2002

Estimation of Standardized Mortality Ratio in Epidemiological Studies

Bingxia Wang

ESTIMATION OF STANDARDIZED MORTALITY RATIO IN EPIDEMIOLOGICAL STUDIES

By

Bingxia Wang

B.S. Central China Normal University, 1995

A THESIS

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Master of Arts

(in Mathematics)

The Graduate School

The University of Maine

August, 2002

Advisory Committee:

Ramesh C. Gupta, Professor of Mathematics and Statistics, Advisor

Pushpa L. Gupta, Professor of Mathematics and Statistics

David M. Bradley, Assistant Professor of Mathematics and Statistics

LIBRARY RIGHTS STATEMENT

In presenting this thesis in partial fulfillment of the requirements for an advanced degree at The University of Maine, I agree that the Library shall make it freely available for inspection. I further agree that permission for "fair use" copying of this thesis for scholarly purposes may be granted by the Librarian. It is understood that any copying or publication of this thesis for financial gain shall not be allowed without my written permission.

Signature: Bingxua Wang

Date: 08/23/02

ESTIMATION OF STANDARDIZED MOTALITY RATIO IN EPIDEMIOLOGICAL STUDIES

By Bingxia Wang

Thesis Advisor: Dr. Ramesh C. Gupta

An Abstract of the Thesis Presented
in Partial Fulfillment of the Requirements for the
Degree of Master of Arts
(in Mathematics)
August, 2002

In epidemiological studies, we are often interested in comparing the mortality rate of a certain cohort to that of a standard population. A standard computational statistic in this regard is the Standardized Mortality Ratio (SMR) (Breslow and Day, 1987), given by

$$\text{SMR}=\frac{O}{E}\times 100$$

where O is the number of deaths observed in the study cohort from a specified cause, E is the expected number calculated from that population.

In occupational epidemiology, the SMR is the most common measure of risk. It is a comparative statistic. It is frequently based on a comparison of the number O in the cohort with the expected value E in a standard population. Our goal is to estimate the value of SMR. Since the expected value E is assumed to be fixed for a certain standard population, what we need to do is to estimate the observed number O, which is traditionally assumed to be Poisson distributed. We are primarily interested in confidence limits for the Poisson parameter.

Many authors have discussed methods for constructing confidence intervals for the SMR. These confidence intervals amount to obtaining more accurate confidence intervals for the Poisson parameter.

In this thesis, by using classic normal approximations, exact confidence intervals based on the chi-square distribution, binomial approximations and shortcut methods, we investigate more accurate methods for the statistical analysis of Poisson distributed data and carry out some simulation studies in order to obtain and compare better estimates of the SMR. These methods will be employed to develop an improved analysis of the SMR with missing death certificates.

ACKNOWLEDGEMENTS

I am grateful to express deep gratitude to Professor Ramesh C. Gupta for his invaluable advice and help at every stage of my graduate study.

I also thank my other committee members, Professor Pushpa L. Gupta and Professor David M. Bradley for their help and participation.

In particular, I thank all my committee members for their help and encouragement, not only in completing this thesis, but also in my pursuit of a Master's degree at the University of Maine.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

INTRODUCTION

In epidemiological studies, interest often lies in comparing the mortality rate of a certain cohort with that of a standard population. The size of the study cohort, for example, members of a certain profession, factory workers or patients, is likely to be relatively small compared to the size of the general population of a state, a province or a country in which the cohort arises. For this, there is a need for a summary measure, or a summary type of rate which can enable us to compare the two populations. Such a rate should be adjusted or standardized. For example, crude adjusted death rate (number of deaths in an area in a year/average population in the area in that year) presents a summary figure for a total population. Since the death rate varies according to the age, age is the variable for which adjustment is most often required because of its marked effect on mortality. So are sex and race.

There is a basic method for holding constant the age composition of a population. The method is used to compare the mortality rate of a certain cohort with that of a standard population. In this method the more stable rates of the larger population are applied to the smaller study group. Comparison of the expected deaths, thus obtained, with the number actually observed in the smaller population yields a standard computational statistical measure known as the Standardized Mortality Ratio, or SMR (Breslow and Day, 1987), given by

$$\text{SMR}=\frac{O}{E},$$

where O is the total number of deaths observed, in the study population, from a specified cause, and E is the total expected number of deaths calculated from that population. An SMR value greater than 1 indicates higher mortality in the study population than in the standard population, and conversely for an SMR value less than 1. In reality, however, the situation is not always so simple. 'As with other summary studies, the SMR depends on the age distribution as well as on mortality patterns in both populations.

In this thesis, our goal is to study several methods for estimating the value of the SMR. Since the observed number of events is traditionally assumed to have a Poisson distribution, and the expected value E is assumed to be fixed for a certain standard population, what we need to do is to estimate the observed number O. That is, we need to estimate the parameter of the Poisson distribution. This will be achieved by investigating different confidence intervals for the Poisson parameter.

Many authors have discussed methods for constructing confidence intervals based on different types of confidence limits for the Poisson parameter. Most of these methods depend on the classic normal approximations. Thus, normal approximations for the Poisson distribution have received some attention in the literature, but to a much lesser extent than the binomial approximation (Molenaar (1973)). In addition, the exact confidence interval for the Poisson parameter can be obtained by using a chi-square distribution based on the relationship $P\left(\chi^2_{2(y_0+1)} > 2n\lambda\right) = \sum_{k=0}^{y_0} \frac{e^{-n\lambda}(n\lambda)^k}{k!}$. However, for comparatively large degrees of freedom, the required critical values for the chi-square distribution may not be readily available, but the excellent approximate values are available. Recently, Schwertman et al. (1993) examined the accuracy of various binomial approximations for the confidence limits for the Poisson parameter. These simple approximations enable statisticians to make a quick evaluation with minimum table values. For use in epidemiological studies, Vandenbrouck (1982), Ury and Wiggins (1985) proposed some shortcut methods for estimating the SMR by using the variance stabilizing square root transformation of a Poisson variable. Ury and Wiggins (1985) claim that their method is quite simple and tends to be more accurate. The confidence intervals for the Poisson parameter enable us to calculate the 95 % confidence interval of the SMR derived by the division of the upper and lower limits of the observed number by the expected number.

As has been said earlier, the evaluation of epidemiological follow-up studies is frequently based on the ratio, SMR. The usual way to follow up persons is to identify the vital status in population registers, which provide precise information on the date and the place of death for the deceased persons with a high degree of completeness. And then, the responsible health offices are asked for the death certificates to obtain the official causes of death. Generally, this works with a high degree of completeness. However, study participants may have died many years or even decades back, and it is an open matter whether the health offices still have death certificates in their files. It is a long-term storage problem. Legally, the health offices are obliged to keep death certificates for 5 or 10 years. In practice, the certificates are usually stored for much longer. But inevitably, the greater the time elapsed, the lower the degree of completeness of the cause of death. This information is needed for historical follow-up studies.

Rittgen and Becker (2000) used the data of a historical follow-up study among foundry workers. In this study, the employees of 37 foundries in Germany were traced back to the 1950s (about 17,700 persons). However, the death certificates could be obtained for only about 70% of all deaths. They used this incomplete data of missing death certificates to create the statistical model, and obtained some confidence intervals for the SMR.

In chapter 2 of this thesis, we review various existing methods of estimating the Poisson parameter, such as classic approximation methods, exact confidence limits by using a chi-square distribution, and binomial approximations. We perform simulation studies to compare some of these confidence intervals in terms of their average length and coverage probability. In chapter 3, we investigate some shortcut methods which are often used in epidemiological studies, and we propose three other new methods for the statistical analysis of Poisson distributed data. Simulation studies are carried out to compare the existing methods and the proposed methods in terms of their average length and coverage probability. It turns out that one of the newly proposed methods outperforms the others. The missing death certificate problem is investigated in section 4. The procedure given by Rettgen & Becker (2000) is modified to accommodate different rates, of the availability of the death certificates, in the disease of interest and otherwise (eg. cancer and noncancer death). The data given in Rettgen & Becker (2000) is reanalyzed using our modification, and the effect of introducing different rates is examined. Finally, some conclusions and remarks are presented in chapter 5.

Chapter 2

REVIEW THE PROPOSED CONFIDENCE INTERVAL FOR POISSON PARAMETER

2.1 Poisson Distribution

Because the observed number of events is assumed to have Poisson distribution, inference procedures for the SMR can be formulated based on those for the Poisson distribution. Now, let's recall the Poisson probability model.

2.1.1 Definition

A random variable X is said to have a Poisson distribution if for some $\lambda > 0$, the probability mass function is $p(x;\lambda) = \dfrac{e^{-\lambda}\lambda^{x}}{x!}$, $\qquad x = 0,1,2,......$

Here, λ is the parameter of the Poisson distribution. The value of λ is frequently a rate per unit time or per unit area.

2.1.2 An Important Property

If X has a Poisson distribution with parameter λ, then

$$E(X) = Var(X) = \lambda$$

Thus, the Poisson distribution has property that the mean and the variance are equal to a common value λ. This important property of the Poisson distribution is, in fact, a characteristic property in a very broad class of discrete distributions; see Gupta (1977). This property is used to obtain classic confidence intervals for λ.

