ATTRIBUTABLE RISK ESTIMATION

IN

MATCHED CASE-CONTROL STUDIES

ATTRIBUTABLE RISK ESTIMATION

IN

MATCHED CASE-CONTROL STUDIES

BY

ISAAC FRIMPONG NUAMAH, B.Sc.

A Project

Submitted to the School of Graduate Studies

in Partial Fulfilment of the Requirements

for the Degree

Master of Science

McMaster University

July, 1987.

MASTER OF SCIENCE (1987)   McMASTER UNIVERSITY
(Statistics)   Hamilton, Ontario

TITLE : Attributable Risk Estimation in Matched Case-Control Studies

AUTHOR : Isaac Frimpong Nuamah, B.Sc. (U of Ghana)

SUPERVISOR : Professor Stephen D. Walter

NUMBER OF PAGES : vi, 64

ABSTRACT

This project discusses some of the methodologies developed over the years to estimate attributable risk among exposed persons and the attributable risk in the entire population (also called Etiologic Fraction). It provides a general framework for estimating attributable risk among the exposed (denoted λ_e). By making use of the recent observation that the two measures of attributable risk can be linked through the prevalence of the risk factor among the cases (denoted V_X), an estimate of population attributable risk (denoted λ) for matched case-control studies is determined. Using the methodology developed recently by Kuritz and Landis (1987), this project provides explicit formulas for estimating the attributable risk among the exposed and the population attributable risk, and their large sample variances. This has been done both in situations where exactly R controls have been matched to a case and for a variable number of controls per case. The methodologies are illustrated with data from some case-control studies reported in the literature. Asymptotic relative efficiencies of different matching designs computed in terms of the costs of gathering cases and controls, are presented, together with some recommendations on what design is considered optimal.

ACKNOWLEDGEMENTS

I wish to express my deepest gratitute to my supervisor, Prof. S. D. Walter, for suggesting this project and for his support, continued guidance and encouragement during the preparation of this project. I would also like to thank him for the many hours he spent with me, from helping me understand the concepts involved, to digging out important references for me, and I am very grateful to him for helping to deepen my interest in the field of Biostatistics.

I would also like to express my deepest appreciation to my parents, my brothers and sisters for their prayers, moral support and encouragement.

I would also like to thank my colleagues Paul, Florence, Rohan and David for their encouragement.

My special thanks to Joseph Torgbor for helping me with the typing of this project, for his encouragement and the many hours (including weekends) that he had to spend with me as a result.

Last, but not the least, I would like to thank McMaster University for their financial assistance and especially to Prof. C. W. Dunnett, for making my dream a reality.

TABLE OF CONTENTS

CHAPTER 1

CHAPTER 2

STATISTICAL FRAMEWORK

CHAPTER 3

MATRIX FORMULATION AND EXAMPLES

CHAPTER 4

EFFICIENCY CONSIDERATIONS

APPENDIX

CHAPTER 1

INTRODUCTION

Usually in the analysis of epidemiologic data it is useful to obtain an estimate of the proportion of some disease that is associated with exposure to a risk factor. This helps in order to meaningfully control diseases since a knowledge of the disease burden that could be prevented by modifying a given risk factor is necessary. Levin's measure of attributable risk (Levin (1953)) was proposed to measure the proportion of disease burden attributable to a risk factor. Various strategies for estimating attributable risk proportion have received increasing attention in the literature (Cole and MacMahon (1971), Miettinen (1974), Walter (1976)). The extension of the technique of attributable risk estimation in the presence of confounding variables (Bruzzi et al (1985), Whittemore (1982)) has been done; and Park (1983) found an estimate for it for recurrent events. Recently, an extension to matched-pair case-control data has been carried out (Kuritz and Landis (1987a)).

A measure of the strength of an association between exposure and a putative risk factor is the odds ratio (Cornfield (1951)). It is the ratio of the odds of a disease when the factor is present to the odds when the factor is absent (see (2.1)). The odds of a disease are defined as the probability of having the disease, divided by the probability of not having the disease. Another relevant measure usually used in epidemiology is the relative risk, which is the ratio of the incidence of a disease in persons exposed to persons not exposed to a risk factor. However, these measures do not provide any information on the actual numbers of affected individuals in the target population. Hence, measures of attributable risk, which attempted to overcome this problem associated with relative risk and odds ratio have been suggested. Walter (1976) has noted that though the procedure for estimating attributable risk may be analogous to those used in the theory of estimation of relative risk, the interpretation of these two risk measures are different. Attributable risk should therefore in no way be regarded as a substitute for relative risk, but rather be regarded as an additional dimension of health hazard appraisal.

In estimating attributable risk, it becomes clear that we are interested in the disease producing role of an etiologic factor. Attributable risk (also called Etiologic Fraction) can be looked upon as the fraction of the disease which would not have occurred had the factor of interest been absent from the population (ie the proportion of the disease attributable to the factor of interest). If we are interested in the fraction of the disease prevented by a beneficial factor (eg an intervention program), the term 'Prevented Fraction' has been suggested (Miettinen (1974)).

Retrospective (or Case-Control) studies are commonly used in epidemiologic research because of the relative ease of gathering disease cases in a short space of time, although it seems to suffer from such biases as those related to the recall of previous exposure to a risk factor.

The odds ratio (which provides an estimate of relative risk in case-control studies under a rare disease assumption (Chapter 2)) can be found for case-control data. It is possible to have a high relative risk, but which may not be an important health problem because very few people are actually exposed to it. On the other hand, a low relative risk may be quite important if a large number of people are exposed to the factor; hence the need of a measure of attributable risk.

Levin (1953) first introduced measures of attributable risk of a specified disease associated with a selected risk factor when both are classified as absent or present. He proposed a measure for the proportion of disease cases associated with the risk factor among members of the target population exposed to that risk factor, denoted here by λ_e (to mean attributable risk among exposed). He also proposed a measure of the proportion of disease cases associated with the risk factor among all members of the population, denoted by λ (to mean population attributable risk). These measures have also been developed by MacMahon and Pugh (1970), Cole and MacMahon (1971), Miettinen (1974) and Walter(1976). Alternative formulations as well as simplification have been produced by Levin and Bartel (1978), Leviton (1973) and Taylor (1977).

In Chapter 2, we shall show how these measures have been derived using case-control data. Methods of finding their large sample variances and confidence intervals will be indicated.

Recently it has been shown that the two measures of attributable risk, λ_e and λ, can be linked through the prevalence of the risk factor among the cases denoted V_x. As noted by Miettinen (1974) and also by

Kleinbaum, Kupper and Morgenstern (1982), it can be shown that

$$\lambda = V_x\lambda_e \tag{1.1}$$

The importance of the above expression is that both λ_e and V_x, and consequently λ, are usually estimable from both matched and unmatched case-control data alone. Kuritz and Landis (1987a), using this formulation, have proposed an estimator of attributable risk for matched-pair case-control data. Their work can therefore be extended to account for multiple matching, and to investigate in terms of cost efficiency the relative merits of multiple matching, if attributable risk estimation is the ultimate goal.

In Chapter 3, an alternative methodology leading to expressions for large sample variances of attributable risks will be shown. We also propose to indicate how these formulations work for some data given in the literaure where hitherto attributable risk measures were not found or were analysed as if matching had been ignored. Though no attempt will be made to discuss the relative merits of matching, it simply means that investigators can now find attributable risk measures for matched case-control data. The decision to match or not to match has been discussed by various authors including McKinlay (1977), Kupper et al (1981) and Schlesselman (1982).

In Chapter 4, we propose to find the asymptotic relative efficiency of different matching designs, in terms of cost of gathering cases and controls. Based on these relative efficiencies, a table of optimal ratios will be provided.

We conclude by discussing computational results, given in the appendix and possible recommendations.

The computations were carried out on the VAX 8600 computer at McMaster University.

CHAPTER 2

STATISTICAL FRAMEWORK

2.1 UNMATCHED ANALYSIS

Consider a set of data from a case-control study where the sample of cases and controls have been randomly selected from a population of interest. This formulation does not take into account confounding factors by extraneous variables, thus it will be carried out as if all confounding factors are under control. Measures of estimating attributable risks in the presence of confounding variables have been dicussed by Whittemore (1982, 1983), Walter (1976), Miettinen (1974) and Bruzzi et al (1985).

The data for such a case-control study has been displayed in the table below.

Table 1.1

Distribution of cases and controls with respect to risk factor status.

		Case	Control	Total
Risk Status	Exposed	a	b	m_1
	Non-exposed	c	d	m_2
Total		n_1	n_2	N

As is appropriate for case-control studies, n_1 and n_2 are considered fixed.

The usual Mantel-Haenszel estimator of the odds ratio (Mantel and Haenszel (1959)) would be given by:

$$\hat{\psi} = ad/bc \tag{2.1}$$

If we make a further assumption that the disease rate is rare, then the risk of the disease and the odds of the disease are virtually identical (Fleiss (1982)). Thus the odds ratio can be used to approximate relative risk in case-control studies, since relative risk cannot be directly estimated from case-control studies.

If we let I_e denote the incidence rate of a disease (proportion of new cases in a group of people who were initially free of the disease) in persons exposed to the risk factor, and I_o the incidence rate of disease persons not exposed to the risk factor, then the relative risk, R, is given by:

$$R = I_e/I_o \tag{2.2}$$

Thus $I_e - I_o$ is the excess risk among persons which may be attributed to exposure. Several methods of estimating attributable risk have been suggested. Berkson (1951) proposed a simple difference between the two incidence rates, $I_e - I_o$, as the measure. Sheps' relative difference (Sheps (1959))

$$(I_e - I_o)/(1 - I_o)$$

considered the component of the incidence among the exposed ascribed to the exposure. MacMahon and Pugh (1970) and Cole and MacMahon (1971) define attributable risk as the proportion of cases among persons exposed which are due to the exposure (denoted here by λ_e) and proposed the measure

$$\lambda_e = (I_e - I_o)/I_e \tag{2.3}$$

Expressed in terms of relative risk, $R=I_e/I_o$, the measure becomes

$$\lambda_e = (R - 1)/R \tag{2.4}$$

The measure developed by Levin (1953) provides an index of attributable risk in the population as opposed to the above, which shows such a measure in the exposed group. The population here refers to those persons who form the common sources of cases and controls in the study. Since controls should be representative of the unaffected (non-diseased) persons in the population, then data available from them can be used to estimate the proportion of the population exposed.

If we denote the proportion of population exposed to the risk by P_e, then the proportion of all cases (in both exposed and non-exposed groups) which are associated with exposure is given by λ (Levin's measure of population attributable risk), where

$$\lambda = P_e(R - 1)/[1 + P_e(R - 1)] \tag{2.5}$$

The rare disease assumption makes it possible to estimate the proportion in the population exposed by the proportion in the control exposed. This means that if the prevalence of disease is sufficiently low, then prevalence of exposure among non-diseased persons is very close to the prevalence of exposure among the selected controls. This becomes a very reasonable assumption if there is no control selection bias.

From our data in Table 1.1,

$$\hat{R} = ad/bc \qquad \text{and} \qquad \hat{P}_e = b/n_2$$

Thus,

$$\hat{\lambda}_e = (ad - bc)/ad \tag{2.6}$$

and,

$$\hat{\lambda} = 1 - cn_2/dn_1 \tag{2.7}$$

Using $\xi = 1 - \lambda$, Walter (1975) showed that ξ has, asymptotically, a log-normal distribution.

If the number of cases equals the number of controls (Taylor (1977)),

$$\hat{\lambda} = 1 - c/d \tag{2.8}$$

Using a different formulation proposed by MacMahon and Pugh (1970), where

$$\lambda' = (I_t - I_0)/I_t$$

and I_t is the incidence rate in the total population and I_0 the incidence rate in the unexposed, Leviton (1973) showed that for $I_t = P_eI_e + (1-P_e)I_0$,

$$\lambda' = \lambda = P_e(R - 1)/(1 + P_e(R - 1))$$

Another simplified formulation has been given by Levin and Bartel (1978), where

$$\hat{\lambda} = (a' - b')/(1 - b')$$

and a' is the proportion of cases exposed and b' is the proportion of controls exposed.

