Mathematical Statistics
Stockholm University

Estimating the Infection and Reinfection risk of Chlamydia trachomatis in Sweden during 1997-2000

Elinore Bengtsson

Examensarbete 2004:10

Postal address:

Mathematical Statistics
Dept. of Mathematics
Stockholm University
SE-106 91 Stockholm
Sweden

Internet:

http://www.math.su.se/matstat

Mathematical Statistics
Stockholm University
Examensarbete **2004:10**,
http://www.math.su.se/matstat

Estimating the Infection and Reinfection risk of *Chlamydia trachomatis* in Sweden during 1997-2000

Elinore Bengtsson*

May 2004

Abstract

In Sweden chlamydia and all other cases of sexually transmitted diseases (STD) must, according to law, be registered at the institute of infectious diseases (SMI). This implies that there is an established database concerning the STD from which one may retrieve information in order to determine the infection and reinfection risk. A complication is that each case of an STD is registered and identified with a code that may be shared with other individuals. This means that if a code appears more than once in data it may be one already infected individual that becomes infected again or an individual that shares the same code as the already infected that becomes infected.

In this thesis a model is constructed, considering this complication of a non-unique identification, in order to estimate the infection and reinfection risk on an individual basis. The model is a likelihood function and the confidence intervals of the estimates are obtained through two different methods; profile likelihood and bootstrap.

*Postal address: Dept of Mathematical Statistics, Stockholm University, SE-106 91 Stockholm, Sweden. E-mail: ronnioellis@hotmail.com. Supervisor: Mikael Andersson.

Contents

Chapter 1

Introduction

Chlamydia is a sexually transmitted disease, that during the years have had a fluctuating number of incidents in Sweden, but from 1997 the number of incidents has been increasing.

Chlamydia is a disease transmitted between people during sexual intercourse or at birth. The disease is caused by a bacteria called *Chlamydia trachomatis*. The bacteria is possible to encounter in the urethra, rectum or in the throat. Chlamydia is an asymptomatic disease which means that it is possible to be infected without developing any symptoms.

1.1 Motivation of the thesis

All diseases that are regarded as dangerous in Swedish society are contained in the law of the infectious diseases (Smittskyddslagen). Since 1988 chlamydia is also contained in this law.

According to this law, chlamydia id a notifiable disease and contact tracing is practiced. It is a duty of the doctor to notify the county medical officer and *Swedish institute for infectious disease control (SMI)* about the case.

For all diseases, except a sexually transmitted (STD), a case of infection is reported and identified in form of a personal number, whereas a case of a sexually transmitted disease is reported and identified in form of a national code. A national code is not unique for one individual, it can be shared by many people.

If there was an individual identification of each case, then it would be trivial to estimate the reinfection risk or to tell what proportion of individuals becomes infected. The complication of the procedure of identifying cases based on national codes forms the basis of this thesis.

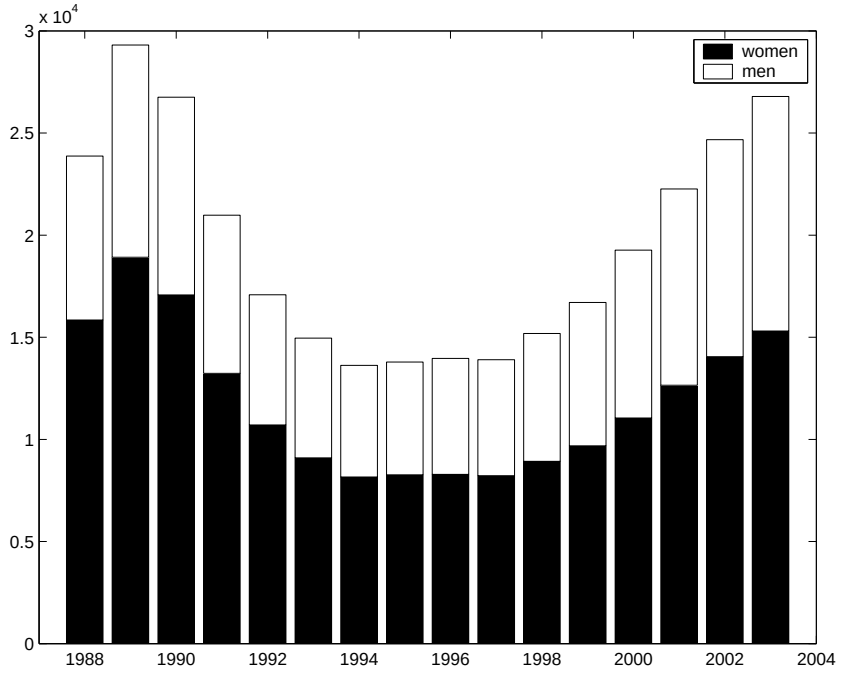


Figure 1.1: Number of incidents of chlamydia

1.2 The aim of this thesis

The aim of this thesis is to develop a method to make it possible to

- Estimate the risk to become infected.
- Estimate the reinfection risk, that is, given that an individual is infected once during a period, what is the probability to get infected again?

Chapter 2

Description of national codes and source data

2.1 National code and personal number

To protect the integrity, a case of a STD is registered at SMI as a national code. A national code contains information about the infected individual, such as gender and age, but does not reveal the identity of the individual. To be able to describe the form of the code, a brief section about the Swedish system of personal numbers follows.

All people that are registered in the national registration have a **personal number** as an identity description. The following personal number $640823 - 3234$ consists of three parts,

- The time of birth 640823, (yy-mm-dd)
- The birth registration number 323
- The control number 4

The birth registration number is odd if you are a man and even if you are a woman. Two persons born the same day have different numbers. Until 1990 this number was split up in different series so that each region in Sweden had a specific series, but now there is only one possible series. The control number is calculated according to the modulus ten principle.

The above example is taken from the national Swedish tax board and is a male that is born the 23 of August 1964 with birth number 323.

The National code consists of the year when the patient is born and the birth registration number together with the control number. Relating to the above example, the national code would be 64-3234 This implies that

a case registered at SMI is not a unique individual. Using the modulus ten algorithm it is possible to calculate the maximum number of possible individuals with the same code.