2.1.3 Classic Poisson Confidence Interval

Method 1. Let $X_1, X_2, \ldots, X_n$ be independent, identically distributed (i.i.d) Poisson random variables, and define $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. We have $E(\overline{X}) = \lambda$, and $Var(\overline{X}) = \frac{\lambda}{n}$. Then $\overline{X} \sim N(\lambda, \frac{\lambda}{n})$ for large n, and the statistic $Z = \frac{\overline{X} - \lambda}{\sqrt{\lambda/n}} \sim N(0,1)$. Since we don't know the value of λ, we replace λ, in the denominator, by its unbiased estimator. Therefore, by the above normal approximation the confidence interval for λ is

$$(\overline{X} - z_{\frac{\alpha}{2}}\sqrt{\frac{\overline{X}}{n}}, \overline{X} + z_{\frac{\alpha}{2}}\sqrt{\frac{\overline{X}}{n}}),$$

where $\Phi(z_{\frac{\alpha}{2}}) = P(z < z_{\frac{\alpha}{2}}) = 1 - \frac{\alpha}{2}$.

Example 1: Suppose $\alpha = 0.05$, $\overline{x} = 50$, $n = 25$. Since $z_{\frac{\alpha}{2}} = 1.96$, the 95% confidence interval for λ is (47.23, 52.77).

Method 2. Instead of replacing λ by its unbiased estimator, as in Method 1, we proceed as follows.

We have

$$P\left(\left|\frac{\overline{X} - \lambda}{\sqrt{\lambda/n}}\right| \le z_{\frac{\alpha}{2}}\right) = 1 - \alpha.$$

Rewriting the previous inequality as $P\left(\left|\frac{\overline{X}-\lambda}{\sqrt{\lambda/n}}\right|^2 \le z_{\frac{\alpha}{2}}^2\right)=1-\alpha$, we form the quadratic inequality $P\left(n\lambda^2-(2n\overline{X}+z_{\frac{\alpha}{2}}^2)\lambda+n\overline{X}^2 \le 0\right)=1-\alpha$ for λ. Solving this quadratic inequality, yields

$$P\left(\frac{2n\overline{X}+z_{\frac{\alpha}{2}}^2-z_{\frac{\alpha}{2}}\sqrt{z_{\frac{\alpha}{2}}^2+4n\overline{X}}}{2n}\le\lambda\le\frac{2n\overline{X}+z_{\frac{\alpha}{2}}^2+z_{\frac{\alpha}{2}}\sqrt{z_{\frac{\alpha}{2}}^2+4n\overline{X}}}{2n}\right)=1-\alpha.$$

Hence, the confidence interval for λ is

$$\left(\frac{2n\overline{X}+z_{\frac{\alpha}{2}}^2-z_{\frac{\alpha}{2}}\sqrt{z_{\frac{\alpha}{2}}^2+4n\overline{X}}}{2n},\frac{2n\overline{X}+z_{\frac{\alpha}{2}}^2+z_{\frac{\alpha}{2}}\sqrt{z_{\frac{\alpha}{2}}^2+4n\overline{X}}}{2n}\right).$$

Example 2: If we choose the same values of α, $\overline{x}$, n, as in Example 1, Method 2 gives the confidence interval as (47.3, 52.84).

Note that the confidence intervals obtained by Method 1 and Method 2 are quite close. Actually, these two classic confidence intervals for the Poisson parameter are derived by using normal approximations. Such approximations have received substantial attention in the statistical literature. In the next section, we will present exact confidence intervals for the Poisson parameter by using the Chi-square distribution.

2.2 Exact Confidence Interval Based on Chi-square Distribution

Exact confidence interval limits for the Poisson can be computed using the Chi-square distribution based on the following relation between the Poisson distribution and the Chi-square distribution:

$$P\left(\chi^2_{2(y_0+1)} > 2n\lambda\right) = \sum_{k=0}^{y_0} \frac{e^{-n\lambda}(n\lambda)^k}{k!}. \qquad (1)$$

Proof: Assuming a random variable $Y \sim Poisson(\lambda)$, the probability mass function is

$$P(y) = \frac{e^{-\lambda}\lambda^y}{y!}, \qquad y = 0,1,2,......$$

And another random variable is $X \sim \chi^2_{2(y_0+1)}$ with $2(y_0+1)$ degrees of freedom. The probability density function (pdf) is

$$f(x; 2(y_0+1)) = \frac{1}{2^{y_0+1}\Gamma(y_0+1)} x^{y_0} e^{-\frac{x}{2}}, \qquad x \geq 0.$$

Then,

$$P(\chi^2_{2(y_0+1)} > 2n\lambda)$$

$$= \int_{2n\lambda}^{\infty} f(x; 2(y_0+1))dx$$

$$= \int_{2n\lambda}^{\infty} \frac{1}{2^{y_0+1}\Gamma(y_0+1)} x^{y_0} e^{-\frac{x}{2}} dx$$

$$= \frac{1}{2^{y_0+1}\Gamma(y_0+1)} \int_{2n\lambda}^{\infty} x^{y_0} e^{-\frac{x}{2}} dx$$

$$= \frac{1}{2^{y_0+1} y_0!}\left(2^{y_0+1}(n\lambda)^{y_0} e^{-n\lambda} + 2y_0 \int_{2n\lambda}^{\infty} x^{y_0-1} e^{-\frac{x}{2}} dx\right)$$

$$= \frac{(n\lambda)^{y_0} e^{-n\lambda}}{y_0!} + \frac{1}{2^{y_0}(y_0-1)!}\int_{2n\lambda}^{\infty} x^{y_0-1} e^{-\frac{x}{2}} dx$$

$$= \frac{(n\lambda)^{y_0} e^{-n\lambda}}{y_0!} + \frac{(n\lambda)^{y_0-1} e^{-n\lambda}}{(y_0-1)!} + \frac{1}{2^{y_0-1}(y_0-1)!}\int_{2n\lambda}^{\infty} x^{y_0-2} e^{-\frac{x}{2}} dx$$

$$= \ ....$$

$$= \frac{(n\lambda)^{y_0} e^{-n\lambda}}{y_0!} + \frac{(n\lambda)^{y_0-1} e^{-n\lambda}}{(y_0-1)!} + \frac{(n\lambda)^{y_0-2} e^{-n\lambda}}{(y_0-2)!} + ... + \frac{1}{2^2}\int_{2n\lambda}^{\infty} x e^{-\frac{x}{2}} dx$$

$$= \frac{(n\lambda)^{y_0} e^{-n\lambda}}{y_0!} + \frac{(n\lambda)^{y_0-1} e^{-n\lambda}}{(y_0-1)!} + \frac{(n\lambda)^{y_0-2} e^{-n\lambda}}{(y_0-2)!} + ... + \frac{(n\lambda)^1 e^{-n\lambda}}{1!} + \frac{(n\lambda)^0 e^{-n\lambda}}{0!}$$

$$= \sum_{y=0}^{y_0} \frac{(n\lambda)^y e^{-n\lambda}}{y!}$$

$$= P(Y \le y_0).$$

This proves the interesting relationship (1) between the Poisson and Chi-square distributions.

We now present the following theorem which enables us to construct an exact confidence interval for the parameter of a discrete random variable.

2.2.1Exact Confidence Interval

Theorem: Let T be a discrete statistic with cdf $F_T(t|\theta) = P(T \le t|\theta)$. Let $0 < \alpha < 1$ be a fixed value. Suppose that for each $t \in T$, if $F_t(t|\theta)$ is a decreasing function of θ, define $\Theta_L(t)$ and $\Theta_U(t)$ by

$$P(T \le t|\Theta_U(t)) = \frac{\alpha}{2} \quad , \quad \text{and} \quad P(T \ge t|\Theta_L(t)) = \frac{\alpha}{2}.$$

Then the random interval $[\Theta_L(T), \Theta_U(T)]$ is a $1-\alpha$ confidence interval for θ. (Casella and Berger (1990))

Applying the above theorem, we can obtain the exact confidence interval for the Poisson parameter as follows.

2.2.2 Exact Confidence Interval for The Poisson Parameter

Let $X_1, X_2, \ldots X_n$ be a random sample from a Poisson population with parameter λ, and define $Y = \sum_i X_i$. Y is a sufficient statistic for λ and $Y \sim Poisson(n\lambda)$. By the above theorem, if $Y = y_0$ is observed, we are led to solve the following equations for λ:

$$\sum_{k=0}^{y_0} e^{-n\lambda} \frac{(n\lambda)^k}{k!} = \frac{\alpha}{2} \quad \text{and} \quad \sum_{k=y_0}^{\infty} e^{-n\lambda} \frac{(n\lambda)^k}{k!} = \frac{\alpha}{2}. \qquad (2)$$

Combining (1) and (2), we have

$$\frac{\alpha}{2} = \sum_{y=0}^{y_0} e^{-n\lambda} \frac{(n\lambda)^y}{y!} = P(Y \le y_0 | \lambda) = P(\chi^2_{2(y_0+1)} > 2n\lambda) \qquad (3)$$

and

$$\frac{\alpha}{2} = \sum_{y=y_0}^{\infty} e^{-n\lambda} \frac{(n\lambda)^y}{y!} = P(Y \ge y_0 | \lambda) = 1 - P(Y < y_0 | \lambda) = 1 - P(Y \le y_0 - 1 | \lambda) = 1 - P(\chi^2_{2y_0} > 2n\lambda) \qquad (4)$$

where $\chi^2_{2(y_0+1)}$, $\chi^2_{2y_0}$ are Chi-square random variables with $2(y_0+1)$, $2y_0$ degrees of freedom, respectively.

We now solve equations (3) and (4).

The upper bound λ_U of the confidence interval is obtained as follows.

