

DIVISION OF THE HUMANITIES AND SOCIAL SCIENCES

# CALIFORNIA INSTITUTE OF TECHNOLOGY

PASADENA, CALIFORNIA 91125

## STATISTICAL ANALYSIS OF THE ADDITIVE AND MULTIPLICATIVE HYPOTHESES FOR COHORT AND CASE-CONTROL STUDIES

Jeffrey A. Dubin



**SOCIAL SCIENCE WORKING PAPER xxx**

July 1999

**Statistical Analysis of the Additive and  
Multiplicative hypotheses for Cohort and  
Case-Control Studies**

Jeffrey A. Dubin

# Contents

<b>1</b>	<b>Introduction</b>	<b>3</b>
<b>2</b>	<b>Case-Control Studies</b>	<b>4</b>
2.1	Hypothesis Tests for Case-Control Studies – Multiplicative Case . . . . .	6
2.2	Woolf’s Method . . . . .	10
2.3	Hypothesis Tests for Case-Control Studies – Additive Case . . . . .	13
2.4	Maximum Likelihood . . . . .	15
2.4.1	Additive Constraint . . . . .	16
2.4.2	Multiplicative Constraint . . . . .	19
2.4.3	Unconstrained Maximum Likelihood . . . . .	20
<b>3</b>	<b>Bonferroni Intervals and Monte Carlo Simulations</b>	<b>20</b>
3.1	Bonferroni Interval . . . . .	20
3.2	Simulation Methods . . . . .	21
<b>4</b>	<b>Synergy Indices</b>	<b>22</b>
4.1	Rothman’s <i>S</i> Index . . . . .	22
4.2	Attributable Proportion . . . . .	23
4.3	Additive-Multiplicative Measure . . . . .	24
<b>5</b>	<b>Cohort Studies</b>	<b>24</b>
5.1	Determination of Death Rates . . . . .	27
<b>6</b>	<b>Prevalance Studies</b>	<b>29</b>
<b>7</b>	<b>Conclusion</b>	<b>29</b>

# 1 Introduction

In epidemiological studies the issue arises whether multiple causes of a particular disease have a synergistic relationship in that their combined effect is greater than that of either activity alone and greater than what one would expect by the sum of their individual risk contributions. Two hypotheses are frequently tested. First, when the sources of disease act independently it is hypothesized that the relative risk of disease given exposure is an additive relationship. Here the relative risk of dying from cause  $A$  adds to the relative risk of dying from cause  $B$  to determine the combined relative risk of dying when exposed to both  $A$  and  $B$ . A second hypothesis states that the relationship between disease and two causal factors is multiplicative. In this case, the combined risk is the product of the individual risks.

Of course synergism is itself a concept that is model dependent. For instance a lack of synergism in a logit model of risk, as demonstrated by the statistical insignificance of an interaction term, leads to a multiplicative model of relative risk. Consider the following example. Consider the following example.

Suppose that the probability of dying from a disease depends on two factors  $A$  and  $B$ . Let  $\delta_A$  denote exposure to  $A$  and  $\delta_B$  denote exposure to  $B$ . Suppose further that the probability of dying is logistic and given by:

$$P[D|\delta_A, \delta_B] = 1 / (1 + e^{-(X_0\beta_0 + \delta_A X_A\beta_A + \delta_B X_B\beta_B + \delta_A\delta_B X_C)})$$

where  $X_A$ ,  $X_B$ ,  $X_C$ , and  $X_0$  are vectors of explanatory factors and  $\beta_j$  are true but unknown coefficient vectors. The presence of the term  $\delta_A\delta_B$  allows for synergism in this model and specifies that the probability of disease may be different when causal factors  $A$  and  $B$  are both present.

Now assume that  $\beta_C$  is zero so that there is no synergistic relationship in the model. The relative odds of dying when exposed to both agents is:

$$\begin{aligned} RO_{AB} &= P[D|\delta_A = 1, \delta_B = 1] / P[\bar{D}|\delta_A = 1, \delta_B = 1] \\ &= \exp(X_0\beta_0 + X_A\beta_A + X_B\beta_B) \end{aligned}$$

Similarly the relative odds of dying when exposed to  $A$  alone is:

$$\begin{aligned} RO_A &= P[D|\delta_A = 1, \delta_B = 0] / P[\bar{D}|\delta_A = 1, \delta_B = 0] \\ &= \exp(X_0\beta_0 + X_A\beta_A) \end{aligned}$$

and

$$\begin{aligned} RO_B &= P[D|\delta_A = 0, \delta_B = 1] / P[\bar{D}|\delta_A = 0, \delta_B = 1] \\ &= \exp(X_0\beta_0 + X_B\beta_B) \end{aligned}$$

and the relative odds of dying from background exposure is

$$\begin{aligned} RO_0 &= P[D|\delta_A = 0, \delta_B = 0] / P[\bar{D}|\delta_A = 0, \delta_B = 0] \\ &= \exp(X_0\beta_0) \end{aligned}$$

The relative risk is defined as the ratio of the relative odds between the exposure group and the baseline:

$$\begin{aligned} RR_{AB} &= \left( \frac{P[D|\delta_A = 1, \delta_B = 1]}{P[\bar{D}|\delta_A = 1, \delta_B = 1]} \right) / \left( \frac{P[D|\delta_A = 0, \delta_B = 0]}{P[\bar{D}|\delta_A = 0, \delta_B = 0]} \right) \\ &= \left( \frac{P[D|\delta_A = 1, \delta_B = 1]}{P[D|\delta_A = 0, \delta_B = 0]} \right) / \left( \frac{P[\bar{D}|\delta_A = 1, \delta_B = 1]}{P[\bar{D}|\delta_A = 0, \delta_B = 0]} \right) \end{aligned}$$

which says that the relative risk from combined exposure is equal to the ratio of the relative odds of dying in the exposed population to the relative odds of dying in the unexposed population.

$$\begin{aligned} \text{Then } RR_{AB} &= (RO_{AB}/RO_0) \\ &= \exp(X_A\beta_A + X_B\beta_B) \quad \text{and} \\ RR_A &= (RO_A/RO_0) = \exp(X_A\beta_A), \quad RR_B = (RO_B/RO_0) = \exp(X_B\beta_B) \end{aligned}$$

We see that  $RR_{AB} = RR_A \cdot RR_B$  even though the model exhibits synergism.

This paper considers several methods for determining the relative odds ratio including the case-control method, the cohort method, and the prevalence method. The particular focus of this paper is the testing of the statistical hypotheses of additivity and multiplicativity of the relative risk measures. While other papers have considered the confidence intervals for relative risk measures no systematic study has been made of the additivity and multiplicativity hypotheses as a matter of statistics. Therefore, while practice in epidemiology may have been to say that one or more studies appear to support the multiplicative model, these have not in general been statistical statements i.e. statements made with attendant levels of confidence.

The paper is divided into six sections. We discuss the case-control method and Wald type tests for the multiplicative and additive hypotheses. We also derive and discuss Woolf's method for determining the variance of log-odds ratios (Woolf (1955)). Then we discuss maximum likelihood methods for optimization subject to constraints following the methods of Gardner and Munford (1980). In Section 3 we discuss other approaches for determining confidence intervals including Bonferroni Intervals and Monte Carlo simulation. In Section 4, we describe various synergy indices and how they relate to tests of hypotheses for additive and multiplicative statistics. In Section 5, we discuss cohort studies and derive hypothesis tests for the additive and multiplicative statistics. In Section 6, we discuss prevalence studies and their relationship to cohort studies. Finally we present conclusions.

## 2 Case-Control Studies

We begin with a table of case-control outcomes at differing levels of exposure:

Exposure					
	None	A	B	A&B	TOTAL
cases	$h_1$	$h_2$	$h_3$	$h_4$	$h$
controls	$k_1$	$k_2$	$k_3$	$k_4$	$k$

We next express the row counts as fractions:

Exposure					
	None	A	B	A&B	TOTAL
cases	$\pi_1$	$\pi_2$	$\pi_3$	$\pi_4$	1
controls	$\theta_1$	$\theta_2$	$\theta_3$	$\theta_4$	1

where  $\hat{\pi}_j = h_j/h$  and  $\hat{\theta}_j = k_j/k$  are consistent estimates of the true cell probabilities.

First, we demonstrate that the retrospective odds-ratio from a case control method provides an approximate estimate of relative risk of being a “case” given exposure. To prove this we examine the odds ratio  $\pi_4 \cdot \theta_1 / \pi_1 \cdot \theta_4$  although the result clearly generalizes to other cases. We show that the odds-ratio approximates the relative risk of being a case given exposure level  $A&B$ .

We denote a case as  $D$  (death from lung cancer for instance) and  $\bar{D}$  a control (death from other causes for instance). The combined exposure  $A&B$  is referred to as  $E$  (exposure) while the exposure level “None” is denoted  $\bar{E}$  (no exposure).