2.2 Data

The data available for further analysis consists of two different data bases:

- Database from SMI containing information on the reported cases as national codes, gender, county where the case received medical attention, clinic, date of case report arrival and year . The database covers the years 1997 to 2000. The number of incidents during this period with respect to gender can be seen in Table 2.1

	gender	
$year$	$male$	$female$
1997	5611	8148
1998	6181	8806
1999	6919	9591
2000	8093	10900

Table 2.1: Number of incidents with respect to gender

- Database containing information on the number of people sharing the same national code. The database originates from SCB, the national statistical office, and consists of information of people born from 1950 until 1985, taken in 2001. The total number of observations is 4182303. A histogram over national code data is shown in Figure 2.1

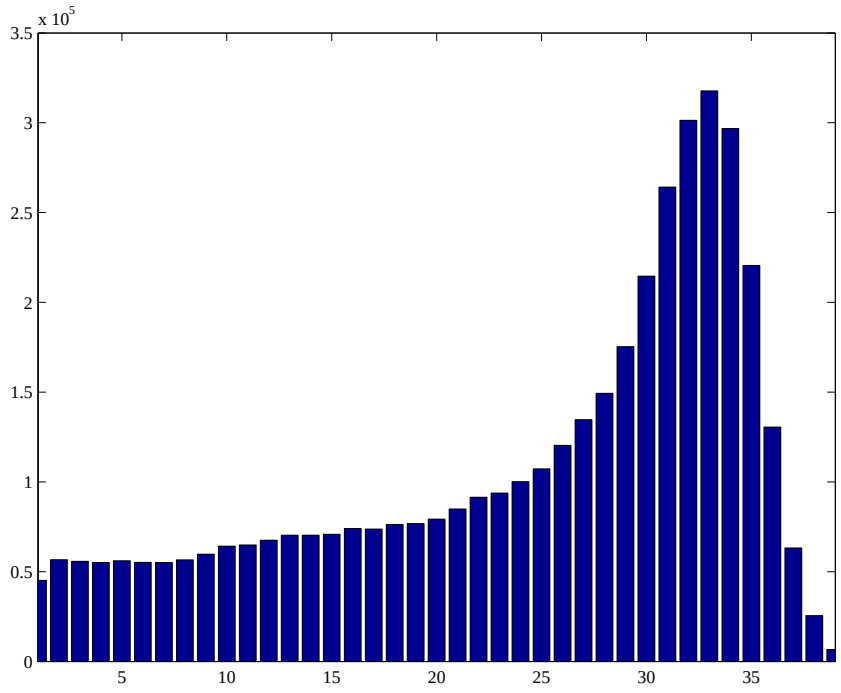


Figure 2.1: The frequency of the number of people sharing the same national-code

Chapter 3

A First Simplified Simulation

Before constructing a more complex model in order to obtain the estimates of interest it would be interesting to get some preliminary result. We may perform a simplified simulation examining the reinfection risk.

First of all, we have to sort out what we expect to encounter in terms of reinfection risk. One might expect that either the risk of infection is increased or decreased given that you already been infected. The motivation to believe that there is a increased risk is that when one have entered the infected set of individuals there is individuals that have a risky behavior, while for the decreased risk one may claim that if an individual becomes infected he/she change the behavior and avoid future exposure to infection. According to earlier studies Ramsedt(1991) and Ritmeier et.al(2001) a increased reinfection risk is to be expected. We proceed this chapter with the expectation that there is a increased risk of infection given that you already been infected.

A simple way to perform this simulation is to simulate the events of infection according to a null hypothesis and then compare the outcome of the null hypothesis with the real data. A suiting null hypothesis is that infection is spread randomly. That is, by selecting individuals that become infected uniformly with replacement.

To keep it as simple as possible, let the time of interest be two consecutive years. Meaning that we will investigate wether the risk of getting infected the second year is increased if you already been infected the first year.

Let X denote the number of times that a national code gets infected.

A simulation of a random variable X from the frequency distribution is achieved through:

- A stochastic variable from the distribution in form of an empirical distribution based on Figure 2.1 is denoted N_j. An outcome from a simulation is denoted

n_j =number of individuals sharing the national-code j,

We simulate r random variables from this distribution. A sufficient number of random variables r is reached when :

$$\sum_{j=1}^{r} N_j \approx T$$

Where T = Total number of individuals in the population. This number and $N_1, \ldots, N_r$ are assumed to be the same for the two consecutive years and is therefore not updated for the second year.

- For each year, randomly infect as many individuals as there are actually observed. Each individual share the same national code according to the simulated distribution. The total number of infected individuals that are to be infected is the total number of real cases during that year.

- For each national code, sum the total number of times the individuals have become infected.

- Finally we compare the simulated number of infected national codes for the two consecutive years with the real data.

It is only meaningful to compare the two distributions if there is at least one national code infected in one of the two consecutive years. This reason for this is that there are only cases in the database, uninfected are not represented here.

The outcome is illustrated as differences between the simulated and the actual outcome of the bivariate distribution of the number of cases the first year and the number of cases the second year . The distribution with respect to gender for 1997/1998 is illustrated in Figure 3.1 and 3.2. The simulations for years 1998/1999 and 1999/2000 with respect to gender are illustrated in Appendix A.

As we can see in Figures 3.1 and 3.2, the number of national codes appearing 0 the first and 1 the second year (0,1), or vice versa (1,0), are the only outcomes where a national code appears more frequently in the simulated material. The reason for this is that simulated infections are not likely to strike the same national code in both years as in the real material. That is, the risk of reinfection for an individual in the simulated material within one year is too low. With this conclusion there is no point in proceeding the work with this simulation. Instead let us turn to the construction of a more informative model.

