Utrecht University

Graduate School of Natural Sciences
Mathematical Sciences

Master's thesis

The forensic statistical analysis of incident series

Supervisors:

Prof. dr. M.J. Sjerps (NFI)
Dr. C. Spitoni (UU)

Second reader:

Prof. dr. ir. J.E. Frank (UU)

Thesis written by:

P.C.M. de Bruin BSc.

Academic year 2021 - 2022

October 24, 2022

Utrecht University

Netherlands Forensic Institute
Ministry of Justice and Security

This version is written after the official submission of the research thesis.
Some minor changes have been made.

Abstract

When a person is involved in a series of incidents, it may be used as evidence in court. An infamous case in the Netherlands is the one of the nurse Lucia de Berk, whose presence during a relatively large number of incidents in hospital wards raised suspicion. Other examples include persons being involved in a high number of fire incidents, or traffic accidents. In this research, we dive into four types of incident series and discuss the different forensic statistical methods that were proposed in the literature to quantify the strength of evidence. We will give comments and recommendations for further improvements of the proposed models. The aim is to write a recommendation to the Netherlands Forensic Institute on how to deal with a series of similar, possibly criminal, events when they are asked by parties such as the police or legal authorities to analyse the statistical evidence.

Keywords: Clusters of unusual events; incident rates; negative binomial distribution; serial crimes; naked statistical evidence; fraud; fire incidents; roster cases.

Acknowledgements

This research thesis was written as part of an internship at the Netherlands Forensic Institute (NFI). Therefore, I want to thank my daily supervisor Marjan Sjerps for giving me the opportunity to be an intern at the NFI. For suggesting this interesting research topic and the weekly meetings full of ideas. Furthermore, I am grateful for Peter Vergeer, Ivo Alberink and Leen van der Ham for making me feel part of the FSM team and the interesting discussions. I also want to mention the FBDA team and other interns at the Netherlands Forensic Institute and thank them for the pleasant and fun times at the NFI. I am thankful for my project supervisor from Utrecht University, Cristian Spitoni, for his enthusiasm, recommendations, and feedback. Also, thanks to my second reader Jason Frank for providing his helpful suggestions. Finally, I want to express my gratitude to my parents and sister for the encouragement during my master's degree and helping me when needed. I am lucky to have you to support me!

Contents

‘Als iets een keer gebeurt, dan is dat een incident. Als dat in korte tijd twee keer plaatsvindt, is dat heel toevallig. Als het drie keer gebeurt, is er sprake van een patroon. En als een vierde geval vrijwel identiek is, dan is dat ronduit verdacht.’

‘When something happens once, it is an incident. When it occurs twice in a short time, this is extremely coincidental. When it happens thrice, there is a pattern. And if a fourth event is nearly identical, then this is downright suspicious.’

Peter R. de Vries (1956 – 2021)

1 Introduction

The reader might have or might not have heard from the citation on the top of this page by Peter R. de Vries, a Dutch crime reporter that unfortunately passed away recently. The kind of phenomenon that he refers to in this quotation may be called an incident series and occurs quite often.

Suppose a person is involved in such a series of incidents and it caught the attention of the police or an insurance company. The involvement in a series of incidents may be used in court as forensic evidence. As an example, one can think about the infamous legal case in the Netherlands of Lucia de Berk, a nurse whose presence at a relatively large number of incidents in hospitals raised suspicion among her colleagues. Other examples include a person being involved in a high number of fire incidents or traffic accidents.

For this type of legal cases, it is often not clear whether there were bad intentions. Therefore, we are interested in the forensic statistical interpretation of these incident series. However, it is a very difficult problem and unfortunately in some legal cases the investigation and determination of the statistical evidence of the series of events went terribly wrong.

For example, the nurse Lucia de Berk was suspected of murdering patients in three hospital wards and would have been one of the greatest serial killers in Dutch history. Law professor Henk Elffers, with a master’s degree in mathematics and statistics, calculated that the probability of her witnessing so many supposedly incidents coincidentally was 1 in 342 million (Elffers, 2002). Though the calculations by Elffers were not used as direct evidence against Lucia de Berk in the end, she was sentenced to life imprisonment. It turned out after a long trail and many mistakes later that she was regarded as innocent. More details can be found further in this thesis and in Schneps and Colmez (2013).

Another example is the case of Sally Clark, a mother of two sons that died at a very young age (Dawid, 2001). Initially, the death of the first child was treated as a case of Sudden Infant Death Syndrome (SIDS). The second child died of similar circumstances and soon after Clark was arrested and tried for murdering both children. At trail, a professor of paediatrics testified that the probability that two babies would die of SIDS in a family like hers would be 1 in 73 million. This number was based on the argument that the probability of a single SIDS death in such a family is 1 in 8500, which was squared because of two children. However, it later turned out that the argument of independence between both deaths does not seem to hold and the research behind the probabilities was questionable (Dawid, n.d.). Sally Clark was convicted of murder, but later released from prison. More details can be found in Schneps and Colmez (2013). Hence it was another case where numbers were abused in the courtroom.

Both cases can be considered as miscarriages of justice. Therefore, it is important to dive further into this problem since clusters of events occur regularly and there is not much literature about how to face the statistical difficulties in a good manner. Often, one only looks at the incidents happening given the scenario that events occurred coincidentally and forget the scenario of there being a systematic cause to balance it with. This happened in both the cases of Lucia de Berk and the one of Sally Clark. Collecting and interpreting data correctly and defining incidents in a unequivocal way appeared to be problematic in these cases as well.

Moreover, articles that are published about this topic are often stand-alone pieces. Now, they do not include an overview of the general dilemma. However, it is worth mentioning that the Statistics and the Law Section Committee of the Royal Statistical Society is currently preparing a report on approaches to interpreting clusters of events (C. Aitken, personal communication, January 25, 2022).[1] Peter Green (personal communication, February 3, 2022) informed us that it will be a policy paper rather than a research paper.

For that reason, the aim of this research thesis is to produce a broad overview of the different types of situations involving incident series in legal cases and the various ways to interpret statistical evidence for these events. We want to discuss the strong and weak points of the methods that are used, and question whether we can apply them to other situations or even improve them. In the end, we wish to give a recommendation what the best approaches are to present this type of evidence and how we should use it in court. Thus, the following research questions can be formulated:

(i) What are the different types of situations involving incident series?

(ii) In which ways can we interpret the statistical evidence for these events?

(iii) What is the best statistical approach in each situation, and why?

To already give a preview on the results, from our research it will follow that a mixed Poisson model, that considers heterogeneity, helps to find the probability of seeing a certain series of events under the hypothesis of coincidence. The main problem is deriving a similar probability under the hypothesis of there being an underlying "systematic" cause. It will need further research.

This research thesis was written as part of an internship at the Netherlands Forensic Institute. The Netherlands Forensic Institute provides forensic analyses that play an essential part in the identification and conviction of suspects and the acquittal of innocent people (Rijksoverheid, 2014). The police and public prosecution service and the courts of the Netherlands use these analyses in solving criminal cases, but also other (inter)national organisations are interested in the work of the Netherlands Forensic Institute.

Hence this research thesis will be of interest for the Netherlands Forensic Institute, since they are regularly asked by the police and insurance companies to make a statement about the forensic statistical interpretation of different situations where someone is involved in a lot of incidents. In this way they have an overview of what has already been done, our recommendation for the best method in each situation, and they can work on their cases even more effectively.

There is a lot of discussion among jurists whether the statistical evidence in these cases, which is a form of so called naked statistical evidence, is admissible in criminal law. According to Pundik (2021), there is no principle objection to using infinitesimal probabilities of a contrasting cause, such as SIDS in the Sally Clark example, as long as it is compared with the frequency of the actual cause alleged by the prosecution (murder). In itself, the probabilities are meaningless. However, it is argued that these frequencies should not be used in court. Not (only) because of the technical and calculation concerns, but also because there is a principled problem with using statistical evidence regarding the frequency of the alleged criminal conduct among people like the accused. Inferences about human conduct rely on causal generalisations and should not be used for the purpose of determining culpability. This is because the inferences are either contradictory or misleading. In the end, the contention that a result was due to a certain cause should remain unaffected by statistical evidence of the extremely low probability of an alternative cause alone, regardless of how reliable the statistical analysis is. For more details, we refer to Pundik (2021).

The particular nature of naked statistical evidence is often illustrated by contrasting it with evidence that is not of the "naked statistical" kind, such as a witness or fingerprint (Dahlman and Pundik, 2021). It is problematic to base a verdict on naked statistical evidence alone, but it is no longer unacceptable to convict a defendant when, e.g., a fingerprint is found at a crime scene. Even though, the fingerprint evidence is also "statistical" in the sense that it is based on statistics about the prevalence of certain fingerprint patterns. There is disagreement on why naked statistical is problematic, pointing to different characteristics of naked statistical evidence as the root of the problem. For more details, see Dahlman and Pundik (2021).

We want to make clear that we are more interested in the statistical analysis behind the legal cases. Therefore, we will focus only on the statistical aspects of the general debate around naked statistical evidence.

[1] See `https://rss.org.uk/membership/rss-groups-and-committees/sections/statistics-law/`.

1.1 Synopsis

We give a brief summary what to expect in the upcoming chapters.

The Preliminary is given in Part I. Chapter 2 shows some notational conventions and gives a summary of all the probability distributions that will be used throughout this document. In Chapter 3 we discuss background knowledge that will be convenient for the statistical analyses that are presented throughout the thesis. We examine how likelihood ratios are defined and applied in the courtroom. Furthermore, confidence intervals for binomial random variables are discussed. Here we focus of the method of Wald and the method of Wilson. Then we discuss the definition of a Poisson process and relevant results. Finally, we consider Bayesian statistics and describe a few numerical methods that can be used to sample from posterior distributions.

Part II presents an overview of the different situations involving incident series and the setup of the investigation. In Chapter 4 we summarise the different characteristics of incident series and present some examples. This chapter formulates an answer to our first research question: What are the different types of situations involving incident series? Chapter 5 describes the different stages of the investigation process when we observe the events. Furthermore, it discusses what questions we like to find an answer to, from the point of view of an expert witness like the Netherlands Forensic Institute experts.

In Part III we explain the methods that are used to interpret statistical evidence in different situations found in the literature. This part attempts to give an answer to the second research question: In which ways can we interpret the statistical evidence for series of events? Chapters 6 and 7 describe the methods that are used for a situation of credit card fraud and arson incidents in one household, respectively. We give a situation description, define the model proposed in the literature and give our comments about it. This analysis is also done in Chapters 8 and 9 for a so called roster case in a hospital and arson incidents happening at multiple households in a town, respectively.

We enhance some of the before mentioned methods in Part IV. In Chapters 10 and 11 the arson and roster cases models are discussed again, and we give our ideas how to adapt and hence improve them. We give some motivation about why we want to modify the original methods and show how we should do this. In Chapter 12 we give a discussion about the improvement of the other arson incidents case.

Part V is considered to be our recommendation to the Netherlands Forensic Institute. In Chapter 13 we first give an overview of the research and in Chapter 14 we present the recommendation. It formulates the response to the question: What is the best statistical approach in each situation involving an incident series, and why? Suggestions for future research are given in Chapter 15.

In Part VI, the Appendix, one can find a few proofs of general statements, an additional chapter about a simulation study that we performed, some figures, listings and data collections. Supplementary material is provided through Github and referred to in the Appendix.[2] We end this document with the Bibliography.

[2] See `https://github.com/PCMDEBRUIN/Incident-series.git`.

Part I

Preliminary

2
Notation

2.1 Notational conventions

We have used the following notation.

Notation	Meaning
$\xrightarrow{d}$	Convergence in distribution
$\xrightarrow{\mathrm{P}}$	Convergence in probability
z_α	The upper α-quantile of the standard normal distribution
$\propto$	Proportional to
$\log(x)$	The natural logarithm of x
$\overline{G}$	Negation of G
$\lvert A\rvert$	The cardinality of the set A
$\mathcal{P}(A)$	Power set of A
I_d	The d-dimensional identity matrix
$\lfloor x \rfloor$	The greatest integer less than or equal to x (entier)
$F_{\alpha,n,m}$	The upper α-quantile of the F distribution with degrees of freedom n and m
$q_\alpha(a, f)$	The upper α percentage points of the Studentized Range Statistic

Table 2.1: Notational conventions.

2.2 Probability distributions

We discuss different discrete and continuous probability distributions that are used throughout this thesis. We assume that the reader is familiar with the different distributions that are introduced, but we still add this section to make clear what notation is used. This section is based on the definitions of Dekking et al. (2010); Weisstein (n.d.b); Cook (2008).

2.2.1 Discrete random variables

The Bernoulli distribution is used to model an experiment with only two possible outcomes, “success” and “failure”, usually encoded as 1 and 0.

Definition 2.1 (Bernoulli)**.** A discrete random variable X has a *Bernoulli distribution* with parameter

p, where $0 \leq p \leq 1$, if its probability mass function is given by

$$P(X = 1) = p, \qquad P(X = 0) = 1 - p. \tag{2.1}$$

We denote this distribution by *Ber*(p).

The binomial distribution is the distribution of the sum of n independent Bernoulli trails.

Definition 2.2 (Binomial)**.** A discrete random variable X has a *binomial distribution* with parameters n and p, where $n = 1, 2, \ldots$ and $0 \leq p \leq 1$, if its probability mass function is given by

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \qquad k = 0, 1, \ldots, n. \tag{2.2}$$

We denote this distribution by *Bin*(n, p).

The geometric distribution describes the number of failures k until the first success in a series of Bernoulli trails.

Definition 2.3 (Geometric)**.** A discrete random variable X has a *geometric distribution* with parameter p, where $0 \leq p \leq 1$, if its probability mass function is given by

$$P(X = k) = p(1-p)^k, \qquad k = 0, 1, 2, \ldots. \tag{2.3}$$

We denote this distribution by *Geo*(p).

Furthermore, the geometric distribution is a special case of the negative binomial distribution with $r = 1$. The negative binomial distribution counts the number of failures k until the rth success in a series of Bernoulli trails.

Definition 2.4 (Negative binomial)**.** A discrete random variable X has a *negative binomial distribution* with parameters r and p, where $r > 0$ and $0 \leq p \leq 1$, if its probability mass function is given by

$$P(X = k) = \binom{k + r - 1}{k} p^r (1-p)^k, \qquad k = 0, 1, 2, \ldots. \tag{2.4}$$

We denote this distribution by *NB*(r, p).

We can extend the above definition for r being any positive real number. Therefore, we have to adjust our typical definition of the binomial coefficient. We define the binomial coefficient as follows for $n, k \in \mathbb{C}$:

$$\binom{n}{k} = \frac{\Gamma(n+1)}{\Gamma(k+1)\Gamma(n-k+1)} \tag{2.5}$$

The quantity $\Gamma(\alpha)$ is the gamma function defined by

$$\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} e^{-t} \, dt. \tag{2.6}$$

When r is a strictly positive integer, the above definition coincides with the usual one for the binomial coefficient. This is because $\Gamma(\alpha + 1) = \alpha\Gamma(\alpha)$ for $\alpha > 0$ and $\Gamma(n) = (n-1)!$ for $n = 1, 2, \ldots$.

For n large and p very small, the Poisson distribution can be used as an approximation for the binomial distribution. It expresses the probability of a given number of events occurring in a fixed time interval, if these events occur with a known constant mean rate μ and independent of time since the last event.

Definition 2.5 (Poisson)**.** A discrete random variable X has a *Poisson distribution* with parameter μ, where $\mu > 0$, if its probability mass function is given by

$$P(X = k) = \frac{\mu^k}{k!} e^{-\mu}, \qquad k = 0, 1, 2, \ldots. \tag{2.7}$$

We denote this distribution by *Pois*(μ).

2.2.2 Continuous random variables

The uniform distribution describes an experiment where the outcome is completely arbitrary, except that we know that it lies between certain bounds.

Definition 2.6 (Uniform)**.** A continuous random variable has a *uniform distribution* on the interval $[a, b]$ if its probability density function f is given by $f(x) = 0$ if x is not in $[a, b]$ and

$$\pi(x) = \frac{1}{b-a}, \qquad \text{for } a \leq x \leq b. \tag{2.8}$$

We denote this distribution by $U(a, b)$.

The exponential distribution is the continuous analogue of the geometric distribution. It describes the time that it takes for a continuous process to change state.

Definition 2.7 (Exponential)**.** A continuous random variable X has an *exponential distribution* with parameter λ if its probability density function π is given by

$$\pi(x) = \begin{cases} 0 & \text{for } x < 0, \\ \lambda e^{-\lambda x} & \text{for } x \geq 0. \end{cases} \tag{2.9}$$

We denote this distribution by $Exp(\lambda)$.

For $\alpha = 1$, the exponential distribution is a special case of the gamma distribution.

Definition 2.8 (Gamma)**.** A continuous random variable X has a *gamma distribution* with parameters $\alpha > 0$ and $\beta > 0$ if its probability density function π is given by

$$\pi(x) = \begin{cases} 0 & \text{for } x < 0, \\ \dfrac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} & \text{for } x \geq 0, \end{cases} \tag{2.10}$$

where the quantity $\Gamma(\alpha)$ is the gamma function defined by (2.6):

$$\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} e^{-t}\, dt$$

We denote this distribution by $Gam(\alpha, \beta)$.

If a continuous random variable X has a $Gam(\alpha, \beta)$ distribution, then $Y = 1/X$ has an $IG(\alpha, \beta)$ distribution.

Definition 2.9 (Inverse gamma)**.** A continuous random variable X has a *inverse gamma distribution* with parameters $\alpha > 0$ and $\beta > 0$ if its probability density function π is given by

$$\pi(x) = \begin{cases} 0 & \text{for } x \leq 0, \\ \dfrac{\beta^\alpha}{\Gamma(\alpha)} x^{-\alpha-1} e^{-\beta/x} & \text{for } x > 0, \end{cases} \tag{2.11}$$

where the quantity $\Gamma(\alpha)$ is the gamma function defined by (2.6):

$$\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} e^{-t}\, dt$$

We denote this distribution by $IG(\alpha, \beta)$.

The beta distribution can be used to model random variables that are defined on a limited interval.

Definition 2.10 (Beta)**.** A continuous random variable X has a *beta distribution* with parameters $\alpha > 0$ and $\beta > 0$ if its probability density function π is given by

$$\pi(x) = \begin{cases} \dfrac{1}{B(\alpha, \beta)} x^{\alpha-1} (1-x)^{\beta-1} & \text{for } 0 < x < 1, \\ 0 & \text{otherwise,} \end{cases} \tag{2.12}$$

where the quantity $B(\alpha, \beta)$ is the beta function defined by

$$B(\alpha, \beta) = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1}\, dt. \tag{2.13}$$

We denote this distribution by $Beta(\alpha, \beta)$.

Finally, the normal distribution is one of the most important distributions, especially for the central limit theorem.

Definition 2.11 (Normal)**.** A continuous random variable X has a *normal distribution* with parameters μ and $\sigma^2 > 0$ if its probability density function π is given by

$$\pi(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \qquad -\infty < x < \infty. \tag{2.14}$$

We denote this distribution by $N(\mu, \sigma^2)$.

3
Background knowledge

We highlight some topics that will be relevant to the material discussed further in this research thesis. A Bayesian approach of deriving the evidential value of forensic evidence, confidence intervals for binomial random variables, Poisson processes and lastly Bayesian statistics and corresponding numerical methods will be discussed.

3.1 Bayesian approach

In criminal cases, multiple pieces of evidence can be found at the crime scene. These have to be evaluated by some forensic experts. The goal is to use the evidence E to cast light on the two competing propositions, H_p and H_d, before the court (Dawid, 2001). The proposition that is put forward by the prosecution is called the *prosecution hypothesis*, which we shall henceforth denote by H_p (Aitken and Taroni, 2004). The proposition that is put forward by the defence is denoted by H_d, and called the *defence hypothesis*.

The two propositions H_p and H_d may be complementary events, e.g., "guilty" versus "innocent". However, this does not always have to be the case. Suppose that we have the propositions "The suspect and one unknown person were present at the crime scene" versus "Two unknown people were present at the crime scene" (Aitken and Taroni, 2004). Then these events are not complementary, since there are many more alternative events that are not covered by the given propositions. For example, there might have been more or less than two people present at the crime scene.

3.1.1 Conditional probability and Bayes' rule

To meet our goal, we first look at the concept of conditional probabilities. Let Ω be a finite outcome space, P a probability measure and $A, B \subset \Omega$ two subsets. Then given the probability measure P, the conditional probability A given B, denoted as $P(A|B)$, is defined (if $P(B) > 0$) by

$$P(A|B) = \frac{P(A \cap B)}{P(B)}. \tag{3.1}$$

This quantity represents the new probability of A once we have learned that B occurs (Meester and Slooten, 2021, Section 1.1).

If we apply (3.1) twice, we obtain the famous *Bayes' rule*:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{3.2}$$

This rule is very significant in forensic and legal affairs, and can be rewritten in the so-called *odds form of Bayes' rule* defined by (Meester and Slooten, 2021, Section 1.1)

$$\frac{P(H_1|E)}{P(H_2|E)} = \frac{P(E|H_1)}{P(E|H_2)} \frac{P(H_1)}{P(H_2)}, \tag{3.3}$$

where H_1 and H_2 are typically unobserved events, often called *hypotheses*. The event E is typically observed and refers to certain evidence.

The left-hand side is the ratio of the probabilities of the hypotheses H_1 and H_2 after observing the evidence E, also called the *posterior odds*. The fraction at the far right is the ratio before knowing this evidence, which are the *prior odds*. Finally, the remaining fraction

$$LR_{H_1,H_2}(E) := \frac{P(E|H_1)}{P(E|H_2)} \tag{3.4}$$

is called the *likelihood ratio*. The likelihood quantifies the evidence that the data E provides for H_1 relative to H_2. Thus, we find that

$$\frac{P(H_1|E)}{P(H_2|E)} = LR_{H_1,H_2}(E) \times \frac{P(H_1)}{P(H_2)}. \tag{3.5}$$

In words, we can read (3.5) as follows:

$$\text{posterior odds} = \text{likelihood ratio} \times \text{prior odds}$$

It holds true that if the likelihood ratio is greater than one, it supports H_1. If it is smaller, it supports H_2. When the likelihood ratio equal one, it is neutral.

If H_1 and H_2 are the only two hypotheses and so each other's negation, it is easy to calculate the posterior probability $P(H_1|E)$ using the posterior odds. Namely, it holds that

$$P(H_1|E) = \frac{\text{posterior odds}}{1 + \text{posterior odds}}. \tag{3.6}$$

This can easily be verified.

3.1.2 Combined evidence

Suppose that we have two pieces of evidence, which we call E_1 and E_2. Now, we like to compute the likelihood ratio of the combined evidence E_1 and E_2, given by $LR_{H_1,H_2}(E_1 \cap E_2)$. We can write (Meester and Slooten, 2021, Subsection 2.2.1)

$$\frac{P(E_1 \cap E_2|H_1)}{P(E_1 \cap E_2|H_2)} = \frac{P(E_1|H_1)}{P(E_1|H_2)} \frac{P(E_2|E_1, H_1)}{P(E_2|E_1, H_2)}. \tag{3.7}$$

If E_1 and E_2 are *conditionally independent* given H_1 and given H_2, we have that

$$\frac{P(E_2|E_1, H_1)}{P(E_2|E_1, H_2)} = \frac{P(E_2|H_1)}{P(E_2|H_2)}.$$

Only then the multiplication of the likelihood ratios on the right-hand side of (3.7) is allowed. In this case, we obtain that

$$LR_{H_1,H_2}(E_1 \cap E_2) = LR_{H_1,H_2}(E_1) \times LR_{H_1,H_2}(E_2).$$

3.1.3 Application in the courtroom

Hence, we can conclude that the posterior odds of the prosecution hypothesis H_p and defence hypothesis H_d from the introduction of this section are given by

$$\frac{P(H_p|E)}{P(H_d|E)} = LR_{H_p,H_d}(E) \times \frac{P(H_p)}{P(H_d)}.$$

Using the information from the previous subsection, one sees that we can combine evidence E by deriving different likelihood ratios for each piece of evidence.

In court, the ultimate question that needs to be answered by the judge is whether the suspect is guilty or innocent. The Bayesian approach of establishing the evidential value of forensic evidence, like in the above, can be helpful to answer part of this question. There is a difference between the definition of guilt from the point of view of a judge and the one of a forensic expert. The forensic expert generally focuses on the underlying question whether the suspect is indeed the culprit. The judge needs to consider in his verdict whether the suspect can be held accountable. In this research thesis, we define the suspect as being guilty when he is indeed the culprit.

Now, following the Bayesian approach, it is the job of a forensic expert to derive the likelihood ratios (Sjerps, 2004). Only the jurist can evaluate the prior odds since the expert only has knowledge about the specific forensic evidence. He may only say something related to his research field, which in turn is often not connected to the prior odds. Consequently, it follows that the expert cannot give a judgement about the posterior probabilities.

The above method now describes how we can reach our previously pronounced goal: use the evidence E to cast light on the two hypotheses, H_p and H_d, before the court.