The odds-ratio  $\pi_4 \cdot \theta_1 / \pi_1 \cdot \theta_4$  is equal to

$$\frac{P[E|D] \cdot P[\bar{E}|\bar{D}]}{P[\bar{E}|D] \cdot P[E|\bar{D}]} \quad (1)$$

since  $\pi_4 = P[E|D]$ ,  $\pi_1 = P[\bar{E}|D]$ ,  $\theta_4 = P[E|\bar{D}]$ , and  $\theta_1 = P[\bar{E}|\bar{D}]$

As the notation implies, the probabilities  $\pi$  and  $\theta$  are conditional probabilities indicating the respective likelihood of having been exposed given an individual’s case-control status. Of interest is the prospective probability of being a case (i.e. dying) given exposure status. <sup>1</sup>

Under a simplifying assumption the odds-ratio approximates the prospective odds:

$$\frac{P[E|D] \cdot P[\bar{E}|\bar{D}]}{P[\bar{E}|D] \cdot P[E|\bar{D}]} = \frac{P[E, D]/P[D] \cdot P[\bar{E}, \bar{D}]/P[\bar{D}]}{P[E, \bar{D}]/P[\bar{D}] \cdot P[\bar{E}, D]/P[D]}$$

<sup>1</sup>Some research studies have used logit analysis to model the conditional probabilities shown above. This allows the introduction of covariates to provide additional controls in the analysis. For example a logit model may be used to specify the conditional probabilities:  $P[A|D]$ ,  $P[\bar{A}|D]$ ,  $P[A|\bar{D}]$ , and  $P[\bar{A}|\bar{D}]$ . A specification of such a model was illustrated in the introduction. The presence of additional covariates complicates the analysis presented below as the variances and covariances become dependent on the assumed probability model and on the precision of the parameter estimation.

$$\begin{aligned}
&= \frac{P[D|E] \cdot P[E]}{P[D|\bar{E}] \cdot P[\bar{E}]} \cdot \frac{P[\bar{D}|\bar{E}] \cdot P[\bar{E}]}{P[\bar{D}|E] \cdot P[E]} \\
&= \frac{P[D|E]}{P[D|\bar{E}]} \cdot \frac{P[\bar{D}|\bar{E}]}{P[\bar{D}|E]} \\
&\doteq \frac{P[D|E]}{P[D|\bar{E}]} \tag{2}
\end{aligned}$$

where the approximation results from the observation that  $P[\bar{D}|\bar{E}]/P[\bar{D}|E]$  is close to one. Case control studies are useful as they provide estimates of the odds  $P[E]/P[\bar{E}]$  i.e. the relative odds of exposure. The relative odds of being a case  $P[D]/P[\bar{D}]$  are irrelevant as they are set by the researcher in the design. They do however have an influence on the confidence of the results.

## 2.1 Hypothesis Tests for Case-Control Studies – Multiplicative Case

The relative risk (prospective) of dying given exposure to contaminant  $A$  is  $\pi_2\theta_1/\pi_1\theta_2$ . The relative risk of dying given exposure to contaminant  $B$  is  $\pi_3\theta_1/\pi_1\theta_3$  where the relative risk of dying if exposed to both contaminants is  $\pi_4\theta_1/\pi_1\theta_4$ . The multiplicative hypothesis states that  $RR_{A\&B} = RR_A \cdot RR_B$  so that:

$$\pi_4\theta_1/\pi_1\theta_4 = (\pi_2\theta_1/\pi_1\theta_2) \cdot (\pi_3\theta_1/\pi_1\theta_3)$$

Taking logarithms this becomes:

$$\log \pi_4 + \log \theta_1 - \log \pi_1 - \log \theta_4 - \log \pi_2 - \log \theta_1 + \log \pi_1 + \log \theta_2 - \log \pi_3 - \log \theta_1 + \log \pi_1 + \log \theta_3 = 0$$

This may be rewritten as

$$\log (\pi_4\theta_2\theta_3\pi_1) - \log (\theta_4\pi_2\pi_3\theta_1) = 0$$

or

$$M = (\log \pi_1 - \log \pi_2 - \log \pi_3 + \log \pi_4) - (\log \theta_1 - \log \theta_2 - \log \theta_3 + \log \theta_4) = 0$$

A consistent estimate of this statistic is obtained by replacing  $\pi_j$  and  $\theta_j$  with  $\hat{\pi}_j$  and  $\hat{\theta}_j$ .

To derive the variance of the resulting statistic is complicated by the fact that  $h_1, h_2, h_3,$  and  $h_4$  form a multinomial probability distribution. Similarly  $k_1, k_2, k_3,$  and  $k_4$  form a multinomial probability distribution, but one which is independent of the joint distribution of the  $h_j k_j$  assumption.

To derive the joint distribution of the  $\log \pi_j$  and  $\log \theta_j$ , we begin with results for the joint distribution of the  $h_j$ . Similar results hold for the outcome of the  $k_j$ . For notational simplicity we present the results using a common symbol  $n_j$  where  $n_1 + n_2 + n_3 + n_4 = n$ .

**Lemma**

Let  $\delta_{jt} = 1$  if outcome  $j$  is realized in observation  $t$ . The probability that  $\delta_{jt} = 1$  is denoted  $\pi_j$ . Let  $n_j$  denote the total number of outcome  $j$ 's that are observed in the sample of  $n$  independent draws, with

$$\begin{aligned} \pi_1 + \pi_2 + \pi_3 + \pi_4 &= 1, & n_1 + n_2 + n_3 + n_4 &= n, \\ n_j &= \sum_{t=1}^n \delta_{jt}, & n &= \sum_{t=1}^n (\delta_{1t} + \delta_{2t} + \delta_{3t} + \delta_{4t}) = \sum_{t=1}^n 1. \end{aligned}$$

Then  $E(n_j) = n\pi_j$ ,  $V(n_j) = n\pi_j(1 - \pi_j)$ , and  $\text{cov}(n_j, n_k) = -n\pi_j\pi_k$  for  $j \neq k$ .

**Proof**

$n_j = \sum_{t=1}^n \delta_{jt}$  implies  $E(n_j) = \sum_{t=1}^n E(\delta_{jt}) = n\pi_j$  since  $E(\delta_j) = 1 \cdot \pi_j + 0 \cdot (1 - \pi_j)$ . Next  $V(n_j) = \sum_{t=1}^n V(\delta_{jt})$ . But  $V(\delta_{jt}) = E(\delta_{jt}) - E(\delta_{jt})^2 = \pi_j - \pi_j^2 = \pi_j(1 - \pi_j)$ . Hence  $V(n_j) = n\pi_j(1 - \pi_j)$ . Finally  $\text{cov}(n_j, n_k) = E[(n_j - n\pi_j)(n_k - n\pi_k)] = E(n_j n_k) - n\pi_j n\pi_k - n\pi_k n\pi_j + n^2 \pi_j \pi_k = E(n_j n_k) - n^2 \pi_j \pi_k$ . Now

$$E(n_j n_k) = E\left[\left(\sum_t \delta_{jt}\right)\left(\sum_t \delta_{kt}\right)\right] = E\left[\sum_t \delta_{jt} \delta_{kt} + \sum_{t \neq s} \delta_{jt} \delta_{ks}\right].$$

But  $\delta_{jt} \delta_{kt} = 0$  if  $j \neq k$  in observation  $t$  (only one unique outcome is realized in each trial) so that the first sum is exactly zero. The second sum consists of  $(n^2 - n)$  terms which are the products of independent random variables (since  $\delta_{jt}$  and  $\delta_{ks}$  are independent when  $t \neq s$ ). The expectation of each term in the second sum is  $E(\delta_{jt} \delta_{ks}) = \pi_j \pi_k$ .

Hence  $E(n_j n_k) = (n^2 - n)\pi_j \pi_k$ . Combining these results we obtain

$$\begin{aligned} \text{cov}(n_j, n_k) &= (n^2 - n)\pi_j \pi_k - n^2 \pi_j \pi_k \\ &= -n\pi_j \pi_k \end{aligned} \tag{3}$$

**Q.E.D.**

Combining these results into the variance covariance matrix for  $n_j$  we obtain:

$$E \begin{pmatrix} n_1 \\ n_2 \\ n_3 \\ n_4 \end{pmatrix} = n \begin{pmatrix} \pi_1 \\ \pi_2 \\ \pi_3 \\ \pi_4 \end{pmatrix}$$

and

$$V \begin{pmatrix} n_1 \\ n_2 \\ n_3 \\ n_4 \end{pmatrix} = n \begin{pmatrix} \pi_1(1 - \pi_1) & -\pi_1\pi_2 & -\pi_1\pi_3 & -\pi_1\pi_4 \\ -\pi_2\pi_1 & \pi_2(1 - \pi_2) & -\pi_2\pi_3 & -\pi_2\pi_4 \\ -\pi_3\pi_1 & -\pi_3\pi_2 & \pi_3(1 - \pi_3) & -\pi_3\pi_4 \\ -\pi_4\pi_1 & -\pi_4\pi_2 & -\pi_4\pi_3 & \pi_4(1 - \pi_4) \end{pmatrix} = n(I - \phi\phi')$$

where  $\phi = \left( \sqrt{\pi_1} \quad \sqrt{\pi_2} \quad \sqrt{\pi_3} \quad \sqrt{\pi_4} \right)'$ .

To derive the variance-covariance matrix for  $\log \hat{\pi}_j = \log (n_j/n)$  we use a Taylor's series expansion to first-order for the logarithm. Then

$$\log \hat{\pi}_j \doteq \log \pi_j + \frac{1}{\pi_j}(\hat{\pi}_j - \pi_j)$$

where we have evaluated the Taylor's expansion around the true but unknown  $\pi_j$ . Then

$$\log \begin{pmatrix} \hat{\pi}_1 \\ \hat{\pi}_2 \\ \hat{\pi}_3 \\ \hat{\pi}_4 \end{pmatrix} = \begin{pmatrix} \log \pi_1 \\ \log \pi_2 \\ \log \pi_3 \\ \log \pi_4 \end{pmatrix} + \begin{pmatrix} 1/\pi_1 & 0 & & \\ & 1/\pi_2 & & \\ 0 & & 1/\pi_3 & \\ & & & 1/\pi_4 \end{pmatrix} \begin{pmatrix} (\hat{\pi}_1 - \pi_1) \\ (\hat{\pi}_2 - \pi_2) \\ (\hat{\pi}_3 - \pi_3) \\ (\hat{\pi}_4 - \pi_4) \end{pmatrix}$$