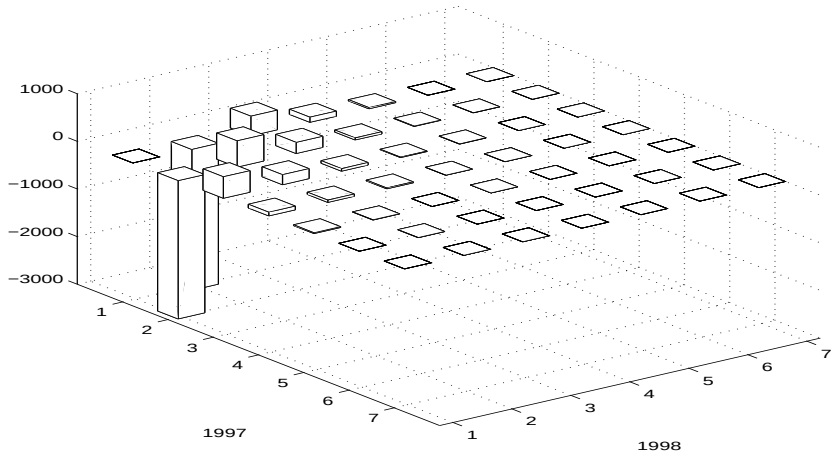


Figure 3.1: Difference between actual outcome and simulation 1997/1998, women

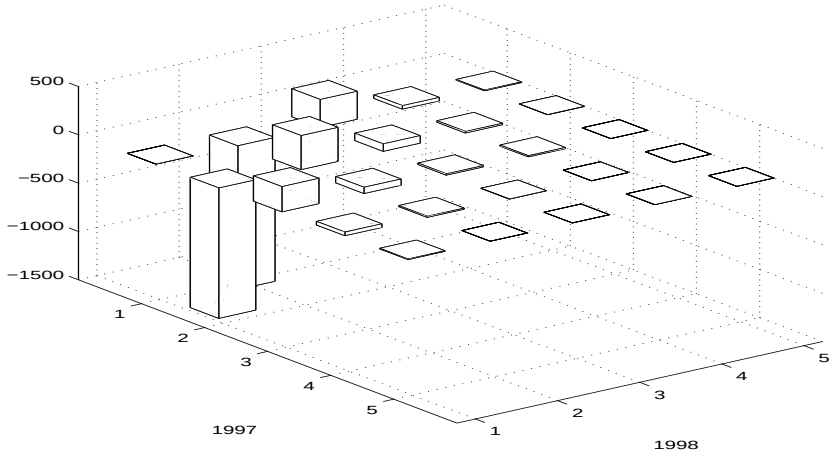


Figure 3.2: Difference between actual outcome and simulation 1997/1998, men

Chapter 4

A Model for estimating the infection risk

Before continuing to define the model, an important distinction is emphasized. We recall that what is observed, concerning the number of cases, is the number of times a national code appears in the material and that what we actually are interested in is the number of times a individual is observed. These two different concepts might be hard to keep apart. It is possible to consider the events of that an individual gets infected (in time) as a stochastic point process and the events of that a national code appears (in time) as a sum of the stochastic point processes of the individuals that share the same national code.

Denote the number of times a national code appears i times in the data as M_i and the probability that an individual becomes infected i times as p_i and denote the vector of these probabilities as $\mathbf{p} = [p_0, p_1, \ldots]$. In the proceeding sections the model is defined and estimates of the parameters of interest are calculated. The main purpose of the model is to estimate p_i using the information that consists of the observations of M_i. The model that will be established yields an ordinary likelihood function and through ML theory the estimates are obtained.

4.1 The Likelihood

Let Y denote the number of times a national code appears, taking one of the values $0, 1, \ldots$.

The likelihood is defined as:

$$L(\mathbf{p}) = \prod_{y=0}^{\infty} P(Y = y)^{M_y} \tag{4.1}$$

One should notice that the likelihood is defined for $y = 0$, while M_y is an observational vector with $y = 1, 2, \ldots$, this since we have no information of national codes that do not appear in the data.

This leads to a problem with the above defined likelihood. One way to avoid this is to to form a conditional likelihood. A likelihood given that a national code appears at least once in the material.

$$\begin{aligned} L(\mathbf{p}|Y > 0) &= \Pi_{y=1}^{\infty} P(Y = y|Y > 0)^{M_y} \\ &= \Pi_{y=1}^{\infty} \left(\frac{P(Y = y)}{1 - P(Y = 0)} \right)^{M_y} \end{aligned}$$

Let N denote the number of individuals that share the same national code. The probability $P(Y = y)$ is presently unknown. It may be fully developed using the law of total probability, which leads to:

$$P(Y = y) = \sum_{n=1}^{\infty} P(Y = y|N = n)P(N = n)$$

$P(Y = y|N = n)$ is the *probability that a national code appears y times given that there are n people sharing the same national code.* In the following sections the details of the different parts of the likelihood (4.1) are uncovered.

4.1.1 The probability that n people share the same code

The probability $P(N = n)$, $n = 1 \ldots m$ (m = maximum number of people sharing the same national code), is estimated by the empirical distribution of the observed frequency distribution. Either we let $P(N = n)$ follow the empirical distribution, or we may simulate sufficiently many random samples from it. In the proceeding the simulated version is used.

4.1.2 The conditional probability, P(Y=y|N=n)

Denote the number of individuals that become infected i times as X_i. Let k be the maximum number of times an individual gets infected, it then follows: $(X_0, \ldots, X_k) \sim Mult(n, p_0, \ldots, p_k)$, Since $y = 0 \ldots 9$ (9 is the maximum number of times the same national code appears in the material) and $n = 1, \ldots 39$ one realizes that there will be many subparts of the probability $P(Y = y|N = n)$, due to many possible combinations of coefficients in the multinomial distribution.

If we return to the probability $P(Y = y)$, it seems that it is going to turn out to be a quite complicated expression. What complicates it is, as mentioned above, all the possible combinations of the multinomial coefficients. Remembering that the coefficients represent the numbers of individuals infected i times given that in total y national codes appear in data. So what we are after is a way to determine all combinations of $X_0, \ldots, X_k$ so that:

$$\sum_{i=0}^{k} X_i = n$$
$$\sum_{i=1}^{k} iX_i = y$$

One straightforward way to determine all possible combinations is to construct an algorithm.

4.1.3 Algorithm

$y = 1$	1				
$y = 2$	11	2			
$y = 3$	111	12	3		
$y = 4$	1111	112	13	22	4
$y = 5$	..	..	..	..	..

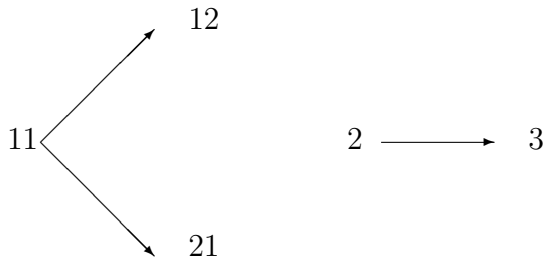
Table 4.1: All possible combinations that sum to y.

Table 4.1 illustrates all combinations that are possible if we want to sum up to y. So for example on the third row one can see that if we want to sum up to 3 we can do it in three ways. Either 1+1+1 or 1+2 or 3. Practically, this means that given that a national code appears 3 times, there are 3 different ways these infection events may be distributed on individuals sharing this code. Either 3 individuals become infected once each, one individual becomes infected twice and one once or one individual becomes infected three times.

Going from stage y to $y + 1$ is either done by letting one new individual get infected or letting an individual already infected become infected once more. Going from stage two to three would be done in the following manner:

One additional individual becomes infected

An already infected individual becomes infected



The final step in this algorithm is to remove all replicates. In the above example, since 12 and 21 represent the same combination, one of them is removed.

The following will work as an illustration on how the probabilities $P(Y = y|N = n)$ turn out for different y and n during a period of time.

(y=0)

$$P(Y = 0|N = 1) = \frac{1!}{1!0!...0!} p_0^1 p_1^0 ... p_k^0 \tag{4.2}$$

$$P(Y = 0|N = 2) = \frac{2!}{2!0!...0!} p_0^2 p_1^0 ... p_k^0 \tag{4.3}$$

and so on for $n = 1..39$.

(y=1)

$$P(Y = 1|N = 1) = \frac{1!}{0!1!...0!} p_0^0 p_1^1 ... p_k^0 \tag{4.4}$$

$$P(Y=1|N=2) = \frac{2!}{1!1!...0!} p_0^1 p_1^1 ... p_k^0 \tag{4.5}$$

and so on for $n = 1..39$.