3.2 Confidence intervals

In this section, we focus on confidence intervals for binomial random variables. We can construct (asymptotic) confidence intervals for the success probability p of a $Bin(n, p)$ distribution, if only the total number of experiments n and the number of successes Y are known (Serra, 2020). We discuss two ways to obtain these confidence intervals for p, which are taken from (Serra, 2020, Subsection 3.1.2).

3.2.1 Method of Wald

The first one that we look at, is the method of Wald. Let $\hat{p}_n = Y/n$. By the central limit theorem (van der Vaart, 1998, Example 1.2) it holds that

$$\sqrt{n}\frac{\hat{p}_n - p}{\sqrt{p(1-p)}} \xrightarrow{d} N(0,1). \tag{3.8}$$

From the weak law of large numbers (van der Vaart, 1998, Example 1.1), we know that $\hat{p}_n \xrightarrow{\text{P}} p$. Hence by Slutsky's lemma (van der Vaart, 1998, Lemma 1.12) we can replace p by $\hat{p}_n$ in the denominator of (3.8). We find

$$\sqrt{n}\frac{\hat{p}_n - p}{\sqrt{\hat{p}_n(1-\hat{p}_n)}} \xrightarrow{d} N(0,1).$$

Asymptotically, we have the equality

$$P\left(-z_{\alpha/2} \leq \sqrt{n}\frac{\hat{p}_n - p}{\sqrt{\hat{p}_n(1-\hat{p}_n)}} \leq z_{\alpha/2}\right) \approx 1-\alpha.$$

If we isolate p, we obtain the (asymptotic) $100(1-\alpha)\%$ confidence interval for p. Hence[1]

$$\left[\hat{p}_n - z_{\alpha/2}\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}},\quad \hat{p}_n + z_{\alpha/2}\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}}\right].$$

The (asymptotic) $100(1-\alpha)\%$ confidence lower bound (and similarly the upper bound) for p can be found by replacing $\alpha/2$ by α in the above lower bound. We obtain the lower bound

$$p \geq \hat{p}_n - z_{\alpha}\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}}. \tag{3.9}$$

3.2.2 Method of Wilson

The first step using the central limit theorem is the same for the method of Wilson. However, we do not make use of Slutsky's lemma to replace p with $\hat{p}_n$ in (3.8). Likewise, its holds true that

$$P\left(-z_{\alpha/2} \leq \sqrt{n}\frac{\hat{p}_n - p}{\sqrt{p(1-p)}} \leq z_{\alpha/2}\right) \approx 1-\alpha.$$

[1] See Appendix A.1, Theorem A.1.

Using some algebra[2], we obtain the following endpoints for the (asymptotic) $100(1-\alpha)\%$ confidence interval for p:

$$\frac{\hat{p}_n + \frac{z_{\alpha/2}^2}{2n} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n} + \frac{z_{\alpha/2}^2}{4n^2}}}{1+\frac{z_{\alpha/2}^2}{n}}$$

In this case, the (asymptotic) $100(1-\alpha)\%$ confidence lower bound for p is given by

$$p \geq \frac{\hat{p}_n + \frac{z_{\alpha}^2}{2n} - z_{\alpha}\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n} + \frac{z_{\alpha}^2}{4n^2}}}{1+\frac{z_{\alpha}^2}{n}}. \tag{3.10}$$

An advantage of this method is that this lower and upper limit cannot exceed 0 and 1, respectively, contrary to the endpoints of Wald's confidence interval. Since $0 \leq p \leq 1$, this should indeed not be possible. Furthermore, Wilson's confidence interval is asymmetric around $\hat{p}_n$, while Wald's confidence interval is symmetric. The observed coverage probability[3] of the confidence interval of Wilson is closer to $1-\alpha$ (Newcombe, 1998). Therefore, we prefer to use the method of Wilson over the method of Wald. However, if the sample size n is large, Wald's confidence interval is a good approximation of Wilson's confidence interval (Windley, n.d.).

3.3 Poisson process

We recall some background knowledge about Poisson processes and their properties. This section is based on (Ross, 2014, Section 5.3) and Dirksen (2019a).

Definition 3.1 (Counting process)**.** A stochastic process $\{N(t) : t \geq 0\}$ taking values in $\mathbb{N} \cup \{0\}$ is said to be a *counting process* if $N(t)$ represents the total number of events that occur by time t. A counting process must satisfy:

(i) $N(t) \geq 0$.

(ii) $N(t)$ is integer valued.

(iii) If $s < t$, then $N(s) \leq N(t)$.

(iv) For $s < t$, $N(t) - N(s)$ equals the number of events that occur in the interval $(s, t]$.

For a given stochastic process $\{N(t) : t \geq 0\}$, we call $N(t) - N(s)$ the *increment* of the process over the time interval $(s, t]$.

Definition 3.2 (Stationary increments)**.** A stochastic process $\{N(t) : t \geq 0\}$ is said to possess *stationary increments* if the distribution of the number of events $N(t+s) - N(s)$ in the time interval $(s, s+t]$ only depends on t, i.e., the distribution of the time interval of any increment only depends on the length of the time interval (and not its position).

Definition 3.3 (Independent increments)**.** A stochastic process $\{N(t) : t \geq 0\}$ is said to possess *independent increments* if increments over (finitely many) disjoint intervals are independent, i.e.,

$$N(t_2) - N(t_1), N(t_3) - N(t_2), \ldots, N(t_n) - N(t_{n-1}), \qquad 0 \leq t_1 \leq t_2 \leq \cdots \leq t_n \tag{3.11}$$

are independent.

Before we turn to the definition of a Poisson process, consider the following definition.

Definition 3.4 (Interarrival times)**.** Consider a counting process and denote the time of the first event by T_1. For $n > 1$, let T_n denote the elapsed time between the $(n-1)$st and the nth event. Then the sequence $\{T_n : n = 1, 2, \ldots\}$ is called the *sequence of interarrival times.*

In the literature, Poisson processes can be defined in multiple ways. One definition can be more suited in a particular situation than another. Therefore, we will first present one and then state a theorem that includes two equivalent definitions.

[2]See Appendix A.1, Theorem A.2.

[3]The coverage probability is the probability that the constructed interval contains the true parameter (Glen, 2022).

Definition 3.5 (Poisson process)**.** The counting process $\{N(t) : t \geq 0\}$, taking values in $\mathbb{N} \cup \{0\}$, is said to be a Poisson process with rate $\lambda > 0$ if the following axioms hold:

(i) $N(0) = 0$

(ii) $\{N(t) : t \geq 0\}$ has independent increments

(iii) As $h \downarrow 0$, uniformly in t,

$$P(N(t+h) - N(t) = 1) = \lambda h + o(h)$$
$$P(N(t+h) - N(t) \geq 2) = o(h)$$

(Recall: $f(h) = o(h)$ as $h \downarrow 0$ means $\lim_{h \downarrow 0} f(h)/h = 0$.)

Theorem 3.1 (Poisson process)**.** *Let $\{N(t) : t \geq 0\}$ be a counting process taking values in $\mathbb{N} \cup \{0\}$ with $N(0) = 0$. Then the following are equivalent:*

(i) $\{N(t) : t \geq 0\}$ is a Poisson process with rate $\lambda > 0$

(ii) $\{N(t) : t \geq 0\}$ has stationary and independent increments, and for all $s > 0, t > 0, N(t+s) - N(s)$ is a Poisson random variable with parameter $t\lambda$

(iii) The interarrival times $T_n, n = 1, 2, \ldots,$ are independent identically distributed exponential random variables having parameter λ

To prove that they are indeed equivalent, one may have a look at, e.g., (Ross, 2014, Theorem 5.1), (Ross, 2014, Proposition 5.1) and (Dirksen, 2019a, Theorem 13.1).

We give the heuristics behind statement (ii) of Theorem 3.1. It says that the number of events in any interval of length t is a Poisson random variable with parameter $t\lambda$. Now, it holds that the probability mass function of the Poisson distribution gives the probability of observing k events in a time period given the length of the period and the average number of events per time unit (Koehrsen, 2019):

$$P(k \text{ events in time period}) = \frac{\left(\text{time period} \cdot \frac{\text{events}}{\text{time}}\right)^k}{k!} \cdot e^{-\text{time period} \cdot \frac{\text{events}}{\text{time}}} \tag{3.12}$$

Hence, in terms of the above defined Poisson process, the parameter $\lambda > 0$ describes the average number of events per time unit and, of course, t describes the time period we are interested in. This will give some insight on what is used later in Chapter 8.

Finally, let us look at statement (iii) of Theorem 3.1. It says that the interarrival times $T_n, n = 1, 2, \ldots,$ of the counting process $\{N(t) : t \geq 0\}$ are independent identically distributed exponential random variables with parameter λ. To prove this, it may be convenient to use the following theorem which proof can be found in (Dirksen, 2019a, Theorem 12.2).

Theorem 3.2. *A random variable X is exponentially distributed if and only if it is memoryless in the sense that*

$$P(X > s + t | X > s) = P(X > t),$$

for all $s, t \geq 0$.

3.4 Bayesian statistics

Here, we give a technical follow up to Subsection 3.1.1. We discuss important concepts regarding Bayesian statistics and in particular the Metropolis-Hasting algorithm, Gibbs sampling algorithm and Hamiltonian Monte Carlo algorithm. Therefore, we will use the contributions of Szabó and van der Vaart (2021); Hoffman and Gelman (2014); Caballero-Cárdenas (2021) to these topics. We just give the definitions and results that are needed, and refer to Szabó and van der Vaart (2021); Hoffman and Gelman (2014); Caballero-Cárdenas (2021) for more details.

3.4.1 Measure theoretic definitions

For completeness, we first consider some measure theoretic definitions. The following definition is taken from (Spreij, 2021, Definition 1.2).

Definition 3.6 (σ-algebra)**.** Let A be a non-empty set. A collection $\mathscr{X} \subset 2^A$ is called a *σ-algebra* (on A) if

(i) $A \in \mathscr{X}$;

(ii) $E \in \mathscr{X}$ implies that $E^c \in \mathscr{X}$;

(iii) $\cup_n E_n \in \mathscr{X}$ as soon as $E_n \in \mathscr{X}$ for all $n = 1, 2, \dots$.

If $\mathscr{X}$ is a σ-algebra on A, then $(A, \mathscr{X})$ is called a *measurable space* and the elements of $\mathscr{X}$ are called *measurable sets* (Spreij, 2021).

Let $\mathscr{X}$ be a σ-algebra on A and $\mu : \mathscr{X} \to [0, \infty]$ a mapping.[4] Then μ is called *countable additive*, if $\mu(\emptyset) = 0$ and if $\mu(\cup_n E_n) = \sum_n \mu(E_n)$ for every sequence (E_n) of disjoint sets of $\mathscr{X}$ (Spreij, 2021). The next definition is based on (Spreij, 2021, Definition 1.4).

Definition 3.7 (Measure)**.** Let $(A, \mathscr{X})$ be a measurable space. A countably additive mapping $\mu : \mathscr{X} \to [0, \infty]$ is called a *measure*. The triple $(A, \mathscr{X}, \mu)$ is called a *measure space*.

We call a measure *finite* if $\mu(A) < \infty$ (Spreij, 2021). It is called *σ-finite*, when we can write $A = \cup_n A_n$, where the A_n are measurable sets and $\mu(A_n) < \infty$. If $\mu(A) = 1$, then μ is called a *probability measure*.

Finally, we consider the definition for a measurable map which is taken from (Spreij, 2021, Definition 3.1).

Definition 3.8 (Measurable map)**.** A mapping $X : A \to \mathbb{R}$ is called *measurable* if $X^{-1}(A) \in \mathscr{X}$ for all $A \in \mathscr{X}$.

3.4.2 Continuous version of Bayes' rule

Now, let X and ϑ be measurable maps on a probability space, with values in measurable spaces $(\mathfrak{X}, \mathscr{X})$ and $(\Theta, \mathscr{B})$ (Szabó and van der Vaart, 2021). Here $\mathscr{X}$ and $\mathscr{B}$ are σ-algebras of subsets of the (arbitrary) sets $\mathfrak{X}$ and Θ, called the *sample space* and the *parameter space*. Then, we consider the following proposition about the continuous version of Bayes' theorem which proof can be found in (Szabó and van der Vaart, 2021, Proposition 1.8).

Proposition 3.1 (Bayes' formula)**.** *If there exists a σ-finite measure μ on the sample space $(\mathfrak{X}, \mathscr{X})$ and jointly measurable maps $(x, \theta) \mapsto p_\theta(x)$ such that $P_\theta(A) = \int_A p_\theta(x)\, d\mu(x)$, for every $A \in \mathscr{X}$, then the formula*

$$\Pi(B|x) = \frac{\int_B p_\theta(x)\, d\Pi(\theta)}{\int p_\theta(x)\, d\Pi(\theta)} \tag{3.13}$$

gives an expression for the posterior distribution.

If the probability density function of the prior distribution Π exists, we can also write (3.13) as follows:

$$\pi(\theta|x) = \frac{p_\theta(x)\pi(\theta)}{\int p_\theta(x)\pi(\theta)\, d\theta} \tag{3.14}$$

Hence we can write $\pi(\theta|x) \propto p_\theta(x)\pi(\theta)$, where the symbol $\propto$ means that $\pi(\theta|x)$ is proportional to $p_\theta(x)\pi(\theta)$ as a function of θ. This convention will become very useful to determine which parametrized family the posterior belongs to.

Despite the simplicity of Bayes' formula given by (3.13), it can be hard to apply. This is because for numerical evaluations, such as computing the posterior mean, it is required to integrate over the whole parameter space. One solution for this are simulation schemes. Examples are given in the upcoming subsection. Another possibility is using so called conjugate priors, whose definition taken from (Szabó and van der Vaart, 2021, Definition 2.1) is given below.

Definition 3.9 (Conjugacy)**.** A parametrized family $(\Pi_\alpha : \alpha \in A)$ of priors is called *conjugate* with respect to a statistical model if the posterior distribution relative to a member of the family is again a member of the family.

[4]Notice that ∞ is allowed as a value.

3.4.3 Numerical methods

As we said before, simulation schemes can be used when it is difficult to evaluate the proportionality constant in Bayes' formula. We discuss the following examples of simulation schemes: the Metropolis-Hastings algorithm, the Gibbs sampler, and Hamiltonian Monte Carlo. These are examples of so-called Markov Chain Monte Carlo methods. In the upcoming part, π and Π denote the posterior density and posterior distribution from which we wish to simulate, not the prior. Hence, we ignore the dependency on the data in the notation.

Intuition

Before we introduce the examples, we try to explain why Markov Chain Monte Carlo methods work based on (Szabó and van der Vaart, 2021, Section 3.1). The goal is to simulate from a given target density or distribution, and in the Bayesian set up this is the posterior distribution, for fixed data. To simulate from this distribution, one wants to construct a Harris recurrent Markov chain with the desired distribution as its stationary distribution.

It holds true that a Markov chain is a sequence of random variables $Y_1, Y_2, \dots$ with values in the measurable space $(\mathfrak{Y}, \mathscr{Y})$ which satisfy the Markov property. The *Markov property* says that for every measurable set B and $y_1, y_2, \dots, y_m \in \mathfrak{Y}$ (Szabó and van der Vaart, 2021):

$$P(Y_{n+1} \in B | Y_n = y_n, Y_{n-1} = y_{n-1}, \dots, Y_1 = y_1) = P(Y_{n+1} \in B | Y_n = y_n) =: Q(y, B) \tag{3.15}$$

The right-hand side of this equation is called the *transition kernel.* This kernel may be given by a *transition density* q relative to a σ-finite measure μ on $(\mathfrak{Y}, \mathscr{Y})$ in that

$$Q(y, B) = \int_B q(y, z)\, d\mu(z), \qquad B \in \mathscr{Y}.$$

Furthermore, a probability distribution is called a *stationary distribution* for Q, if for every measurable set B we have that (Szabó and van der Vaart, 2021)

$$\int_{\mathfrak{Y}} Q(y, B)\, d\Pi(y) = \Pi(B).$$

If Q allows a transition density q and Π a density π relative to some σ-finite measure μ, then this is equivalent to, for μ-almost every $z \in \mathfrak{Y}$,

$$\int_{\mathfrak{Y}} q(y, B)\pi(y)\, d\mu(y) = \pi(z).$$

For the definition of Harris recurrence, see (Szabó and van der Vaart, 2021, Section 3.1).

Then we consider the following theorem from (Szabó and van der Vaart, 2021, Theorem 3.2).

Theorem 3.3 (Ergodic theorem)**.** *If the Markov chain $Y_1, Y_2, \dots$ is Harris recurrent with stationary probability distribution Π, then for every Π-integrable f and $n \to \infty$,*

$$\frac{1}{n}\sum_{i=1}^{n} f(Y_i) \to \int f\, d\Pi, \qquad a.s.. \tag{3.16}$$

A sequence of random variables $Y_1, Y_2, \dots$ that satisfies Theorem 3.3 is called *ergodic.* It holds true that a Harris recurrent Markov chain that has a stationary distribution is ergodic, and hence by the ergodic theorem the average over $Y_1, \dots, Y_n$ converge to the target value, as $n \to \infty$. Nevertheless, one typically throws away the earlier values $Y_1, Y_2, \dots, Y_m$ and use only $Y_{m+1}, Y_{m+2}, \dots$ for sufficiently large m. This is called the "burn in".

Lastly, we have that a transition density q is said to satisfy the *detailed balance* relationship to a density π if for every $y, z \in \mathfrak{Y}$,

$$\pi(y)q(y, z) = \pi(z)q(z, y). \tag{3.17}$$

If the probability density π and transition density q satisfy (3.17), then π is a stationary distribution for the transition kernel with density q. See (Szabó and van der Vaart, 2021, Lemma 3.5). The relationship given by (3.17) is the key to constructing a transition density q for the target distribution π that is a stationary density which will be the posterior distribution that we want to sample from.

Metropolis-Hastings algorithm

The Metropolis-Hastings algorithm is a Markov chain Monte Carlo method that can be used to draw random samples from a probability distribution π from which direct sampling is difficult (Szabó and van der Vaart, 2021, Section 3.2).

Let q be a transition density such that it easy to sample from the density $z \mapsto q(y, z)$ for a given y. Then we require that the transition density q is known up to a multiplicative constant, or that it is symmetric, i.e., $q(y, z) = q(z, y)$ for all y, z. We define the *Metropolis-Hastings acceptance probability* as

$$\alpha(y, z) = \min\left\{\frac{\pi(z)q(z, y)}{\pi(y)q(y, z)}, 1\right\}. \tag{3.18}$$

For the calculation of $\alpha(y, z)$, it suffices to know both π and q up to a constant (in case of a symmetric q this is not necessary).

The *Metropolis-Hastings algorithm* proceeds, from a given initial value Y_0, recursively for $n = 0, 1, 2, \dots$ by the following steps:

```
Given $Y_n$
Generate $Z_{n+1}$ from the distribution with density $q(Y_n, \cdot)$
Generate $U_{n+1}$ from the uniform distribution on $[0,1]$
If $U_{n+1} < \alpha(Y_n, Z_{n+1})$, set $Y_{n+1} := Z_{n+1}$
else set $Y_{n+1} := Y_n$
```

In the *independent Metropolis-Hastings algorithm* the transition density q is chosen equal to $q(y, z) = g(z)$. The proposals are independent of the current states. The acceptance probability becomes

$$\alpha(y, z) = \min\left\{\frac{\pi(z)g(y)}{\pi(y)g(z)}, 1\right\}$$

Then the algorithm reduces to:

```
Given $Y_n$
Generate $Z_{n+1}$ from $g$
Generate $U_{n+1}$ from the uniform distribution on $[0,1]$
If $U_{n+1} < \alpha(Y_n, Z_{n+1})$, set $Y_{n+1} := Z_{n+1}$
else set $Y_{n+1} := Y_n$
```

The following example is taken from (Szabó and van der Vaart, 2021, Example 3.8).

Example 3.1 (Independent Metropolis-Hastings)**.** Suppose that we want to know the expected value of the variable $f(Y) = \sqrt{|Y|}$, where Y follows the density π that satisfies $\pi(y) \propto e^{-y^4}$. As a transition kernel, we can take the standard normal distribution. Given the initial value $y_0 = 0$, we iterate the following steps:

- Draw $z_{n+1} \sim N(0, 1)$ and $u_{n+1} \sim U(0, 1)$, independently.
- Set $y_{n+1} := z_{n+1}$ if

  $$u_{n+1} < -e^{y_n^4 + \frac{1}{2}z_{n+1}^2 - z_{n+1}^4 - \frac{1}{2}y_n^2},$$

  or else set $y_{n+1} := y_n$.

Then we can estimate $E(\sqrt{|Y|}$ after n iterations by $n^{-1}\sum_{i=1}^{n}\sqrt{|y_i|}$.

In the *random walk Metropolis-Hastings algorithm* we assume that the transition density q takes the form $q(y, z) = f(y - z)$. When f is symmetric around zero, then the acceptance probability reduces to

$$\alpha(y, z) = \frac{\pi(z)}{\pi(y)}.$$

The algorithm becomes:

Given Y_n
Generate Z_{n+1} from $f(\cdot - Y_n)$
Generate U_{n+1} from the uniform distribution on $[0,1]$
If $U_{n+1} < \pi(Z_{n+1})/\pi(Y_n)$, set $Y_{n+1} := Z_{n+1}$
else set $Y_{n+1} := Y_n$

The next example is taken from (Szabó and van der Vaart, 2021, Example 3.10).

Example 3.2 (Random walk Metropolis-Hastings)**.** Suppose that we want to sample from the following probability density satisfying

$$\pi(y) \propto y^2 e^{-(y-1)^2} + e^{-|y|}.$$

Then we may use the random walk Metropolis-Hastings algorithm with a normal proposal density with mean zero and variable τ^2. The parameter τ controls the step sizes of the random walk. Starting from an initial value y_0, we iterate the following steps:

- Draw $z_{n+1} \sim N(y_n, \tau^2)$ and $u_{n+1} \sim U(0,1)$, independently.
- Set

$$y_{n+1} := \begin{cases} z_{n+1} & \text{if } u_{n+1} < v_{n+1}, \\ y_n & \text{if } u_{n+1} \geq v_{n+1}, \end{cases}$$

where

$$v_{n+1} = \frac{z_{n+1}^2 e^{(z_{n+1}-1)^2} + e^{-|z_{n+1}|}}{y_n^2 e^{(y_n-1)^2} + e^{-|y_n|}}.$$

After sufficiently long burn-in, the chain $y_m, y_{m+1}, \ldots$ can be used as an approximate sample from π.

Gibbs sampling algorithm

The Gibbs sampler can be used to reduce the problem of sampling from a high-dimensional prior to sampling from lower-dimensional distributions (Szabó and van der Vaart, 2021, Section 3.4). Suppose that π is a density relative to a product measure $\mu_1 \times \cdots \times \mu_m$ on a product space and let $y = (y_1, \ldots, y_m) \in \mathfrak{Y}$. Furthermore, assume that we can generate variables from each of the conditional densities,

$$\pi_i(y_i | y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_m) = \frac{\pi(y)}{\int \pi(y)\, d\mu_i(y_i)}, \tag{3.19}$$

for $i = 1, \ldots, m$. The distributions given by the above equation are called *full conditional distributions*. It is very useful when these can be determined analytically, since then it will be easier to sample from the full conditional distributions in the algorithm.

The *Gibbs sampling algorithm* proceeds, from a given initial value $Y_0 = (Y_{0,1}, \ldots, Y_{0,m})$, recursively for $n = 0, 1, 2, \ldots$ by the following steps:

Given $Y_n = (Y_{n,1}, \ldots, Y_{n,m})$
Generate $Y_{n+1,1}$ from $\pi_1(\cdot | Y_{n,2}, \ldots, Y_{n,m})$
Generate $Y_{n+1,2}$ from $\pi_2(\cdot | Y_{n+1,1}, Y_{n,3}, \ldots, Y_{n,m})$
$\vdots$
Generate $Y_{n+1,m}$ from $\pi_m(\cdot | Y_{n+1,1}, \ldots, Y_{n+1,m-1})$

Thus, we update each of the coordinates in turn, and condition every time on the last available value of the other coordinates. The result is that the Gibbs sampler simulates from the probability density function π.

The following example is taken from (Szabó and van der Vaart, 2021, Example 3.16).