From equation (3), we have

$$P(\chi^2_{2(y_0+1)} > 2n\lambda_U) = \frac{\alpha}{2}$$

$$\Longrightarrow \quad 2n\lambda_U = \chi^2_{2(y_0+1),\ \frac{\alpha}{2}}$$

$$\Longrightarrow \quad \lambda_U = \frac{1}{2n}\chi^2_{2(y_0+1),\ \frac{\alpha}{2}}\ .$$

On the other hand, the lower bound of the confidence interval is obtained by solving equation (4).

We have $$1 - P(\chi^2_{2y_0} > 2n\lambda_L) = \frac{\alpha}{2}$$

$$\Longrightarrow \quad P(\chi^2_{2y_0} > 2n\lambda_L) = 1 - \frac{\alpha}{2}$$

$$\Longrightarrow \quad 2n\lambda_L = \chi^2_{2y_0,\ 1-\frac{\alpha}{2}}$$

$$\Longrightarrow \quad \lambda_L = \frac{1}{2n}\chi^2_{2y_0,\ 1-\frac{\alpha}{2}}\,.$$

Therefore, the $1-\alpha$ confidence interval for λ is

$$\left(\frac{1}{2n}\chi^2_{2y_0,1-\frac{\alpha}{2}}, \frac{1}{2n}\chi^2_{2(y_0+1),\frac{\alpha}{2}}\right).$$

At $y_0 = 0$, we define $\chi^2_{0,1-\frac{\alpha}{2}} = 0$.

Now, we are taking a numerical example.

<u>Example:</u> Let $n = 10$ and $y_0 = \sum_i x_i = 6$. A 95% confidence interval for λ is given by $\left(\frac{1}{20}\chi^2_{12,0.975}, \frac{1}{20}\chi^2_{14,0.025}\right) = (0.22,\ 1.306)$.

However, if y_0 is large, say $y_0 > 50$, then the required critical value for the Chi-square distribution may not be readily obtained. Nevertheless excellent approximate values are available. As an alternative, however, it may be convenient to use a simple highly accurate binomial approximation that is not based on the Chi-square critical value. In the next section, we present some confidence limits of the Poisson parameter which are based on the approximation of the binomial distribution by the Poisson distribution.

2.3 Binomial Approximate Confidence Limits

The Binomial approximate confidence limits of the Poisson parameter are based on the following basic principle.

2.3.1 Principle

The binomial can be approximated by the Poisson. In other words, suppose $X \sim Binomial(n,p)$. Let $n \to \infty$ and $p \to 0$ in such a way that $np = \lambda > 0$ remains fixed. Then $Binomial(n,p) \to Poisson(\lambda)$.

We now present six Binomial approximated confidence intervals for the Poisson parameter.

2.3.2 Binomial Approximate Confidence Limits

Let $p_{(1-\frac{\alpha}{2})}, p^{(1-\frac{\alpha}{2})}$ be the respective lower and upper confidence limits for p. Then the corresponding confidence limits for the Poisson parameter λ are (Blyth 1986):

Lower: $$\lambda_{(1-\frac{\alpha}{2})} = \lim_{n\to\infty} np_{(1-\frac{\alpha}{2})}$$

Upper: $$\lambda^{(1-\frac{\alpha}{2})} = \lim_{n\to\infty} np^{(1-\frac{\alpha}{2})}.$$

We now present six methods of constructing confidence limits for a Poisson parameter based on binomial approximate confidence limits.

Method 1 For a binomial random variable X with probability of success p, we have $E(X) = np$, $Var(X) = np(1-p)$. Thus for large n, the statistic $Z = \frac{X-np}{\sqrt{np(1-p)}}$ is a approximately normally distributed $N(0,1)$. Assuming p is unknown, we replace p in the denominator of Z by its unbiased estimator $\hat{p} = \frac{X}{n}$ and obtain

$$Z = \frac{n\hat{p}-np}{\sqrt{n\hat{p}(1-\hat{p})}} = \frac{\hat{p}-p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \sim N(0,1).$$

The confidence limits for p are then given by

$$p_{(1-\frac{\alpha}{2})} = \hat{p} - z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \frac{X}{n} - \frac{z_{\frac{\alpha}{2}}}{\sqrt{n}}\sqrt{\frac{X}{n}(1-\frac{X}{n})},$$

$$\text{and} \quad p^{(1-\frac{\alpha}{2})} = \hat{p} + z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \frac{X}{n} + \frac{z_{\frac{\alpha}{2}}}{\sqrt{n}}\sqrt{\frac{X}{n}(1-\frac{X}{n})}.$$

The corresponding confidence limits for λ are

$$\lambda_{(1-\frac{\alpha}{2})} = \lim_{n\to\infty} n\left(\frac{X}{n} - \frac{z_{\frac{\alpha}{2}}}{\sqrt{n}}\sqrt{\frac{X}{n}(1-\frac{X}{n})}\right) = X - z_{\frac{\alpha}{2}}\sqrt{X}$$

and

$$\lambda^{(1-\frac{\alpha}{2})} = \lim_{n\to\infty} n\left(\frac{X}{n} + \frac{z_{\frac{\alpha}{2}}}{\sqrt{n}}\sqrt{\frac{X}{n}(1-\frac{X}{n})}\right) = X + z_{\frac{\alpha}{2}}\sqrt{X}.$$

Method 2 The second binomial approximated confidence limits are the same as those in Method 1 except that they include the continuity correction factor.

The confidence limits for p are

$$p_{(1-\frac{\alpha}{2})} = \frac{X-0.5}{n} - \frac{z_{\frac{\alpha}{2}}}{\sqrt{n}}\sqrt{\frac{X-0.5}{n}(1-\frac{X-0.5}{n})}$$

and

$$p^{(1-\frac{\alpha}{2})} = \frac{X+0.5}{n} + \frac{z_{\frac{\alpha}{2}}}{\sqrt{n}}\sqrt{\frac{X+0.5}{n}(1-\frac{X+0.5}{n})}.$$

The corresponding confidence limits for λ are therefore given by

$$\lambda_{(1-\frac{\alpha}{2})} = \lim_{n\to\infty} np_{(1-\frac{\alpha}{2})} = X - 0.5 - z_{\frac{\alpha}{2}}\sqrt{X-0.5}$$

$$\lambda^{(1-\frac{\alpha}{2})} = \lim_{n\to\infty} np^{(1-\frac{\alpha}{2})} = X + 0.5 + z_{\frac{\alpha}{2}}\sqrt{X+0.5}.$$

Method 3 The third binomial approximated confidence limits are based on the lower and upper limits obtained by solving a quadratic equation in p.

Specifically, suppose $X \sim Binomial(n,p)$. For large n, $X \sim N(np, np(1-p))$, and then $Z = \frac{X-np}{\sqrt{np(1-p)}} \sim N(0,1)$.

Since $P(|Z| \le z_{\frac{\alpha}{2}}) = 1-\alpha$, i.e. $P(Z^2 \le z^2_{\frac{\alpha}{2}}) = 1-\alpha$, we have the following process:

$$P\left((X-np)^2 \le z^2_{\frac{\alpha}{2}}(np(1-p))\right) = 1-\alpha,$$

i.e. $P\left((n^2 + z_{\frac{\alpha}{2}}^2 n)p^2 - (2nX + z_{\frac{\alpha}{2}}^2 n)p + X^2 \le 0\right) = 1-\alpha$

Therefore the solutions for p are

$$p_{(1-\frac{\alpha}{2})} = \frac{X + \frac{z_{\frac{\alpha}{2}}^2}{2} - z_{\frac{\alpha}{2}}\sqrt{X - \frac{X^2}{n} + \frac{z_{\frac{\alpha}{2}}^2}{4}}}{n + z_{\frac{\alpha}{2}}^2} \quad \text{and} \quad p^{(1-\frac{\alpha}{2})} = \frac{X + \frac{z_{\frac{\alpha}{2}}^2}{2} + z_{\frac{\alpha}{2}}\sqrt{X - \frac{X^2}{n} + \frac{z_{\frac{\alpha}{2}}^2}{4}}}{n + z_{\frac{\alpha}{2}}^2},$$

whence $$P(p_{(1-\frac{\alpha}{2})} \le p \le p^{(1-\frac{\alpha}{2})}) = 1-\alpha .$$

The confidence limits for λ are

$$\lambda_{(1-\frac{\alpha}{2})} = \lim_{n\to\infty} np_{(1-\frac{\alpha}{2})} = X + \frac{z_{\frac{\alpha}{2}}^2}{2} - z_{\frac{\alpha}{2}}\sqrt{X + \frac{z_{\frac{\alpha}{2}}^2}{4}}$$

and $$\lambda^{(1-\frac{\alpha}{2})} = \lim_{n\to\infty} np^{(1-\frac{\alpha}{2})} = X + \frac{z_{\frac{\alpha}{2}}^2}{2} + z_{\frac{\alpha}{2}}\sqrt{X + \frac{z_{\frac{\alpha}{2}}^2}{4}} .$$

Method 4 This method is as the same as the above method but includes the correction factor. That are

$$p_{(1-\frac{\alpha}{2})} = \frac{X - 0.5 + \frac{z_{\frac{\alpha}{2}}^2}{2} - z_{\frac{\alpha}{2}}\sqrt{X - 0.5 - \frac{(X-0.5)^2}{n} + \frac{z_{\frac{\alpha}{2}}^2}{4}}}{n + z_{\frac{\alpha}{2}}^2},$$

and $$p^{(1-\frac{\alpha}{2})} = \frac{X + 0.5 + \frac{z_{\frac{\alpha}{2}}^2}{2} + z_{\frac{\alpha}{2}}\sqrt{X + 0.5 - \frac{(X+0.5)^2}{n} + \frac{z_{\frac{\alpha}{2}}^2}{4}}}{n + z_{\frac{\alpha}{2}}^2}.$$

