2004

Estimation of Standardized Mortality Ratio in Geographic Epidemiology

Anna Kettermann

ESTIMATION OF STANDARDIZED MORTALITY RATIO IN GEOGRAPHIC EPIDEMIOLOGY

By

Anna Kettermann

Diploma, University of Kaiserslautern, Germany, 2001

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, 2004

Advisory Committee:

Ramesh Gupta, Professor of Mathematics, Advisor

Pushpa Gupta, Professor of Mathematics

Henrik Bresinsky, Professor of Mathematics

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:

Date: June 4, 2004

ESTIMATION OF STANDARDIZED MORTALITY RATIO IN GEOGRAPHIC EPIDEMIOLOGY

By Anna Kettermann

Thesis Advisor: Dr. Ramesh Gupta

An Abstract of the Thesis Presented
in Partial Fulfillment of the Requirements for the
Degree of Master of Arts
(in Mathematics)
August, 2004

The analysis of geographic variation of disease and its representation on a map form an important topic of research in epidemiology and in public health in general. Identification of spatial heterogeneity of relative risk using morbidity and mortality data is required.

The usual technique of disease atlas generation consists of data collection (observed number of disease cases). These data are collected during a continuous period of time (5 to 10 years). The second aspect of atlas creation relates to the analysis of these data. A traditional measure of the spatial variation is usually taken as a ratio of the number of observed disease cases to the number of the expected disease cases for the given region. This measure is called the Standardized Mortality (morbidity) ratio (SMR). Our interest is to estimate the spatial variation, i.e. to estimate the mean and the variance of the SMR.

In this paper we will focus on the developments that avoid the pitfalls of the crude SMR. We will compare the results of nonparametric and parametric approaches to the SMR estimation. More specifically, we present a mixture model to evaluate the heterogeneity in estimating SMR. Simulation studies are carried out and the results are analyzed.

Acknowledgements

I would like to thank the Department of Mathematics and Statistics of the University of Maine for giving me the opportunity not only to do the research in the field of Statistics but also to teach and tutor university courses in areas of Mathematics and Statistics.

I would like to thank my thesis advisor, Dr. Ramesh Gupta, for his invaluable advice, help, support and encouragement at every stage of my graduate study.

I would like to thank the members of my committee, Dr. Pushpa Gupta and Dr. Bresinsky for the encouragement, support and advice in my pursuit of a Master's degree at the University of Maine.

TABLE OF CONTENTS

Chapter

List of Tables

List of Figures

Chapter 1

INTRODUCTION

An essential problem in the construction of disease maps is the variation of population density between urban and rural areas. Epidemiological investigation of factors such as climate, environmental pollution, and prosperity, which may relate to disease prevalence, must take into account the spatial heterogeneity of the population, which is both a denominator for any number of cases in a region, and itself an indicator of physical proximity which may be a correlate of disease.

In this paper we are going to investigate spatial heterogeneity in disease map construction, expanding on the work of Boehning and Sarol, which demonstrated a nonparametric approach. They were able to estimate the mean Standard Mortality Ratio (SMR) for disease maps, as the non-parametric estimators for the mean are unbiased, but found that estimation of variance of the SMR becomes difficult, as this estimator does become biased. We will here investigate the estimation of variation of the SMR in a fixed model, the mapping of hepatitis cases in Berlin, where the distribution of data is already known. Using a mixed-model approach, variance of SMR will be estimated by both Maximum Likelihood and non-parametric techniques. In order to confirm our results, we have run computer simulations of the data using S-PLUS routines. One of the chief advantages of the suggested approach is that in addition to the parameter estimation it also gives us an opportunity to estimate the confidence interval of the variance of the heterogeneity parameter.

In the beginning of the first chapter we are going to introduce the basic concepts of disease mapping, discuss the methods and analyze the problems of disease map construction.

In the second chapter we will look at the nonparametric approach to the estimation of the spatial variation. We will study the nonparametric approach in case of the observed mortality with a particular distribution (cases of Poisson and Binomial distributions). A general nonparametric approach and its application to the given data set (Hepatitis cases in Berlin) will be presented in the end of chapter 2.

In the third chapter, we are going to apply a fixed model to the data. The estimation of the spatial variation will be done with the help of the Maximum Likelihood approach. We are going to compare the outcomes of nonparametric and parametric calculations.

A simulation study based on the parametric approach may be found in the chapter four.

We conclude with some final remarks and recommendations.

Now, let us come to the main definitions.

Definition 1 ***Disease mapping*** *is a method of displaying the spatial distribution of disease occurrence. It is widely used in geographic epidemiology, especially in creation of disease atlases.*

One of the most important uses of disease mapping may be seen in disease surveillance and health outcome research. It has become widely accepted that a potentially fruitful way to monitor the disease status of a community is to look at the health data in time and space. The Centers for Disease Control (CDC) defines public surveillance as the ongoing systematic collection, analysis and interpretation of the health data essential to the planning, implementation and evaluation of public health practice.

Definition 2 *The* ***Standardized Mortality Ratio (SMR)*** *is widely used in epidemiology as a measure of disease occurrence. Generally, SMR is defined as the ratio of observed mortality to expected mortality for the given region.*

Let us think of a map divided into N regions. A formal definition of SMR for the region i is:

$$SMR_i = \frac{O_i}{E_i}, i = 1, ..., N. \tag{1.1}$$

Both values in this ratio represent the data for the same region i where:

O_i represents the number of observed deaths and E_i represents the expected number of deaths calculated from the reference population.

According to Boehning (2003), there are two most frequently used methods in disease map construction:

1. A classification is based upon a certain percentile of the empirical.
SMR-distribution

2. The classification is based upon the Poisson distribution

$$\begin{aligned} Poisson(o_i \quad &| \quad \lambda E_i) = \frac{\exp((-\lambda E_i)(\lambda E_i)^{o_i})}{o_i!} \qquad (1.2) \\ o_i \quad &= \quad 0, 1, 2, ... \qquad (1.3) \end{aligned}$$

Conventional methods of atlas comparisons have two problems:

1. No account is taken of varying population size over the map. Therefore, the SMR estimation based only on a few cases is not precise.
2. There is no common presentation format of the atlases.

The goal of this project is to find better results for the estimators of the mean and variance of the Standardized Mortality Ratio by comparing both parametric and nonparametric approaches.

Chapter 2

SPATIAL HETEROGENEITY AND ITS MEASURES. A NONPARAMETRIC APPROACH.

An important aspect of disease mapping is the concept of **heterogeneity**. For example, we have a sampling model with X being a variable of interest. X has density $p(x, \lambda)$, where λ is a scalar parameter. For a given subpopulation, the density $p(x, \lambda)$ might be a good fit, but the value of λ is not able to cover the whole population of interest. In this case, we have **heterogeneous population**. The parameter of interest. λ, varies from one geographical region to another, so it is impossible to verify which subpopulation is generating the variation. As the data are drawn from distinct geographical regions, the value of λ is not able to capture all characteristics of the data. This means that λ is not a constant. Thus it can be regarded as a random variable with some possibly unknown distribution. We are interested in the estimation of the first two moments of this distribution. For these we present the following general approach.

Let

$$E(O_i \mid \Theta) = a\Theta,$$

where Θ is a parameter and a is a real constant and

$$Var(O_i \mid \Theta) = a_1 + a_2\Theta + a_3\Theta^2 \tag{2.1}$$

where a_1, a_2, a_3 are real constants. Thus

$$E(O_i) = E[E(O_i \mid \Theta)] = aE(\Theta) = a\mu, \tag{2.2}$$

where $\mu = E(\Theta)$

Looking at (2.2), a natural estimator of μ can be taken as $\hat{\mu} = \frac{\bar{O}}{a}$, where $\bar{O} = \frac{\sum_{i=1}^{N} O_i}{N}$. Now, using the formula

$$Var(O_i) = E[Var(O_i \mid \Theta)] + Var[E(O_i \mid \Theta)], \tag{2.3}$$

we have

$$Var(O_i) = a_1 + a_2\mu + a_3[\tau^2 + \mu^2] + Var(a\Theta) = a_1 + a_2\mu + a_3\mu^2 + (a_3 + a^2)\tau^2 \tag{2.4}$$

Thus τ^2 is given by

$$\tau^2 = \frac{Var(O_i) - a_1 - (a_2\mu + a_3\mu^2)}{a_3 + a^2} \tag{2.5}$$

A natural estimator of τ^2 can be taken as:

$$\hat{\tau}^2 = \frac{S^2 - a_1 - (a_2\frac{\bar{O}}{a} + a_3^2\frac{\bar{O}^2}{a^2})}{a_3 + a^2} \tag{2.6}$$

where S^2 is the sample variance of O_i' s. We now apply the above approach to our problem.

We consider two cases with random variable O_i having a) A Poisson distribution; b) A Binomial distribution. The goal is to be able to estimate the variance of spatial heterogeneity, i.e. the variance of the parameter Θ.

CASE 1: (O_i has a Poisson distribution). In this case:

$E(O_i \mid \Theta) = \Theta$, $Var(O_i \mid \Theta) = \Theta$ thus gives $\hat{\mu} = \bar{O}$, and $\hat{\tau}^2 = S^2 - \bar{O} = S^2 - \hat{\mu}$.

CASE 2: (O_i has a Binomial distribution). In this case:

$$E(O_i \mid \Theta) = N\Theta \text{ and } Var(O_i \mid \Theta) = N\Theta(1-\Theta) = N\Theta - N\Theta^2 \tag{2.7}$$

This gives $\hat{\mu} = \frac{\bar{O}}{N}$ and

$$\hat{\tau}^2 = \frac{S^2}{N[N-1]} - \frac{\frac{\bar{O}}{N}(1 - \frac{\bar{O}}{N})}{[N-1]}. \tag{2.8}$$

2.1 Mixture model approach

In this approach we treat λ as a random variable. The goal is to find the estimator of the mean and the variance of the spatial heterogeneity parameter, i.e. the mean and the variance of the parameter λ.

2.1.1 Example for the Poisson distribution

Example 3 *We consider a random variable O_i having a Poisson distribution, given as:*

$$\begin{aligned} O_i &\backsim Poisson(o_i \mid \lambda E_i). & (2.9) \\ \textit{with probability mass function. } f(o_i) &= \int_0^{+\infty} Poisson(o_i \mid \lambda E_i) p(\lambda) d\lambda. & (2.10) \end{aligned}$$

Assumption: λ has a distribution P with density $p(\lambda)$, mean μ and variance τ^2 Since $O_i \geqslant 0$ (O_i represents count data and therefore can not be negative), we use integration on the interval $[0, +\infty)$, i.e. $f(o_i \mid \lambda E_i) = \int_0^{+\infty} Poisson(o_i) p(\lambda) d\lambda$

First, let us determine the variance and mean of the variable O_i for this particular case. Using formula (2.3), we have

$$E(O_i) = E[E(O_i \mid \lambda)] = E[\lambda E_i] = E_i E(\lambda) = \mu E_i \tag{2.11}$$

$$Var(O_i) = E_i E(\lambda) + E_i^2 Var(\lambda) = E_i \mu + E_i^2 \tau^2. \tag{2.12}$$

where E_i is a constant.

The above results are incorporated into the calculations of the mean and variance of the Standardized Mortality Ratio of the region i (SMR_i). Thus we have:

$$E(SMR_i) = E(\frac{O_i}{E_i}) = \frac{\mu E_i}{E_i} = \mu. \tag{2.13}$$

and

$$Var(SMR_i) = Var(\frac{O_i}{E_i}) = \frac{1}{E_i^2}Var(O_i) = \frac{1}{E_i^2}[\mu E_i + E_i^2\tau^2] = \frac{\mu}{E_i} + \tau^2 \tag{2.14}$$

This leads us to the following expression:

$$\tau^2 = Var(SMR_i) - \frac{\mu}{E_i} = \frac{Var(O_i)}{E_i^2} - \frac{\mu}{E_i}. \tag{2.15}$$

Thus yields the following estimator of τ^2

$$\hat{\tau}^2 = \frac{1}{N}\sum_{i=1}^{N}(SMR_i - \mu)^2 - \frac{\mu}{N}\sum_{i=1}^{N}\frac{1}{E_i}. \tag{2.16}$$

(assuming that μ is known.).

We now show the following:

RESULT:

$\hat{\tau}^2$ is unbiased for τ^2, i.e. $E(\hat{\tau}^2) = \tau^2$

Proof.

$$E(\hat{\tau}^2) = \tau^2 + \frac{1}{N}\sum_{i=1}^{N}\frac{\mu}{E_i} - \frac{1}{N}\sum_{i=1}^{N}\frac{\mu}{E_i} = \tau^2. \tag{2.17}$$

■

2.1.2 Example for the Binomial distribution

Example 4 *Now, let us take another example, where $O_1, \ldots O_N$ is a random sample from a binomial distribution with pmf $p(o_i, \lambda) = \binom{N}{o_i}(\lambda E_i)^{o_i}(1 - \lambda E_i)^{N-o}$.*

In this case:

$$\begin{aligned} E(O_i \quad &| \quad \lambda E_i) = N\lambda E_i, \qquad (2.18) \\ Var(Oi \quad &| \quad \lambda E_i) = N\lambda E_i(1 - \lambda E_i), \end{aligned}$$

Now using the formula (2.3), we get

$$E(O_i) = E[E(O_i \mid \lambda E_i)] = NE_iE(\lambda) = NE_i\mu.$$

and

$$Var(O_i) = E(Var(Oi \mid \lambda E_i)) + Var(E(O_i \mid \lambda E_i)) = N(N-1)E_i^2\tau^2 + NE_i\mu(1-\mu E_i). \quad (2.19)$$

This implies:

$$\tau^2 = \frac{Var(O_i) - NE_i\mu(1 - E_i\mu)}{N[N-1]E_i^2} = \frac{Var(O_i)}{N[N-1]E_i^2} - \frac{E_i\mu(1 - E_i\mu)}{[N-1]E_i^2}. \quad (2.20)$$

An estimator of μ can be taken as $\hat{\mu} = \frac{\bar{O}}{N}\sum_{i=1}^{N}\frac{1}{E_i}$, similarly, an estimator of $\tau^2$is

$$\hat{\tau}^2 = \frac{1}{N^2[N-1]}\sum_{i=1}^{N}\frac{Var(O_i)}{E_i^2} - \frac{\frac{1}{N}\bar{O}(1 - \frac{\bar{O}}{N})}{N[N-1]}\sum_{i=1}^{N}\frac{1}{E_i^2}. \quad (2.21)$$

Example 5 *Applying this result to the SMR_i estimation gives:*

$$\begin{aligned} E(SMR_i) &= E(\frac{O_i}{E_i}) = N\mu & (2.22) \\ Var(SMR_i) &= Var(\frac{O_i}{E_i}) = \frac{Var(O_i)}{E_i^2} = N(N-1)\tau^2 + \frac{N\mu}{E_i} - N\mu^2 \end{aligned}$$

2.2 A nonparametric way to estimate the variance of the heterogeneity distribution P.

Our goal is to estimate the variation τ^2 of the heterogeneity parameter.