Example 3.3 (Gibbs)**.** Suppose that we are interested in the joint distribution of the pair of variables (θ, τ^2), where

$$\begin{aligned} \theta|\tau &\sim N(\mu, \tau^2), \\ \frac{1}{\tau^2} &\sim Gam(a, b). \end{aligned}$$

Then the joint distribution has density

$$\pi(\theta, \tau^2) \propto \tau^{-2a+1} e^{-\frac{(\theta-\mu)^2+2b}{2\tau^2}}, \qquad \theta \in \mathbb{R}, \tau^2 > 0.$$

The conditional density of τ^2 given θ is proportional to the above density as a function of τ^2. It follows that the conditional distribution satisfies

$$\tau^2|\theta \sim Gam\left(a + \frac{1}{2}, \frac{1}{2}(\theta - \mu)^2 + b\right).$$

Hence the Gibbs sampling algorithm to simulate from the distribution of (θ, τ^2) proceeds by the sequence of updates

$$\begin{aligned} \theta_{n+1} &\sim N(\mu, \tau_n^2), \\ \frac{1}{\tau_{n+1}^2} &\sim Gam\left(a + \frac{1}{2}, \frac{1}{2}(\theta_{n+1} - \mu)^2 + b\right). \end{aligned}$$

Finally, we want to share the following result taken from (Szabó and van der Vaart, 2021, Lemma 3.22).

Lemma 3.1. *Every coordinate update of the Gibbs sampler is a Metropolis-Hastings sampler with acceptance probability equal to one.*

Hamiltonian Monte Carlo

The Metropolis-Hastings algorithm and Gibbs sampling algorithm may take a long time to converge to the target distribution, which is due to the tendency of these methods to explore the parameter space via inefficient random walks (Hoffman and Gelman, 2014). Hamiltonian Monte Carlo is a Markov chain Monte Carlo method that avoids this random walk behaviour. Therefore, it may converge much faster than simpler methods such as the random walk Metropolis-Hastings algorithm or the Gibbs sampling algorithm from before. When model parameters are continuous rather than discrete, the Hamiltonian Monte Carlo algorithm is able to avoid the random walk behaviour by introducing an extra variable to transform the problem of sampling from a target distribution into the problem of simulating Hamiltonian dynamics.

Now, we explain how the Hamiltonian Monte Carlo method works. The upcoming definitions are based on Caballero-Cárdenas (2021). Let $\pi : \mathbb{R}^d \to \mathbb{R}$ be the probability density function that we want to sample from. Let $\theta \in \mathbb{R}^d$ be the parameter vector of π. Then we introduce an auxiliary momentum vector $r \in \mathbb{R}^d$ for the model vector $\theta \in \mathbb{R}^d$. These momentum variables are usually drawn independently from the standard normal distribution.

Then we define the *potential energy* as

$$\mathcal{L}(\theta) = -\log(\pi(\theta)),$$

and the *force* as minus the gradient of the potential energy:

$$F(\theta) = -\nabla\mathcal{L}(\theta)$$

The *kinetic energy* equals

$$\mathcal{T}(r) = \frac{1}{2} r \cdot r,$$

where $x \cdot y$ denotes the inner product of the vectors x and y. Then the *total energy* is given by $\mathcal{H} = \mathcal{T} + \mathcal{L}$. Hamilton's equations, that is the equations of motion, are

$$\frac{d}{dt}\theta = r, \qquad \frac{d}{dt} r = F(\theta).$$

Using this notation, this yields the (unnormalized) joint density

$$\pi(\theta, r) \propto e^{-\mathcal{L}(\theta) - \frac{1}{2} r \cdot r}. \tag{3.20}$$

This is called the *Boltzmann-Gibbs distribution* and is invariant by the flow of Hamilton's equations. It holds true that the marginal distribution with respect to θ is our target distribution π, and the marginal distribution with respect to r is standard normal distributed.

The Hamiltonian Monte Carlo method generates elements $(\theta_n, r_n) \in \mathbb{R}^{2d}$ that have invariant distribution $\pi(\theta, r)$. Therefore, the elements θ_n for $n \in \mathbb{N}$ will have the target distribution π as their invariant distribution.

We define the acceptance probability

$$\alpha(\theta_n, \xi_{n+1}) = \min\left\{ e^{-(\mathcal{H}(\Psi_t(\theta_n, \xi_{n+1})) - \mathcal{H}(\theta_n, \xi_{n+1}))}, 1 \right\}.$$

Now, let $t > 0$ and Ψ_t be the numerical approximation of the solution to Hamilton's equations at time t. Then the *Hamiltonian Monte Carlo algorithm* proceeds, from a given initial value $(\theta_0, r_0) \in \mathbb{R}^{2d}$, recursively for $n = 1, 2, \ldots$ by the following steps:

Given (θ_n, r_n)
Generate ξ_{n+1} from $N(0, I_d)$
Generate U_{n+1} from the uniform distribution on $[0, 1]$
Compute $\Psi_t(\theta_n, \xi_{n+1})$
If $U_{n+1} < \alpha(\theta_n, \xi_{n+1})$, set $(\theta_{n+1}, r_{n+1}) := \Psi_t(\theta_n, \xi_{n+1})$
else set $(\theta_{n+1}, r_{n+1}) := (\theta_n, -\xi_{n+1})$

The numerical integrator of the Hamiltonian Monte Carlo algorithm is called the *leapfrog function*, given by Ψ_t. Like we said before, this gives a numerical approximation of the solution to Hamilton's equations. Hence it simulates the evolution of the Hamiltonian dynamics over time. This function is (strongly) dependent on two parameters: the step size ε and the number of steps L. We can define it as follows:

$$\Psi_\varepsilon^L = \psi_{\varepsilon/2}^{(2)} \circ \underbrace{\left(\psi_\varepsilon^{(1)} \circ \psi_\varepsilon^{(2)}\right) \circ \cdots \circ \left(\psi_\varepsilon^{(1)} \circ \psi_\varepsilon^{(2)}\right)}_{L-1\,\text{steps}} \circ \psi_\varepsilon^{(1)} \circ \psi_{\varepsilon/2}^{(2)} \tag{3.21}$$

This simulates the dynamics for $\varepsilon \times L$ time steps. At every step, we have the following two updates

$$\psi_\varepsilon^{(1)}(\theta, r) = (\theta + \varepsilon r, r),$$

and

$$\psi_\varepsilon^{(2}(\theta, r) = (\theta, r + \varepsilon F(\theta)).$$

We refer to (Bou-Rabee and Sanz-Serna, 2018, Section 3.3) for more information about the leapfrog function and the Hamiltonian Monte Carlo algorithm in general.

No-U-Turn sampler

The No-U-Turn sampler is a modification of the original Hamiltonian Monte Carlo method. The aim is to refrain from specifying the parameter L that models the number of steps taken in the leapfrog function. The idea behind it is that we would like to avoid any additional computational costs. It holds that if we run additional time steps and this not results in a higher distance between the initial state θ and the proposal state $\tilde{\theta}$, then this does not lead to a better proposal (Caballero-Cárdenas, 2021).

Therefore, we need some stopping criterion when we have simulated the dynamics long enough. We can define this criterion based on the dot product between the current momentum $\tilde{r}$ and the vector from our initial position to our current position $\tilde{\theta} - \theta$ (Hoffman and Gelman, 2014). This is the derivative with respect to time (in the Hamiltonian system) of half the squared distance between θ and $\tilde{\theta}$:

$$\frac{d}{dt}\frac{1}{2}(\tilde{\theta} - \theta) \cdot (\tilde{\theta} - \theta) = (\tilde{\theta} - \theta) \cdot \frac{d}{dt}(\tilde{\theta} - \theta) = (\tilde{\theta} - \theta) \cdot \tilde{r}$$

This quantity is proportional to the progress made away from the starting point θ if the simulation were to run for an infinitesimal amount of additional time. If $(\tilde{\theta} - \theta) \cdot \tilde{r} > 0$, we keep on simulating the system's dynamics, and if $(\tilde{\theta} - \theta) \cdot \tilde{r} < 0$ we stop the simulation. For the implementation and pseudo code of the No-U-Turn sampler, we refer to Hoffman and Gelman (2014).

Part II

Problem definition

4
Situations

In the introduction we already gave some brief examples of cases involving incident series. In this chapter, we want to dive deeper into the different situations and give an overview of all types of (legal) cases related to a series of incidents. Different situations have different characteristics, and therefore it is important to know how we can distinguish between them.

A literature review on what models to apply to cases of incident series gave us an idea of what types of incident series there actually are.

4.1 Situation settings

By an incident series we mean that a lot of similar, possibly criminal, events took place in a short period of time. But how many is "a lot"? And what is a "short period"? This is not always clear. Insurance companies model the frequency of claims and based on this information have an idea what situations are peculiar. The police also keep statistics of crimes, and look if they can find crime trends in different areas. For example, during some period there were a lot of ATM robberies or car fires.

We discuss some situations and questions that we can ask about the incidents that will help us to distinguish various types of cases. It will hopefully in the end give us an idea what method to use to determine the forensic statistical evidence. We start with the question whether we are sure there was criminal intent.

4.1.1 Intent

We distinguish different possible situations:

(i) We are sure that all of the incidents that took place were intentional, and we are looking for the culprit.

(ii) We are sure that some of the incidents were intentional. We want to know for the other incidents of the series whether there was also intent. Sometimes we are still looking for a culprit.

(iii) We are not sure whether a crime took place, and we want to know if the series of incidents was a coincidence or that the events happened on purpose.

Often, it is unclear whether there were criminal intentions. Therefore, we find ourselves more frequently in the second and last situation compared to the first one.

Example 4.1 (Insulin)**.** A situation that fits description (ii), has to do with hospital incidents. Suppose that a lot of patients die unexpectedly in a short period of time. After careful forensic investigation, it is discovered that a few of the dead patients had an elevated amount of insulin in their blood while insulin was not prescribed to them. The presence of insulin, which is a naturally occurring hormone, does not immediately equate to murder (Burks, 2021). It can also indicate a medical condition. However, it is sometimes used as a murder weapon. Hence someone of the working medical staff could have injected it, and we want to know who this person is.

For the patients where the insulin was discovered, we are sure that a crime or some medical error happened: e.g., injection marks were found, and it was concluded that the patient died of this cause. However, there might be more patients that died before the investigation, for which the hospital initially thought that they died a natural death or of some other cause. Also, some of the investigated patients could have been injected by the culprit as well, but it was unobserved. So, in this case, it is certain that only some of the incidents were intentional. We want to know whether the rest was also intent or that the patients died from some other cause.

To give another example, sometimes it can be proven that fire incidents were caused by arson. Hence these kinds of situations fit description (i) or (ii), depending on the setting of the events.

The third situation suggests there are only two possible causes for an incident. However, this is not entirely true. If we say that an incident was just mere coincidence, we can mean that it happened accidentally without an underlying cause. However, it is also possible that someone is at a higher risk to experience them. E.g., someone is clumsy and induces a lot of accidents at work, or thinking about car accidents, this person is a bad driver. These are just a few of the alternative causes that fall under the incidents being a "coincidence". Being at a higher risk does not mean that someone meant to do it intentionally.

4.1.2 Definition of incident

Sometimes, we do not have a consistent definition of an incident. In these situations, it is not clear what is and what is not an incident. This idea becomes clearer if we look at an example.

Example 4.2 (Arson)**.** Suppose that someone experienced a lot of fires in his household. In some situations, it may seem clear what we mean by a fire, but in a forensic setting it needs to be defined in a very specific way. Do we focus on all the fires that a household experienced during a predetermined period, hence also small kitchen fires? Or are we only interested in the fires where the police and the fire department were involved? This definition needs to be completely clear before we further investigate the situation. It sometimes influences the evidence against the suspect.

A legal case, where the definition of an incident was not clear and led to great misunderstandings, was the case of Lucia de Berk. She was a nurse in two Dutch hospitals and was suspected of murdering her patients. This case will be discussed in Chapter 8, and we give more details on what precisely went wrong in Subsection 8.3.2.

Hence there are multiple possibilities regarding the precision of the definition. Firstly, the definition of the incident of interest is clearly stated and all events meet the criteria that are formulated. Secondly, for some of the events it is difficult to determine if it concerns an incident or not. And finally, we have situations where it is unknown whether we are dealing with incidents at all.

4.1.3 Suspect

Suppose there is no doubt that a crime took place. Then there are some possibilities about who the culprit is. Are we dealing with one person who committed the crimes all by himself? Or are there multiple who caused the incidents together? In the second case, we can again distinguish two alternatives. Firstly, we may have a fixed group of perpetrators that caused all the incidents together. The alternative suggests there are altering subgroups who each committed some crime.

Example 4.3 (Credit card fraud)**.** An example where it is certain that a crime took place, but we are not sure who the culprit is, is a case of credit card fraud in England in the early part of 2000 (Lucy and Aitken, 2004). In short, members of staff working at a retail outlet were suspected of making copies of customers credit card details and using these to draw upon those accounts.

Since none of the assistants was present during all incidents, a group of them should have done it together. Hence, we do not have only one assistant that helped each customer, and therefore there must be an alternating group of culprits. More details can be found in an (unpublished) article of Lucy and Aitken (2004) and in Chapter 6.

4.1.4 Information

Certain information about the role of the suspect in the incidents and his whereabouts can be important evidence. We can discriminate some cases:

(i) In work-related incidents, the roster data of the suspect can be made available as evidence. Hence, we know when the suspect was present or absent at incidents during working hours.

(ii) Sometimes using (police) data, it can be confirmed that the suspect was involved in the crime. For example, he was caught at security footage or registered in the police database.

(iii) We know nothing about the whereabouts of the suspect.

Example 4.4 (Roster case)**.** When roster data is available as in the first situation, the legal case is often called a *roster case*. This type of case arises in hospitals or nursing homes, where it can happen that someone was present at a lot of incidents. The legal cases of Lucia de Berk, Ben Geen and Daniela Poggiali (Fenton et al., 2022) fall into this category, and even more examples can be found in an article of Lucy and Aitken (2002). More details on the case of Lucia de Berk, can also be found in Chapter 8.

For the second situation, it is important to know how involvement in a crime is defined. Can we define involvement, e.g., as being 0 for not being involved and 1 for being involved? Or does it depend on time and distance? The first case we call *discrete* involvement, and the second one *continuous* involvement. The next situation is a nice example where time and distance play an important role.

Example 4.5 (Neighbourhood)**.** Someone can be seen in the neighbourhood prior to or after that an incident took place. If it happens many times, it may seem suspicious. "Being seen" can be defined in multiple ways, e.g., it is confirmed by witnesses or camera footage like we suggested before. Time and distance play an important role here, because how long before or after the incident does one have to be seen? And how close does one need to be to the crime scene for it to be wary?

Bolviken and Egeland (1995) described in their article a situation where a fireman was seen in the neighbourhood of 24 out of 37 cases of forest fire prior to their onset. One of the points they dive into is the concept of distance to a crime scene. We refer to the article for more details.

Situations where it is certain that someone had a more definite role in the crime, are given below.

Example 4.6 (Insurance fraud)**.** Another example when it is confirmed that someone is involved in a series of incidents, is the suspicion of insurance fraud. A person was involved a lot of accidents in a short period of time and reported the damage at his insurance company. This can be called discrete involvement since a claim is registered at the insurance company under the name of the suspect.

To illustrate an example, suppose that someone just bought an electric bike and closed an insurance policy. In a short period of time, he reports multiple claims of damage to his bike. This seems suspicious since he recently bought the bike. Did he just have bad luck, or is there more going on?

Example 4.7 (Arson)**.** The Netherlands Forensic Institute studied a few cases where someone experienced multiple fires in his household. When the police is involved, the role of people engaged in this incident is registered in the police database. As an example, such a role can be being the occupant of the residence that burned down or being the person that called the police after discovering the fire. This played an important part in the case that is discussed in Chapter 7, which was investigated by the Netherlands Forensic Institute (NFI, 2019).

4.1.5 Victim

Finally, we have to study how many victims are involved in the crime. In some of the situations, the crime only concerns a single individual like the example of arson that we discussed above. But there is also a possibility that the victim is a group of people, which is shown in the next example.

Example 4.8 (Town fires)**.** In 2013, there happened in approximately six months a series of fifteen arson incidents in a small town in the Netherlands (Alkemade, 2015). They happened very close to each other, and the inhabitants of the town were scared that they would be the next victim. The question was whether there was a serial arsonist operating in this town , and whom this possibly was. It had to be considered that there were multiple culprits. There were two suspects, A and B, that also lived in this town. More details on this situation can be found in Alkemade (2015) and are also further examined in Chapter 9.

4.2 Overview

Now that we discussed different situations, we can make an overview of all the examples of incident series and their characteristics. One can find this overview in Table 4.1. For each example that we discussed above, we look at whether it is certain that the events were intentional, if the definition of the incident is clear, how many suspects there are, what information about the case is available and how many victims are involved.

We want to note that we only integrated the characteristics of the specific examples in the overview and not characteristics of general cases. Namely, in some cases we can have more than one type of aspect that holds for a case. Therefore, we will discuss some examples that differ from the situations that we talked about before.

If we look at insurance fraud it is possible that the suspect is not one person but a family. In this case we have a fixed group of suspects instead of one person. This is also possible if we consider arson cases and suspect that a variable group committed different crimes in a certain area. See Chapter 9. Or in hospital cases there were groups of nurses worked together to murder patients.

We can also come up with alternative situations regarding intent. For example, someone is already convicted for arson, and we want to know if he is also guilty for more cases. Hence there are many situations of incident series, and they all depend on different circumstances.

	Insulin	Customer fraud	Roster cases	Arson	Neighbourhood	Insurance fraud	Town fires
Intent							
Sure (all)		■					■
Sure (partially)	■						
Not sure			■	■	■	■	
Definition of incident							
Clear (all)		■					■
Clear (partially)	■						
Not clear			■	■	■	■	
Suspect							
One	■		■	■	■	■	
Group (fixed)							
Group (variable)		■					■
Information							
Roster data	■	■	■				
Involvement (discrete)				■		■	
Involvement (continuous)					■		■
Victim							
One				■		■	
Multiple	■	■	■		■		■

Table 4.1: Overview situations involving a series of incidents.

5 Investigation process

Suppose that the police, an insurance company, or some other party observes a series of similar, possibly criminal, events. They are interested in whether it was a coincidence that many comparable incidents occurred or there is some underlying reason that these events took place. Therefore, an expert is asked by this party to give their opinion about this so-called incident series. It triggers a process of examination of the events. During the investigation that is started, there are some stages that we must go through before we can say something about this series of incidents at all.

5.1 Stages

Below we describe the underlying stages of the investigation. We will highlight the different questions that have to be asked during the analysis and the various actions that should be taken. We describe what parties are involved and what their role is in the examination. The investigation can globally be divided into the following three stages: the start of the investigation, how the evidential value of the forensic evidence is obtained and the decision of the court.

5.1.1 Start of the investigation

In the beginning of the case, we must decide whether we start an investigation in the first place. Is seeing this number of similar events a unique occasion, or is it not very peculiar to witness it? We must determine whether something than just mere coincidence is going on. Hence, do they have some systematic underlying cause? E.g., do the incidents of interest have some connection to a person?

To decide whether to proceed the examination, we must make a choice which is based on different aspects. We can have a look at the data that is included in the case, and maybe already perform a rough calculation that can give us some idea of how rare the event is. Hence at this point, the expert is already involved in the process. Thus, the most important question that we need to answer during this stage is: Do we need to start an investigation, when we observe the incident series?

After we decided that we want to further examine this incident series, it is time to collect data that will help us to establish the evidential value of the forensic evidence (the series of events). Furthermore, we must determine the explicit number of occurrences that are suspicious to us. This is quite important since it influences the definition of the evidence. Therefore, we must specify the definition of what we call an incident and hence look which events meet the criterion that we set.

As we already explained in the previous chapter, it is crucial to know what is and what is not an incident. Also, the process of data collecting must be transparent, otherwise big mistakes can be made. As we mentioned before, an example of a legal case where this went completely wrong was the case of Lucia de Berk. We will more thoroughly discuss this case in Chapter 8. She was a nurse that was suspected of murdering patients in two hospitals in the Netherlands. Some of the so defined incidents that occurred during her working hours were classified as “suspicious”, simply because she was present (Schneps and Colmez, 2013). No further suspicious incidents were included in a list of unusual events, that occurred when Lucia was not present. Incidents that were not considered suspicious at the time but had similar features. There were also inconsistencies in the timing of several incidents (Gill et al., 2018). The collected data was very biased, since witnesses knew what was looked for (and some of them may have been convinced of the guilt of the suspect) and it party relied on memory (Gill et al., 2010). It is important to be aware of the difference between scientific data gathering and criminal investigation, two research disciplines that came into conflict during this legal case.

5.1.2 Value of the evidence

Once the forensic evidence is defined and the data is collected, the expert has to come up with some method to determine the evidential value of the evidence. Based on how strong this evidence is, it has to be decided whether to prosecute the suspect. This is not the duty of the expert, but of the (Netherlands) public prosecution service. Before the determination of the evidential value of the forensic evidence, the public prosecution service plays a more neutral part in the investigation. The defence is often also interested in this stage of the investigation since they have the option to come up with counterarguments. We deduce that the main question during this stage of the investigation is: What is the evidential value of the forensic evidence (against the suspect)?

One of the goals of this thesis is to investigate different models to determine the statistical evidence of an incident series. Hence this stage of the investigation is the one that we are keen on. In literature, the evidential value is determined in several ways for various situations. A few of them are discussed in Part III about statistical evidence.

5.1.3 Decision of the court

In the end, it has to be decided what the probability is that the suspect is guilty given the forensic evidence. As we discussed in Subsection 3.1.3, the forensic expert cannot give his or her judgement about it. The judiciary is responsible for making this decision. They have to remain neutral during the whole process of the legal case, and only at the end give their verdict about whether the suspect is guilty and what the judicial penalty will be.

5.2 Role of the expert

From the above, we conclude that the expert (that ultimately gives their opinion about occurrence of the series of event) is only involved in two stages of the investigation: in the beginning when one has to decide whether to further examine the activities, and when they have to determine the value of the evidence. Experts have no say in the decision of guilt versus innocence or about the punishment of the perpetrator.

Taking this into account, we can say that dr. Frans J.M. Alkemade has a special role in comparison to other experts in the Netherlands. He is a part time senior advisor at the Dutch Ministry of Justice and writes reports for criminal cases (Alkemade, n.d.). In these reports, he writes how in his opinion the judge should reason in order to reach a verdict on guilt or innocence (M.J. Sjerps, personal communication, April 19, 2022). He uses his own methods as an example to advise the judge. Hence, we see that his role is different compared to other experts, that are only involved in the beginning of the investigation and the stage of deciding the evidential value of the forensic evidence. They have no say in the ultimate verdict. There has been a lot of debate about Alkemade's role as an advisor to the court. In this research thesis, we do not adopt a position in this discussion.

Hereinafter, we will discuss different methods that were published in the literature to interpret forensic statistical evidence involving a series of incidents. The various methods that are proposed, will give answer to the different stages that we discussed before where the expert is involved. Some can help to decide whether to start an investigation, but others explicitly determine the evidential value of the forensic evidence of interest.

The following situations will be considered in the upcoming part:

- Chapter 6: Credit card fraud.
- Chapter 7: Arson in one household.
- Chapter 8: A roster case in a hospital.
- Chapter 9: Arson at several inhabitants of a town.

Part III

Statistical evidence

6
Credit card fraud

We start with a situation regarding credit card fraud. At first glance this situation seems like a simple case, since we are almost sure that a crime took place, and we have a closed population of suspects. However, it will turn out some subtleties have to be considered.

An unpublished draft written by Lucy and Aitken (2004), henceforth also called the authors, is in circulation about situations of incident series regarding unauthorised withdrawals from credit card accounts and deaths in an intensive care unit. This article is also referred to in Meester et al. (2006). In Lucy (2006) it is explained that it was not published because of some weaknesses in the argument regarding the hospital incidents in the second part of the article. It is not mentioned why the first part about credit card fraud was not made public.

In this chapter we still want to discuss the model explained in this article, because it describes an interesting dilemma and a promising way of solving the problem. However, it will turn out that the issue needs some follow-up study to better understand the question at hand. Once more, we stress that we deal with a draft and that the authors chose not to publish it for a reason. We begin this chapter with a specification of the situation at hand.

6.1 Situation description

The upcoming scene is taken from Lucy and Aitken (2004). In the early part of 2000, there had been multiple complaints from customers of a particular retail outlet in England about unauthorised withdrawals from their credit card accounts. These unauthorised transactions had occurred distant from the retail outlet; hence it was taken into consideration that the staff members working for the company who owned the retail outlet had made copies of the customer's credit card details and used them to draw upon those accounts. Since the actual fraudulent acts had not taken place in the retail outlet made it impossible to deduce from the records alone who was responsible. Nevertheless, a full record of when each defrauded customer had visited the retail outlet and which assistants had been on duty during each visit was available.

Due to the fact that most customers had visited the outlet more than once and had been served by several of the outlet's assistants, the culprit was not uniquely identifiable. If only one assistant is responsible for the fraud, then this assistant would have served every defrauded customer at least once. However, no single assistant satisfied this condition of having served each customer at least once. However, it was an option that two or more assistants had been operating together.

This chapter illustrates the authors' method to answer the question: *'What is the value of evidence of the roster data?'* Using the record of the customer's visits and the roster data of the assistants, the aim is to compute likelihood ratios for all possible combinations of assistants given the hypotheses of a specific group of assistants being responsible for the fraud versus there being no causal dependence between the presence of the assistants and the unauthorised transactions. This way, the authors try to show which combination of assistants explains the data the best and might be responsible for all the occurrences of fraud at the retail outlet.