Inherent in all these formulations is the fact that the factors are associated with increased risk (ie $R > 1$ or $\psi > 1$). If the factors are associated with decreased risk ($\psi < 1$), we can talk about the 'Prevented Fraction' as being the proportion of the potential disease experience prevented by the factor (presumably a beneficial one). Walter (1976) has shown that such a prevented fraction of the risk denoted by λ_P, could be represented as

$$\lambda_P = \Phi(1 - \psi)$$

where Φ and ψ can be estimated by $\hat{\Phi} = b/n_2$ and $\hat{\psi} = ad/bc$ respectively. Thus

$$1 - \lambda_P = (1 - \lambda)^{-1}.$$

It is possible to construct confidence intervals for λ and λ_e. This is based on the assumption that the estimators or some transformations of them are approximately normally distributed.

Since ψ has been used to estimate R, we could find a confidence interval for λ_e based on the logarithm of ψ, where

$$\log_e(1 - \hat{\lambda}_e) = -\log_e\hat{\psi}$$

Thus

$$\mathrm{Var}(\log_e(1 - \hat{\lambda}_e)) = \mathrm{Var}(\log_e\hat{\psi}) \tag{2.9}$$

where

$$\mathrm{Var}(\log_e \hat{\psi}) = 1/a + 1/b + 1/c + 1/d \tag{2.10}$$

We could derive their asymptotic variance by using methods shown by Kendall and Stuart (1969) and also by Bishop, Fienberg and Holland (1975), concerning the asymptotic distribution of a smooth function F of multinomial proportions. Let $n_1, n_2, \ldots\ldots, n_K$ be the observed multinomial frequencies, $n = \sum n_K$ the total sample, $p_K = n_K/n$ the kth observed proportion, and P_K the expectation of p_K. If $F = F(p_1, p_2, \ldots, p_K)$ is a regular function of $p_1, p_2, \ldots, p_K$, and if n is large, then F is asymptotically normal with mean $F(P_1,P_2,\dots,P_K)$ and variance

$$Var(F) = (\sum p_K d_K^2 - (\sum p_K d_K)^2)/n$$

where

$$d_K = \partial F(p_1,p_2,\dots,p_K)/\partial p_K$$

In fact, Fleiss (1982) has shown that if

$F(p_1,p_2,\dots,p_K) = F(n_1,n_2,\dots,n_K)$;   or

$F(cx_1,cx_2,\dots,cx_K) = F(x_1,x_2,\dots,x_K)$, (for all nonzero c), then the above variance formula can be reduced to

$$Var(F) = (\sum p_K d_K^2)/n$$

since $\sum p_K d_K = 0$ in this case. The odds ratio, attributable risk and smooth functions of them are all examples of such a function F (Fleiss (1982)). It can then be shown that

$$Var(\hat{\psi}) = (ad/bc)^2(1/a + 1/b + 1/c + 1/d) \tag{2.11}$$

$$\text{or} \qquad Var(\log_e\hat{\psi}) = (1/\hat{\psi})^2 Var(\hat{\psi}) \tag{2.12}$$

Walter (1978) has shown that

$$Var(\hat{\lambda}) = (cn_2/dn_1)^2(a/cn_1 + b/dn_2) \tag{2.13}$$

We have indicated that the two measures of attributable risk, λ_e and λ, can be linked.

Using the case-control data from Table 1.1, we have

$$\lambda = P_e(R-1)/(1+P_e(R-1)) = ((R-1)/R)P_eR/(1+P_e(R-1)) = \lambda_e P_e R/(1+P_e(R-1))$$

Using the following estimates

$$\hat{R} = ad/bc \quad \text{and} \quad \hat{P}_e = b/n_2$$

we can show that

$$PeR/(1+Pe(R-1)) = a/n_1 \tag{2.14}$$

Since cases are assumed to be a random sample of disease cases from some population of interest, a/n_1 represents the sample exposure prevalence among the cases. Denoting a/n_1 by V_x, we have

$$\hat{\lambda} = \hat{V}_x\hat{\lambda}_e$$

where V_x and λ_e are both estimable from the case-control data under reasonable assumptions.

Confidence intervals could be found for λ_e and λ, (or their log-transformation). We note that ψ is the maximum likelihood estimator of the odds ratio. Leung and Kupper (1981) have shown that when the actual attributable risk is between 0.21 and 0.79, the width of the log-transformation based interval is less than that for the maximum likelihoood based interval. Similar results were also obtained by Whittemore (1982).

2.2 MATCHED ANALYSIS

Kuritz and Landis (1987a) using the fact that $\lambda = V_x\lambda_e$, have given explicit formulas for estimating attributable risk among the exposed and the population attributable risk.

Consider data from a matched-pair case-control study given in the table below.

Table 2.2

		Control		
		Exposed	Non-exposed	Total
Cases	Exposed	a	b	a+b
	Non-exposed	c	d	c+d
Total		a+c	b+d	n

Using the fact that attributable risk among exposed is a direct function of the odds ratio, and population attributable risk is also a direct function of odds ratio and exposure prevalence among the cases only, we find these two measures of attributable risk.

The assumptions necessary in this regard are that:

(a) the prevalence of the disease among the population is rare;

(b) the cases constitute a random sample of cases in the population.

For a matched pair design, the estimator of odds ratio depends only on the exposure-discordant pairs. The odds ratio estimator has been given by Mantel and Haenszel (1959) as

$$\hat{\psi} = b/c \qquad (2.15)$$

and its asymptotic variance has been shown by Ejigou and McHugh (1977) to be

$$\mathrm{Var}(\hat{\psi}) = (\hat{\psi})^2(1/b + 1/c)^2 \qquad (2.16)$$

or $\mathrm{Var}(\log_e \hat{\psi}) = (1/\hat{\psi})^2\mathrm{Var}(\hat{\psi})$

In this case, ψ is both the maximum likelihood estimator and the Mantel-

Haenszel estimator for the odds ratio. It follows from above that

$$\hat{\lambda}_e = 1 - 1/\hat{R} = (b - c)/b \tag{2.17}$$

Since the n cases among the matched pairs is assumed to constitute a random sample of X cases from the target population, V_X can be estimated by

$$\hat{V}_X = (a + b)/n \tag{2.18}$$

Hence an estimate of the population attributable risk is given by

$$\hat{\lambda} = (a + b)(b - c)/bn \tag{2.19}$$

Generalisation to multiple controls per case

With the availability of an estimator for odds ratio in multiple matching (Mantel and Haenszel (1959), Miettinen (1970) among others), it is possible to extend the idea to take care of matching with multiple controls per case.

Consider the situation where two controls are matched to a case. Following the notation of Connett et al (1982), we can represent the data as:

Table 2.3

		Number of Controls exposed			Total
		2	1	0	
Cases	Exposed(1)	Z_{12}	Z_{11}	Z_{10}	$\sum_t Z_{1t}$
	Non-exposed(0)	Z_{02}	Z_{01}	Z_{00}	$\sum_t Z_{0t}$
					N

The Mantel-Haenszel estimator could be generalised as

$$\hat{\psi} = \sum_t (R - T) Z_{1t} \;/\; \sum_t T Z_{0t} \qquad (2.20)$$

where R is the number of controls matched to a case;

T is the number of units from R matched to a selected unit; thus $0 \leq T \leq R$.

N is the total number of cases.

Z_{ij} is the number of matched groups (in each $(1 + R)$ tuple).

The basic design considered here is that of a case-control study with multiple matching (with fixed matching ratio) on a confounding variable of healthy control to diseased cases.

For example, consider the following eight outcomes for a 2-to-1 matching (Schlesselman (1982)), presented as:

Table 2.4

Case	Control		Frequency
	1	2	# of matched groups
+	+	+	n_0
+	+	-	n_1
+	-	+	n_2
+	-	-	n_3
-	+	+	n_4
-	+	-	n_5
-	-	+	n_6
-	-	-	n_7

with dichotomous exposure (exposed +), (non-exposed -).

By regarding each triplet as a separate subgroup, the Mantel-Haenszel estimate of the odds ratio could be found.

It is given by

$$\hat{\psi} = (n_1 + n_2 + 2n_3)/(2n_4 + n_5 + n_6)$$

or in the Connett et al (1982) notation

$$\hat{\psi} = (2Z_{10} + Z_{11})/(Z_{01} + 2Z_{02}) \qquad (2.21)$$

Thus, by ignoring ordering of the two matched controls, we would only have six (6) possible 2x2 tables for each triplet.These could be put in a single table as the one already shown in the Table 2.3, where $n_0 = Z_{12}$, $n_1 + n_2 = Z_{11}$, $n_3 = Z_{10}$, $n_4 = Z_{01}$, $n_5 + n_6 = Z_{02}$ and $n_7 = Z_{00}$. Miettinen (1970) obtained a conditional maximum likelihood estimator of this estimate as

$$\hat{\rho} = (4Z_{10}-Z_{01}+Z_{11}-4Z_{02})/4(Z_{01}+Z_{02}) + \sqrt{}\{[(4Z_{10}-Z_{01}+Z_{11}-Z_{02})/4(Z_{01}+Z_{02})]^2$$
$$+ (Z_{10}+Z_{11})/(Z_{01}+Z_{02})\}$$

Ejigou and McHugh (1981) also gave an unconditional estimate of the odds ratio as

$$\hat{\psi}_e = [Z_{01}Z_{10}(Z_{11}+Z_{02}) + Z_{11}Z_{02}(Z_{10}+Z_{01})]/[Z_{01}^2(Z_{11}+Z_{02})/2 + 2Z_{02}^2(Z_{10}+Z_{01})]$$

In a recent simulation study, Donner and Hauck (1986) found that the Mantel-Haenszel estimator of odds ratio, ψ, compares favourably to both the odds ratio estimator obtained by the maximum likelihood method and the conditional maximum likelihood method with respect to bias and precision over a wide range of fixed stratum designs likely to occur in practice. For example, when $\psi \leqslant 5$, they found that the relative efficiency never falls below 0.93. We have therefore employed the ψ in the calculation of measures for attributable risk. Despite the fact that ψ may fare badly in certain situations against the other estimators (for example, when the odds ratio ψ is large (approaching 10) and $V_x \leqslant 0.3$), ψ has proved useful in many situations as is evident by its wide usage over the years in the medical literature. The Mantel-Haenszel estimator is easy to use and has been shown to be a reasonable and efficient estimate, and thus a good alternative to the maximum likelihood estimator (MLE). The only problem with MLE is calculation, since it has to be obtained by iteration (Breslow (1981).

Estimators for attributable risk measures, λ_e and λ, from data obtained through case-control studies where one or more controls have been matched to each case can now be obtained. According to Kuritz and Landis (1987b) , the sampling design for obtaining these matched data could be conceptualised as a simple random sample of cases being equivalent to a random sample of matched sets. Thus, by combining information across strata determined by the matched sets, this approach provides for the attributable risk estimate all the benefits associated with the Mantel-Haenszel procedure.

We can therefore obtain for the case of 2-to-1 matching, an estimate of attributable risk among exposed persons as

$$\hat{\lambda}_e = (\hat{R} - 1)/\hat{R} = (2Z_{10} + Z_{11} - Z_{01} - 2Z_{02})/(2Z_{10} + Z_{11}) \qquad (R>1)$$

An estimate of exposure prevalence among cases only is given by:

$$\hat{V}_X = (Z_{10} + Z_{11} + Z_{12})/N$$

Hence

$$\hat{\lambda} = \hat{V}_X\hat{\lambda}_e = (Z_{10}+Z_{11}+Z_{12})(2Z_{10}+Z_{11}-Z_{01}-2Z_{02})/N(2Z_{10}+Z_{11})$$

We can thus show that for R-to-1 matching and with dichotomous exposure levels, the attributable risk measures generalise to:

$$\hat{\lambda}_e = \{ \sum_t(R - T)Z_{1t} - \sum_t TZ_{0t} \} / \sum_t(R - T)Z_{1t} \qquad (2.22)$$

$$\hat{\lambda} = \{ \sum_t Z_{1t} \}\{ \sum_t(R - T)Z_{1t} - \sum_t TZ_{0t} \} / N\sum_t(R - T)Z_{1t} \qquad (2.23)$$

Generalisation to variable number of controls matched per case

By making use of Fleiss' (1984) method of determining ψ when matching a varying number of controls per case, we could still obtain estimates of λ_e and λ.