Then the confidence limits for λ are given by

$$\lambda_{(1-\frac{\alpha}{2})} = \lim_{n\to\infty} np_{(1-\frac{\alpha}{2})} = X - 0.5 + \frac{z^2_{\frac{\alpha}{2}}}{2} - z_{\frac{\alpha}{2}}\sqrt{X - 0.5 + \frac{z^2_{\frac{\alpha}{2}}}{4}},$$

and
$$\lambda^{(1-\frac{\alpha}{2})} = \lim_{n\to\infty} np^{(1-\frac{\alpha}{2})} = X + 0.5 + \frac{z^2_{\frac{\alpha}{2}}}{2} + z_{\frac{\alpha}{2}}\sqrt{X + 0.5 + \frac{z^2_{\frac{\alpha}{2}}}{4}}.$$

<u>Method 5</u> The fifth approximated confidence limits are based on the Molenaartype approximation for the binomial. The lower bound and upper bound are obtained from Blyth (1986) equation C. They are

$$p_{(1-\frac{\alpha}{2})} = \frac{(X-1)(1+\frac{1-c^2}{3n}) + \frac{2+c^2}{3} + \frac{1-c^2}{6n} - c\sqrt{\frac{X(n-X+1)}{n}(1+\frac{7-c^2}{18n}) - (n+1)\frac{7-c^2}{18n}}}{n+\frac{2+c^2}{3}}$$

$$p^{(1-\frac{\alpha}{2})} = \frac{X(1+\frac{1-c^2}{3n}) + \frac{2+c^2}{3} + \frac{1-c^2}{6n} + c\sqrt{\frac{(X+1)(n-X)}{n}(1+\frac{7-c^2}{18n}) - (n+1)\frac{7-c^2}{18n}}}{n+\frac{2+c^2}{3}}$$

where $c = z_{\frac{\alpha}{2}}$.

Then the corresponding confidence limits for λ are

$$\lambda_{(1-\frac{\alpha}{2})} = \lim_{n\to\infty} np_{(1-\frac{\alpha}{2})} = X - 1 + \frac{2+z^2_{\frac{\alpha}{2}}}{3} - z_{\frac{\alpha}{2}}\sqrt{X - \frac{7-z^2_{\frac{\alpha}{2}}}{18}}$$

and
$$\lambda^{(1-\frac{\alpha}{2})} = \lim_{n\to\infty} np^{(1-\frac{\alpha}{2})} = X + 1 + \frac{2+z^2_{\frac{\alpha}{2}}}{3} + z_{\frac{\alpha}{2}}\sqrt{X + 1 - \frac{7-z^2_{\frac{\alpha}{2}}}{18}}.$$

Method 6 The final binomial approximated confidence limits that we wish to present are based on the Pauson-Camp-Pratt approximate confidence limits for the binomial, see equation D in Blyth (1986). These are

$$p_{(1-\frac{\alpha}{2})} = \left\{ 1 + \left(\frac{X}{n-X+1} \right)^2 \left[\frac{81X(n-X+1) - 9n - 8 + 3z_{\frac{\alpha}{2}} \sqrt{9X(n-X+1)(9n+5-z_{\frac{\alpha}{2}}^2) + n + 1}}{81X^2 - 9X(2 + z_{\frac{\alpha}{2}}^2) + 1} \right]^3 \right\}^{-1}$$

$$p^{(1-\frac{\alpha}{2})} = \left\{ 1 + \left(\frac{X+1}{n-X} \right)^2 \left[\frac{81(X+1)(n-X) - 9n - 8 - 3z_{\frac{\alpha}{2}} \sqrt{9(X+1)(n-X)(9n+5-z_{\frac{\alpha}{2}}^2) + n + 1}}{81(X+1)^2 - 9(X+1)(2 + z_{\frac{\alpha}{2}}^2) + 1} \right]^3 \right\}^{-1} .$$

The corresponding confidence limits for λ are:

$$\lambda_{(1-\frac{\alpha}{2})} = \lim_{n\to\infty} np_{(1-\frac{\alpha}{2})} = \frac{\left(9X - 1 - 3z_{\frac{\alpha}{2}} \sqrt{X} \right)^3}{729X^2}$$

and $$\lambda^{(1-\frac{\alpha}{2})} = \lim_{n\to\infty} np^{(1-\frac{\alpha}{2})} = \frac{\left(9(X+1) - 1 + 3z_{\frac{\alpha}{2}} \sqrt{X+1} \right)^3}{729(X+1)^2} .$$

Schwertman N.C. et.al (1994) gave us some examples to display the confidence limits for each approximation (1) through (6).

2.3.3 Example

Let $\alpha = 0.05$, $x = 25$. Using the above methods, we get the confidence intervals for λ:

Method 1: (15.2, 34.8)

Method 2: (14.8,35.4)

Method 3: (16.9, 36.9)

Method 4: (16.53, 37.5)

Method 5: (16.18, 36.907)

Method 6: (16.174, 36.906)

All of these results are very close. In order to study the performance of the above six methods, we conduct a simulation study in the next sub-section.

2.3.4 Simulation

The simulation study is carried out as follows:

(1) Generate 1000 samples.

(2) For each sample, we set a sample size of $n = 25$.

(3) We repeat the above process for several different values of the parameter λ.

Normally, average length and coverage probability are used as scales to measure the goodness of a confidence interval. The length of the interval is the difference between the lower and upper confidence limits, and coverage is the probability that the random interval covers the actual value. Naturally, we want small average length and large coverage probability. In our case, we want a smaller length and 95% coverage probability.

The results are presented in the following table.

TABLE 2.1 Simulation Results for Binomial Approximations

		METHODS				
	1	2	3	4	5	6
λ =2						
avg length	5.504	6.45	6.716	7.691	6.905	6.962
coverage	1		1	1	1	1
λ =3						
avg length	6.756	7.73	7.774	8.758	8.095	8.131
coverage	1		1	1	1	1
λ =4						
avg length	7.818	8.80	8.713	9.701	9.118	9.144
coverage	1		1	1	1	1
λ =5						
ave length	8.722	9.71	9.532	10.523	9.994	10.016
coverage	1		1	1	1	1
λ =6						
avg length	9.587	10.57	10.329	11.322	10.836	10.855
coverage	1		1	1	1	1
λ =7						
avg length	10.33	11.32	11.022	12.017	11.563	11.579
coverage	1		1	1	1	1
λ =8						
avg length	11.064	12.05	11.712	12.708	12.282	12.297
coverage	1		1	1	1	1

Since the coverage probability in all these cases does not conform to 0.95, the above confidence interval methods are not of much use in terms of coverage. Therefore, we need to find other ways to obtain the confidence interval for the Poisson parameter λ.

Chapter 3

CONFIDENCE INTERVALS IN EPIDEMIOLOGICAL LITERATURE

In this chapter, we study the shortcut methods which are often used in epidemiological studies. In addition to these methods, we propose some new methods for estimating the Poisson parameter.

3.1 The Square Root Transformation Theorem

Before proceeding further, we present the square root transformation of the Poisson random variable on which these methods are based. We present the following square root transformation theorem which stabilizes the variance of the Poisson random variable.

The Square Root Transformation Theorem

For the Poisson distribution, it can be shown for "reasonably large λ", say $\lambda \geq 30$, that if $X \sim Poisson(\lambda)$, then $Var(\sqrt{X}) \approx 0.25$.

Proof: If a function f has continuous derivatives up to $(n+1)^{th}$ order at a point a, then by Taylor's theorem, a can be expanded about a

$$f(x) = f(a) + f^{'}(a)(x-a) + \frac{f^{''}(a)(x-a)^2}{2!} + ... + \frac{f^{(n)}(a)(x-a)^n}{n!} + R_n \ ,$$

where R_n, remainder after $n+1$ terms, is given by

$$R_n = \int_a^x f^{(n+1)} \frac{(x-u)^n}{n!} du = \frac{f^{(n+1)}(\xi)(x-a)^{n+1}}{(n+1)!}, \qquad a < \xi < x,$$

where

$$\lim_{n\to\infty} R_n = \lim_{n\to\infty} \frac{f^{(n+1)}(\xi)(x-a)^{n+1}}{(n+1)!} = 0.$$

In general, we will not be concerned with the explicit form of the remainder. Since we are interested in approximations, we are just going to ignore the remainder. Therefore, the function $f(x)$ has the following Taylor's approximation:

$$f(x) \approx f(a) + f'(a)(x-a) + \frac{f''(a)(x-a)^2}{2!} + \ldots\ldots + \frac{f^{(n)}(a)(x-a)^n}{n!}.$$

For the statistical application of Taylor's Theorem, we are most concerned with the `first-order` Taylor series, that is, an approximation using just the first derivative:

$$f(x) = f(a) + \sum_{i=1}^{k} f_i'(a)(x_i - a_i) + \mathrm{Re}\,mainder.$$

In our case, we have just one parameter λ, $f'(\lambda) = \frac{\partial}{\partial x} f(x)\big|_{x=\lambda}$, then

$$f(x) = f(\lambda) + f'(\lambda)(x-\lambda) + \mathrm{Re}\,mainder$$

We can re-write this by using approximation:

$$f(x) \approx f(\lambda) + f'(\lambda)(x-\lambda).$$

As we know, if X is a Poisson random variable with the parameter λ, we have

$$E(X) = \lambda \qquad \text{and} \qquad Var(X) = \lambda.$$

This gives

$E(f(X)) \approx f(\lambda) + f'(\lambda)(E(X) - \lambda) = f(\lambda)$, and

$$\begin{aligned} Var(f(X)) &= E(f(X) - E[f(X)])^2 \\ &\approx E(f(X) - f(\lambda))^2 \end{aligned}$$

$$\approx E\left(f(\lambda)+f'(\lambda)(X-\lambda)-f(\lambda)\right)^2$$

$$=\left(f'(\lambda)\right)^2 E(X-\lambda)^2$$

$$=\left(f'(\lambda)\right)^2 Var(X).$$

If we set $f(X)=\sqrt{X}$, we havė

$$E\left(\sqrt{X}\right)\approx\sqrt{\lambda}$$

and

$$Var\left(\sqrt{X}\right)\approx\frac{1}{4\lambda}\lambda=\frac{1}{4}=0.25.$$

Therefore, for reasonably large λ, $\sqrt{X}$ is approximately normally distributed. That is $\sqrt{X}\sim N\left(\sqrt{\lambda},0.25\right)$.