First, we give an overview of the assumptions that are made and then explain the model also using an example. Secondly, we discuss how the evidence can be interpreted. Finally, we give our comments about the validity of assumptions and some other remarks.

6.1.1 Definitions

Before the discussion of the assumptions, we want to note that in the draft there is no clear discrimination between the definitions of a “fraudulent transaction” and an “unauthorised transaction”. To stay close to the original, we use the same formulation as the authors.

However, in our opinion a fraudulent transaction is defined as the event that credit card details of the customers are copied by someone and used to draw upon the accounts. For unauthorised transactions, the cause is unknown.

6.1.2 Assumptions

In the article it is assumed that

- there is precisely one fraudulent transaction per defrauded customer; and
- it is equally probable that an unauthorised transaction occurred whilst any member of the subset of assistants was in attendance, under the hypothesis that the evidence was entirely coincidental; and
- customers are independent of each other.

The following notation is adapted:

- n: the number of defrauded customers who visited the retail outlet
- m: the number of dates on which the customers visited the retail outlet
- p: the number of assistants that worked at the retail outlet
- $C = \{c_1, c_2, \ldots, c_n\}$: set of defrauded customers
- $D = \{d_1, d_2, \ldots, d_m\}$: set of dates
- $A = \{a_1, a_2, \ldots, a_p\}$: set of assistants
- A_r: combination of assistants under suspicion
- q : particular transaction, assuming precisely one fraudulent transaction per customer
- c_q: particular customer who complained of an unauthorised credit card deduction
- A_{c_q}: subset of assistants that served customer c_q throughout the period of interest

As explained in the draft, the data that is used to determine the forensic statistical evidence consists of information about the customers C, the dates D and the assistants A. It is given by an $n \times m$ matrix, where the rows represent the n defrauded customers and the columns the m dates. For customer c_i that visited the retail outlet there is a coded entry indicating the assistant a_k who had served them at date d_j, for $i = 1, \ldots, n$ and $j = 1, \ldots, m$ and $k = 1, \ldots, p$. It is not necessary to present the data in this $n \times m$ matrix format. The authors explained in the article that the underlying structure of the original situation was a ragged edge array of defrauded customers, each with a list of assistants by whom they had been served. An example data set is shown in Table 6.1.

Remark. We stress that in the original article it was not mentioned whether there was information available about the time when the defrauded transactions distant from the retail outlet took place and how many unauthorised transactions there were per customer.

$$\begin{array}{c} \\ c_1 \\ c_2 \\ c_3 \\ c_4 \\ c_5 \\ c_6 \\ c_7 \\ c_8 \\ c_9 \\ c_{10} \end{array} \begin{array}{c} \begin{array}{cccccccc} d_1 & d_2 & d_3 & d_4 & d_5 & d_6 & d_7 & d_8 \end{array} \\ \begin{pmatrix} a_1 & & & & a_3 & & & \\ & a_3 & & & & a_2 & & \\ & & & a_5 & & a_2 & & \\ & & a_4 & & a_1 & & & a_2 \\ a_2 & & & & & & a_3 & \\ & a_1 & & & a_4 & & a_3 & \\ & & & a_2 & & & & \\ & & & a_1 & & & a_2 & a_5 \\ & a_5 & & & & a_2 & & \\ & & & & a_3 & & & \end{pmatrix} \end{array}$$

Table 6.1: Example matrix of combinations of dates, customers and assistants. For each date a customer had been to the retail outlet there is a coded entry indicating which assistant had served them.

6.2 Model of Lucy and Aitken

As stated before, the authors wanted to determine the evidential value of the roster data for all possible combinations of assistants given the hypotheses of a specific group of assistants being responsible for the fraud versus there being no causal dependence between the presence of the assistants and the unauthorised transactions. Therefore, we must first explicitly define the hypotheses and evidence. Then we show how the evidential value of roster data is determined.

Remark. Notice that in the upcoming subsection the evidence is ill-defined, which affects the understanding of the results. We chose to first discuss the method described in the article to determine the value of evidence of roster data. In Subsection 6.2.3 we present our ideas how the evidence should be interpreted to obtain the same results as in the article.

6.2.1 Hypotheses and evidence

The two hypotheses that are considered, are that one, or more than one, of the assistants of the retail outlet were responsible for deliberately defrauding customers versus the evidence being entirely coincidental, and instead some other cause gave rise to the unauthorised removals. The authors wrote that situations we can think of considering the second scenario are, e.g., that the apparatus for dealing with the credit cards is at fault, or that some other person, or persons, that have nothing to do with the retail outlet made fraudulent transactions.

Suppose that we consider a particular combination of assistants A_r under suspicion. Then the authors reformulate these hypotheses as follows:

- G: The subset A_r of assistants deliberately removed monies[1] from the credit card accounts of customers C.
- $\overline{G}$: The evidence is entirely coincidental, and there is no association between the defrauded customers and the assistants A_r.

The hypothesis G can also be considered as the prosecution hypothesis and $\overline{G}$ as the defence hypothesis.

It holds true that *'(...) the evidence E comprised $C = \{c_1, c_2, \dots, c_n\}$ customers, $D = \{d_1, d_2, \dots, d_m\}$ dates and $A = \{a_1, a_2, \dots, a_p\}$ assistants'* (Lucy and Aitken, 2004).

An extension to the definition of a likelihood ratio already seen in (3.4), is given by (Aitken and Taroni, 2004, Subsection 3.5.1)

$$LR_{H_1,H_2}(E|I) := \frac{P(E|H_1, I)}{P(E|H_2, I)}. \tag{6.1}$$

Hence, we condition on some extra background information I. In the draft this ratio is also called the *value of evidence.*

In the context of the credit card fraud case, one wants to determine the likelihood ratios

$$LR_{G,\overline{G}}(E|C, A_r) = \frac{P(E|G, C, A_r)}{P(E|\overline{G}, C, A_r)}, \tag{6.2}$$

[1]This sentence is transcribed from Lucy and Aitken (2004).

for all *allowed* combinations of assistants A_r under suspicion. By an allowed combination we mean that at least one assistant from subset A_r has served each defrauded customer at least once. From (6.2) it follows that the background information I includes information about who the group of assistants under suspicion A_r is and about the defrauded customers C.

In the upcoming subsection we determine the two important components of the above ratio: the probability of the evidence given that the prosecution hypothesis is true, and the probability of the evidence given that the defence hypothesis is true.

6.2.2 Value of evidence of roster data

Under the prosecution hypothesis, hence one or more than one of the assistants were responsible for deliberately defrauding customers, the authors reason there would be a strong affinity between combinations of assistants and those assistants having served the customers. Thus, under the hypothesis G that a subset A_r of assistants deliberately removed monies from the credit card accounts of customers C, the authors state that the probability of seeing the evidence is

$$P(E|G, C, A_r) = 1, \tag{6.3}$$

for all allowed combinations A_r.

Under the defence hypothesis, the evidence is entirely coincidental and instead some cause gave rise to the unauthorised removal. In the article it is said that in this case it would be expected that the appearance of the subset of assistants A_r under suspicion is entirely at random. Hence there would be no association between the defrauded customers and any assistants.

To remind us, we denoted any particular transaction by q assuming precisely one fraudulent transaction per customer and any particular customer who complained of an unauthorised credit card deduction by c_q. Then A_{c_q} denotes the group of assistants that served c_q throughout the period. Then given the scenario of the unauthorised transaction being due to a technical error or non-criminal event, one can say that it is equally probable that an unauthorised transaction occurred whilst any of the subset of assistants A_{c_q} was in attendance according to the authors.

Based on this argument, the authors conclude that

$$P(E|\overline{G}, c_q, A_{c_q}, A_r) = \frac{|A_{c_q} \cap A_r|}{|A_{c_q}|}. \tag{6.4}$$

Here $|A_{c_q} \cap A_r|$ is the number of assistants that served customer c_q who also appear in the subset A_r under suspicion.[2]

Under the defence hypothesis, there is no causal dependence between the presence of A_{c_q} and the occurrence of an unauthorised transaction. Therefore, by independence of the customers, the authors say it is true that

$$P(E|\overline{G}, C, A_{c_1}, \ldots, A_{c_n}, A_r) = \prod_{q=1}^{n} \frac{|A_{c_q} \cap A_r|}{|A_{c_q}|}. \tag{6.5}$$

Combining (6.3) and (6.5), the value of evidence (6.2) is computed for each combination of assistants A_r. If we assume that there are p assistants working at the retail outlet, then it holds true that all subsets of assistants are given by the power set $\mathcal{P}(A)$ of $A = \{a_1, \ldots, a_p\}$. We have that $|\mathcal{P}(A)| = 2^{|A|} = 2^p$. Hence there are $2^p - 1$ combinations of one or more assistants.

Now, it is important to note that we only have to look at all the possible combinations of assistants for which every customer is attended by at least one assistant from the subset. All the others do not account for every fraudulent transaction, in line with the authors.

The largest possible group, including all assistants, must be associated with a likelihood ratio of one. This is because at least one of all assistants must have served each defrauded customer at least once. Hence for all defrauded customers c_q it holds true that the probability given by (6.4) equals one. Thus, the value of evidence is neutral. In fact, if it was the only allowed combination then it would be of no evidential value according to the article.

In the article, the above method is applied to a data set of customers, dates and assistants, but this data set was not included as an appendix. Therefore, we examine the next self-conceived example, to

[2] In the original article, the authors give no further explanation for (6.4). In our opinion, the evidence is not properly defined and therefore (6.4) does not seem to hold. We elaborate in Subsection 6.2.3 how we think the evidence should be interpreted to obtain these exact results.

give a feeling how this approach works to obtain the value of evidence of roster data. The outcome of this example shows some similarities with the one given in the article.

Example 6.1. Suppose that we have $n = 10$ defrauded customers, that visited the retail outlet at $m = 8$ dates and $p = 5$ assistants worked at the retail outlet. We assume that the evidence is given by the 10×8 matrix presented in Table 6.1. We see that every customer visited the retail outlet one, two or three times and every date one, two or three assistants were on duty.

We denoted A_r by the group of assistants under suspicion. Then it holds true that there are $2^5 - 1 = 31$ possible combinations of one or more assistants that can be our culprits. Using Table 6.1 one can check that all allowed combinations A_r, that is the groups of assistants where at least one member of the subset had attended every customer at least once, are

- $A_{r_1} = \{a_2, a_3\}$,
- $A_{r_2} = \{a_1, a_2, a_3\}$,
- $A_{r_3} = \{a_2, a_3, a_4\}$,
- $A_{r_4} = \{a_2, a_3, a_5\}$,
- $A_{r_5} = \{a_1, a_2, a_3, a_4\}$,
- $A_{r_6} = \{a_1, a_2, a_3, a_5\}$,
- $A_{r_7} = \{a_2, a_3, a_4, a_5\}$,
- $A_{r_8} = \{a_1, a_2, a_3, a_4, a_5\}$.

For only these groups of assistants we will compute the likelihood ratio since all the others do not account for every fraudulent transaction.

We have that the subsets A_{c_q} of assistants that served customer c_q throughout the period of interest, for $q = 1, \ldots, 10$, are

- $A_{c_1} = \{a_1, a_3\}$,
- $A_{c_2} = \{a_2, a_3\}$,
- $A_{c_3} = \{a_2, a_5\}$,
- $A_{c_4} = \{a_1, a_2, a_4\}$,
- $A_{c_5} = \{a_2, a_3\}$,
- $A_{c_6} = \{a_1, a_3, a_4\}$,
- $A_{c_7} = \{a_2\}$,
- $A_{c_8} = \{a_1, a_2, a_5\}$,
- $A_{c_9} = \{a_2, a_5\}$,
- $A_{c_{10}} = \{a_3\}$.

They can be derived from the rows of the 10×8 matrix given in Table 6.1.

Then we know that for all possible combinations of assistants A_r it holds true, under the hypotheses that subset A_r deliberately removed monies from the credit card accounts of customers C,

$$P(E|G, C, A_{r_i}) = 1,$$

for $i = 1, \ldots, 8$. Combining (6.4) and (6.5), we obtain

$$\begin{aligned}
P(E|\overline{G}, C, A_{c_1}, \ldots, A_{c_{10}}, A_{r_1}) &= \frac{1}{216}, & P(E|\overline{G}, C, A_{c_1}, \ldots, A_{c_{10}}, A_{r_5}) &= \frac{1}{6}, \\
P(E|\overline{G}, C, A_{c_1}, \ldots, A_{c_{10}}, A_{r_2}) &= \frac{2}{27}, & P(E|\overline{G}, C, A_{c_1}, \ldots, A_{c_{10}}, A_{r_6}) &= \frac{4}{9}, \\
P(E|\overline{G}, C, A_{c_1}, \ldots, A_{c_{10}}, A_{r_3}) &= \frac{1}{54}, & P(E|\overline{G}, C, A_{c_1}, \ldots, A_{c_{10}}, A_{r_7}) &= \frac{4}{27}, \\
P(E|\overline{G}, C, A_{c_1}, \ldots, A_{c_{10}}, A_{r_4}) &= \frac{1}{54}, & P(E|\overline{G}, C, A_{c_1}, \ldots, A_{c_{10}}, A_{r_8}) &= 1.
\end{aligned}$$

The details are presented in Appendix A.5, Lemma A.1. Note that for all combinations of assistants A_r under suspicion, which do not meet the condition that at least one member of the subset had attended every customer, we have

$$P(E|\overline{G}, C, A_{c_1}, \ldots, A_{c_{10}}, A_r) = 0,$$

according to (6.4).

We find using (6.2) that[3]

$$LR_{G,\overline{G}}(E|C,\{a_2,a_3\}) = 216, \qquad LR_{G,\overline{G}}(E|C,\{a_1,a_2,a_3,a_4\}) = 6,$$
$$LR_{G,\overline{G}}(E|C,\{a_1,a_2,a_3\}) = 13.5, \qquad LR_{G,\overline{G}}(E|C,\{a_1,a_2,a_3,a_5\}) = 2.25,$$
$$LR_{G,\overline{G}}(E|C,\{a_2,a_3,a_4\}) = 54, \qquad LR_{G,\overline{G}}(E|C,\{a_2,a_3,a_4,a_5\}) = 6.75,$$
$$LR_{G,\overline{G}}(E|C,\{a_2,a_3,a_5\}) = 54, \qquad LR_{G,\overline{G}}(E|C,\{a_1,a_2,a_3,a_4,a_5\}) = 1.$$

From the above, we can conclude that following the model mentioned in the article the combination of assistants a_2 and a_3 explains the evidence E the best. Looking at the remaining likelihood ratios, it seems likely that assistant a_1 is innocent. The presence of assistants a_4 and a_5 leads to higher likelihood ratios compared to others if we have a group of three or four suspects.

It holds true that the highest likelihood ratio is associated with the smallest group of two assistants. From the allowed combinations $A_{r_1},\ldots,A_{r_8}$, we derive that there are no combinations of assistants which do not contain this "core group".

6.2.3 Interpretation of the evidence

It was stated in the original article that the evidence E comprised $C = \{c_1, c_2, \ldots, c_n\}$ customers, $D = \{d_1, d_2, \ldots, d_m\}$ dates and $A = \{a_1, a_2, \ldots, a_p\}$ assistants. Hence to our view the evidence is vaguely formulated. They seem to refer to the record of when each defrauded customer had visited the retail outlet and which assistants had been on duty during each visit. However, to obtain the exact probabilities (6.3), (6.4) and (6.5) from the article the evidence E needs to be redefined. We will do this below. Unfortunately, the evidence needs to include a statement about what one wants to find out. This cannot be done. Therefore, we will reformulate the evidence again at the end of this subsection.

To find the desired probabilities, the evidence needs to be defined in the following way:

- E: One, or more than one, of the assistants in A_r served the customers C when they visited the retail outlet, and as a result of these visits the unauthorised transactions took place.

Indeed, given that the assistants A_r were responsible for deliberately defrauding the customers, the probability of E given by (6.3) is equal to one. This is because the assistants must have served the customers in order to obtain their credit card details. For other probability, let us first consider one defrauded customer c_q. It is assumed that, under the defence hypothesis, it is equally probable that an unauthorised transaction occurred whilst any member of the subset of assistants A_{c_q} was in attendance.[4] So given the hypothesis that the evidence was coincidental, and assuming that an unauthorised transaction took place after one or more than one of the assistants in A_r helped customer c_q, each assistant from A_{c_q} has a probability of

$$\frac{1}{|A_{c_q}|} \tag{6.6}$$

that by chance an unauthorised transaction happened when they helped customer c_q. Then the probability of E, specified for one defrauded customer c_q, is equal to the sum of the probabilities (6.6) for each of the assistants from A_{c_q} that are also in A_r. Hence

$$\frac{|A_{c_q} \cap A_r|}{|A_{c_q}|},$$

in line with (6.4). Finally, to obtain (6.5) one needs to multiply the probabilities (6.4) for all customers C under the assumption that customers are independent.

In the above defined evidence E, we assume that a result of the visits of the customers to the retail outlet the unauthorised transactions took place. However, this is what we want to find out. Therefore, E should be rephrased as

- E': One, or more than one, of the assistants in A_r served the customers C when they visited the retail outlet.

[3] Again, we emphasise that we do not report the likelihood ratios of the other possible combinations of assistants.
[4] Note that the possible differences between the number of shifts of the assistants is not taken into account.

Then (6.3) still holds. But (6.4) is not true anymore since it will depend on the attendance data of all assistants working at the retail outlet on the days that the defrauded customers visited. If we focus on one customer c_q, then we can determine what the probability is that none of the assistants in A_r served the customer c_q when he visited the outlet using this attendance data. One minus this probability is the probability that one, or more than one, of the assistants in A_r indeed served the customer c_q, under the hypothesis there is no causal dependence between the presence of A_r and the occurrence of the unauthorised transaction. By independence, we can multiply the probabilities for all customers C.

6.3 Comments

We discuss the validity of assumptions and other remarks about the method described in Section 6.2 to find the value of evidence of roster data.

6.3.1 Validity of assumptions

To remind us, it was assumed in the article that

- there is precisely one fraudulent transaction per customer; and
- it is equally probable that an unauthorised transaction occurred whilst any member of the subset of assistants was in attendance, under the hypothesis that the evidence was entirely coincidental; and
- customers are independent of each other.

The first one, the assumption of having precisely one fraudulent transaction per customer, is a strong assumption and does not necessarily have to be true. Still, we have no information about how many unauthorised withdrawals took place per customer and on what date(s). If there is a customer that experienced, e.g., two unauthorised withdrawals then it is important to know when they precisely took place and hence take this information into account for our calculations. However, it is quite complex to integrate this into the model. Hence, this assumption was presumably made to make the calculations easier but may not be realistic.

The second assumption is reasonable, only if we assume that the evidence E is formulated as in Subsection 6.2.3. Hence the evidence is defined as the event that one, or more than one, of the assistants in A_r served the customers C when they visited the retail outlet, and because of these visits the unauthorised transactions took place. Then under $\overline{G}$, either assistant could have been present during the visits of the customers. Hence they all have equal probability to have been present, which is the basis of equation (6.4). If the evidence is not formulated in this way and we assume that the evidence equals the event that one, or more than one, of the assistants in A_r served the customers C when they visited the retail outlet, then it is not a reasonable assumption. It would depend on the attendance of all assistants.

Finally, to our knowledge the customers have nothing to do with each other and hence independently visited the retail outlet. Therefore, the last assumption which is used in (6.5) seems sensible. However, under the hypothesis of coincidence we reasoned that the unauthorised removals must have another cause. For example, there is a defect apparatus for dealing with the credit cards. It might induce dependence.

6.3.2 Other remarks

At a glance, the method described in Section 6.2 seems like a simple and insightful method to determine the value of evidence of roster data. However, in the original article the evidence was vaguely formulated. To find the same results as in the draft one should have defined the evidence differently. If we reformulate it, as we did in Subsection 6.2.3, it should include the statement that the unauthorised transactions happened because of the visits of the customers to the retail outlet, which is exactly what we wanted to know. Hence what can be concluded from this ill-defined evidence is that one needs to be very precise in formulation.

In Subsection 6.2.3 we already gave an advise how to improve the method explained in the article. If the evidence is defined as the event that one, or more than one, of the assistants under suspicion served the defrauded customers during their visits at the retail outlet, then for the calculation of the probability (6.4) one needs information about the attendance of all of the assistants that worked at the outlet. Unfortunately, we do not have access to such information and therefore cannot perform these suggested calculations. Consequently, this problem needs to be investigated more thoroughly.

It holds true that for a likelihood ratio as defined in (6.1), we need to condition both probabilities in the numerator and denominator on the same events. When we have a close look at the probabilities (6.3) and (6.5), then the first one is not conditioned on the subsets $A_{c_1}, \ldots, A_{c_n}$ while the second one is. The value of evidence is computed as the ratio of these two probabilities. However, it does not equal the likelihood ratio given in (6.2). Hence something went wrong here as well. To correct it we need to, for example, condition the probability (6.3) on the subsets $A_{c_1}, \ldots, A_{c_n}$ or in some way leave this information out.

Finally, assume that the method mentioned in the article was indeed correct. Then it holds that when a situation is more complex, e.g., there are a lot of assistants and customers, it becomes more difficult to perform the calculations by hand like we did in Example 6.1. It would make sense to implement this method using a programming language of choice. To give an example why it is reasonable to use an implementation of the model, in the original article they considered $p = 11$ assistants which led to $2^{11} - 1 = 2047$ possible combinations of one or more assistants. Then we need to decide what combinations are allowed, as at least one assistant of the subset needs to have attended every customer at least once. It followed that there were 256 possibilities. Even though the total number of customers and dates were not mentioned in the article, it already gives an indication how complex the computations become if we perform them unstructured. It holds for either the original model, or our proposed improvement.

We chose not to investigate this option of implementation further, because we wanted to explore other topics as well. Moreover, we did not have access to a large data set and found it more beneficial to only examine a simple example. Yet we encourage the reader to look further into it.

7
Arson

In this chapter, we consider a legal case where someone is suspected of arson. We can distinguish two types of legal cases. First, someone is seen at multiple scenes of fire prior to their onset, and we wonder whether this person was the arsonist. On the other hand, someone can also be involved in a fire in different way. His or her properties are burned down. If it happens frequently, we may ask ourselves whether it was coincidence or not. In this chapter, we dive into the latter problem. The first case is modelled in the article of Bolviken and Egeland (1995).

We give a description of a legal case where there was suspicion of an arson series, and the model that was used to obtain statistical evidence. After this, we comment on the model and discuss the strong and weak points of it.

7.1 Situation description

We give an example of the legal case of our interest, that is based on a report of the Netherlands Forensic Institute, described in NFI (2019). Someone was suspected of a series of seven fires over a period of four years. At four of the fires, he was registered in the so called BVH police database of the Netherlands as occupant of the burned down residence. These four fires happened over a period of two years. At the rest of the fires, he was registered as the reporter of the fire, or he was merely involved, or he was the suspect.

Now, we may wonder whether it was a coincidence that this person was involved in so many fires. Therefore, the Netherlands Forensic Institute was asked to determine the probability that the fires were just mere coincidence or that there was a systematic cause, given that the suspect had a role (reporter, involved, suspect) in the seven fires that took place over a period of four years. Their focus was to look at the evidence that in two years time, the suspect was involved in at least four fires.

In this chapter, we want to show the model that was used to investigate this legal case. This method was originally invented by the Netherlands Forensic Institute and also described in their report (NFI, 2019). First, we give the definitions and assumptions that are used to determine the desired probabilities. Then we explain the model. Our goal is to determine the probability that at least i fires that occurred had a systematic cause, given that suspect was involved in at least four fires in two years, for $i = 1, \dots, 4$.

7.1.1 Definitions

We only distinguish two types of fires: *accidental* fires and *systematic* fires. According to NFI (2019), an accidental fire can be understood as a fire that was started by a cause that does not significantly decrease or increase the chance that we experience a second fire. These types of fires are for example caused by short-circuit, or a device that is subsequently replaced or lightning strike.

A systematic fire is defined as a non-accidental fire. They have a systematic cause in a particular household. E.g., arson by a person that is related to the suspect or arson by the suspect himself. In this case, the arsonist has a connection to the suspect. Note that there are also situations where fires did have a systematic cause, but there was no question of arson.

7.1.2 Assumptions

We assume that (NFI, 2019)

- accidental fires are independent of systematic fires, and are independent of each other; and
- without any further information, accidental fires are uniformly distributed among all households. Hence, households are indistinguishable.

Additionally, we adapt the following notation:

- N: the number of households in the Netherlands where there may occur a fire
- K: the number of fires in the Netherlands in two years
- K_1: the number of fires with an accidental cause in the Netherlands
- K_2: the number of fires with a systematic cause in the Netherlands
- X: the number of households where there have been at least four fires

Per definition, it holds that $K = K_1 + K_2$. We assume that the quantities N and K are known. A lower bound for X is enough. The number of accidental fires K_1 and systematic fires K_2 are unknown. Finally, an ad hoc maximum for the possible number of fires in a household is needed.