Method 1 Boehning in his article in (2003) introduced the following variables:

$$W_i = \frac{(O_i - \mu E_i)^2 - \mu E_i}{E_i^2} = (SMR_i - \mu)^2 - \frac{\mu}{E_i}. \; i = 1, .., N. \quad (2.23)$$

W_i is an unbiased estimator of τ^2 can be seen as follows:

$$E(W_i) = E[(SMR_i - \mu)^2] - \frac{\mu}{E_i} = Var(SMR_i) - \frac{\mu}{E_i} = \frac{\mu}{E_i} + \tau^2 - \frac{\mu}{E_i} = \tau^2 \qquad (2.24)$$

As was suggested in the paper referenced above, we can now define a function combining all W_i's:

$$T_\alpha(w) = \frac{\sum_{i=1}^{N} \alpha_i W_i}{\sum_{i=1}^{N} \alpha_i} \qquad (2.25)$$

RESULT:

$T_\alpha(w)$ is an unbiased estimator of τ^2

Proof. $E(T_\alpha(w)) = \frac{E(\sum_{i=1}^{N} \alpha_i W_i)}{\sum_{i=1}^{N} \alpha_i} = \frac{\sum_{i=1}^{N} \alpha_i E(W_i)}{\sum_{i=1}^{N} \alpha_i} = \tau^2$ ■

Different forms of $\alpha_i's$ have been suggested in the literature. Since $T_\alpha(w)$ is unbiased, in order to evaluate the choice of $\alpha_i's$, we will need to compare the variances of $T_\alpha(w)$ corresponding to different choices of $\alpha_i's$. We are interested to find the set of $\alpha_i's$ which would minimize the variance of $T_\alpha(w)$.

Case 1.(Boehning)

$$\alpha_i = \frac{1}{N} \qquad (2.26)$$

In this case, $T_\alpha(w)$ has the variance:

$$Var(T_\alpha(w)) = Var\left[\frac{1}{N}\sum_{i=1}^{N} W_i\right] = \frac{1}{N^2} Var(\sum W_i). \qquad (2.27)$$

Case 2. (Bautista)

$$\alpha_i = E_i^2. \qquad (2.28)$$

Now $T_\alpha(w)$ has the variance:

$$Var(T_\alpha(w)) = \frac{1}{(\sum_{i=1}^{N} E_i^2)^2}\left[E_1^4 Var(W_1) + ... + E_N^4 Var(W_N)\right] \qquad (2.29)$$

Case 3.

$$\alpha_i = \frac{1}{Var(W_i)}. \tag{2.30}$$

and the variance of $T_\alpha(w)$ is:

$$\begin{aligned} Var(T_\alpha(w)) &= Var\left[\frac{\frac{W_1}{Var(W_1)} + ... + \frac{W_N}{Var(W_N)}}{\sum_{i=1}^{N} \frac{1}{Var(W_i)}}\right] = \left[\frac{\sum_{i=1}^{N} \frac{Var(Wi)}{[Var(W_i)]^2}}{\left[\sum_{i=1}^{N} \frac{1}{Var(W_i)}\right]^2}\right] \qquad (2.31) \\ &= \left[\frac{\sum_{i=1}^{N} \frac{1}{Var(W_i)}}{\left[\sum_{i=1}^{N} \frac{1}{Var(W_i)}\right]^2}\right] = \frac{1}{\sum_{i=1}^{N} \frac{1}{Var(W_i)}}. \end{aligned}$$

RESULT:

We now show that $T_\alpha(w)$ with $\alpha_i = \frac{1}{Var(W_i)}$ yields a best linear unbiased estimator (BLUE).

Proof:

We need to minimize $Var(T_\alpha(w))$, i.e. minimize the the variance of the numerator of $T_\alpha(w)$:

$$Var(\sum_{i=1}^{N} \alpha_i W_i) = \sum_{i=1}^{N} \alpha_i^2 Var(W_i) = \sigma^2 \sum_{i=1}^{N} \alpha_i^2.$$

Where $\sigma^2 = Var(W_i)$. Now

$$\begin{aligned} \sum_{i=1}^{N} \alpha_i^2 &= \sum_{i=1}^{N} \left[\alpha_i - \frac{1}{Var(W_i)} + \frac{1}{Var(W_i)}\right]^2 \\ &= \sum_{i=1}^{N} (\alpha_i - \frac{1}{Var(W_i)})^2 + 2(\sum_{i=1}^{N} (\alpha_i - \frac{1}{Var(W_i)})(\frac{1}{Var(W_i)}) + \sum_{i=1}^{N} (\frac{1}{Var(W_i)})^2. \end{aligned}$$

Thus this expression is minimized when $\alpha_i = \frac{1}{Var(W_i)}$

Finally, we can estimate τ^2 in the following way:
(using $\alpha_i = \frac{1}{N}$, an estimator of τ^2 is):

$$\hat{\tau}^2 = \frac{1}{N}\left[\sum_{i=1}^{N} \frac{(O_i - E_i\mu)^2}{E_i^2} - \mu \sum_{i=1}^{N} \frac{1}{E_i}\right] = \frac{\sum_{i=1}^{N} W_i}{N} = \bar{W}. \tag{2.32}$$

Resulting in an unbiased estimator since we know that $E(W_i) = \tau^2$, i.e. $E(\hat{\tau}^2) = \tau^2$

Method 2. An alternative way of nonparametric estimation of μ and τ^2

For the rest of this section we will be working with nonparametric estimators of the spatial heterogeneity which were suggested by Boehning and Sarol (2000).

A nonparametric estimator of the variance of λ could be written as it is shown in (2.33), using the fact that $SMR_i = \frac{O_i}{E_i}$ and applied to the formula (2.32) to arrive at (2.33). The essence of the second method is contained in the following:

$$\hat{\tau}^2_{\mu} = \frac{1}{N}\left[\sum_{i=1}^{N}(SMR_i - \hat{\mu})^2 - \hat{\mu}\sum_{i=1}^{N}\frac{1}{E_i}\right]. \tag{2.33}$$

There are two nonparametric estimators presented for μ, the expected value of λ. Those estimators could later be applied to the expression of the variance (2.33) and bring us to $\hat{\tau}^2_{simple}$ and $\hat{\tau}^2_{pooled}$:

1. Simple mean

$$\hat{\mu}_{simple} = \frac{1}{N}\sum_{i=1}^{N}\frac{O_i}{E_i} \tag{2.34}$$

2. Pooled mean:

$$\hat{\mu}_{pooled} = \frac{\sum_{i=1}^{N} O_i}{\sum_{i=1}^{N} E_i} \tag{2.35}$$

RESULT:

$\hat{\mu}_{simple}$ is unbiased and the variance $Var(\hat{\mu}_{simple}) = \mu\frac{1}{N^2}\sum_{i=1}^{N}\frac{1}{E_i} + \tau^2\frac{1}{N}$

Proof.

$$E(\hat{\mu}_{simple}) = E(\frac{1}{N}\sum_{i=1}^{N}\frac{O_i}{E_i}) = \frac{1}{N}\sum_{i=1}^{N}\frac{E(O_i)}{E_i} = \frac{1}{N}\sum_{i=1}^{N}\frac{\mu E_i}{E_i} = \mu.$$

$$Var(\hat{\mu}_{simple}) = Var(\frac{1}{N}\sum_{i=1}^{N}\frac{O_i}{E_i}) = \frac{1}{N^2}Var(\sum_{i=1}^{N}\frac{O_i}{E_i}) = \frac{1}{N^2}(\sum_{i=1}^{N}\frac{1}{E_i^2}Var(O_i))$$

$$= \frac{1}{N^2}[\sum_{i=1}^{N}\frac{1}{E_i^2}(E_i\mu + E_i^2\tau^2)] = \frac{1}{N^2}[\sum_{i=1}^{N}\frac{\mu}{E_i} + \tau^2] = \frac{1}{N^2}[\mu\sum_{i=1}^{N}\frac{1}{E_i} + N\tau^2] = \frac{\mu}{N^2}\sum_{i=1}^{N}\frac{1}{E_i} + \frac{\tau^2}{N}.$$

■

RESULT:

$\hat{\mu}_{pooled}$ is also unbiased with the variance $Var(\hat{\mu}_{pooled}) = \mu\frac{1}{\sum\limits_{i=1}^{N} E_i} + \tau^2\frac{\sum\limits_{i=1}^{N} E_i^2}{(\sum\limits_{i=1}^{N} E_i)^2}$.

Proof.

$$E(\hat{\mu}_{pooled}) = \frac{\sum\limits_{i=1}^{N} \mu E_i}{\sum\limits_{i=1}^{N} E_i} = \mu\frac{\sum\limits_{i=1}^{N} E_i}{\sum\limits_{i=1}^{N} E_i} = \mu, \tag{2.36}$$

$$Var(\hat{\mu}_{pooled}) = Var(\frac{\sum\limits_{i=1}^{N} O_i}{\sum\limits_{i=1}^{N} E_i}) = \frac{1}{(\sum\limits_{i=1}^{N} E_i)^2}\sum_{i=1}^{N} Var(O_i) = \frac{1}{(\sum\limits_{i=1}^{N} E_i)^2}[\sum_{i=1}^{N}(E_i\mu + E_i^2\tau^2)]$$

$$= \frac{1}{(\sum\limits_{i=1}^{N} E_i)^2}[\mu\sum_{i=1}^{N} E_i + \tau^2\sum_{i=1}^{N} E_i^2] = \frac{\mu}{\sum\limits_{i=1}^{N} E_i} + \tau^2\frac{\sum\limits_{i=1}^{N} E_i{}^2}{(\sum\limits_{i=1}^{N} E_i)^2} \quad \blacksquare$$

Let us now look at three possible ways to estimate the variance of parameter λ. Using the general formula for the variance (2.33), and inserting $\hat{\mu}_{simple}$ instead of $\hat{\mu}$,we get:

$$\hat{\tau}^2_{simple} = \frac{1}{N-1}\sum_{i=1}^{N}\left(SMR_i - \hat{\mu}_{simple}\right)^2 - \frac{1}{N}\hat{\mu}_{simple}\sum_{i=1}^{N}\frac{1}{E_i} \tag{2.37}$$

RESULT:

$\hat{\tau}^2_{simple}$ is an unbiased estimator of τ^2

Proof. using the fact that

$$Var(SMR_i) = \frac{\mu}{E_i} + \tau^2 \tag{2.38}$$

and

$$Var(\hat{\mu}_{simple}) = \mu\frac{1}{N^2}\sum_{i=1}^{N}\frac{1}{E_i} + \tau^2\frac{1}{N}. \tag{2.39}$$

$$E(\hat{\tau}^2_{simple}) = E[\frac{1}{N-1}\sum_{i=1}^{N}\left(\frac{O_i}{E_i}\underbrace{-\mu+\mu}_{\text{add and subtract}} - \hat{\mu}_{simple}\right)^2 - \frac{1}{N}\hat{\mu}_{simple}\sum_{i=1}^{N}\frac{1}{E_i}]$$

$$
\begin{aligned}
&= E[\frac{1}{N-1}\sum_{i=1}^{N}\left((\frac{O_i}{E_i}-\mu)^2+2(\mu-\hat{\mu}_{simple})(\frac{O_i}{E_i}-\mu)+(\mu-\hat{\mu}_{simple})^2\right)-\frac{1}{N}\hat{\mu}_{simple}\sum_{i=1}^{N}\frac{1}{E_i}],\\
&= \frac{1}{N-1}[\sum_{i=1}^{N}(\frac{\mu}{E_i}+\tau^2)+2\sum_{i=1}^{N}E\{(\mu-\hat{\mu}_{simple})(\frac{O_i}{E_i}-\mu)\}+E[N(\mu-\hat{\mu}_{simple})^2]-\frac{1}{N}E(\hat{\mu}_{simple})\sum_{i=1}^{N}\frac{1}{E_i}\\
&= \frac{1}{N-1}[\mu\sum_{i=1}^{N}\frac{1}{E_i}+N\tau^2-2NE\{(\mu-\hat{\mu}_{simple})^2\}+NE[(\mu-\hat{\mu}_{simple})^2]-\frac{1}{N}\mu\sum_{i=1}^{N}\frac{1}{E_i}\\
&=\frac{1}{N-1}[\mu\sum_{i=1}^{N}\frac{1}{E_i}+N\tau^2-NE\{(\mu-\hat{\mu}_{simple})^2]\}-\frac{1}{N}\mu\sum_{i=1}^{N}\frac{1}{E_i}\\
&=\frac{1}{N-1}[\mu\sum_{i=1}^{N}\frac{1}{E_i}+N\tau^2-N\{\mu\frac{1}{N^2}\sum_{i=1}^{N}\frac{1}{E_i}+\tau^2\frac{1}{N}\}]-\frac{1}{N}\mu\sum_{i=1}^{N}\frac{1}{E_i}\\
&=\frac{1}{N-1}[\mu\sum_{i=1}^{N}\frac{1}{E_i}(1-\frac{1}{N})+N\tau^2-\tau^2]-\frac{1}{N}\mu\sum_{i=1}^{N}\frac{1}{E_i}\\
&=\frac{1}{N-1}\mu\sum_{i=1}^{N}\frac{1}{E_i}(1-\frac{1}{N})+\tau^2-\frac{1}{N}\mu\sum_{i=1}^{N}\frac{1}{E_i}=\tau^2 \quad \blacksquare
\end{aligned}
$$

A similar strategy is used for the $\hat{\tau}^2_{pooled}$. This is given by

$$
\hat{\tau}^2_{pooled}=\frac{1}{N-1}\sum_{i=1}^{N}\left(SMR_i-\hat{\mu}_{pooled}\right)^2-\frac{1}{N}\hat{\mu}_{pooled}\sum_{i=1}^{N}\frac{1}{E_i}. \qquad (2.40)
$$

We shall now show that $\hat{\tau}^2_{pooled}$ is biased:

Proof. For this, we will use the following formula:

$$
Cov(O_i,\hat{\mu}_{pooled}E_i) \qquad (2.41)
$$

$$
\begin{aligned}
Cov\left(O_i,\frac{E_i\sum_{i=1}^{N}O_i}{\sum_{i=1}^{N}E_i}\right) &= E_iCov\left(O_i,\frac{\sum_{i=1}^{N}O_i}{\sum_{i=1}^{N}E_i}\right)=\frac{E_i}{\sum_{i=1}^{N}E_i}Cov\left(O_i,\sum_{i=1}^{N}O_i\right)\\
&= \frac{E_i}{\sum_{i=1}^{N}E_i}Var\left(O_i\right)=\frac{E_i}{\sum_{i=1}^{N}E_i}(\mu E_i+E_i^2\tau^2).
\end{aligned}
$$

now

$$
\begin{aligned}
E(\hat{\tau}^2_{pooled}) &= E[\frac{1}{N-1}\sum_{i=1}^{N}\left(\frac{(O_i-\hat{\mu}_{pooled}E_i}{E_i}\right)^2-\frac{1}{N}\hat{\mu}_{pooled}\sum_{i=1}^{N}\frac{1}{E_i}\\
&=E[\frac{1}{N-1}\sum_{i=1}^{N}\left(\frac{(O_i\underbrace{-\mu E_i+\mu E_i}-\hat{\mu}_{pooled}E_i)^2}{E_i^2}\right)-\frac{1}{N}\hat{\mu}_{pooled}\sum_{i=1}^{N}\frac{1}{E_i}]\\
&=E[\frac{1}{N-1}\sum_{i=1}^{N}\left(\frac{(O_i-\mu E_i)^2+2(O_i-\mu E_i)(\mu E_i-\hat{\mu}_{pooled}E_i)+(\mu E_i-\hat{\mu}_{pooled}E_i)^2}{E_i^2}\right)-\frac{1}{N}\hat{\mu}_{pooled}\sum_{i=1}^{N}\frac{1}{E_i}]
\end{aligned}
$$

$$=\frac{1}{N-1}\left[\sum_{i=1}^{N}\frac{Var(O_i)-2Cov(O_i,\hat{\mu}_{pooled}E_i)+Var(\hat{\mu}_{pooled}E_i)}{E_i^2}\right]-\frac{1}{N}\mu\sum_{i=1}^{N}\frac{1}{E_i}$$

$$=\frac{1}{N-1}\sum_{i=1}^{N}\left[\frac{(\mu E_i+E_i^2\tau^2)-2Cov(O_i,\hat{\mu}_{pooled}E_i)+E_i^2\{\frac{\mu}{\sum\limits_{i=1}^{N}E_i}+\tau^2\frac{\sum\limits_{i=1}^{N}{E_i}^2}{(\sum\limits_{i=1}^{N}E_i)^2}\}}{E_i^2}\right]$$

$$-\frac{1}{N}\mu\sum_{i=1}^{N}\frac{1}{E_i}$$

$$=\frac{1}{N-1}\sum_{i=1}^{N}\left[\frac{(\mu E_i+E_i^2\tau^2)-2\frac{E_i}{\sum\limits_{i=1}^{N}E_i}(\mu E_i+E_i^2\tau^2)+E_i^2\{\frac{\mu}{\sum\limits_{i=1}^{N}E_i}+\tau^2\frac{\sum\limits_{i=1}^{N}{E_i}^2}{(\sum\limits_{i=1}^{N}E_i)^2}\}}{E_i^2}\right]-\frac{1}{N}\mu\sum_{i=1}^{N}\frac{1}{E_i}$$

$$=\frac{1}{N-1}\left[\mu\sum_{i=1}^{N}\frac{1}{E_i}+N\tau^2-\frac{2N\mu}{\sum_{i=1}^{N}E_i}-2\tau^2+\frac{N\mu}{\sum_{i=1}^{N}E_i}+N\tau^2\frac{\sum_{i=1}^{N}{E_i}^2}{(\sum_{i=1}^{N}E_i)^2}\right]-\frac{1}{N}\mu\sum_{i=1}^{N}\frac{1}{E_i}$$

$$=\mu\left[-\frac{N}{N-1}\frac{1}{\sum\limits_{i=1}^{N}E_i}+\sum_{i=1}^{N}\frac{1}{E_i}\{\frac{1}{N-1}-\frac{1}{N}\}\right]+\tau^2\left[\frac{N-2}{N-1}+\frac{N}{N-1}\frac{\sum_{i=1}^{N}{E_i}^2}{(\sum_{i=1}^{N}E_i)^2}\right]$$

$$=\mu\left[\frac{1}{N(N-1)}\sum_{i=1}^{N}\frac{1}{E_i}-\frac{1}{\sum\limits_{i=1}^{N}E_i}\frac{N}{N-1}\right]+\tau^2\left[\frac{N-2}{N-1}+\frac{N}{N-1}\frac{\sum\limits_{i=1}^{N}{E_i}^2}{(\sum\limits_{i=1}^{N}E_i)^2}\right].\ \blacksquare$$