Let r denote the number of controls matched to a particular case, where r may vary from 1 (matched pair) to a high of R (R-to-1 matching). The analysis begins with stratification of the cases according to a particular value of r. Thus for each value of r ($1 \leqslant r \leqslant R$), we look at the table as if there were only r controls per case, where r is fixed.

For a particular r, we could present the table as:

Table 2.5

	No. of controls exposed			
Status of case	r	r-1	1	0
Exposed (1)	$Z_{1r}^{(r)}$	$Z_{1r-1}^{(r)}$	$Z_{11}^{(r)}$	$Z_{10}^{(r)}$
Non-exposed(0)	$Z_{0r}^{(r)}$	$Z_{0r-1}^{(r)}$	$Z_{01}^{(r)}$	$Z_{00}^{(r)}$

where $Z_{1j}^{(r)}$ refers to the number of matched sets with r controls in which both the case and exactly j of the controls are exposed;

and $Z_{0j}^{(r)}$ refers to the number of matched sets with case unexposed and exactly j controls exposed.

Following Fleiss' (1984) notation, if we define

$$A^{(r)} = \sum_{j=0}^{r} (r-j) Z_{1j}^{(r)}/(r+1)$$

and

$$B^{(r)} = \sum_{j=0}^{r} j Z_{0j}^{(r)}/(r+1)$$

then

$$\hat{\psi} = \sum_{r=1}^{R} A^{(r)} / \sum_{r=1}^{R} B^{(r)} \qquad (2.24)$$

Hence, for a variable number of controls per case

$$\hat{\lambda}_e = (\sum_{r=1}^{R} A^{(r)} - \sum_{r=1} \overset{R}{B}^{(r)} \}/ \sum_{r=1}^{R} A^{(r)}, \qquad \psi > 1 \qquad (2.25)$$

and

$$\hat{\lambda} = \hat{V}_x \hat{\lambda}_e = \sum_{r=1}^{R} \sum_{j=0}^{r} Z_{1j}^{(r)} (\sum_{r=1}^{R} A^{(r)} - \sum_{r=1}^{R} B^{(r)})/N \sum_{r=1}^{R} A^{(r)}$$

$$\text{where } \hat{V}_x = \sum_{r=1}^{R} \sum_{j=0}^{r} Z_{ij}^{(r)}/N \qquad (2.26)$$

2.3 Large Sample Variance Estimation

As already indicated in Section 2.1, attributable risk estimators for matched set case-control data could be regarded as regular functions of a set of multinomial proportions (Kuritz and Landis (1987b)). It is therefore possible to obtain direct expressions for their large sample variances by applying multivariate version of the delta method as described in Section 14.6 of Bishop, Fienberg and Holland (1975).

In general, these asymptotic variances can be expressed as

$$\text{Var}(F) = (\sum p_i d_i^2 - (\sum p_i d_i)^2)/n$$

where F is a regular function of the multinomial proportions and

$$d_i = \partial F(p_1, p_2, \ldots, p_K)/\partial p_i$$

is the partial derivative of the function with respect to the i-th multinomial proportion.

Estimators for the Mantel-Haenszel odds ratio and attributable risks are part of the wide class of functions where

$$\sum p_i d_i = 0$$

Then, as noted by Fleiss (1982), the estimated large sample variance for these functions simplifies to

$$\mathrm{Var}(F) = (\sum p_i d_i^2)/n$$

For a fixed matching ratio (ie matching with exactly k controls to a case) with dichotomous exposure level, Connett et al (1982) have shown that

$$\mathrm{Var}(\log_e \hat{\psi}) = \{ \sum (R - T)^2 Z_{1t} + \hat{\psi}^2 \sum T^2 Z_{0t} \} / \{ \sum (R - T) Z_{1t} \}^2 \qquad (2.27)$$

Breslow (1981) gave a similar expression. Consider a data set as being given as:

Table 2.6

	Exposed	Non-exposed	Total
Cases	A_i	B_i	N_{1i}
Controls	C_i	D_i	N_{2i}
Total	M_{1i}	M_{2i}	N_i

Thus $\psi = \sum U_i / \sum V_i$, $i = 1,2,\ldots,n$; where

$U_i = A_i D_i / N_i$ and $V_i = B_i C_i / N_i$.

When the number of tables is large in comparison with each of the individual table totals N_i, then Breslow showed that an appropriate estimator of the variance of ψ is

$$\mathrm{Var}(\hat{\psi}) = \sum (U_i - \hat{\psi}V_i)^2 / (\sum V_i)^2$$

and

$$\mathrm{Var}(\log_e \hat{\psi}) = \sum(U_i - \hat{\psi}V_i)^2 / (\sum U_i)^2 \tag{2.28}$$

For a R-to-1 matching, $N_{1i}=1$, $N_{2i}=R$ for $i=1,2,\ldots,n$. There are Z_{0t} matched sets for which $A_i=0$, $C_i=T$ and for each of these $U_i=0$ and $V_i=T/(R+1)$. Similarly, there are Z_{1t} matched sets for which $A_i=1$, $C_i=T$, $U_i=(R-T)/(R+1)$ and $V_i=0$. Substitution of these values in the above expression (2.28) gives the previous result (2.27).

Fleiss (1984) showed that for matched case-control data with a variable number of controls per case, an estimator for the variance of the Mantel-Haenszel estimator of odds ratio is given by:

$$\mathrm{Var}(\log_e \hat{\psi}) = \sum C^{(r)} / (\sum A^{(r)})^2 \tag{2.29}$$

where

$C^{(r)} = (1/r+1)^2\{ \sum(r-j)^2 Z_{1j}^{(r)} + \hat{\psi}^2 \sum j^2 Z_{0j}^{(r)} \}$ and $A^{(r)}$ as defined in (2.24)

By noting that

$$\log_e(1-\hat{\lambda}_e) = -\log_e \hat{\psi} \tag{2.30}$$

we have

$$\mathrm{Var}(\log_e(1-\hat{\lambda}_e)) = \mathrm{Var}(\log_e \hat{\psi}) \tag{2.31}$$

which gives

$$\mathrm{Var}(\hat{\lambda}_e) = (1-\hat{\lambda}_e)^2 \mathrm{Var}(\log_e \hat{\psi}) \tag{2.32}$$

As already indicated, the estimator for λ is the product of two terms, V_x and λ_e. Since V_x is a binomial proportion , its variance is given as:

$$\mathrm{Var}(\hat{V}_x) = V_x(1-V_x)/N \qquad (2.33)$$

Thus a large sample variance formula for the estimate of population attributable risk can be formed by writing λ as a regular function of the multinomial proportions and using the delta method to obtain it. Kuritz and Landis (1987b) have shown that

$$\mathrm{Var}(\hat{\lambda}) = (V_x)^2\mathrm{Var}(\lambda_e) + (\lambda_e)^2\mathrm{Var}(V_x) + 2\lambda(1 - \lambda_e)/N \qquad (2.34)$$

As a by-product it is useful to note that

$$\mathrm{Cov}(\hat{V}_x,\hat{\lambda}_e) = (1 - \lambda_e)/N$$

recognising that the expression for $\mathrm{Var}(\lambda)$ is in the usual form of a Taylor series expansion for the product of the two terms of the expression.

Since the delta method assumes that F (the function of multinomial proportions) is asymptotically normally distributed, confidence intervals for these measures of attributable risk can readily be computed once their standard errors have been found.

For instance, an approximate 95% confidence interval for λ_e would be given by:

$$\hat{\lambda}_e \pm 1.96\mathrm{se}(\hat{\lambda}_e)$$

and that for λ:

$$\hat{\lambda} \pm 1.96\mathrm{se}(\hat{\lambda})$$

We shall illustrate the methods described in this chapter with examples from the literature in Section 3.2.

CHAPTER 3

MATRIX FORMULATION AND EXAMPLES

3.1 MATRIX FORMULATION FOR VARIANCES OF ATTRIBUTABLE RISKS

An alternative method for finding large sample variances for λ_e and λ, based on matrix formulation when these measures of attributable risk are expressed as compounded functions of sample multinomial proportions, can be found using the method for analysing multivariate categorical data outlined in the Appendix of Koch et al (1977). This method leads to explicit expressions for Var(λ_e) and Var(λ) after simplification, up to matching with two controls per case, but the algebra gets messy for more than two controls per case and for matching with variable numbers of control per case. However, a computer could be used in such situations.

Consider a sample of multinomial frequencies arising from a case-control study with R controls to a case, where the data set could be represented as:

Table 3.1

	No. of controls exposed				
Status of case	R	R-1	1	0	Total
Exposed (1)	Z_{1R}	Z_{1R-1}	Z_{11}	Z_{10}	$\sum_t Z_{1t}$
Non-exposed(0)	Z_{0R}	Z_{0R-1}	Z_{01}	Z_{00}	$\sum_t Z_{0t}$
					N

The vector $\mathbf{Z_{rs}}$ will be assumed to follow the multinomial distribution with parameters **N** and Π_{ij}. A vector of multinomial proportions, **p**, for the data set above will be given by:

$$p' = [Z_{1R},....,Z_{10},Z_{0R},...,Z_{00}]/N$$

where **p** is a 2(R+1)x1 column vector and is the maximum likelihood estimator of Π and $'$ denotes the matrix transpose.

As outlined in the appendix of Koch et al (1977), a consistent estimator for the covariance matrix of **p**, is given by a (2R+2)x(2R+2) covariance matrix **Var(p)**, such that

$$\mathbf{Var(p)} = [\mathbf{D_p} - \mathbf{pp'}]/N \qquad (3.1)$$

where $\mathbf{D_p}$ represents a diagonal matrix with the vector **p** on the main diagonal.

By taking F(p) as a compounded logarithmic—exponential—linear function of observed proportions leading to relationships of interest (λ_e and λ), then an estimate of the variance of F could be found.

Consider a class of functions that can be expressed in terms of a sequence of the matrix operations:

(i) Linear transformation of the type

$$F_1(p) = \mathbf{A_1 p} = a_1$$

where A_1 is a matrix of known constants.

(ii) Logarithmic transformation of the type

$$F_2(p) = \log_e (p) = a_2$$

(iii)Exponential transformation of the type

$$F_3(p) = \exp(p) = a_3.$$

Thus a linearized Taylor-series based estimate of the variance of F (Koch et al (1977)) is given by

$$\mathrm{Var}(F) = \mathbf{H}[\mathrm{Var}(\mathbf{p})]\mathbf{H}' \qquad (3.2)$$

where **H** is the first derivative matrix for the corresponding compounded function and F is assumed to have a continuous partial derivative through order 2 with respect to **p**.

For example, let

$$\mathbf{a}_1 = \mathbf{A}_1\mathbf{p} \qquad \mathbf{a}_2 = \exp\{\mathbf{A}_2(\log_e \mathbf{a}_1)\}$$

then a consistent estimate of the variance (3.2) is obtained by the application of the chain rule for matrix differentiation leading to

$$\mathbf{H} = \mathbf{D}\mathbf{a}_2\mathbf{A}_2\mathbf{D}\mathbf{a}_1^{-1}\mathbf{A}_1$$

where $\mathbf{Da}_1$ and $\mathbf{Da}_2$ are diagonal matrices with the vectors $\mathbf{a}_1$ and $\mathbf{a}_2$ on their main diagonals respectively.