We focus on a specific time period (2013 - 2019). From the report of the Netherlands Forensic Institute (NFI, 2019) we know that[1]

$$
\begin{aligned}
N &\approx 7.7 \times 10^6, \\
K &\approx 52000, \\
X &\geq 10.
\end{aligned}
$$

The ad hoc maximum for the possible number of fires in a household is 20.

7.2 Model of the Netherlands Forensic Institute

The idea behind the model explained below, is that it is quite rare that someone had a series of four fires in his household over a period of two years. We wonder what the underlying cause may be. Is it merely a coincidence, or is there a systematic cause? Therefore, we define the two types of fires as above and characterise the hypotheses and evidence as follows.

7.2.1 Hypotheses and evidence

We will apply the Bayesian approach that is explained in Section 3.1. We define for $i = 1, \ldots, 4$ the following two hypotheses (NFI, 2019):

- $H_{1,i}$: At least i of the fires in the household of the suspect had a systematic cause
- $H_{2,i}$: Less than i of the fires in the household of the suspect had a systematic cause

Our evidence is, as was mentioned before,

- E: In two years time, the suspect was involved in at least four fires.

[1] The number of fires (BVH registered) in the Netherlands in one year is about 26000. Hence, we have approximately $2 \times 26000 = 52000$ fires in two years.

To remind us, our goal is to determine the posterior probability that at least i fires had a systematic cause, given that suspect was involved in at least four fires in two years, for $i = 1, \ldots, 4$. Therefore, we want to determine the ratio (NFI, 2019)

$$\begin{aligned}\frac{P(H_{1,i}|E)}{P(H_{2,i}|E)} &= \frac{P(H_{1,i} \cap E)}{P(H_{2,i} \cap E)} \\ &= \frac{P(\text{at least 4 fires, at least } i \text{ systematic cause})}{P(\text{at least 4 fires, less than } i \text{ systematic cause})} \\ &:= \frac{q_1}{q_2}.\end{aligned}$$

Once we obtained these posterior odds, we can easily determine the posterior probability $P(H_{1,i}|E)$ for $i = 1, \ldots, 4$ using the method described in Subsection 3.1.1. It turns out that we can only obtain a lower bound for the posterior probability. In the upcoming subsections, we are going to find a lower bound for the probability $p := q_1 + q_2$, and an upper bound for the probability q_2.

7.2.2 Probability of having at least 4 fires

Notice that

$$\begin{aligned}q_1 + q_2 &= P(\text{at least 4 fires, at least } i \text{ systematic cause}) + P(\text{at least 4 fires, less than } i \text{ systematic cause}) \\ &= P(\text{at least 4 fires}) \\ &=: p.\end{aligned}$$

Hence p is the probability of having at least four fires in a household. We denoted the number of households where there have been at least four fires by X and assumed that without further information households are indistinguishable. Hence clusters of fires are uniformly distributed among households. Therefore, we model X as a binomial distributed random variable with parameters N and p. Using the method of Wilson, we can compute a 95% confidence lower bound for the probability p, given that $X \geq 10$. As described in Subsection 3.2.2, we prefer this method compared to the method of Wald, since $0 \leq p \leq 1$ and the lower bound of Wilson's method is always nonnegative.

In our example, we take the number of successes, that is the number of households that experienced at least four fires, equal to $Y = 10$. It holds that $n = N$. Hence, if we plug $\hat{p}_n = 10/N$ in (3.10), we obtain that

$$p := q_1 + q_2 \geq 7.764 \times 10^{-7}. \tag{7.1}$$

Remark. For sake of comparison, if we use the method of Wald to obtain a 95% confidence lower bound, we find that

$$p := q_1 + q_2 \geq 6.231 \times 10^{-7}.$$

Hence Wilson's lower bound is slightly sharper.

7.2.3 Probability of having j accidental fires

Now, we are going to compute the probability that in an arbitrary household we have at least four fires, from which j have an accidental cause. Since we assumed that accidental fires are uniformly distributed among all households in the Netherlands, the number of accidental fires in a household follow a binomial distribution with parameters K_1 and $1/N$. Because fires with an accidental cause are independent of those with a systematic cause, we have the equality (NFI, 2019)

$$\begin{aligned}P(\text{at least 4 fires}, j \text{ accidental cause}) &= P(j \text{ accidental cause, at least } 4 - j \text{ systematic cause}) \\ &= P(j \text{ accidental cause}) \times P(\text{at least } 4 - j \text{ systematic cause}) \\ &= \binom{K_1}{j} \left(\frac{1}{N}\right)^j \left(1 - \frac{1}{N}\right)^{K_1 - j} \times P(\text{at least } 4 - j \text{ systematic cause}).\end{aligned}$$

So we only have to determine the last probability.

We look at the probability of having at least k fires with systematic cause in a household, for $k = 1, 2, \ldots$. There are K_2 systematic fires, which are distributed over N households in the Netherlands.

They are distributed according to a certain partition, which we call $\mathcal{P} = \{p_1, \ldots, p_N\}$. By definition, it holds that

$$\sum_{x=1}^{N} p_x = K_2.$$

Let A be the number of households having at least k systematic fires. It equals the number of elements in the set $\{p_x : p_x \geq k,\, x = 1, \ldots, N\}$. We explain why A is at most equal to $\lfloor K_2/k \rfloor$. Because we are interested in the number of households having at least k systematic fires, we can make clusters of k fires and distribute those uniformly among the N households. The total number of clusters we have in this case equals $\lfloor K_2/k \rfloor$ since the total number of systematic fires was K_2. We round down, because none of our chosen households can have strictly less than k systematic fires. Thus, it is possible that all chosen households have k systematic fires, but it can also be the case that some have strictly more than k fires. According to this distribution most households will have at least k fires compared to other possible distributions of the fires. Notice that it is the worst-case scenario. Hence now we have an upper bound for the number of households having at least k systematic fires, which is $A \leq \lfloor K_2/k \rfloor$.

The total number of households in the Netherlands is N. Given that the clusters of at least k fires are uniformly distributed among the households, the probability that one household has at least k systematic fires is $1/N$. So we find that the probability of having at least k fires with systematic cause in a random household, for $k = 1, 2, \ldots$, is bounded from above by (NFI, 2019)

$$P(\text{at least}\, k\, \text{systematic cause}) = \frac{A}{N} \leq \frac{\lfloor \frac{K_2}{k} \rfloor}{N}.$$

Combining the above results, we conclude that the probability of having at least four fires in a household, from which j have an accidental cause, is

$$P(\text{at least}\, 4\, \text{fires}, j\, \text{accidental cause}) \leq \binom{K_1}{j} \left(\frac{1}{N}\right)^j \left(1 - \frac{1}{N}\right)^{K_1 - j} \frac{\lfloor \frac{K_2}{4-j} \rfloor}{N}, \tag{7.2}$$

for $j = 0, \ldots, 3$. From the above, we see this result does not hold for $j \geq 4$. Therefore, for $j \geq 4$ we use the following inequality[2]:

$$\begin{aligned} P(\text{at least}\, 4\, \text{fires}, j\, \text{accidental cause}) &\leq P(j\, \text{accidental cause}) \\ &= \binom{K_1}{j} \left(\frac{1}{N}\right)^j \left(1 - \frac{1}{N}\right)^{K_1 - j} \end{aligned} \tag{7.3}$$

7.2.4 Posterior probability of having at least i systematic fires

We are almost there. We only need to find an upper bound for the probability q_2 of having at least four fires in a household, from which less than i have a systematic cause, where $i = 1, \ldots, 4$. Therefore, we are going to use the results from the previous subsection. First, we observe that the set {"at least 4 fires, less than i systematic cause"} is contained in the set {"at least 4 fires, more than $4 - i$ accidental cause"}.[3] Beforehand, we assumed that the ad hoc maximum for the possible number of fires in a household is 20. Consequently, using (7.2) and (7.3) we find that (NFI, 2019)

$$\begin{aligned} q_2 &= P(\text{at least}\, 4\, \text{fires, less than}\, i\, \text{systematic cause}) \\ &\leq P(\text{at least}\, 4\, \text{fires, more than}\, 4 - i\, \text{accidental cause}) \\ &= \sum_{j=5-i}^{20} P(\text{at least}\, 4\, \text{fires}, j\, \text{accidental cause}) \\ &\leq \sum_{j=5-i}^{3} \binom{K_1}{j} \left(\frac{1}{N}\right)^j \left(1 - \frac{1}{N}\right)^{K_1 - j} \frac{\lfloor \frac{K_2}{4-j} \rfloor}{N} + \sum_{j=4}^{20} \binom{K_1}{j} \left(\frac{1}{N}\right)^j \left(1 - \frac{1}{N}\right)^{K_1 - j}. \end{aligned}$$

[2] It holds that $P(A \cap B) \leq \min\{P(A), P(B)\}$.

[3] We explain why this statement holds. Let k be the total number of fires that the suspect experienced in his household. Let k_1 and k_2 be the number of accidental and systematic fires, respectively. Then we have that $k = k_1 + k_2$. Now, if $k_2 < i$ then it follows that $k - k_1 < i$. This implies that $k_1 > k - i$. Since $k \geq 4$, we obtain $k_1 > 4 - i$ as desired.

Since the number of accidental fires K_1 and systematic fires K_2 are unknown, we must look for an upper bound for the last expression. What we do know, is that $K_2 = K - K_1$. If we plug this equation into the above equation, it yields

$$q_2 \leq \sum_{j=5-i}^{3} \binom{K_1}{j} \left(\frac{1}{N}\right)^j \left(1 - \frac{1}{N}\right)^{K_1 - j} \frac{\lfloor \frac{K-K_1}{4-j} \rfloor}{N} + \sum_{j=4}^{20} \binom{K_1}{j} \left(\frac{1}{N}\right)^j \left(1 - \frac{1}{N}\right)^{K_1 - j}. \tag{7.4}$$

We plot this number for $i = 1, \ldots, 4$ and all values of K_1 . The results can be found in Appendix C.1, Figure C.1. From the figures, we derive the following upper bounds for (7.4):

$$\begin{aligned} i = 1: \quad q_2 &\leq \quad 8.618 \times 10^{-11} \\ i = 2: \quad q_2 &\leq \quad 8.619 \times 10^{-11} \\ i = 3: \quad q_2 &\leq \quad 1.141 \times 10^{-8} \\ i = 4: \quad q_2 &\leq \quad 3.797 \times 10^{-6} \end{aligned}$$

Using the above result and (7.1), we are going to bound the probability $q_1 := p - q_2$ of having at least four fires in a household, from which at least i have a systematic cause, where $i = 1, \ldots, 4$. Now, observe that we cannot derive a lower bound for the case $i = 4$. It becomes negative, which does not make sense since we talk about probabilities. For the other cases, we find the following results:

$$\begin{aligned} i = 1: \quad q_1 &\geq \quad 7.764 \times 10^{-7} - 8.618 \times 10^{-11} &= \quad 7.763 \times 10^{-7} \\ i = 2: \quad q_1 &\geq \quad 7.764 \times 10^{-7} - 8.619 \times 10^{-11} &= \quad 7.763 \times 10^{-7} \\ i = 3: \quad q_1 &\geq \quad 7.764 \times 10^{-7} - 1.141 \times 10^{-8} &= \quad 7.650 \times 10^{-7} \end{aligned}$$

Finally, we are going to derive the posterior odds and compute the probability that at least i of the fires in the household of the suspect had a systematic cause, given that in two years time the suspect was involved in at least four fires, for $i = 1, \ldots, 4$. Unfortunately, we cannot derive this probability for the case $i = 4$ as there is no reasonable upper bound for q_1. We discuss the results for the cases $i = 1, 2, 3$ one by one.

Using (3.6), for $i = 1$ we find that

$$\begin{aligned} \frac{P(H_{1,1}|E)}{P(H_{2,1}|E)} = \frac{q_1}{q_2} \geq \frac{7.763 \times 10^{-7}}{8.618 \times 10^{-11}} &\approx 9007.89, \\ P(H_{1,1}|E) \geq \frac{9007.89}{1 + 9007.89} &\approx 0.999. \end{aligned}$$

For $i = 2$ we derive that

$$\begin{aligned} \frac{P(H_{1,2}|E)}{P(H_{2,2}|E)} = \frac{q_1}{q_2} \geq \frac{7.763 \times 10^{-7}}{8.619 \times 10^{-11}} &\approx 9006.85, \\ P(H_{1,2}|E) \geq \frac{9006.85}{1 + 9006.85} &\approx 0.999. \end{aligned}$$

Lastly, for $i = 3$ we have that

$$\begin{aligned} \frac{P(H_{1,3}|E)}{P(H_{2,3}|E)} = \frac{q_1}{q_2} \geq \frac{7.650 \times 10^{-7}}{1.141 \times 10^{-8}} &\approx 67.05, \\ P(H_{1,3}|E) \geq \frac{67.05}{1 + 67.05} &\approx 0.985. \end{aligned}$$

To summarise, we found that

$$\begin{aligned} P(\text{at least one of the fires in the household of the suspect had a systematic cause}|E) &\geq 0.999, \\ P(\text{at least two of the fires in the household of the suspect had a systematic cause}|E) &\geq 0.999, \\ P(\text{at least three of the fires in the household of the suspect had a systematic cause}|E) &\geq 0.985. \end{aligned}$$

The probability that at least four of the fires in the household of the suspect had a systematic cause, given that in two years time the suspect was involved in at least four fires could not be derived.

Number of systematic fires	**Probability at least ($N \approx 7.7 \times 10^{-6}$)**	**Probability at least ($N/2 \approx 3.85 \times 10^{-6}$)**
At least one	0.9998889853	0.9982333319
At least two	0.9998889768	0.9982331963
At least three	0.9853063783	0.8824535354
At least four	0	0

Table 7.1: Comparison $P(H_{1,i}|E)$ for different values of N, for $i = 1, \ldots, 4$.

7.3 Comments

Now, we give our comments on this model that was used to determine the probability that some of the fires in the household of the suspect had a systematic cause, given that in two years time the suspect was involved in at least four fires. First, we look at the validity of the assumptions, and then perform a sensitivity analysis on two of the parameters. Finally, we highlight some other points that we think are important to discuss.

7.3.1 Validity of assumptions

To remind us, our two most important assumptions were that

- accidental fires are independent of systematic fires, and are independent of each other; and
- without any further information, accidental fires are uniformly distributed among all households. Households are indistinguishable.

In our opinion, the first assumption is reasonable. From the definition of an accidental fire, it follows that these types are indeed independent of each other. Since accidental and systematic fires depends on different causes, it is acceptable to assume that they are also independent.

The assumption about indistinguishability is questionable. Naturally, some households are more vulnerable to fire than others. E.g., people with physical or mental disabilities and elderly, but also people that smoke (Brandweeracademie, 2019). However, we indeed have no information about what households are more at risk. In the next subsection we therefore carry out a sensitivity analysis on the number of households in the Netherlands where there may occur a fire, which was denoted by the parameter N.

7.3.2 Sensitivity analysis

Suppose that half of the households in the Netherlands are not at risk and the other half are at full risk. So we only look at half of the population, and our parameter becomes $N/2$ instead of N. We again calculate the posterior probability that at least i of the fires in the household of the suspect had a systematic cause, for $i = 1, \ldots, 4$, and compare it to the findings from before. The results can be found in Table 7.1, where the zeros denote that some of the probabilities could not be determined. This time we did not round the probabilities.

From Table 7.1, we can conclude that adjusting the parameter leads to smaller probabilities. Nevertheless, for $i = 1$ and $i = 2$ the difference is minimal. The case for $i = 3$ shows the greatest distinction, i.e., it shifts from approximately 98.5% to 88.2%.

However, we can ask ourselves if this sensitivity analysis is enough to illustrate that considering differences between households does not significantly influence our results. Therefore, we maybe want to look at an adapted or even an alternative model that (better) integrates the idea of heterogeneity between households. See Chapter 10.

Besides looking at the number of households in the Netherlands, we can also consider performing a sensitivity analysis on the number of fires in two years K. The publication of Verbond Van Verzekeraars (2019) tells us that the total number of fire claims in one year is actually higher (approximately 95000 per year) than the 26000 fires that was given in the report of the Netherlands Forensic Institute (NFI, 2019).[4]

[4]The number of fires in the report of the Netherlands Forensic Institute only includes those that were registered in the BVH police database. Therefore, the 26000 fires in one year do not include small fires where the police were not involved. For example, (small) kitchen fires or improper use of a fire pit.

Number of systematic fires	**Probability at least** ($K \approx 52000$)	**Probability at least** ($2K \approx 104000$)	**Probability at least** ($4K \approx 208000$)
At least one	0.9998889853	0.9982331629	0.9720328691
At least two	0.9998889768	0.9982330951	0.9720323341
At least three	0.9853063783	0.8824518275	0.0596875300
At least four	0	0	0

Table 7.2: Comparison $P(H_{1,i}|E)$ for different values of K, for $i = 1, \ldots, 4$.

The number of fires per year was approximately 95000 during the period of 2015 to 2018 according to Verbond Van Verzekeraars (2019). Therefore, we look at what happens if the number of fires in the Netherlands in two years K is two times and even four times greater. The result is shown in Table 7.2.

As one sees, the result in the last column is a quite surprising. When the number of fires in two years equals 208000, the probability that at least three of the fires in the household of the suspect had a systematic cause becomes almost 6%. It a clear difference with the probabilities for the cases that the number of fires in two years equals 104000 (88.2%) or our original number 52000 (98.5%). We conclude that adjusting the parameter K leads to smaller, and in some cases almost negligible, probabilities.

7.3.3 Other remarks

To obtain our probability of interest, we used the posterior odds without looking at the likelihood ratio and prior odds. It is different from the way that we described in Subsection 3.1.3. Here, we said that experts only determine the likelihood ratio and do not look at the posterior odds. In this specific case, determining the posterior odds instead of the likelihood ratio seems to work quite well and appears to be easier. It seems (almost) impossible to determine the likelihood of the evidence that the suspect was involved in at least four fires in two years time, given that a certain number of them were systematic fires. In the end, the posterior probabilities that we obtained answer the question that one is interested in. Therefore, we think that it is allowed to use posterior odds instead of a likelihood ratio.

In the report of the Netherlands Forensic Institute (NFI, 2019), it is also highlighted that we focus on the posterior probabilities that the events took place at a certain household in the Netherlands. It is concluded that the probabilities can therefore be applied to the situation of the suspect, since it is treated as a random household in the Netherlands. Deciding to look at the situation from the perspective of households in the Netherlands, influences the prior probabilities. If we have more information about the situation of the suspect, this perspective can be changed which influences the probabilities for him.

To follow up, as we can conclude the calculation in Subsection 7.2.4 we cannot compute the posterior probability that at least four of the fires in the household of the suspect had a systematic cause. In the ideal situation, we would like to determine these as well. How can we calculate it? What is the cause that we cannot determine them now? To answer the second question, it appears that it is due to the use of all upper and lower bounds. If we can make these more precise, then it is perhaps possible to also determine this final probability. However, now we did not come up with a solution to improve the estimating steps.

Finally, we wonder whether we can apply this model to other types of incident series besides arson. For example, if we look at roster cases and a nurse is exceptionally often present at incidents in a hospital. We can be interested in the cause of this incidents. Is it merely coincidence, or is there a systematic cause? It is the same question we asked ourselves at the beginning of this chapter in case of arson. It turns out that in the roster cases that we will discuss next, we have access to other kinds of information than is explicitly needed for this arson model. E.g., in the Lucia de Berk case that is considered we do not know the population size N, i.e., total number of nurses. Therefore, it is difficult to apply it. We can conclude that it is important before making a choice for a particular approach, that we have an overview of what is already known and base our next steps on this information.

8
Roster cases

Next, we discuss so called roster cases. A physician or nurse is remarkably often associated with incidents in a hospital or nursing home. They are sometimes even suspected of murdering their patients. We are interested in the evidence of attendance. To find this kind of evidence, we need access to roster data to establish whether a member of medical staff was present during the incident.

8.1 Situation description

In this chapter, we focus on the infamous legal case of Lucia de Berk. However, worldwide there were a lot more legal cases that are very similar to this one, e.g., Ben Geen and Daniela Pogglia (Fenton et al., 2022). Lucia de Berk was a Dutch nurse that worked at the Juliana Children's Hospital and at two wards of the Red Cross Hospital (Schneps and Colmez, 2013). On September 4, 2001, a baby died unexpectedly during one of De Berk's shifts in the Juliana Children's Hospital. In response, one of her colleagues went to the supervisor and explained that Lucia de Berk was present at a lot more resuscitations compared to other nurses that worked in the hospital. The situation was also brought to the attention of the hospital's director, and information was gathered from the two hospitals to find out at how many incidents Lucia de Berk was present. Some informal calculations were made, to find out what the probability was of De Berk being present at so many resuscitations and deaths by sheer misfortune. The results were shocking, and the information was turned over to the police and the press was contacted. On December 31, 2001, Lucia de Berk was arrested and charged with thirteen murders and four attempted murders.

Law professor Henk Elffers, with a master's degree in mathematics and statistics, was contacted to analyse the information. He concluded there was a chance of 1 in 342 million that Lucia de Berk could coincidentally have been present at so many unnatural deaths (Elffers, 2002). Later, it turned out that the information collected by the hospitals was incorrect and the computations by Henk Elffers were questionable. The model that he used was criticised by, for example, Meester et al. (2006) and de Vos (2004).

One of the criticisms he got, was that the differences between nurses were not considered. We focus in this chapter on the model proposed by Gill et al. (2018).[1] They wanted to add a modest amount of heterogeneity between nurses and shed some new light on this legal case.

Like in the previous chapter, we first give the definitions and assumptions that the model was based on and then discuss the details of the model. The aim of Gill et al. (2018) was to calculate the probability that a nurse coincidentally experienced at least as many incidents as Lucia de Berk.

8.1.1 Definitions

We start with a difficult one: the definition of an incident. From the data that was initially used, it is not entirely clear what they meant with an "incident". It also played an important role in data collecting.

[1]For the discussion of the model, we also made use an earlier published version of this article: Gill et al. (2010).

One was not sure what to look for. The incidents were mostly resuscitations and (unnatural) deaths. Therefore, we just assume that the incidents marked by the hospital were indeed incidents (whatever is meant with it) and postpone the discussion about the definition of incident and collecting of the data to Section 8.3.

Furthermore, we need some precision of what a shift is. Shifts in the hospitals where Lucia de Berk worked, were eight hours long. One day contained three shifts: a day shift, an evening shift, and a night shift. Using this information, we can find out how many shifts there were in total during the period that De Berk worked in the two hospitals.

8.1.2 Assumptions

The model proposed by Gill et al. (2010) is meanly based on the two assumptions,

- the incidents that a nurse experiences can be modelled as a homogeneous Poisson process on the positive half line, where the positive half line coincides with time; and
- the intensity of this Poisson process, that is the intensity of nurses seeing or reporting incidents, is nurse-dependent. It is assumed to be $Gam(\rho, \rho/\mu)$ distributed.

Henceforth, we call it a mixed Poisson model.

We adapt the following notation:

- $N(t)$: the number of incidents that a nurse witnesses in time period t (Poisson process)
- t: time period of interest
- Λ: nurse-dependent intensity of the Poisson process $\{N(t) : t \geq 0\}$

Moreover, we adapt a time unit of one shift.

Hence, using this model our sample of nurses consists of realisations of the random vector (Gill et al., 2010)

$$(\Lambda, N(t)).$$

So, if for example we have n nurses, we deal with a sample

$$(\Lambda_1, N_1(t_1)), \ldots, (\Lambda_n, N_n(t_n)),$$

of independent random vectors, all having the same distribution $(\Lambda, N(t))$.

The original data that was used for the legal case of Lucia de Berk is given in Appendix E.1, Table E.1. For example, the time period of interest for the Juliana Children's Hospital was from October 1, 2000, to September 9, 2001. This period consists of 343 days. Like we said before, one day contained three shifts of eight hours. Hence, we obtain a total of $343 \times 3 = 1029$ shifts for the Juliana Children's Hospital. The other periods of interest for the two wards at the Red Cross Hospital can be found in the technical report of Elffers (2002). In the table, we can also find the total number of incidents that happened in the two hospitals, and the number of incidents where Lucia de Berk was present.

Now, we look at the values of the parameters ρ and μ of the gamma distribution used for the nurse-dependent intensity of the Poisson process. Gill et al. (2010) assume that $\rho = 1$, so that the intensity Λ is actually $Exp(1/\mu)$ distributed. They say that *'this implies that it can easily happen that one nurse has twice the incident rate of another nurse'*. In fact, if we look at an arbitrary pair of nurses, the probability that one of them has twice the incident rate of the other is 2/3. More generally, the probability of a nurse's incident rate being of a factor k that of another nurse is $2/(k + 1)$.[2]

Thus, the model boils down to the estimation of the parameter μ. It holds that μ is the overall probability of an incident per shift and assumed to be the ratio of total number of incidents to total number of shifts. From Table E.1, it follows that the total number of shifts was 1734 and the total number of incidents was 27. Hence μ can be estimated as $\mu = 27/1734$.