Therefore, if we use $\hat{\mu}_{pooled}$ instead of $\hat{\mu}$ in (2.33) for $\hat{\tau}^2$, our estimator becomes biased. In order to avoid biasedness of $\hat{\tau}^2$, we adjust this estimation as follows:

$$\hat{\tau}^{2*}_{corrected}=\frac{(\hat{\tau}^2-\alpha_{corr}\hat{\mu})}{\beta_{corr}},\qquad(2.42)$$

where

$$\alpha_{corr} = \frac{1}{N(N-1)}\sum_{i=1}^{N}\frac{1}{E_i} - \frac{N}{N-1}\frac{1}{\sum_{i=1}^{N}E_i} \tag{2.43}$$

$$\beta_{corr} = \frac{N-2+N\left[\frac{\sum_{i=1}^{N}E_i^2}{(\sum_{i=1}^{N}E_i)^2}\right]}{N-1}$$

A measure of heterogeneity

We need to test whether the variance of SMR is homogeneous, i.e. check to see if $\tau^2 = 0$ (homogeneous case). If $\tau^2 = 0$, then the expression of the variance of the standardized mortality ratio could be calculated as

$$Var(SMR_i) = \frac{\mu}{E_i}. \tag{2.44}$$

In order for us to evaluate the possibility of the spatial variation in the given data set, we introduce the **proportion of spatial heterogeneity (PSH),** which is defined as

$$PSH = \frac{\hat{\tau}^2}{\frac{1}{N}\sum_{i=1}^{N}(SMR_i - \mu)^2}. \tag{2.45}$$

If PSH is much larger than zero, that would imply that we have heterogeneity of variance

If PSH is very close to zero, that means that the spatial heterogeneity is relatively small and could possibly be negligible.

Here are some special features of PSH:

- 1) $0 \leq PSH \leq 1$.
- 2) $E(\text{numerator of } PSH) = \tau^2$
- 3) $E(\text{denominator of } PSH) = \frac{1}{N}\sum_{i=1}^{N}\frac{\mu}{E_i} + \tau^2$
- 4) $E(PSH) \approx \frac{\tau^2}{E(\text{denominator of } PSH)}$.

Proof. 1) is obvious, using the expression of $\hat{\tau}^2$ given by (2.33) ∎

2)$E(\text{numerator of } PSH) = \tau^2$, already shown

Proof. already shown ■

3) $E(\text{denominator of } PSH) = \frac{\mu}{N}\sum_{i=1}^{N}\frac{1}{E_i} + \tau^2$

Proof. $E\{\frac{1}{N}\sum_{i=1}^{N}(SMR_i-\mu)^2\} = \frac{1}{N}\sum_{i=1}^{N} Var(SMR_i) = \frac{1}{N}\sum_{i=1}^{N} Var(\frac{\mu}{E_i}+\tau^2) = \frac{\mu}{N}\sum_{i=1}^{N}\frac{1}{E_i} + \tau^2$ ■

4) $E(PSH) \approx \cfrac{1}{1 + \frac{\mu}{N\tau^2}\sum_{i=1}^{N}\frac{1}{E_i}}$

EXAMPLE (Hepatitis data)

We are given a data set which represents the number observed and expected cases of hepatitis in 23 regions of Berlin.

Hepatitis data

Area i	O_i	E_i	Area i	O_i	E_i
1	29	10.7121	13	25	8.3968
2	26	17.9929	14	11	15.6438
3	54	18.1699	15	11	11.8289
4	30	19.2110	16	2	9.9513
5	16	21.9611	17	2	10.8313
6	15	14.6268	18	9	18.3403
7	6	9.6220	19	2	5.1758
8	35	17.2671	20	3	10.9543
9	17	18.8230	21	11	20.0121
10	7	18.2705	22	5	13.8389
11	43	32.1823	23	2	12.7996
12	17	24.5929			

Table 2.1: Observed and Expected Hepatitis Cases in 23 City Regions of Berlin. Source: Berlin Census Bureau, 1995.

First, let's consider a graphical analysis of the data presented in table 2.1.

Due to the nature of the data, the observed cases of hepatitis are represented only by integer values, because we are actually measuring a variable with a binomial outcome (sick or healthy) and we record the sick cases as our objective. O_i represents the count data. Considering expected values, they can be either integer or noninteger real values, since they are taken from the reference population and also could be a result of numerical manipulations. Thus, the observed and expected values in this example do indeed have different distributions.

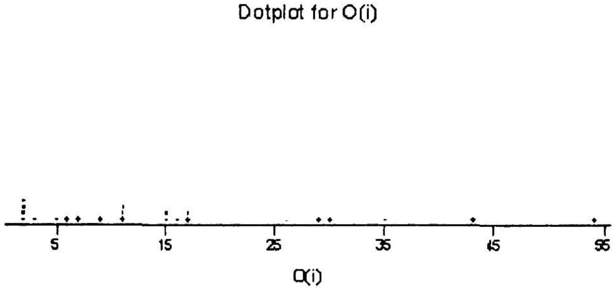


Figure 2-1: Dotplot of observed values

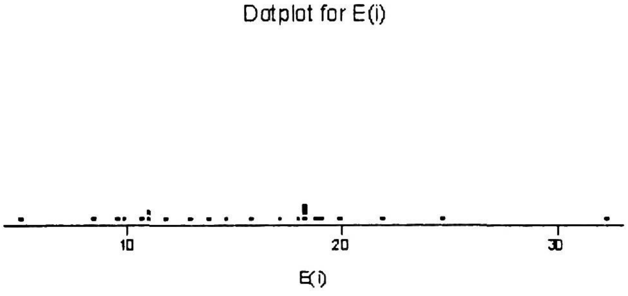


Figure 2-2: Dotplot of expected values

Looking at the dotplots of both data sets (figure 2-1 and 2-2), we see that the O_i (observed data) is primarily concentrated on the interval [2, 18], with the E_i (expected data) concentrated on the interval [8,20].

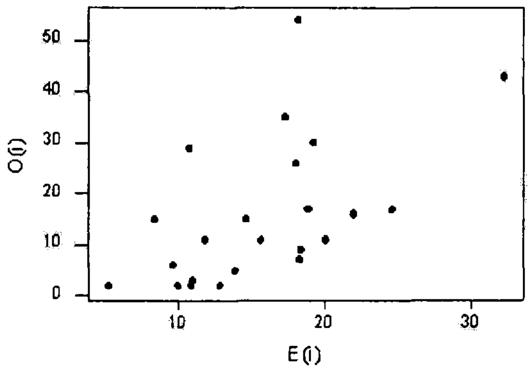


Figure 2-3: Scatterplot of observed vs expected values

If we view the data as a set of ordered pairs (figure 2-3). (Each region has two data points assigned to it, i.e. observed and expected number of disease cases). We see that data is mostly concentrated in the left lower corner, implying that we do not have observations with large absolute value. Two of the points can be considered as being outside of the data cluster. A better visual data representation can be presented by the graphs of boxplots.

Looking at the graph of the parallel bcxplots of our data [figure 2-4], we see that the observed values have a larger range than expected values. Looking at the height of the box, the distance between the third (75th percentile of the data) and the first (25th percentile of the data) quantiles is larger than the same points for the expected values. Examining the boxplot, it is apparent that in the case of observed values, the box is located in the lower part of the data, with the median shifted towards the bottom of the box. This tells us that the data is mostly concentrated in the beginning of the scale, i.e. the data is skewed to the right. For the expected values, our box is relatively centered in the data and the median is slightly shifted towards the top of the box. This tells us that expected value data seems to have greater symmetry than the observed data. Additionally, we see that the expected

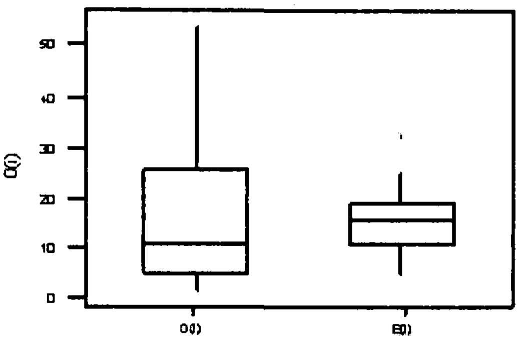


Figure 2-4: Parallel boxplots

Variable	N	Mean	Median	TrMean	StDev	SE Mean	Minimum	Maximum
O(i)	23	16.00	11.00	14.86	14.11	2.94	2.00	54.00
E(i)	23	15.70	15.64	15.42	6.00	1.25	5.18	32.18

Table 2.2: Statistical analysis of the samples

data also has an outlier. For a better visualization, let us consider the histograms of both data sets [figures 2-5 and 2-6]

Neither histogram contradicts our initial observation about skewness and data distribution.

The table (table 2.2) represents the statistical data analysis of both samples of the data set.

A test for homogeneity is based on the statistic:

$$\chi^2_{N-1} = \frac{\sum_{i=1}^{N} (O_i - \hat{\mu} E_i)}{\hat{\mu} E_i}. \tag{2.46}$$

This is a 1-sided test. Using our data, we get $\chi^2_{22} = 193.52$, using a pooled estimate of μ and $\chi^2_{22} = 202.92$, using a simple estimate of μ. Both values of the test statistic are much larger than the table value (33.924) which definitely indicates heterogeneity and therefore we can imply that $\tau^2 \neq 0$.

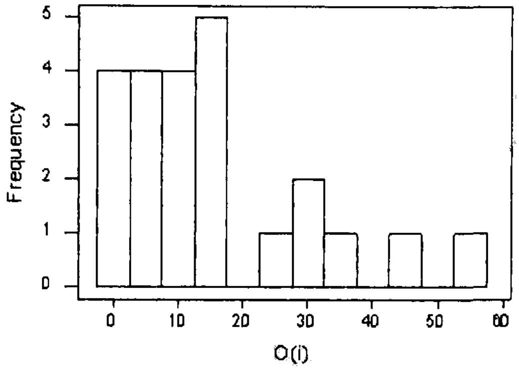


Figure 2-5: A histogram of observed values

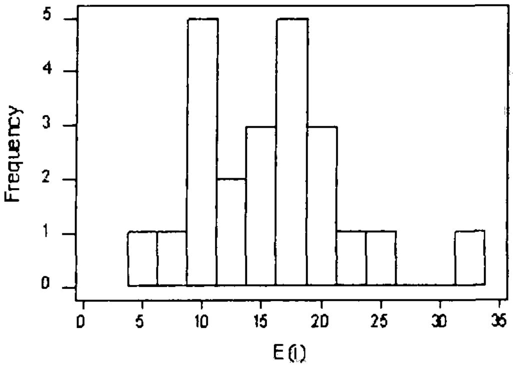


Figure 2-6: A histogram of expected values

Applying the nonparametric estimation technique presented above to the given data we see how to we estimate the mean and variance of the unknown parameter λ.

$\hat{\mu}_{pooled}$	1.018813
$\hat{\mu}_{simple}$	0.9751049
$\hat{\tau}^2_{pooled}$	0.5476439
$\hat{\tau}^2_{simple}$	0.5488984
$\hat{\tau}^2_{pooled.corrected}$	0.5437004
$\hat{\alpha}$	0.0004872763
$\hat{\beta}$	1.00634
$Var(SMR)_{pooled}$	0.5504645
$Var(SMR)_{simple}$	0.551598
$Var(SMR)_{pooled.corrected}$	0.546521
PSH_{pooled}	0.994876
PSH_{simple}	0.9951059
$PSH_{pooled.corrected}$	0.994839

Table 2.3: Results of nonparametric estimation

Analyzing the results, presented in the table 2.3, we could conclude that the numerical values of the estimates of the mean and the variance are not significantly affected by the choice of the estimator of the mean. That means that the values of the parameters in case of simple and pooled mean coming out to be close to each other. A numerical value of the PSH in all three cases are close to one. That would implicate a relatively high spatial heterogeneity.

Chapter 3

MAXIMUM LIKELIHOOD APPROACH.

In this chapter, we present the estimation of the heterogeneity parameter by the MLE method and investigate the performance of the resulting confidence intervals.

Let us assume that

$$O_i \mid \lambda \sim Poisson(O_i \mid \lambda E_i) \tag{3.1}$$

with conditional probability mass function

$$p(O_i \mid \lambda) = \frac{1}{O_i!}(\lambda E_i)^{O_i} \exp(-\lambda E_i) \tag{3.2}$$

and

$$p(\lambda) \sim Gamma(\alpha, \beta) \tag{3.3}$$

where the density has the following form:

$$p(\lambda \mid \alpha, \beta) = \frac{1}{\beta^\alpha \Gamma(\alpha)} \lambda^{\alpha-1} \exp(-\lambda/\beta). \tag{3.4}$$

The unconditional mass function of O_i is given by

$$g(o_i) \sim \int_0^\infty \underbrace{\frac{1}{o_i!}(\lambda E_i)^{o_i} \exp(-\lambda E_i)}_{p(O_i|\lambda)} \underbrace{\frac{1}{\beta^\alpha \Gamma(\alpha)} \lambda^{\alpha-1} \exp(-\lambda/\beta) d\lambda}_{p(\lambda)} = \tag{3.5}$$

$$= \binom{o_i + \alpha - 1}{\alpha - 1} \left(\frac{1}{1 + \beta E_i}\right)^{\alpha} \left(\frac{\beta E_i}{1 + \beta E_i}\right)^{o_i}$$

$$= \frac{1}{\Gamma(\alpha)\beta^{\alpha}} \frac{E_i^{o_i}}{o_i!} \int_0^{\infty} \lambda^{\alpha + o_i - 1} \exp[-\lambda(E_i + \frac{1}{\beta})] d\lambda$$

$$= \binom{o_i + \alpha - 1}{\alpha - 1} \left(\frac{1}{1 + \beta E_i}\right)^{\alpha} \left(\frac{\beta E_i}{1 + \beta E_i}\right)^{o_i}. \tag{3.6}$$

This new density resembles the negative binomial density.

