$$\ln\left(\mathbf{OR}_{\mathbf{F}_i\,\mathbf{F}_0(\mathbf{F}_i)\,|\,\mathbf{X}}\right) = \beta'\left(\mathbf{F}_i - \mathbf{F}_0(\mathbf{F}_i)\right)$$

where $\mathbf{F}_i$ is the (vector of) observed exposure level(s) for individual i and $\mathbf{F}_0(\mathbf{F}_i)$ is that individual's targeted exposure level.

This provides a very general context for estimating **OR** using the logistic model. The risk factor, $\mathbf{F}$, may represent something as simple as a single dichotomous factor or may incorporate multiple, interrelated risk factors of different forms, including interactions among the risk factors and/or interactions between risk factors and confounding factors.

3.3. *Logistic Regression and Odds Ratios in Case-Control Designs*

The logistic regression model described above accomodates both the unmatched, pair-matched and matched set case-control designs. (For more detail on these issues, see Breslow and Day, 1980.) Here we give a simple outline of how logistic regression is used to estimate **OR** in these designs.

3.3.1. *Unmatched Case-Control Design*

For an unmatched case-control design the estimated odds ratios for each case can be obtained through the previously described logistic regression models. This follows since

$$\begin{aligned}\mathbf{OR} &= \frac{P(D=1\,|\,\mathbf{F},\mathbf{X})\,P(D=0\,|\,\mathbf{F}_0,\mathbf{X})}{P(D=0\,|\,\mathbf{F},\mathbf{X})\,P(D=1\,|\,\mathbf{F}_0,\mathbf{X})} \\ &= \frac{P(\mathbf{F}\,|\,D=1,\mathbf{X})\,P(\mathbf{F}_0\,|\,D=0,\mathbf{X})}{P(\mathbf{F}\,|\,D=0,\mathbf{X})\,P(\mathbf{F}_0\,|\,D=1,\mathbf{X})}.\end{aligned}$$

The last expression employs probabilities (pdf's) that are accessible in a case-control design since data are gathered conditional on disease status. Because of the above equality, **OR** is estimable using logistic regression with disease status as the response variable even though it is also the variable on which the data are stratified.

3.3.2. Pair-matched Case-Control Design

In a pair-matched case-control design a case, an individual with the disease, is matched with a control, an individual without the disease, according to some possibly confounding variables, such as age and gender. (For example, see section 2.2.4.) In this case, there is no information concerning the risk of disease associated with the exposure(s) for case-control pairs who have the same exposure status. Hence, we estimate **OR** by conditioning on case-control pairs with differing exposure status. This is done via conditional logistic regression.

Suppose that for the i^{th} case-control pair we have two vectors, $\mathbf{x}_{\text{case,i}}$ and $\mathbf{x}_{\text{control,i}}$, which may contain both *confounder* and *exposure* information. In particular, $\mathbf{x}_{\text{case,i}}$ and $\mathbf{x}_{\text{control,i}}$ may contain interactions between matching variables and confounder or exposure variables that were not used to match cases and controls. Further, suppose we have some target distribution for the exposure(s) so that for each case $\mathbf{F}_{0i} \equiv \mathbf{F}_0(\mathbf{x}_{\text{case,i}})$ is a vector, like $\mathbf{x}_{\text{case,i}}$, with the target exposure level(s) substituted for the individual's original exposure level(s). Then an estimator of **OR** for the case of the i^{th} pair, $\widehat{\mathbf{OR}}(\mathbf{x}_{\text{case,i}}, \mathbf{F}_{0i})$, can be found by

a) estimating $\hat{\beta}$ in a logistic regression model for the pairs as

$$\text{logit}(P(D = 1 | \mathbf{x}_{\text{case,i}}, \mathbf{x}_{\text{control,i}})) = \alpha + \beta' d_i ,$$

where $d_i \equiv \mathbf{x}_{\text{case,i}} - \mathbf{x}_{\text{control,i}}$, and each outcome is 1,

b) using $\hat{\beta}$ to get,

$$\widehat{\mathbf{OR}}(\mathbf{x}_i, \mathbf{F}_{0i}) = e^{\hat{\beta}'(\mathbf{x}_i - \mathbf{F}_{0i})} .$$

This is easily done in S-PLUS (Becker, Chambers, and Wilks, 1988) using the `glm` command or in SAS (SAS/STAT User's Guide, 1990) using `proc logistic`.

3.3.3. N_i:M_i Matched Sets Case-Control Design

In a matched sets case-control design, possibly several (N_i) cases are matched with possibly several (M_i) controls according to some possibly confounding variables. In this design, there is no information concerning the risk of disease associated with the exposure(s) for case-control sets in which all members have the same exposure status. Hence, we estimate **OR** by conditioning on case-control sets for which at least some cases have differing exposure status from some controls. As before, this is done via conditional logistic regression.

Suppose that the i^{th} of n matched sets contains N_i cases and M_i controls. In most such studies $N_i = 1$ so that each set contains a single case. Denote by $\mathbf{x}_{ij0}$ the covariate vector of exposures and

confounders for the cases, $j = 1, \ldots, N_i$, and $\mathbf{x}_{ij1}$ the covariate vector of exposures and confounders for the j^{th} control, $j = 1, \ldots, M_i$. Note that $\mathbf{x}_{ij0}$ and $\mathbf{x}_{ij1}$ may contain interactions between matching variables and confounder or exposure variables that were not used to match cases and controls. Further, suppose we have some target distribution for the exposure(s) so that for each case $\mathbf{F}_{ij0} \equiv \mathbf{F}(\mathbf{x}_{ij0})$ is a vector, like $\mathbf{x}_{ij0}$, with the target exposure level(s) substituted for the individual's original exposure level(s). An estimator of **OR** for the case of the i^{th} set, $\widehat{\mathbf{OR}}(\mathbf{x}_{ij0}, \mathbf{F}_{ij0})$, is obtained as follows. First, estimate β with $\hat{\beta}$ which maximizes the conditional likelihood function (Breslow and Day, 1980)

$$L(\beta) = \prod_{i=1}^{n} \frac{\sum_{j=1}^{N_i} \exp\left(\beta' \mathbf{x}_{ij0}\right)}{\sum_{j=1}^{N_i} \exp\left(\beta' \mathbf{x}_{ij0}\right) + \sum_{j=1}^{M_i} \exp\left(\beta' \mathbf{x}_{ij1}\right)} .$$

Then

$$\widehat{\mathbf{OR}}(\mathbf{x}_i, \mathbf{F}_{ij0}) = \mathrm{e}^{\left(\hat{\beta}' \mathbf{x}_{ij0} - \mathbf{F}_{ij0}\right)} .$$

Since this likelihood function is equivalent to that used in a stratified Cox proportional hazards model, where the matched sets are the strata, this is easily done in S-PLUS (Becker, Chambers, and Wilks, 1988) using `coxph`.

In SAS (SAS/STAT User's Guide, 1990) these estimates are computed using `proc phglm`.

3.4. Generalization of **AR**

Now that the definition and estimation of **RR** has been generalized using **OR** and logistic regression, it is necessary to generalize the definition of **AR** to accomodate these more complex and realistic structures.

The generalization discussed here is due to Bruzzi, et al (1985) and has been discussed by Benichou, et al (1991) in more detail.

Recall that for a dichotomous risk factor where D represents a case and $\overline{D}$ represents a control,

$$\begin{aligned} \mathbf{AR} &= P(F|D)\left(1 - \frac{1}{\mathbf{RR}}\right) \\ &= 1 - P\left(\overline{F}|D\right) - \frac{P(F|D)}{\mathbf{RR}} \\ &= 1 - \left(\frac{P(F_0|D)}{\mathbf{RR}_{00}} + \frac{P(F_1|D)}{\mathbf{RR}_{10}}\right) . \end{aligned}$$

where $\mathbf{RR}_{00} = 1$ and $\mathbf{RR}_{10} = \frac{\mathbf{P}(D \mid F=1)}{\mathbf{P}(D \mid F=0)}$. Using the notation of section 3.1.2 for a polychotomous risk factor, the above expression for $\mathbf{AR}$ can be generalized as

$$\mathbf{AR} = 1 - \sum_{i=0}^{K} \frac{P(F_i \mid D)}{\mathbf{RR}_{i0}} .$$

To generalize this slightly, suppose we have an exposure variable with $K+1$ levels denoted by $F_1, \ldots, F_K$, and a baseline level denoted by F_0. Further, suppose we have discrete-valued covariates denoted by $\mathbf{X} = (X_1, \ldots, X_m)$. Then we define the *attributable risk* by

$$\mathbf{AR} = 1 - \sum_{\mathbf{x}} \sum_{i=0}^{K} \frac{P(F_i, \mathbf{X} = \mathbf{x} \mid D)}{\mathbf{RR}_{i0 \mid \mathbf{x}}}$$

where the summations extend over all possible values of each covariate.

More generally, suppose we have any exposure variable (i.e. risk factor) denoted by $\mathbf{F}$. (Note that now $\mathbf{F}$ can be vector-valued, discrete, continuous, or some mixture.) Also, let $\mathbf{F}_0$ be the single baseline value to which all levels of the exposure will be referred. Further, suppose we have covariates (these too may be discrete, continuous, or a mixture) denoted by $\mathbf{X} = (X_1, \ldots, X_m)$. Then the *attributable risk* associated with the entire vector $\mathbf{F}$ is defined by

$$\mathbf{AR} = 1 - \int \frac{1}{\mathbf{RR}_{\mathbf{F}\mathbf{F}_0 \mid \mathbf{x}}} dF(\mathbf{F}, x \mid D) .$$

where F represents the conditional joint distribution function of the random variates $\mathbf{F}$ and $\mathbf{X}$ among the diseased. A semi-parametric estimator of $\mathbf{AR}$ is obtained by estimating $\mathbf{RR}$ by $\widehat{\mathbf{OR}}$ from the appropriate logistic regression and by estimating F with the empirical (conditional) distribution function, $\hat{F}$.

Most generally, instead of referring each exposure level to a single value, suppose we allow for an arbitrary target distribution (see section 5). Let $\mathbf{F}_0(\mathbf{F})$ be the targeted level of exposure for an individual currently exposed at level $\mathbf{F}$ and suppose the covariates are denoted by $\mathbf{X} = (X_1, \ldots, X_m)$ Then the *attributable risk* (this version has also been called the *general impact fraction*) is defined by

$$\mathbf{AR} = 1 - \int \frac{1}{\mathbf{RR}\left(\mathbf{F}_0(\mathbf{F}), \mathbf{F} \mid \mathbf{x}\right)} dF(\mathbf{F}, x \mid D) , \tag{19}$$

where F represents the conditional joint distribution function of the random variates $\mathbf{F}$ and $\mathbf{X}$ among the diseased and

$$\mathbf{RR}\left(\mathbf{F}_0(\mathbf{F}), \mathbf{F} \mid \mathbf{x}\right) = \mathbf{P}(D \mid \mathbf{F}, \mathbf{x}) / \mathbf{P}(D \mid \mathbf{F}_0(\mathbf{F}), \mathbf{x}) .$$

To obtain an estimate of $\mathbf{AR}$ let $\mathbf{F}_{0i} \equiv \mathbf{F}_0(\mathbf{F}_i)$ be the target exposure level for individual i who has an observed exposure level $\mathbf{F}_i$. Then an estimator of (19) is

$$\widehat{\mathbf{AR}} = 1 - \frac{1}{n_1} \sum_{i=1}^{n_1} e^{-\hat{\beta}(\mathbf{F}_i - \mathbf{F}_{0i})} = 1 - \frac{1}{n_1} \sum_{i=1}^{n_1} \frac{1}{\mathbf{OR}\left(\mathbf{F}_{0i}, \mathbf{F}_i ; \mathbf{x}_i , \hat{\beta}\right)} , \tag{20}$$

where n_1 is the number of cases, summation extends over all cases and $\hat{\beta}$ is a likelihood or conditional likelihood estimate of the parameter vector, β, in a logistic model as previously discussed. Note that when interest is in a specific component of $\mathbf{F}$ the remaining components may be treated as part of $\mathbf{X}$. In the case of multiple exposures β, $\mathbf{F}_i$ and $\mathbf{F}_{0i}$ are vector-valued. Also note that (18) allows one to include interactions between risk factors and confounding factors.