(y=2)

$$P(Y=2|N=1) = \frac{1!}{0!0!1!...0!} p_0^0 p_1^0 p_2^1 ... p_k^0 \tag{4.6}$$

$$P(Y=2|N=2) = \frac{2!}{1!0!1!...0!} p_0^1 p_1^0 p_2^1 ... p_k^0 + \frac{2!}{0!2!0!...0!} p_0^0 p_1^2 p_2^0 ... p_k^0 \tag{4.7}$$

We have now completely determined the form of $P(Y = y)$ and we may proceed with estimation of the parameters.

4.2 Estimating Parameters

4.2.1 Solving the ML equations

After defining the likelihood we may proceed to calculate the estimates based on the ML equations. But before doing so, we define the optimization problem:

$$\max L(\mathbf{p}) \; so \; that \begin{cases} \sum_{i=0} p_i = 1 \\ \sum_{i=0} i p_i = T_1 \\ p_i \geq 0, \forall i = 0..k \end{cases}$$

$$\begin{aligned} T_0 &= Total\ population \\ T_1 &= \frac{number\ of\ infection\ events}{T_0} \end{aligned}$$

As one can see this is an optimization problem with 3 constraints. Since there are linear restrictions to this problem, it may be re-parametrized into a lower dimensional problem. The following re-parameterization leaves two arbitrarily chosen variables out, in this case p_0 and p_1, and express them in terms of the defined parameters.

$$\psi_i = \log(\tfrac{p_i}{1-p_0}), i = 2 \dots k$$

Or reversed we have that

$$p_i = (1 - p_0)e^{\psi_i}$$

p_0 and p_1 are expressed in terms of the rest of the parameters:

$$\begin{cases} p_0 = T_1 - \frac{1}{1+e^{\psi_1}+2e^{\psi_2}+\dots(n-2)e^{\psi_{n-2}}} \\ p_1 = 1 - p_0 - \sum_{i=2}^{n} p_i \end{cases}$$

The re-parametrized problem:

$$\max L(\mathbf{p}) \; so \; that \left\{ \; \forall_i p_i \geq 0 \right.$$

In the proceeding the log likelihood will be considered, but will of course yield the same result. Even if we have reduced the optimization problem by two dimensions we still have a constrained problem. This is not theoretically a problem, but practically it leads to that either we have to use an optimization method with constraints or we put constraints on the problem by ourselves. This is done by imposing a penalty to the log-likelihood every time the optimization algorithm tries to make a step outside the boundaries of the problem. Before proceeding a brief section containing the principal conditions for an optimal solution of an unconstrained problem is presented.

4.2.2 Conditions for an optimal solution

Even if we use a pre-implemented optimization algorithm of some program one should not be too sure that it really is an optimum delivered by the program. And even if it is an optimum it might be a local optimum. In order to make sure that we really obtain a global optimum, we should check if the optimality conditions are fulfilled.

Denote the Hessian of the function as $\mathbf{H}$.

Definition 4.2.1 *Suppose that f is a twice differentiable function on a convex set* $\boldsymbol{X}$*. An allowed solution point* $x \in \boldsymbol{X}$ *is then a global minimum if f is a convex function. f is a convex function if* $\boldsymbol{H}$ *is positive definite or semi positive-definite*

Definition 4.2.2 *A set is convex if for every choice of* $x^{(1)}$ *and* $x^{(2)} \; \epsilon \boldsymbol{X}$ $\boldsymbol{x}$$=\lambda \; x^{(1)} + (1 - \lambda)x^{(2)} \in \boldsymbol{X}, \quad \lambda \in [0, 1]$

Finally, we must not forget that a third condition of optimality is that the partial derivatives in the optimal estimations of the likelihood should be zero.

4.2.3 Optimization Algorithm

The optimization algorithm chosen is the pre-implemented optimization algorithm in Matlab without constraints, fminsearch. We use the method to assign a penalty to the likelihood. If the optimization algorithm tries to take a step outside the boundary, we assign a value to the log-likelihood that is very small. In this case we have to use a minimization algorithm so instead of a small value we assign a very large value and use the minus log-likelihood. The procedure fminsearch uses an algorithm called Nelder-Mead simplex method. The details of the method will not be explained in this paper, but are to be found for example in Lagarias (1998).

4.2.4 Estimating the Partial Derivatives and the Hessian

According to the optimality conditions we must make sure that the Hessian is negative-definite, since it is a maximum, and that the partial derivatives are zero. The analytical derivatives and Hessian of the log-likelihood is very complicated to derive so in this situation the numerical approximations of the derivatives and the Hessian has to be used.

A numerical approximation of the partial derivatives are obtained by a Taylor series expansion of the first partial derivatives. The form is,

$$\frac{\partial f}{\partial x_i} \approx \frac{f(x_1, \ldots, x_i + \Delta_{x_i}, \ldots, x_n) - f(x_1, \ldots, x_n)}{\Delta_{x_i}}$$

A numerical approximation of the Hessian is also derived from a Taylor second order approximation, the form is,

$$\begin{cases} \frac{\partial^2 f}{\partial x_i^2} \approx & \frac{f(x_1,\ldots,x_i+\Delta_{x_i},\ldots,x_n)+f(x_1,\ldots,x_i-\Delta_{x_i},\ldots,x_n)}{\Delta_{x_i}^2} \\ & \frac{-2f(x_1,\ldots,x_n)}{\Delta_{x_i}^2} \\ \frac{\partial^2 f}{\partial x_i \partial x_j} \approx & \frac{f(x_1,\ldots,x_i+\Delta_{x_i},\ldots,x_j+\Delta_{x_j},\ldots,x_n)+f(x_1,\ldots,x_n)}{\Delta_{x_i}\Delta_{x_j}} \\ & \frac{-f(x_1,\ldots,x_i+\Delta_{x_i},\ldots,x_n)-f(x_1,\ldots,x_j+\Delta_{x_j},\ldots,x_n)}{\Delta_{x_i}\Delta_{x_j}} \end{cases}$$

The approximation of the Hessian is obtained by letting Δ_{x_i} and Δ_{x_j} assume very small values.

Chapter 5

Representation of Source Data.

5.1 A brief discussion about the partitioning of the data

The model is now defined. We want to estimate the risk to become infected with chlamydia and the reinfection risk. To be able to obtain these two estimates we have to consider the partitioning of the data. The infection risk is only a matter of definition, we only split up the data in subgroups that are of interest. The probability to become infected and the reinfection probability would then be possible to estimate through the parameters $p_0, \ldots, p_k$ in the model.

We have to remember that gender plays an important role in the incidence of chlamydia. So partition for gender is obvious. One possible way to represent data would be according to Table 2.1. A different approach is to partition according to gender and to consider a two-year period, that is male-female and three different time-periods,1997-1998,1998-1999,1999-2000. The result of this representation is the frequency Table 5.1 and 5.2. One should notice that with this representation of data the time periods overlap.