A general negative binomial density with parameters r and p has of the following form:

$$p(x) = \binom{r + x - 1}{x} (p)^r (1 - p)^x, x = 0, 1, 2, \ldots.$$

where r corresponds to our α and x corresponds to o_i. Finally, $p = \frac{1}{1 + \beta E_i}$.

Application to our model:

Since we are interested in the estimation of the variance of the spatial heterogeneity parameter λ, without loss of generality, let us assume the mean of the Gamma distribution (3.6) to be equal to one and the variance to be equal to a, where a is a positive value. This implies that

$$\begin{aligned} \beta &= a, \\ \alpha &= \frac{1}{a}. \end{aligned} \tag{3.7}$$

Then, $p(\lambda) = \frac{1}{a^{\frac{1}{a}}\Gamma(\frac{1}{a})} \lambda^{\frac{1}{a} - 1} \exp(-\frac{\lambda}{a})$
and the probability mass function of O_i is given by

$$g(o_i) = \frac{\Gamma(o_i + \frac{1}{a})}{\Gamma(\frac{1}{a}) o_i!} \left[\frac{aE_i}{1 + aE_i}\right]^{O_i} \left[\frac{1}{1 + aE_i}\right]^{\frac{1}{a}}. \tag{3.8}$$

Simplifying a portion of the previous expression, we get:

$$\begin{aligned}
&\frac{\Gamma(o_i+\frac{1}{a})}{\Gamma(\frac{1}{a})} && (3.9)\\
&= \frac{(o_i+\frac{1}{a}-1)(o_i+\frac{1}{a}-2)(o_i+\frac{1}{a}-3).....(\frac{1}{a})\Gamma(\frac{1}{a})}{\Gamma(\frac{1}{a})} &&\\
&= (o_i+\frac{1}{a}-1)...\frac{1}{a} &&\\
&= \frac{1(1+a)(1+a).....(1+a(o_i-1))}{a^{o_i}}. && (3.10)
\end{aligned}$$

Applying this expression to our model(3.8) we arrive at

$$g(o_i) = \left[\prod_{j=0}^{o_i-1} \frac{1}{o_i!}(1+aj)\frac{1}{a^{o_i}}\right] \left[\frac{aE_i}{1+aE_i}\right]^{o_i} \left[\frac{1}{1+aE_i}\right]^{\frac{1}{a}}$$

Our likelihood function will be:

$$L = \prod_{i=1}^{N}\left[\prod_{j=0}^{o_i-1} \frac{1}{o_i!}(1+aj)\frac{1}{a^{o_i}} \left[\frac{aE_i}{1+aE_i}\right]^{o_i} \left[\frac{1}{1+aE_i}\right]^{\frac{1}{a}}\right], \qquad (3.11)$$

where L is always positive. Taking the natural logarithm of L:

$$\begin{aligned}
\ln L &= \sum_{i=1}^{N}\left[\sum_{j=1}^{o_i}\ln(1+aj) - \ln o_i! + o_i\ln E_i - o_i\ln(1+aE_i) - \frac{1}{a}\ln(1+aE_i)\right] \qquad (3.12)\\
&= \sum_{i=1}^{N}\left[\sum_{j=1}^{o_i}\ln(1+aj) - \ln o_i! + o_i\ln E_i - (o_i+\frac{1}{a})\ln(1+aE_i)\right].
\end{aligned}$$

Our goal is to estimate the variance a and find its confidence interval.

Differentiate this expression with respect to a :

$$\frac{d\ln L}{da} = \sum_{i=1}^{N}\sum_{j=1}^{O_i}\frac{j}{1+aj} + \frac{1}{a^2}\left(\sum_{i=1}^{N}\ln(1+aE_i)\right) - \sum_{i=1}^{N}\frac{(O_i+\frac{1}{a})E_i}{1+aE_i}$$

Thus the MLE of a is given by the solution of the following non-linear equation

$$\sum_{i=1}^{N}\sum_{j=1}^{O_i}\frac{j}{1+aj}+\frac{1}{a^2}\left(\sum_{i=1}^{N}\ln(1+aE_i)\right)-\sum_{i=1}^{N}\frac{(O_i+\frac{1}{a})E_i}{1+aE_i}=0. \tag{3.13}$$

Applications of the Newton-Raphson algorithm. Example of hepatitis in Berlin (revisited).

Coming back to the hepatitis data presented in the previous chapter, we assume the negative binomial model defined above. We have an equation,

$$\sum_{i=1}^{N}\sum_{j=1}^{O_i}\frac{j}{1+aj}+\frac{1}{a^2}\left(\sum_{i=1}^{N}\ln(1+aE_i)\right)-\sum_{i=1}^{N}\frac{(O_i+\frac{1}{a})E_i}{1+aE_i}=0.$$

Applying the $O_i's$ and $E_i's$ from the hepatitis example, we can estimate parameter a. Since our function is differentiable and smooth, we can use the Newton-Raphson method.

Since the calculations become computationally complex, the best way to overcome this problem is to write an S-PLUS routine using the Newton-Raphson method as a way to estimate the parameter a. This code is presented in the appendix. In order to have a starting value for a, we use the result that we had for the nonparametric case. The basic formulas to be used in the algorithm are presented below:

$$f(a)=\sum_{i=1}^{N}\sum_{j=1}^{O_i}\frac{j}{1+aj}+\frac{1}{a^2}\left(\sum_{i=1}^{N}\ln(1+aE_i)\right)-\sum_{i=1}^{N}\frac{(O_i+\frac{1}{a})E_i}{1+aE_i} \tag{3.14}$$

$$f'(a)=-\sum_{i=1}^{N}\sum_{j=1}^{O_i}\left(\frac{j}{1+aj}\right)^2-\frac{2}{a^3}\left(\sum_{i=1}^{N}\ln(1+aE_i)\right)+\frac{2}{a^2}\sum_{i=1}^{N}\frac{E_i}{1+aE_i}+\sum_{i=1}^{N}\frac{E_i^2(O_i+\frac{1}{a})}{(1+aE_i)^2} \tag{3.15}$$

Beginning with the setup of the algorithm, initial value of a, setup of the iteration counter and setting the tolerance of the algorithm, we consider the algorithm to have converged if the (k+1)th iteration result is equal to the kth iteration up to the 5th position after the decimal point. To begin, we will set the maximum number of iterations to 45.

After three iterations we have reached convergence at the desired level, so $\hat{a}_{MLE}$ =0.483947179095742

Information Matrix and Parameter Variances

In order to obtain the variance of $\hat{a}$, we compute

$$\frac{d^2\ln L}{da^2}=-\sum_{i=1}^{N}\sum_{j=1}^{O_i}\left(\frac{j}{1+aj}\right)^2-\frac{2}{a^3}\left(\sum_{i=1}^{N}\ln(1+aE_i)\right)+\frac{2}{a^2}\sum_{i=1}^{N}\frac{E_i}{1+aE_i}+\sum_{i=1}^{N}\frac{E_i^2(O_i+\frac{1}{a})}{(1+aE_i)^2}.$$

So that $Var(\hat{a})$ is given by

$$Var(\hat{a}) = \frac{1}{-E\left[\frac{d^2 \ln L}{da^2}\right]}. \tag{3.16}$$

The S-PLUS code for this matrix is presented in the appendix.

Thus an asymptotic confidence intervals for the variance of λ, is given by:

$$\hat{a} \pm z_{\frac{\alpha}{2}} \sqrt{var(\hat{a})}. \tag{3.17}$$

For the case of a confidence level $\alpha = 0.05$, we have $z_{\frac{\alpha}{2}} = 1.96$. Then the expression for the confidence interval becomes $\hat{a} \pm 1.96\sqrt{var(\hat{a})}$.

Hepatitis data revisited

Using the nonparametric estimate of the parameter a =0.5476439 as a starting value, we are getting the following output:

Iteration 1	0.483072211758264
Iteration 2	0.483944644900506
Iteration 3	0.483947179095742

Let us return to the hepatitis data presented before. We apply the $O_i's$ and $E_i's$ from the data, so that we can estimate the parameter, a. Since our function is differentiable and smooth, we can use the Newton-Raphson method.

Comparing parametric and nonparametric results: the final results for the given data set are presented in the table below

$\hat{a}_{pooled}$	0.5476439
$\hat{a}_{simple}$	0.5488984
$\hat{a}_{pooled.corrected}$	0.5437004
$\hat{a}_{mixture \text{ model}}$	0.483947179095742
$Var(\hat{a})$	0.02589858

The numerical values of the estimates look relatively close, but the nonparametric procedures overestimate heterogeneity.

CONFIDENCE INTERVAL

confidence interval for a=(0.1685236 0.7993707)
length of the confidence interval for a= 0.6308471

Chapter 4

SIMULATION STUDIES

In order to evaluate performance of the estimated confidence intervals, we have to simulate the data with knowledge of the mean and variance of λ in advance. Then, we execute our program and see how frequently the estimated confidence interval contains the true value of the parameter. We vary three aspects of our data set:

1. Distribution of the expected values, the $E_i^{'}$s.
2. Sample size.
3. The value of the variance parameter a.

4.1 Data generation

In order to generate observed values (negative binomial random observations) in S-Plus, we do the following:

- create the set of $E_i^{'}$s, say N times, as $E_i \sim Poisson(\theta)$

- generate K samples each of size N

- assume that the above are the given values of $E_i^{'}$s

- calculate the $p_i = \left(\frac{1}{1+\beta E_i}\right)$ (the number of $p_i^{'}$s equal the number of $E_i^{'}$s).

We have K sets of $E_i^{'}$s, so that this construction leads us to an NxK matrix of $E_i^{'}$s. This matrix has sets of $E_i^{'}$s as its columns.

- get K sets of $O_i^{'}$s created as $O_i \backsim NegativeBin(n, p_i)$

- apply our algorithm (Newton -Raphson, information matrix, confidence intervals) to each set of $E_i^{'}$s and corresponding $O_i^{'}$s. Thus get K confidence intervals.

- check whether the confidence interval covers the true parameter.

We want to generate K=5000 samples of

sample size N	10	20	30	40	50	100

variance a	0.167	0.25	0.5	1

Keeping in mind that $a = \alpha\beta$ and $\alpha > 0$ and α has to be an integer, since we have an expression of $\Gamma(\alpha)$ in the distribution function of our model.

4.2 Results

In order to evaluate the performance of our estimation, it is reasonable to measure two aspects of the confidence intervals:

I. Coverage probability

II. Length of the confidence interval

The goal is to get a small confidence interval with coverage probability.equal to the nominal value.

Now, let us come to the first aspect.

In our attempt to simulate the real data, for each of our experimental trials, generate 5000 samples of $E_i^{'}$s and corresponding $O_i^{'}$s respectively.

- Start with a relatively small sample size, say 10.

- Run our program (parameter estimation, confidence interval, etc.) on each pair of $E_i^{'}$s and $O_i^{'}$s.

- Take $E_i^{'}$s$\sim$ $Poisson(\theta)$, where $\theta = 10, 20, 30, 40, 50$.

- x-axis of each of the graphs represents the variance.

- y-axis shows the scale of the coverage probability.
- for a better visual evaluation weather the coverage probability of a given parameter estimation is acceptable, draw a line y=0.95, which represents the level of coverage probability equal to 95%.
- repeat this experiment using a larger sample size, say 20,30,40,50 and 100.
- After all experiments were done, plot the values of coverage probability versus the set of the values of a, the variance of heterogeneity parameter.

Let us call the connected values of the coverage probabilities *a coverage probability graph.*

- Put the coverage probability graphs of different sample sizes and the same distribution of the $E_i^{'}$s together.
- The values on the graph lines represent the sample size.

The goal is to check if the initial distribution of $E_i^{'}$s makes a significant impact on the outcome of the experiment (see the figures presented below).

Observations:

For each of the coverage probability graphs, we could observe:

1. Coverage probability tends to increase after the starting point, $a = 0.167$

2. After the coverage probability line reaches its maximum, it either stays constant or slightly goes down.

3. The coverage probability graph of the sample size 10 could take the values below the 95% line. The final result shows that the coverage probability of the sample size 100 has the best performance.

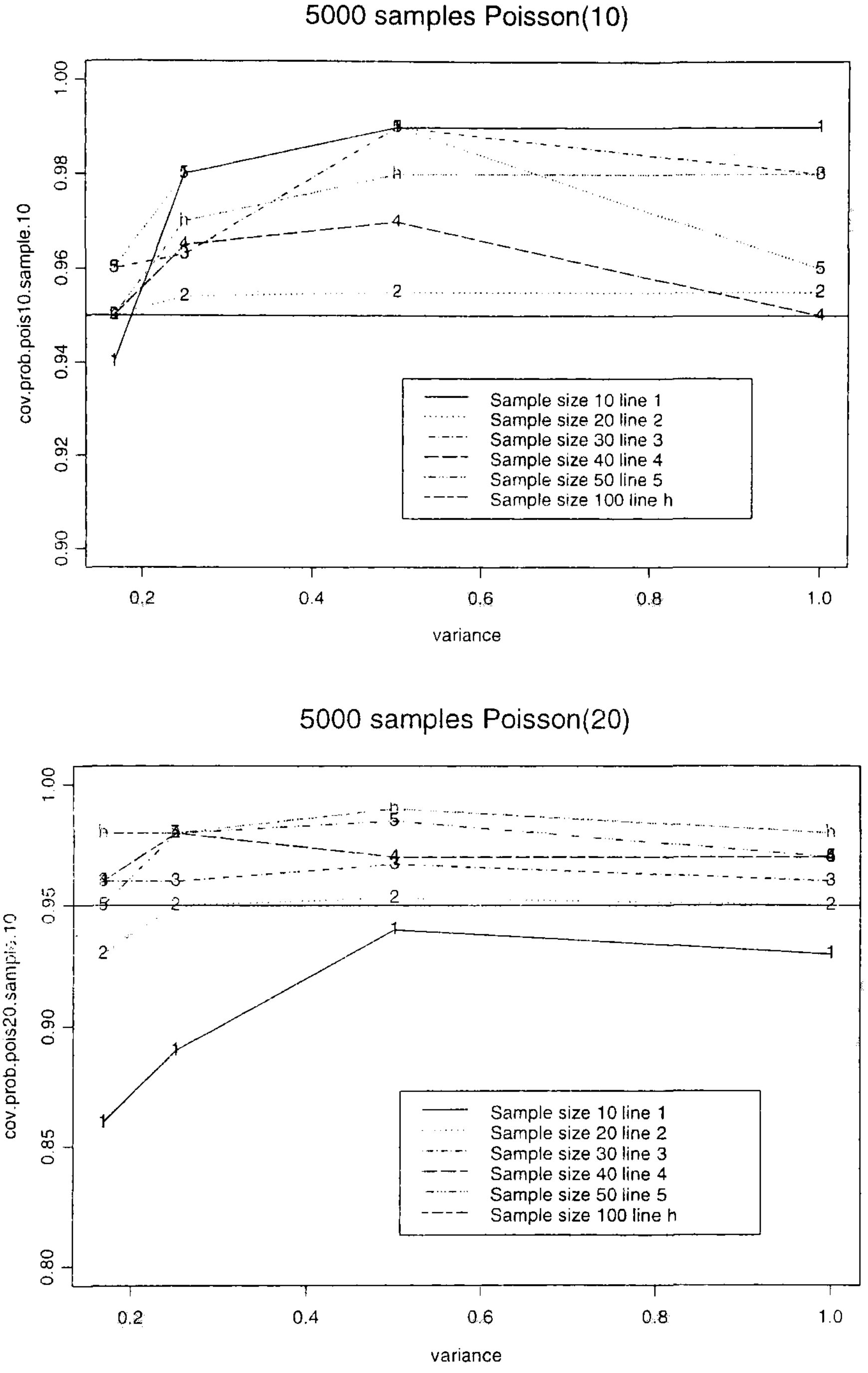
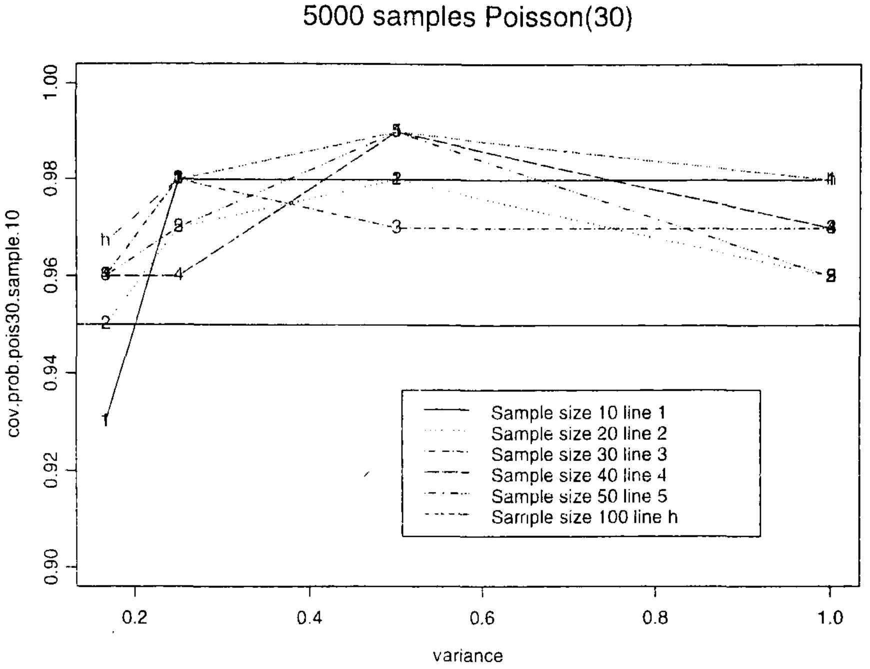