The heart of this technical report lies in calculating an estimate of the standard error of $\widehat{\mathbf{AR}}$. Benichou and Gail (1990) provide the most general expression for the variance of $\widehat{\mathbf{AR}}$, but it is only applicable in unmatched case-control designs with discrete exposures referred to a single-valued target exposure. Further, no one has been able to generate software to implement the variance expression given in Benichou and Gail, 1990. In section 3.5 we prove the asymptotic normality of $\widehat{\mathbf{AR}}$ for the general case of a vector-valued (possibly discrete, continuous, or mixed) exposure variable and an arbitrary target exposure distribution. In the proof, we rederive an asymptotic variance expression which clearly shows where the difficulty in the computations lie.

Due to the intractibility of computing the standard error of $\widehat{\mathbf{AR}}$, we resort to resampling methods, namely the bootstrap and jackknife. We find both methods to be reliable, accurate and reasonably efficient computationally. The jackknife is preferred to the bootstrap as it is typically much faster and nearly equivalent to the bootstrap in this problem, though care should always be taken in any application of interest to make sure the two methods give similar answers. The bootstrap is preferred in any case where the two methods give grossly dissimilar results. In the next section we review why this is the case, as well as the basic background for bootstrap and jackknife methods for estimating the standard error of $\widehat{\mathbf{AR}}$.

3.5. *Asymptotic Normality of* $\widehat{\mathbf{AR}}$

In this section, we prove that $\widehat{\mathbf{AR}}$ is asymptotically normal in both the unmatched and pair-matched cases. To do this, let the number of cases be n and the number of controls be m. In the unmatched case-control design assume that $0 < \lim_{m,n\to\infty} \frac{n}{m} < \infty$. Further, in the unmatched design, let $N \equiv n + m$ In the pair-matched design, let $N \equiv n$. Then N represents the sample size from which (conditional) logistic regression parameters are estimated.

Let F be the cumulative conditional distribution function of the exposures among the diseased, and $\hat{F}$ the empirical conditional distribution function of the exposures among the diseased, then

$$\sqrt{n}\left(\hat{F} - F\right) \xrightarrow{d} \text{Brownian Bridge process} \tag{21}$$

and, for any n, this process is bounded with mean zero.

Further, if $\hat{\beta} \in \mathbb{R}^k$ is the (conditional) logistic regression estimator yielding the odds ratio estimate of relative risk, then

$$\sqrt{N}\left(\hat{\beta} - \beta\right) \xrightarrow{d} \mathrm{N}_k\left(\mathbf{0}, \Sigma_\beta\right), \tag{22}$$

where Σ_β is the inverse of the information matrix at β. From this we have, for a fixed $\mathbf{x}$ and target exposure $\mathbf{b(x)}$,

$$\sqrt{N}\left(\mathrm{e}^{-\hat{\beta}(\mathbf{x}-\mathbf{b}(\mathbf{x}))} - \mathrm{e}^{-\beta(\mathbf{x}-\mathbf{b}(\mathbf{x}))}\right) \xrightarrow{d} \mathrm{N}\left(0, (\mathbf{x}-\mathbf{b}(\mathbf{x}))' \Sigma_\beta (\mathbf{x}-\mathbf{b}(\mathbf{x}))\right). \tag{23}$$

Also note that, under the assumption that $\mathbf{OR}$ and $\widehat{\mathbf{OR}}$ exceed 1 for any N, this process is bounded between $-\sqrt{N}$ and $\sqrt{N}$.

Theorem : $\sqrt{N}\left(\widehat{\mathbf{AR}} - \mathbf{AR}\right) \xrightarrow{d} \mathrm{N}(0, V)$, where V is the asymptotic variance of $\widehat{\mathbf{AR}}$ calculated in the manner of Benichou and Gail (1991).

To show this, rewrite

$$\sqrt{N}\left(\widehat{\mathbf{AR}} - \mathbf{AR}\right) = \sqrt{N}\int \left(\mathrm{e}^{-\hat{\beta}'(\mathbf{x}-\mathbf{b}(\mathbf{x}))} - \mathrm{e}^{-\beta'(\mathbf{x}-\mathbf{b}(\mathbf{x}))}\right) d\left(\hat{F} - F\right) \tag{24}$$

$$+ \sqrt{N}\int \mathrm{e}^{-\beta'(\mathbf{x}-\mathbf{b}(\mathbf{x}))}\, d\left(\hat{F} - F\right) \tag{25}$$

$$+ \sqrt{N}\int \left(\mathrm{e}^{-\hat{\beta}'(\mathbf{x}-\mathbf{b}(\mathbf{x}))} - \mathrm{e}^{-\beta'(\mathbf{x}-\mathbf{b}(\mathbf{x}))}\right) dF \tag{26}$$

From (21) and (23) we see that (24) converges in probability to 0 since the integrand in (24) is going to a mean 0, finite variance process at a rate of $\sqrt{N}$ and the measure is going to a mean 0, finite variance process at a rate of $\sqrt{n}$. Thus, a multiplier of $\sqrt{Nn}$ is needed to obtain a nondegenerate limiting distribution. So, we can ignore the first term, (24), for large N.

The second term, (25), also has an asymptotic normal distribution. This follows from (21) and the fact that $0 < \lim_{m,n\to\infty} \frac{N}{n} = K < \infty$. Then

$$\sqrt{\frac{N}{n}}\int \mathrm{e}^{-\beta'(\mathbf{x}-\mathbf{b}(\mathbf{x}))}\, d\left(\sqrt{n}\left(\hat{F} - F\right)\right) \xrightarrow{d} \mathrm{N}\left(0, K\,\mathbf{Var}\left(\int \mathrm{e}^{-\beta'(\mathbf{x}-\mathbf{b}(\mathbf{x}))}\, dW^o(\mathbf{x})\right)\right),$$

where $\lim_{m,n\to\infty} \frac{N}{n} = K$ and W^o represents a Brownian Bridge process.

The variance term above is simply the variability of $1/\mathbf{RR}(\mathbf{x})$ among the diseased individuals. Note that as in section 3 of Benichou and Gail (1990), this is easily estimated by

$$\frac{N}{n(n-1)}\left\{\sum_{i=1}^{n} \mathrm{e}^{-2\hat{\beta}'(\mathbf{x}_i-\mathbf{b}(\mathbf{x}_i))} - n^{-1}\left(\sum_{i=1}^{n} \mathrm{e}^{-\hat{\beta}'(\mathbf{x}_i-\mathbf{b}(\mathbf{x}_i))}\right)^2\right\}.$$

Finally, consider the limit of the third term, (26). The integrands in (26) are uniformly integrable since we restrict $\mathbf{RR} \geq 1 \Leftrightarrow 0 < 1/\mathbf{RR} \leq 1$ which implies that the integrands,

$\sqrt{N}\left(e^{-\hat{\beta}'(\mathbf{x}-\mathbf{b}(\mathbf{x}))} - e^{-\beta'(\mathbf{x}-\mathbf{b}(\mathbf{x}))}\right)$ have as their support $\left(-\sqrt{N}, \sqrt{N}\right)$. As N gets large the tails of the distribution of these integrands decay exponentially which implies that the integrands are uniformly integrable. So, the limit can be taken inside the integral yielding

$$(26) \to \mathrm{E}\left(Z\sqrt{(\mathbf{x}-\mathbf{b}(\mathbf{x}))'\,\Sigma_\beta\,(\mathbf{x}-\mathbf{b}(\mathbf{x}))} \mid \mathrm{D}\right)$$

where Z is a standard normal random variable independent of D. Hence, the third term, (26), has asymptotic mean 0 and asymptotic variance $W = \mathrm{E}\left((\mathbf{x}-\mathbf{b}(\mathbf{x}))'\,\Sigma_\beta\,(\mathbf{x}-\mathbf{b}(\mathbf{x})) \mid \mathrm{D}\right)$, which is easily estimated by

$$\frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i-\mathbf{b}(\mathbf{x}_i))'\,\hat{\Sigma}_{\hat{\beta}}\,(\mathbf{x}_i-\mathbf{b}(\mathbf{x}_i))\ ,$$

where the summation extends over all cases.

Together these give us that $\sqrt{N}\left(\widehat{\mathbf{AR}}-\mathbf{AR}\right)$ is asymptotically normal with mean 0 and variance given by $W + K\,\mathbf{Var}\left(\int e^{-\beta'(\mathbf{x}-\mathbf{b}(\mathbf{x}))}\,dW^o(\mathbf{x})\right) + 2\,\mathrm{Cov}\left((25),(26)\right)$. The covariance term can be estimated using a delta method for implicitly defined random variables as described in Benichou and Gail, 1989. However, it is exactly this covariance calculation that makes these computations so difficult to implement efficiently.

For an unmatched design with a single-valued target distribution, this asymptotic variance calculation coincides exactly with that given in Benichou and Gail (1990). In any setting other than the most restrictive, unrealistic cases, calculation of the asymptotic variance is too complicated for generally applicable, swift running implementation, due almost exclusively to the covariance term. Furthermore, Benichou and Gail's delta method approach is approximately the infinitesimal jackknife method for calculating variance (Efron and Tibshirani, 1993). In the case that the estimator is linear, as is nearly the case for **AR**, the infinitesimal jackknife, jackknife and bootstrap yield the same variance estimates. In the vast majority of cases, the jackknife is the most efficient computationally.

4. The bootstrap and jackknife estimates of variance

The ideal, frequentist notion of the variance (or the sampling distribution) of an estimator arises from imagining that the experiment was repeated infinitely, observing a value of the estimator from each "experiment". The distribution of all these hypothetical values is, by definition, the sampling distribution of the estimator. In many cases this yields tools by which one could find the variance, or confidence intervals, etc. for the corresponding parameter. In cases where the sampling distribution is either incalculable or intractable other techniques must be used for finding the variance (or distribution) of an estimator. Below we outline several techniques for obtaining the variance (or distribution) of an estimator.

1) *Theory.*

a) For some models, one can find an expression for the exact variance (sampling distribution) of the estimator. That is, under the model assumptions we can describe mathematically what would result if we repeated the experiment infinitely.

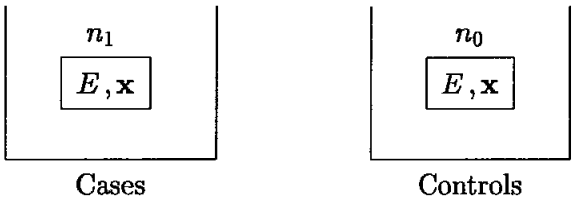
b) In many settings, one can derive an asymptotic approximation to the variance or sampling distribution (e.g. delta method, central limit theorem).

2) *Resample.*

Use the observed data to "repeat" the experiment "infinitely". This includes the bootstrap, jackknife and the Bayesian bootstrap.

4.1. How does the bootstrap work in the unmatched case-control designs?

In this case, the data obtained for each subject are level of exposure (denote this by E), data on confounders/stratifiers, say $\mathbf{x}$, and knowledge of the disease status (i.e. case or control), denote this by D. As in Freedman, et al, 1991, consider having written this information down on a slip of paper, one slip per subject. Then the design and data are described by the following "box model":



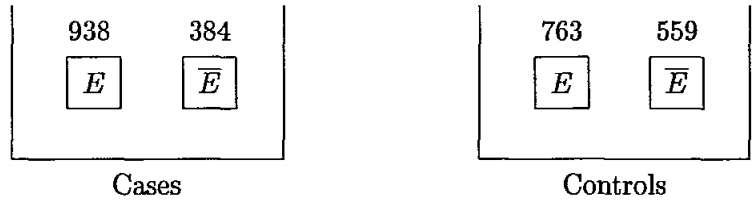
The diagram indicates that there are two boxes; one for "Cases" containing n_1 tickets, one for each of the cases, and another box for "Controls" containing n_0 tickets, one for each of the controls. From these observed data we can calculate $\widehat{\mathbf{AR}}$.

To carry out the bootstrap for estimating the standard error of $\widehat{\mathbf{AR}}$, draw a sample *with replacement* of size n_1 from the *Cases* box and a sample *with replacement* of size n_0 from the *Controls* box. This gives a "new" data set with n_1 cases and n_0 controls. From this new data set we calculate $\widehat{\mathbf{AR}}$. Repeating this procedure (resampling from the original data to obtain "new" data sets with n_1 cases and n_0 controls) yields a sequence of $\widehat{\mathbf{AR}}$'s. We obtain an approximation to the sampling distribution of $\widehat{\mathbf{AR}}$ through this sequence.