Frequency of national codes

year	1	2	3	4	5	6	7	8
1997-1998	9424	2247	666	196	42	12	3	0
1998-1999	9735	2523	762	233	56	23	3	1
1999-2000	10241	2878	870	295	98	22	10	2

Table 5.1: Female

Frequency of national codes

year	1	2	3	4	5	6	7	8	
1997-1998	8084	1382	230	57	6	0	0	0	0
1998-1999	8698	1547	317	63	9	4	0	0	0
1999-2000	9484	1866	408	110	20	1	3	0	1

Table 5.2: Male

It is also possible to widen the time frame and consider a three-year periods,1997-1999 and 1998-2000, or to partition for different ages.

5.2 Is there a preferable representation of data for further analysis?

A time frame of one year might be too narrow in the sense that in order to observe repeated infections, the reinfections must occur within that time frame. This would justify a further analysis with a time frame at least longer than one year. If the data are split up in three year periods it will be national codes that appear more times than with a two year frame. This will increase the time for the calculations. Since the calculations are very computer intensive, a time frame of two years is chosen. Partitioning for age is one of the most attractive ideas, but is not performed in the scope of this thesis.

Chapter 6

Confidence Intervals for the Estimated Parameters

It is also of interest to be able to decide the precision of the estimates and in order to do so we must choose some measure of the variability of the estimated parameters. Let us now look at Tables 5.1 and 5.2. It is obvious that there are very few observations in some parts of the table. One usual procedure in order to obtain a confidence interval is to apply the asymptotic results of ML-theory, where the parameters $\hat{\mathbf{p}}$, have approximate distribution $\mathrm{N}(\mathbf{p}, I(\mathbf{p})^{-1})$. In this situation it is not feasible. We have to find a different approach to this problem. Two different approaches are profile-likelihood and bootstrapping. In the following section we will briefly go through the details of both methods.

6.1 Profile Likelihood

The basic idea behind profile likelihood is to keep one parameter fixed and optimize the likelihood for all other parameters. More formally,

Definition 6.1.1 *Let $\boldsymbol{p}$ denote the full vector of the parameters $p_0, \ldots, p_k$ and $\mathbf{p}_i$ denote the partial vector when the parameter of interest p_i is excluded. Given the joint likelihood L($\boldsymbol{p}$), the profile likelihood is*

$$L_{profile}(p_i) = \max_{\mathbf{p}_i} L(\boldsymbol{p})$$

The profile likelihood is considered as a normal likelihood and shares all the characteristics of a likelihood. Let us first construct the log likelihood-ratio for a general parameter θ:

$$W = 2 \log \frac{L(\hat{\theta})}{L(\theta)} \tag{6.1}$$

We derive the asymptotic distribution of the ratio (6.1) by first performing a second-order Taylor-expansion of a multi dimensional θ around $\hat{\theta}$ of $L(\theta)$.

$$\log L(\theta) \approx \log L(\hat{\theta}) + S(\hat{\theta})(\hat{\theta} - \theta) - \frac{1}{2}(\hat{\theta} - \theta)' I(\hat{\theta})(\hat{\theta} - \theta)$$

Where $S(\hat{\theta})$ is the score function and $I(\hat{\theta})$ is the information matrix. Since $S(\hat{\theta})$ is zero, the remaining series may be written as

$$L(\theta) = ke^{-\frac{1}{2}(\hat{\theta}-\theta)' I(\hat{\theta})(\hat{\theta}-\theta)} \tag{6.2}$$

According to (6.2) we see that this is the likelihood for a single observation $\hat{\theta}$ from a $\mathrm{N}(\theta, I(\hat{\theta})^{-1})$

The ratio (6.1) is now approximated by

$$\begin{aligned} W &= 2 \log \frac{L(\hat{\theta})}{L(\theta)} \\ &= (\hat{\theta} - \theta)' I(\hat{\theta})(\hat{\theta} - \theta) \rightarrow^d \chi^2(r) \end{aligned} \tag{6.3}$$

That is, W is approximately χ^2 distributed with r degrees of freedom, r is the degrees of freedom in the nominator - degrees of freedom in the denominator.

Let us Reformulate the ratio in terms of profile likelihood,

$$W_i = 2 \log \frac{L(\hat{\mathbf{p}})}{L_{profile}(\hat{p}_i)}$$

In this case $r = 1$, as the profile likelihood contains only one free parameter less than the likelihood since one parameter is fixed in the profile likelihood.

A 100(1-α)% confidence region for p_i is defined as

Definition 6.1.2

$$CR : \{p_i, W_i \leqslant \chi^2_\alpha(1)\}$$

Where $\chi^2_\alpha(1)$ is the $1-\alpha$ percentile of the χ^2 distribution. The distribution of W_i is obtained as mentioned before as an asymptotic result. In order to obtain good estimates the number of observations must be sufficiently large. Initially, that was the main motivation to use the approach of a profile likelihood instead of estimating the whole parameter vector $\mathbf{p}$ and use the information matrix to estimate the confidence intervals of all the parameters. What observations that must be sufficiently large is out of the scope of this thesis. But what can be concluded is that if the profile likelihood has a nice quadratic shape, then the normal approximation should be alright.

6.2 Bootstrap

Bootstrapping is one of many different re-sampling methods. The motivation to the method is that one wants to estimate a statistic θ from a unknown distribution $F(x)$. Assume that given is an observed sample $x_1, \ldots, x_n$, the parameter θ is estimated from the sample $\hat{\theta} = \hat{\theta}(x_1, \ldots, x_n)$. The bootstrap approach is to consider the empirical distribution of the sample, $F_n(x) = \frac{1}{n}\sum_{i=1}^{n} 1_{\{x_i \leq x\}}$ as a true distribution and then re-sample from this considered true distribution.

The procedure of estimating a 95% confidence interval is as follows:

- calculate the "true" empirical distribution as $F_n(x) = \frac{1}{n}\sum_{i=1}^{n} 1_{\{x_i \leq x\}}$
- simulate N bootstrap samples $X_1^*, \ldots, X_n^*$ from F_n.
- calculate the statistic $\hat{\theta}_k^*(X_1^*, \ldots, X_n^*)$ for sample k.
- A 95% confidence interval for $\hat{\theta}$ is most easily achieved by the percentile method which is CI:$[\hat{\theta}^*_{[0.025N]}, \hat{\theta}^*_{[0.975N]}]$

For further reading see Efron and Tibshirani (1993).

Chapter 7

Results

Both the bootstrapping and the profile likelihood method are very computer intensive, even if bootstrapping is the most. Considering that there are six different representations of data, and for each partitioning two different parameter estimates are of interest, it is going to take some time to derive the estimates. Because of this, all parameter estimates will not be retrieved. The profile likelihood will be used to estimate the infection probability and the reinfection probability, while the bootstrap method will only be used as a comparison to the profile likelihood.