Figure 4-1: A collection of the coverage probability figures corresponding to the fixed distribution of the expected values (Poisson(10) and Poisson(20))



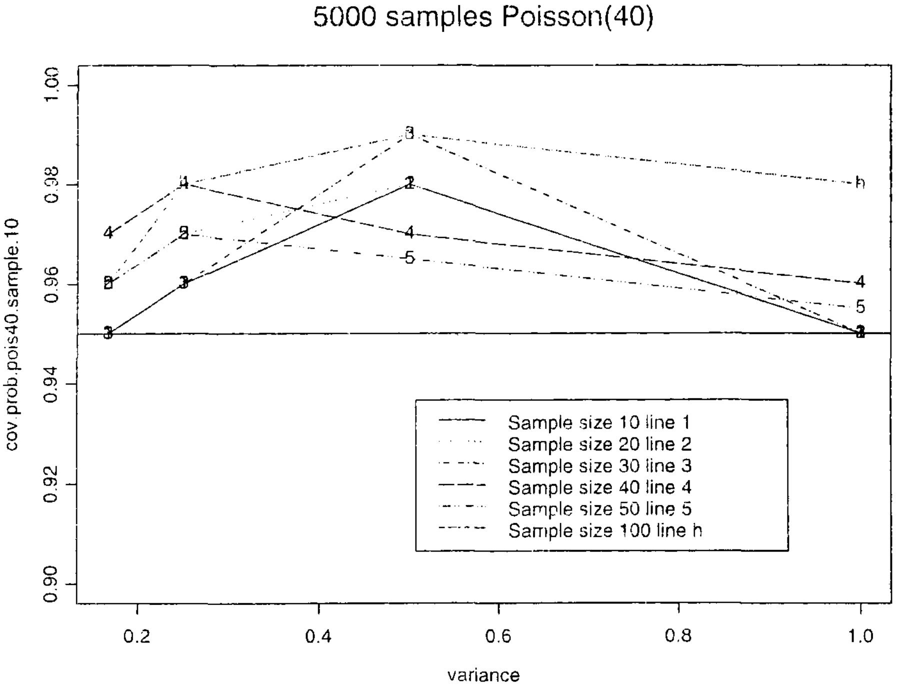


Figure 4-2: A collection of the coverage probability figures corresponding to the fixed distribution of the expected values (Poisson(30) and Poisson(40))

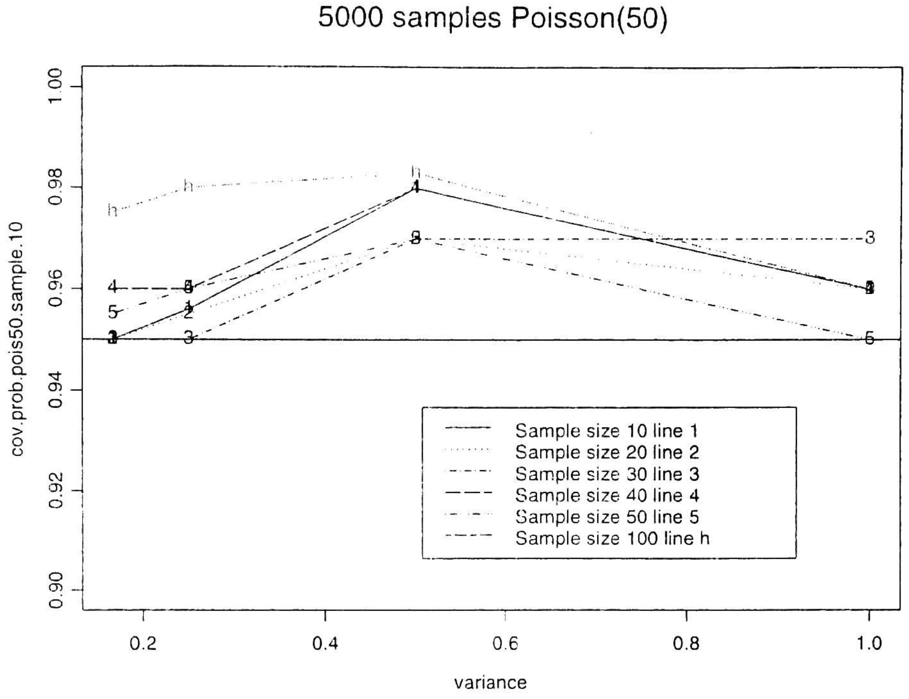


Figure 4-3: A collection of the coverage probability figures corresponding to the fixed distribution of the expected values (Poisson(50))

Conclusion:

1. The graphs do have an analogous shape independent of the distribution of the expected values. That would imply that the distribution of $E_i^{'}$s does not have an influence the outcome.

2. The values of the coverage probability improve as the sample size is becoming larger.

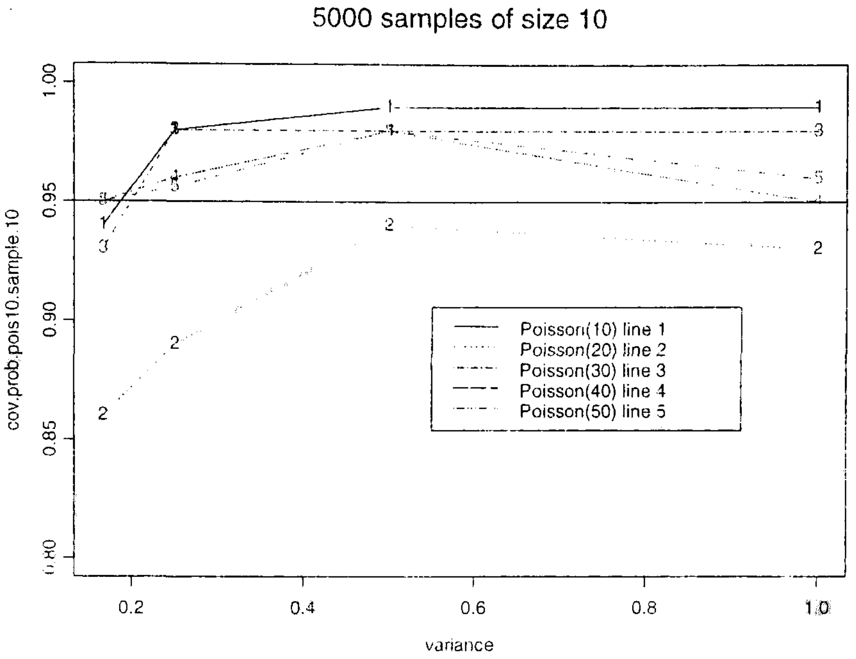
3. The comparison between the set of graphs with distinct value of the parameter θ shows that the coverage probability becomes higher as the value of θ is growing.

4. In most cases, the coverage probability is above the nominal value 0.95.

Now, let us rearrange our results in the following way:

- overlay the graphs according to their sample size, i.e. put the graphs with different distributions of expected values of the same sample size together.
- The values on the graph lines represent the distribution parameter of $E_i^{'}$s.

The goal is to check the effect of the sample size on the outcome of the experiment (see the figures below)



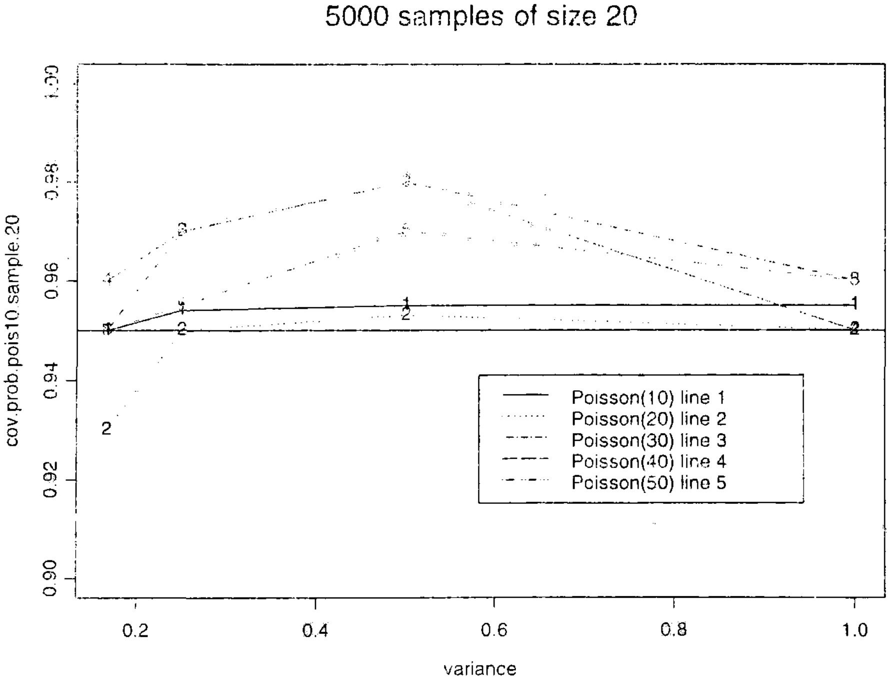


Figure 4-4: A collection of the coverage probability figures corresponding to the fixed sample size (sample sizes 10 and 20)

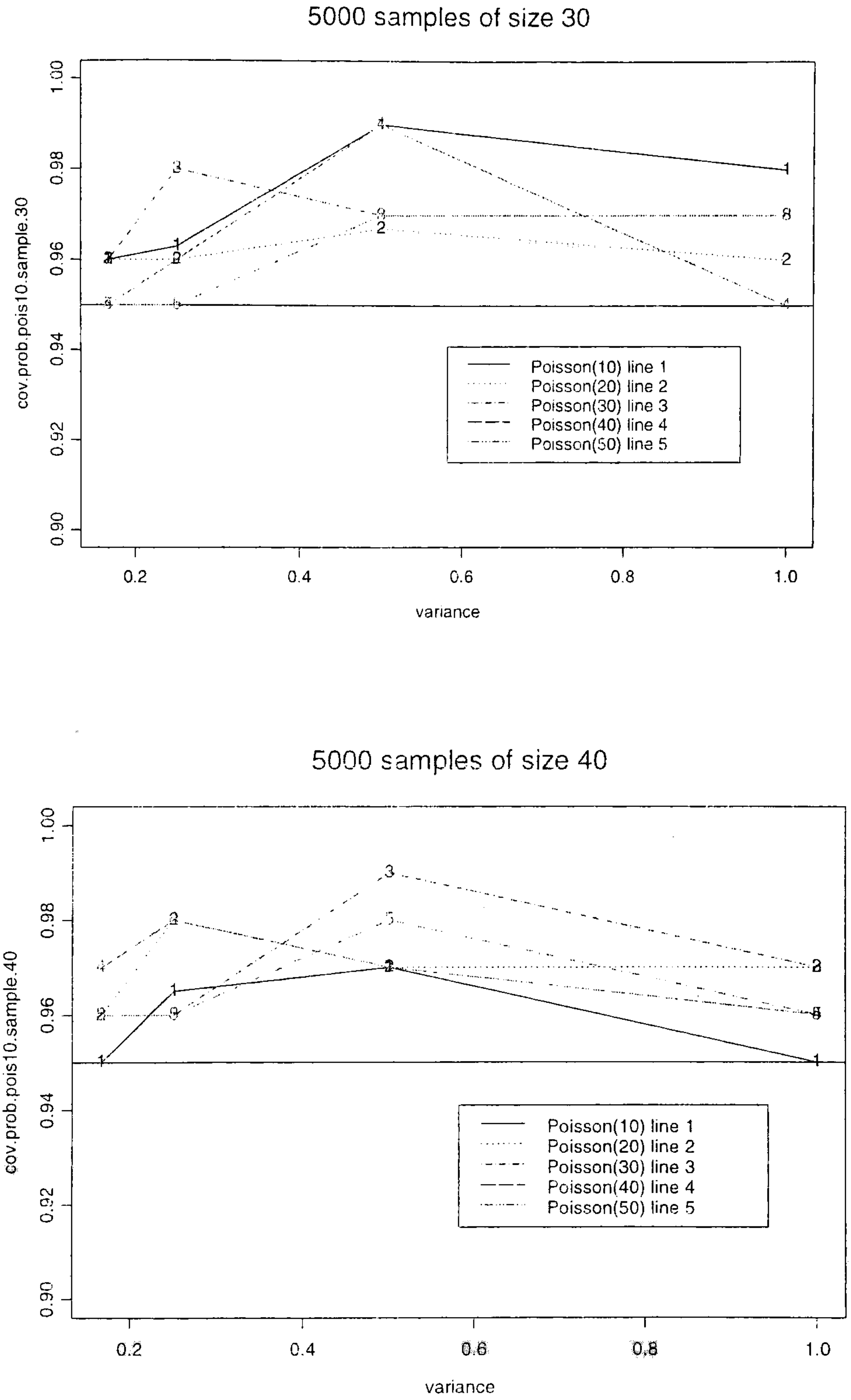


Figure 4-5: A collection of the coverage probability figures corresponding to the fixed sample size (sample sizes 30 and 40)