Hence

$$\begin{aligned} V(\log \hat{\pi}_j) &= \begin{pmatrix} 1/\pi_1 & 0 & & \\ & 1/\pi_2 & & \\ 0 & & 1/\pi_3 & \\ & & & 1/\pi_4 \end{pmatrix} \text{Var}(\hat{\pi}_j - \pi_j) \begin{pmatrix} 1/\pi_1 & & & \\ & 1/\pi_2 & & \\ & & 1/\pi_3 & \\ & & & 1/\pi_4 \end{pmatrix}' \\ &= \frac{1}{n} \begin{pmatrix} 1/\pi_1 & 0 & & \\ & 1/\pi_2 & & \\ 0 & & 1/\pi_3 & \\ & & & 1/\pi_4 \end{pmatrix} \begin{pmatrix} \pi_1(1-\pi_1) & -\pi_1\pi_2 & -\pi_1\pi_3 & -\pi_1\pi_4 \\ -\pi_2\pi_1 & \pi_2(1-\pi_2) & -\pi_2\pi_3 & -\pi_2\pi_4 \\ -\pi_3\pi_1 & -\pi_3\pi_2 & \pi_3(1-\pi_3) & -\pi_3\pi_4 \\ -\pi_4\pi_1 & -\pi_4\pi_2 & -\pi_4\pi_3 & \pi_4(1-\pi_4) \end{pmatrix} \\ &\quad \begin{pmatrix} 1/\pi_1 & 0 & & \\ & 1/\pi_2 & & \\ 0 & & 1/\pi_3 & \\ & & & 1/\pi_4 \end{pmatrix}' \end{aligned}$$

$$\text{since } \text{Var}(\hat{\pi}_j) = \frac{1}{n^2} \text{Var}(n_j) = \frac{1}{n} \begin{pmatrix} \pi_1(1-\pi_1) & -\pi_1\pi_2 & -\pi_1\pi_3 & -\pi_1\pi_4 \\ -\pi_2\pi_1 & \pi_2(1-\pi_2) & -\pi_2\pi_3 & -\pi_2\pi_4 \\ -\pi_3\pi_1 & -\pi_3\pi_2 & \pi_3(1-\pi_3) & -\pi_3\pi_4 \\ -\pi_4\pi_1 & -\pi_4\pi_2 & -\pi_4\pi_3 & \pi_4(1-\pi_4) \end{pmatrix}.$$

Next  $nV(\log \hat{\pi}_j)$

$$\begin{aligned} &= \begin{pmatrix} 1/\pi_1 & 0 & & \\ & 1/\pi_2 & & \\ 0 & & 1/\pi_3 & \\ & & & 1/\pi_4 \end{pmatrix} \begin{pmatrix} (1-\pi_1) & -\pi_1 & -\pi_1 & -\pi_1 \\ -\pi_2 & (1-\pi_2) & -\pi_2 & -\pi_2 \\ -\pi_3 & -\pi_3 & (1-\pi_3) & -\pi_3 \\ -\pi_4 & -\pi_4 & -\pi_4 & (1-\pi_4) \end{pmatrix} \\ &= \begin{pmatrix} (1-\pi_1)/\pi_1 & -1 & -1 & -1 \\ -1 & (1-\pi_2)/\pi_2 & -1 & -1 \\ -1 & -1 & (1-\pi_3)/\pi_3 & -1 \\ -1 & -1 & -1 & (1-\pi_4)/\pi_4 \end{pmatrix} \quad (4) \end{aligned}$$

## Theorem

For the multiplicative hypothesis,

$$\text{Var}(M) = \left( \frac{1}{h_1} + \frac{1}{h_2} + \frac{1}{h_3} + \frac{1}{h_4} \right) + \left( \frac{1}{k_1} + \frac{1}{k_2} + \frac{1}{k_3} + \frac{1}{k_4} \right)$$

## Proof

The multiplicative hypothesis may be written as

$$M = \begin{pmatrix} 1 & -1 & -1 & 1 \end{pmatrix} \begin{pmatrix} \log \hat{\pi}_1 \\ \log \hat{\pi}_2 \\ \log \hat{\pi}_3 \\ \log \hat{\pi}_4 \end{pmatrix} - \begin{pmatrix} 1 & -1 & -1 & 1 \end{pmatrix} \begin{pmatrix} \log \hat{\theta}_1 \\ \log \hat{\theta}_2 \\ \log \hat{\theta}_3 \\ \log \hat{\theta}_4 \end{pmatrix}$$

Hence

$$\begin{aligned} \text{Var}(M) &= \frac{1}{h} \begin{pmatrix} 1 & -1 & -1 & 1 \end{pmatrix} \\ &\quad \begin{pmatrix} (1 - \pi_1)/\pi_1 & -1 & -1 & -1 \\ -1 & (1 - \pi_2)/\pi_2 & -1 & -1 \\ -1 & -1 & (1 - \pi_3)/\pi_3 & -1 \\ -1 & -1 & -1 & (1 - \pi_4)/\pi_4 \end{pmatrix} \begin{pmatrix} 1 \\ -1 \\ -1 \\ 1 \end{pmatrix} \\ &+ \frac{1}{k} \begin{pmatrix} 1 & -1 & -1 & 1 \end{pmatrix} \\ &\quad \begin{pmatrix} (1 - \theta_1)/\theta_1 & -1 & -1 & -1 \\ -1 & (1 - \theta_2)/\theta_2 & -1 & -1 \\ -1 & -1 & (1 - \theta_3)/\theta_3 & -1 \\ -1 & -1 & -1 & (1 - \theta_4)/\theta_4 \end{pmatrix} \begin{pmatrix} 1 \\ -1 \\ -1 \\ 1 \end{pmatrix} \\ &= \frac{1}{h} \left[ \begin{array}{l} (1) \quad [(1 - \pi_1)/\pi_1 + 1 + 1 - 1] \quad + \\ (-1) \quad [-1 - (1 - \pi_2)/\pi_2 + 1 - 1] \quad + \\ (-1) \quad [-1 + 1 - (1 - \pi_3)/\pi_3 - 1] \quad + \\ (1) \quad [-1 + 1 + 1 + (1 - \pi_4)/\pi_4] \end{array} \right] + \text{similar terms in } \theta \\ &= \frac{1}{h} \left[ \frac{(1 - \pi_1)}{\pi_1} + 1 + \frac{(1 - \pi_2)}{\pi_2} + 1 + \frac{(1 - \pi_3)}{\pi_3} + 1 + \frac{(1 - \pi_4)}{\pi_4} + 1 \right] + \\ &\quad \text{similar terms in } \theta \\ &= \frac{1}{h} \left[ \frac{1}{\pi_1} + \frac{1}{\pi_2} + \frac{1}{\pi_3} + \frac{1}{\pi_4} \right] + \text{similar terms in } \theta \end{aligned} \tag{5}$$

Hence

$$\begin{aligned} Var(M) &= \left[ \left( \frac{1}{h\pi_1} + \frac{1}{h\pi_2} + \frac{1}{h\pi_3} + \frac{1}{h\pi_4} \right) + \left( \frac{1}{k\theta_1} + \frac{1}{k\theta_2} + \frac{1}{k\theta_3} + \frac{1}{k\theta_4} \right) \right] \\ &= \left( \frac{1}{h_1} + \frac{1}{h_2} + \frac{1}{h_3} + \frac{1}{h_4} \right) + \left( \frac{1}{k_1} + \frac{1}{k_2} + \frac{1}{k_3} + \frac{1}{k_4} \right). \end{aligned} \quad (6)$$

Q.E.D.

## 2.2 Woolf's Method

A similar result for the variance of a log odds-ratio itself, is derived as follows. Consider  $\log(\pi_4\theta_1/\pi_1\theta_4)$ , the log-odds ratio for the relative risk at the combined exposure level in a case control study. We have

$$\log RR_{A\&B} = \log(\pi_4\theta_1/\pi_1\theta_4) = \log(\pi_4/\pi_1) - \log(\theta_4/\theta_1)$$

Next, without loss of generality assume that  $\pi_1$  and  $\pi_4$  have been normalized so that  $\pi_1 + \pi_4 = 1$  (This may be accomplished by setting  $\pi'_1 = \pi_1/(\pi_1 + \pi_4)$  and  $\pi'_4 = \pi_4/(\pi_1 + \pi_4)$ ). Now  $\pi'_1 + \pi'_4 = 1$  and the log odds-ratio remains unchanged since

$$\log RR_{A\&B} = \log(\pi'_4\theta'_1/\pi'_1\theta'_4) = \log(\pi_4\theta_1/\pi_1\theta_4).$$

The expression for  $\log(\pi_4/\pi_1)$  is in the form  $\log\left(\frac{\rho}{1-\rho}\right)$  where  $\rho = \pi_4$  and  $(1-\rho) = \pi_1$ . A Taylor's series expansion of  $\log\left(\frac{\rho}{1-\rho}\right)$  demonstrates that:

$$\begin{aligned} \log \frac{\rho}{1-\rho} &= \log \frac{\rho_0}{1-\rho_0} + \frac{1-\rho}{\rho} \left[ \frac{(1-\rho) + \rho}{(1-\rho)^2} \right] \Bigg|_{\rho_0} \cdot (\rho - \rho_0) \\ &= \log \frac{\rho_0}{1-\rho_0} + \frac{\rho - \rho_0}{\rho_0(1-\rho_0)}. \end{aligned} \quad (7)$$

Next

$$\begin{aligned} Var\left(\log \frac{\hat{\rho}}{(1-\hat{\rho})}\right) &= \left( \frac{1}{\hat{\rho}(1-\hat{\rho})} \right)^2 \frac{\hat{\rho}(1-\hat{\rho})}{N} \\ &= \frac{1}{N\hat{\rho}(1-\hat{\rho})} \end{aligned} \quad (8)$$

where  $\hat{\rho} = \frac{1}{N} \sum_{t=1}^N \delta_t$  is the unbiased estimator of  $\rho$ ,  $E(\hat{\rho}) = \rho$  and  $Var(\hat{\rho}) = \hat{\rho}(1-\hat{\rho})/N$  and  $N$  is the number of independent trials resulting in  $\sum_{t=1}^N \delta_t$  exposure cases (as compared to non-exposure cases). Similar expressions follow for the theta distribution. Now

$$\begin{aligned} Var(\log(\hat{\pi}_4/\hat{\pi}_1)) &= \frac{1}{h\hat{\pi}_4\hat{\pi}_1} = \frac{h}{(h\hat{\pi}_4)(h\hat{\pi}_1)} \\ &= \frac{h_1 + h_4}{h_1 h_4} \\ &= \frac{1}{h_4} + \frac{1}{h_1}. \end{aligned} \quad (9)$$

Then

$$\text{Var}(\log RR_{A\&B}) = \frac{1}{h_1} + \frac{1}{h_4} + \frac{1}{k_1} + \frac{1}{k_4}$$

Note that the repeated application of this result (assuming independence) to the multiplicative hypothesis would not produce the correct result in a case control setting because  $RR_{A\&B}$ ,  $RR_A$  and  $RR_B$  are mutually correlated.