For example, suppose the observed data are:

	D	$\overline{D}$
E	938	763
$\overline{E}$	384	559
	1322	1322

Then $\widehat{\mathbf{AR}} \approx .313$ and the boxes full of tickets look like



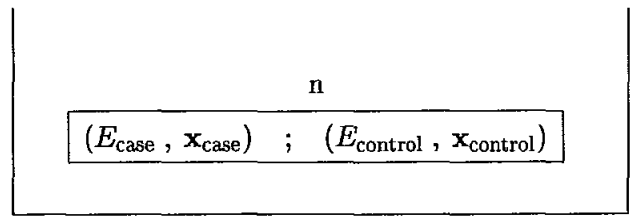
Draw a sample with replacement of size 1322 from the *Cases* box and a sample with replacement of size 1322 from the *Controls* box and calculate $\widehat{\mathbf{AR}}$. For example, we might get

	D	$\overline{D}$
E	963	733
$\overline{E}$	359	589
	1322	1322

with $\widehat{\mathbf{AR}} \approx .390$. Carrying out this resampling repeatedly yields an approximation to numerous realizations from the sampling distribution of $\widehat{\mathbf{AR}}$, say $\widehat{\mathbf{AR}}^{(1)}, \ldots, \widehat{\mathbf{AR}}^{(B)}$ (e g. in our example $\widehat{\mathbf{AR}}^{(1)} = .390$) and an estimate of the standard error of $\widehat{\mathbf{AR}}$ is the standard deviation of $\widehat{\mathbf{AR}}^{(1)}, \ldots, \widehat{\mathbf{AR}}^{(B)}$, namely $\sqrt{\frac{1}{B-1}\sum_{b=1}^{B}\left(\widehat{\mathbf{AR}}^{(b)} - \overline{\widehat{\mathbf{AR}}}\right)^2}$, where $\overline{\widehat{\mathbf{AR}}}$ is the mean of $\widehat{\mathbf{AR}}^{(1)}, \ldots, \widehat{\mathbf{AR}}^{(B)}$.

4.2. How does the bootstrap work in pair-matched case-control designs?

In the pair-matched case-control design, simply resample the pairs with replacement. That is, for the pair-matched design the box model will look like



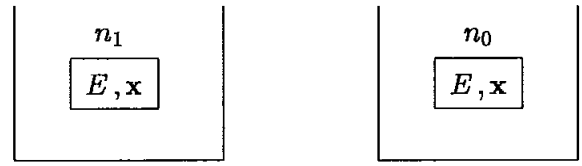
There is one ticket for each of the n case-control pairs. Each ticket has the exposure, E, and confounder/stratifier, $\mathbf{x}$, for the case and for the control in that pair. The bootstrap procedure then draws a sample of size n with replacement from this box obtaining an estimate of $\mathbf{AR}$. As before, repeat this to obtain an approximation to the sampling distribution of $\widehat{\mathbf{AR}}$.

4.3. How does the bootstrap work in matched sets case-control designs?

In the matched sets case-control design, simply resample the sets with replacement. That is, for the matched sets design the box model will contain one ticket for each of the n case-control sets. Each ticket has the exposure, E, and confounder/stratifier, $\mathbf{x}$, for the cases and for the controls in that set. The bootstrap procedure then draws a sample of size n with replacement from this box obtaining an estimate of $\mathbf{AR}$. As before, repeat this to obtain an approximation to the sampling distribution of $\widehat{\mathbf{AR}}$.

4.4. How does the jackknife work in unmatched case-control designs?

As in section 4.1, suppose that we have gathered data according to an unmatched case-control design and that we represent this by the following box model:



where E represents the exposure status of the subject and $\mathbf{x}$ represents confounder/stratifier information.

The i^{th} *jackknife* value of the estimator ($\widehat{\mathbf{AR}}$) is obtained by deleting (i.e. temporarily removing from the data set) the i^{th} subject and calculating the estimator, say $\widehat{\mathbf{AR}}^{(-i)}$. Doing this for each i yields the jackknife distribution of the estimator.

For example, suppose the observed data are:

	D	$\overline{D}$	
E	938	763	
$\overline{E}$	384	559	
	1322	1322	

$\widehat{\mathbf{AR}} \approx .313$. Then one jackknife value for $\widehat{\mathbf{AR}}$ (the one obtained when deleting a patient who is both exposed and diseased) is obtained by estimating $\widehat{\mathbf{AR}}$ from the table

	D	$\overline{D}$	
E	937	763	
$\overline{E}$	384	559	
	1321	1322	

$\widehat{\mathbf{AR}}^{(-1)} \approx .312539$.

There will be 938 of these, one for each exposed case. Similarly, we get 763 jackknife values of 0.3135787 corresponding to the deletion of an exposed control, 384 jackknife values of 0.3143293 and 559 jackknife values of 0.3123485. From this we get a an estimated standard error of 0.03689. using the formula for computing the standard error from a jackknife sample found in section 4.7.

4.5. How does the jackknife work in a pair-matched case-control design?

In a pair-matched case-control design we get $\widehat{\mathbf{AR}}^{(-i)}$ by deleting the i^{th} pair and computing $\widehat{\mathbf{AR}}$ from the remaining $n-1$ pairs. This yields $\widehat{\mathbf{AR}}^{(-1)}, \ldots, \widehat{\mathbf{AR}}^{(-n)}$ from which we can compute the jackknife standard error estimate of $\widehat{\mathbf{AR}}$.

4.6. How does the jackknife work in a matched set case-control design?

In a matched set case-control design we get $\widehat{\mathbf{AR}}^{(-i)}$ by deleting the i^{th} set and computing $\widehat{\mathbf{AR}}$ from the remaining $n-1$ sets. This yields $\widehat{\mathbf{AR}}^{(-1)}, \ldots, \widehat{\mathbf{AR}}^{(-n)}$ from which we compute the jackknife standard error estimate of $\widehat{\mathbf{AR}}$.

4.7. Standard Results for Bootstrap and Jackknife Estimates of Standard Error

In general, if $\hat{\theta}^{(-i)}$ is the i^{th} jackknife value of some estimator, $\hat{\theta}$, where $i = 1, \ldots, n$, then the jackknife estimate of variance is given by

$$\mathbf{Var}_{JK}(\hat{\theta}) = \frac{n-1}{n} \sum_{i=1}^{n} \left(\hat{\theta}^{(-i)} - \bar{\hat{\theta}} \right)^2$$

where $\bar{\hat{\theta}}$ is the mean of $\hat{\theta}^{(-1)}, \ldots, \hat{\theta}^{(-n)}$, Efron, 1987.

Result: (Efron, 1987) (Relationship between jackknife and bootstrap estimates of variance.)

If $\hat{\theta}$ is a *linear* statistic, then

$$\mathbf{Var}_{BOOT}(\hat{\theta}) = \left(\frac{n-1}{n} \right) \mathbf{Var}_{JK}(\hat{\theta}) \, .$$

The simplest and most natural definition of a *linear* statistic is that $\hat{\theta}$ can be written as

$$\gamma + \frac{1}{n}\sum_{i=1}^{n}\alpha(x_i)\ .$$

For example,

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n}x_i$$

and

$$7 + \frac{1}{n}\sum_{i=1}^{n}(x_i)^2$$

are linear statistics.

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

is not a linear statistic since $(x_i - \overline{x})^2$ *cannot* be written as $\alpha(x_i)$, a function of a single one of the x's.

We address the issue of linearity for $\widehat{\mathbf{AR}}$ in the next section.

4.7.1. Problems with the jackknife.

1) *Non-linearity*

The bootstrap accounts for the curvature in a non-linear statistic while the jackknife corresponds to a specific linear approximation of the statistic. Unless the non-linearity of the statistic is drastic, there is typically not a big discrepancy between the two estimates of variance.

2) *Lack of smoothness*

If the statistic is not smooth (e.g. discontinuous, nondifferentiable) then the jackknife estimate of variance can breakdown entirely. For example, because the median changes discontinuously as you change a single data point, a linear approximation to the median cannot be expected to work well, if at all.

Claim :

$$\widehat{\mathbf{AR}} = 1 - \frac{1}{n}\sum_{j=1}^{n}\frac{1}{\widehat{\mathbf{OR}}(\mathbf{x}_j\ ,\ \mathbf{e}_j\ ;\ \hat{\beta})}\ ,$$

is smooth, bounded and "almost" linear.

The smoothness results from the fact that $\widehat{\mathbf{AR}}$ is a simple function of the average of the reciprocals of the odds ratios. The non-linearity results from the fact that $\hat{\beta}$ is a function of all the $\mathbf{x}_j$'s. Along the

lines of Efron and Tibshirani (1992), we are currently working on a diagnostic for automatically measuring and graphically diagnosing the extent of this nonlinearity in any of the designs discussed in this report.

We have tested these techniques on small and large data sets from unmatched, matched-pair and matched-set designs, simple and complicated exposure structure, and large and small $\widehat{\mathbf{AR}}$. We conjecture that the cases in which the jackknife estimate of standard error for $\widehat{\mathbf{AR}}$ breaks down are pathological enough as to be of little practical consequence.

Practically, we suggest analyses begin by exploring models and obtaining interval estimates of $\widehat{\mathbf{AR}}$ using just the jackknife, not the bootstrap, since the jackknife is faster, especially in problems with many subjects and/or many logistic regression parameters. However, any final inferences and interpretations should be withheld until the results from a bootstrap are obtained and compared with results using the jackknife. If the results from the jackknife and the bootstrap disagree substantially, thought must be given to the data and the model.

For example, in model 15 of Benichou (1991) it is noted that no standard error estimate can be obtained since the model is saturated and the observed information matrix is non-invertible at the MLE. Our results (Appendix) found that the jackknife gives an estimate for the standard error of $\widehat{\mathbf{AR}}$ in this model, but the bootstrap fails due to the sparsity of the corresponding data matrix. Some of the bootstrap samples yielded no cases in high-risk categories which leads to a bootstrap estimate that is negative (i.e. $\widehat{\mathbf{AR}} < 0$) indicating a protective effect for the exposure variable's presumed high-risk categories. This happened in both of the saturated models (models 10 and 15) since the data are simply spread too thin. However, in these cases if we simply focus on all the non-negative bootstrap estimates the results are quite close to those using the jackknife.

We now consider a number of examples, including brief descriptions of S-plus software for carrying out these procedures.

5. Examples using the software, arhat

In this section we carry out some of the computations described previously by way of an S-PLUS program, `arhat`. The syntax of `arhat` is similar to the modelling functions in S-PLUS such as `lm`, `glm`, `coxph`, etc. The unique aspect of the concept of attributable risk, and hence, `arhat`, is that there is a clear distinction between explanatory variables considered to be *exposure* variables and explanatory variables that are are not considered *exposure* variables. For `arhat`, this distinction is made in the formula of the call to `arhat` by including exposure variables in the special function `expos`. A simple example is

```
arhat( stroke ~ expos(hbp) + age)
```

which tells the program to calculate the estimated attributable risk of the exposure variable hbp on the outcome variable stroke, adjusting for the confounding variable age.

What follows are more detailed explanations of this and numerous other features of arhat in the context of several examples.

By convention, we will use > as the S prompt, and + as the S continuation-of-a-line prompt, so that any text following a > or a + is typed by the user. Any other text is output from the issued S command. A detailed help section pertaining to arhat is given in appendix B.

5.1. Dichotomous exposures with no confounders

5.1.1. An unmatched case-control design

These are the data from *Whisnant, et al* 1996 (see section 2.3.3) concerning cerebral infarction (CI) and high blood pressure (HBP) treated as an unmatched design.

	CI	$\overline{CI}$
HBP	938	763
$\overline{HBP}$	384	559
	1322	1322

As seen earlier (section 2.3.3)

$$\widehat{\mathbf{OR}} = (938 * 559) / (763 * 384) \approx 1.79 \approx e^{.582} \quad , \quad \widehat{\mathbf{AR}} \approx .313 \ , \ \mathbf{SE}\left(\widehat{\mathbf{AR}}\right) = .037 \, .$$

The following are the S commands for analyzing these data with the arhat program we have developed. The syntax of the arhat command mimics that of the glm command for generalized linear models, most specifically for logistic regression. As with glm one needs to specify a *formula*, but with arhat *at least one of the independent variables must be specified as an exposure (risk factor) variable.* This is done using the special function expos.