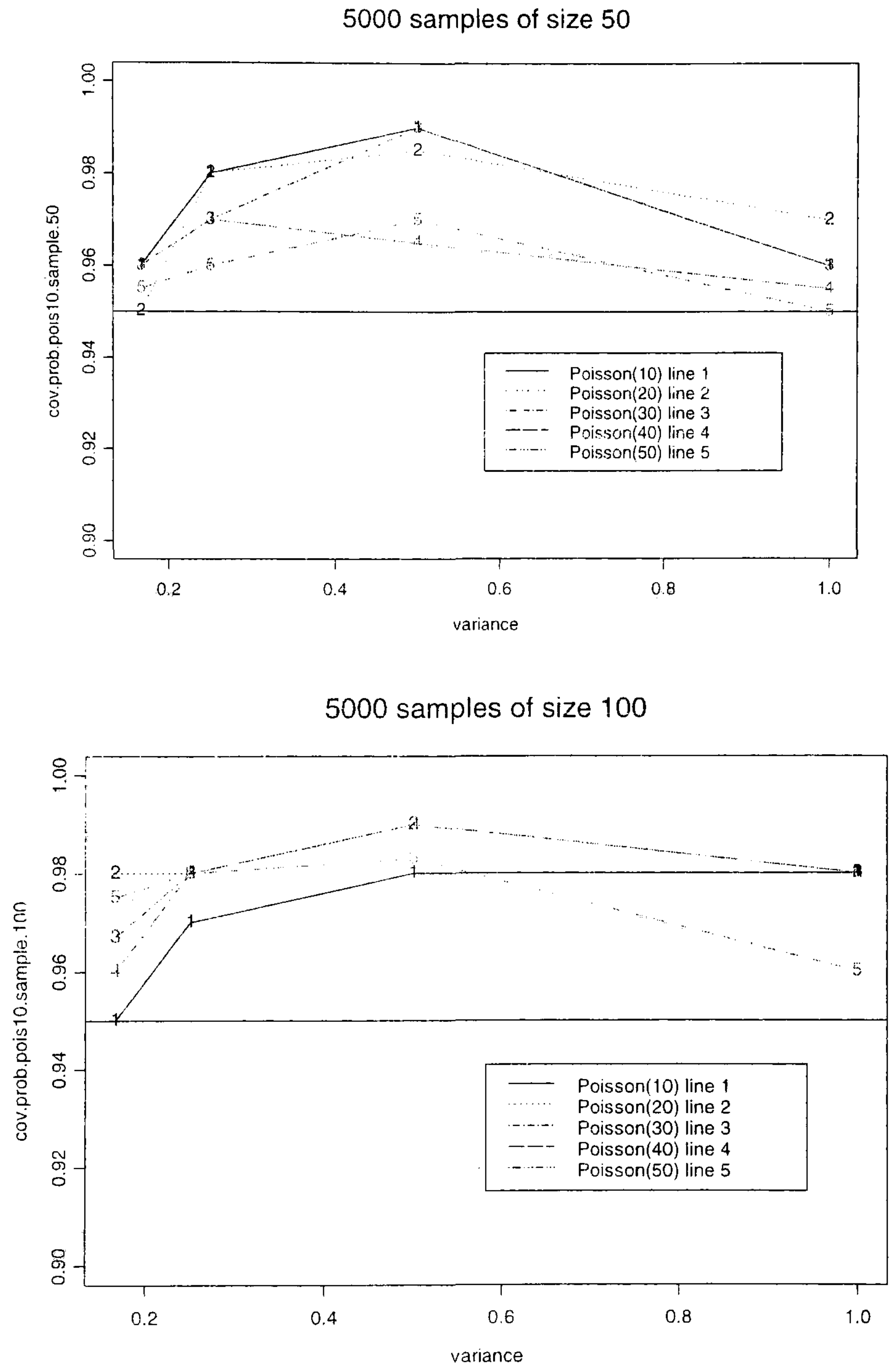


Figure 4-6: A collection of the coverage probability figures corresponding to the fixed sample size (sample sizes 50 and 100)

Observations:

1. The coverage probability graphs of the samples with distinct distributions of $E_i^{'}$s are becoming closer as the sample size is growing.

2. The smallest spread between the coverage probability graphs could be found in case of the sample size 100, and the largest spread is in case of the sample size equal to 10.

Conclusion:

The coverage probability could be affected by the sample size, but the choice of the distribution of $E_i^{'}$s does not have a major influence on the final outcome.

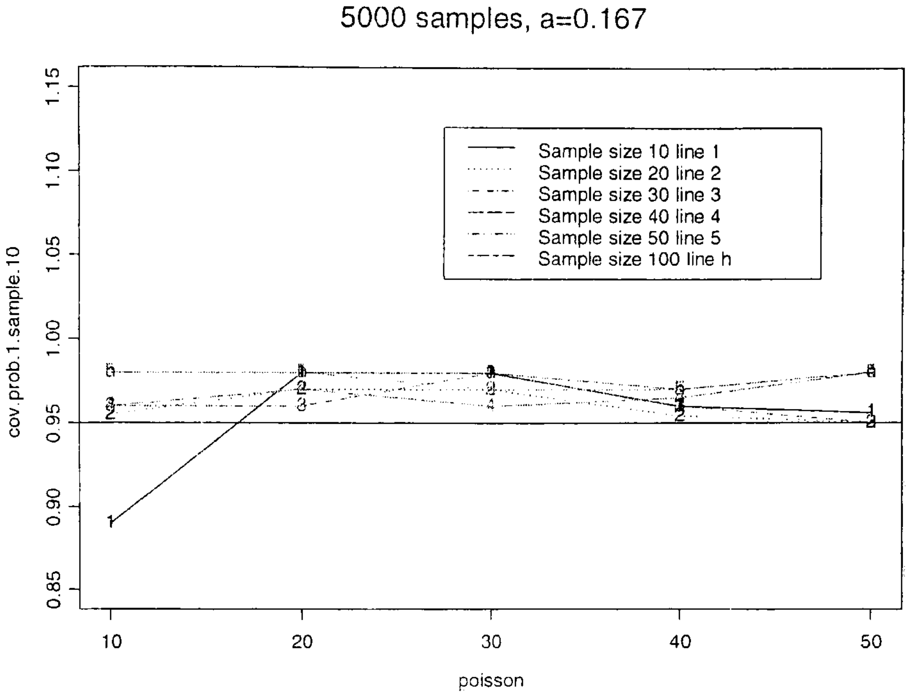
The coverage probability for Poisson with mean 20 is below the nominal value of 0.95 for the small sample size and a small value of a.However, as the value of a increases, and the sample size increases, the coverage probability is above the nominal value.

In most cases, the coverage probability is maximum for the values of a between 0.4 and 0.6.

Now, let us have a look at our result from a slightly different angle. Let us take the same set of simulation results and rearrange it in the following way:

- x-axis represent the values of the parameter θ
- y-axis represent the coverage level
- Graph a separate figure for each value of the parameter a.
- In this case, the numbers on the coverage probability graphs represent the sample size.

The goal is to check the effect of the variance of the spatial heterogeneity parameter.



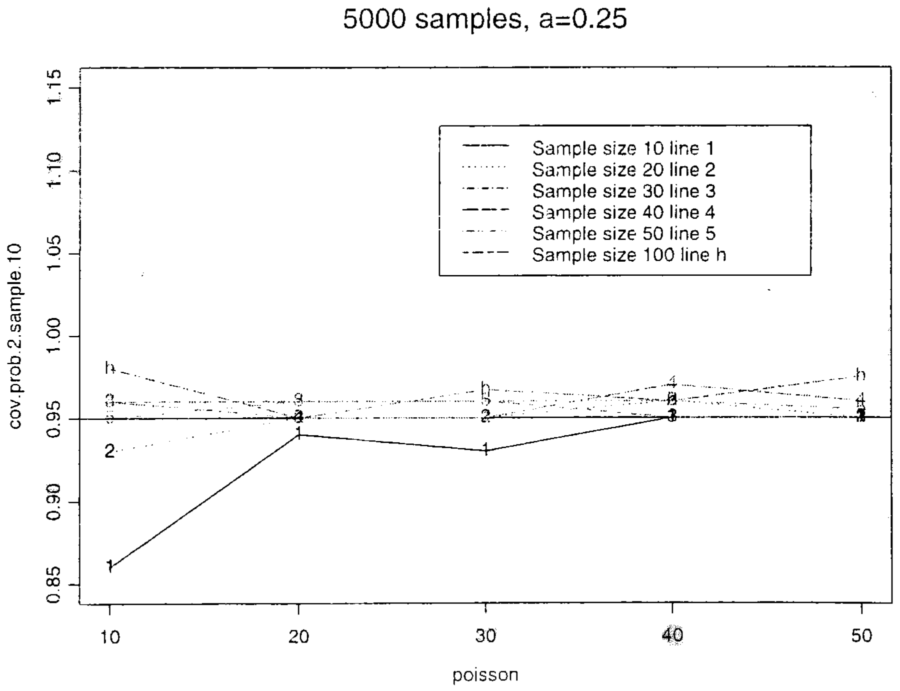


Figure 4-7: A collection of the coverage probability figures corresponding to the fixed variance (a=0.167 and a=0.25)

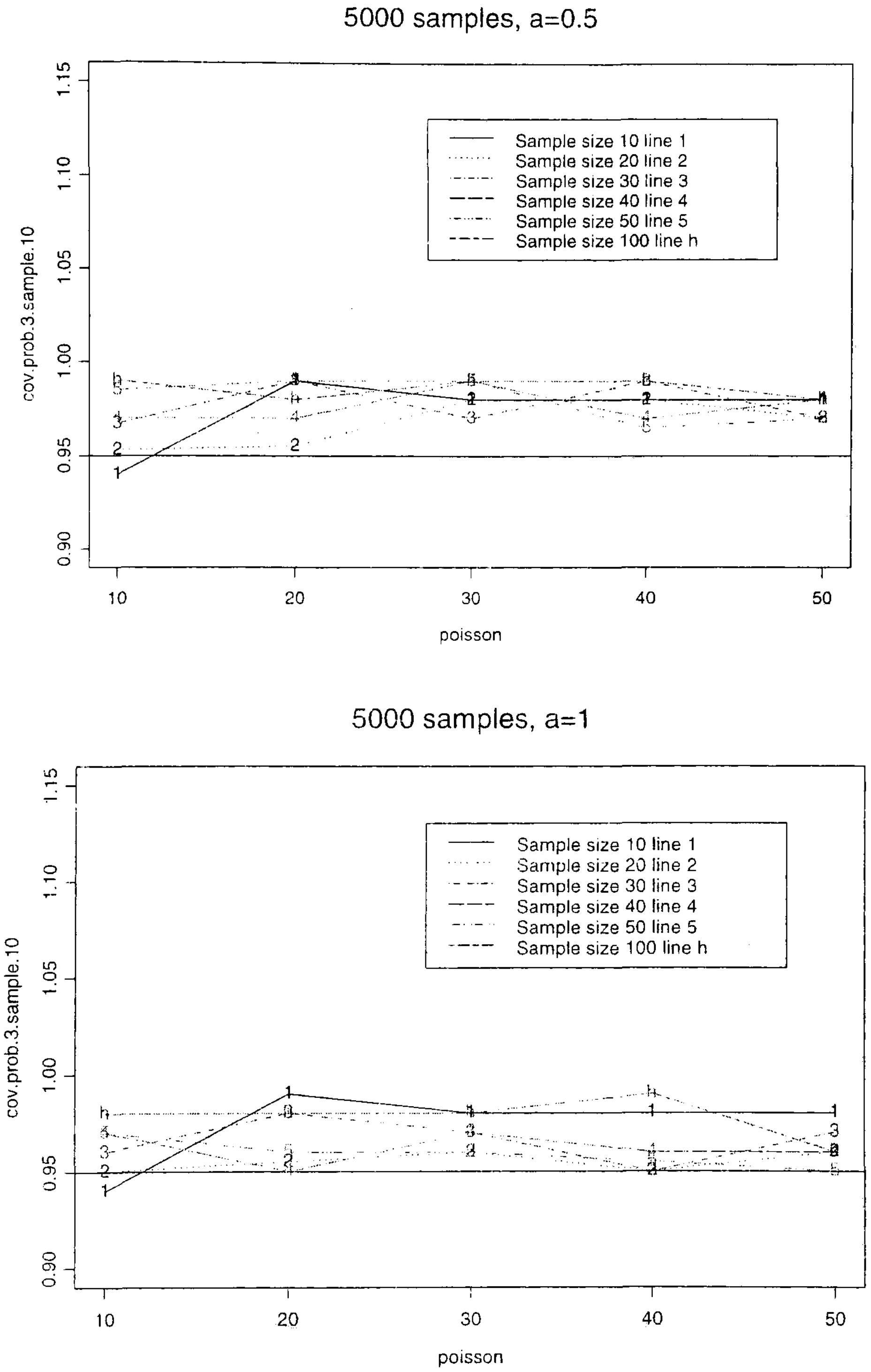


Figure 4-8: A collection of the coverage probability figures corresponding to the fixed variance (a=0.5 and a=1)

Observations:

1. The spread between the coverage probability graphs gets smaller as the value of the variance parameter grows.

2. From the last set of figures, we could observe that the sets of size 10 have the lowest coverage probability and the sets of the sample size 100 are the most acceptable in this aspect.

Conclusion:

The influence of the sample size on the outcome becomes smaller as the value of a is increases. Therefore, we may conclude that the value of the variance of the parameter of the variance of the spatial heterogeneity has a significant impact on the final outcome.

RESULT:

The analysis of coverage probability graphs shows that the sample size and the variance of the parameter of the spatial heterogeneity are significant for the outcome of the experiment. The choice of the distribution of the expected values does not demonstrate such a big impact on the outcome.

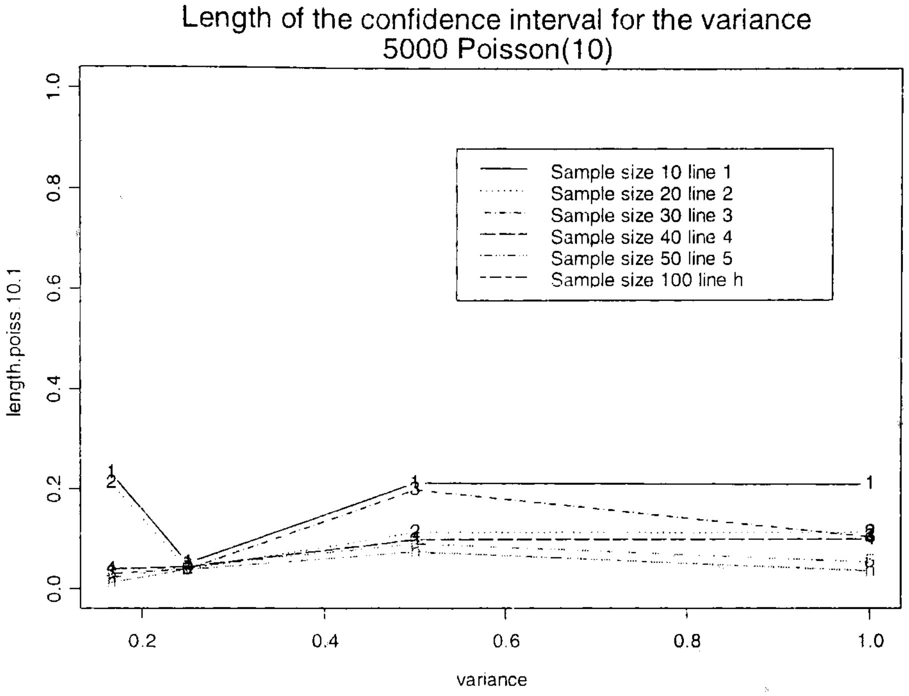
For the part II, we do a graphical analysis of the length of the estimated confidence intervals.

- Plot the length of the confidence intervals versus fixed values of

$a = 0.167$, 0.25, 0.5 and 1.