8.2 Model of Gill, Groeneboom and de Jong

We are interested in the probability that a nurse coincidentally experienced at least as many incidents as Lucia de Berk. Therefore, we first need to find the unconditional distribution of the number of incidents.

[2] See Appendix A.2, Theorem A.3.

8.2.1 Distribution of mixed Poisson model

We determine the distribution of our mixed Poisson model. The result can be found in the following theorem, which proof is partly based on (Brouwer et al., 2020, Chapter 5).

Theorem 8.1. *The above defined Poisson process is* $NB(\rho, (1 + t\mu/\rho)^{-1})$ *distributed.*

Proof. Since the number of incidents is a homogeneous Poisson process $\{N(t) : t \geq 0\}$, we know from statement (ii) of Theorem 3.1 that $N(t)$ is Poisson distributed with parameter $t\lambda$. Here λ should be treated as a random variable Λ. Then its probability mass function equals

$$P(N(t) = n|\Lambda = \lambda) = \frac{(t\lambda)^n}{n!}e^{-t\lambda}, \qquad n = 0, 1, 2, \ldots.$$

By assumption, it holds that $\Lambda \sim Gam(\rho, \rho/\mu)$. Hence its probability density function π is given by

$$\pi(\lambda) = \frac{(\rho/\mu)^\rho}{\Gamma(\rho)}\lambda^{\rho-1}e^{-(\rho/\mu)\lambda}, \qquad \lambda \geq 0,$$

where $\Gamma(\rho)$ is the gamma function defined as in (2.6). We know that $\Gamma(\alpha + 1) = \alpha\Gamma(\alpha)$ for $\alpha > 0$ and $\Gamma(n) = (n - 1)!$ for $n = 1, 2, \ldots$ (Dekking et al., 2010).

We derive the unconditional distribution of this Poisson process. By the law of total probability,

$$\begin{aligned} P(N(t) = n) &= \int_0^\infty P(N(t) = n|\Lambda = \lambda)\pi(\lambda)\, d\lambda \\ &= \int_0^\infty \frac{(t\lambda)^n}{n!}e^{-t\lambda}\frac{(\rho/\mu)^\rho}{\Gamma(\rho)}\lambda^{\rho-1}e^{-(\rho/\mu)\lambda}\, d\lambda \\ &= \frac{t^n}{n!}\frac{(\rho/\mu)^\rho}{\Gamma(\rho)}\int_0^\infty e^{-(t+\rho/\mu)\lambda}\lambda^{n+\rho-1}\, d\lambda. \end{aligned}$$

Now, let $x = (t + \rho/\mu)\lambda$. Then it follows that

$$\begin{aligned} P(N(t) = n) &= \frac{t^n}{n!}\frac{(\rho/\mu)^\rho}{\Gamma(\rho)}\int_0^\infty e^{-x}\left(\frac{x}{t + \rho/\mu}\right)^{n+\rho-1}\frac{1}{t + \rho/\mu}\, dx \\ &= \frac{t^n}{n!}\frac{(\rho/\mu)^\rho}{\Gamma(\rho)}\frac{1}{(t + \rho/\mu)^{n+\rho}}\int_0^\infty e^{-x}x^{n+\rho-1}\, dx \\ &= \frac{t^n}{n!}\frac{(\rho/\mu)^\rho}{\Gamma(\rho)}\frac{1}{(t + \rho/\mu)^{n+\rho}}\Gamma(n + \rho) \\ &= \frac{\Gamma(n + \rho)}{n!\Gamma(\rho)}\left(\frac{\rho/\mu}{t + \rho/\mu}\right)^\rho\left(\frac{t}{t + \rho/\mu}\right)^n \\ &= \frac{\Gamma(n + \rho)}{\Gamma(n + 1)\Gamma(\rho)}\left(\frac{\rho/\mu}{t + \rho/\mu}\right)^\rho\left(\frac{t}{t + \rho/\mu}\right)^n \\ &= \binom{n + \rho - 1}{n}\left(\frac{1}{1 + t\mu/\rho}\right)^\rho\left(\frac{t\mu/\rho}{1 + t\mu/\rho}\right)^n. \end{aligned}$$

Probability mass functions are uniquely determined, so from the above expression we can conclude that the above defined process is $NB(\rho, (1 + t\mu/\rho)^{-1})$ distributed. ■

In general, the negative binomial distribution can be viewed as a Poisson distribution, where the Poisson rate is a gamma distributed random variable and the parameters have a specific form. See Appendix A.3, Theorem A.4, for the proof which is very similar to the one above.

Using the above, we have a look at the expectation and variance of the Poisson process $N(t)$. For Poisson distributions, it holds that the expectation and variance are equal. However, from the well-known relations (Gill et al., 2018)

$$\begin{aligned} E(X) &= E(E(X|Y)), \\ Var(X) &= E(Var(X|Y)) + Var(E(X|Y)), \end{aligned}$$

it follows that the variance of the mixed Poisson model is larger than its expectation. As Gill et al. (2018) say in our context: *'If some nurses experience more or less incidents than others, the end result in all cases is* overdispersion[3] *caused by* heterogeneity*'*. We find that

$$\begin{aligned} E(N(t)) &= E(E(N(t)|\Lambda)) \\ &= tE(\Lambda) \\ &= t\frac{\rho}{\rho/\mu} \\ &= t\mu. \end{aligned}$$

Moreover,

$$\begin{aligned} Var(N(t)) &= E(Var(N(t)|\Lambda)) + Var(E(N(t)|\Lambda)) \\ &= tE(\Lambda) + t^2 Var(\Lambda) \\ &= t\frac{\rho}{\rho/\mu} + t^2\frac{\rho}{(\rho/\mu)^2} \\ &= t\mu + (t\mu)^2/\rho. \end{aligned}$$

In the above, we used that the expectation and variance of a gamma distribution with parameters α and β are given by α/β and α/β^2, respectively (Dekking et al., 2010). Hence the heterogeneity between nurses, and so overdispersion, is controlled by the parameter ρ. As ρ becomes arbitrarily large, we find ourselves in the homogeneous case. By choosing $\rho = 1$, we thus add a modest amount of heterogeneity.

For $\rho = 1$, the mixed Poisson model has a geometric distribution with parameter $(1 + t\mu)^{-1}$.

8.2.2 Probability that a nurse experiences at least n incidents

Now, our aim was to find the probability $P(N(t) \geq n)$. Using the result from Theorem 8.1, the probability that a nurse experiences at least n incidents is

$$\begin{aligned} P(N(t) \geq n) &= \sum_{k=n}^{\infty} \binom{k+\rho-1}{k} \left(\frac{1}{1+t\mu/\rho}\right)^{\rho} \left(\frac{t\mu/\rho}{1+t\mu/\rho}\right)^{k} \\ &= 1 - \sum_{k=0}^{n-1} \binom{k+\rho-1}{k} \left(\frac{1}{1+t\mu/\rho}\right)^{\rho} \left(\frac{t\mu/\rho}{1+t\mu/\rho}\right)^{k}. \end{aligned} \tag{8.1}$$

Lucky for us, this probability simplifies when we plug in $\rho = 1$. We find that

$$\begin{aligned} P(N(t) \geq n) &= \sum_{k=n}^{\infty} \frac{1}{1+t\mu} \left(\frac{t\mu}{1+t\mu}\right)^{k} \\ &= \left(\frac{t\mu}{1+t\mu}\right)^{n}. \end{aligned} \tag{8.2}$$

We estimated that $\mu = 27/1734$ and Lucia de Berk worked $t = 201$ shifts. So, on average, Lucia de Berk should have experienced $E(N(t)) = t\mu = 201 \times (27/1734) \approx 3.130$ incidents. The probability that the number of incidents in 201 shifts for one nurse is at least n, for $n = 1, 2, \ldots$, is illustrated in Figure 8.1.

Lucia de Berk was present at fourteen incidents. Hence the probability that a nurse coincidentally experienced at least as many incidents as De Berk, is

$$P(N(t) \geq 14) = \left(\frac{201 \cdot (27/1734)}{1 + 201 \cdot (27/1734)}\right)^{14} \approx 0.0206. \tag{8.3}$$

This is about 1 in 49. One can conclude that *'(...) a modest amount of heterogeneity turns an almost impossible occurrence into something merely mildly unusual, (...)'* (Gill et al., 2018).

[3]Overdispersion is the presence of greater variability in a data set than would be expected on a given statistical model (Wikipedia, 2021). The observed variance of the data is higher than the variance of a theoretical model.

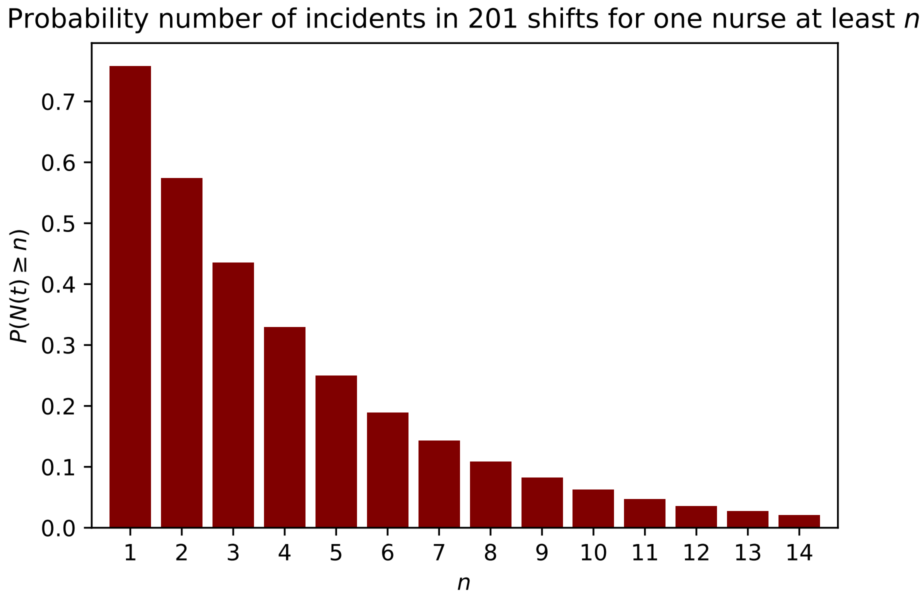


Figure 8.1: The probability that the number of incidents in 201 shifts for one nurse is at least n, for $n = 1, 2, \dots,$ if $\mu = 27/1734$.

8.3 Comments

Now, we give our comments about this approach to model the incidents that a nurse experiences. We discuss, inter alia, the validity of the assumptions and give a sensitivity analysis on the parameters ρ and μ.

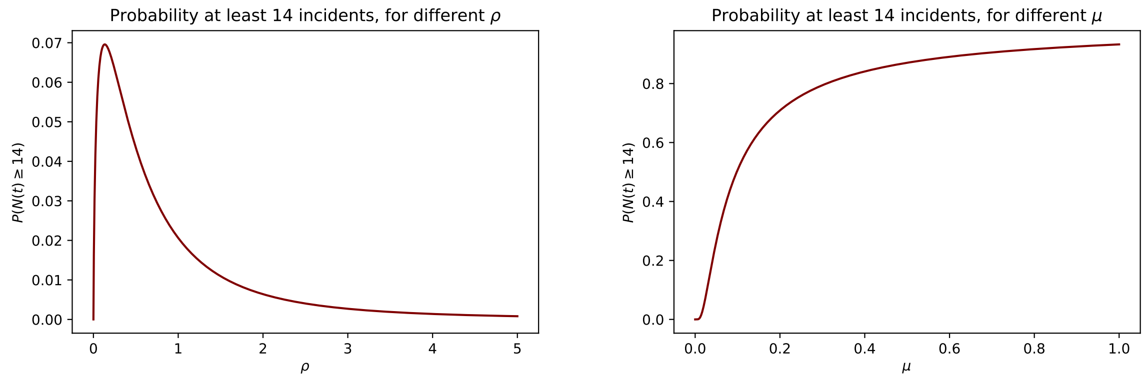
8.3.1 Validity of assumptions

This model was based on the assumptions that the incidents that a nurse experiences can be modelled as a homogeneous Poisson process, and that its intensity is nurse-dependent and $Gam(\rho, \rho/\mu)$ distributed.

To begin, it is well-known that Poisson processes are for example used to model incoming phone calls during non-busy hours, fires in a big city and many more (Gill et al., 2018). These can be thought of as rare events. Now, incidents in a hospital are also considered to be rare, so a Poisson process to model the incoming of incidents is an obvious choice.

Besides this intuition, we discuss more formally whether we can model the incidents as a Poisson process. First, note that we are dealing with a counting process. Then statement (iii) of Theorem 3.1 explains that for $\{N(t) : t \geq 0\}$ to be a Poisson process, we need that $N(0) = 0$ and the interarrival times are independent identically distributed exponential random variables. It is obvious that the first property holds, i.e., the number of incidents equal zero when the nurse did not work a shift yet. However, for the second property we should assume that independence between the interarrival times holds. There are situations when one cause induces multiple incidents, such as a flu outbreak in a hospital (Green, 2006) or a power cut (M. Sjerps, personal communication, November 19, 2021). However, since these circumstances are occasional exceptions, it is still reasonable to assume that the independence property holds (Green, 2006). Though, note that the incidents themselves may not be independent. Then we should assume that the occurrence of an incident during one shift is independent of an incident in the subsequent shift. However, e.g., a successful reanimation may be followed by death (Meester et al., 2006).

There is left to show that the interarrival times are exponentially distributed. Therefore, we use the memoryless property of the exponential distribution described in Theorem 3.2. In our context, we can translate this property as: how much time has passed since the last incident does not influence how much time it takes until the next incident (Collins, 2005). One can argue that more hospital incidents occur during night shifts than during day shifts (Meester et al., 2006), and therefore the average arrival rate is time-dependent. Also, some days may be busier than others. However, overall a Poisson process may be seen as a good model for a shorter time interval over which the arrival rate is fairly constant (Green, 2006). The exponential distribution is often used to model customers entering a store with a



(a) $\rho \in (0, 5]$

(b) $\mu \in [0, 1]$

Figure 8.2: The probability that the number of incidents in 201 shifts for one nurse is at least fourteen, for different values of ρ and μ. We set $t = 201$. In addition, we set $\mu = 27/1734$ in (a), and $\rho = 1$ in (b).

certain frequency. If we wait for the next customer to enter our store, it does not matter how much time has passed since the last customer entered the store. It does not give us information how long it takes until the next customer arrives. One example, for which the memoryless property is not reasonable is the longevity of a mechanical device (Kim, 2019). The older the device, the more likely it is to break down. But for modelling the time until a car accident, the exponential distribution is appropriate. E.g., if no one has hit you in the past five hours, it does not increase of decrease the chance of a car accident (Kim, 2019). We conclude that the memoryless property is reasonable and that the interarrival times are indeed exponentially distributed.

Next, we look at the nurse-dependent intensity. There is a lot of discussion whether nurses are interchangeable or not (Gill et al., 2010). Medical specialists say that nurses are completely interchangeable with respect to the occurrence of medical emergencies among their patients. However, according to nursing staff this is not the case at all. To cite Gill et al. (2010): *'Different nurses have different styles and different personalities, and it can and does have a medical impact on the state of their patients'*. Moreover, ill persons tend to die preferable in the presence of a nurse whom they feel more comfortable. There has been no research on this phenomenon. Consequently, the results of this model give strong support to explore whether and if so in what form heterogeneity plays a role in healthcare. We agree with the above nurses' arguments, and hence adding a modest amount of heterogeneity via the nurse-dependent intensity is acceptable.

Finally, we may wonder why one chooses a gamma distribution to model the nurse-dependent intensity. Gill et al. (2018) say that other distributions are also possible, but one might have noticed that using this particular distribution leads to closed expressions. All we needed to model the heterogeneity between nurses was an extra parameter in the Poisson process, so why not choose this nice one? We do not have information about the distribution of the intensities. We would need data regarding the other nurses and their witnesses of incidents. Also, in case of overdispersion the negative binomial distribution is often used (David and Jemna, 2015).

8.3.2 Sensitivity analysis

Gill et al. (2010) choose to use the parameter $\rho = 1$, so that the intensity is $Exp(1/\mu)$ distributed. But what happens when we use a different value for ρ? We plotted the probability that a nurse coincidentally experienced at least fourteen incidents for $\rho \in (0, 5]$, see Figure 8.2a. We see that $P(N(t) \geq 14)$ peaks for small ρ and then goes to zero as ρ becomes larger. The parameter ρ controls the amount of heterogeneity. As ρ becomes arbitrarily large, we find ourselves in the homogeneous case and all nurses are regarded as the same. In this case the number of incidents is Poisson distributed with parameter $t\mu$ where $t = 201$ and $\mu = 27/1734$, and we see that the probability $P(N(t) \geq 14)$ is almost negligible (approximately 5.46×10^{-6}). It corresponds with the results from Figure 8.2a. Furthermore, the small probabilities before the peak in Figure 8.2a may be caused by some so called "incident magnets" that experience a large number of incidents, which affects the probability of experiencing at least fourteen incidents for another person to be small (P. Vergeer, personal communication, November 30, 2021). When ρ is very small, there are more outliers in the population.

We also looked at what happens when we vary the values of μ from zero to one, see Figure 8.2b. We set $\rho = 1$ but note that the result does not differ much for different values of ρ. When μ becomes larger, $P(N(t) \geq 14)$ also grows. This makes sense, because μ was the probability that an incident occurs during a shift. If we are almost certain that an incident occurs during a shift ($\mu \to 1$), then the probability that a lot of incidents occur over a certain time period will be large.

To continue the discussion about the definition of incidents, there was a lot of disagreement how the data was gathered in the first place. The data collection given in Appendix E.1, Table E.1, was flawed (Gill et al., 2010). Research disciplines came into conflict: criminal investigation and scientific data gathering are very different. For meaningful statistics we need clear definitions and uniformity of the data collection. In the legal case of Lucia de Berk, the lack of it proved to be disastrous. The data collection was incomplete and inconsistent. For example, incidents without De Berk were dropped and incidents were declassified without clear reasons (Gill et al., 2010). Extra shifts were discovered without incidents and incidents outside De Berk's shifts. Moreover, the data collection partly relied on memory and the witnesses were aware of what facts were looked for (confirmation bias).

A statistician should be concerned about the data quality and point out what the effects are of inconsistent and incomplete data (Gill et al., 2010). Therefore, we want to also look at the outcome of this model if we apply it to the "corrected" data collection[4] given in Appendix E.1, Table E.2. We see that in this case, Lucia de Berk was present at nine incidents instead of fourteen. The total number of shifts remains 1734, but the total number of incidents reduces to 26. The number of shifts of De Berk is 203. Therefore, our parameters become $\mu = 26/1734$ and $t = 203$. We find that

$$P(N(t) \geq 9) = \left(\frac{203 \cdot (26/1734)}{1 + 203 \cdot (26/1734)}\right)^9 \approx 0.078,$$

which is about 1 in 13. Hence changing the data causes a major impact. Here, we also looked at what happened to this probability when we vary ρ and μ. The results can be found in Appendix C.2, Figures C.2a and C.2b.

8.3.3 Other remarks

The probability that a nurse experienced at least as many incidents as Lucia de Berk, can be seen as a probability (to be precise a p-value) under the hypothesis of chance. We do not look at the probability of being present at so many incidents under the hypothesis of an alternative cause, for example murder or De Berk being a serial killer nurse. Following Pundik (2021), the use of statistical evidence to prove the probability of a natural cause (in our case coincidence) is meaningless, without contrasting it with the frequency of the alleged cause (murder) among a group of people similar to the accused.

To illustrate this idea, assume that the probability of an event under the hypothesis H_1 is very small. Then we miss some information. It is also important to look at an alternative hypothesis H_2. Maybe the probability of the event under the hypothesis H_2 is even smaller or rather large. If we look at a so called "contrasting probability" we can conclude under which hypothesis the data (event) is explained better.

To go back to the legal case of interest, suppose that after all Lucia de Berk was a murderous nurse. Then Gill et al. (2018) already notices that we have to take into account that the incidents can be partly not murders. In this case, we have a combination of coincidence and intent. It seems to be a difficult task to model the phenomenon. Therefore, Gill et al. (2018) choose to present their results in the way explained in Section 8.2. Pundik (2021) criticises this by saying that *'(...), even in the absence of statistical evidence of the alternative cause, it is unhelpful and potentially misleading to consider this evidence on its own'*. It is objectionable, regardless of how reliable the statistical analysis is.

So, what do we have to do in this case? Do we disregard the statistical evidence that we obtained, or still use it? A p-value only gives us information that something extraordinary had happened, but not what the cause was of this event. We think that Gill et al. (2018) wanted to present a contrasting probability to Elffers' result and therefore chose to give a p-value in their article.

We will see that this problem, of presenting no probability under the alternative hypothesis, is very common. In Chapters 9 and 12 we will discuss a legal case where there happened a large number of arson incidents in some town. There appears to be a similar problem in this situation, because of the suspicion of there being a serial arsonist. We will explain in Subsection 12.2.2 how information about

[4]Our intention is to look at the behaviour of the model, if we adapt our data. As Gill et al. (2018) explain in their article, the given data collection is still flawed since the incidents that the court finally decided not to count as provably caused by Lucia de Berk were removed from the data collection. However, to illustrate the effect we still use it.

how many serial arsonists there are for the prior and the distribution of the number of fires over the serial arsonists can hopefully give us an idea about the order of magnitude of the probability under the alternative hypothesis. In the context of this chapter, we shall need information on serial killer nurses. Fenton et al. (2022) explained in their article that cases of serial killer nurses are incredible rare, and the incidence is perhaps roughly one in two million. In the Netherlands, it comes down to about once every ten years (R. Gill, personal communication, March 22, 2022).

The given p-value by Gill et al. (2018) also shows that the incidents that Lucia de Berk witnessed during her shifts was not necessarily rare, if we compare it to a group of similar nurses. From the report of Elffers (2002), we know there worked a maximum of 27 nurses on the ward of the Juliana's Children Hospital during the period that Lucia de Berk worked there. If we assume that a similar number of nurses worked at the Red Cross Hospital during this time, than the probability of 1 in 49 that Gill et al. (2018) found is actually to be expected.

Another point of discussion can be deduced from the data collections given in Appendix E.1, Tables E.1 and E.2. Namely, we have no information about the other nurses. The data collections only tell us something about Lucia de Berk and the "rest". It was already a problem when Henk Elffers investigated this legal case and will be a problem in many legal cases. We can ask ourselves what will happen to this model, when information about the other nurses and their witnesses of incidents is available. Maybe it turns out that the nurse-dependent intensity does not follow a gamma distribution at all, though we said that using a gamma distribution is a common choice.

Furthermore, in our opinion Gill et al. (2018) could not properly estimate the parameters ρ and μ if the gamma distribution was indeed a good representation of the intensity distribution. It holds especially for the parameter ρ, which represents the amount of heterogeneity between the nurses. It is interesting to look at what happens if we estimate the parameters based on data of a group of nurses that is big enough. In Chapter 11 we will dive into this.

9
Linear Bayes

Yet again, we investigate a situation where there is suspicion of arson. This time we consider a real legal case where for the major part it is certain that a crime took place, and one is more interested in who the culprit is than what the cause was. As we will see, this situation is somewhat different than the legal case that we discussed in Chapter 7. Instead of the fires happening in one household, arson is committed at different households but still in the same town.

As before, we will first give a situation description and discuss how the statistical evidence of interest is derived. We give our comments on the method at the end of this chapter.

9.1 Situation description

During the first half of 2013, fifteen fires took place at different locations in town X in the Netherlands, which varied in size and danger (Alkemade, 2015).[1] It had a great impact on the inhabitants of the town, since there was a constant threat that one could be the next victim. Two inhabitants of town X were suspected of committing these crimes. For convenience, we call them suspect A and suspect B. Suspect A recently moved to town X, and suspect B already lived there for some time. The question is whether one, both or none of them were the arsonist(s) behind the large number of fires.[2]

The above situation is examined by dr. Frans J.M. Alkemade, who is a part time senior advisor at the Dutch Ministry of Justice and writes reports for criminal cases (Alkemade, n.d.). Alkemade is sometimes called by the prosecution or by defence lawyers to give his advice about legal cases. He was involved as an expert in what we will call the "Town X fires". He used a different kind of approach in this legal case than one encounters most of the time. Alkemade proposed a so called "linear Bayesian" approach to evaluate the legal case as a whole, instead of reporting likelihood ratios about separate pieces of evidence as many experts do. See for example Dahlman (2020). According to Meester and Slooten (2021, Subsection 5.5.2), this approach ignores the dependency structures that one comes across in complex or complicated cases.

Initially, all the crimes were treated separately by the court, hence they assumed there was no cohesion between all the incidents of arson (Alkemade, 2015). However, it is possible to link the evidence of multiple crimes to investigate whether they can be attributed to the same suspect(s). Therefore, we need to think about the concept of *crime linkage*. The linear Bayesian approach of Alkemade (2015) is one of the methods that aims to do this. With the permission of Alkemade, we looked into his report about this legal case which is given in Alkemade (2015). Both the verdict of the court, see footnote 2, and of the appeal court mention explicitly that the report written by Alkemade had no contribution to the judgement of the case.