In the following example a dichotomous risk factor, high blood pressure (hbp), is analyzed in an unmatched case-control design. The variable cases indicates whether the subject had a cerebral infarction (CI) (cases = 1) or not (cases = 0) and hbp is a categorical variable (i.e. factor) with levels HBP-0 and HBP-1 that indicate whether the subject had high blood pressure (hbp = HBP-1) or not (hbp = HBP-0). hbp is considered the exposure variable and is a categorical data type. Further, we assume that a data frame called chapter.dat exists in which the relevant variables, in this example cases and hbp, and possibly other variables reside.

```
> example1 <- arhat(cases ~ expos(hbp) , data=chapter.dat)
```

Next is a brief printout of the results from fitting this model in which we simply ask for the estimated attributable risk. We explain how to get the SE for $\widehat{\mathrm{AR}}$ in this example later.

```
> example1
Call:
arhat(formula = cases ~ expos(hbp), data  =  chapter.dat)

Estimate of AR = 0.313 .

Coefficients:
 (Intercept)       hbp
  -0.3755052 0.5819971
```

The `Coefficients` are the estimated logistic regression parameters; $\hat{\beta} = 0.5819971$.

In this case, `chapter.dat`, is a data frame with 2644 rows, the first and last of which look like

```
> chapter.dat[c(1,2644),]
```

	hbp	dm	ihd	cases	match.id
1	HBP-0	DM-0	IHD-0	0	1
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
2644	HBP-1	DM-1	IHD-1	1	1322

Note that this data frame contains lists (variables) named `hbp`, `dm`, `ihd`, `cases` and `match.id`. The original table describing the relationship between `hbp` and `cases` is given by

```
> table(chapter.dat$hbp,chapter.dat$cases)
```

	0	1
HBP-0	559	384
HBP-1	763	938

The logistic regression coefficients are given by

```
> example1$coef
Coefficients:
```

```
  (Intercept)          hbp
   -0.3755052    0.5819971
```

and the table of relative risks is given by

```
> table(example1$rel.risk)

  1   1.7896
943     1701
```

This indicates that the 943 individuals with low blood (HBP-0) pressure have a relative risk of 1 and the 1701 individuals with high blood pressure (HBP-1) have an estimated odds ratio of $1.7896 = e^{0.5819971}$

Using the above data set, we now show how to obtain standard error estimates for $\widehat{\mathbf{AR}}$. The bootstrap, using B resamples, and jackknife estimates of standard error are obtained by including the following options `B = B` and `jackknife =T`. In this example we use `B = 1000` resamples.

```
> example1 <- arhat(cases ~ expos(hbp) , data=chapter.dat ,
                    B = 1000 , jackknife = T)
```

The simplest output from this is

```
> example1
Call:
arhat(formula = cases ~ expos(hbp), data  =  chapter.dat,
      B = 1000 , jackknife = T)

Estimate of AR = 0.313 .
The mean of the jackknifed AR's = 0.313 .
The standard error of the jackknifed AR's = 0.037 .
The mean of the bootstrapped AR's = 0.312 .
The standard error of the bootstrapped AR's = 0.036 .

Coefficients:
 (Intercept)       hbp
  -0.3755052 0.5819971
```

Note that we report the mean of the jackknife and the bootstrap "samples" only for diagnostic purposes. They should not be used as estimates of $\widehat{\mathbf{AR}}$.

A host of other options are detailed in appendix B. These include options for calculation of jackknife, percentile, bias-corrected and accelerated bias-corrected bootstrap confidence intervals (Efron and Tibshirani 1992), as well as diagnostic information concerning both the bootstrap and the jackknife resamples.

The confidence intervals are obtained simply with the summary command in S.

```
> summary(example1)
Call: arhat(formula = cases ~ expos(hbp), data = chapter.dat, B = 1000,
jackknife = T)

Estimate of AR = 0.3130583 .

Jackknife CI                              ( 0.241 , 0.385 )
Percentile Bootstrap CI                   ( 0.241 , 0.380 )
Percentile.t Bootstrap CI                 ( 0.243 , 0.383 )
Bias-corrected Bootstrap CI               ( 0.241 , 0.380 )
Accelerated Bias-corrected Bootstrap CI ( 0.239 , 0.378 )
```

Even in small-sample simulations we have seen little difference among these confidence intervals. Typically, the jackknife interval is quickest to calculate and we suggest its use for exploratory purposes. Nonetheless, final results should be calculated using each of these intervals. Substantial differences indicate skewness, bias, or non-linearity in the problem that ought to be investigated. For more on this see Efron, 1987.

5.1.2. A pair-matched case-control design

In this section we use `arhat` to obtain an estimate of attibutable risk and its standard error in a matched case-control design. We use the same data set as above, but we now make use of the fact that the 2644 subjects were matched pairs of cases and controls (see section 2.3.4). The pairs were matched based on age and gender. In this case, it is necessary to have a variable which indicates which case(control) is matched to which control(case). In `chapter.dat` the variable `match.id` is a matching index. For example, since rows 1 and 1323 both have `match.id = 1`, these are a matched pair.

```
> example1a <-
+    arhat(cases ~ strata(match.id) + expos(hbp),
```

```
+          data = chapter.dat , B = 1000 , jackknife =T)

Estimate of AR = 0.331 .
The mean of the jackknifed AR's = 0.331 .
The standard error of the jackknifed AR's  = 0.037 .
The mean of the  1000  bootstrapped AR's = 0.328 .
The standard deviation of the  1000  bootstrapped AR's = 0.037 .
```

Note that the expression `strata(match.id)` is used in the formula of `arhat` to tell the function that a matched case-control design is to be used and how the subjects are to be matched. This is done on the "explanatory" side of the formula (i.e. the right side of the ~) using the special function `strata`. For example, the formula `cases ~ expos(hbp) + strata(match.id)` indicates that the attributable risk of `hbp` on `cases` is to be estimated in a matched case-control design where `match.id` is the variable that indexes subjects to be matched. The `strata(match.id)` term must be *added* to the remaining formula (i.e. it must look like a simple linear term) though its position in the formula is arbitrary so that `cases ~ strata(match.id) + expos(hbp)` would yield the same results as before. However, the following formula **is not valid**; `cases ~ expos(hbp) * strata(match.id)`.

The unmatched analysis of the same data resulted in an **AR** estimate of 0.313 with a standard error of approximately 0.037, not appreciably different from an estimated **AR** of 0.331 with standard error of 0 037 obtained by this, more proper, analysis.

5.1.3. A matched-set case-control design

The last example in this section is from a matched-set case-control design. We revisit the study concerning risk factors for temporal arteritis (Machado, et al 1989) in which there were four controls matched to each case, where a case is a patient diagnosed with temporal arteritis. The risk factors considered in this example are whether the patient ever smoked (yes/no) and whether the patient had a history of angina (yes/no). We consider history of angina to be a confounding variable and the patient's smoking status to be the exposure variable. In this case, the sets were matched based on age and gender. As in the pair-matched design, it is necessary to have a variable which indicates which cases and controls are matched. In `arteritis.dat` the variable `set` is a matching index. For example, since rows $1, 2, 3, 4$ and 5 all have `set = 1`, these form a matched set.

```
ex.arteritis <- arhat(case ~ expos(evrsmk) + angina + strata(set), data = arteritis,
                      jackknife = T)
```

```
Call: arhat(formula = case ~ expos(evrsmk) + angina + strata(set), data = arteritis,
 jackknife = T, coxph = T)

Estimate of AR = 0.235 .
The mean of the jackknifed AR's = 0.235 .
The standard error of the jackknifed AR's = 0.056 .
```

We reiterate that the expression `strata(set)` is used in the formula of `arhat` to tell the function that a matched case-control design is to be used and how the subjects are to be matched. This is done on the "explanatory" side of the formula (i.e. the right side of the ~) using the special function `strata`.

This analysis suggests that after adjusting for history of angina, and assuming we could adjust the risk of temporal arteritis among smokers to that of non-smokers, we could expect a 25% reduction in temporal arteritis incidence, with a standard error of approximately 5.6 percentage points.

5.2. More complicated target distributions

In the next two sections we provide examples with more general data structure, as well as complicated and realistic target distributions.

The two examples come from the same study of risk factors for *cerebral infarctions* (i.e *stroke*) described in sections 2.3.3 and 2.3.4. In these examples we consider the following covariates for each patient; age at study entry (`age`), *diastolic blood pressure* (`diastolic`) and *smoking level* (`smoke`). The age of the individual is accounted for since older individuals are more likely to suffer a stroke, though **age** is a confounder and not thought of as an exposure variable in that we can not imagine any prevention for getting older. Consider each of the other risk factors as exposure variables. In this case, if both `smoke` and `diastolic` are dichotomous then the natural target exposure level is the absence of both risk factors. More realistically, if we describe smoking status with 5 levels, *Current, Former, Never , Unknown, Uncertain* and blood pressure by the last measured diastolic blood pressure, which is considered as a continuous exposure, then there is no natural target distribution for describing, hypothetically, the idealized distribution of the risk factors among exposed individuals. We describe two possible analyses.

5.2.1. No one ever smokes and everyone lowers DBP by 10%

In the first analysis we describe the use of `arhat` in an example in which the reference distribution is that "no one ever smoked" and "everyone lowers blood pressure by 10%". We consider **diastolic**

blood pressure as a continuous exposure, **smoke** as a polychotomous exposure with 5 levels, "Never","Current","Former","Uncertain","Unknown", and we adjust for the confounding variable **age**.

There are several new components in this model. First, note that we have both discrete and continuous exposure variables. Second, note that the reference distribution is not a single value/category for each individual but, instead, the target exposure is determined by that individual's observed exposure level. Finally, note that we obtain a multivariate estimate of **AR** based on two exposure variables.

The reference distribution is described by

reduce observed **diastolic** by 10% ;
put everyone into the "Never" smoked category.

In this case, we call arhat by

```
> example2 <-
+ arhat(formula = cases ~ age + expos(smoke) + expos(diastolic),
+        data = stroke.dat, B = 1000, jackknife = F,
+        categorical = F, baseline = stroke.target)

> example2
Call:
arhat(formula = cases ~ age + expos(smoke) + expos(diastolic),
      data = stroke.dat, B = 1000, jackknife = F,
      categorical = F, baseline = stroke.target)

Estimate of AR = 0.428 .
The mean of the  1000  bootstrapped attributable risks = 0.428 .
The standard deviation of the  1000  bootstrapped attributable risks = 0.037 .

Coefficients:
 (Intercept)         age smokeCurrent smokeFormer smokeUncertain smokeUnknown
   -1.925598 0.00837275    0.9876965   0.3684065       1.497466     1.019093
  diastolic
 0.01195578
```

One new component in this invocation of arhat is baseline = stroke.target which is itself a data frame containing the targeted values for each of the exposure variables from the original data frame. In this example, stroke.target contains two variables smoke and diastolic. In stroke.target, smoke looks like the variable smoke in stroke.dat except that every individual has a value of "Never". This is easily created with

```
> smoke.target <- stroke.dat$smoke
> smoke.target[smoke.target != "Never"] <- "Never"
```

Similarly, in stroke.target, diastolic looks like the variable diastolic in stroke.dat except that each value is 90% of the observed value. This is created as follows

```
> DBP.target <- .90*stroke.dat$diastolic
```

Once these variables are constructed, stroke.target (with the variables named diastolic and smoke) is created with a simple call to data.frame;

```
> stroke.target <- data.frame(diastolic=DBP.target , smoke=smoke.target)
```

Also note that the categorical = F option is passed to arhat since we are considering diastolic to be a continuous-valued exposure.

Finally, in this case the estimate of **AR** is $\widehat{\mathbf{AR}} \approx 0.43$ with a standard error of about 0.038. Hence, an approximate 95% confidence interval for **AR** is $(.354, .506)$. This suggests that if the entire population had never started smoking and had 10% lower blood pressure than they do currently, there would be between 35% and 51% fewer strokes.