- Overlay graphs with equal distribution of $E_i^{'}$s

- The values on the graph lines represent the sample size



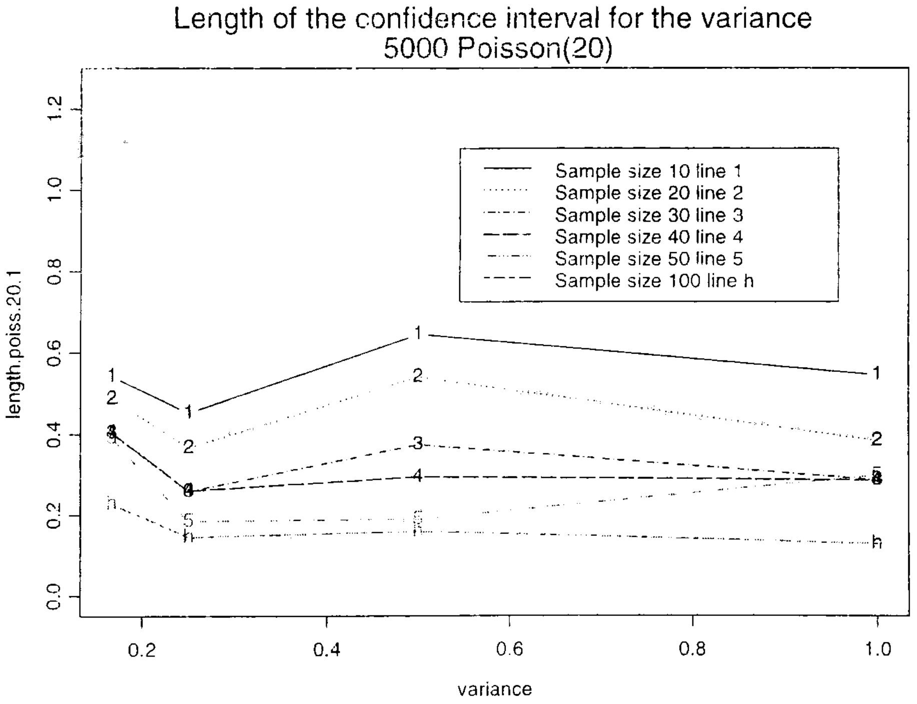
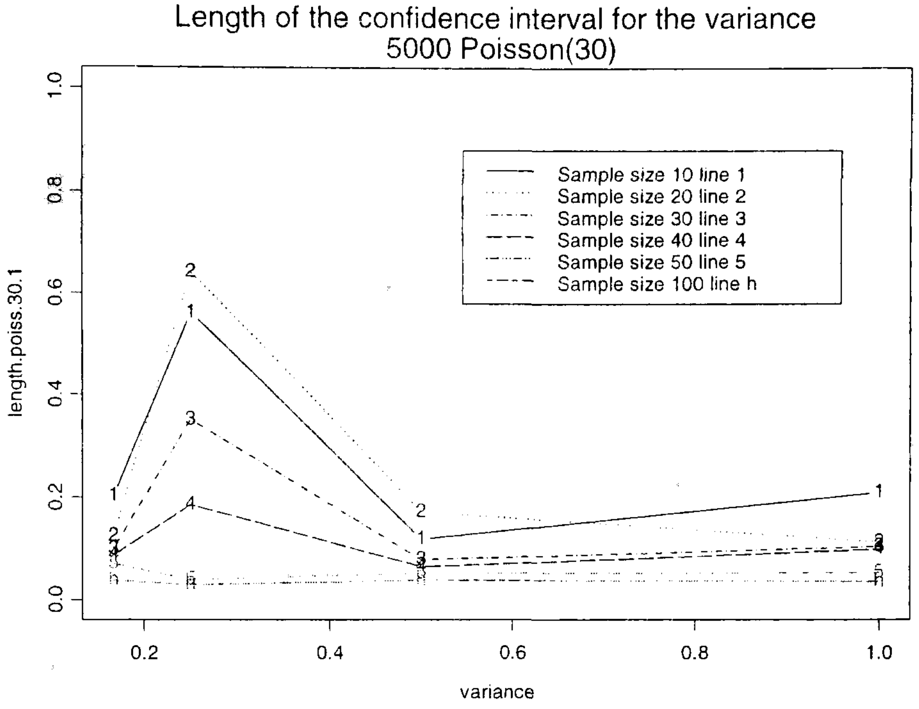


Figure 4-9: A collection of the confidence interval length figures corresponding to the fixed distribution of the expected values (Poisson(10) and Poisson(20))



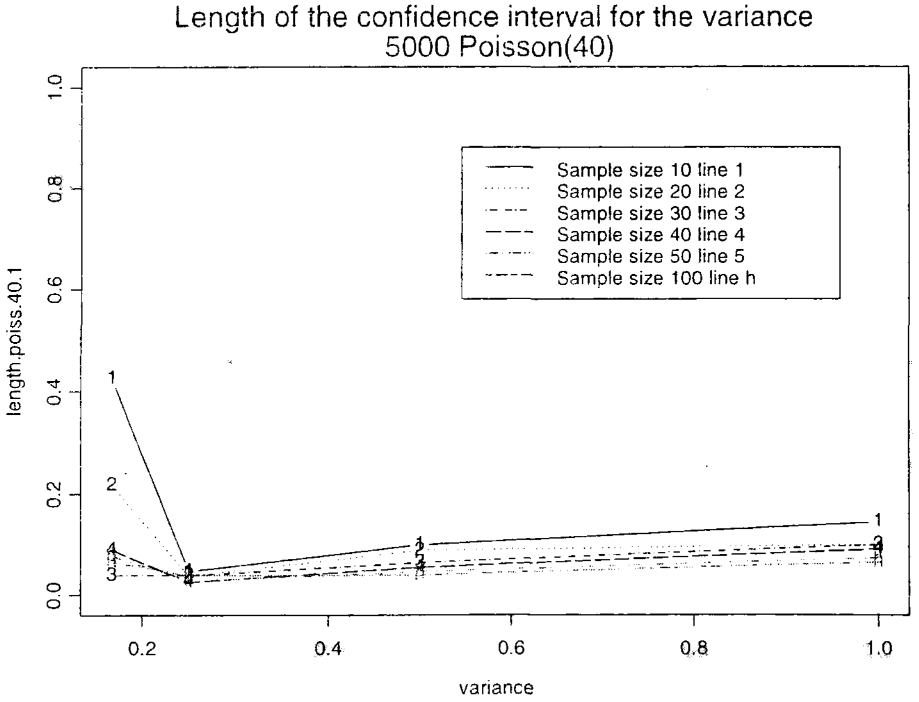


Figure 4-10: A collection of the confidence interval length figures corresponding to the fixed distribution of the expected values (Poisson(30) and Poisson(40))

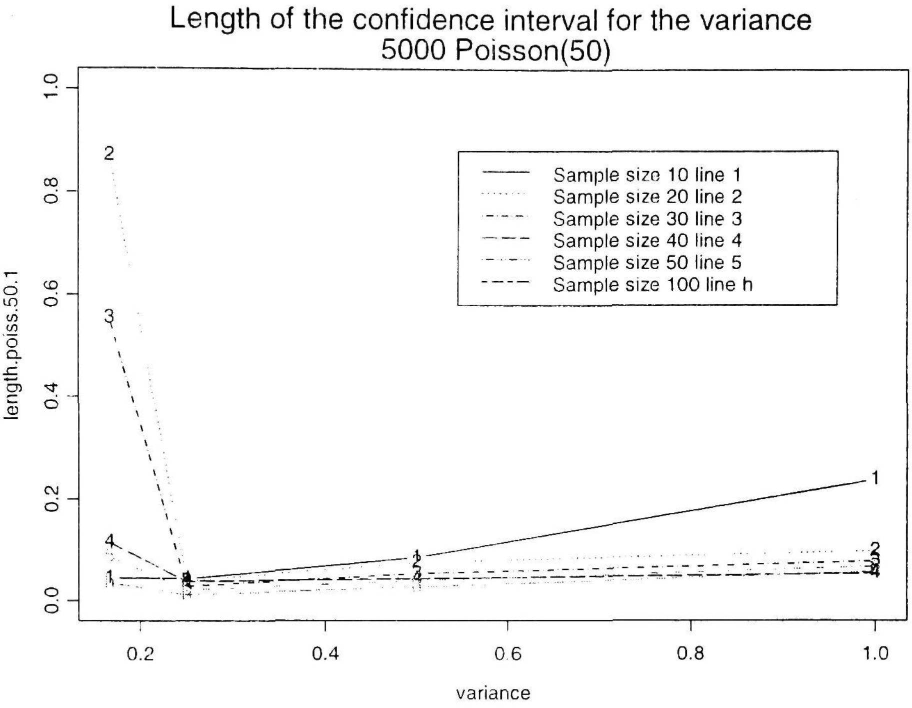


Figure 4-11: A collection of the confidence interval length figures corresponding to the fixed distribution of the expected values (Poisson(50))

Observation:

The graphs of the length of the confidence intervals have the same pattern independent from the value of the parameter θ.

Conclusion:

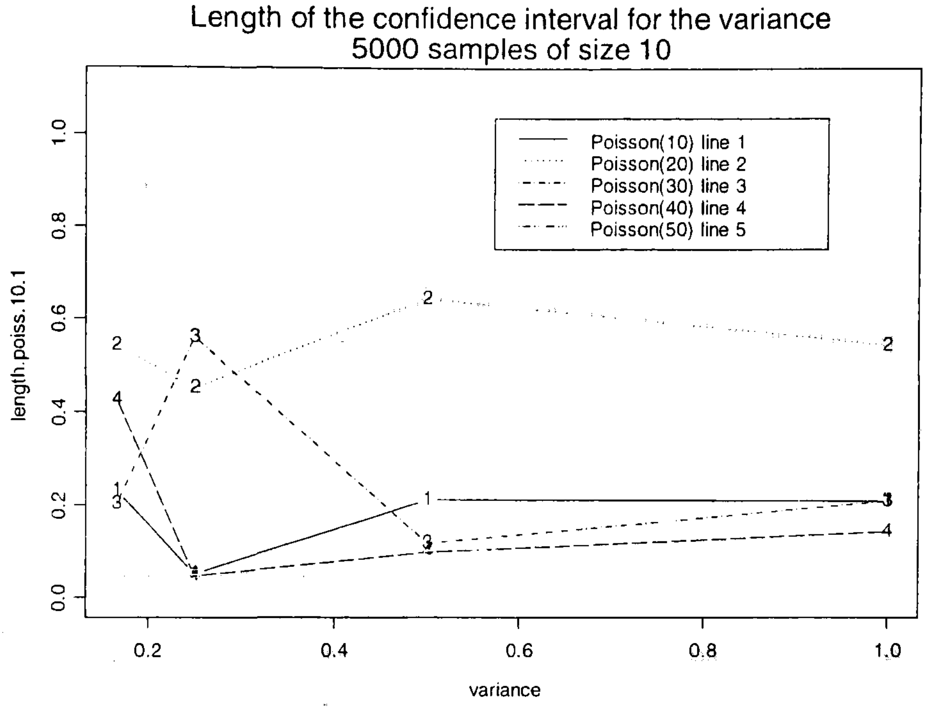
Examining the set of the graphs presented above, we could say that there is no influence of the distribution of $E_i^{'}$s.

The length of the confidence interval is quite large for the the samples of size 10 and 20 and for small values of a. As the sample size increases, the confidence interval becomes narrower. For the Poisson with mean 50 (sample size $\geqslant$ 30), the length of the confidence interval converge to the same numerical value.

Now, let us rearrange our graphs:

- Plot the lines of the same sample size together, keeping the initial distribution of the observed values distinct.

- Keep the labeling of the parameter of the distribution of $E_i^{'}$s on the confidence interval length lines.



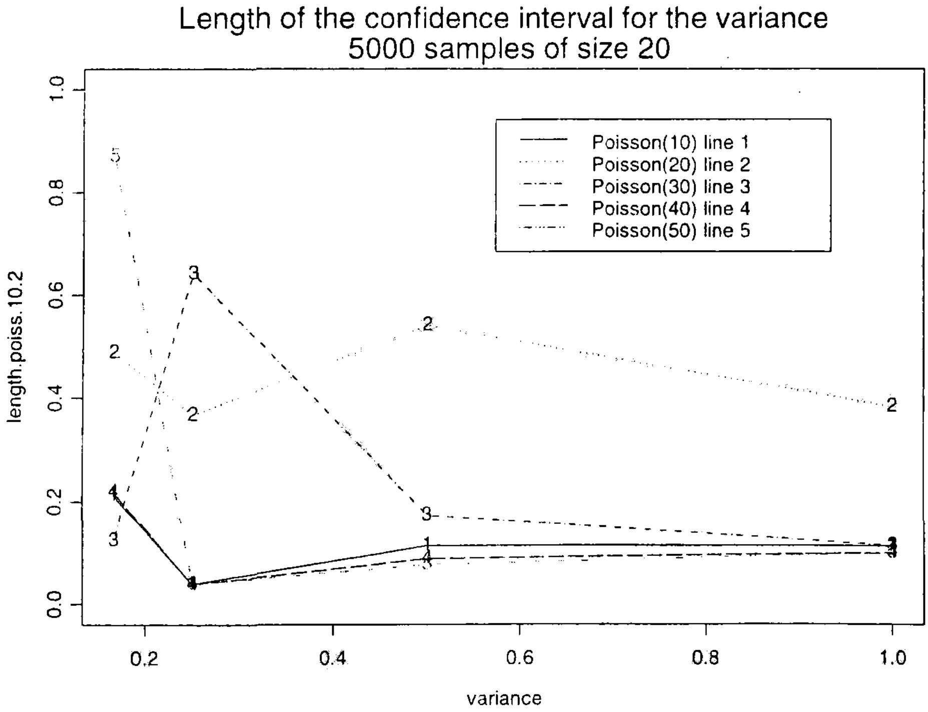
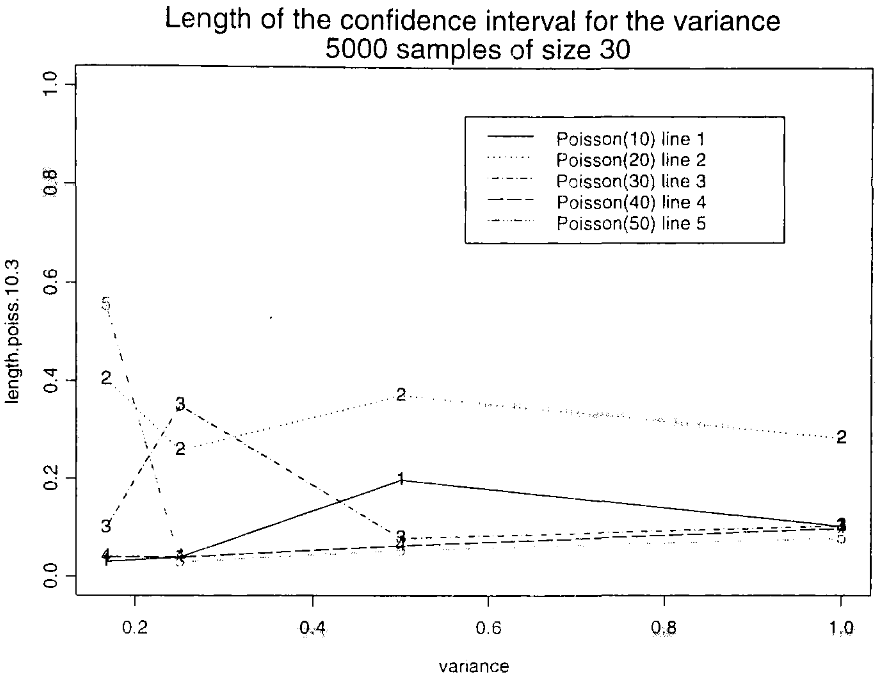


Figure 4-12: A collection of the confidence interval length figures corresponding to the fixed sample size (sample sizes 10 and 20)