The result we have just shown is also known as Woolf's method and is sometimes written

$$\text{Var} \left( \log \frac{AD}{BC} \right) = \frac{1}{A} + \frac{1}{B} + \frac{1}{C} + \frac{1}{D}$$

where  $RR = AD/BC$  and  $A$  denotes the number of cases with exposure,  $B$  denotes cases without exposure,  $C$  denotes controls with exposure, and  $D$  denotes controls without exposure.

It is also possible to derive the covariances of the relative risk measures. Consider  $RR_A = \pi_2\theta_1/\pi_1\theta_2$  and  $RR_{A\&B} = \pi_4\theta_1/\pi_1\theta_4$ . Then

$$\log RR_A = (\log \pi_2 - \log \pi_1) - (\log \theta_2 - \log \theta_1)$$

and

$$\log RR_{A\&B} = (\log \pi_4 - \log \pi_1) - (\log \theta_4 - \log \theta_1).$$

Clearly these are correlated because the common components. Consider the  $\pi$  components first (analogous results apply to the  $\theta$  components). Recall that

$$\text{Var}(\log \hat{\pi}) = \frac{1}{h} \begin{pmatrix} (1 - \pi_1)/\pi_1 & -1 & -1 & -1 \\ -1 & (1 - \pi_2)/\pi_2 & -1 & -1 \\ -1 & -1 & (1 - \pi_3)/\pi_3 & -1 \\ -1 & -1 & -1 & (1 - \pi_4)/\pi_4 \end{pmatrix}$$

But  $\log \pi_2 - \log \pi_1 = \begin{pmatrix} -1 & 1 & 0 & 0 \end{pmatrix} \begin{bmatrix} \log \pi_1 \\ \log \pi_2 \\ \log \pi_3 \\ \log \pi_4 \end{bmatrix}$  so that

$$\text{Var}(\log \pi_2 - \log \pi_1) =$$

$$\begin{aligned} & \left( \frac{1}{h} \right) \begin{pmatrix} -1 & 1 & 0 & 0 \end{pmatrix} \\ & \begin{pmatrix} (1 - \pi_1)/\pi_1 & -1 & -1 & -1 \\ -1 & (1 - \pi_2)/\pi_2 & -1 & -1 \\ -1 & -1 & (1 - \pi_3)/\pi_3 & -1 \\ -1 & -1 & -1 & (1 - \pi_4)/\pi_4 \end{pmatrix} \begin{pmatrix} -1 \\ 1 \\ 0 \\ 0 \end{pmatrix} \\ & = \left( \frac{1}{h} \right) \begin{pmatrix} -1 & 1 & 0 & 0 \end{pmatrix} \begin{bmatrix} -(1 - \pi_1)/\pi_1 - 1 \\ 1 + (1 - \pi_2)/\pi_2 \\ 0 \\ 0 \end{bmatrix} \end{aligned}$$

$$\begin{aligned}
&= \frac{1}{h} \left( \frac{(1 - \pi_1)}{\pi_1} + 1 + \frac{(1 - \pi_2)}{\pi_2} + 1 \right) \\
&= \left( \frac{1}{h} \right) \left( \frac{1}{\pi_1} \right) + \left( \frac{1}{h} \right) \left( \frac{1}{\pi_2} \right) = \frac{1}{h_1} + \frac{1}{h_2}
\end{aligned} \tag{10}$$

Combining this with the analogous result for  $\log \theta_2 - \log \theta_1$ , we obtain:

$$\text{Var}(\log RR_A) = \frac{1}{h_1} + \frac{1}{h_2} + \frac{1}{k_1} + \frac{1}{k_2}$$

This is exactly the Woolf result shown above. Similarly:

$$\text{Var}(\log RR_{A\&B}) = \frac{1}{h_1} + \frac{1}{h_4} + \frac{1}{k_1} + \frac{1}{k_4}$$

and

$$\text{Var}(\log RR_B) = \frac{1}{h_1} + \frac{1}{h_3} + \frac{1}{k_1} + \frac{1}{k_3}$$

Next consider the covariance between  $\log RR_A$  and  $\log RR_{A\&B}$ . Again we consider the  $\pi$  terms first. Using the fact that  $\text{cov}(t'x, s'x) = t'\text{Var}(x)s$  for conformable column vectors, we have, (for the  $\pi$  terms only)

$$\begin{aligned}
\text{cov}[\log RR_A, \log RR_B] &= \left( \frac{1}{h} \right) \begin{pmatrix} -1 & 1 & 0 & 0 \end{pmatrix} \cdot \\
&\begin{pmatrix} (1 - \pi_1)/\pi_1 & -1 & -1 & -1 \\ -1 & (1 - \pi_2)/\pi_2 & -1 & -1 \\ -1 & -1 & (1 - \pi_3)/\pi_3 & -1 \\ -1 & -1 & -1 & (1 - \pi_4)/\pi_4 \end{pmatrix} \begin{pmatrix} -1 \\ 0 \\ 0 \\ 1 \end{pmatrix} \\
&= \left( \frac{1}{h} \right) \begin{pmatrix} -1 & 1 & 0 & 0 \end{pmatrix} \begin{bmatrix} -(1 - \pi_1)/\pi_1 - 1 \\ 1 - 1 \\ 1 - 1 \\ 1 + (1 - \pi_4)/\pi_4 \end{bmatrix} \\
&= \left( \frac{1}{h} \right) \left( \frac{1}{\pi_1} \right) = \frac{1}{h_1}
\end{aligned} \tag{11}$$

A similar covariance term can be derived for the  $\theta$  terms.

$$\text{Thus } \text{cov}[\log RR_A, \log RR_{A\&B}] = \frac{1}{h_1} + \frac{1}{k_1}.$$

Combining analogous results for all log-odds ratios we obtain:

$$\text{Var} \begin{bmatrix} \log RR_A \\ \log RR_B \\ \log RR_{A\&B} \end{bmatrix} = \begin{bmatrix} \frac{1}{h_1} + \frac{1}{h_2} + \frac{1}{k_1} + \frac{1}{k_2} & \frac{1}{h_1} + \frac{1}{k_1} & \frac{1}{h_1} + \frac{1}{k_1} \\ \frac{1}{h_1} + \frac{1}{k_1} & \frac{1}{h_1} + \frac{1}{h_3} + \frac{1}{k_1} + \frac{1}{k_3} & \frac{1}{h_1} + \frac{1}{k_1} \\ \frac{1}{h_1} + \frac{1}{k_1} & \frac{1}{h_1} + \frac{1}{k_1} & \frac{1}{h_1} + \frac{1}{h_4} + \frac{1}{k_1} + \frac{1}{k_4} \end{bmatrix}$$

We now apply these results to derive the variance of the multiplicative statistic,  $M$ . We have

$$\begin{aligned} M &= \log RR_{A\&B} - \log RR_A - \log RR_B \\ &= \begin{pmatrix} -1 & -1 & 1 \end{pmatrix} \begin{bmatrix} \log RR_A \\ \log RR_B \\ \log RR_{A\&B} \end{bmatrix}. \end{aligned}$$

Hence  $Var(M) =$

$$\begin{aligned} &\begin{pmatrix} -1 & -1 & 1 \end{pmatrix} \begin{bmatrix} \frac{1}{h_1} + \frac{1}{h_2} + \frac{1}{k_1} + \frac{1}{k_2} & \frac{1}{h_1} + \frac{1}{k_1} & \frac{1}{h_1} + \frac{1}{k_1} \\ \frac{1}{h_1} + \frac{1}{k_1} & \frac{1}{h_1} + \frac{1}{h_3} + \frac{1}{k_1} + \frac{1}{k_3} & \frac{1}{h_1} + \frac{1}{k_1} \\ \frac{1}{h_1} + \frac{1}{k_1} & \frac{1}{h_1} + \frac{1}{k_1} & \frac{1}{h_1} + \frac{1}{h_4} + \frac{1}{k_1} + \frac{1}{k_4} \end{bmatrix} \begin{pmatrix} -1 \\ -1 \\ 1 \end{pmatrix} \\ &= \begin{pmatrix} -1 & -1 & 1 \end{pmatrix} \begin{bmatrix} -(\frac{1}{h_1} + \frac{1}{h_2} + \frac{1}{k_1} + \frac{1}{k_2}) \\ -(\frac{1}{h_1} + \frac{1}{h_3} + \frac{1}{k_1} + \frac{1}{k_3}) \\ (\frac{1}{h_1} + \frac{1}{h_4} + \frac{1}{k_1} + \frac{1}{k_4}) - (\frac{1}{h_1} + \frac{1}{k_1}) - (\frac{1}{h_1} + \frac{1}{k_1}) \end{bmatrix} \\ &= \left( \frac{1}{h_1} + \frac{1}{h_2} + \frac{1}{h_3} + \frac{1}{h_4} \right) + \left( \frac{1}{k_1} + \frac{1}{k_2} + \frac{1}{k_3} + \frac{1}{k_4} \right) \end{aligned} \quad (12)$$

Hence this formula for  $Var(\log M)$  agrees with our previous derivation.

To test the multiplicative hypothesis, we note that  $\log M$  should be zero if the multiplicative hypothesis is true. Therefore we can perform a Wald test using the ratio of  $\log(M)$  to its standard error  $\sqrt{Var(\log M)}$ . This will have an asymptotic normal distribution. (Rao, *Linear Statistical Inference and its Applications*).