5.2.2. Current smokers quit and higher DBP implies greater DBP reduction

Finally, we briefly describe what is necessary for using arhat in another example. In this example the reference distribution is that "Current smokers quit" and "blood pressure is lowered according to current level". We consider **diastolic blood pressure** as a continuous exposure, **smoke** as a polychotomous exposure with 5 levels, "Never","Current","Former","Uncertain","Unknown", and we adjust for the confounding variable *age*

The reference distribution is described by

if **diastolic** > 150 , reduce **diastolic** by 25% ;
if $120 <$ **diastolic** ≤ 150 , reduce **diastolic** by 15% ;
if $100 <$ **diastolic** ≤ 120 , reduce **diastolic** by 10% ;
if $85 <$ **diastolic** ≤ 100 , reduce **diastolic** by 5% ;
if **diastolic** ≤ 85 , leave **diastolic** as is ,
suppose all "Current" smokers become "Former" smokers.

In this case, we call arhat by

```
> example3 <-
+   arhat(cases ~ age + expos(smoke) + expos(diastolic) ,
+         data = stroke.dat , categorical = F ,
+         baseline = stroke.target3 , B = 1000 ,
+         jackknife = T)

> example3
Call:
arhat(cases ~ age + expos(smoke) + expos(diastolic) ,
data = stroke.dat , categorical = F , baseline = stroke.target3 ,
B = 1000 , jackknife = T)

Estimate of AR = 0.117 .
The mean of the jackknifed AR's = 0.117 .
The standard deviation of the jackknifed AR's = 0.02 .
The mean of the  1000  bootstrapped attributable risks = 0.117 .
The standard deviation of the  1000  bootstrapped attributable risks = 0.02 .

Coefficients:
 (Intercept)         age smokeCurrent smokeFormer smokeUncertain smokeUnknown
   -1.925598 0.00837275    0.9876965   0.3684065       1.497466 1.019093

  diastolic
 0.01195578
```

`stroke.target3` is itself a data frame containing the targeted values for each of the exposure variables from the original data frame. In this example, `stroke.target3` contains two variables

smoke and diastolic. In stroke.target3, smoke looks like the variable smoke in stroke.dat except that where the original value was "Current", the stroke.target3 value of smoke is "Former". This is easily created with

```
> smoke.target3 <- stroke.dat$smoke
> smoke.target3[smoke.target3 == "Current"] <- "Former"
```

Similarly, in stroke.target3, diastolic looks like the variable diastolic in stroke.dat except that where the original value was above 150 the stroke.target3 value of diastolic is reduced by 25%, and so on. This is created as follows

```
> DBP.target3 <- ifelse(stroke.dat$diastolic > 150,
+                       .75*stroke.dat$diastolic , stroke.dat$diastolic)
> DBP.target3 <-
+      ifelse(stroke.dat$diastolic > 120 & stroke.dat$diastolic <= 150,
+             .85*stroke.dat$diastolic , DBP.target3)
> DBP.target3 <-
+      ifelse(stroke.dat$diastolic > 100 & stroke.dat$diastolic <= 120,
+             .90*stroke.dat$diastolic , DBP.target3)
> DBP.target3 <-
+      ifelse(stroke.dat$diastolic > 85 & stroke.dat$diastolic <= 100,
+             .95*stroke.dat$diastolic , DBP.target3)
```

Once these variables are constructed, stroke.target3 (with the variables named diastolic and smoke) is created with a simple call to data.frame;

```
> stroke.target3 <- data.frame(diastolic=DBP.target3 , smoke=smoke.target3)
```

In this example we see that even after adjusting for age, if we could get all current smokers to quit and all individuals with high diastolic blood pressure to reduce their diastolic blood pressure as described above, a 95% confidence interval for **AR** implies that we would expect there to be between 7.5% and 15.5% fewer strokes.

Unsurprisingly, the substantively different target distributions yield substantially different estimates of the attributable risk for stroke, 0.43 compared with 0.12.

In the first example one might argue that this provides a hypothetical, limiting prevalence of stroke as if noone ever started smoking. The second is an attempt to understand the impact on prevalence

of stroke from an impossibly successful campaign for convincing people to stop smoking and to reduce their blood pressure by an amount commensurate with their current hypertensive trouble.

For our purposes, the point to be made here is that the methods described in this report allow either unmatched or matched study designs, general modelling procedures and flexible target exposure levels to be used for estimating attributable risk.

6. Conclusion

Benichou and Gail (1990) state that "Although the theory for inference on the attributable risk has been presented in considerable detail, it remains to develop easily used computer programs to implement these methods." Though we have not implemented the delta-method variances presented by Benichou and Gail, we have developed software for drawing inference concerning attributable risk in a greater variety of settings with equivalent standard errors. Further, the delta-method approach taken by Benichou and Gail is very nearly the infinitesimal jackknife approximation for estimating the variance of $\mathbf{AR}$. The infinitesimal jackknife is based on a slightly different linear approximation of the $\widehat{\mathbf{AR}}$, than the jackknife. Since $\widehat{\mathbf{AR}}$ is nearly linear, all three methods, the jackknife, infinitesimal jackknife and the bootstrap give nearly equivalent standard errors. The only reason for implementing Benichou and Gail's delta-method calculations would be to increase the speed of the calculations. In the simpler settings we found the analytic expressions to be only slightly more efficient computationally than the jackknife and in the more realistic settings we found these expressions intractable. The difficulty arises from the nearly closed-form expression for the asymptotic variance which involves a complicated covariance calculation, $\mathbf{Cov}\left(T(\hat{\beta}), \int e^{\beta\,(x-b)}\, d\hat{F}\right)$, where $T(\hat{\beta}) = \int e^{\hat{\beta}(x-b(x))} - e^{\beta(x-b(x))}\, dF$. Even when a straight-forward, analytic expression for this is obtained, the computations are non-trivial and, more importantly, this formulation does not allow diagnostics concerning the accuracy of the linear approximation used in its derivation.

More generally, the methods described in this paper yield point and interval estimates of $\mathbf{AR}$ for unmatched, pair-matched and matched set case-control designs. They allow for models which include any number of confounders and exposures, which can be discrete or continuous, and, further, they allow for arbitrary reference distributions, not just a single, categorical baseline. S-plus software is available for carrying out these computations.

It has been our experience based on examples reported in this article, as well as a wide variety of applications not reported, that the jackknife and bootstrap methods are nearly identical. In most cases the jackknife is preferred due to its speed. When the "nonlinearity" diagnostic is finished we hope to have a way of detecting those settings in which the bootstrap is preferred. Of course, before

reporting final results both the bootstrap and the jackknife should be calculated for diagnostic purposes.

The work on issues related to **AR** continues. We continue to work on diagnostic tools for determining the validity of the bootstrap and jackknife extimates of variance, extending the notion of **AR** to survival/censored data settings, and partitioning **AR** so as to describe that part of the risk that is attributable to a particular subcategory (subset) of the exposure(s).

References

Becker, Chambers, and Wilks (1988) The New S Language *Wadsworth*

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

Benichou, J. and Gail, M. (1989) A Delta Method for Implicitly Defined Random Variables *The American Statistician* 43:41–44

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

Breslow, N.E. and N.E. Day (1980) Statistical Methods in Cancer Research Vol 1: The Analysis of Case-Control Studies International Agency for Research on Cancer Scientific Publications, No. 32, Lyon

Bruzzi, P, Green SB, Byar DP, et al (1985) Estimating the population attributable risk for multiple risk factors using case-control data *American Journal of Epidemiology* 122:904–14

Coughlin, S.S., Nass, C.C., Pickle, L.W., Trock, B. and Bunin, G. (1991) Regression methods for estimating attributable risk in population-based case-control studies: a comparison of additive and multiplicative models *American Journal of Epidemiology* 133:305–313

Cole, P. and MacMahon, B. (1971) Attributable risk percent in case-control studies *British Journal of Preventive and Social Medicine* 25:242–244

Davis, P.H., Dambrosia, J.M., Schoenberg, B.S., et al, (1987) Risk factors for ischemic stroke: a prospective study in Rochester, Minnesota *Annals of Neurology* 22:319–327

Denman, D.W. III and Schlesselman, J.J. (1983) Interval estimation of the attributable risk for multiple exposure levels in case-control studies *Biometrics* 39:185–192

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

Efron, B. (1987) Better bootstrap confidence intervals (with discussion) *J. Amer. Statist. Assoc.* 82:171–200

Efron, B. and Tibshirani, R. (1993) An Introduction to the Bootstrap *Chapman and Hall, New York*

Ejigou, A. (1979) Estimation of attributable risk in the presence of confounding *Biometrical Journal* 21:155–165

Freedman, D., Pisani, Purves and Adhikari (1991) Statistics, 2nd ed. *W. W. Norton*

Garguillo, P., R. Rothenberg, H. Wilson (1995) Confidence intervals, hypothesis tests and sample sizes for the prevented fraction in cross-sectional samples *Statistics in Medicine* 14:51–72

Kleinbaum, D.G., Kupper, L.L. and Morgenstern, H. (1982) Epidemiologic Research: Principles and Quantitative Methods *Lifetime Learning Publications* Belmont, California

Kooperberg, C. and Petitti, D.B. (1991) Using logistic regression to estimate the adjusted attributable risk oflow birthweight in an unmatched case-control study *Epidemiology* 2:363–366

Kuritz, S.J. and Landis, J.R. (1987) Attributable risk ratio estimation from matched-pairs case-control data *American Journal of Epidemiology* 125:324–328

Kuritz, S.J. and Landis, J.R. (1988) Attributable risk estimation from matched case-control data case-control studies *Biometrics* 44:355-367

Kuritz, S.J. and Landis, J.R. (1988) Summary attributable risk estimation from unmatched case-control studies *Statistics in Medicine* 7:507–517

Leung, H.M. and Kupper, L.L. (1981) Comparisons of confidence intervals for attributable risk *Biometrics* 37:293–302

Levin, M. L. (1953) The occurrence of lung cancer in man *Acta Un Intern Cancer* 19:531–541

Machado, E.B., Gabriel, S. Beard, C.M., Michet, C., O'Fallon, W.M., Ballard, D. (1989) A population-based case-control study of temporal arteritis: evidence for an association between temporal arteritis and degnerative vascular disease? *International Journal of Epidemiology* 18:836–841

Markush, R E. (1977) Levin's attributable risk statistic for analytic studies and vital statistics *American Journal of Epidemiology* 105:401–406

Miettinen, O.S. (1974) Proportion of disease caused or prevented by a given exposure, trait or intervention *American Journal of Epidemiology* 99:325–332

Morgenstern, H. and Bursic, E.S. (1982) A method for using epidemiologic data to estimate the potential impact of an intervention on the health status of a target population *Journal of Community Health* 7.292–309

O'Fallon, W.M. and Sicks, J. (1993) *Stroke: Populations, Cohorts and Clinical Trials* 111–134 Edited by J. Whisnant , Publ. Butterworth-Heinemann

Sandok, B.A., Whisnant, J.P., Furlan, A.J. and Mickell, J.L. (1982) Carotid artery bruits: prevalence survey and differential diagnosis *Mayo Clinic Proceedings* 57:227–230

SAS/STAT User's Guide (1990) version 6, vol. 2, Cary NC; SAS Institute

Tuyns, A.J., G. Pequignot and O.M. Jensen (1977) Le cancer de l'oesophage en Ille-et Vilaine en fonction des niveaux de consommation d'alcool et de tabac Bulletin of Cancer 64:45–60

Walter, S. D. (1975) The distribution of Levin's measure of attributable risk *Biometrika* 62:371–374

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

Walter, S. D. (1978) Calculation of Attributable Rısks from Epidimiological Data *Int'l J. of Epid.* 7:175–182

Walter, S. D. (1980) Prevention for multifactorial diseases *American Journal of Epidemiology* 112:409–416

Walter, S. D. (1983) Effects of interaction, confounding and observational error on attributable risk estimation *American Journal of Epidemiology* 117:598–604

Whisnant, J.P., Wiebers, D.O., O'Fallon, W.M., Sicks, J.D., Frye, R.L., (1996) A population-based model of risk factors for ischemic stroke: Rochester, Minnesota *Neurology* 47:1420-1428

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

Whittemore, A.S. (1983) Estimating attributable risk from case-control studies retrospective data *American Journal of Epidemiology* 117:76–85

Wiebers, D.O., Whisnant, J.P., Sandok, B.A. and O'Fallon, W.M. (1990) Prospective comparison of a cohort with asymptomatic carotid bruit and a population-based cohort without carotid bruit *Stroke* 21:984–988

A. Application of software and comparison of results with Benichou (1991)

For purposes of comparison, we reconsider data from case-control study reported by Tuyns, et al (1977) concerning oesophageal cancer. The data have become a staple since Breslow and Day (1980) and were considered as a motivating and illustrative example by Benichou (1991). The data from table 1 of Benichou (1991) are listed in the table on the following page. Using the software and techniques described in this report we refit all of the models from table 3 in Benichou (1991).