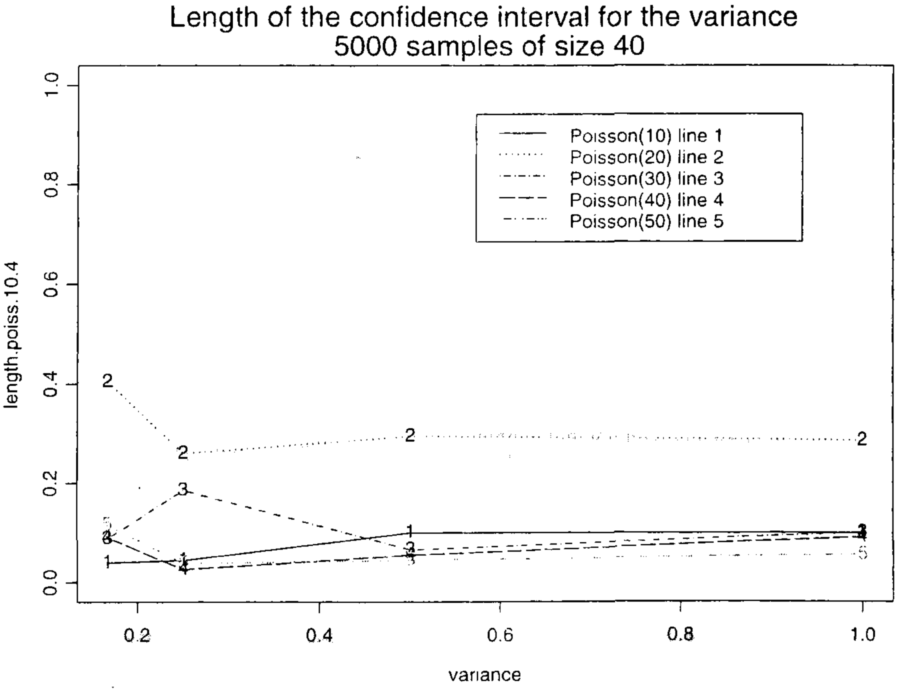
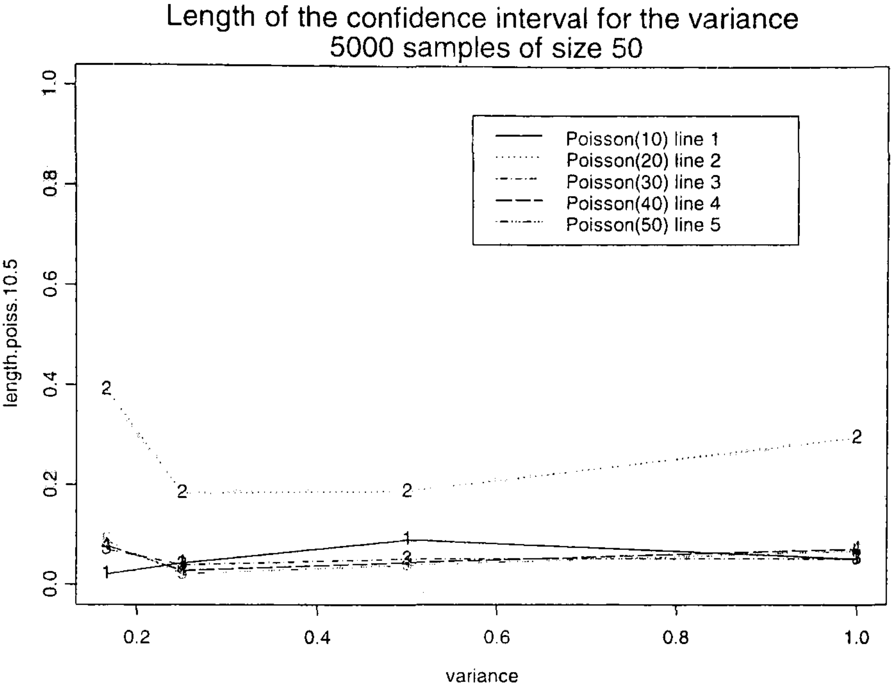


Figure 4-13: A collection of the confidence interval length figures corresponding to the fixed sample size (sample sizes 30 and 40)



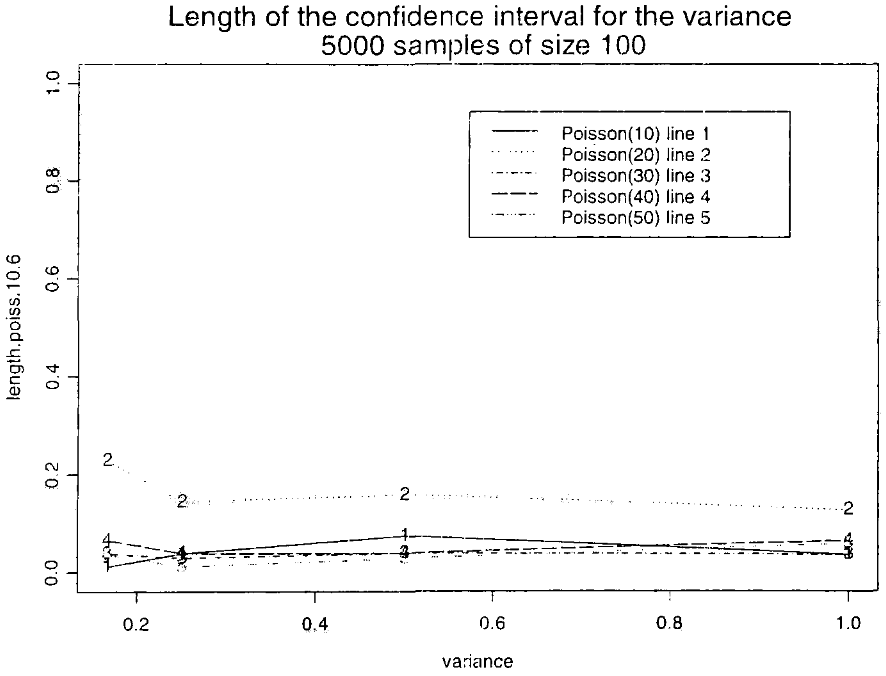


Figure 4-14: A collection of the confidence interval length figures corresponding to the fixed sample size (sample sizes 50 and 100)

Observation:

The figures of distinct distributions of E_i's are becoming closer as the sample size is growing

Conclusion:

1. The sample size makes a significant impact on the result of our experiment.

2. The distribution of E_i's does not play the most important role.

For sample size 20 and 30, the lengths of the confidence intervals are quite large. The least value of the length is obtained when sample size is larger or equal to 50 and the value of a is close to 1.

4.3 Final remarks and recommendations

We have proposed an alternative method of the estimation of spatial heterogeneity which occurs in estimation of the SMR.

The essence of the method:

1. Randomization of parameter λ

2. Assumption: λ has a unit mean and variance a, $a \in R_+$

3. The resulting distribution resembles the modified form of the negative binomial distribution.

4. Estimation of the heterogeneity parameter a using maximum likelihood approach.

5. Derivation of asymptotic confidence intervals for the estimated parameter.

Conclusion:

As a result of the simulation studies, we could conclude

1.The coverage probability becomes higher as the sample size and the Poisson mean are increasing

2. The value of variance a also has an effect on the final outcome, i.e. the coverage probability grows as the value of a becomes larger.

3. In case of the large sample size, the value of the mean of the Poisson distribution and the value of a are not significant for the final outcome.

4. The length of the confidence intervals becomes smaller as the sample size grows.

5. The algorithm is very sensitive to the initial parameter estimation (in this case Newton-Raphson or analogous procedures).

Due to the fact that the variance has a relatively small numerical value, the confidence interval becomes very short and a slight miscalculation in parameter estimation could lead us to a low rate of coverage probability.

Bibliography

[1] *Boehning, D., Sarol, J.* (2000) A Nonparametric Estimator of Heterogeneity Variance with Applications to SMR- and Proportion Data. *Biometrical Journal* 42 , (2000) 3, 321-334

[2] *Boehning, D.* (2003) Empirical Bayes estimators and nonparametric mixture models for space and time-space disease mapping and surveillance. *Environmetrics* 14, (2003): 431-451

Appendix A

S-PLUS CODES

Newton Raphson algorithm

```
x<-c(29,26,54,30,16,15,6,35,17,7,43,17,15,11,11,2,2,9,2,3,11,5,2)
y<-c(10.7121, 17.9929,18.1699, 19.211, 21.9611, 14.6268, 9.622, 17.2671, 18.823, 18.2705, 32.1823, 24.5929, 8.3968, 15.6438, 11.8289, 9.9513, 10.8313, 18.3403, 5.1758, 10.9543, 20.0121, 13.8389, 12.7996)
a<-0.5
it<-0
f.over.derivative<-1
while(abs(f.over.derivative)>0.0001 && (it<-it+1)<100){
{
{
{bottom.calc<-function(x,y){
    bottom<-matrix(0,length(x),1)
        for (i in 1:length(x)){
        for (j in 1:(x[i]-1)){
                        bottom[i]<-bottom[i]+(j/(1+a*j))
}
                    }
bottom}
sum(bottom.calc(x,y))}
{
    one.plus.a.ei<-(1+a*y)
    lgone.plus.a.ei<-log(one.plus.a.ei)
    summation<-sum(lgone.plus.a.ei)
    result<-((1/a^2)*(summation))
    result
final<-( sum(bottom.calc(x,y))+result)
}
sum(bottom.calc(x,y))
result
{top<-(x+1/a)*y
    denominator<-1+a*y
    ratio<-top/denominator
    sum.ratio<-sum(ratio)
}
all.together<-final-sum.ratio
all.together
{
{new.bottom.calc<-function(x,y){
    bottom<-matrix(0,length(x),1)
        for (i in 1:length(x)){
            for (j in 1:(x[i]-1)){
            bottom[i]<-bottom[i]+((-1)*((j^2)/((1+a*j)*(1+a*j))))
}
                    }
bottom
}
sum(new.bottom.calc(x,y))}
}
new.one.plus.a.ei<-(1+a*y)
    new.lgone.plus.a.ei<-log(one.plus.a.ei)
    new.summation<-sum(new.lgone.plus.a.ei)
    new.result<-((2/a^3)*(new.summation))
```

```
     new.result
  new.final<-( sum(new.bottom.calc(x,y))-new.result)
  {new.top<-(x+1/a)*y^2
     new.denominator<-(1+a*y)^2
     new.ratio<-new.top/new.denominator
     new.sum.ratio<-sum(new.ratio)
  }
  {
     third.term<-y/(1+a*y)
     third.term<-sum(third.term)
     third.term<-2/a^2*third.term

     third.term
  }
  new.all.together<-new.final+third.term+new.sum.ratio
  new.all.together
  sum(new.bottom.calc(x,y))
  new.result
  new.final
  f.over.derivative<-all.together/new.all.together
  f.over.derivative }
  a<-(a-f.over.derivative)
  cat(it, a,"\n")}}
```

Information matrix and invertted information matrix

```
  x<-c(29,26,54,30,16,15,6,35,17,7,43,17,15,11,11,2,2,9,2,3,11,5,2)
  y<-c(10.7121, 17.9929,18.1699, 19.211, 21.9611, 14.6268, 9.622, 17.2671, 18.823, 18.2705,
32.1823, 24.5929, 8.3968, 15.6438, 11.8289, 9.9513, 10.8313, 18.3403, 5.1758, 10.9543, 20.0121,
13.8389, 12.7996)
  a<-0.5
  it<-0
  f.over.derivative<-1
  while(abs(f.over.derivative)>0.0001 && (it<-it+1)<100){
```

```
{
{
{bottom.calc<-function(x,y){
    bottom<-matrix(0,length(x),1)
        for (i in 1:length(x)){
        for (j in 1:(x[i]-1)){
                        bottom[i]<-bottom[i]+(j/(1+a*j))
}
                    }
bottom}
sum(bottom.calc(x,y))}
{
    one.plus.a.ei<-(1+a*y)
    lgone.plus.a.ei<-log(one.plus.a.ei)
    summation<-sum(lgone.plus.a.ei)
    result<-((1/a^2)*(summation))
    result
final<-( sum(bottom.calc(x,y))+result)
}
sum(bottom.calc(x,y))
result
{top<-(x+1/a)*y
    denominator<-1+a*y
    ratio<-top/denominator
    sum.ratio<-sum(ratio)
}
all.together<-final-sum.ratio
all.together
{
{new.bottom.calc<-function(x,y){
    bottom<-matrix(0,length(x),1)
        for (i in 1:length(x)){
            for (j in 1:(x[i]-1)){
            bottom[i]<-bottom[i]+((-1)*((j^2)/((1+a*j)*(1+a*j))))
}
                    }
bottom
}
sum(new.bottom.calc(x,y))}
}
new.one.plus.a.ei<-(1+a*y)
    new.lgone.plus.a.ei<-log(one.plus.a.ei)
    new.summation<-sum(new.lgone.plus.a.ei)
    new.result<-((2/a^3)*(new.summation))
    new.result
new.final<-( sum(new.bottom.calc(x,y))-new.result)
{new.top<-(x+1/a)*y^2
    new.denominator<-(1+a*y)^2
    new.ratio<-new.top/new.denominator
    new.sum.ratio<-sum(new.ratio)
}
{
    third.term<-y/(1+a*y)
    third.term<-sum(third.term)
    third.term<-2/a^2*third.term

    third.term
}
new.all.together<-new.final+third.term+new.sum.ratio
new.all.together
sum(new.bottom.calc(x,y))
new.result
new.final
f.over.derivative<-all.together/new.all.together
```

```
f.over.derivative }
a<-(a-f.over.derivative)
cat(it, a,"\n")}}
{new1.bottom.calc<-function(x,y){
    bottom<-matrix(0,length(x),1)
        for (i in 1:length(x)){
            for (j in 1:(x[i]-1)){
            bottom[i]<-bottom[i]+((-1)*((j^2)/((1+a*j)*(1+a*j))))
}
                    }
bottom
}
sum(new1.bottom.calc(x,y))}
{new1.one.plus.a.ei<-(1+a*y)
    new1.lgone.plus.a.ei<-log(one.plus.a.ei)
    new1.summation<-sum(new1.lgone.plus.a.ei)
    new1.result<-((2/a^3)*(new1.summation))
    new1.result
new1.final<-( sum(new1.bottom.calc(x,y))-new1.result)
{new1.top<-(x+1/a)*y^2
    new1.denominator<-(1+a*y)^2
    new1.ratio<-new1.top/new1.denominator
    new1.sum.ratio<-sum(new1.ratio)
}
{
    new.third.term<-y/(1+a*y)
    new.third.term<-sum(new.third.term)
    new.third.term<-2/a^2*new.third.term


    new.third.term
}
new1.all.together<-new1.final+new.third.term+new1.sum.ratio
```

```
new1.all.together }
information<-new1.all.together
variance.a<-(1/new1.all.together)
{
          upper.bound.C.I.a.hat<-a+1.96*sqrt(variance.a)
}
{
     lower.bound.C.I.a.hat<-a-1.96*sqrt(variance.a)
}
{
     length.a.int<-upper.bound.C.I.a.hat-lower.bound.C.I.a.hat


     conf.interval.a.hat<-matrix(c(lower.bound.C.I.a.hat,upper.bound.C.I.a.hat), ncol=2)}
     length.a.int
     information
     variance.a
```

BIOGRAPHY OF THE AUTHOR

Anna Kettermann was born in St. Petersburg, Russia on April 18, 1974.

She entered St. Petersburg State Marine Technical University in 1991. In 1993 she received a scholarship from the University of Kaiserslautern, Germany. In 2001 she obtained an MS degree in Mathematics and Economics. In September 2001 she was enrolled in the Department of Statistics at the University of Connecticut and served as a Teaching Assistant. In January 2003 she was enrolled in the Department of Mathematics and Statistics at the University of Maine and served as a Teaching Assistant.

She is a candidate for the Master of Arts degree in Mathematics from The University of Maine in August, 2004.