### 2.3 Hypothesis Tests for Case-Control Studies – Additive Case

We next consider the additive hypothesis which may be stated:

$$A = RR_{A\&B} - (RR_A + RR_B - 1) = 0$$

i.e. that the relative risk of dying from contaminants  $A\&B$  is equal to the sum of the relative risks from  $A$  and  $B$  separately less one. To derive a variance for the statistic  $A$ , we note that

$$\begin{aligned} Var(A) &= Var(RR_{A\&B}) + Var(RR_A) + Var(RR_B) \\ &\quad - 2cov(RR_{A\&B}, RR_A + RR_B) \\ &= Var(RR_{A\&B}) + Var(RR_A) + Var(RR_B) \\ &\quad + 2cov(RR_A, RR_B) - 2cov(RR_{A\&B}, RR_A) \\ &\quad - 2cov(RR_{A\&B}, RR_B) \end{aligned} \quad (13)$$

In the derivations presented thus far, we have found expressions for the variances and covariances of log relative risks. Clearly the additive hypothesis requires variances and covariances of the relative risks themselves. One approach is to develop confidence intervals for the log relative risks and translate them into confidence intervals for the relative risks by exponentiating the terms in the confidence interval inequality. In the presence of correlation, however, the best one can achieve with this technique are broad intervals based on the Bonferroni inequalities. A second approach uses the fact that if the log relative risks are approximately normal than the relative risks are approximately log normally distributed. Again the joint distribution of log normal random variables is not straightforward and this approach similarly becomes unworkable.

Instead we follow Rothman (1976) and rely on a Taylor's series expansion. Specifically, we approximate the logarithm using:

$$\log y \doteq \log y_0 + \frac{1}{y_0}(y - y_0) \quad \text{so that} \quad \text{Var}(\log y) \doteq \frac{1}{y_0^2} \text{Var}(y) . \quad \text{Hence:}$$

$$\text{Var}(y) = y_0^2 \text{Var}(\log y)$$

The accuracy of the approximation improves for  $y$  close to  $y_0$  which we will achieve by taking  $y$  to be a consistent estimate of  $y_0$ .

Collecting the terms required for the variance of the additive statistic,  $\text{Var}(A)$ , we have:

$$\begin{aligned} \text{Var}(RR_{A\&B}) &\doteq (RR_{A\&B})^2 \cdot \left[ \frac{1}{h_1} + \frac{1}{h_4} + \frac{1}{k_1} + \frac{1}{k_4} \right] \\ \text{Var}(RR_A) &\doteq (RR_B)^2 \cdot \left[ \frac{1}{h_1} + \frac{1}{h_2} + \frac{1}{k_1} + \frac{1}{k_2} \right] \\ \text{Var}(RR_B) &\doteq (RR_b)^2 \cdot \left[ \frac{1}{h_1} + \frac{1}{h_3} + \frac{1}{k_1} + \frac{1}{k_3} \right] \end{aligned}$$

For the covariance terms we employ similar Taylor's expansions. Specifically let:

$$\log y \doteq \log y_0 + \frac{1}{y_0}(y - y_0) \quad \text{and}$$

$$\log z \doteq \log z_0 + \frac{1}{z_0}(z - z_0) . \quad \text{Then}$$

$$\text{cov}(\log y, \log z) \doteq \frac{1}{y_0 z_0} \text{cov}(y - y_0, z - z_0) \quad \text{so that}$$

$$\text{cov}(y, z) \doteq (y_0 z_0) \cdot \text{cov}(\log y, \log z) . \quad \text{Then:}$$

$$\text{cov}(RR_A, RR_B) = (RR_A \cdot RR_B) \cdot \left( \frac{1}{h_1} + \frac{1}{k_1} \right)$$

$$\text{cov}(RR_{A\&B}, RR_A) = (RR_{A\&B} \cdot RR_A) \cdot \left(\frac{1}{h_1} + \frac{1}{k_1}\right)$$

$$\text{cov}(RR_{A\&B}, RR_B) = (RR_{A\&B} \cdot RR_B) \cdot \left(\frac{1}{h_1} + \frac{1}{k_1}\right)$$

Then we have:

$$\begin{aligned} \text{Var}(A) &\doteq (RR_{A\&B})^2 \cdot \left[\frac{1}{h_1} + \frac{1}{h_4} + \frac{1}{k_1} + \frac{1}{k_4}\right] \\ &\quad + (RR_A)^2 \cdot \left[\frac{1}{h_1} + \frac{1}{h_2} + \frac{1}{k_1} + \frac{1}{k_2}\right] \\ &\quad + (RR_B)^2 \cdot \left[\frac{1}{h_1} + \frac{1}{h_3} + \frac{1}{k_1} + \frac{1}{k_3}\right] \\ &\quad + 2(RR_A \cdot RR_B) \cdot \left(\frac{1}{h_1} + \frac{1}{k_1}\right) \\ &\quad - 2(RR_{A\&B} \cdot RR_A) \cdot \left(\frac{1}{h_1} + \frac{1}{k_1}\right) \\ &\quad - 2(RR_{A\&B} \cdot RR_B) \cdot \left(\frac{1}{h_1} + \frac{1}{k_1}\right) \end{aligned} \tag{14}$$

A Wald test may be conducted using the ratio of  $A$  to its standard error  $\sqrt{\text{Var}(A)}$ . Asymptotically this will be standard normal, given the limiting distribution of the joint multinomial probabilities for  $\pi$  and  $\theta$ .

Since the Wald tests are valid only asymptotically we also consider a likelihood ratio approach.

## 2.4 Maximum Likelihood

The likelihood function for the case control study is  $\prod_{i=1}^4 \pi_i^{h_i} \theta_i^{k_i}$  and is maximized subject to the constraint  $\sum \pi_i - 1 = \sum \theta_i - 1 = 0$ . The log likelihood function is

$$F = \sum h_i \log \pi_i + \sum k_i \log \theta_i .$$

This is maximized subject to the constraints:

$$F_1 = \sum \pi_i - 1 = 0 \tag{15}$$

$$F_2 = \sum \theta_i - 1 = 0 \tag{16}$$

$$\text{and } F_3 = \frac{\pi_4 \theta_1}{\pi_1 \theta_4} - \frac{\pi_2 \theta_1}{\pi_1 \theta_2} - \frac{\pi_3 \theta_1}{\pi_1 \theta_3} + 1 = 0 \text{ "additivity" or}$$

$$F_4 = \log(\pi_1 \theta_2 \theta_3 \pi_4) - \log(\theta_1 \pi_2 \pi_3 \theta_4) = 0 \text{ "multiplicativity"}$$

Note that  $F_3$  may be rewritten:

$$\begin{aligned} F_3 &= \frac{\pi_4}{\theta_4} - \frac{\pi_2}{\theta_2} - \frac{\pi_3}{\theta_3} + \frac{\pi_1}{\theta_1} \\ &= \frac{\pi_1}{\theta_1} - \frac{\pi_2}{\theta_2} - \frac{\pi_3}{\theta_3} + \frac{\pi_4}{\theta_4} = 0 \end{aligned} \quad (17)$$

### 2.4.1 Additive Constraint

For the additive model we maximize the Lagrangian

$$F_A = F + \mu_1 F_1 + \mu_2 F_2 + \mu_3 F_3$$

where  $\mu_1$ ,  $\mu_2$ , and  $\mu_3$  are Lagrange multipliers. The first order conditions are:

$$\frac{\partial F_A}{\partial \pi_i} = \frac{h_i}{\pi_i} + \mu_1 + \frac{\delta_i \mu_3}{\theta_i} = 0 \quad (18)$$

$$\frac{\partial F_A}{\partial \theta_i} = \frac{k_i}{\theta_i} + \mu_2 - \frac{\delta_i \pi_i \mu_3}{\theta_i^2} = 0 \quad \text{and} \quad (19)$$

$$\frac{\partial F_A}{\partial \mu_i} = F_i = 0 \quad i = 1, 2, 3 \quad (20)$$

where  $\delta_1 = -\delta_2 = -\delta_3 = \delta_4 = 1$ .

It follows that:

$$\sum \pi_i \frac{\partial F_A}{\partial \pi_i} = h + \mu_1 = 0 \quad \text{and} \quad \sum \theta_i \frac{\partial F_A}{\partial \theta_i} = k + \mu_2 = 0$$

Hence  $\hat{\mu}_1 = -h$  and  $\hat{\mu}_2 = -k$  and the remaining conditions may be written:

$$(h_i - h \hat{\pi}_i) \hat{\theta}_i + \delta_i \hat{\pi}_i \hat{\mu}_3 = 0$$

$$(k_i - k \hat{\theta}_i) \hat{\theta}_i - \delta_i \hat{\pi}_i \hat{\mu}_3 = 0$$

$$\text{and} \quad \sum \delta_i \frac{\hat{\pi}_i}{\hat{\theta}_i} = 0$$

Writing  $x_i = \hat{\pi}_i / \hat{\theta}_i$  and solving the first order conditions implies:

$$x_i = \frac{(\delta_i k_i h - k \hat{\mu}_3) \pm \sqrt{(\delta_i k_i h - k \hat{\mu}_3)^2 - 4 \delta_i h_i h k \hat{\mu}_3}}{2 \hat{\mu}_3 h} \quad (21)$$

Since  $\sum_i \delta_i x_i = 0$  it follows that:

$$\begin{aligned} 0 &= \sum_i \left[ (k_i h - k \hat{\mu}_3 \delta_i) \pm \delta_i \sqrt{(\delta_i k_i h - k \hat{\mu}_3)^2 - 4 \delta_i h_i h k \hat{\mu}_3} \right] \\ &= k h + \sum_i \pm \delta_i \sqrt{(\delta_i k_i h - k \hat{\mu}_3)^2 - 4 \delta_i h_i h k \hat{\mu}_3} \end{aligned} \quad (22)$$

This equation in  $\hat{\mu}_3$  may be solved for each of 16 possible sign combinations (+ or - for each of the four terms in the sum).