The following tables provide summaries of the 18 models fit in Benichou (1991) tables III, IV and V, along with the corresponding $\widehat{\mathbf{AR}}$, the Benichou and Gail estimate of standard error $\widehat{SD}_{BG}$, and the jackknife estimate of standard error $\widehat{SD}_{JK}$.

Model	log(OR)	$\widehat{\mathbf{AR}}$	$\widehat{SD}_{BG}$	$\widehat{SD}_{JK}$
1	$\alpha + \beta Al$	.395	.042	.042
2	$\alpha_1 Ag + \alpha_2 S + \alpha_3 Ag \cdot S + \beta Al$	.382	.044	.044
3	$\alpha_1 Ag + \alpha_2 S + \alpha_3 Ag \cdot S + \beta Al + \gamma AL \cdot Ag$	.380	.044	.044
4	$\alpha_1 Ag + \alpha_2 S + \alpha_3 Ag \cdot S + \beta Al + \gamma AL \cdot S$	.381	.044	.045
5	$\alpha_1 Ag + \alpha_2 S + \alpha_3 Ag \cdot S + \beta Al + \gamma AL \cdot S \cdot Ag$	.380	.044	.045

In these five models, alcohol consumption was considered a binary factor (0 - 79 ; 80+ g/day), and nine parameters were used for the main effect of age, smoking and their interaction in models 2 - 5. In model 5, eight parameters were used to model the interactions of alcohol consumption with smoking and age.

Model	log(OR)	$\widehat{\mathbf{AR}}$	$\widehat{SD}_{BG}$	$\widehat{SD}_{JK}$
6	$\alpha + \beta Al$	.709	.051	.051
7	$\alpha_1 Ag + \alpha_2 S + \alpha_3 Ag \cdot S + \beta Al$	.719	.050	.050
8	$\alpha_1 Ag + \alpha_2 S + \alpha_3 Ag \cdot S + \beta Al + \gamma AL \cdot Ag$	.723	.050	.050
9	$\alpha_1 Ag + \alpha_2 S + \alpha_3 Ag \cdot S + \beta Al + \gamma AL \cdot S$	.703	.054	.053
10	$\alpha_1 Ag + \alpha_2 S + \alpha_3 Ag \cdot S + \beta Al + \gamma AL \cdot S \cdot Ag$	.700	056	.056
11	$\alpha + \beta Al$	.709	.051	.051
12	$\alpha_1 Ag + \alpha_2 S + \alpha_3 Ag \cdot S + \beta Al$	.721	.050	.050
13	$\alpha_1 Ag + \alpha_2 S + \alpha_3 Ag \cdot S + \beta Al + \gamma AL \cdot Ag$	.726	.050	.049
14	$\alpha_1 Ag + \alpha_2 S + \alpha_3 Ag \cdot S + \beta Al + \gamma AL \cdot S$	.703	.055	.054
15	$\alpha_1 Ag + \alpha_2 S + \alpha_3 Ag \cdot S + \beta Al + \gamma AL \cdot S \cdot Ag$	.701	NA	.056

In models 6 - 10, alcohol consumption was considered a binary (0-39 , 40+ g/day) factor, and nine parameters were used for the main effects of age, smoking and their interaction in models 7-10. In

model 10, eight parameters were used to model the interactions of alcohol consumption with smoking and age.

In models 11 - 15, alcohol consumption was considered a polychotomus factor (0-39 ,40-79, 80-119, 120+ g/day) factor, and nine parameters were used for the main effects of age, smoking and their interaction in models 12-15. In model 15, twenty-four parameters were used to model the interactions of alcohol consumption with smoking and age.

Model	log(OR)	$\widehat{\mathbf{AR}}$	$\widehat{SD}_{BG}$	$\widehat{SD}_{JK}$
16	$\alpha + \beta\mathbf{X}$	.862	.046	.043
17	$\alpha Ag + \beta\mathbf{X}$	.866(.867)	.045	.043
18	$\alpha Ag + \beta\mathbf{X} + \gamma Ag \cdot \mathbf{X}$	.868(.872)	.044	.041

In models 16-18 the main effect of alcohol consumption and smoking was modelled with one parameter corresponding to one binary variable $\mathbf{X}$ with baseline defined by 0-39 g/day and 0-9 g/day of alcohol consumption and smoking, respectively. Three parameters were used to model the interaction between $\mathbf{X}$ and Ag.

Note that the $\widehat{\mathbf{AR}}$ values given in the previous table in parentheses are from our software and differ slightly from Benichou's estimates. The `arhat` S-plus program duplicates all of the remaining $\widehat{\mathbf{AR}}$ estimates from Benichou's review article and even gives a reasonable estimate of variance in model 15 where the delta method cannot provide an estimate of the standard error of $\widehat{\mathbf{AR}}$.

Alcohol Consumption (g/day)	Age (years)	Smoking (g/day)	Number of cases	Number of controls
0-39	25-44	0-9	0	100
		10-29	1	36
		30+	0	13
	45-64	0-9	1	45
		10-29	0	28
		30+	0	4
	65+	0-9	8	107
		10-29	14	47
		30+	5	6
40-79	25-44	0-9	0	62
		10-29	4	44
		30+	0	15
	45-64	0-9	6	32
		10-29	9	27
		30+	5	2
	65+	0-9	28	51
		10-29	19	44
		30+	4	3
80-119	25-44	0-9	0	13
		10-29	0	9
		30+	0	3
	45-64	0-9	3	13
		10-29	7	12
		30+	2	2
	65+	0-9	16	16
		10-29	18	19
		30+	5	0
120+	25-44	0-9	2	2
		10-29	3	6
		30+	0	2
	45-64	0-9	4	0
		10-29	5	2
		30+	4	0
	65+	0-9	10	6
		10-29	11	3
		30+	6	1

What follows are the results from the arhat function which give the results from Benichou's 1991 review article No new features of the software are used, but it shows the ease with which one can fit and choose among various models. This is also used as an error check for the software.

From table III in Benichou (1991) we have

```
> model1 <- arhat(cases ~ expos(alcohol80), data = benichou,
+                 jack = T, B = 500)
Call:
arhat(formula = cases ~ expos(alcohol80), data = benichou,
jackknife = T, B = 500)

Estimate of AR = 0.395 .
The mean of the jackknifed AR's =  0.395 .
The standard error of the jackknifed AR's =  0.042 .
The mean of the  500  bootstrapped AR's =  0.396 .
The standard deviation of the  500  bootstrapped AR's =  0.042 .

> summary(model1)
Call:
arhat(formula = cases ~ expos(alcohol80), data = benichou, B = 500, jackknife
 = T)

Estimate of AR =  0.395  .

Jackknife CI                              ( 0.312 , 0.478 )
Percentile Bootstrap CI                   ( 0.310 , 0.473 )
Percentile.t Bootstrap CI                 ( 0.312 , 0.478 )
Bias-corrected Bootstrap CI               ( 0.308 , 0.471 )
Accelerated Bias-corrected Bootstrap CI ( 0.308 , 0.471 )

> model2 <- arhat(cases ~ age*smoke + expos(alcohol80) ,
+                 data = benichou , jack = T, B = 500)
Call:
arhat(formula = cases ~ age * smoke + expos(alcohol80),
      data = benichou, jackknife = T, B = 500)
```

```
Estimate of AR = 0.382 .
The mean of the jackknifed AR's =  0.382 .
The standard error of the jackknifed AR's =  0.044 .
The mean of the  500  bootstrapped AR's =  0.383 .
The standard deviation of the  500  bootstrapped AR's =  0.045 .

> summary(model2)
Call:
arhat(formula = cases ~ age * smoke + expos(alcohol80), data = benichou, B =
500, jackknife = T)

Estimate of AR =  0.382  .

Jackknife CI                              ( 0.295 , 0.469 )
Percentile Bootstrap CI                   ( 0.291 , 0.468 )
Percentile.t Bootstrap CI                 ( 0.294 , 0.469 )
Bias-corrected Bootstrap CI               ( 0.296 , 0.474 )
Accelerated Bias-corrected Bootstrap CI ( 0.297 , 0.474 )

> model3 <- arhat(cases ~ age*(smoke + expos(alcohol80)) ,
+                 data = benichou , jack = T, B = 500)
Call:
arhat(formula = cases ~ age * (smoke + expos(alcohol80)) ,
      data = benichou, jackknife = T, B = 500)

Estimate of AR = 0.38 .
The mean of the jackknifed AR's =  0.38 .
The standard error of the jackknifed AR's =  0.044 .
The mean of the  500  bootstrapped AR's =  0.381 .
The standard deviation of the  500  bootstrapped AR's =  0.042 .

> summary(model3)
Call:
arhat(formula = cases ~ age * (smoke + expos(alcohol80)), data = benichou, B =
```

```
500, jackknife = T)

Estimate of AR =  0.38  .

Jackknife CI                                ( 0.293 , 0.467 )
Percentile Bootstrap CI                     ( 0.301 , 0.459 )
Percentile.t Bootstrap CI                   ( 0.297 , 0.463 )
Bias-corrected Bootstrap CI                 ( 0.297 , 0.456 )
Accelerated Bias-corrected Bootstrap CI ( 0.297 , 0.457 )

> model4 <- arhat(cases ~ smoke*(age + expos(alcohol80)) ,
+                 data = benichou , jack = T, B = 500)
Call:
arhat(formula = cases ~ smoke * (age + expos(alcohol80)), data = benichou, B =
500, jackknife = T)

Estimate of AR = 0.381 .
The mean of the jackknifed AR's =  0.381 .
The standard error of the jackknifed AR's =  0.045 .
The mean of the  500  bootstrapped AR's =  0.381 .
The standard deviation of the  500  bootstrapped AR's =  0.046 .

> summary(model4)
Call:
arhat(formula = cases ~ smoke * (age + expos(alcohol80)), data = benichou, B =
500, jackknife = T)

Estimate of AR =  0.381  .

Jackknife CI                                ( 0.294 , 0.469 )
Percentile Bootstrap CI                     ( 0.296 , 0.469 )
Percentile.t Bootstrap CI                   ( 0.292 , 0.471 )
Bias-corrected Bootstrap CI                 ( 0.296 , 0.469 )
Accelerated Bias-corrected Bootstrap CI ( 0.296 , 0.469 )
```

```
> model5 <- arhat(cases ~ age*smoke + expos(alcohol80)*(smoke:age) ,
+                 data = benichou , jack = T, B = 500)
Call:
arhat(formula = cases ~ age * smoke + expos(alcohol80) * (smoke:age), data =
benichou, B = 500, jackknife = T)

Estimate of AR = 0.38 .
The mean of the jackknifed AR's =  0.38 .
The standard error of the jackknifed AR's =  0.045 .
The mean of the  500  bootstrapped AR's = 0.379 .
The standard deviation of the  500  bootstrapped AR's =  0.044 .

> summary(model5)
Call:
arhat(formula = cases ~ age * smoke + expos(alcohol80) * (smoke:age), data =
benichou, B = 500, jackknife = T)

Estimate of AR =  0.38  .

Jackknife CI                              ( 0.292 , 0.467 )
Percentile Bootstrap CI                   ( 0.284 , 0.467 )
Percentile.t Bootstrap CI                 ( 0.292 , 0.467 )
Bias-corrected Bootstrap CI               ( 0.281 , 0.457 )
Accelerated Bias-corrected Bootstrap CI ( 0.282 , 0.457 )
```