In the following three sections, results of the two main aims of the thesis, as declared in the beginning and as well a secondary product from the bootstrapping method. An estimate that is not available is denoted *na*

7.1 What is the probability to get infected with chlamydia?

In the following sections we denote a confidence interval obtained with the bootstrap method CI_B and obtained with profile likelihood CI_P.

The probability to becoming infected with chlamydia is $1 - p_0$. The optimal solution to the ML equations is the point estimate $1 - \hat{p}_0$. Point estimates of $1 - p_0$ and the corresponding confidence intervals are displayed in Tables 7.1, 7.2, 7.3 and 7.4

7.2 Reinfection risk

A reinfection risk is defined (as before) as *the probability to become infected again given that you already been infected once.* More strictly:

	$1-\hat{p}_0$	$CI_B(1-p_0)$	
1997-1998	0.001134	0.001122	0.001140
1998-1999	0.001238	0.001223	0.001242
1999-2000	0.001379	na	na

Table 7.1: Confidence interval for individual infection probability, bootstrap ,men

	$1-\hat{p}_0$	$CI_P(1-p_0)$	
1997-1998	0.001134	0.001125	0.001143
1998-1999	0.001238	0.001225	0.001245
1999-2000	0.001379	0.001370	0.001392

Table 7.2: Confidence interval for individual infection probability, profile likelihood,men

	$1-\hat{p}_0$	$CI_B(1-p_0)$	
1997-1998	0.001433	0.001417	0.001462
1998-1999	0.001518	na	na
1999-2000	0.001638	na	na

Table 7.3: Confidence interval for individual infection probability, bootstrap, women

	$1-\hat{p}_0$	$CI_P(1-p_0)$	
1997-1998	0.001433	0.001420	0.001446
1998-1999	0.001518	0.001506	0.001533
1999-2000	0.001638	0.001632	0.001657

Table 7.4: Confidence interval for individual infection probability, profile likelihood, women

$$R = \frac{1-p_0-p_1}{1-p_0}$$

The point estimates of R and the corresponding confidence intervals are displayed in Tables 7.5, 7.6, 7.7 and 7.8.

The profile likelihood of the reinfection risk may be displayed visually. If we re-scale the profile likelihood it is possible to declare a cutoff point for

	$\hat{R}$	$CI_B(R)$	
1997-1998	0.1593	0.1585	0.1627
1998-1999	0.1692	0.1622	0.1746
1999-2000	0.1872	na	na

Table 7.5: Confidence interval for reinfection risk, bootstrap ,men

	$\hat{R}$	$CI_P(R)$	
1997-1998	0.1593	0.1517	0.1667
1998-1999	0.1692	0.1619	0.1759
1999-2000	0.1872	0.1802	0.1947

Table 7.6: Confidence interval for reinfection risk, profile likelihood,men

	$\hat{R}$	$CI_B(R)$	
1997-1998	0.2364	0.2281	0.2462
1998-1999	0.2545	na	na
1999-2000	0.2735	na	na

Table 7.7: Confidence interval for reinfection risk, bootstrap, women

	$\hat{R}$	$CI_P(R)$	
1997-1998	0.2364	0.2297	0.2447
1998-1999	0.2545	0.2477	0.2621
1999-2000	0.2735	0.2682	0.2808

Table 7.8: Confidence interval for reinfection risk, profile likelihood, women

the confidence interval. Recalling the confidence region in Definition 6.1.2, the cutoff point yields the limits of the confidence region. In the case of fixating one parameter the degree of freedom are 1 and it follows that the $\chi^2_{0.05}(1) = 3.843$. This derives the cutoff point as 0.152.

Two graphs will be displayed as examples in Figure, 7.1 and 7.2. In the graphs the cut off points and the corresponding confidence region are marked.

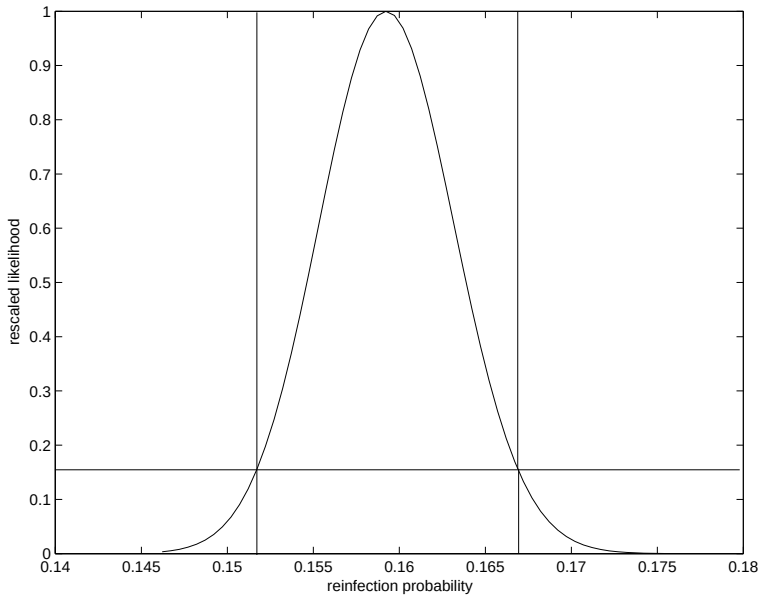


Figure 7.1: Re-scaled Profile likelihood for men 1997-1998

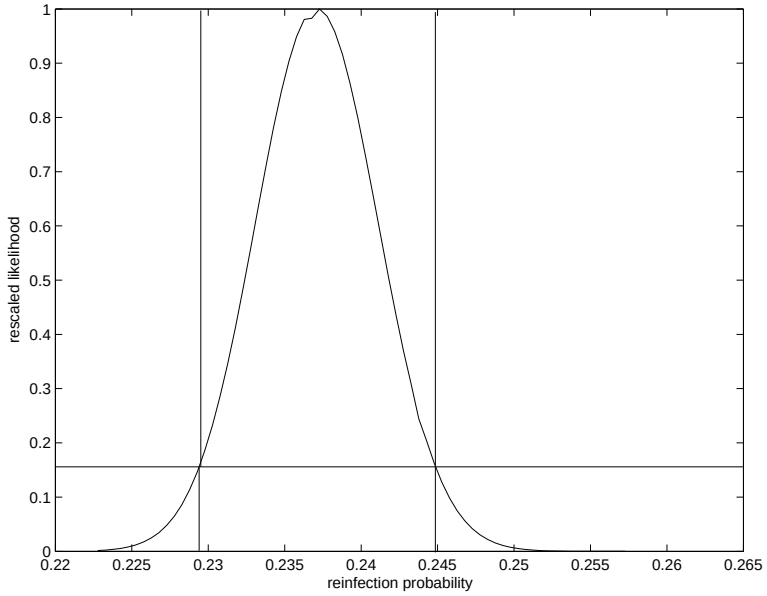


Figure 7.2: Re-scaled Profile likelihood for women 1997-1998

7.3 What is the probability to get infected with chlamydia i times?

One extra bonus from using the bootstrap method is that confidence intervals of all the separate parameters $p_0, \ldots, p_n$ are obtained. The results are displayed in Appendix B in Tables B.1, B.2.