3.2 Shortcut Methods in Epidemiological Studies

In epidemiological studies, two shortcut methods have been proposed to construct the confidence intervals for the Poisson parameter.

Shortcut Method 1

This method was given by Vandenbroucke J.P. in 1982.

By using the Square Root Transformation Theorem, we know $\overline{\sqrt{X}}\sim N\left(\sqrt{\lambda},\frac{0.25}{n}\right)$, and the statistic $Z=\frac{\overline{\sqrt{X}}-\sqrt{\lambda}}{0.5/\sqrt{n}}\sim N(0,1)$. Therefore the confidence interval for $\sqrt{\lambda}$ is given by $\left(\overline{\sqrt{X}}-\frac{0.5}{\sqrt{n}}z_{\frac{\alpha}{2}},\overline{\sqrt{X}}+\frac{0.5}{\sqrt{n}}z_{\frac{\alpha}{2}}\right)$.

Example E1: Let $X=23$, and $\alpha=0.05$. Then the 95% confidence interval for λ is (14.56, 33.64).

Shortcut Method 2

This shortcut was given by Ury H.K. and Wiggins A.D. in 1985. Actually, it is a quick and simple normal approximation by adding 1 to the lower limit and 2 to the upper limit of the classic 95% confidence interval obtained earlier. See Method 1 of the section 2.1.3. Thus, the shortcut 95% confidence interval is given by ($\overline{X} - z_{\frac{\alpha}{2}}\sqrt{\frac{\overline{X}}{n}} + 1, \overline{X} + z_{\frac{\alpha}{2}}\sqrt{\frac{\overline{X}}{n}} + 2$). With the same data as in Example E1, the confidence interval for λ is (14.6, 34.4).

In the following section we obtain some new shortcut methods for estimating the Poisson parameter.

3.3 Some New Methods for Estimating The Poisson Parameter

Shortcut Method 3

We combine the Square Root Transformation and Vandenbroucke's method as follows:

Suppose the random variables $X_i^{'}$s have i.i.d. Poisson distribution with parameter λ, $i = 1, 2, \ldots\ldots n$ and $Y = \sum_{i=1}^{n} X_i$. Then Y also has Poisson distribution but with parameter $n\lambda$.

By the Square Root Transformation, $\sqrt{Y} = \sqrt{\sum_i X_i} \sim N\left(\sqrt{n\lambda}, 0.25\right)$, and the

95% confidence interval for $n\lambda$ is given by $\left((\sqrt{\sum_i X_i} - 0.5z_{\frac{\alpha}{2}})^2, (\sqrt{\sum_i X_i} + 0.5z_{\frac{\alpha}{2}})^2 \right)$.

Hence, the corresponding confidence interval for λ is

$$\left(\frac{(\sqrt{\sum_i X_i} - 0.5z_{\frac{\alpha}{2}})^2}{n}, \frac{(\sqrt{\sum_i X_i} + 0.5z_{\frac{\alpha}{2}})^2}{n} \right).$$

Here, using the data as in example E1, the confidence interval is (14.56, 33.64).

Shortcut Method 4

In this case, we add the correction term1 to the statistic involved in the upper limit of Method 3.3.a and obtain the following second confidence interval as

$$\left(\frac{(\sqrt{\sum_i X_i} - 0.5z_{\frac{\alpha}{2}})^2}{n}, \frac{(\sqrt{\sum_i X_i + 1} + 0.5z_{\frac{\alpha}{2}})^2}{n} \right).$$

With the same data as in Example E1, the confidence interval is (14.56, 34.57).

Shortcut Method 5

In this case, we modify the Ury-Wiggins shortcut method, presented in the previous section.

Since $Y \sim Poisson(n\lambda)$, the confidence interval for $n\lambda$ is given by $\left(Y - z_{\frac{\alpha}{2}}\sqrt{Y} + 1, Y + z_{\frac{\alpha}{2}}\sqrt{Y} + 2 \right)$. This gives a confidence interval for λ as

$$\left(\frac{Y - z_{\frac{\alpha}{2}}\sqrt{Y} + 1}{n}, \frac{Y + z_{\frac{\alpha}{2}}\sqrt{Y} + 2}{n} \right).$$

The above confidence interval can be written as

$$\left(\frac{\sum_i X_i - z_{\frac{\alpha}{2}} \sqrt{\sum_i X_i + 1}}{n}, \frac{\sum_i X_i + z_{\frac{\alpha}{2}} \sqrt{\sum_i X_i + 2}}{n} \right).$$

The confidence interval for the same data in Example E1 is (14.6, 34.4).

We find that the confidence intervals obtained in our examples are very close.

3.4 Comparison and Simulation Studies

We next compare the above five methods (methods in section 3.2 and 3.3) by carrying out some simulation studies. For this purpose we generate 1000 samples of size 25 for different values of the parameter and examine the lengths of the 95% confidence intervals and their coverage probabilities. Results are presented in the following table.

TABLE 3.1 Simulation Results for Shortcut Methods

	VJP	U-W	GB1	GB2	GB3
λ =50					
avg length	5.511	6.533	5.533	5.575	5.573
coverage	0.885	0.91	0.94	0.94	0.94
λ =51					
avg length	5.568	6.591	5.591	5.632	5.631
coverage	0.878	0.903	0.941	0.943	0.943
λ =52					
avg length	5.613	6.639	5.639	5.68	5.679
coverage	0.855	0.901	0.902	0.906	0.906
λ =53					
avg length	5.677	6.699	5.699	5.74	5.739
coverage	0.891	0.916	0.943	0.943	0.943
λ =54					
avg length	5.725	6.749	5.749	5.79	5.789
coverage	0.883	0.915	0.931	0.934	0.934
λ =55					
avg length	5.775	6.798	5.798	5.839	5.838
coverage	0.875	0.908	0.915	0.921	0.921
λ =56					
avg length	5.832	6.855	5.855	5.896	5.895
coverage	0.881	0.928	0.935	0.936	0.936

Where VJP and U-W present the shortcut method 1, 2, respectively. GBi (i=1,2,3) represent the newly proposed shortcut Methods.

From the Table 3.1, we notice that in terms of the coverage probability, GB1, GB2, GB3 are closer to the 95% nominal value than VJP and U-W in all cases. In terms of the average length, VJP outperforms the other procedures. Comparing the lengths of VJP and GB1, we notice that GB1 is slightly longer than VJP, but has appreciably closer coverage probability to the nominal value of 0.95 than the VJP.

Overall the new method GB1 gives the best result in terms of the average length and coverage probability.

Chapter 4

THE PROBLEM OF MISSING DEATH CERTIFICATES

4.1 Rettgen &Becker Model

4.1.1 Background

The comparative statistic can be used for the SMR evaluation of epidemiological follow-up studies. In epidemiological studies, the usual way to follow up persons is to identify the vital status in population registers, which are compulsory and provide precise information on date and place of death for deceased persons with a high degree of completeness. In a second step, the responsible health offices are asked for the death certificates to obtain the official causes of death. In practice, the certificates are usually stored for much longer. But inevitably, the greater the time elapsed, the lower the degree of completeness of cause-of-death information.

4.1.2 Problem of Missing Death Certificates

As an example, we use the data of a historical follow-up study among foundry workers. In this study, the employees of 37 foundries were traced back to the 1950's (approximately 17,700 persons). The vital status could also be traced sufficiently completely over the decades by means of the population registers (loss to follow-up of 6.2%). However the death certificates could only be obtained for about 70% of all deaths. Table 1 shows selected SMRs from a preliminary evaluation of these data (Adzersen et al. 1997).

TABLE 4.1

SMR calculated with empirically observed numbers of deaths *O* and confidence limits

Cause of death	O	E	SMR	CL
All causes	3972	3441.	115.4	111.9-119.1
All known causes	2896	3441.	84.2	81.1-87.3
Malignant neoplasms	831	881.	94.3	88.0-100.9
Lip, oral cavity, and pharynx	36	30.	117.5	82.3-162.7
Liver and intrahepatic bile ducts	28	12.	225	149.5-325.2
Larynx	20	14.	140.1	85.5-216.4
Trachea, bronchus, lung	322	253.	127.2	113.7-141.8
Respiratory system	199	185.	107.1	92.7-123.0

CL=95% confidence limits caculated with methods described in Breslow and Day (1987).

We can just think about one disease: Malignant neoplasms, and simply call it "cancer". So we get a 2×2 table which is easier to analyze.