Consider a matched-pair case-control data where the data are represented as:

		Control	
		+	-
Case	+	a	b
	-	c	d
			n

exposed (+), non-exposed (-).

where 'a' denotes the number of pairs in which both members are exposed, 'b' the number of pairs where only the case is exposed, etc.

In this case the attributable risk among exposed, λ_e, and population attributable risk, λ, could be expressed as compounded functions of $\mathbf{p}$, where

$$p' = [a,b,c,d]/n$$

as

$$\hat{\lambda}_e = \exp\{\mathbf{A}_2(\log_e \mathbf{a}_1)\}$$

and

$$\hat{\lambda} = \exp\{\mathbf{B}_2(\log_e \mathbf{b}_1)\}$$

where $\mathbf{a}_1 = \mathbf{A}_1\mathbf{p}$, $\mathbf{A}_2 = [1,-1]$, $\mathbf{b}_1 = \mathbf{B}_1\mathbf{p}$, $\mathbf{B}_2 = [1,1,-1]$

and

$$\mathbf{A}_1 = \begin{bmatrix} 0 & 1 & -1 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} \qquad \mathbf{B}_1 = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}$$

Thus, by using appropriate matrix products arising from multivariate Taylor-series methods, the large sample variances for these estimators are obtained as

$$\mathrm{Var}(\hat{\lambda}_e) = (\hat{\lambda}_e)^2\mathbf{A}_2\mathbf{D}\mathbf{a}_1{}^{-1}\mathbf{A}_1\mathbf{Var(p)}\mathbf{A}_1{}'\mathbf{D}\mathbf{a}_1{}^{-1}\mathbf{A}_2{}' \tag{3.3}$$

with H here being defined as $\mathbf{H} = \hat{\lambda}_e\mathbf{A}_2\mathbf{D}\mathbf{a}_1{}^{-1}\mathbf{A}_1$; and

$$\mathrm{Var}(\hat{\lambda}) = (\hat{\lambda})^2\mathbf{B}_2\mathbf{D}_{b1}{}^{-1}\mathbf{B}_1\mathbf{Var(p)}\mathbf{B}_1{}'\mathbf{D}_{b1}{}^{-1}\mathbf{B}_2{}' \tag{3.4}$$

where $\mathbf{H} = \hat{\lambda}\mathbf{B}_2\mathbf{D}_{b1}{}^{-1}\mathbf{B}_1$ and $\mathbf{D_p}$ is a diagonal matrix with the vector $\mathbf{p}$ on the main diagonal.

Thus for

$$\hat{\lambda}_e = (b-c)/c \text{ and } \hat{\lambda} = (a+b)(b-c)/bn$$

we have equations (8) and (9) of Kuritz and Landis (1987a),

$$\mathrm{Var}(\hat{\lambda}_e) = (1/b)^2[c(b+c)/b] \tag{3.5}$$

$$\mathrm{Var}(\hat{\lambda}) = (1/bn)^2[a(b-c) + (b^2+ac)/b + c(a+b)^2 - (a+b)^2(b-c)^2/n] \qquad (3.6)$$

In the case of matching 2 controls to a case, if the data are represented as

Table 3.2

		No. of controls exposed		
		2	1	0
Case	Exposed	a	b	c
	Non-exposed	d	e	f
Total				n

It can be shown from (2.22) and (2.23) of Chapter 2 that,

$$\hat{\lambda}_e = (b+2c-2d-e)/(b+2c)$$

$$\hat{\lambda} = (a+b+c)(b+2c-2d-e)/(b+2c)n$$

and by simplifying (A.1) and (A.2) of Appendix A, we can show that

$$\mathrm{Var}(\hat{\lambda}_e) = (1/(b+2c))^4[(b+2c)^2(e+4d) + (e+2d)^2(b+4c)] \qquad (3.7)$$

$$\begin{aligned}\mathrm{Var}(\hat{\lambda}) = {} & [1/(n(b+2c))]^2\{(a+b+c)(b+2c-2d-e)^2 - (a+b+c)^2(b+2c-2d-e)^2/n \\ & + (a+b+c)^2(e+4d) + (a+b+c)^2(2d+e)^2(b+4c)/(b+2c)^2 \\ & + 2(a+b+c)(2d+e)(b+2c-2d-e)\} \qquad (3.8)\end{aligned}$$

(See Appendix A)

As noted by Kuritz and Landis (1987a), further empirical work is necesary to investigate the small to moderate sample size behaviour of these estimators. We realise that these asymptotic formulas have a critical link to the magnitude of the frequency of exposure-discordant cells (ie b and c for the matched-pair situation, and b+2c and 2d+e for the 2 controls per case.)

Based on the fact that these estimators are asymptotically normally distributed, approximate 95 % confidence intervals for these attributable risk estimates can be found as:

$$\hat{\lambda}_e \pm 1.96se(\hat{\lambda}_e)$$

$$\hat{\lambda} \pm 1.96se(\hat{\lambda})$$

It is again worthwhile to point out that these formulas were obtained on the assumption that the relative risk (or odds ratio in this case) was greater than unity. Otherwise, we would have to define a 'Prevented Fraction' for a relative risk less than 1. If we have reason to believe that the actual relative risk is greater than 1, but very close to 1, and sampling variation causes our estimate to be less than 1, then we could define the attributable risk estimate as zero.

The above formulation could be applied in the case where variable number of controls have been matched to the cases. Such a situation may arise out of the study protocol (Walter (1980)) eg. if all siblings are to be used as controls; or from practical difficulties (eg information on one of the controls is not available (refused interview, moved away etc).

As indicated in Chapter 2, let R be the maximum number of controls matched to a case. Let the vector of multinomial proportions be given by p such that:

$p' = [p_1^{(1)}, p_1^{(2)},..., p_1^{(r)}, p_0^{(1)}, p_0^{(2)},..., p_0^{(r)}]/n$ for a particular r, $1 \leq r \leq R$; where

$p_1^{(r)} = [Z_{1j}^{(r)}]/n$ and $p_0^{(r)} = [Z_{0j}^{(r)}]/n$, $j=1,2,....,r$; and n is the total number of matched sets.

For example, consider a variable number of controls per case situation, where the maximum number of controls matched to a case is 3. The data for such a situation could be represented as follows:

Table 3.3

		1 control observed		2 controls observed			3 controls observed			
		No. exposed		No. exposed			No. exposed			
		1	0	2	1	0	3	2	1	0
Case	1	$Z_{11}^{(1)}$	$Z_{10}^{(1)}$	$Z_{12}^{(2)}$	$Z_{11}^{(2)}$	$Z_{10}^{(2)}$	$Z_{13}^{(3)}$	$Z_{12}^{(3)}$	$Z_{11}^{(3)}$	$Z_{10}^{(3)}$
	0	$Z_{01}^{(1)}$	$Z_{00}^{(1)}$	$Z_{02}^{(2)}$	$Z_{01}^{(2)}$	$Z_{00}^{(2)}$	$Z_{03}^{(3)}$	$Z_{02}^{(3)}$	$Z_{01}^{(3)}$	$Z_{00}^{(3)}$

Here $p_1' = [Z_{11}^{(1)}, Z_{10}^{(1)}, Z_{12}^{(2)}, Z_{11}^{(2)}, Z_{10}^{(2)}, Z_{13}^{(3)}, Z_{12}^{(3)}, Z_{11}^{(3)}, Z_{10}^{(3)}]/n$

and $p_0' = [Z_{01}^{(1)}, Z_{00}^{(1)}, Z_{02}^{(2)}, Z_{01}^{(2)}, Z_{00}^{(2)}, Z_{03}^{(3)}, Z_{02}^{(3)}, Z_{01}^{(3)}, Z_{00}^{(3)}]/n$

where $p' = [p_1' \vdots p_0']$.

Define

$$a_1 = A_1 p, \quad \text{where} \quad A_1 = \begin{bmatrix} C_1 - C_0 \\ C_1 \end{bmatrix}$$

and $C_1 = [M_1 \vdots 0]$ with M_1 having elements $(r-j)/(r+1)$ for each r, $r=1,2,\ldots,R$, and $j=0,1,\ldots,r$. Also, $C_0 = [0 \vdots M_0]$ with M_0 having elements $r/(r+1)$ for each r, $r=1,2,\ldots,R$, and $j=0,1,2,\ldots,r$.

Here the row vectors 0, M_0 and M_1 have the same number of columns as both p_1' and p_0'. 0 is a row vector where each element is zero.

We thus realise that C_1p is the numerator of the odds ratio as defined by Fleiss (1984), and C_0p is the denominator.

From the example above,

C_1 = $[0, {}^1/_2, 0, {}^1/_3, {}^2/_3, 0, {}^1/_4, {}^2/_4, {}^3/_4, 0, 0, 0, 0, 0, 0, 0, 0, 0]$

C_0 = $[0, 0, 0, 0, 0, 0, 0, 0, 0, {}^1/_2, 0, {}^2/_3, {}^1/_3, 0, {}^3/_4, {}^2/_4, {}^1/_4, 0]$

We can therefore write attributable risk among cases as a compounded function

$$\hat{\lambda}_e = \exp[\mathbf{A_2}\log_e(\mathbf{A_1}\mathbf{p})]$$

Similarly, the estimator for population attributable risk, λ, could be found by defining:

$$\mathbf{B_1} = \begin{bmatrix} C_2 \\ C_1 - C_0 \\ C_1 \end{bmatrix}$$

$\mathbf{b_1} = \mathbf{B_1}\mathbf{p}$; $C_2 = [1 \vdots 0]$; and $\mathbf{B_2} = [1, 1, -1]$. Both 0 and 1 have the same number of columns as $\mathbf{p_1}'$ and $\mathbf{p_0}'$.

Thus, the compounded function of population attributable risk is given by:

$$\hat{\lambda} = \exp[\mathbf{B_2}\log_e(\mathbf{B_1}\mathbf{p})]$$

We also realise from the above expression that $V_x = \mathbf{C_2 p}$

Asymptotic large sample variances can now be obtained by using appropriate matrix products arising from multivariate Taylor series method. As already indicated, the large sample variance estimators are given by:

$$\mathrm{Var}(\hat{\lambda}_e) = \mathbf{H_1}\mathbf{Var(p)}\mathbf{H_1}'$$

$$\mathrm{Var}(\hat{\lambda}) = \mathbf{H_2}\mathbf{Var(p)}\mathbf{H_2}'$$

where in this case

$$\mathbf{H_1} = \hat{\lambda}_e \mathbf{A_2}\mathbf{D}\mathbf{a_1}^{-1}\mathbf{A_1} \qquad \text{and} \qquad \mathbf{H_2} = \hat{\lambda}\mathbf{B_2}\mathbf{D_{b1}}^{-1}\mathbf{B_1}$$

and $\mathbf{D_p}$ is a diagonal matrix with the vector $\mathbf{p}$ on the main diagonal.

3.2 EXAMPLES

We shall illustrate the methodolgy developed in both Chapter 2 and Section 3.1 with examples from the literature.

1. Using Matched-pair Case-Control Data

Consider the data in Table 3.4 which were obtained from a matched-pair case-control study on the exposure to oral conjugated estrogens among cases of endometrial cancer and their matched controls (on the basis of sex, race, date of admission, and hospital of admission) reported by Autunes et al (1979) and used by Schlesselman (1982) as an example.

Table 3.4

Frequency of Exposure among cases and their matched controls.

		Control		
		+	-	Total
Cases	+	12	43	55
	-	7	121	128
Total		19	164	183

[exposed (+), non-exposed (-)].