Using  $x_i = \pi_i / \theta_i$ , the first two first order conditions may be written

$$(k_i - k \theta_i) = \delta_i x_i \mu_3 \quad \text{and} \quad (h_i - h \pi_i) = -\delta_i x_i \mu_3$$

Hence  $(k_i - k \theta_i) = -(h_i - h \pi_i)$  which implies

$$k_i + h_i = k \theta_i + h \pi_i = \theta_i (k + h x_i) \quad \text{or}$$

$$\theta_i = \frac{k_i + h_i}{k + h x_i}. \quad (23)$$

Now substitute into the first order condition:

$$k_i - k \left[ \frac{k_i + h_i}{k + h x_i} \right] = \delta_i x_i \mu_3 \quad \text{or}$$

$$k_i (k + h x_i) - k (k_i + h_i) = (k + h x_i) \delta_i x_i \mu_3$$

$$k_i k + k_i x_i h - k (k_i + h_i) = k \delta_i x_i \mu_3 + x_i^2 h \delta_i \mu_3$$

$$-k h_i = x_i (-k_i h + k \delta_i \mu_3) + x_i^2 h \delta_i \mu_3$$

$$-k h_i = x_i (k \delta_i \mu_3 - k_i h) + x_i^2 h \delta_i \mu_3$$

$$-k h_i \delta_i = x_i (k \delta_i^2 \mu_3 - k_i h \delta_i) + x_i^2 h \delta_i^2 \mu_3$$

Now use  $\delta^2 = 1$  as  $\delta = 1$  or  $-1$ . Then:

$$-k h_i \delta_i = x_i (k \mu_i - k_i h \delta_i) + x_i^2 (h \mu_3) \quad \text{so that}$$

$$0 = x_i^2 (h \mu_3) + x_i (k \mu_3 - k_i h \delta_i) + k h_i \delta_i$$

The last equation establishes a bound on  $x_i$  since the discriminant of the quadratic equation must be positive. The discriminant is:

$$(k\mu_3 - k_i h \delta_i)^2 - 4(h\mu_3)(k h_i \delta_i) \geq 0$$

$$k_i^2 h^2 - 2\delta_i k_i h k \mu_3 + k^2 \mu_3^2 - 2 \cdot 2\delta_i h_i h k \mu_3 \geq 0$$

$$\mu_3^2 - 2\delta_i \left(\frac{h}{k}\right) \mu_3 (k_i + 2h_i) + k_i^2 \left(\frac{h}{k}\right)^2 \geq 0$$

Next solve this quadratic at the point of equality to zero for  $\mu_3$ . We obtain:

$$\begin{aligned} \mu_{3i}^* &= \frac{2\delta_i \left(\frac{h}{k}\right) (k_i + 2h_i) \pm \sqrt{4\delta_i^2 \frac{h^2}{k^2} (k_i + 2h_i)^2 - 4k_i^2 \left(\frac{h}{k}\right)^2}}{2} \\ &= \delta_i \left(\frac{h}{k}\right) (k_i + 2h_i) \pm \sqrt{\frac{h^2}{k^2} (k_i + 2h_i)^2 - k_i^2 \left(\frac{h}{k}\right)^2} \\ &= \left(\frac{h}{k}\right) \left[ (k_i + 2h_i)\delta_i \pm \sqrt{k_i^2 + 4k_i h_i + 4h_i^2 - k_i^2} \right] \\ &= \left(\frac{h}{k}\right) \left[ \delta_i (k_i + 2h_i) \pm 2\sqrt{h_i (h_i + k_i)} \right] \end{aligned} \quad (24)$$

Since the quadratic has a positive second derivative, the inequalities are  $\mu_3 \leq \min \mu_{3i}^*$  and  $\mu_3 \geq \max \mu_{3i}^*$ .

Setting,  $a_i = \left(\frac{h}{k}\right) \left[ k_i + 2h_i - 2\sqrt{h_i^2 + h_i k_i} \right]$ , Gardner and Munford (1980) show that  $-\min(a_2, a_3) \leq \hat{\mu}_3 \leq \min(a_1, a_4)$ .

Unfortunately, while these bounds bracket the true value of  $\hat{\mu}_3$  they are not guaranteed to produce sign changes in the equation of interest so that an iterative solution is required to bracket each of the solutions for  $\hat{\mu}_3$ . We have found that  $\hat{\mu}_3 = 0$  will always be a trivial solution to the equation above and should be ignored.

Once  $\hat{\mu}_3$  is found  $\hat{\pi}_i$  and  $\hat{\theta}_i$  are found from the first order conditions.

## 2.4.2 Multiplicative Constraint

For the multiplicative model we maximize the Lagrangian:

$$F_M = F + \lambda_1 F_1 + \lambda_2 F_2 + \lambda_4 F_4$$

$$\begin{aligned} \text{with } F &= \sum h_i \log \pi_i + \sum k_i \log \theta_i \quad \text{and} \\ F_1 &= \sum \pi_i - 1 \\ F_2 &= \sum \theta_i - 1 \\ F_4 &= \sum \delta_i \log \pi_i - \sum \delta_i \log \theta_i \end{aligned}$$

we have:

$$\begin{aligned} \frac{\partial F_M}{\partial \pi_i} &= \frac{h_i}{\pi_i} + \lambda_1 + \lambda_4 \frac{\delta_i}{\pi_i} = 0 \\ \frac{\partial F_M}{\partial \theta_i} &= \frac{k_i}{\theta_i} + \lambda_2 + \lambda_4 \frac{-\delta_i}{\theta_i} = 0. \quad \text{Then:} \\ \sum \pi_i \frac{\partial F_M}{\partial \pi_i} &= \sum h_i + \lambda_1 + \lambda_4 \sum \delta_i = 0 \Rightarrow \lambda_1 = -h \\ \sum \theta_i \frac{\partial F_M}{\partial \theta_i} &= \sum k_i + \lambda_2 + \lambda_4 \sum -\delta_i = 0 \Rightarrow \lambda_2 = -k \\ \frac{h_i}{\pi_i} - h + \frac{\lambda_4 \delta_i}{\pi_i} &= 0 \Rightarrow h_i - h\pi_i + \lambda_4 \delta_i = 0 \\ h\pi_i &= h_i + \lambda_4 \delta_i \\ \pi_i &= \frac{h_i + \delta_i \lambda_4}{h} \\ \frac{k_i}{\theta_i} + \lambda_2 + \lambda_4 \left( \frac{-\delta_i}{\theta_i} \right) &= 0 \Rightarrow \\ \frac{k_i}{\theta_i} - k + \lambda_4 \left( \frac{-\delta_i}{\theta_i} \right) &= 0 \Rightarrow \\ k_i - k\theta_i + \lambda_4 (-\delta_i) &= 0 \\ -k\theta_i &= \lambda_4 \delta_i - k_i \\ \theta_i &= \frac{\lambda_4 \delta_i - k_i}{-k} = \frac{k_i - \lambda_4 \delta_i}{h} \end{aligned}$$

Finally, substituting into the constraint implies:

$$\left[ \frac{h_1 + \delta_1 \lambda_4}{h} \right] \left[ \frac{k_2 - \delta_2 \lambda_4}{k} \right] \left[ \frac{k_3 + \delta_3 \lambda_4}{k} \right] \left[ \frac{h_4 + \delta_4 \lambda_4}{h} \right] - \text{similar terms} = 0$$

which implies  $(h_1 + \lambda_4)(k_2 + \lambda_4)(k_3 + \lambda_4)(h_4 + \lambda_4) - \text{similar terms} = 0$ .

### 2.4.3 Unconstrained Maximum Likelihood

The log likelihood under the constraint of additivity or multiplicativity is  $\sum h_i \log \hat{\pi}_i + \sum k_i \log \hat{\theta}_i$ . For the unconstrained case we maximize the Lagrangian

$$L = \sum h_i \log \pi_i + \sum k_i \log \theta_i + \psi_1 \left[ \sum \pi_i - 1 \right] + \psi_2 \left[ \sum \theta_i - 1 \right]$$

The first order conditions are

$$\frac{\partial L}{\partial \pi_i} = \frac{h_i}{\pi_i} + \psi_1 = 0 \quad \text{and} \quad \frac{\partial L}{\partial \psi_1} = \sum \pi_i - 1 = 0$$

$$\frac{\partial L}{\partial \theta_i} = \frac{k_i}{\theta_i} + \psi_2 = 0 \quad \text{and} \quad \frac{\partial L}{\partial \psi_2} = \sum \theta_i - 1 = 0$$

These equations imply that  $\hat{\pi}_i = h_i/h$  and  $\hat{\theta}_i = k_i/k$  for the unconstrained maximum likelihood.

Hypothesis tests may be based on  $-2(\log \text{likelihood unconstrained} - \log \text{likelihood constrained})$  which has a  $\chi^2$  distribution with one degree of freedom.

While the additive and multiplicative models are non-nested, a comparison of the log likelihood values provides a basis for a non-nested hypothesis test.