As we already said, Alkemade (2015) evaluates the legal case as a whole in his report. Therefore, he looks at different types of findings which are defined as every form of evidence and treated as clues.

[1] Because this situation is based on a real legal case, we decided to not include the real name of the town in this research thesis. Also because of duty of confidentiality, the names of the suspects are not mentioned.

[2] The verdict is given in ECLI:NL:GHSHE:2016:5165 (Gerechtshof 's-Hertogenbosch, 2016).

E.g., interrogations with suspects and victims, or objects found at the crime scene. In this thesis, we are interested in the forensic statistical interpretation of a series of incidents. Hence we want to make clear that we only consider the finding that a large number of fires happened in a short period of time. We do not analyse the other findings that Alkemade (2015) discussed in his report.

The goal is to determine the evidential value of experiencing fifteen fires in approximately six months (in town X) given different hypotheses about whether there was a serial arsonist and whom this possibly was. We start our analysis by listing some of the underlying assumptions that Alkemade (2015) used for his method to determine this evidential value of the forensic evidence, and then discuss his calculations in detail.

9.1.1 Definitions

To avoid confusion, we want to specify that throughout this chapter the concepts arson and fire are interchangeable.

9.1.2 Assumptions

It is assumed that

- under the scenario that there is no serial arsonist operating in town X, fires are independent of each other; and
- residents in the Netherlands are indistinguishable.

The following notions will be important throughout this chapter:

- N: the number of residents in the Netherlands
- n: the number of inhabitants in town X
- K: the number of fires in the Netherlands caused by arson
- t: time period of interest

We adapt a time unit of one year.

From Alkemade (2015), we know that

$$N \approx 17 \times 10^6,$$
$$n \approx 2400.$$

Hence the Netherlands approximately has seventeen million residents, of which 2400 live in town X. To determine K, fire statistics of the Netherlands in 2013 reported in CBS (2014) were used. In 2013, there were 36100 fires in the Netherlands. The fire statistics distinguish two types of fires: indoor fires and outdoor fires. Indoor fires are defined to be those that occurred in a (semi) closed space, and outdoor fires occurred in open-air spaces. In general, no buildings are involved in outdoor fires. Alkemade (2015, Subsection 3.4.1) assumed that the fifteen fires can be seen as indoor fires. The total number of indoor fires in 2013 was fifteen thousand. From (CBS, 2014, Figure 2.3.1) it follows that 19.2% of the indoor fires was caused by arsonism.[3] Hence Alkemade (2015) concluded that

$$K \approx 3000.$$

It is assumed that the period of interest is about $t = 0.5$ years, or six months.

9.2 Model of Alkemade

Now, we give a description of the calculations that Alkemade (2015) made in his report to interpret the statistical evidence of having fifteen fires in a town similar to town X during a time period of six months. Therefore, we must first define the hypotheses and evidence from the report. In the next section, we will give our comments about his method.

[3] Excluding other and unknown causes.

9.2.1 Hypotheses and evidence

In Section 3.1, we already discussed how Bayes' rule can be used in criminal proceedings. In the simplest case, we only have two hypotheses or scenarios of interest: the suspect is the culprit of a crime, or he is not. For some evidence we can compute the probability of observing this evidence given each of the hypotheses.

The situation of the "Town X fires" is a bit more complex. Not only did multiple incidents of arson take place instead of just one, but for some of these events we also must take into account that there may be multiple culprits. We have two suspects, called A and B. It is questioned whether they caused the fires, and so are serial arsonists. Therefore, the following eight possible hypotheses are considered:

- H_1: There was no serial arsonist operating in town X.
- H_2: At least one unknown person was operating as a serial arsonist.
- H_3: Only suspect A was operating as a serial arsonist.
- H_4: Suspect A and at least one unknown person were operating as serial arsonists.
- H_5: Only suspect B was operating as a serial arsonist.
- H_6: Suspect B and at least one unknown person were operating as serial arsonists.
- H_7: Only suspect A and suspect B were operating as serial arsonists.
- H_8: Suspect A, suspect B and at least one unknown person were operating as serial arsonists.

In his report, Alkemade (2015) computed the evidential value of multiple findings, given each one of these hypotheses. Mainly, he included findings in his analysis that he considered reliable. Since we have eight instead of two hypotheses, computing the posterior probabilities of each hypothesis works a little bit different than we saw in Section 3.1. Note first that we need to assign a value to each prior probability of the eight scenarios, and that they all need to add up to one.

Then for the evidential value, instead of looking at the likelihood ratio of a finding given two hypotheses, we must come up with a ratio between eight hypotheses. We use the same notation as Alkemade (2015). In mathematical notation, the evidential value can be presented by the following ratio between probabilities:

$$P(E|H_1) : P(E|H_2) : P(E|H_3) : P(E|H_4) : P(E|H_5) : P(E|H_6) : P(E|H_7) : P(E|H_8)$$

Note that they do not have to add up to one. We can also display the above as a ratio between numbers instead of probabilities. For example, if we think that a finding is ten times more likely under H_2 and H_4 than under the remaining six hypotheses, one can denote it as

$$1 : 10 : 1 : 10 : 1 : 1 : 1 : 1,$$

but also as

$$0.1 : 1 : 0.1 : 1 : 0.1 : 0.1 : 0.1 : 0.1.$$

They are equivalent since the ratio between them is the same. Using these proportions, we can update the prior probabilities to the correct ratio of posterior probabilities. Using Bayes' rule, we find that the posterior probabilities equal (Meester and Slooten, 2021, Equations (2.2) and (2.3))

$$P(H_i|E) = \frac{P(E|H_i)P(H_i)}{P(E)} = \frac{P(E|H_i)P(H_i)}{\sum_{j=1}^{8} P(E|H_j)P(H_j)},$$

for $i = 1, \ldots, 8$.

As mentioned before, we are only interested in the fact that many similar events took place. Therefore, the evidence is defined as

- E: In approximately six months, fifteen fires took place in town X.

Hence we want to find the likelihood of E given each one of the eight hypotheses.

9.2.2 Probability of experiencing 15 fires

Before looking into the likelihood ratio, we want to derive the probability that in a similar town as town X there were fifteen incidents of arson in a time period of approximately six months.

It is assumed that in 2013 there happened approximately $K \approx 3000$ arson incidents in the Netherlands similar to those that occurred in town X, and the number of residents in the Netherlands is $N \approx 17 \times 10^6$. Then it follows that the average number of arson incidents per capita per year is

$$\frac{K}{N} = \frac{3000}{17000000} \approx 0.000176. \tag{9.1}$$

In 2013, town X had $n \approx 2400$ inhabitants. Hence the expected number of arson incidents per year in town X is

$$n \times 0.000176 = 2400 \times 0.000176 \approx 0.42. \tag{9.2}$$

It means there is approximately one fire every 2.4 years.

Now, we describe how Alkemade (2015) determined the probability that in some town of similar size as town X, fifteen fires happened in six months given there is no serial arsonist active. Approximately 0.42 incidents of arson happen every year in town X. Then it follows that every $t = 0.5$ years, the average number of arson incidents in town X is

$$t \times 0.42 = 0.5 \times 0.42 \approx 0.21. \tag{9.3}$$

Alkemade (2015) interprets this result as the probability per six months of experiencing an incident of arson in a town similar to town X under the hypothesis of chance (sic.).[4] Then under the assumption of independence, it holds true that the probability of having fifteen fires in six months in a town like town X becomes

$$(0.21)^{15} \approx 7 \times 10^{-11}, \tag{9.4}$$

as reported by Alkemade (2015). This is about one in ten billion.

Now, suppose that the Netherlands consists of 10000 towns like town X, hence each with 2400 inhabitants. Then according to Alkemade (2015) the probability equals

$$10000 \times 10^{-10} = 10^{-6}. \tag{9.5}$$

Alkemade (2015) reasons that, based on this result, one expects under the hypothesis of having no serial arsonist, these fifteen fires happen somewhere in one of those towns similar to town X is less than one in every million years (sic.). Thus, his conclusion is that the probability that in 2013 this series of fifteen fires occurred somewhere in the Netherlands, without assuming there is some underlying cause such as a serial arsonist, is less than one in a million.

9.2.3 Likelihood ratio of experiencing 15 fires

Now, using the above probabilistic arguments, we want to determine the value of the evidence that in six months fifteen fires took place in town X. Alkemade (2015) reasons that we expect to encounter this series of events under H_1 less often than under the remaining hypotheses. Based on the above reasoning, Alkemade (2015) states that it may easily be one million times less than in the other cases. He argues that, at first glance, there does not seem a reason for the evidence to discriminate between hypotheses H_2 through H_8. Hence a ratio of

$$10^{-6} : 1 : 1 : 1 : 1 : 1 : 1 : 1 \tag{9.6}$$

is obtained.

9.3 Comments

We now give our comments about Alkemade's approach to compute the probability that in six months there were fifteen incidents of arson in town X, and hence determine the likelihood ratio of this event given each of the eight hypotheses. We start with looking at the assumptions that he (implicitly) made and the specific steps in his calculations. Then we look at the influence of the data on the outcomes and give some final thoughts on his method.

[4] Here, "chance" means there is no serial arsonist operating in town X at the moment.

9.3.1 Validity of assumptions

First, Alkemade implicitly assumed that residents of the Netherlands are interchangeable and hence have the same vulnerability to fires. In Subsection 7.3.1 we already made clear that this assumption does not seem right to us. For example, from (CBS, 2014, Table A2.13), it already follows that the number of arson incidents that happened in different types of houses, e.g., terraced houses and apartment buildings, vary quite a lot. It may be the cause of poor households living in small apartments, and there is more criminal activity in that area. Hence the argument of Alkemade does not seem to hold.

We also want to give some comments on the data that Alkemade (2015) used during his calculations. We had a look at the document CBS (2014) that he referred to in the report, and came to the conclusion that he made a mistake during his data extraction. In the fire statistics document, it indeed states that 19.2% of the indoor fires are caused by arson. However, as we already highlighted in a footnote, it holds true that this percentage excludes other and unknown causes. From (CBS, 2014, Table A2.7) it follows that almost 75% of the indoor fires have other or unknown causes. It also states that about 700 indoor fires were caused by arson. Therefore, Alkemade should have used the number $K \approx 700$ instead of $K \approx 3000$. These numbers differ a factor four, and seem to have quite some influence on the number of arson incidents per capita per year given in (9.1). Note that some of the fires having other or unknown causes mentioned in the fire statistics of CBS (2014) may also be induced by arson. Therefore, we are not certain whether $K \approx 700$ represents the reality better. From L. van der Ham (personal communication, March 2022) we know that the information for the fire statistics from CBS (2014) are obtained by the firemen that were present at the scene. They report what they think the cause of the fire was. If the cause turns out to be different sometime later, then it is often not corrected in the statistics. The restrictions of the fire statistics in CBS (2014) are also confirmed in (Schoenmakers and van Ham, 2013, Section 4.3). Therefore, the fire statistics from CBS (2014) are not really trustworthy.

It is assumed that all the incidents of arson in town X were indoor fires. The reason behind it is that, although some of the fires started outdoors, the goal was to start a fire indoors (Alkemade, 2015). Hence the number of fires in the Netherlands caused by arson K, that are like those that happened in town X, is also influenced by this choice. Alkemade (2015) argues that if we also include outdoor fires in his original numbers, then the final probability might be larger. Though, he also states that this probability can be even smaller since all fires happened very close to each other.

Furthermore, we think it is a bit curious that Alkemade (2015) used the total number of residents in the Netherlands instead of the number of households in his calculations, like the Netherlands Forensic Institute (NFI, 2019) did. See Chapter 7. It holds true that, if a fire takes place in a house, often multiple residents of the house are victim of the event. Therefore, it makes more sense to look at households. We note that, when he computes the number of arson incidents per year in town X, in (9.2), the ratio of inhabitants of town X to total number of residents in the Netherlands is computed. If we translate it to the number of households, the ratio will likely not change much. However, the scale level of the number changes. Instead of obtaining the number of arson incidents per year *at an inhabitant* of town X, we get the number of arson incidents per year *at a household* of town X. The rest of the calculations stays the same, but the perspective alters. In our opinion, the choice of Alkemade for taking the number of residents in the Netherlands was on that account a bit odd.

In (9.3) we determined the number of arson incidents in town X every half year. Now, Alkemade (2015) interprets this result as the probability per six months of experiencing arson in a town similar to town X, under the hypothesis that there was no serial arsonist. He changes an expected value into a probability. Suppose that the number of inhabitants of town X was, e.g., $n \approx 24000$ instead of $n \approx 2400$. Then the "probability" of 0.21 in (9.3) becomes 2.1, which is larger than one. The number obtained in the case can, however, be regarded as a reasonable approximation.

We agree that overall fires can be independent of each other, under the hypothesis that there is no serial arsonist operating in town X. Namely, under this hypothesis we have that each fire is caused by an arsonist that only acted once. One can still think of extreme scenarios, e.g., we have a gang of criminals that committed arson or there was a revenge on someone. Nevertheless, the independence assumption is reasonable under H_1.

Alkemade (2015) looks at the end of his calculations at the possibility of there being 10000 towns like town X, each with 2400 inhabitants. He corrects his probability of one in ten billion, to one in one million by multiplying it by 10000. What Alkemade does here, falls under so called selection effects. He tries to correct for the scale level of seeing the evidence. This correction seems to come a bit out of the blue, and there is not given support for the number 10000. Maybe this number should be much smaller, since the number of residents in the Netherlands is only seventeen million instead of 24 million.[5] It is difficult to choose some sort of scale, hence this last step is unfounded and completely arbitrary. One can also ask themselves why Alkemade only corrects for looking specifically at town X, and not for the fact that exactly fifteen fires happened. Correction for selection effects in the likelihood ratio is unnecessary. The prior odds correct for possible selection effects (Sjerps, 2004; Sjerps and Meester, 2009). This way, the posterior odds will remain the same.

We note that in his report Alkemade (2015, Section 3.4.2) states that we expect under the hypothesis of chance that in a town like town X, these fifteen fires happen less than one in every million years. However, from the context it follows that we talk about the probability of the event happening every six months and not every year. Hence his conclusion is incorrect.

Finally, we have a look at the likelihood ratios that are derived at the end. Alkemade (2015, Section 3.8.3) argues that the evidence is one million times less explainable under hypothesis H_1 than under the remaining hypotheses, based on his computations presented in Subsection 9.2.2. He sees no reason to discriminate between the other hypotheses and concludes that the ratio between the eight scenarios is

$$10^{-6} : 1 : 1 : 1 : 1 : 1 : 1 : 1. \tag{9.7}$$

We disagree. One has to take into account that the probability of having exactly fifteen incidents of arson in six months is small under each of the hypotheses, and not just under H_1 (though it might be the smallest). We think that considering the number of serial arsonists operating in town X already discriminates between the probabilities.

Alkemade draws his conclusion based on his computations. However, his calculated probability says nothing about the ratio between the probabilities of having fifteen fires in six months given the hypotheses. He states that the absolute probability of having fifteen fires given no serial arsonist is 10^{-6}, and concludes from this result that the evidential value is given by (9.7). It would mean that $P(E|H_1) = 10^{-6}$ and $P(E|H_i) = 1$ for $i = 2, \dots, 8$, which is not true. We are not certain that we obtain exactly fifteen fires, if there is at least one serial arsonist operating in town X.

9.3.2 Sensitivity analysis

We want to look at the effects on the probability of having fifteen fires when adapting the following numbers:

- K: the number of fires in the Netherlands caused by arson
- t: the time period of interest

Alkemade (2015) states in his report that he tries to make reasonable decisions, mostly in favour of the suspect(s). By doing a sensitivity analysis, we can see whether this is indeed the case. For now, assume that all the steps that Alkemade (2015) made in Subsection 9.2.2 are allowed.

First we explain why the effects of these quantities are important to us. We discussed that the number of arson incidents in the Netherlands is not $K \approx 3000$, but $K \approx 700$ based on CBS (2014). These numbers differ about a factor four and may have some influence on the outcome of the probability. Therefore, we look at the probabilities if the number of arson incidents in the Netherlands is two times and even four times smaller.

The time period of interest is also a discussion point in the report of Alkemade (2015). From the information about the legal case, it follows that all fifteen fires actually happened during two and a half months (which is about 0.21 years) instead of six months. Now, Alkemade (2015) chooses to look at the period from the moment that suspect A moved to town X until the last fire and to round this number up to six months. He argues that it works in favour of the innocence scenario. We want to see whether he is right and look at the effect when we adapt the time period to $t = 1$ and $t = 0.25$ years.

During our analysis, we keep the other factors constant and the same as in the previous calculations. In Tables 9.1a and 9.1b one finds the results of the adaptation of the number of arson incidents in the Netherlands K and the time period of interest t, respectively. We can conclude that the choice for K and t have some major impact of the probability of having fifteen fires in t years. It follows that Alkemade's choice for the time period of interest indeed does work in favour of the suspects. Though, it is a bit alarming that using the correct number of arson incidents in the Netherlands causes the order of magnitude of the probability to be that much smaller than in the original case (10^{-19} relative to 10^{-10}). His choice for taking $K \approx 3000$ seems to work in favour of the suspects as well.

[5]For 10000 towns of each 2400 inhabitants, we obtain $10000 \times 2400 = 24 \times 10^6$ residents.

Values of K	**Probability**
3000	7×10^{-11}
1500	2×10^{-15}
750	7×10^{-20}

(a) Fixed $t = 0.5$

Values of t	**Probability**
1	2×10^{-6}
0.5	7×10^{-11}
0.25	2×10^{-15}

(b) Fixed $K \approx 3000$

Table 9.1: Comparison of probability of having fifteen fires in t years in a town similar to town X for different values of K and t.

9.3.3 Other remarks

We conclude that Alkemade's method to obtain the evidential value of a series of arson incidents in a town contains multiple (interpretation) errors and is based on unreliable assumptions. We think that the model of Gill et al. (2018) might be suited to find a better answer to the problem of modelling incidents given the hypothesis of chance. However, it will still be very difficult to determine the probability under the hypothesis that at least one serial arsonist was operating in town X.

In Chapter 12 we will try to improve Alkemade's approach. We will discuss how to obtain the probabilities of a series of arson incidents under both the hypotheses of having no or (at least) one serial arsonist operating in town X.

Remark. Alkemade could not agree with our comments in Section 9.3 and emphasises that the goal of his report was not to calculate exact probabilities but rather to provide a upper limit favourable to the defence. He presented a linear Bayesian analysis of the complete criminal case in which the statistical analysis of the evidential value for the observed fifteen arson incidents only played a minor role.

Part IV

Improvements

10
Arson (revised)

After the discussion of the models for arson and roster cases in Chapters 7 and 8, we will have another look at the arson model and try to improve it by adapting one of the assumptions.

10.1 Motivation

In Subsection 7.3.1 we noticed that the assumption about indistinguishability of households is questionable. Some households are more vulnerable to fire than others. Therefore, we wish to incorporate the differences between households in our model. The model proposed by Gill et al. (2018) for roster cases includes heterogeneity between nurses. Now, we want to use the idea of the roster cases model in the arson model and look what the effect of integrating heterogeneity between households where there may occur a fire is on our previous calculated probabilities for the suspect having at least i fires in his household given that he was involved in at least four fires in two years, for $i = 1, \dots, 4$.

The model proposed by Gill et al. (2018) works under the hypothesis of chance. It is intended to determine the probability that a nurse coincidentally experiences at least as many incidents as Lucia de Berk. Therefore, we use the distribution of the number of incidents that a nurse witnesses to model the accidental fires defined in the arson model.

In the original model, we modelled the fires with an accidental cause according to a binomial distribution with parameters K_1 and $1/N$, where K_1 was the number fires with an accidental cause in the Netherlands and N was the number of households in the Netherlands where there may occur a fire. From the law of rare events, see Appendix A.4, Theorem A.5, it follows that we can approximate a binomial distribution by a Poisson distribution under certain conditions. Therefore, it is a natural choice to model the accidental fires as a Poisson process. To add heterogeneity between households, we again model the rate of the Poisson process as a gamma distributed random variable.

10.1.1 Assumptions

The assumptions given in Subsection 7.1.2 become

- accidental fires are independent of systematic fires; and
- the fires with an accidental cause that a household experiences can be modelled as a homogeneous Poisson process on the positive half line, where the positive half line coincides with time; and
- the intensity of this Poisson process, that is the intensity of households experiencing fires, is household-dependent. It is assumed to be $Gam(\rho, \rho/\mu)$ distributed.

We adapt similar notation as in Chapters 7 and 8. To summarise and put some quantities into the correct context:

- $N(t)$: the number of accidental fires that one household experiences in time period t (Poisson process)
- t: time period of interest
- Λ: household-dependent intensity of the Poisson process $\{N(t) : t \geq 0\}$
- N_h: the number of households in the Netherlands where there may occur a fire
- K: the number of fires in the Netherlands in t years
- K_1: the number of fires with an accidental cause in the Netherlands
- K_2: the number of fires with a systematic cause in the Netherlands
- X: the number of households where there have been at least four fires

It holds that $K = K_1 + K_2$, where K_1 and K_2 are unknown. In this case, we adapt a time unit of one year. From Chapter 7, we know that

$$\begin{aligned} N_h &\approx 7.7 \times 10^6, \\ K &\approx 52000, \\ X &\geq 10. \end{aligned}$$

We note that in this case we do not need an ad hoc maximum for the possible number of fires in a household, which was equal to 20 in our original case. This follows because the distribution that we use to model the number of accidental fires is not limited anymore. Its range is from zero to infinity, instead of from zero to K_1.

Then we only have to estimate the parameters for the mixed Poisson model. The time period of interest is $t = 2$ years. For simplicity, we set $\rho = 1$ for the parameter of the gamma distribution that models the heterogeneity. Then we still need an estimation of μ: the overall probability of an accidental fire per year.[1] This parameter can be evaluated by dividing the total number of fires (with an accidental cause) in one year by the total number of households where there may occur a fire. So, it holds true that $\mu = (K_1/t)/N_h$.

10.2 Model adaptations

To recite, we have the following two hypotheses for $i = 1, \ldots, 4$ (NFI, 2019):

- $H_{1,i}$: At least i of the fires in the household of the suspect had a systematic cause.
- $H_{2,i}$: Less than i of the fires in the household of the suspect had a systematic cause.

The evidence is

- E: In two years, the suspect was involved in at least four fires.

The posterior odds that we need to compute in order to determine the probabilities for the suspect having at least i fires in his household, for $i = 1, \ldots, 4$, are equal to

$$\begin{aligned} \frac{P(H_{1,i}|E)}{P(H_{2,i}|E)} &= \frac{P(\text{at least 4 fires, at least } i \text{ systematic cause})}{P(\text{at least 4 fires, less than } i \text{ systematic cause})} \\ &:= \frac{q_1}{q_2}. \end{aligned}$$

We will go through the same stages as in the original arson model described in Section 7.2, and will change some of the steps according to our new assumptions.

[1] In the context of the arson case from Chapter 7, it is the translation of "the overall probability of an incident per shift".

10.2.1 Probability of having at least 4 fires

We still model the quantity X, which is the number of households where these has been at least four fires, as a binomial distributed random variable with parameters N and p. Here $p := q_1 + q_2$ is the probability of having at least four fires in a household. Thus it holds true that by the method of Wilson, the 95% confidence lower bound for the probability p is given by

$$p := q_1 + q_2 \geq 7.764 \times 10^{-7}. \tag{10.1}$$

10.2.2 Probability of having j accidental fires

In this subsection, we are going to see some differences with the reasoning in Subsection 7.2.3 for computing the probability of having at least four fires from which j have an accidental cause. We assumed that accidental fires are independent of systematic fires, but in this case accidental fires are modelled as a homogeneous Poisson process with a household-dependent intensity that is $Gam(\rho, \rho/\mu)$ distributed. From Theorem 8.1, we know that $N(t) \sim NB(\rho, (1 + t\mu/\rho)^{-1})$. From the independence of the accidental and systematic fires, we know have the following equality:

$$\begin{aligned}
P(\text{at least 4 fires}, j \text{ accidental cause}) &= P(j \text{ accidental cause}, \text{at least } 4 - j \text{ systematic cause}) \\
&= P(j \text{ accidental cause}) \times P(\text{at least } 4 - j \text{ systematic cause}) \\
&= P(N(t) = j) \times P(\text{at least } 4 - j \text{ systematic cause}) \\
&= \binom{j + \rho - 1}{j} \left(\frac{1}{1 + t\mu/\rho}\right)^{\rho} \left(\frac{t\mu/\rho}{1 + t\mu/\rho}\right)^{n} \\
&\times P(\text{at least } 4 - j \text{ systematic cause}).
\end{aligned}$$

It still holds, for $k = 1, 2, \ldots,$ that

$$P(\text{at least } k \text{ systematic cause}) \leq \frac{\lfloor \frac{K_2}{k} \rfloor}{N_h}.$$

Therefore, the original probabilities of having at least four fires from which j have an accidental cause given in (7.2) and (7.3) become

$$\begin{aligned}
P(\text{at least 4 fires}, j \text{ accidental cause}) &\leq P(N(t) = j) \times \frac{\lfloor \frac{K_2}{4-j} \rfloor}{N_h} \\
&= \binom{j + \rho - 1}{j} \left(\frac{1}{1 + t\mu/\rho}\right)^{\rho} \left(\frac{t\mu/\rho}{1 + t\mu/\rho}\right)^{j} \frac{\lfloor \frac{K_2}{4-j} \rfloor}{N_h}
\end{aligned} \tag{10.2}$$

for $j = 0, \ldots, 3$ and

$$\begin{aligned}
P(\text{at least 4 fires}, j \text{ accidental cause}) &\leq P(N(t) = j) \\
&= \binom{j + \rho - 1}{j} \left(\frac{1}{1 + t\mu/\rho}\right)^{\rho} \left(\frac{t\mu/\rho}{1 + t\mu/\rho}\right)^{j}
\end{aligned} \tag{10.3}$$

for $j \geq 4$, respectively.