7.3.1 A comparison between the number of infected individuals and the number of reported national codes

After obtaining the estimates that an individual get infected i times during a period, it would be interesting to compare the *proportion of actual reported number of cases* with the *estimated probability that an individual becomes infected*. We define the proportion of cases as *the total number of cases divided by the total population*. What we expect is that the proportion of actual reported cases is larger than the estimated probability that an individual becomes infected. The reason for this is that in the proportion of actual reported cases it is not taken into account that one individual might represent multiple cases.

Denote the proportion of actual reported cases with $1 - p_0^a$ and the estimated (as before) with $1-\hat{p}_0$, we would like to compare $1-p_0^a$ with $1-\hat{p}_0$. First we take a look at the estimations obtained with the *bootstrapping method*. Tables 7.9 and 7.10 present a 95% confidence interval for $1-\hat{p}_0$ and the probability $1 - p_0^a$. A brief look in the tables reveal that the probability $1 - p_0^a$ is , as we expected, in all cases larger than the probability $1 - \hat{p}_0$.

	year	95 % CI		$1 - p_0^a$
men	1997-1998	0.001122	0.001134	0.001352
	1998-1999	0.001223	0.001238	0.001498
	1999-2000	*na*	*na*	0.001711
women	1997-1998	0.001417	0.001462	0.001898
	1998-1999	*na*	*na*	0.002058
	1999-2000	*na*	*na*	0.002284

Table 7.9: Estimated individual infection probability versus actual reported proportion of cases, bootstrapping method

	year	95 % CI		$1 - p_0^a$
men	1997-1998	0.001125	0.001143	0.001352
	1998-1999	0.001225	0.001245	0.001498
	1999-2000	0.001370	0.001392	0.001711
women	1997-1998	0.001420	0.001446	0.001898
	1998-1999	0.001506	0.001533	0.002058
	1999-2000	0.001632	0.001657	0.002284

Table 7.10: Estimated individual infection probability versus actual reported proportion of cases, profile likelihood

Chapter 8

Previous studies and remarks

8.1 Previous studies

To round off, a brief comparison with two other studies on the reinfection risk are performed. The first study is made by Ramstedt(1991). In this report there is a retrospective study made at Sahlgrenska Hospital in Gothenburg, Sweden. Data were collected in two 15 month periods, the first from January 1979 to March 1980 and the second from January 1983 to March 1984. Out of 2181 observations 156 were reinfected within 12 months, that is the probability to become reinfected is 0,0715. The second study was performed in Denver, USA by Ritmaier er al.(2001). A retrospective cohort study that took place at a STD clinic during one connected 30 month period, between January 1997 and June 1998. The Study-population were at baseline 3568 individuals, (notice that not all were infected at baseline). At the end of the study the number of infections and reinfections are summarized in Table 8.1

		follow-up	
		yes	no
baseline	yes	99	392
	no	286	2791

Table 8.1: number of incidents

If we would like to compare the outcome of this study with the above, one could conclude that the probability of reinfection given that at the first control the patient is infected is much higher, 0,202.

8.2 Remarks

8.2.1 Partition of data

The decision to partition data the way that was done was mostly due to time consuming optimizations. It would be very interesting to partition the data into smaller subsets. For example, to take into account that there is a higher incidence of chlamydia during the younger years and partition data into, let's say four different age classes.

8.2.2 Bootstrap and profile likelihood

A result of this thesis was confidence intervals obtained from two different approaches, bootstrapping and profile likelihood. Let us first take a look at the probability to become infected, in Tables 7.1, 7.2, 7.3 and 7.4. The width of the confidence intervals for the two different methods seems more or less the same, they are slightly wider in the profile likelihood method. One reason for this might be that the chosen step-length to calculate the profile likelihood is too rough. This decreases the precision of the profile likelihood. This can also be seen in the Figure 7.2, where the profile likelihood at one point is irregular, this depending on a too rough step-length. It is possible to narrow the step length, but then it will take longer time to calculate. Now, lets turn to the confidence intervals of the reinfection risk illustrated in Tables 7.5, 7.6, 7.7 and 7.8. The width of the confidence intervals of the reinfection risk is also wider for the profile likelihood method, except for the subpopulation "female 1997/1998".

With both methods we obtain acceptably narrow confidence intervals. One divergence from that is the interval of the separate probabilities to become infected i times, obtained with the bootstrapping method for women 1997/1998, see Appendix B Table B.3. One can see that the confidence interval of parameter p_4 looks a bit "strange". One should notice that it is not strange, it is only the lower limits that are a bit wider not the upper.

The two methods does build on assumptions which don't have to be true. The main drawback with the confidence intervals of the profile likelihood method is that we have to assume an asymptotic distribution, which might not be fulfilled. With the confidence intervals obtained with the bootstrapping method, we do assume that the observed outcome is the real distribution, which also is questionable.

Finally, both methods do assume doubtful initial approximations, still, both methods exhibit almost the same confidence intervals and cover the point estimates. This is a good indication that the confidence intervals ob-

tained are of an acceptable quality.

Appendix A

Figures

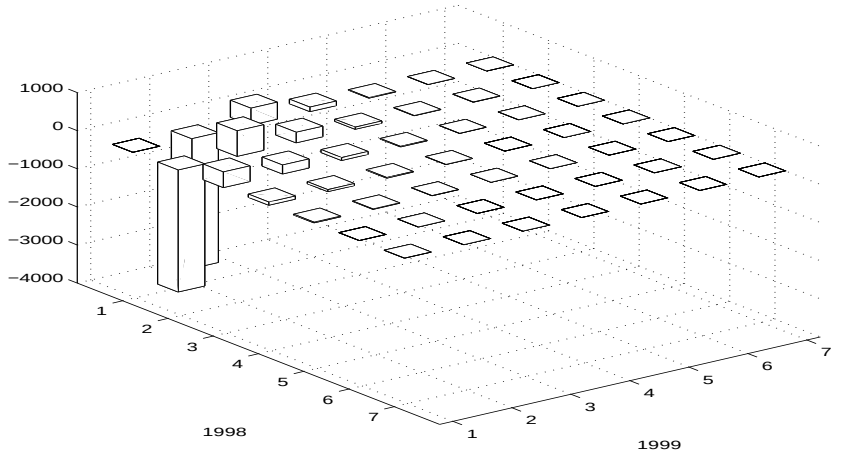


Figure A.1: Difference between actual outcome and simulation 1998/1999, women.