## 3 Bonferroni Intervals and Monte Carlo Simulations

### 3.1 Bonferroni Interval

The additive statistic  $A = RR_{A\&B} - RR_A - RR_B - 1$  is composed of three random variables. A confidence interval for each component may be established using the variance of the log-odds ratio. Set at appropriate levels these confidence intervals may be combined using basic results from probability theory. For a 95 percent confidence interval, chose a significance level such that one third of one half of 5 percent probability is in each tail of a normal distribution. Then:

$$\text{prob}[-2.39 \leq N(0, 1) \leq 2.39] = 1 - \frac{.05}{6} = 0.98334$$

Since  $(\log \hat{RR} - \log RR)/\sigma \sim^A N(0, 1)$  we have

$$\text{prob}[-2.39\sigma \leq \log \hat{RR} - \log RR \leq 2.39\sigma] = 0.98334$$

or

$$\text{prob}[-2.39\sigma + \log \hat{RR} \leq \log RR \leq 2.39\sigma + \log \hat{RR}] = 0.98334$$

so that

$$\text{prob}[\hat{RR}e^{-2.39\sigma} \leq RR \leq \hat{RR}e^{2.39\sigma}] = 0.98334$$

Similarly,

$$\text{prob}[\hat{RR}_{A\&B}e^{-2.39\sigma_{RR_{A\&B}}} \leq RR \leq \hat{RR}_{A\&B}e^{2.39\sigma_{RR_{A\&B}}}] = 0.98334$$

and so forth for  $\hat{RR}_A$  and  $\hat{RR}_B$ . Similarly:

$$\text{Prob}[C_{low}^{A\&B} \leq RR_{A\&B} \leq C_{high}^{A\&B}] = .98334$$

$$\text{Prob}[C_{low}^A \leq RR_A \leq C_{high}^A] = .98334$$

$$\text{Prob}[C_{low}^B \leq RR_B \leq C_{high}^B] = .98334$$

Denoting the intervals within square brackets as  $A, B$ , and  $C$  we have by the Bonferroni inequality:

$$\text{prob}[A \cap B \cap C] \geq 1 - (P(A^c) + P(B^c) + P(C^c))$$

Then

$$\begin{aligned} &\text{prob}[C_{low}^{A\&B} \leq RR_{A\&B} \leq C_{high}^{A\&B} \cap \\ &\quad -C_{high}^A \leq -RR_A \leq -C_{low}^A \cap \\ &\quad -C_{high}^B \leq -RR_B \leq -C_{low}^B] \geq 1 - .05 = .95 \end{aligned} \quad (25)$$

so that

$$\text{prob}[C_{low}^{A\&B} - C_{high}^A - C_{high}^B \leq RR_{A\&B} - RR_A - RR_B \leq C_{high}^{A\&B} - C_{low}^A - C_{low}^B] \geq .95$$

and

$$\text{prob}[C_{low}^{A\&B} - C_{high}^A - C_{high}^B - 1 \leq A \leq C_{high}^{A\&B} - C_{low}^A - C_{low}^B - 1] \geq .95$$

As noted before these intervals should be rejected in favor of Wald or Likelihood Ratio tests given the tendency for the intervals to be broad and imprecise.

### 3.2 Simulation Methods

Consistent estimates of the  $\pi_j$  and  $\theta_j$  are formed using  $h_j/h$  and  $k_j/k$  respectively. A Monte Carlo technique draws a random multinomial deviate with marginal probabilities  $\pi_j$  and  $\theta_j$ . Then the empirical distribution of the statistics  $M$  and  $A$  are formed using repeated simulations. The empirical distributions establish confidence intervals centered around the realized value of the statistic. If these confidence intervals contain zero then the hypothesis is not rejected.

## 4 Synergy Indices

### 4.1 Rothman's $S$ Index

Rothman (1976) considers the independently acting agents  $A$  and  $B$  and a background affect  $C$ .  $C$  is assumed to act independently of  $A$  and  $B$ .

Let  $P_T$  denote the probability that disease develops when both  $A$  and  $B$  are present in addition to the background  $C$ .  $P_A$  is probability that disease develops if  $A$  were to act in isolation (without background). We define  $P_B$  similarly.  $P_C$  is the probability of getting disease from background only. Then

$$\begin{aligned} P_T &= P[A \cup B \cup C] \\ &= P[A] + P[B] + P[C] - P[A \cap B] - P[A \cap C] - \\ &\quad P[B \cap C] + P[A \cap B \cap C] \end{aligned} \tag{26}$$

Now under independence we have:

$$P_T = P[A] + P[B] + P[C] - P[A]P[B] - P[A]P[C] - P[B]P[C] + P[A]P[B]P[C]$$

Let  $R_{AB} = P_T$  denote the combined risk.

$$\text{Let } R_A = P[A \cup C] = P[A] + P[C] - P[A]P[C]$$

$$\text{Let } R_B = P[B \cup C] = P[B] + P[C] - P[B]P[C]$$

$$\text{Let } R_0 = P[C]$$

Then under independence:

$$\begin{aligned} R_{AB} - R_0 &= (R_A - R_0) + (R_B - R_0) - \frac{P_A P_B (1 - P_C)(1 - P_C)}{(1 - P_C)} \\ &= (R_A - R_0) + (R_B - R_0) - \frac{(R_A - R_0)(R_B - R_0)}{(1 - R_0)} \end{aligned} \tag{27}$$

Rothman's synergy index is defined as the ratio of the left-hand side of this equation to the right-hand side.

$$S = \frac{(R_A - R_0)}{(R_A - R_0) + (R_B - R_0) + \frac{(R_A - R_0)(R_B - R_0)}{(1 - R_0)}}$$

Under independence, the numerator and denominator will be equal and the synergy index will equal one. Ignoring the product terms in the denominator which are likely to be small, Rothman's index becomes:

$$S = \frac{(R_{AB} - R_0)}{(R_A - R_0) + (R_B - R_0)} = \frac{RR_{AB} - 1}{RR_A + RR_B - 2}$$

where  $RR_{AB} = RR_{AB}/R_0$  etc. When  $S = 1$ , we obtain:

$$RR_{AB} - 1 = RR_A + RR_B - 2 \quad \text{or}$$

$$RR_{AB} = RR_A + RR_B - 1 \quad (28)$$

which we recognize as the additive hypothesis.

An alternative expression for Rothman's  $S$  index is

$$S = \frac{ERR_{AB}}{ERR_A + ERR_B}$$

where  $ERR_{AB} = RR_{AB} - 1$  and  $ERR_A = RR_A - 1$  etc. Here  $ERR$  denotes excess relative risk.

## 4.2 Attributable Proportion

The attributable proportion is defined as the excess relative risk compared to the additive model divided by the combined relative risk. Formally,

$$\begin{aligned}
 AP &= \frac{ERR_{AB} - (ERR_A + ERR_B)}{(ERR_{AB} + 1)} \\
 &= \frac{(R_{AB}/R_0 - 1) - ((R_A/R_0) + (R_B/R_0 - 1))}{[R_{AB}/R_0 - 1 + 1]} \\
 &= \frac{R_{AB} - R_0 - (R_A + R_B - 2R_0)}{R_{AB}} \\
 &= \frac{R_{AB} - (R_A + R_B - R_0)}{R_{AB}} \\
 &= \frac{RR_{AB} - (RR_A + RR_B - 1)}{RR_{AB}} \quad (29)
 \end{aligned}$$

When the additive model is correct,  $AP = 0$ .

Rothman's index  $S$  and the attributable proportion  $AP$  measure departure from additivity. They do not include the multiplicative hypothesis as a natural alternative. Therefore we consider an alternative which nests both hypotheses.

### 4.3 Additive-Multiplicative Measure

Define

$$\begin{aligned}\gamma &= \frac{[RR_{A\&B} - 1] - [RR_A - 1] - [RR_B - 1]}{[RR_A - 1][RR_B - 1]} \\ &= \frac{RR_{A\&B} - RR_A - RR_B + 1}{(RR_A - 1)(RR_B - 1)}\end{aligned}\tag{30}$$

Note that when  $\gamma = 0$  the additive hypothesis is true. When  $\gamma = 1$  we have:

$$RR_{A\&B} = RR_A - RR_B + 1 = RR_A RR_B - RR_A - RR_B + 1$$

which implies:  $RR_{A\&B} = RR_A \cdot RR_B$ , i.e. the multiplicative hypothesis.

While difficult, a confidence interval may be derived by examining the distribution of  $\log \gamma$ . Note that

$$\log \gamma = \log A - (\log (RR_A - 1) + \log (RR_B - 1))$$

where  $A$  is the additive statistic. Then

$$\begin{aligned}Var(\log \gamma) &= Var(\log A) - Var((\log (RR_A - 1)) + Var(\log (RR_B - 1))) \\ &\quad + 2cov[\log (RR_A - 1), \log (RR_B - 1)] \\ &\quad - 2cov[\log A, \log (RR_A - 1)] \\ &\quad - 2cov[\log A, \log (RR_B - 1)].\end{aligned}\tag{31}$$

For case-control studies, we have previously derived these components. However, the utility of the expansion is questionable given that when the additive hypothesis is true, the log transformation is not defined.

## 5 Cohort Studies

The cohort method compares the death rates between groups for those exposed to contaminant  $A$  (with or without exposure to  $B$ ) and for those not exposed to contaminant  $A$  (with or without exposure to  $B$ ). For present purposes contaminant  $A$  will be smoking while contaminant  $B$  will be asbestos. Death rates are calculated and given in the following  $2 \times 2$  table:

	non-smoking	smoking
asbestos	$d_A^{NS}$	$d_A^S$
non-asbestos	$d_{NA}^{NS}$	$d_{NA}^S$

The cohort method follows a group of individuals with some exposure to asbestos. Death rates are determined over time for this cohort. A sample of individuals from a non-asbestos exposed population is matched to the exposed population at the aggregate level (i.e. there is a similar number of individuals of each age group).

Before discussing the derivation of the death rates  $d_j^i$ , we note that cohort studies make each cell of the  $2 \times 2$  table independent by design. This greatly simplifies the hypothesis testing and determination of confidence intervals. First relative risks are determined as follows:

$$RR_A = \text{relative risk of asbestos exposure} = d_A^{NS} / d_{NA}^{NS}$$

$$RR_S = \text{relative risk of smoking exposure} = d_{NA}^S / d_{NA}^{NS}$$

$$RR_{AS} = \text{the relative risk of combined exposure} = d_A^S / d_{NA}^{NS}$$

The additive hypothesis is stated as:

$$RR_{AS} - RR_A - RR_S + 1 = 0$$

or

$$\frac{d_A^S}{d_{NA}^{NS}} = \frac{d_A^{NS}}{d_{NA}^{NS}} + \frac{d_{NA}^S}{d_{NA}^{NS}} - 1$$

or

$$d_A^S = d_A^{NS} + d_{NA}^S - d_{NA}^{NS}$$

or

$$A^* = d_A^{NS} + d_{NA}^S - d_A^S - d_{NA}^{NS}$$

Under additivity  $A^* = 0$ .