Using the methodology described in Chapter 2, an estimate of the odds ratio (which in this case happens to be the estimate of the relative risk) is given by (2.15):

$$\hat{\psi} = 6.143$$

The estimate of attributable risk among exposed is given by (2.17):

$$\hat{\lambda}_e = 0.837$$

The estimate of the exposure prevalence among cases only is given by (2.18):

$$\hat{V}_x = 0.3005$$

and thus the estimate of population attributable risk would be:

$$\hat{\lambda} = \hat{V}_x\hat{\lambda}_e = 0.252$$

Estimates of their large sample variances could be obtained by using the formulas indicated in Chapter 2.

From (2.16) $\quad \mathrm{Var}(\log_e \hat{\psi}) = 0.1666666.$

From (2.32) $\quad \mathrm{Var}(\hat{\lambda}_e) = 0.0044021.$

Using (2.33) and (2.34), $\quad \mathrm{Var}(\hat{\lambda}) = 0.0016505.$

To compare with the method developed from using the matrix method, we have from (3.5) $\quad \mathrm{Var}(\hat{\lambda}_e) = 0.0044021$;

and from (3.6) $\quad \mathrm{Var}(\hat{\lambda}) = 0.0016505.$

The two methods yield the same estimates for $\mathrm{Var}(\lambda_e)$ and $\mathrm{Var}(\lambda)$.

It is, however, interesting to compare the results of the data above to the case when it was analysed as if matching had been ignored as reported by Schlesselman (1982) in an illustrative example.

Table 3.5

Use of oral conjugated estrogens (OCE) for cases of endometrial cancer and controls.

		Cases	Controls	Total
OCE	Yes	55	19	74
	No	128	164	292
		183	183	366

The following results are obtained:

$\hat{\psi} = 3.709$ from (2.1)

$\hat{\lambda}_e = 0.730$ from (2.4)

$\hat{\lambda} = 0.220$ from (2.5)

$\mathrm{Var}(\hat{\psi}) = 1.165492$ from (2.11)

$\mathrm{Var}(\hat{\lambda}_e) = 0.0061591$ from (2.32)

$\mathrm{Var}(\hat{\lambda}) = 0.0018160$ from (2.13)

$\mathrm{Var}(\log_e \hat{\psi}) = 0.0847235$ from(2.10)

We realise that the odds ratio in the unmatched case is about half that obtained in the matched case, which highlights the importance of matched analysis in this case. Estimates of attributable risk were also underestimated in the unmatched analysis. For example, we would conclude that the sample attributable risk estimate of 0.837 (matched analysis) among exposed suggests that among women who took oral conjugated estrogens (OCE), 83.7 % of endometrial cancer were associated with such risk. We would however conclude that 73 % of the cancer were associated with the risk of using OCE in the unmatched analysis. This is due to the sizable reduction of the exposure-disease association (as measured by the odds ratio) when we do not take matching into account.

2. The case of matching 2 controls to a case.

Consider the data below which were obtained from a matched case-control study of the smoking habits (exposure) of some bladder cancer patients and their matched controls (each case matched to 2 controls on the basis of sex and age within 10 years) reported by Miller et al (1978). Writing the data to conform to our formulation in Table 2.3, we have :

Table 3.6

Smoking Habit in cases	Both controls 20+ a day	One of 2 controls ⩽ 20 a day	Both controls ⩽ 20 a day	Total
20+ per day (exposed)	31	42	17	90
⩽ 20 per day (not/less exposed)	11	23	12	46
				136

From (2.21), an estimate of the Mantel-Haenszel odds ratio for this data set is given by

$$\hat{\psi} = 1.689$$

and from (2.22) $\hat{\lambda}_e = 0.408$

and from (2.23) $\hat{\lambda} = 0.270$

where an estimate of the exposure prevalence among cases only from (2.26) is $\hat{V}_X = 0.662$.

We shall now find their asymptotic large sample variances using the methodology previously described. Using the formula given by Connett et al (1982), we have from (2.27) that

$$\text{Var}(\log_e \hat{\psi}) = 0.05213074$$

It follows from (2.31) that $\text{Var}(\hat{\lambda}_e) = 0.01827645$

From (2.33) $\text{Var}(\hat{V}_X) = 0.00164582$

and it follows from (2.34) that $\text{Var}(\hat{\lambda}) = 0.01062808$

We could also obtain estimates of the large sample variance using the multivariate Taylor series methods described by Koch et al (1977).

We could write the vector of multinomial proportions, **p** , as

$$\mathbf{p}' = [31,\ 42,\ 17,\ 11,\ 23,\ 12]/136$$

Then a consistent estimator of covariance of **p** is given

$$\textbf{Var(p)} = 1/N[\mathbf{D_p} - \mathbf{pp}']$$

which could be written as

$$\textbf{Var(p)} = \frac{1}{(136)^3}\begin{bmatrix} 3255 & -1302 & -527 & -341 & -713 & -372 \\ -1302 & 3948 & -714 & -462 & -966 & -504 \\ -524 & -714 & 2023 & -187 & -391 & -204 \\ -324 & -462 & -187 & 1375 & -253 & -132 \\ -714 & -966 & -391 & -253 & 2599 & -276 \\ -372 & -504 & -204 & -132 & -276 & 1488 \end{bmatrix}$$

Here

$$\mathbf{A_1} = \begin{bmatrix} 0 & 1 & 2 & -2 & -1 & 0 \\ 0 & 1 & 2 & 0 & 0 & 0 \end{bmatrix} \qquad \mathbf{A_2} = [\ 1,\ -1\ ]$$

$$\mathbf{B_1} = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 2 & -2 & -1 & 0 \\ 0 & 1 & 2 & 0 & 0 & 0 \end{bmatrix} \qquad \mathbf{B_2} = [\ 1,\ 1,\ -1\ ]$$

Thus $\mathbf{a_1} = \mathbf{A_1p} = [31,\ 76]'$ and $\mathbf{b_1} = \mathbf{B_1p} = [90,\ 31,\ 76]'$

We can thus express both λ_e and λ as compounded functions such that

$$\hat{\lambda}_e = \exp\{A_2(\log_e a_1)\} \quad \text{and} \quad \hat{\lambda} = \exp\{B_2(\log_e b_1)\}$$

We have seen that their variances turn out to be given by

$$\mathrm{Var}(\hat{\lambda}_e) = (\hat{\lambda}_e)^2 \mathbf{A_2 Da_1{}^{-1} A_1 Var(p) A_1{}' Da_1{}^{-1}} A_2{}'$$

and $$\mathrm{Var}(\hat{\lambda}) = (\hat{\lambda})^2 \mathbf{B_2 D_{b1}{}^{-1} B_1 Var(p) B_1{}' D_{b1}{}^{-1} B_2{}'}$$

where in this example:

$$\mathbf{Da_1} = \frac{1}{136}\begin{bmatrix} 31 & 0 \\ 0 & 76 \end{bmatrix} \qquad \mathbf{D_{b1}} = \frac{1}{136}\begin{bmatrix} 90 & 0 & 0 \\ 0 & 31 & 0 \\ 0 & 0 & 76 \end{bmatrix}$$

By multiplying the matrix products using a computer, we obtain

$$\mathrm{Var}(\hat{\lambda}_e)=0.01827645 \qquad \mathrm{Var}(\hat{\lambda})= 0.01062808$$

Alternatively, we could simply use the algebraic expressions obtained in (3.7) and (3.8). As is expected, the results are

$$\mathrm{Var}(\hat{\lambda}_e) = 0.01827645 \qquad \mathrm{Var}(\hat{\lambda}) = 0.01062808$$

Again, the two methods lead to the same estimates of the variances.

3 The case of variable number of controls per case

Consider the example of a case-control of the association between Hodgkins disease and tonsillectomy used by Walter (1980). The 104 cases considered below are part of the 153 cases reported by Walter. For illustrative purposes, only the data on completely observed triples (ie cases with 2 controls observed) and cases with only one control observed are used. The controls (up to 2 per case) were sampled from other patients at the same hospital and were matched on age, admission date, sex and race. Tonsillectomy status (exposure) was determined when possible, only medical record of each patient. The table below shows the data.

Table 3.7

Previous tonsillectomy status for Hodgkins disease cases

and matched controls.

Cases	Controls		Number of matched groups
1	1	1	4
1	1	0	8
1	0	0	17
0	1	1	9
0	1	0	21
0	0	0	18
1	1	-	3
1	0	-	5
0	1	-	6
0	0	-	13

1=Previous tonsillectomy ; 0=no previous tonsillectomy ; — =missing observation.

To conform to our formulation in Table 3.3, we shall represent the data as:

Table 3.8

	1 control observed		2 controls observed			
	No. exposed		No. exposed			
Case Status	1	0	2	1	0	Total
Exposed	3	5	4	8	17	37
Non-exposed	6	13	9	21	18	67
						104

Then from (2.24), an estimate of Mantel-Haenszel odds ratio (Fleiss (1984)) would be obtained as:

$$\hat{\psi} = 1.03125$$

and an estimate of attributable risk among exposed from (2.25) would be:

$$\hat{\lambda}_e = 0.0303$$

Exposure prevalence among the exposed only would be $\hat{V}_x$=0.3558

and thus the estimate of population attributable risk is given by

$$\hat{\lambda} = \hat{V}_x\hat{\lambda}_e = 0.0108$$

From Fleiss (1984), we can find variance of $\log_e \psi$ given in equation (2.29). Using the data, we obtain

$$\mathrm{Var}(\log_e \hat{\psi}) = 0.06620757 \qquad \text{and thus}$$

$$\mathrm{Var}(\hat{\lambda}_e) = (1-\hat{\lambda}_e)^2\mathrm{Var}(\log_e \hat{\psi}) = 0.0622558$$

Using $\mathrm{Var}(\hat{V}_x) = 0.00220382$ we find from (2.34) that

$$\mathrm{Var}(\hat{\lambda}) = 0.0080828$$

Alternatively, we could use the multivariate Taylor series (matrix products) method already discussed in Section 3.1.

Here, the vector of multinomial proportions, p , could be written as

$$p' = [3, 5, 4, 8, 17, 6, 13, 9, 21, 18]/104$$

Thus $C_1 = [0, {}^1/_2, 0, {}^1/_3, {}^2/_3, 0, 0, 0, 0, 0]$

and $C_2 = [0, 0, 0, 0, 0, {}^1/_2, 0, {}^2/_3, {}^1/_3, 0]$

Here, we have

$$A_1 = \begin{bmatrix} 0 & {}^1/_2 & 0 & {}^1/_3 & {}^2/_3 & -{}^1/_2 & 0 & -{}^2/_3 & -{}^1/_3 & 0 \\ 0 & {}^1/_2 & 0 & {}^1/_3 & {}^2/_3 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

from which we can show that $C_1p = ({}^5/_2 + {}^8/_3 + {}^{34}/_3)/104 = 16.5$

and $C_0p = ({}^6/_2 + {}^{18}/_3 + {}^{21}/_3)/104 = 16,$ the ratio of which gives the odds ratio.