From table IV in Benichou (1991) we have

```
> model6 <- arhat(cases ~ expos(alcohol40), data = benichou, jack = T, B = 500)
Call:
arhat(formula = cases ~ expos(alcohol40), data = benichou, B = 500, jackknife
 = T)

Estimate of AR = 0.709 .
The mean of the jackknifed AR's =  0.709 .
The standard error of the jackknifed AR's =  0.051 .
The mean of the  500  bootstrapped AR's =  0.710 .
```

```
The standard deviation of the  500  bootstrapped AR's =  0.051 .

> summary(model6)
Call:
arhat(formula = cases ~ expos(alcohol40), data = benichou, B = 500, jackknife
 = T)

Estimate of AR =  0.709  .

Jackknife CI                              ( 0.609 , 0.809 )
Percentile Bootstrap CI                   ( 0.602 , 0.804 )
Percentile.t Bootstrap CI                 ( 0.608 , 0.809 )
Bias-corrected Bootstrap CI               ( 0.601 , 0.798 )
Accelerated Bias-corrected Bootstrap CI ( 0.583 , 0.781 )

> model7 <- arhat(cases ~ age*smoke + expos(alcohol40) ,
+                 data = benichou , jack = T, B = 500)
Call:
arhat(formula = cases ~ age * smoke + expos(alcohol40), data = benichou, B =
500, jackknife = T)

Estimate of AR = 0.719 .
The mean of the jackknifed AR's =  0.719 .
The standard error of the jackknifed AR's =  0.05 .
The mean of the  500  bootstrapped AR's =  0.718 .
The standard deviation of the  500  bootstrapped AR's =  0.052 .

> summary(model7)
Call:
arhat(formula = cases ~ age * smoke + expos(alcohol40), data = benichou, B =
500, jackknife = T)

Estimate of AR =  0.719  .

Jackknife CI                              ( 0.621 , 0.818 )
```

```
Percentile Bootstrap CI                  ( 0.608 , 0.810 )
Percentile.t Bootstrap CI                ( 0.618 , 0.821 )
Bias-corrected Bootstrap CI              ( 0.605 , 0.810 )
Accelerated Bias-corrected Bootstrap CI ( 0.564 , 0.795 )

> model8 <- arhat(cases ~ age*(smoke + expos(alcohol40)) ,
+                 data = benichou , jack = T)
Call:
arhat(formula = cases ~ age * (smoke + expos(alcohol40)), data = benichou, B =
500, jackknife = T)

Estimate of AR = 0.723 .
The mean of the jackknifed AR's =  0.723 .
The standard error of the jackknifed AR's =  0.05 .
The mean of the  500  bootstrapped AR's =  0.722 .
The standard deviation of the  500  bootstrapped AR's =  0.052 .

> summary(model8)
Call:
arhat(formula = cases ~ age * (smoke + expos(alcohol40)), data = benichou, B =
500, jackknife = T)

Estimate of AR =  0.723  .

Jackknife CI                             ( 0.626 , 0.821 )
Percentile Bootstrap CI                  ( 0.622 , 0.821 )
Percentile.t Bootstrap CI                ( 0.621 , 0.825 )
Bias-corrected Bootstrap CI              ( 0.623 , 0.824 )
Accelerated Bias-corrected Bootstrap CI ( 0.619 , 0.821 )

> model9 <- arhat(cases ~ smoke*(age + expos(alcohol40)) ,
+                 data = benichou , jack = T , B = 500)
Call:
arhat(formula = cases ~ smoke * (age + expos(alcohol40)), data = benichou, B =
500, jackknife = T)
```

```
Estimate of AR = 0.703 .
The mean of the jackknifed AR's =  0.703 .
The standard error of the jackknifed AR's =  0.053 .
The mean of the  500  bootstrapped AR's =  0.699 .
The standard deviation of the  500  bootstrapped AR's =  0.057 .

> summary(model9)
Call:
arhat(formula = cases ~ smoke * (age + expos(alcohol40)), data = benichou, B =
500, jackknife = T)

Estimate of AR =  0.703  .

Jackknife CI                              ( 0.598 , 0.808 )
Percentile Bootstrap CI                   ( 0.578 , 0.802 )
Percentile.t Bootstrap CI                 ( 0.592 , 0.814 )
Bias-corrected Bootstrap CI               ( 0.576 , 0.800 )
Accelerated Bias-corrected Bootstrap CI ( 0.576 , 0.796 )

> model10 <- arhat(cases ~ age*smoke + expos(alcohol40)*(smoke:age) ,
+                  data = benichou , jack = T , B = 500)
Call:
arhat(formula = cases ~ age * smoke + expos(alcohol40) * (smoke:age), data =
benichou, B = 500, jackknife = T)

Estimate of AR = 0.7 .
The mean of the jackknifed AR's =  0.699 .
The standard error of the jackknifed AR's =  0.056 .
The mean of the  500  bootstrapped AR's =  -4.36 .
The standard deviation of the  500  bootstrapped AR's =  74.1 .

> summary(model10)
Call:
arhat(formula = cases ~ age * smoke + expos(alcohol40) * (smoke:age), data =
```

```
benichou, B = 500, jackknife = T)

Estimate of AR =  0.7  .

Jackknife CI                              ( 0.590 , 0.809 )
Percentile Bootstrap CI                   ( 0.558 , 0.787 )
Percentile.t Bootstrap CI                 ( -144  , 146   )
Bias-corrected Bootstrap CI               ( 0.580 , 0.794 )
Accelerated Bias-corrected Bootstrap CI ( 0.583 , 0.799 )

> model11 <- arhat(cases ~ expos(alcohol) , data = benichou , jack = T , B = 500)
Call:
arhat(formula = cases ~ expos(alcohol), data = benichou, B = 500, jackknife = T)

Estimate of AR = 0.709 .
The mean of the jackknifed AR's =  0.709 .
The standard error of the jackknifed AR's =  0.051 .
The mean of the  500  bootstrapped AR's =  0.706 .
The standard deviation of the  500  bootstrapped AR's =  0.051 .

> summary(model11)
Call:
arhat(formula = cases ~ expos(alcohol), data = benichou, B = 500, jackknife = T)

Estimate of AR =  0.709  .

Jackknife CI                              ( 0.609 , 0.809 )
Percentile Bootstrap CI                   ( 0.598 , 0.801 )
Percentile.t Bootstrap CI                 ( 0.610 , 0.808 )
Bias-corrected Bootstrap CI               ( 0.626 , 0.807 )
Accelerated Bias-corrected Bootstrap CI ( 0.626 , 0.807 )

> model12 <- arhat(cases ~ age*smoke + expos(alcohol) ,
+                  data = benichou , jack = T , B = 500)
Call:
```

```
arhat(formula = cases ~ age * smoke + expos(alcohol), data = benichou, B = 500,
jackknife = T)

Estimate of AR = 0.721 .
The mean of the jackknifed AR's =  0.721 .
The standard error of the jackknifed AR's =  0.05 .
The mean of the  500  bootstrapped AR's =  0.721 .
The standard deviation of the  500  bootstrapped AR's =  0.047 .

> summary(model12)
Call:
arhat(formula = cases ~ age * smoke + expos(alcohol), data = benichou, B = 500,
jackknife = T)

Estimate of AR =  0.721  .

Jackknife CI                              ( 0.623 , 0.819 )
Percentile Bootstrap CI                   ( 0.626 , 0.804 )
Percentile.t Bootstrap CI                 ( 0.629 , 0.812 )
Bias-corrected Bootstrap CI               ( 0.616 , 0.798 )
Accelerated Bias-corrected Bootstrap CI ( 0.604 , 0.792 )

> model13 <- arhat(cases ~ age*(smoke + expos(alcohol)) ,
+                  data = benichou , jack = T , B = 500)
Call:
arhat(formula = cases ~ age * (smoke + expos(alcohol)), data = benichou, B =
500, jackknife = T)

Estimate of AR = 0.726 .
The mean of the jackknifed AR's =  0.726 .
The standard error of the jackknifed AR's =  0.049 .
The mean of the  500  bootstrapped AR's =  0.726 .
The standard deviation of the  500  bootstrapped AR's =  0.053 .

> summary(model13)
```

```
Call:
arhat(formula = cases ~ age * (smoke + expos(alcohol)), data = benichou, B =
500, jackknife = T)

Estimate of AR =  0.726  .

Jackknife CI                                ( 0.629 , 0.822 )
Percentile Bootstrap CI                     ( 0.616 , 0.827 )
Percentile.t Bootstrap CI                   ( 0.622 , 0.830 )
Bias-corrected Bootstrap CI                 ( 0.594 , 0.819 )
Accelerated Bias-corrected Bootstrap CI ( 0.594 , 0.827 )

> model14 <- arhat(cases ~ smoke*(age + expos(alcohol)) ,
+                  data = benichou , jack = T , B = 500)
Call:
arhat(formula = cases ~ smoke * (age + expos(alcohol)), data = benichou, B =
500, jackknife = T)

Estimate of AR = 0.703 .
The mean of the jackknifed AR's =  0.703 .
The standard error of the jackknifed AR's =  0.054 .
The mean of the  500  bootstrapped AR's =  0.688 .
The standard deviation of the  500  bootstrapped AR's =  0.278 .

> summary(model14)
Call:
arhat(formula = cases ~ smoke * (age + expos(alcohol)), data = benichou, B =
500, jackknife = T)

Estimate of AR =  0.703  .

Jackknife CI                                ( 0.598 , 0.808 )
Percentile Bootstrap CI                     ( 0.583 , 0.811 )
Percentile.t Bootstrap CI                   ( 0.158 , 1.250 )
Bias-corrected Bootstrap CI                 ( 0.588 , 0.813 )
```

```
Accelerated Bias-corrected Bootstrap CI ( 0.583 , 0.813 )

> model15 <- arhat(cases ~ age*smoke + expos(alcohol)*(smoke:age) ,
+                  data = benichou , jack = T , B = 500)
Call:
arhat(formula = cases ~ age * smoke + expos(alcohol) * (smoke:age), data =
benichou, B = 500, jackknife = T)

Estimate of AR = 0.701 .
The mean of the jackknifed AR's =  0.701 .
The standard error of the jackknifed AR's =  0.056 .
The mean of the  500  bootstrapped AR's =  -17.1 .
The standard deviation of the  500  bootstrapped AR's =  187 .

> summary(model15)
Call:
arhat(formula = cases ~ age * smoke + expos(alcohol) * (smoke:age), data =
benichou, B = 500, jackknife = T)

Estimate of AR =  0.701  .

Jackknife CI                              ( 0.591 , 0.810 )
Percentile Bootstrap CI                   ( 0.517 , 0.800 )
Percentile.t Bootstrap CI                 ( -365  , 367 )
Bias-corrected Bootstrap CI               ( 0.536 , 0.802 )
Accelerated Bias-corrected Bootstrap CI ( -118 , 0.800 )
```

Note that in Benichou (1991), the estimate of AR for model 15 is 0.701 though no standard error is given.