TABLE 4.2

Cancer& Noncancer Data (1)

	Death Certificate Available	Death Certificate not Available	Total
Cancer	831	?	?
Noncancer	2065	?	?
Total	2896		3972

4.1.3 The Statistical Model

Now we are setting up the statistical model for the problem of missing death certificates. First, we like to introduce several parameters, which can be identified by the follow-up in the population registers and can be observed.

K: the Poisson-distributed random variable with parameter k_0, which presents the total number of deaths from the disease of interest in the cohort, which we call "cancer" in the following. K is unknown because some death certificates are not available.

L: the Poisson variable with parameter λ representing the number of all noncancer deaths, which is also unknown.

Z: $Z = \mathrm{K} + L$, the Poisson random variable with parameter $k_0 + \lambda$. It represents the total number of deceased persons in the cohort.

As we know, a particular cause of death can be identified by an obtainable death certificate can be considered by a series of i.i.d. Bernoulli random variables. Let $\{X_i\}$ be the i.i.d. Bernoulli random variables that represent the cancer deaths for which the death certificate is available, with the probability p, i.e., $P(X_i = 1) = p$, $P(X_i = 0) = 1 - p$. Similarly, let $\{Y_i\}$ be the i.i.d. Bernoulli random variables, independent of $\{X_i\}$ and having the same parameter p, which represents the noncancer deaths for which the death certificates are available. Now, we observe:

$M = \sum_{i=1}^{K} X_i$, where $X_i = 1$, then $M \sim Poisson(\mu = pk_0)$, and

$N = \sum_{j=1}^{L} Y_j$, where $Y_i = 1$, then $N \sim Poisson(\upsilon = p\lambda)$.

we may present the above notations in the following table, Table 4.3.

TABLE 4.3

Cancer & Noncancer Data (2)

	Death Certificate Available	Death Certificate not Available	Total
Cancer	$M(831) \sim$ $Poisson(\mu = pk_0)$		$\text{K} \sim Poisson(k_0)$
Noncancer	$N(2065) \sim$ $Poisson(\upsilon = p\lambda)$		$L \sim Poisson(\lambda)$
Total	$M + N(2896)$		$Z(3972) \sim$ $Poisson(k_0 + \lambda)$

The probability distribution of the observed numbers M and K is given by

$$P(M=m, \mathrm{K}=k) = P(M=m|\mathrm{K}=k)P(\mathrm{K}=k)$$

$$= \binom{k}{m} p^m (1-p)^{k-m} \frac{e^{-k_0} k_0^k}{k!}$$

$$= \frac{k!}{m!(k-m)!} p^m (1-p)^{k-m} \frac{e^{-k_0} k_0^k}{k!}. \qquad (4.1)$$

Similarly, the probability distribution of the observed numbers N and L is given by

$$P(N=n, L=l) = P(N=n|L=l)P(L=l)$$

$$= \binom{l}{n} p^n (1-p)^{l-n} \frac{e^{-\lambda} \lambda^l}{l!}$$

$$= \frac{l!}{n!(l-n)!} p^n (1-p)^{l-n} \frac{e^{-\lambda} \lambda^l}{l!} . \qquad (4.2)$$

Therefore, the probability distribution of the factually observed numbers M ,N and Z is

$$P(M=m, N=n, \mathrm{K}+L=z)$$

$$= \sum_{k=m}^{z-n} P(M=m, N=n, \mathrm{K}=k, L=z-k)$$

$$= \sum_{k=m}^{z-n} P(M=m, \mathrm{K}=k)P(N=n, L=z-k)$$

$$= \sum_{k=m}^{z-n} \frac{k!}{m!(k-m)!} p^m (1-p)^{k-m} \frac{e^{-k_0} k_0^k}{k!} \frac{l!}{n!(l-n)!} p^n (1-p)^{l-n} \frac{e^{-\lambda} \lambda^l}{l!}$$

$$= \frac{p^m k_0^m}{m!} \frac{p^n \lambda^n}{n!} (1-p)^{z-m-n} e^{-(k_0+\lambda)} \sum_{k=m}^{z-n} \frac{k_0^{k-m}}{(k-m)!} \frac{\lambda^{z-k-n}}{(z-k-n)!} \qquad (4.3)$$

$$= \frac{1}{(z-m-n)!}\sum_{k=m}^{z-n}\binom{z-m-n}{k-m}k_0^{k-m}\lambda^{z-k-n}$$

$$= \frac{1}{(z-m-n)!}\sum_{t=0}^{z-m-n}\binom{z-m-n}{t}k_0^{k-m}\lambda^{z-m-n-t} \qquad (t = k-m)$$

$$= \frac{1}{(z-m-n)!}(k_0+\lambda)^{z-m-n}. \qquad (4.4)$$

Now, we are defining likelihood function with unknown parameters $p,\ k_0$ and λ:

$$L(p,\ k_0,\ \lambda) = P(M = m, N = n, \mathrm{K} + L = z)$$

$$= \frac{p^m k_0^m}{m!}\frac{p^n \lambda^n}{n!}e^{-(k_0+\lambda)}\frac{((1-p)(k_0+\lambda))^{z-m-n}}{(z-m-n)!}\ . \qquad (4.5)$$

In the terminology used before, the number of empirically observed cases is just m, i.e., $O = m$, but the actually relevant number is K, the unknown true number of cancer cases in the cohort.

4.1.4 Maximum Likelihood Estimation

From the probability model, or likelihood function (4.5), we get the log likelihood function:

$$\ln L = m\ln p + m\ln k_0 + n\ln p + n\ln\lambda + (z-m-n)\ln(1-p) + (z-m-n)\ln(k_0+\lambda) - (k_0+\lambda)$$

$$-\ln(m!n!(z-m-n)!).$$

The likelihood equations are

$$\frac{\partial}{\partial p}\ln L = \frac{m}{p} + \frac{n}{p} - \frac{z-m-n}{1-p} = 0,$$
$$\frac{\partial}{\partial k_0}\ln L = \frac{m}{k_0} + \frac{z-m-n}{k_0+\lambda} - 1 = 0,$$
$$\frac{\partial}{\partial \lambda}\ln L = \frac{n}{\lambda} + \frac{z-m-n}{k_0+\lambda} - 1 = 0.$$

The maximum likelihood estimators (MLE) for the parameters are

$$\begin{aligned} \hat{p} &= \frac{m+n}{z}, \\ \hat{\lambda} &= \frac{nz}{m+n} = \frac{n}{\hat{p}}, \\ \hat{k}_0 &= \frac{mz}{m+n} = \frac{m}{\hat{p}}. \end{aligned} \tag{4.6}$$

The information matrix, J, is given by

$$J = \begin{pmatrix} -\frac{\partial^2}{\partial p^2}\ln L & -\frac{\partial^2}{\partial p \partial k_0}\ln L & -\frac{\partial^2}{\partial p \partial \lambda}\ln L \\ -\frac{\partial^2}{\partial k_0 \partial p}\ln L & -\frac{\partial^2}{\partial k_0^2}\ln L & -\frac{\partial^2}{\partial k_0 \partial \lambda}\ln L \\ -\frac{\partial^2}{\partial \lambda \partial p}\ln L & -\frac{\partial^2}{\partial \lambda \partial k_0}\ln L & -\frac{\partial^2}{\partial \lambda^2}\ln L \end{pmatrix}$$

$$= \begin{pmatrix} \frac{z^3}{(m+n)(z-m-n)} & 0 & 0 \\ 0 & \frac{zm+mn+n^2}{mz^2} & \frac{z-m-n}{z^2} \\ 0 & \frac{z-m-n}{z^2} & \frac{zn+mn+m^2}{nz^2} \end{pmatrix}.$$

As we know, J^{-1} is variance-covariance matrix. Next, we find J^{-1}.

The determinant of J is $Det(J) = \dfrac{(m+n)^2}{mn(z-m-n)}$.

The cofactor matrix of J is

$$J_0 = \begin{pmatrix} \dfrac{(m+n)^3}{mnz^3} & 0 & 0 \\ 0 & \dfrac{z(zn+mn+m^2)}{n(m+n)(z-m-n)} & -\dfrac{z}{m+n} \\ 0 & -\dfrac{z}{m+n} & \dfrac{z(zm+mn+n^2)}{m(m+n)(z-m-n)} \end{pmatrix}$$

and the transpose matrix of $J_0, J_0^{'} = J_0$, since J_0 is symmetric. Hence, the inverse matrix of J is

$$J^{-1} = \frac{J_0^{'}}{Det(J)} = \begin{pmatrix} \dfrac{(m+n)(z-m-n)}{z^3} & 0 & 0 \\ 0 & \dfrac{mz(zn+mn+m^2)}{(m+n)^3} & -\dfrac{zmn(z-m-n)}{(m+n)^3} \\ 0 & -\dfrac{zmn(z-m-n)}{(m+n)^3} & \dfrac{zn(zm+mn+n^2)}{(m+n)^3} \end{pmatrix}.$$

In particular, we have the variance of $\hat{k}_0$, which is $\left(J^{-1}\right)_{22}$.

$$\left(J^{-1}\right)_{22} = Var\left(\hat{k}_0\right) = \frac{mz}{m+n}\left(\frac{(m+n)^2}{(m+n)^2} + \frac{n}{m+n}\frac{z-m-n}{m+n}\right)$$

$$Var\left(\hat{k}_0\right) = \hat{k}_0\left(1 + \frac{n}{m+n}\frac{1-\hat{p}}{\hat{p}}\right). \qquad (4.7)$$