We can thus express the measures of attributable risk as compounded functions of these matrices such that

$$\hat{\lambda}_e = \exp\{A_2 \log_e (\mathbf{A_1 p})\} \text{ and } \qquad \hat{\lambda} = \exp\{B_2 \log_e (\mathbf{B_1 p})\}$$

where $\mathbf{A_2} = [\ 1,\ -1\ ]$, $\qquad \mathbf{B_2} = [\ 1,\ 1,\ -1\ ]$

$$B_1 = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & {}^1/_2 & 0 & {}^1/_3 & {}^2/_3 & -{}^1/_2 & 0 & -{}^2/_3 & -{}^1/_3 & 0 \\ 0 & {}^1/_2 & 0 & {}^1/_3 & {}^2/_3 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

The consistent estimate of the covariance of **p**, would thus be given by

$$\mathbf{Var(p)} = 1/N[\mathbf{D_p} - \mathbf{pp'}]$$

From our example,

$$\mathbf{Var(p)} = \frac{1}{(104)^3} \begin{bmatrix} 303 & -15 & -12 & -24 & -51 & -18 & -39 & -27 & -63 & -54 \\ -15 & 495 & -20 & -40 & -85 & -30 & -65 & -45 & -105 & -90 \\ -12 & -20 & 400 & -32 & -68 & -24 & -52 & -36 & -84 & -72 \\ -24 & -40 & -32 & 768 & -136 & -40 & -104 & -72 & -168 & -144 \\ -51 & -85 & -68 & -136 & 1479 & -102 & -221 & -153 & -357 & -306 \\ -18 & -30 & -24 & -40 & -102 & 588 & -78 & -54 & -126 & -108 \\ -39 & -65 & -52 & -104 & -221 & -78 & 1183 & -117 & -273 & -234 \\ -27 & -45 & -36 & -72 & -153 & -54 & -117 & 855 & -189 & -162 \\ -63 & -105 & -84 & -168 & -357 & -126 & -273 & -189 & 1743 & -378 \\ -54 & -90 & -72 & -144 & -306 & -108 & -234 & -162 & -378 & 1548 \end{bmatrix}$$

As already indicated, the large sample variance formulas are given as in (3.3) and (3.4), where by multiplying the appropriate matrices we obtain

$$\text{Var}(\hat{\lambda}_e) = 0.0622558 \qquad \text{and} \qquad \text{Var}(\hat{\lambda}) = 0.0080816$$

We realise that the two methods yield the almost the same estimates of the variances and the difference is only due to rounding error.

CHAPTER 4

EFFICIENCY CONSIDERATIONS

4.1 INTRODUCTION

The use of case-control studies with multiple controls per case is widespread in epidemiologic research. Frequently, one has only a limited number of cases available or they are expensive to obtain, whereas the controls are readily available. However, situations do exist where the controls are just as hard to obtain as are the cases, and under these circumstances it might be unwise to take a large number of controls per case.

Many authors have looked at what they consider to be a good choice for the number of controls when there is a single dichotomous exposure variable. Ury (1975) considered the asymptotic relative efficiency of two designs with a different number of controls per case. He used the number of cases required for a given power in order to compare the efficiencies of alternative matching ratios. He showed that the efficiency of a design with R_1 controls relative to a design with R_2 controls is given by

$$R_1(R_2+1)/R_2(R_1+1)$$

Miettinen (1969) performed a cost analysis. He looked at the cost of choosing R controls per case for n cases, where c_1 is the cost per case and c_2 is the cost per control, making the total cost $n(c_1+Rc_2)$. For fixed R, he found the sample size that gives a certain power against a local alternative, and then by minimising the total cost with respect to R, he showed that the best choice of R is $\sqrt{}(c_1/c_2)$. Gail et al (1973) reported similar results and calls it the 'square root rule'.

Taylor (1986) looked at a simple example of a matched case-control study, where there is a single binary exposure variable. By looking at the power of the usual test against a specific alternative he showed that there is a rapidly diminishing return with an increase in the number of controls per case. He thus recommended that, with equal difficulty in obtaining cases and controls, it appears that R = 1 is the best choice and that rarely is it worth having more than 3 controls per case.

Walter (1980) considered the case where a variable number of controls may be matched to a case, in the situation where the design results from censoring or an interim analysis. Suppose that censoring of cases and controls occurs independently but with possibly different probabilities π_1 and π_2, and for simplicity the cost of enrolling a case is unity and that of a control is c. He showed that the asymptotic relative efficiency of a design using R_1 controls per case as compared to R_2 controls per case for a binary response is given by

$$RE_b(R_1,R_2) = \frac{R_1(CR_2+1)[R_2(1-\pi_2)+1+\pi_2]}{R_2(CR_1+1)[R_1(1-\pi_2)+1+\pi_2]}$$

He showed that the optimal ratio (maximising efficiency) is

$$R_{opt} = \sqrt{}\{(1+\pi_2)/[C(1-\pi_2)]\}$$

which reduces to $1/\sqrt{C}$ (square root rule) when $\pi_2=0$.

We shall look at cost-efficiency considerations when attributable risk estimation is the ultimate goal of the analysis. Consider the total cost of choosing R controls per case for N cases, where C_1 is the cost per case (assumed to be the same for all cases) and C_2 is the cost per control (also assumed to be the same for each control), thus making the total cost $N(C_1+RC_2)$. If in addition, it is assumed that it costs k times the cost of obtaining one control to obtain a case (ie $C_1 = kC_2$), then the total cost is $NC_2(k+R)$. We note that the actual total cost would include an overhead cost but this is ignored in this formulation. We can therefore determine the total number of matched sets corresponding to any matched design at a fixed total cost, and consequently determine which matched design results in the smallest variance of the attributable risk. Since the variance depends on the relative risk (which in case-control studies is the odds ratio ψ) and the exposure prevalence among cases, V_X (as is appropriate for a matched design), we shall specify some values of ψ and V_X.

For example, for a fixed total cost $300C_2$ (ie $N(k+R)=300$), we would have the following number of matched sets, N, corresponding to the number of controls, R, matched to each case, where $C_1 = C_2$.

N	R	Cost
150	1	$300C_2$
100	2	$300C_2$
75	3	$300C_2$
60	4	$300C_2$
50	5	$300C_2$

For various values of ψ and V_X, we could determine the 'cell frequencies' corresponding to each matched design, for any N, and hence obtain an estimate of the attributable risk and its large sample variance. In this case the expected cell frequencies are those expected from random sampling.

As is appropriate for a multinomial table, we could determine the probability that a matched set falls in any cell.

Let P_0 = P(exposure/control), and P_1 = P(exposure/case).

Consider a matched-pair case-control data represented as:

		Control		
		1	0	Total
Case	1	Z_{11}	Z_{10}	
	0	Z_{01}	Z_{00}	
				N

Let P_{ij} be the probability that a matched set falls in the (ij)th cell, i = 0,1 and j = 0,1. Thus

$$P_{11} = P_1 P_0 \qquad P_{10} = P_1(1-P_0)$$

$$P_{01} = (1-P_1)P_0 \qquad P_{00} = (1-P_1)(1-P_0)$$

Then the expected cell frequencies are given by $E(Z_{ij}) = NP_{ij}$.

Generally, for R controls per case represented as

		No. of controls exposed					
		R	...	2	1	0	Total
Case	1	Z_{1R}	...	Z_{12}	Z_{11}	Z_{10}	
	0	Z_{0R}	...	Z_{02}	Z_{01}	Z_{00}	
							N

The probablity that a matched set falls in the (ij)th cell, P_{iJ}, i=0,1 and J=0,1,2,..., R will be given by:

$$P_{1J} = \frac{R\,!}{J!(R-J)!} P_1 P_0^{J}(1 - P_0)^{R-J}, \; J = 0,1,2,\ldots, R$$

$$P_{0J} = \frac{R\,!}{J!(R-J)!}\,(1-P_1)P_0^{\,J}(1-P_0)^{R-J}, J = 0,1,2,\ldots, R$$

where the expected cell frequencies are given by $E(Z_{iJ}) = NP_{iJ}$.

We have already indicated that for matched case-control data, provided the case represents a random sample of cases from the target population, an estimate of exposure prevalence among cases, V_X , could be found. This estimate is what is being referred to as P_1, here. We could also determine P_0 (proportion of matched controls who are exposed) by using the equation

$$P_0=P_1/[P_1+\psi(1-P_1)] = V_X/[V_X+\psi(1-V_X)]$$

where ψ (odds ratio) is an estimate of the relative risk. The above equation is obtained from the fact that

$$P_1 = P_0\psi/(1+P_0(\psi-1)) \qquad \text{(cf 2.14)}$$

Thus by specifying values for N, ψ, and V_X, we can find the expected cell frequencies, and consequently find λ and λ_e, and their asymptotic variances using methods already outlined in Chapters 2 and 3.
Using the variances (determined at fixed cost) for various matching designs, we can find for example, the asymptotic relative efficiency of multiple controls relative to matched pairs, defined here as

$$RE(R_r,R_1) = Var(\lambda_1)/Var(\lambda_r)$$

where $Var(\lambda_1)$ refers to the variance of the population attributable risk when 1 control is matched per case; and $Var(\lambda_r)$ refers to variance of the population attributable risk when r controls are matched to a case.

Similarly, we could determine

$$RE(E_r,E_1) = Var(\lambda_{e1})/Var(\lambda_{er})$$

where $Var(\lambda_{e1})$ and $Var(\lambda_{er})$ refer to the variances of attributable risk among exposed when 1 and r controls, respectively, are matched to a case.

We could therefore determine relative efficiency of one matching design to another (up to matching with 5 controls per case) at different costs of obtaining a control when it costs k times to obtain a case.
By means of a simple computer program, we found relative efficiency under these instances:

(i) ψ ranges from 1.5 to 10.

(ii) V_X ranges from 0.05 to 0.95.

(iii) matching R = 1, 2, 3, 4, 5 controls per case.

(iv) for some total cost (arbitrarily chosen) depending on k = 1, 2, or 5;

(for example, total cost $300C_2$ for k = 1, $420C_2$ for k = 2 and $2520C_2$ for k = 5).The total cost chosen does not make any difference in the relative efficiency obtained, as it only determines the value of N (the total number of matched sets) to be used in the large-sample variance calculation.

Some important results are presented in Tables B.1, B.2 and B.3 of Appendix B.

4.2 DISCUSSION OF RESULTS

Table B.1 of Appendix B gives the asymptotic relative efficiency for various of ψ (the odds ratio used here to estimate relative risk) and V_x (the exposure prevalence among the cases) when $C_1 = C_2$ (equal cost for cases and controls). We realise that R = 1 (matched pair) is the best choice in such a situation for estimating population attributable risk (λ) in the sense that it has the smallest variance of λ. In the case of estimating attributable risk among the exposed (λ_e), we realise that for some situations ($\psi < 2.5$), R = 1 is the best choice, though for large ψ ($\psi \geqslant 3$) and small V_x ($V_x < 0.5$), matching with 2 controls per case (R = 2) might seem more appropriate.

For example, consider a chemical plant where the relative risk of some disease (eg. carcinoma of the lung) is very high ($\psi \geqslant 10$) among workers exposed to certain chemicals, compared to those not exposed, but only a small fraction of cases of the disease reported are exposed to the particular industrial chemical (eg. $V_x \leqslant 0.1$). In such a situation, it might be more appropriate to match 3 controls to a case. To help in deciding what is the optimal matching ratio, we have provided a table of optimal R's in Table 4.1. formed from Table B.1.

Table 4.1

Optimal R (C_1 = C_2)

	For λ_e					For λ				
V_x	0.1	0.3	0.5	0.7	0.9	0.1	0.3	0.5	0.7	0.9
ψ										
1.5	1	1	1	1	1	1	1	1	1	1
2.0	1	1	1	1	1	1	1	1	1	1
2.5	2	1	1	1	1	1	1	1	1	1
3.0	2	2	1	1	1	1	1	1	1	1
5.0	2	2	2	1	1	1	1	1	1	1
10.0	3	3	2	2	1	1	1	1	1	1

This table becomes particularly useful, when estimates of ψ and V_x are available from other sources. If these estimates are not already available as might be the case before most studies, R = 1 (the matched-pair design) offers the best choice, and rarely do we need to consider R > 3. This is similar to the conclusion reached by Taylor (1986).

Table B.2 looks at the situation where it costs 2 times to obtain a case. We realise from Table B.2 and also the table of optimal matching ratios obtained from it (Table 4.2) that R = 1 would be the best design if only population attributable risk were to be determined. However, the estimate of population attributable risk (λ) depends on both the exposure prevalence among the cases (V_x) and the attributable risk among the exposed (λ_e), so we need to consider the optimal matching ratio for λ_e

as well. We see that for most designs likely to occur in practice, R = 2 is the best choice, except for situations with very high V_x ($V_x > 0.7$) or very high ψ ($\psi > 5$) where other designs with a higher R would seem appropriate.