From table V in Benichou (1991) we have

```
> model16 <- arhat(cases ~ expos(smoke)*expos(alcohol) ,
+                  data = benichou , B = 500
+                  baseline = base2 , jack = T)
Call:
arhat(formula = cases ~ expos(smoke) * expos(smoke), data = benichou, B = 500,
```

```
jackknife = T, baseline = base2, categorical = F)

Estimate of AR = 0.862 .
The mean of the jackknifed AR's =  0.862 .
The standard error of the jackknifed AR's =  0.043 .
The mean of the  500  bootstrapped AR's =  0.863 .
The standard deviation of the  500  bootstrapped AR's =  0.045 .

> summary(model16)
Call:
arhat(formula = cases ~ expos(smoke) * expos(alcohol), data = benichou, B = 500,
jackknife = T, baseline = base2, categorical = F)

Estimate of AR =  0.862  .

Jackknife CI                                ( 0.776 , 0.947 )
Percentile Bootstrap CI                     ( 0.759 , 0.942 )
Percentile.t Bootstrap CI                   ( 0.773 , 0.950 )
Bias-corrected Bootstrap CI                 ( 0.749 , 0.932 )
Accelerated Bias-corrected Bootstrap CI ( 0.743 , 0.942 )

> model17 <- arhat(cases ~ age + expos(smoke)*expos(alc) , B = 500,
+                  data = benichou , baseline = base2 , jack = T)
Call:
arhat(formula = cases ~ age + expos(smoke) * expos(alcohol), data =
benichou, B = 500, jackknife = T, baseline = base2)

Estimate of AR = 0.867 .
The mean of the jackknifed AR's =  0.867 .
The standard error of the jackknifed AR's =  0.043 .
The mean of the  500  bootstrapped AR's =  0.865 .
The standard deviation of the  500  bootstrapped AR's =  0.043 .

> summary(model17)
Call:
```

```
arhat(formula = cases ~ age + expos(smoke) * expos(alcohol), data =
benichou, B = 500, jackknife = T, baseline = base2)

Estimate of AR =  0.867  .

Jackknife CI                                ( 0.782 , 0.951 )
Percentile Bootstrap CI                     ( 0.776 , 0.938 )
Percentile.t Bootstrap CI                   ( 0.782 , 0.951 )
Bias-corrected Bootstrap CI                 ( 0.776 , 0.938 )
Accelerated Bias-corrected Bootstrap CI ( 0.734 , 0.917 )

> model18 <- arhat(cases ~ age*expos(smoke)*expos(alcohol) ,
+                  data = benichou , baseline = target1 , jack = T)
Call:
arhat(formula = cases ~ age * expos(smoke) * expos(alcohol), data =
benichou, B = 500, jackknife = T, baseline = base2)

Estimate of AR = 0.872 .
The mean of the jackknifed AR's =  0.872 .
The standard error of the jackknifed AR's =  0.041 .
The mean of the  500  bootstrapped AR's =  0.871 .
The standard deviation of the  500  bootstrapped AR's =  0.039 .

> summary(model18)
Call:
arhat(formula = cases ~ age * expos(smoke) * expos(alcohol), data =
benichou, B = 500, jackknife = T, baseline = base2)

Estimate of AR =  0.872  .

Jackknife CI                                ( 0.792 , 0.951 )
Percentile Bootstrap CI                     ( 0.792 , 0.943 )
Percentile.t Bootstrap CI                   ( 0.795 , 0.949 )
Bias-corrected Bootstrap CI                 ( 0.789 , 0.942 )
Accelerated Bias-corrected Bootstrap CI ( 0.720 , 0.916 )
```

B. Help function for arhat

For completeness we include a listing of the help function for arhat. This can be obtained in S, as usual, with

```
> help(arhat)
```

```
Calculates (semiparametric) MLE, standard errors and confidence
intervals for attributable risk in unmatched, pair-matched or
set-matched case-control designs, with respect to any
reference(baseline) distribution, for any number of exposure and
confounder variables, either categorical or continuous. Also,
calculates jackknife and bootstrap estimates of variance and four
bootstrap confidence intervals (percentile-t, percentile, bias-
corrected and accelerated bias-corrected). There are also options
for monitoring the bootstrap iterations and other diagnostic
information. When you are dealing with the matched set case-
control design, you have to use the coxph method to get the
attributable risks, estimates of variance etc.

DESCRIPTION:
        Estimates population (etiological) attributable risk for
        unmatched, pair-matched or set-matched case-control
        designs and returns a list containing the estimated
        attributable risk, estimates of coefficients, and their
        standard errors, from the (conditional, if necessary)
        logistic regression used for estimating the relative risk.

USAGE:
        arhat(formula=formula(data), family=binomial(link = logit),
                data=sys.parent(), weights, subset, na.action =
                na.omit, start=NULL, control=glm.control(...),
                model=F, x=F, y=F, B=0, jackknife=F, coxph=F,
                contrasts=NULL, baseline=NULL, diagnostics=F,
                conf.level=0.95, categorical=T, ...)

REQUIRED ARGUMENTS:
```

formula: a formula expression as for other regression models, of the form response ~ predictors . See the documentation of lm and formula for more details.

OPTIONAL ARGUMENTS:

family: a family object - a list of functions and expressions for defining the link and variance functions, initialization and iterative weights. Families supported are gaussian, binomial, poisson, Gamma, inverse.gaussian and quasi. Functions like binomial produce a family object, but can be given without the parentheses. Family functions can take arguments, as in binomial(link=probit).

data: an optional data frame in which to interpret the variables occurring in the formula.

weights: the optional weights for the fitting criterion.

subset: expression saying which subset of the rows of the data should be used in the fit. This can be a logical vector (which is replicated to have length equal to the number of observations), or a numeric vector indicating which observation numbers are to be included, or a character vector of the row names to be included. All observations are included by default.

na.action: a function to filter missing data. This is applied to the model.frame after any subset argument has been used. The default, na.omit, deletes observations that contain one or more missing values. Also, see na.fail which returns an error if there are any missing values.

start: a vector of initial values on the scale of the linear predictor.

control: a list of iteration and algorithmic constants. See glm.control for their names and default values. These can also be set as arguments to glm itself.

model: if TRUE, the model.frame is returned. If this argument is itself a model.frame, then the formula and data argu-

ments are ignored, and model is used to define the model (default is FALSE).

x: logical flag: if TRUE, the model.matrix is returned (default is FALSE).

y: logical flag: if TRUE, the response variable is returned (default is FALSE).

B: the number of bootstrap iterations desired (default is 0, no bootstrap).

categorical: logical flag: if FALSE, the covariates are not tabled and the estimate of the attributable risk is gotten by using the empirical distribution function of the exposure and confounder levels among the cases. If TRUE, all confounders and exposure covariates are assumed to be categorical ("factor") variables and appropriate tables are calculated. (Default is TRUE).

jackknife: logical flag: if TRUE, the vector of jackknife attributable risks is calculated. (Default is FALSE).

coxph: logical flag: if TRUE, the coxph method will be used to calculate conditional logistic regression estimates of the parameters (Default is FALSE). In the unmatched case-control design, coxph = F. In the matched pair case-control design, coxph can be either T or F. If F, then conditional logistic regeression estimates are calculated using glm . In the matched set case-control design, coxph = T. If you set this argument wrong during the call, the function will give you warning message, and it will automatically try to use the apropriate method.

contrasts: in this function we REQUIRE treatment coding, see page 36 of Statistical Modelling in S.

baseline: baseline must be either NULL, or a data frame with variables (names attribute) which form a superset of the exposure variables exposure variables expressed in the formula, with the desired baseline value for each individual. The default is NULL. If categorical = F, then a non-NULL baseline is required.

diagnostics: logical variable for choosing to return a variety of diagnostics for monitoring the bootstrap iterations. If TRUE, arhat returns logistic regression coefficients and standard errors, deviances, relative risks and prevalences for *each* bootstrap sample. There is useful information here when the exposure is only mildly deleterious or when the confounder-by-exposure table contains cells with small counts as the estimated attributable risk can be far outside the range of (0,1). If FALSE, the only values returned are the call to the function and the bootstrap iterations' arhats.

conf.level: the confidence level at which the bootstrap confidence intervals are calculated.

...: control arguments may be given directly, see the control argument.

VALUE:

an object of class arhat is returned, which inherits from glm, which inherits from lm. The object returns the following components :

call: the call to arhat

arhat: a vector of length B+1 containing the estimate of attributable risk based on the original data and the bootstrapped attributable risks, if any.

match.action: a vector of indeces useful from pair-matched and set-matched designs. This vector indicates, in the appropriate order, which rows of the original data were used, ordering so that the match.id is in asccending order and within pairs the control precedes the case. This is the order of the relative risks, residuals, etc. when a pair-matched or a set-matched design is fit. So, when plotting relative risks, one must use plot(example$rel.risk , exposure[match.action]) .

na.action: for an umatched case-control design, indicates which rows were omitted in the analysis. Hence, for a

relative risk plot from an unmatched design, one must use plot(example$rel.risk , exposure[!na.action]) .

coefficients: the (conditional, if necessary) logistic regression coefficients corresponding to the formula

std.errs: the standard errors of the logistic regression coefficients

deviance: the deviance corresponding to the logistic regression model

rel.risk: a vector of the relative risks for each individual used in the analysis after omitting NA's

risk.table: if categorical = T, a table with the relative risk for each combination of the exposure and confounder levels

bootstrap.tables: if diagnostics = T and categorical = T, the case-by-exposure-by-confounder tables for each bootstrap sample

orig.table: the original case-by-exposure-by-confounder table

prevalence.table: if categorical = T, the table of prevalences for each exposure level *among the cases* .

jackknife: the vector of jackknifed attributable risks, one for each data point, as if that data point were omitted from the fit

influences: a vector of jackknife estimates of the influence functionals. The i-th element is the jackknife approximation to the directional derivative of attributable risk in the direction of the i-th data point.

variance.jack: the jackknife estimate of variance of the estimate of attributable risk

rank: the rank of the design matrix corresponding to the formula

residuals: the residuals of the (conditional) logistic regression

df.resid: the error degrees of freedom corresponding to the (conditional) logistic regression

percentile.t: if B > 0, the percentile-t bootstrap confidence interval based on the B bootstrap iterations.

percentile: if B > 0, the percentile method bootstrap confidence interval based on the B bootstrap iterations.

bias.corrected: if B > 0, the bias-corrected bootstrap confidence interval based on the B bootstrap iterations.

acc.bias.corrected: if B > 0, the accelerated bias-corrected bootstrap confidence interval based on the B bootstrap iterations.

DETAILS:

Fits the specified logistic regression model for an unmatched case-control, or conditional logistic regression for a pair-matched or a set-matched case-control design. Calculates the appropriate odds ratio (which, for rare diseases is a good approximation to the relative risk), prevalence of confounder-by-exposure levels among the diseased and a composite summary measure of all of these, the attributable risk. This function also calculates a bootstrap estimate of the variance of the estimated attributable risk, four different bootstrap confidence intervals (percentile-t, percentile, bias-coreected and accelerated bias-corrected), as well as the jackknife estimate of variance. Bootstrap resampling is done keeping each indvidual's disease status and covariate information together.

SEE ALSO:

glm, print.arhat, summary.arhat, print.summary.arhat

Background:

Attributable risk (AR) is a relevant quantity in epidimiology (sometimes called etiological risk, population attributable risk, etc.) and has been around at least since, Levin (1953). AR can be thought of as the public health impact of an exposure on some disease. This implementation most closely follows the review article by

Benichou (1991).

SIDE EFFECTS:

No known side effects.

REFERENCES:

Benichou, J. (1991) Methods of Adjustment for Estimating Attributable Risk in Case-Control Studies: A Review Statistics in Medicine, Vol. 10, pp. 1753-1773

Levin, M.L. (1953) The Occurrence of Lung Cancer in Man Acta Unio Internationalis contra Cancrum, Vol. 9, pp. 531-541

EXAMPLES:

```
# For the continuous exposure variable  diastolic , find the
# attributable risk for stroke (i.e. the disease) while adjusting
# for individuals' smoking levels. baseline is defined to be
# "reduce all diastolic blood pressures by 10%. Do the jackknife and
# 1000 bootstrap iterations.

arhat(cases ~ smoke * expos(diastolic) ,
      data = stroke.dat , categorical = F ,
      baseline = data.frame(diastolic = .90*stroke.dat$diastolic) ,
      jackknife = T , B = 1000)

# Same as above, but now consider  smoke  as an exposure
#  variable and consider moving all  'Current'  smokers to
# the  'Former'  category.

smoke.base <- stroke.dat$smoke

smoke.base [smoke.base == "Current"] <- "Former"

stroke.base <-
```

```
      data.frame(
           diastolic = .90*stroke.dat$diastolic ,
           smoke = smoke.base)

arhat(cases ~ expos(smoke) * expos(diastolic) ,
       data = stroke.dat , categorical = F ,
       baseline = stroke.base ,
       jackknife = T , B = 1000)

# For categorical data, we fit the model that adjusts  for
# age while assesing the attributable risk associated with
# smoking and alcohol consumption.  Ask for  5000   itera-
# tions and calculate the jackknife.

arhat(cases ~ age*expos(smoke)*expos(alcohol) ,
      data = benichou , B = 5000 , jackknife = T)

# This is a set-matched case-control design, we use  coxph
#  method  to  calculate the attributable risk and related
# results.

arhat(case ~ expos(evrsmk) +  expos(hang)  +  strata(set),
               data =  all, coxph = T)

Estimate of attributable risk = 0.3063403 .

Coefficients:
    evrsmk      hang
 0.7630788 0.5998175
```