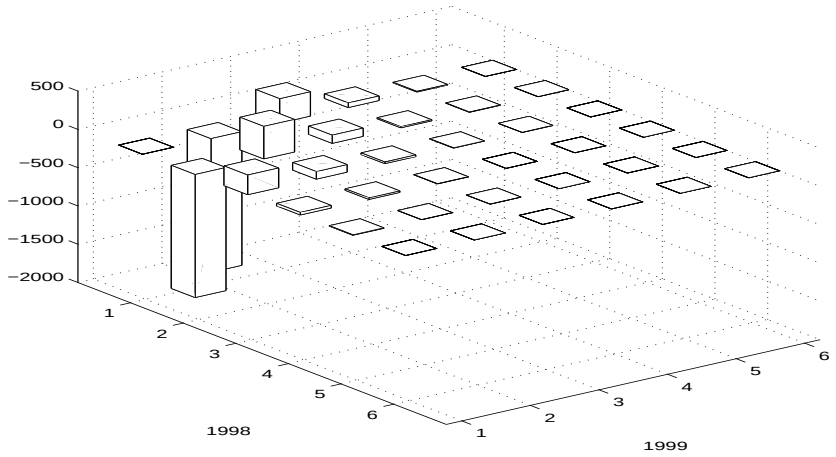


Figure A.2: Difference between actual outcome and simulation 1998/1999, men.

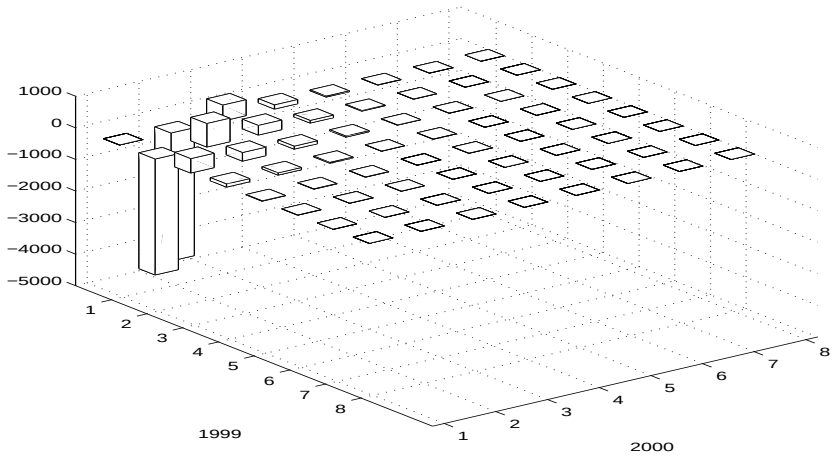


Figure A.3: Difference between actual outcome and simulation 1999/2000, women

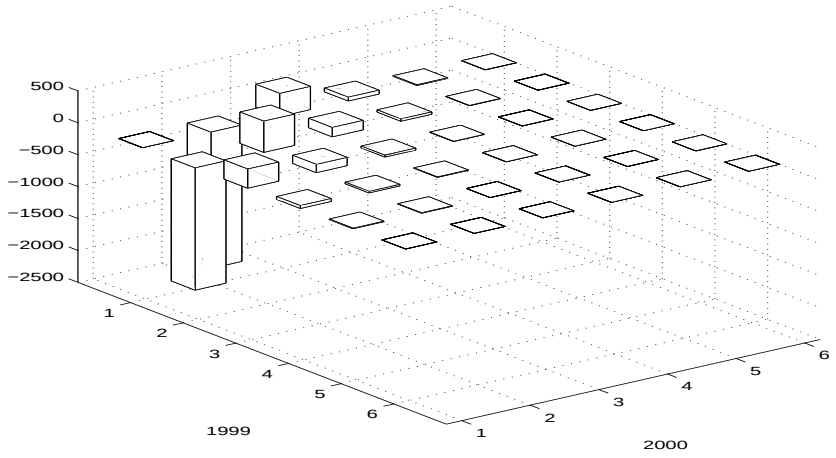


Figure A.4: Difference between actual outcome and simulation 1999/2000, men

Appendix B

Tables

i	95% CI(p_i)	
0	0.9988	0.9988
1	$0.9380 * 10^{-3}$	$0.9689 * 10^{-3}$
2	$0.1434 * 10^{-3}$	$0.1582 * 10^{-3}$
3	$0.2005 * 10^{-4}$	$0.2692 * 10^{-4}$
4	$0.4494 * 10^{-5}$	$0.7955 * 10^{-5}$
5	0	$0.1011 * 10^{-5}$

Table B.1: estimated probability that a man becomes infected i times 1997-1998

i	95% CI(p_i)	
0	0.9987	0.9987
1	$0.1013 * 10^{-2}$	$0.1045 * 10^{-2}$
2	$0.1605 * 10^{-3}$	$0.1762 * 10^{-3}$
3	$0.2886 * 10^{-4}$	$0.3633 * 10^{-4}$
4	$0.5410 * 10^{-5}$	$0.9274 * 10^{-5}$
5	$0.1900 * 10^{-12}$	$0.1425 * 10^{-5}$
6	$0.1086 * 10^{-12}$	$0.8514 * 10^{-6}$

Table B.2: estimated probability that a man becomes infected i times 1998-1999

i	95% CI(p_i)	
0	0.9985	0.9985
1	$0.1069 * 10^{-2}$	$0.1126 * 10^{-2}$
2	$0.2347 * 10^{-3}$	$0.2567 * 10^{-3}$
3	$0.6503 * 10^{-4}$	$0.8693 * 10^{-4}$
4	0	$0.2286 * 10^{-4}$
5	$0.2135 * 10^{-5}$	$0.5517 * 10^{-5}$
6	$0.1151 * 10^{-8}$	$0.1956 * 10^{-5}$
7	0	$0.7088 * 10^{-6}$

Table B.3: estimated probability that a woman becomes infected i times 1997-1998

Bibliography

[1] Kristina Ramstedt, Lars Forsman , Gunnar Johannison. Contact tracing in the control of genital *Chlamydia trachomatis* infection. *International journal of STD & Aids* 1991; 2:116-118.

[2] Cornelis A.Ritmeier, Rogier van Bemmelen, Franklym N.Judson, John M.Douglas. Incidence and Repeat Infection Rates of *Chlamydia trachomatis* Among Male and Female Patients in an STD Clinic. *Sexually transmitted diseases* 2002; 29:65-72.

[3] Jeffrey C.Lagarias, James A.Reeds, Margaret H.Wright, Paul E.Wright. Convergence properties of the nelder-mead simplex method in low dimensions. *Siam J.Optim* 1998; 9:112-147.

[4] Jan Lundgren, Mikael Rönnkvist, Peter Värbrand. *Linjär och icke linjär optimering*. Studentlitteratur 2001.

[5] Yudi Pawitan. *In all likelihood*. Oxford university press 2001.

[6] B.Efron and R.Tibshirani. *An introduction to the bootsrap*. Chapman & Hall 1993.

[7] Swedish taxboard (Skatteverket). www.skatteverket.se