For large sample, $\hat{k}_0$ is approximately normally distributed. That is

$$\hat{k}_0 \sim N\left(k_0, \hat{k}_0(1 + \frac{n}{m+n}\frac{1-\hat{p}}{\hat{p}})\right).$$

Then the $(1-\alpha)100\%$ confidence interval for k_0 is

$$\left(\hat{k}_0 - z_{\frac{\alpha}{2}}\sqrt{\hat{k}_0\left(1+\frac{n}{m+n}\frac{1-\hat{p}}{\hat{p}}\right)},\ \hat{k}_0 + z_{\frac{\alpha}{2}}\sqrt{\hat{k}_0\left(1+\frac{n}{m+n}\frac{1-\hat{p}}{\hat{p}}\right)}\right). \qquad (4.8)$$

4.1.5 The Confidence Interval for SMR

$SMR^* = \frac{O^*}{E}\times 100$, where SMR^* is the calculation of a corrected SMR, O^* is the total number of cancer deaths. Then

$$\hat{O}^* = \hat{k}_0 = \frac{m}{\hat{p}} = \frac{O}{\hat{p}} \qquad \text{and} \qquad \hat{SMR}* = \frac{\hat{O}^*}{E} = \frac{O}{\hat{p}E}.$$

(1). Denoting by $\underline{k_0}$, $\overline{k_0}$, the lower and upper confidence limits for k_0 respectively, the first confidence interval (CL1) for SMR^* is $(\underline{SMR^*}, \overline{SMR^*}) = \left(\frac{\underline{k_0}}{E}, \frac{\overline{k_0}}{E}\right)$.

(2). The second confidence interval (CL2) for SMR^* is based on the binomial parameter p. We have already got the confidence interval for this parameter in section 2.3.2

$\left(\underline{p}, \overline{p}\right) = \left(\hat{p} - z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{z}},\ \hat{p} + z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{z}}\right)$, then the corresponding confidence limits for SMR^*(CL2) are

$$\text{Lower bound: } \underline{SMR^*} = \frac{\underline{O^*}}{E} = \frac{\underline{k_0}}{E} = \frac{\underline{O}}{\overline{p}E} = \frac{\underline{p}}{\overline{p}}\frac{\underline{O^*}}{E}$$

$$\text{Upper bound: } \overline{SMR^*} = \frac{\overline{O^*}}{E} = \frac{\overline{k_0}}{E} = \frac{\overline{O}}{\underline{p}E} = \frac{\overline{p}}{\underline{p}}\frac{\overline{O^*}}{E}.$$

where $\left(\underline{O^*}, \overline{O^*}\right) = \left(\underline{k_0}, \overline{k_0}\right)$.

For the data presented in Table 4.2, the results are

TABLE 4.4

SMR^* calculated with estimated numbers of deaths O^* and confidence limits

Cause of death	O*	SMR*	CL1	CL2
Malignant neoplasms	1139.8	129.3	120.9-137.8	117.3-142.5
Lip, oral cavity, and pharynx	49.4	161.1	108.6-213.7	105.0-237.2
Liver and intrahepatic bile ducts	38.4	308.6	194.4-422.8	189.3-476.4
Larynx	27.4	192.1	108.0-276.3	106.8-319.2
Trachea, bronchus, lung	441.6	174.4	155.7-193.2	150.6-201.4
Respiratory system	272.9	146.8	126.6-167.0	122.2-175.3

CL1=95% approximate confidence limits calculated with formula 1

CL2=95% approximate confidence limits calcutaled with formula 2

From the above table, it is very clear that CL1 is performing better than CL2 in terms of lengths of the confidence intervals.

4.2 Modification of Rittgen & Becker Model

In this section, we modify the model presented earlier. In the original model, it was assumed that the probability of the availability of the death certificates, in both cancer and noncancer deaths, is the same, p. However, we feel that this assumption is too restrictive. Hence we assume that the probability of availability of death certificate in the noncancer deaths is ap, where a can be less (more) than 1. The modified model is exhibited in the following table.

TABLE 4.5

Cancer & Noncancer Data (3)

	Death Certificate Available	Death Certificate not Available	Total
Cancer	$M(831) \sim$ $Poisson(\mu = pk_0)$		$\mathrm{K} \sim Poisson(k_0)$
Noncancer	$N(2065) \sim$ $Poisson(\upsilon = ap\lambda)$		$L \sim Poisson(\lambda)$
Total	$M + N(2896)$		$Z(3972) \sim$ $Poisson(k_0 + \lambda)$

For cancer case,

$$P(M = m, \mathrm{K} = k) = P(M = m|\mathrm{K} = k)P(\mathrm{K} = k)$$

$$= \binom{k}{m} p^m (1-p)^{k-m} \frac{e^{-k_0} k_0^k}{k!}$$

$$= \frac{k!}{m!(k-m)!} p^m (1-p)^{k-m} \frac{e^{-k_0} k_0^k}{k!}. \qquad (4.9)$$

For noncancer case,

$$P(N = n, L = l) = P(N = n | L = l)P(L = l)$$

$$= \binom{l}{n}(ap)^n(1 - ap)^{l-n}\frac{e^{-\lambda}\lambda^l}{l!}$$

$$= \frac{l!}{n!(l-n)!}(ap)^n(1-ap)^{l-n}\frac{e^{-\lambda}\lambda^l}{l!}. \qquad (4.10)$$

Therefore, the probability of the factually observed numbers of deaths M, N, and Z is

$$P(M = m, N = n, \mathrm{K} + L = z)$$

$$= \sum_{k=m}^{z-n} P(M = m, N = n, \mathrm{K} = k, L = z - k)$$

$$= \sum_{k=m}^{z-n} P(M = m, \mathrm{K} = k)P(N = n, L = z - k)$$

$$= \sum_{k=m}^{z-n} \frac{k!}{m!(k-m)!}p^m(1-p)^{k-m}\frac{e^{-k_0}k_0^k}{k!}\frac{l!}{n!(l-n)!}(ap)^n(1-ap)^{l-n}\frac{e^{-\lambda}\lambda^l}{l!}$$

$$= \frac{p^m k_0^m}{m!}\frac{a^n p^n \lambda^n}{n!}e^{-(k_0+\lambda)}\sum_{k=m}^{z-n}\frac{k_0^{k-m}\lambda^{z-k-n}}{(k-m)!(z-k-n)!}(1-ap)^{z-k-n}(1-p)^{k-m}$$

$$= \frac{p^m k_0^m}{m!}\frac{a^n p^n \lambda^n}{n!}e^{-(k_0+\lambda)}\sum_{k=m}^{z-n}\frac{k_0^{k-m}\lambda^{z-k-n}}{(k-m)!(z-k-n)!}(1-ap)^{z-k-n}(1-p)^{k-m}\frac{(1-ap)^{k-m}}{(1-ap)^{k-m}}$$

$$= \frac{p^m k_0^m}{m!}\frac{a^n p^n \lambda^n}{n!}e^{-(k_0+\lambda)}(1-ap)^{a-m-n}\sum_{k=m}^{z-n}\frac{\left[k_0\left(\frac{1-p}{1-ap}\right)\right]^{k-m}\lambda^{z-k-n}}{(k-m)!(z-k-n)!}$$

$$= \frac{p^m k_0^m}{m!}\frac{a^n p^n \lambda^n}{n!}e^{-(k_0+\lambda)}(1-ap)^{a-m-n}\frac{\left[k_0\left(\frac{1-p}{1-ap}\right)+\lambda\right]^{z-m-n}}{(z-m-n)!}. \qquad (4.11)$$

For $a = 1$, the model (4.11)reduces to the original model.

Then the new loglikelihood function is

$$\ln L = m\ln p + m\ln k_0 + n\ln p + n\ln a + n\ln\lambda - (k_0 + \lambda) + (z - m - n)\ln\left[k_0(1-p) + \lambda(1-ap)\right]$$
$$- \ln\left[\ m!n!(z-m-n)!\ \right].$$

The likelihood equations are

$$\frac{\partial}{\partial p}\ln L = \frac{m}{p} + \frac{n}{p} - \frac{(z-m-n)(k_0 + a\lambda)}{k_0(1-p) + \lambda(1-ap)} = 0,$$
$$\frac{\partial}{\partial k_0}\ln L = \frac{m}{k_0} - 1 + \frac{(z-m-n)(1-p)}{k_0(1-p) + \lambda(1-ap)} = 0,$$
$$\frac{\partial}{\partial \lambda}\ln L = \frac{n}{\lambda} - 1 + \frac{(z-m-n)(1-ap)}{k_0(1-p) + \lambda(1-ap)} = 0,$$
$$\frac{\partial}{\partial a}\ln L = \frac{n}{a} - \frac{(z-m-n)(\lambda p)}{k_0(1-p) + \lambda(1-ap)} = 0\ .$$

We get the MLE by solving the above equations for the parameters as

$$\hat{k}_0 = \frac{amz}{am+n},$$
$$\hat{p} = \frac{am+n}{az}, \qquad (4.12)$$
$$\hat{\lambda} = \frac{nz}{am+n}\ .$$

Note that the estimation of a is not feasible because of the relationship $\frac{\lambda}{k_0} = \frac{n}{ma}$ between the parameters. So we assume that a is known.