Table 4.2

Optimal R (C_1 = 2X C_2)

	For λ_e					For λ				
V_x	0.1	0.3	0.5	0.7	0.9	0.1	0.3	0.5	0.7	0.9
ψ										
1.5	2	2	2	1	1	1	1	1	1	1
2.0	2	2	2	1	1	1	1	1	1	1
2.5	2	2	2	2	1	1	1	1	1	1
3.0	2	2	2	2	1	1	1	1	1	1
5.0	3	3	2	2	1	1	1	1	1	1
10.0	4	4	3	3	2	1	1	1	1	1

Table B.3 considers the situation where it costs much more (5 times) to obtain a case. For population attributable risk (λ), R = 2 is better for $\psi < 3$, but R = 1 is more appropriate if $\psi \geq 5$. However, for attributable risk among exposed (λ_e), R = 3 is the optimal matching ratio (Table 4.3) when $\psi < 3$ and $V_x < 0.5$, but for large V_x ($V_x > 0.5$) R = 2 seems a better choice. It also has the interesting feature that for significantly high relative risks ($\psi > 5$) and low to moderate exposure prevalence among cases ($V_x < 0.5$), the number of controls to match to a case that has the potential of increasing efficiency increases (R = 4 or even 5 in some cases). Table 4.3 provides the optimal matching ratios under some of these circumstances. Since it is very likely that values of V_X and ψ may not be available before a study, we suggest that R = 3 be taken as the optimal matching ratio when it costs 5 times to obtain a case.

Table 4.3

Optimal R (C_1 = 5X C_2)

	For λ_e					For λ				
V_X	0.1	0.3	0.5	0.7	0.9	0.1	0.3	0.5	0.7	0.9
ψ										
1.5	3	3	2	2	2	2	2	2	2	2
2.0	3	3	3	2	2	2	2	2	2	2
2.5	3	3	3	2	2	2	2	2	2	2
3.0	4	3	3	2	2	1	1	2	2	2
5.0	5	4	4	3	2	1	1	1	1	1
10.0	5	5	5	4	2	1	1	1	1	1

4.3 CONCLUSION

We have looked at the various methodologies for estimating attributable risks in case-control studies, and for matched case-control studies in particular. Until recently, such estimators, their large-sample standard errors as well as their interval estimates have not been available. Kuritz and Landis (1987a) have provided explicit expressions for estimating attributable risks and their large-sample variances (given in equations (3.5) and (3.6) of the last chapter) for matched-pair case-control data, and we have also derived simple algebraic expressions for the large-sample variance of attributable risks (both λ_e and λ) when exactly 2 controls are matched to a case (given in equations (3.7) and (3.8)). Since the algebra gets very messy when working with data obtained from a case-control study with more than 2 controls matched to a case, the alternative expressions given by Kuritz and Landis (1987b) should be used (given in equations (2.32) and (2.34)). Our examples indicate that these alternative expressions are as good as those obtained through the matrix approach. Both sets of results are asymptotic. These expressions mean that researchers can make use of simple algebraic formulas to find attributable risks for matched case-control data, their asymptotic standard errors and interval estimates using only a pocket calculator.

Kleinbaum et al (1982) noted that the selection of more controls than cases helps to insure that there will be controls for cases at all relevant levels of the confounding variables, so that adequate comparisons could be made. On the other hand, if the cost of obtaining the study information is high, the greatest efficiency could be obtained by having equal number of cases and controls, where in this case matching will assure comparable distribution with respect to confounding variables. The cost benefit of matching in terms of efficiency and validity of study depends on the degree of confounding, and matching on an unrelated factor could result in overmatching, and hence loss in efficiency.

We considered relative efficiency under different matching designs (on the assumption that matching was on relevant confounding variables) and with different costs of obtaining cases and controls when attributable risk is the ultimate goal. Our results led to tables of optimal matching ratios (Tables 4.1, 4.2 and 4.3). These results have led us to believe that when it costs more to obtain a case for a matched study, more controls should taken, but rarely do we need to take more than 3 controls, except when ancillary information suggests that the risk of disease is much higher ($\psi > 10$) among the exposed persons as compared to those not exposed. The table of optimal matching ratios could be used whenever possible to help us make decisions about how many controls to match per case.

As noted by Kuritz and Landis (1987a), further empirical work is necessary to investigate the small to moderate sample size behaviour of these estimators. We have not particularly considered cases where the factors are associated with decreased risk (ie when $\psi < 1$) and further work could be carried in such situations. The ideas presented here could also be extended to situations with multiple exposure levels (not exposed, mildly exposed, severely exposed, etc.).

APPENDIX A

The sample multinomial proportions arising from Table 3.1.1 can be formed as

$$p' = [\ a,\ b,\ c,\ d,\ e,\ f\ ]/n$$

with the estimated covariance matrix being

$$\mathbf{Var(p)} = (\mathbf{D_p} - \mathbf{pp'})/n$$

where $\mathbf{D_p}$ is a diagonal matrix with vector $\mathbf{p}$ on the main diagonal. If we let

$$\mathbf{A_1} = \begin{bmatrix} 0 & 1 & 2 & -2 & -1 & 0 \\ 0 & 1 & 2 & 0 & 0 & 0 \end{bmatrix}$$

$$\mathbf{B_1} = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 2 & -2 & -1 & 0 \\ 0 & 1 & 2 & 0 & 0 & 0 \end{bmatrix}$$

then $\mathbf{a_1} = \mathbf{A_1}p$ and $\mathbf{b_1} = \mathbf{B_1}p$ are the linear functions of $\mathbf{p}$ necessary to give the required numerators and denominators.

The attributable risk estimates can be expressed as compounded functions of $\mathbf{p}$ by writing

$$\hat{\lambda}_e = \exp\{\mathbf{A_2}\ (\log_e \mathbf{a_1})\ \}$$

$$\hat{\lambda} = \exp\{\mathbf{B_2}\ (\log_e \mathbf{b_1})\ \}$$

where $\mathbf{A_2} = [\ 1,\ -1\ ]$ and $\mathbf{B_2} = [\ 1,\ 1,\ -1\ ]$

Thus, by appropriate matrix products arising from multivariate Taylor series methods [Kuritz and Landis (1987a)], the large-sample variance estimators for these functions are given by

$$\mathrm{Var}(\hat{\lambda}_e) = (\hat{\lambda}_e)^2 \mathbf{A_2 Da_1{}^{-1} A_1 Var(p) A_1{}' Da_1{}^{-1} A_2{}'} \qquad (A.1)$$

$$\mathrm{Var}(\hat{\lambda}) = (\hat{\lambda})^2 \mathbf{B_2 D_{b1}{}^{-1} B_1 Var(p) B_1{}' D_{b1}{}^{-1} B_2{}'} \qquad (A.2)$$

APPENDIX B

Table B.1

Relative Efficiency for various values of ψ and V_X. ($C_1 = C_2$) under different matching designs.

	RE(E_2,E_1)					RE(R_2,R_1)				
V_X	0.1	0.3	0.5	0.7	0.9	0.1	0.3	0.5	0.7	0.9
ψ										
1.5	0.943	0.923	0.901	0.876	0.848	0.835	0.836	0.838	0.838	0.836
2.0	0.987	0.958	0.923	0.881	0.830	0.802	0.807	0.811	0.812	0.808
2.5	1.023	0.989	0.947	0.893	0.822	0.781	0.788	0.794	0.798	0.792
3.0	1.052	1.016	0.970	0.908	0.821	0.765	0.773	0.782	0.788	0.782
5.0	1.129	1.093	1.043	0.967	0.836	0.731	0.739	0.751	0.763	0.765
10.0	1.212	1.186	1.144	1.070	0.902	0.701	0.708	0.718	0.734	0.754

	RE(E_3,E_2)					RE(R_3,R_2)				
V_X	0.1	0.3	0.5	0.7	0.9	0.1	0.3	0.5	0.7	0.9
ψ										
1.5	0.870	0.860	0.849	0.838	0.825	0.819	0.820	0.820	0.820	0.819
2.0	0.893	0.878	0.860	0.840	0.817	0.805	0.807	0.809	0.809	0.807
2.5	0.912	0.894	0.872	0.846	0.813	0.795	0.798	0.801	0.803	0.800
3.0	0.929	0.909	0.884	0.853	0.812	0.789	0.779	0.783	0.798	0.796
5.0	0.975	0.954	0.924	0.883	0.819	0.775	0.779	0.783	0.788	0.789
10.0	1.031	1.013	0.985	0.939	0.850	0.773	0.766	0.770	0.776	0.784

Table B.1 (continued)

Relative Efficiency for various values of ψ and V_x ($C_1 = C_2$)

	RE(E_4,E_3)					RE(R_4,R_3)				
V_x	0.1	0.3	0.5	0.7	0.9	0.1	0.3	0.5	0.7	0.9
ψ										
1.5	0.870	0.863	0.857	0.850	0.842	0.838	0.839	0.839	0.839	0.839
2.0	0.884	0.874	0.863	0.851	0.837	0.830	0.831	0.832	0.833	0.832
2.5	0.897	0.885	0.871	0.855	0.835	0.825	0.827	0.828	0.829	0.828
3.0	0.908	0.895	0.878	0.859	0.835	0.821	0.823	0.825	0.826	0.825
5.0	0.941	0.926	0.905	0.877	0.839	0.813	0.816	0.818	0.821	0.821
10.0	0.985	0.970	0.949	0.916	0.857	0.807	0.808	0.811	0.814	0.819

	RE(E_5,E_4)					RE(R_5,R_4)				
V_x	0.1	0.3	0.5	0.7	0.9	0.1	0.3	0.5	0.7	0.9
ψ										
1.5	0.879	0.875	0.870	0.866	0.861	0.858	0.858	0.859	0.859	0.858
2.0	0.890	0.883	0.875	0.867	0.857	0.852	0.853	0.854	0.854	0.854
2.5	0.899	0.890	0.880	0.869	0.856	0.849	0.850	0.851	0.852	0.851
3.0	0.907	0.897	0.885	0.872	0.856	0.847	0.848	0.849	0.850	0.849
5.0	0.932	0.920	0.904	0.885	0.858	0.842	0.843	0.845	0.847	0.847
10.0	0.967	0.955	0.938	0.913	0.870	0.838	0.839	0.840	0.842	0.845

Table B.2

Relative Efficiency for various values of ψ and V_X (C_1 = 2X C_2) under different matching designs.