The multiplicative hypothesis is stated as:

$$RR_{AS} = RR_A \cdot RR_S = 0$$

or

$$\frac{d_A^S}{d_{NA}^{NS}} - \frac{d_A^{NS}}{d_{NA}^{NS}} \cdot \frac{d_{NA}^S}{d_{NA}^{NS}} = 0$$

or

$$d_A^S \cdot d_{NA}^{NS} - d_A^{NS} \cdot d_{NA}^S = 0 \quad (32)$$

or

$$\log d_A^S + \log d_{NA}^{NS} - \log d_A^{NS} - \log d_{NA}^S = 0$$

or

$$M^* = \log d_A^{NS} + \log d_{NA}^S - \log d_A^S - \log d_{NA}^{NS} = 0$$

We note that the multiplicative statistic is similar to the additive statistic with the exception that it is stated as a sum of logarithms. This suggests that the two hypotheses may be nested using a Box-Cox transformation.

It is worth noting that (32) implies

$$\frac{d_A^{NS}}{d_A^S} = \frac{d_{NA}^{NS}}{d_{NA}^S}$$

which states that the columns in the table are proportional to one another. Similarly, the rows are in proportion if the multiplicative hypothesis is correct. These are common statements of independence and can be tested via Pearson Chi-squared statistics for such tables. Finally, given the relationship between contingency tables and the log-linear model we should expect a direct test of the multiplicative hypothesis from the log-linear model.

Suppose  $\log (P[Y_1, Y_2]) = \mu_0 + \mu_1 Y_1 + \mu_2 Y_2 + \mu_{12} Y_1 \cdot Y_2$  Then

$$\log (P(0, 0)) = \mu_0$$

$$\log (P(0, 1)) = \mu_0 + \mu_2$$

$$\log (P(1, 0)) = \mu_0 + \mu_1$$

$$\log (P(1, 1)) = \mu_0 + \mu_1 + \mu_2 + \mu_{12}$$

If  $P(0, 0)$  is estimated by  $d_{NA}^{NS}$ ,  $P(1, 0)$  by  $d_A^{NS}$ ,  $P(0, 1)$  by  $d_{NA}^S$ , and  $P(1, 1)$  by  $d_A^S$  (after suitable normalization), then the multiplicative hypothesis may be stated as:

$$M^* = (\mu_0 + \mu_1) + (\mu_0 + \mu_2) - (\mu_0 + \mu_1 + \mu_2 + \mu_{12}) - (\mu_0) = -\mu_{12}$$

Then  $M^* = 0$  (the multiplicative hypothesis) if and only if the interaction parameter  $\mu_{12} = 0$  in the log-linear model.

## 5.1 Determination of Death Rates

The death rate is defined as the number of deaths per 100,000 person years. This is typically measured by the number of deaths observed in the cohort divided by the number of person years multiplied by 100,000.

For example, suppose that a particular cohort has  $N_A^{NS}$  individuals who are non-smokers but who are exposed to asbestos. Suppose that these  $N_A^{NS}$  individuals are followed for  $Y_A^{NS}$  person years (on average  $Y_A^{NS}/N_A^{NS}$  years per person). Suppose that  $h_A^{NS}$  of these individuals die during the period of observation. Then

$$d_A^{NS} = \left( \frac{h_A^{NS}}{N_A^{NS}} \right) \left( \frac{N_A^{NS}}{Y_A^{NS}} \right) \cdot 100,000.$$

The stochastic component in the expression is the binomially distributed random variable  $h_A^{NS}$  which denotes the number of observed deaths in  $N_A^{NS}$  trials. Let  $P_A^{NS}$  denote the true but unobserved probability of dying. Then  $\hat{P}_A^{NS} = h_A^{NS}/N_A^{NS}$  is a consistent estimate of  $P_A^{NS}$ .

$$\text{Now } E(\hat{P}_A^{NS}) = P_A^{NS} \text{ and } \text{Var}(\hat{P}_A^{NS}) = \frac{P_A^{NS}(1 - P_A^{NS})}{N_A^{NS}}.$$

Then

$$\text{Var}(d_A^{NS}) = \text{Var}(\hat{P}_A^{NS}) \cdot \left[ \frac{N_A^{NS}}{Y_A^{NS}} \right]^2 \cdot 100,000^2$$

When logarithmic transformations are employed we have

$$\log d_A^{NS} = \log \hat{P}_A^{NS} + \log \left[ \frac{N_A^{NS}}{Y_A^{NS}} \right] + \log(100,000).$$

Recall that a Taylor's series expansion shows that  $\log \hat{P} \doteq \log P_0 + \frac{1}{P_0}(\hat{P} - P_0)$  so that

$$\text{Var}(\log \hat{P}) = \frac{1}{P_0^2} \frac{P_0(1 - P_0)}{N} = \frac{(1 - P_0)}{P_0 N}$$

Then

$$\text{Var}(\log(d_A^{NS})) \doteq \frac{(1 - \hat{P}_A^{NS})}{\hat{P}_A^{NS} N_A^{NS}}$$

Before proceeding with the formula for the variance of the  $A^*$  and  $M^*$  statistics, we note that replacing  $P_A^{NS}$  by  $\hat{P}_A^{NS}$  in the variance formula is valid asymptotically. Some researchers have noted that it may be more accurate in small samples to use a chi-square approximation.

To do this we set  $\chi^2 = \frac{(\hat{P}-P)}{P(1-P)/N}$ . Then we set the  $\chi^2$  value to a critical level for the appropriate size test. Let  $\chi_r^2$  be the critical value. Then

$$\chi_r^2 = \frac{(\hat{P} - P)^2}{P(1 - P)/N}$$

so that

$$\begin{aligned} \hat{P} - 2\hat{P}P + P^2 &= \chi_r^2 P(1 - P)/N \\ &= \frac{P}{N}\chi_r^2 - \frac{P^2}{N}\chi_r^2 \end{aligned} \quad (33)$$

Then

$$P^2 \left( \frac{\chi_r^2}{N} + 1 \right) + P \left( \frac{-\chi_r^2}{N} - 2\hat{P} \right) + \hat{P}^2 = 0$$

is a quadratic equation that may be solved for  $P$ . A confidence bound is derived using the two solutions of the quadratic equation.

Next we derive the variance of the additive and multiplicative statistics for cohort studies. Recall

$$A^* = (d_A^{NS} - d_A^S) - (d_{NA}^{NS} - d_{NA}^S)$$

For the non-asbestos exposed cohort, the rates  $d_{NA}^{NS}$  and  $d_{NA}^S$  are determined from large samples and are considered non-stochastic. Therefore the variance is determined from the components  $d_A^{NS}$  and  $d_A^S$  which are stochastic but independent. In this case,

$$\begin{aligned} Var(A^*) &= Var(d_A^{NS}) + Var(d_A^S) \\ &= \left[ \frac{\hat{P}_A^{NS}(1 - \hat{P}_A^{NS})}{N_A^{NS}} \right] \left( \frac{N_A^{NS}}{Y_A^{NS}} \right)^2 \cdot (100,000)^2 + \\ &\quad \left[ \frac{\hat{P}_A^S(1 - \hat{P}_A^S)}{N_A^S} \right] \left( \frac{N_A^S}{Y_A^S} \right)^2 \cdot (100,000)^2 \end{aligned} \quad (34)$$

For the multiplicative statistic,

$$M^* = [(\log d_A^{NS}) - (\log d_A^S)] - [(\log d_{NA}^{NS}) - (\log d_{NA}^S)]$$

so that

$$Var(M^*) = \frac{(1 - \hat{P}_A^{NS})}{\hat{P}_A^{NS} N_A^{NS}} + \frac{(1 - \hat{P}_A^S)}{\hat{P}_A^S N_A^S}$$

These variances are used to calculate standard errors, confidence intervals, and Wald tests for the additive and multiplicative hypotheses.

For instance,  $M^*/\sqrt{Var(M^*)}$  is asymptotically standard normally distributed under the null hypothesis that  $M^* = 0$ .

## 6 Prevalance Studies

Some studies pertain to populations that all have a common exposure level of some contaminant. For instance in studies of asbestosis it is necessary that all subjects by definition have the same exposure to asbestos. In such cases the issues of additivity and multiplicativity are not germane as one can consider the separate effect of each each causal agent. A similar situation occurs in a cohort setting where a companion population is not used as a reference group. These situations are nevertheless illuminating in discerning the relative contribution of a second contaminant as it affects the probability of contracting or dying from a disease. Another example is the analysis of the prevalence of a disease attribute (such as pleural plaques) in an exposed population.

A prevalence model may be fitted with a logistic functional form. The outcome variable is usually the presence or absence of a disease characteristic where the explanatory factors will include control variables and an indicator for the level of contaminant. If the cohort provides some level of variance in the level of exposure of both contaminants, then an interactive term can be used as illustrated in the introduction to test for synergy even if this does not provide a test of additivity or multiplicativity.

## 7 Conclusion

Case-control, cohort, and prevalence studies provide varying types of information to determine relative risks and attendant confidence levels. We have considered several methods for testing additivity and multiplicativity hypotheses using Wald and likelihood ratio techniques. In these cases we have relied on asymptotic expectation for which the small sample populations are unknown. Our empirical results are reported in a companion paper and generally we find agreement in our conclusions regarding the additivity or multiplicativity hypothesis whether the analysis is conducted using Wald or likelihood ratio methods.

## References

- [1] Gardner, MJ, Munford, AG, (1980): The combined effect of two factors on disease in a case-control study. *Applied Statistics*, 29: No.3, 276-281.
- [2] Rao, CR, (1973). *Linear Statistical Inference and its Applications*, Wiley, New York.
- [3] Rothman, KJ, (1976): The estimation of synergy or antagonism. *American Journal of Epidemiology*, 103: No. 6, 506-511.
- [4] Woolf, B (1955): On estimating the relation between blood group and disease. *Annals of Human Genetics*, 19: 251-253.