The information matrix for the MLE is given by

$$J^{*}=\begin{pmatrix}\frac{(am+n)^{2}(z-m-n)+m(za-zm-n)^{2}}{az^{2}m(z-m-n)} & \frac{nz(1-a)}{(am+n)(z-m-n)} & \frac{(az-am-n)(z-am-n)}{az^{2}(a-m-n)}\\ \frac{nz(1-a)}{(am+n)(z-m-n)} & \frac{a^{2}z^{3}(m+n)}{(am+n)^{2}(z-m-n)} & \frac{amz(a-1)}{(am+n)(z-m-n)}\\ \frac{(az-am-n)(z-am-n)}{az^{2}(z-m-n)} & \frac{amz(a-1)}{(am+n)(z-m-n)} & \frac{(z-m-n)(am+n)^{2}+n(z-am-n)^{2}}{nz^{2}(z-m-n)}\end{pmatrix}$$

Therefore, $Var(\hat{k}_0)=\left[B_{11}-B_{12}B_{22}^{-1}B_{21}\right]^{-1}$,

where $B_{11}=\dfrac{(am+n)^2(z-m-n)+m(az-am-n)^2}{az^2m(z-m-n)}$,

$$B_{12}=\begin{pmatrix}\dfrac{nz(1-a)}{(am+n)(z-m-n)} & \dfrac{(z-am-n)(az-am-n)}{az^2(z-m-n)}\end{pmatrix},$$

$$B_{21}=\begin{pmatrix}\dfrac{nz(1-a)}{(am+n)(z-m-n)}\\ \dfrac{(z-am-n)(az-am-n)}{az^2(z-m-n)}\end{pmatrix},$$

$$B_{22}=\begin{pmatrix}\dfrac{a^2z^3(m+n)}{(am+n)^2(z-m-n)} & \dfrac{amz(a-1)}{(am+n)(z-m-n)}\\ \dfrac{amz(a-1)}{(am+n)(z-m-n)} & \dfrac{(z-m-n)(am+n)^2+n(z-am-n)^2}{nz^2(z-m-n)}\end{pmatrix},$$

$$B_{12}B_{22}^{-1}B_{21}=\frac{n\left[nz(1-a)^2(am+n)+\left(az(m+n)-(am+n)^2\right)(az-am-n)(a-zm-n)\right]}{a^2z^2(z-m-n)(a^2m^3+2anm^2+mnz+mn^2+zn^2)}.$$

The confidence limits for the parameter k_0 are

Lower Bound: $\underline{k_0}=\hat{k}_0-z_{\frac{\alpha}{2}}\sqrt{Var(\hat{k}_0)}$, Upper Bound: $\overline{k_0}=\hat{k}_0+z_{\frac{\alpha}{2}}\sqrt{Var(\hat{k}_0)}$.

The corresponding confidence interval for SMR is $\left(\dfrac{\underline{k_0}}{E},\dfrac{\overline{k_0}}{E}\right)$.

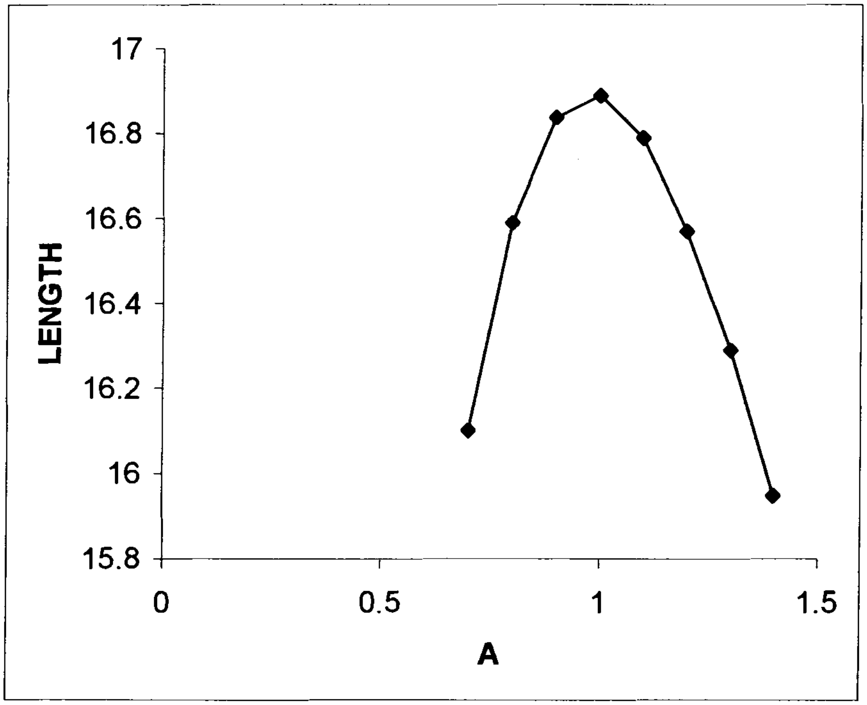
In this case assuming that a is known, the length of the confidence interval for SMR would vary with a. Thus, we have the following conclusions:

(1). The minimum value of a is around 0.66 for malignant neoplasms, because for $a < 0.66$, p is greater than 1. For other cases, the minimum values of a are around 0.725.

(2). For neoplasms, when a increases, the length of the confidence interval for SMR goes up to a maximum value, and then goes down. The following graph shows the effect of varying 'a' in the case of neoplasms. It is evident that the maximum value of the length of the confidence interval occurs for $a = 1$. Thus, in this case, by assuming $a \neq 1$, we are led to shorter confidence intervals for SMR. For other cases, the behavior of the length of confidence interval is different.

FIGURE 4.1

The Graph of The Confidence Interval for SMR (Malignant Neoplasms)



Chapter 5

CONCLUSIONS AND REMARKS

The discussion in this thesis shows that to estimate the Standardized Mortality Ratio, we have to estimate the observed value, which is assumed to be Poisson distributed. There are several ways to estimate the Poisson parameter:

(1). Normal approximation

(2). Exact confidence interval by using chi-square distribution

(3). Binomial approximation

(4). Shortcut methods used in epidemiological studies

Our simulation studies demonstrate that the binomial approximation methods are not of much use because the coverage probability for every one of them is 1, while the nominal value is 0.95. Comparing the shortcut methods and the newly proposed methods, we notice that one of our methods performs better than the others in terms of the length of the intervals and the coverage probability.

The problem of missing certificates is quite natural in follow-up studies. The problem can arise with the nonaccessibility of the causes of death of all the deceased study participants. In this thesis, a statistical model for this situation is developed to derive a maximum likelihood estimator (MLE) for the true unknown number of death from a specified cause. The model assumes that the probability of the availability of the death certificate in both the disease of interest and otherwise is the same.

In addition to the procedures presented in this thesis, we tried to develop a new statistical model by not assuming that the probability of the availability of death certificate is the same for the disease of interest and otherwise. The probability for the noncancer is modified to ap, where a can be different from 1. We re-estimate the true (but unknown) number of death from a specified cause. We find that the length for the confidence interval of SMR would change when a varies. In the case of neoplasms, the maximum value of the length of the confidence interval occurs when $a = 1$. Thus, in this case, by assuming $a \neq 1$, we get shorter confidence intervals for SMR. For other cases, the behavior of the length of confidence interval is different.

As has been noticed before, we could not estimate the value of a due to certain constraints. In further work, we would like to find ways to estimate the value of a instead of assuming that a is known. By means of simulation studies, we would like to compare the estimates of SMR obtained by the original method and the modified procedure.

REFERENCES

[1] Adzersen K.H., Becker N., Steindorf K., and Frentzel-Beyme R. (1997), *Cancer Mortality in a Germen Cohort of Male Iron Foundry Workers*, Heidelberg: German Cancer Research Center.

[2] Blyth C.R. (1986), *Approximate Binomial Confidence Limits*, Journal of the American Statistical Association, 81,843-855.

[3] Breslow N.E. and Day N.E. (1987), *Statistical Model in Cancer Research*, Volume II; *The design and Analysis of Cohort Studies*, IARC Scientific Publications 82. Lyon:IARC.

[4] Casella G. and Berger R. (1990), Statistical Inference, Duxbury Press, Belmont, California.

[5] Gupta R.C. (1977), *On Characterizing Distribution by The Ratio of Variance and Mean,* Math. Operationsforsch. Statist.,Ser. Statistics, Vol. 8 No. 4, 523-527.

[6] Molenaar W. (1973), *Approximations to the Poisson, Binomial and Hypergeometric Distribution Functions,* Amsterdam: Mathematisch Centrum.

[7] Rittgen W. and Becker N. (2000), *SMR Analysis of historical Follow-Up Studies with Missing Death Certificates,* Biometrics 56, 1164-1169.

[8] Schwertman N.C. and Martinez R. A. (1994), *Approximate Poisson Confidence Limits,* Marcel Dekker.

[9] Ury H.K. and Wiggins A.D. (1985), *Another Shortcut Method for Calculating The Confidence Interval of A Poisson Variable (or of A Standardized Mortality Ratio),* Am J Epidemiol, 122, 197-198.

[10] Vandenbroucke J.P. (1982), *A Shortcut Method for Calculating The 95 Per Cent Confidence Interval of The Standardized Mortality Ratio,* Am J Epidemol, 115, 303-304.

BIOGRAPHY OF THE AUTHOR

Bingxia Wang was born on October 6, 1972 in Shanghai, China. She was raised in her hometown and attended high school in Wuhan High School, Hubei. She attended Central China Normal University in 1991 and graduated in 1995 with a Bachelor of Science in Mathematics. In the fall of 2000, after teaching in Tong Ji Medical University for four and half years, she enrolled in the graduate program of the Department of Mathematics and Statistics at the University of Maine with a concentration in Statistics. She is a candidate for the Master of Arts Degree in Mathematics from The University of Maine in August, 2002.