	$RE(E_2,E_1)$					$RE(R_2,R_1)$				
V_X	0.1	0.3	0.5	0.7	0.9	0.1	0.3	0.5	0.7	0.9
ψ										
1.5	1.061	1.038	1.013	0.986	0.954	0.939	0.941	0.943	0.943	0.940
2.0	1.110	1.078	1.038	0.992	0.934	0.903	0.909	0.913	0.915	0.909
2.5	1.150	1.112	1.065	1.005	0.925	0.879	0.886	0.894	0.897	0.891
3.0	1.183	1.143	1.091	1.021	0.923	0.861	0.870	0.879	0.886	0.879
5.0	1.270	1.230	1.174	1.088	0.940	0.822	0.832	0.845	0.859	0.861
10.0	1.364	1.333	1.287	1.203	1.015	0.789	0.796	0.808	0.826	0.848

	$RE(E_3,E_2)$					$RE(R_3,R_2)$				
V_X	0.1	0.3	0.5	0.7	0.9	0.1	0.3	0.5	0.7	0.9
ψ										
1.5	0.928	0.918	0.906	0.894	0.880	0.873	0.875	0.875	0.875	0.874
2.0	0.953	0.936	0.918	0.896	0.871	0.858	0.861	0.862	0.863	0.861
2.5	0.973	0.954	0.930	0.902	0.868	0.849	0.852	0.854	0.856	0.853
3.0	0.991	0.969	0.943	0.910	0.867	0.842	0.845	0.849	0.851	0.849
5.0	1.040	1.017	0.986	0.942	0.874	0.826	0.830	0.835	0.841	0.842
10.0	1.100	1.081	1.051	1.003	0.907	0.814	0.817	0.821	0.828	0.836

Table B.2 (continued)

Relative efficiency for various values of ψ and V_X (C_1 = 2X C_2)

	RE(E_4,E_3)					RE(R_4,R_3)				
V_X	0.1	0.3	0.5	0.7	0.9	0.1	0.3	0.5	0.7	0.9
ψ										
1.5	0.906	0.899	0.893	0.885	0.877	0.874	0.874	0.874	0.874	0.874
2.0	0.921	0.911	0.899	0.887	0.872	0.865	0.866	0.867	0.868	0.866
2.5	0.934	0.922	0.907	0.890	0.870	0.859	0.861	0.863	0.864	0.862
3.0	0.946	0.932	0.914	0.895	0.870	0.856	0.857	0.859	0.861	0.860
5.0	0.981	0.964	0.943	0.914	0.874	0.847	0.849	0.852	0.855	0.855
10.0	1.026	1.010	0.988	0.954	0.893	0.841	0.842	0.845	0.848	0.852

	RE(E_5,E_4)					RE(R_5,R_4)				
V_X	0.1	0.3	0.5	0.7	0.9	0.1	0.3	0.5	0.7	0.9
ψ										
1.5	0.904	0.900	0.895	0.890	0.885	0.882	0.883	0.883	0.883	0.883
2.0	0.915	0.908	0.900	0.891	0.882	0.877	0.878	0.879	0.879	0.878
2.5	0.924	0.915	0.905	0.894	0.880	0.874	0.874	0.876	0.876	0.875
3.0	0.933	0.923	0.911	0.897	0.880	0.871	0.872	0.874	0.875	0.874
5.0	0.959	0.946	0.930	0.910	0.883	0.866	0.867	0.869	0.871	0.871
10.0	0.995	0.982	0.965	0.938	0.895	0.862	0.863	0.864	0.866	0.869

Table B.3

Relative Efficiency for various values of ψ and V_X (C_1 = 5X C_2) under different matching designs.

	$RE(E_2,E_1)$					$RE(R_2,R_1)$				
V_X	0.1	0.3	0.5	0.7	0.9	0.1	0.3	0.5	0.7	0.9
ψ										
1.5	1.213	1.187	1.158	1.127	1.091	1.073	1.076	1.078	1.078	1.075
2.0	1.269	1.231	1.187	1.133	1.067	1.031	1.035	1.043	1.045	1.039
2.5	1.315	1.271	1.217	1.149	1.058	1.006	1.012	1.021	1.026	1.018
3.0	1.352	1.306	1.247	1.167	1.055	0.984	0.994	1.005	1.012	1.005
5.0	1.451	1.406	1.341	1.243	1.075	0.941	0.952	0.966	0.982	0.984
10.0	1.559	1.524	1.469	1.376	1.159	0.903	0.911	0.923	0.943	0.969

	$RE(E_3,E_2)$					$RE(R_3,R_2)$				
V_X	0.1	0.3	0.5	0.7	0.9	0.1	0.3	0.5	0.7	0.9
ψ										
1.5	1.015	1.004	0.991	0.978	0.962	0.955	0.956	0.957	0.957	0.956
2.0	1.042	1.024	1.004	0.980	0.953	0.938	0.941	0.944	0.944	0.941
2.5	1.064	1.043	1.018	0.987	0.949	0.927	0.931	0.934	0.936	0.933
3.0	1.084	1.060	1.031	0.995	0.948	0.921	0.924	0.928	0.931	0.928
5.0	1.138	1.113	1.078	1.030	0.955	0.903	0.908	0.913	0.919	0.920
10.0	1.203	1.182	1.149	1.095	0.991	0.888	0.893	0.897	0.905	0.915

Table B.3 (continued)

Relative Efficiency for various values of ψ and V_x (C_1 = 5X C_2)

	$RE(E_4,E_3)$					$RE(R_4,R_3)$				
V_x	0.1	0.3	0.5	0.7	0.9	0.1	0.3	0.5	0.7	0.9
ψ										
1.5	0.966	0.959	0.952	0.944	0.936	0.933	0.932	0.933	0.933	0.932
2.0	0.983	0.972	0.959	0.946	0.930	0.923	0.924	0.925	0.925	0.924
2.5	0.997	0.983	0.968	0.949	0.928	0.917	0.919	0.920	0.921	0.919
3.0	1.009	0.994	0.976	0.954	0.927	0.912	0.915	0.917	0.918	0.917
5.0	1.046	1.028	1.005	0.975	0.932	0.904	0.905	0.909	0.912	0.912
10.0	1.094	1.077	1.056	1.017	0.952	0.896	0.898	0.901	0.905	0.909

	$RE(E_5,E_4)$					$RE(R_5,R_4)$				
V_x	0.1	0.3	0.5	0.7	0.9	0.1	0.3	0.5	0.7	0.9
ψ										
1.5	0.950	0.945	0.940	0.935	0.929	0.925	0.926	0.927	0.927	0.927
2.0	0.961	0.953	0.945	0.936	0.926	0.920	0.921	0.923	0.923	0.922
2.5	0.979	0.961	0.951	0.939	0.924	0.915	0.919	0.919	0.920	0.919
3.0	0.979	0.969	0.956	0.941	0.924	0.915	0.916	0.917	0.918	0.917
5.0	1.007	0.994	0.977	0.955	0.927	0.910	0.911	0.913	0.914	0.915
10.0	1.045	1.031	1.013	0.987	0.927	0.906	0.906	0.907	0.910	0.913

BIBLIOGRAPHY

1. Autunes, C.M.F., Stolley, P.D., Rosenshein, N.B., Davis, J.L., Tonascia, J.A., Brown, C., Burnett, L., Rutledge, A., Pokempner, M. and Garcia, R. (1979). Endometrial cancer and estrogen use:report of a large case-control study. *N. Eng. J. Med.* 300,9-13.

2. Berkson, J. (1958). Smoking and Lung Cancer: Some observations on 2 recent reports *J. Amer. Statist. Assoc.* 53, 28-38.

3. Bishop, Y., Fienberg, S.E. and Holland, P.W. (1975). *Discrete Multivariate Analysis.* Cambridge. Mass. M.I.T. Press.

4. Breslow, N. (1981). Odds ratio when data is sparse. *Biometrika* 68, 73-84.

5. Bruzzi, P., Green, S.B., Byar, D.P., Brinton, L.A. and Schairer, C. (1985). Estimating the population attributable risk for multiple risk factors using case-control data. *Amer. J. Epid.* 122, 904-914.

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

7. Connett, J., Ejigou, A., McHugh, R. and Breslow, N. (1982). Precision of the Mantel-Haenszel Estimator in Case-Control Studies with multiple matching. *Amer. J. Epid.* 116, 875-877.

8. Cornfield, J. (1951). A method of estimating comparative rates from clinical data. Applications to cancer of the lung, breast and cervix.

J. Nat. Cancer Inst. **11**,1269-1275.

9. Donner, A. and Hauck, W.W. (1986). The large-sample relative efficiency of the Mantel-Haenszel estimator in the fixed-strata case. *Biometrics* 42, 537-545.

10. Ejigou, A. and McHugh, R. (1977). Estimation of relative risk from matched pairs in epidemiologic research. *Biometrics* 33, 552-556.

11. —————————————— (1981). Relative risk estimation under multiple matching. *Biometrika* **68**, 85-91.

12. Fleiss, J.L. (1981). *Statistical Methods for Rates and Proportions.* Second Edition, Wiley, New York.

13. ————————(1982). Simplification of the classic large-sample standard error of a function of multinomial proportions. *Amer. Statist.* 36, 377-378.

14. ————————(1984). Mantel-Haenszel Estimator in Case-Control Studies with varying number of controls matched to each case. *Amer. J. Epid.* 120, 1-3.

15. Gail, M., William, R., Byar, D.P. and Brown, C. (1976). 'How many controls'. *J. Chron. Dis.* **29**, 723-731.

16. Kendall, M.G. and Stuart, A. (1969). *The Advanced Theory of Statistics.* Third Edition. Griffin. London.

17. Kleinbaum, D.G., Kupper, L.L. and Morgenstein, H. (1982). *Epidemiologic Research: Principles and Quantitative Methods.* Boston.

18. Koch, G.G., Landis, J.R., Freeman, J.L., Freeman, D.H. and Lehnen, R.G. (1977). A general methodology for the analysis of experiments with repeated measurement of categorical data. *Biometrics* 33, 133-158.

19. Kupper, L.L., Karon, J.M., Kleinbaum, D.G., Morgenstein, H. and Lewis, D.K. (1981). Matching in Epidemiologic Studies: Validity and Efficiency Considerations. *Biometrics* 37, 271-279.

20. Kuritz, S.J. and Landis, J.R. (1987a). Attributable risk ratio estimation from matched-pair case-control data. *Amer. J. Epid.* 127 324-328.

21. ----------------------------- (1987b). Attributable risk ratio estimation from matched set case-control data. *Submitted to Biometrics*

22. Leung, H.M. and Kupper, L.L. (1981). Comparison of confidence intervals for attributable risk. *Biometrics* 37, 293-302.

23. Levin, M.L. (1953). The occurance of lung cancer in man. *Acta Unio Int. Cancer* 19, 531-541.

24. Levin, M.L. and Bartel, R. (1978). Re: Simple estimation of population attributable risk from case-control studies. *Amer. J. Epid.* 108, 78-79.

25. Leviton, A. (1973). Definations of attributable risk. *Amer. J. Epid.* 98, 231.

26. MacMahon, B. and Pugh, T.P. (1970). *Epidemiology, Principles, and Methods.* Little Brown and Co.

27. Mantel, N. and Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *J. Nat. Cancer Inst.* 22, 719-748.

28. McKinlay, S.M. (1977). Pair-matching - a reappraisal of a popular technique. *Biometrics* 33, 725-735.

29. Miettinen, O.S. (1969). Individual matching with multiple controls in the case of all-or-none responses. *Biometrics* 25, 339-355.

30. ———————— (1970). Estimation of relative risk from individually matched series. *Biometrics* 26, 75-86.

31. ———————— (1974). Proportion of diseases caused or prevented by a given exposure, trait or intervention. *Amer. J. Epid.* 99, 325-332.

32. Miller, C.T., Neutel, C.I., Nair, R.C., Marrett, L.D., Last, J.M. and Collins, W.E. (1978). Relative importance of risk factors in bladder carcinogenesis. *J. Chron. Dis.* 31, 51-56.

33. Park, C.B. (1981). Attributable risk for recurrent events: An extension of Levin's measure. *Amer. J. Epid.* 113, 491-493.

34. Schlesselman, J.J. (1982). *Case-Control Studies.* New York. Oxford U. Press.

35. Shep, M.C. (1959). An extension of some methods of comparing several rates or proportions. *Biometrics* 15, 87-97.

36. Taylor, J.W. (1977). Simple estimation of population attributable risk from case-control studies. *Amer. J. Epid.* 106, 260.

37. Taylor, J.M.G. (1986). Choosing the number of controls in a matched case-control study, some sample size, power and efficiency considerations. *Stat. Med.* 5, 29-36.

38. Ury, H.K. (1975). Efficiency of case-control studies with multiple controls per case: continuous or dichotomous data. *Biometrics* 31, 643-649.

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

40. ————————— (1976). The estimation and interpretation of attributable risk in health research. *Biometrics* 32, 829-849.

41. ————————— (1978). Calculation of attributable risks from epidemiologic data. *Int. J. Epid.* 7, 175-185.

42. ————————— (1980). Matched case-control studies with a variable number of controls per case. *J. Roy. Statist. Soc. Ser. C* 29, 172-179.

43. Whittemore, A.S. (1982). Statistical methods for estimating attributable risk from retrospective data. *Stat. Med.* 1, 229-243.

44. ————————— (1983). Estimating attributable risk from case-control studies. *Amer. J. Epid.* 117, 76-83.