10.2.3 Posterior probability of having at least i systematic fires

Using the exact same reasoning given in the beginning of Subsection 7.2.4, we can find an upper bound for the probability q_2 of having at least four fires in a household from which less than i have a systematic cause, where $i = 1, \ldots, 4$. The only difference is that we will sum to infinity, since the range of the negative binomial distribution is not limited. Though if we do sum until 20 instead of infinity, the difference between the obtained upper bounds will be infinitesimal small.

It holds true that

$$
\begin{aligned}
q_2 &= P(\text{at least 4 fires, less than } i \text{ systematic cause}) \\
&\leq P(\text{at least 4 fires, more than } i \text{ accidental cause}) \\
&= \sum_{j=5-i}^{\infty} P(\text{at least 4 fires}, j \text{ accidental cause}) \\
&\leq \sum_{j=5-i}^{3} P(N(t)=j)\frac{\lfloor \frac{K_2}{4-j} \rfloor}{N} + \sum_{j=4}^{\infty} P(N(t)=j) \\
&= P(N(t)\geq 4) + \sum_{j=5-i}^{3} P(N(t)=j)\frac{\lfloor \frac{K_2}{4-j} \rfloor}{N_h}.
\end{aligned}
$$

If we plug in our estimated parameters $\rho = 1$ and $\mu = (K_1/t)/N_h$ and use the equality $K_2 = K - K_1$, we find using (8.2) that

$$
q_2 \leq \left(\frac{K_1/N_h}{1+K_1/N_h}\right)^4 + \sum_{j=5-i}^{3} \frac{1}{1+K_1/N_h}\left(\frac{K_1/N_h}{1+K_1/N_h}\right)^j \frac{\lfloor \frac{K-K_1}{4-j} \rfloor}{N_h}. \tag{10.4}
$$

The upper bound for q_2 still depends on the number of fires with an accidental cause K_1. Like in Chapter 7 we therefore have to plot (10.4) and determine the upper bounds for $i = 1, \ldots, 4$. The result can be found in Appendix C.3, Figure C.3.

Using the found upper bounds and (10.1) we can determine a lower bound for the probability $q_1 := p - q_2$ of having at least four fires in one household from which less than i have a systematic cause, where $i = 1, \ldots, 4$. Then we can compute a lower bound for the posterior odds

$$
\frac{P(H_{1,i}|E)}{P(H_{2,i}|E)}.
$$

This number can again be utilised to find the probabilities for the suspect having at least i fires in his household given that he was involved in at least four fires in two years, given by $P(H_{1,i}|E)$. This time, we skip the formal calculations and just give the results. These can be found in Table 10.1. Note that we yet cannot determine the posterior odds and posterior probability for the case that $i = 4$.

Number of systematic fires	**Posterior odds at least (Revised)**	**Probability at least (Revised)**	**Probability at least (Original)**
At least one	382.48	0.997	0.999
At least two	382.47	0.997	0.999
At least three	32.58	0.970	0.985
At least four	0	0	0

Table 10.1: Results for the improved arson model.

10.3 Comparison

Now, we compare the obtained results with the results from our original model. Table 10.1 shows that the probabilities that at least one or two fires in the household of the suspect had a systematic cause, given that in two years he was involved in at least four fires, change from approximately 99.9% to 99.7%. The posterior probability that at least three fires in the household of the suspect had a systematic cause also decreases from 98.5% to 97.0%. Thus, if we include some heterogeneity between the households in the Netherlands, we see that the posterior probabilities that at least i fires in the household of the suspect had a systematic cause, for $i = 1, 2, 3$, will decrease slightly.

The results here can be contrasted to those given in Table 7.1. They were what gave us a first indication to look whether heterogeneity between households affect the desired probabilities. It shows that for $i = 1$ and $i = 2$ the results are almost the same, but for $i = 3$ we have a bigger difference. This difference may be due to our assumption that $\rho = 1$, the parameter that controls the amount of heterogeneity in the model. As ρ goes to zero, households vary more and more in their risk of experiencing a fire. The probabilities in the improved model will decrease as ρ becomes smaller.

This behaviour is also shown in Appendix C.3, Figure C.4. It holds that negative values for the posterior probability indicate that q_1 is negative valued, and hence the given "probabilities" are meaningless. We see that when $\rho < 0.085$ the posterior odds and posterior probabilities cannot be determined anymore, since q_1 becomes negative for $i = 1, 2, 3$. (If $\rho > 168$, q_2 cannot be determined as well.) For $i = 4$, the posterior odds and posterior probabilities are not valid for all values of ρ as we expected from before.

In the next chapter, we consider once more the roster cases model proposed by Gill et al. (2018) and come up with different methods to estimate the parameters ρ and μ of the gamma distribution. One of those methods is the method of moments, what turns out to be most useful when much data is available. We already estimated μ as $(K_1/t)/N_h$, which coincides with the moment estimator for μ. For some fixed time period t, it will follow that ρ can be approximated as

$$\hat{\rho}_{MOM} = \frac{\overline{x}^2}{\overline{x^2} - \overline{x}^2 - \overline{x}}, \tag{10.5}$$

where x_i is the number of fires in household i in t years. Since the total number of accidental fires K_1 is unknown, we must estimate ρ using the distribution of the total number of BVH registered fires K in two years over the total number of households N_h in the Netherlands. Hence we need to know how many Dutch households experienced $0, 1, 2, \ldots$ fires in two years, which can be given by, e.g., a frequency table. Notice that the above moment estimator (10.5) does not depend directly on time. Hence the distribution of the number of fires over the households can also be used for a representable shorter or longer period. We just need an indication how the number of fires is spread among the households.

Unfortunately for us, this information was not available in the report of the Netherlands Forensic Institute (NFI, 2019). Nevertheless, we wanted to describe this method to invite others who have access to the asked frequency table to check whether our given value $\rho = 1$ is representable to model the amount of heterogeneity in a population of households that are susceptible to fires.

11
Roster cases (revised)

In this chapter, we revise the method proposed by Gill et al. (2018) to model the incidents that a nurse experiences in a hospital. In Subsection 8.3.3 we already addressed some points that may be adapted. We will discuss one of them below.

11.1 Motivation

During the trial of Lucia de Berk, there was no data publicly available about the other nurses that worked in the same hospitals as De Berk. The only things that we knew were the total number of shifts that the nurses worked and the total number of incidents they experienced. Therefore, it is unknown how the shifts and incidents were distributed among them. Since this data was not available, in our opinion Gill et al. (2010) could not properly estimate the parameters μ and ρ that describe the overall probability of an incident per shift and the amount of heterogeneity between nurses, respectively. The parameter μ was estimated as the ratio of the total number of incidents to the total number of shifts, which seems reasonable. However, the estimation of ρ was taken quite arbitrarily. The choice of Gill et al. (2010) for $\rho = 1$ was only based on the idea that in this case it was plausible that the intensity of one nurse can be twice as big as the intensity of another nurse.

We want to know what happens to the results when we estimate the parameters based on data of a group of nurses that is large enough. Namely, if we observe another case like Lucia de Berk's, we think it is possible to obtain the right data that can be used to model the number of incidents in a hospital even better. In the upcoming, we must use simulated data to obtain the incidents that the different nurses experience and their number of worked shifts since this data is not available.

To estimate the parameters ρ and μ, we can make use of different methods. Which ones we think will be most fitting, is discussed in the upcoming section. In the end we will apply the estimation methods to a data set about wave damage to cargo ships, which shows similar properties to the actual data that we need but that is not accessible.

11.2 Model adaptations

The aim is to determine the probability that a nurse coincidentally experienced at least as many incidents as Lucia de Berk, hence $P(N(t) \geq 14)$, given the number of incidents that other nurses than De Berk actually witnessed. Remember that we assumed that $N(t)$ given $\Lambda = \lambda$ is $Pois(t\lambda)$ distributed and Λ is $Gam(\rho, \rho/\mu)$ distributed.

11.2.1 Gibbs sampling algorithm

We can simulate ρ and μ using a Gibbs sampler. The idea behind this subsection is taken from (Collins, 2005, Appendix A), and applied to the problem at hand. The Gibbs sampling algorithm is defined in

Subsection 3.4.3, and often used when it is hard to sample from the posterior distribution and the prior is multidimensional (Szabó and van der Vaart, 2021, Section 3.4). In the mixed Poisson model, we deal with two parameters, and therefore we think this method will be helpful for the estimation of these parameters.

Now, suppose that we have n nurses other than Lucia de Berk. For each nurse i it holds that

$$\begin{aligned} N(t_i) \mid \Lambda_i = \lambda_i &\sim \mathit{Pois}(t_i\lambda_i), \\ \Lambda_i &\sim \mathit{Gam}(\rho, \rho/\mu), \end{aligned} \tag{11.1}$$

where $i = 1, \dots, n$. The quantity t_i is the number of shifts that nurse i worked, and λ_i is the intensity of nurse i witnessing or reporting an incident. We denote by x_i the number of incidents that nurse i experienced during t_i shifts. Hence it is the observation that belongs to $N(t_i)$. Since the number of incidents is time-dependent (the number of worked shifts), it can also be denoted by $x_i(t_i)$. Finally, define $\lambda = (\lambda_1, \dots, \lambda_n)$ as the vector of all intensities.

To be able to sample from the joint distribution of (ρ, μ), we need to know the full conditional distributions of $\rho|(\mu, \lambda_1, \dots, \lambda_n)$ and $\mu|(\rho, \lambda_1, \dots, \lambda_n)$. We assume that ρ and μ are distributed according to some prior density functions $\pi_0(\rho)$ and $\pi_0(\mu)$, respectively. Furthermore, assume that for the prior joint density function it holds that

$$\pi(\rho, \mu) = \pi_0(\rho)\pi_0(\mu).$$

Hence the prior distributions of the parameters ρ and μ are independent. From (11.1), we have the following density functions for the number of incidents that nurse i experiences and the intensity of witnessing or reporting an incident:

$$\begin{aligned} p_{\lambda_i}(x_i) &= \frac{(t_i\lambda_i)^{x_i}}{x_i!} e^{-t_i\lambda_i} \\ \pi(\lambda_i) &= \frac{(\rho/\mu)^\rho}{\Gamma(\rho)} \lambda_i^{\rho-1} e^{-(\rho/\mu)\lambda_i} \end{aligned}$$

We assume that the intensities $\lambda_1, \dots, \lambda_n$ of the nurses witnessing or reporting incidents are independent.

Now, we determine the full conditionals. First, we find the posterior distribution of λ_i given x_i for $i = 1, \dots, n$. It holds true by Bayes' formula, Theorem 3.1, that

$$\begin{aligned} \pi(\lambda_i|x_i) &\propto p_{\lambda_i}(x_i)\pi(\lambda_i) \\ &= \frac{(t_i\lambda_i)^{x_i}}{x_i!} e^{-t_i\lambda_i} \frac{(\rho/\mu)^\rho}{\Gamma(\rho)} \lambda_i^{\rho-1} e^{-(\rho/\mu)\lambda_i} \\ &\propto \lambda_i^{x_i+\rho-1} e^{-(t_i+\rho/\mu)\lambda_i}. \end{aligned}$$

Hence the posterior is $\mathit{Gam}(x_i + \rho, t_i + \rho/\mu)$ distributed, for $i = 1, \dots, n$. This was to be expected, since the gamma distribution is a conjugate prior of the Poisson distribution.

To determine the full conditional of μ, we need the joint distribution of $\lambda = (\lambda_1, \dots, \lambda_n)$. Using the independence assumption, it follows that

$$\begin{aligned} \pi(\lambda_1, \dots, \lambda_n|\rho, \mu) &= \prod_{i=1}^n \pi(\lambda_i|\rho, \mu) \\ &= \prod_{i=1}^n \frac{(\rho/\mu)^\rho}{\Gamma(\rho)} \lambda_i^{\rho-1} e^{-(\rho/\mu)\lambda_i} \\ &= \frac{(\rho/\mu)^{n\rho}}{\Gamma(\rho)^n} e^{-(\rho/\mu)\sum_{i=1}^n \lambda_i} \left[\prod_{i=1}^n \lambda_i^{\rho-1}\right]. \end{aligned}$$

Since we interpret μ as the overall probability of an incident per shift, we assume at this point that the prior of μ is $\mathit{Beta}(\alpha, 1)$ distributed. Setting the second parameter in the beta distribution equal to one leads to conjugacy. Hence

$$\pi_0(\mu) = \alpha\mu^{\alpha-1}, \qquad 0 < \mu < 1.$$

$$\begin{tikzcd}
\rho \sim \mathit{Exp}(1) \arrow[dr] & & \mu \sim \mathit{Beta}(0.01,1) \arrow[dl] & \\
& \lambda \sim \mathit{Gam}(\rho,\rho/\mu) \arrow[dr] & & (x,t) \sim \mathit{data} \arrow[dl] \\
& & N(t) \sim \mathit{Pois}(t\lambda) &
\end{tikzcd}$$

Figure 11.1: Hierarchical structure of the Bayesian model for the Gibbs sampler and No-U-Turn sampler.

Using Bayes' formula, Theorem 3.1, again and the assumption that $\pi(\rho,\mu) = \pi_0(\rho)\pi_0(\mu)$ we derive that

$$\begin{aligned}
\pi(\mu|\rho,\lambda_1,\dots,\lambda_n) &\propto \pi(\lambda_1,\dots,\lambda_n|\rho,\mu)\pi(\rho,\mu) \\
&= \frac{(\rho/\mu)^{n\rho}}{\Gamma(\rho)^n} e^{-(\rho/\mu)\sum_{i=1}^n \lambda_i} \left[\prod_{i=1}^n \lambda_i^{\rho-1}\right] \alpha\mu^{\alpha-1}\pi_0(\rho) \\
&\propto \mu^{\alpha-n\rho-1} e^{-(\rho/\mu)\sum_{i=1}^n \lambda_i}.
\end{aligned}$$

Hence the full conditional of μ given ρ and $\lambda = (\lambda_1,\dots,\lambda_n)$ is $\mathit{IG}(n\rho-\alpha, \rho\sum_{i=1}^n \lambda_i)$ distributed.[1] Thus μ^{-1} given ρ and $\lambda = (\lambda_1,\dots,\lambda_n)$ is $\mathit{Gam}(n\rho-\alpha, \rho\sum_{i=1}^n \lambda_i)$ distributed.

Now, we look at the full conditional of ρ. By Bayes' formula, Theorem 3.1, we find

$$\begin{aligned}
\pi(\rho|\mu,\lambda_1,\dots,\lambda_n) &\propto \pi(\lambda_1,\dots,\lambda_n|\rho,\mu)\pi(\rho,\mu) \\
&= \frac{(\rho/\mu)^{n\rho}}{\Gamma(\rho)^n} e^{-(\rho/\mu)\sum_{i=1}^n \lambda_i} \left[\prod_{i=1}^n \lambda_i^{\rho-1}\right] \alpha\mu^{\alpha-1}\pi_0(\rho) \\
&\propto \pi_0(\rho)\frac{(\rho/\mu)^{n\rho}}{\Gamma(\rho)^n} e^{-(\rho/\mu)\sum_{i=1}^n \lambda_i} \left[\prod_{i=1}^n \lambda_i^{\rho-1}\right].
\end{aligned}$$

Unfortunately, the above probability density function is quite difficult to draw random variables from. Hence to sample from it, we must turn to other methods than we are accustomed to. Usually, one uses methods like the inverse transform method or the acceptance-rejection method (Jianhua, n.d.).[2] However, another way to sample from the distribution of interest is by discretising the probability density function like Collins (2005) did in her thesis. We let ρ range from 0 to 1.5 with a step size of 0.005 and in this way create bins. We allocate a probability to each bin, equal to full conditional distribution at the value of ρ corresponding to the bin.

Let C be the proportionality constant of the above full conditional distribution of ρ. Then we obtain the following full conditional distributions for μ and ρ, respectively,

$$\begin{aligned}
\mu \mid \rho,\lambda_1,\dots,\lambda_n &\sim \mathit{IG}(n\rho-\alpha, \rho\sum_{i=1}^n \lambda_i), \\
\rho \mid \mu,\lambda_1\dots,\lambda_n &\sim \mathit{Discrete}(R,P),
\end{aligned}$$

where $R = (R_j)$ ranges from 0 to 1.5 with step size 0.005 and $P = (P_j)$ with

$$P_j = C^{-1}\cdot\pi_0(R_j)\frac{(R_j/\mu)^{nR_j}}{\Gamma(R_j)^n} e^{-(R_j/\mu)\sum_{i=1}^n \lambda_i} \left[\prod_{i=1}^n \lambda_i^{R_j-1}\right]. \tag{11.2}$$

As follows from the above, we have to specify the prior distribution of the parameter ρ and the parameter α of the prior distribution of μ. We choose an $\mathit{Exp}(1)$ distribution for the prior of ρ, which is

[1]Note that μ actually needs to be restricted to the interval $[0,1]$, but using the above will not form a problem since μ is almost surely smaller than one.

[2]We did look into the before mentioned methods, but the probability density function that we are dealing with is of such complex form because of the gamma function, that it is very difficult to apply them.

a somewhat arbitrary choice but should not have too much influence on the outcome. Furthermore, we choose $\alpha = 0.01$ such that the order of magnitude will be proper.

The hierarchical structure of the model in given in Figure 11.1. Since we added another layer of randomness in our model to be able to apply the Gibbs sampler, we will estimate the parameter ρ and μ by the posterior mean. It is the mean of the simulated values of the parameters.

Remark. The posterior distribution of a parameter represents the uncertainty of it, after combining the information in the data (the likelihood) with what we knew before collecting the data (the prior) (Stephens, 2017). Instead of the uncertainty of a parameter, we are more interested in the so-called best guess of the parameters. Therefore, we use point estimates. Examples of point estimates are the posterior mean and the posterior median, which are both suited for continuous distributions. One can also think of the posterior mode, but it is often only used for discrete random variables.

The posterior mean is the most popular point estimate, and depends on the shape of the distribution (Meredith, 2021a). We prefer to use the posterior median over the posterior mean when the posterior distribution is not symmetric, e.g., when the distribution has a right tail but is left truncated (Meredith, 2021b). The median is not influenced by the tail, but the mean indeed is.

Now, as we will see later it holds true that the posterior distribution of ρ and μ are quite symmetric, and therefore the posterior mean works well. During the application to an example, we also had a look at the results for the posterior median. It turns out that the differences between the two point estimates are small, and the posterior mean only works a little bit better.

From the report of Elffers (2002), we know that a maximum of 27 nurses worked on the ward of the Juliana's Children Hospital during the period that Lucia de Berk worked there. Unfortunately, we do not have such information about the other wards of the Red Cross Hospital. Therefore, we must estimate the number of nurses n in another way.

De Vos (2004) says that one needs information from approximately ten similar hospitals, to say something about the variation coefficient of the incident rate. Only 27 nurses will be insufficient to be sure that we estimate the parameters in a correct way. However, at the time that Lucia de Berk worked at the Juliana Children's Hospital and Red Cross Hospital, around 2001, there were only seven hospitals specialised in children's care (Volksgezondheidenzorg.info, 2021). Therefore, de Vos' assumption is somewhat ambitious. From (11.2) it also follows that n has quite some influence on the full conditional distribution of ρ, e.g., the gamma function denominator grows rapidly with n as $R_j < 1$ and decreases as $R_j > 1$. After some test runs of our algorithm, we know that information from three similar hospitals, hence from $27 \times 3 = 81$ nurses is sufficient and ensures no errors during the performance. Thus we choose to simulate data from $n = 81$ nurses. Note that it is also representative in the context of our problem since De Berk worked at three hospital wards.

Application: Gibbs sampler

Since we do not have access to real data, we simulate $n = 81$ nurses from a $NB(\rho_0, (1 + r\mu_0/\rho_0)^{-1})$ distribution with parameters $\rho_0 = 1, \mu_0 = 27/1734$ and the number of shifts equal to $r = 201$. (It simplifies to a geometric distribution.) Thus, we assume that all nurses work the same number of shifts, which is the same number of shifts that Lucia de Berk worked. Note that it does not represent reality very much, but it will make our model simpler to verify using it as a ground truth.[3] The data set with simulated incidents of $n = 81$ nurses is given in Appendix E.2.1, Table E.3. The corresponding distribution of the data set compared with the true probability mass function is given in Appendix C.4.1, Figure C.5.

We are aware that using a simulated data set can have some influence on the outcome of our algorithm. It is possible that one simulated set contains more outliers than another set, and therefore performs worse.[4] However, we would still like to show the results of the Gibbs sampler for our simulated data, since we think that they show the purpose of it nicely.

To remind us, our goal is to use the Gibbs sampling algorithm to compute the probability $P(N(t) \geq 14)$ given the number of incidents $x_1, \ldots, x_n$ that the n nurses experienced. Using the above information, our Gibbs sampler should give us back the values of the parameters ρ and μ close to the true values ρ_0 and μ_0, respectively. Both parameters are estimated as the average of the simulated values post burn-in. The estimated value of $P(N(t) \geq 14)$, also defined as the average of the simulated values post burn-in, should also be somewhat close to the true value. This probability was approximately 0.0206, see (8.3). Hence using the simulated data, we can check if our Gibbs sampler works decently.

[3] It was one of the assumptions that Collins (2005) made in her model as well.

[4] We will dive further into this phenomenon below when we discuss the sensitivity to outliers of our Gibbs sampler.

We take the initial values equal to $\rho^0 = 1$ and $\mu^0 = 0.015$ and run our algorithm $T = 10000$ times with a burn in period of $m = 3000$ iterations. The initial values are chosen such that they are somewhat close to the true values from which we sample. In Appendix D.1.1, Listing D.1, one can find our algorithm, programmed in R. The corresponding pseudo code is given below. In the pseudo code, the superscripts identify the time.

For $i = 1, \ldots, n$ do

  Generate x_i from $NB\left(\rho_0, (1 + t\mu_0/\rho_0)^{-1}\right)$

End for

Given $\rho^0 = 1$

Given $\mu^0 = 0.015$

For $t = 1, \ldots, T$ do

  For $i = 1, \ldots, n$ do

    Generate λ_i^t from $Gam\left(x_i + \rho^{t-1}, r + \rho^{t-1}/\mu^{t-1}\right)$

  End for

  Generate μ^t from $IG\left(n\rho^{t-1} - \alpha, \rho^{t-1}\sum_{i=1}^{n}\lambda_i^t\right)$

  Generate ρ^t from $Discrete(R, P)$ with $R = (R_j)$ ranging

   from 0 to 1.5 with step size 0.005 and $P = (P_j)$

   with $P_j = C^{-1} \cdot \pi_0(R_j)\dfrac{(R_j/\mu^t)^{nR_j}}{\Gamma(R_j)^n}e^{-(R_j/\mu^t)\sum_{i=1}^{n}\lambda_i}\left[\prod_{i=1}^{n}\lambda_i^{R_j-1}\right]$

  Compute $P(N^t \geq 14)$ with $N^t \sim NB\left(\rho^t, (1 + t\mu^t/\rho^t)^{-1}\right)$

End for

Appendix C.4.1, Figures C.6a and C.6b, present the traceplots of the parameters ρ and μ. They show us the evolution of the parameters over time. The figures imply that the values of the parameters do fluctuate a lot over the course of the algorithm, where the values of ρ are more spread than those of μ. We can also conclude that the estimated value of the parameter μ lies close to the true value $\mu_0 = 27/1734$. The estimated value of ρ deviates more from the true value $\rho_0 = 1$.

Figure C.7 shows the course of $P(N(t) \geq 14)$. Even though not all parameters are close to their true values, the probability $P(N(t) \geq 14)$ is estimated quite well by the average of the simulated values. Yet the (relative) standard deviation is quite large compared to the mean.

Table 11.1 shows a comparison of the true values of $P(N(t) \geq 14)$, ρ and μ with their estimated values and standard deviations. Additionally, Appendix C.4.1, Figures C.8a and C.8b show the conditional distributions of the parameters ρ and μ. From Table 11.1 we can conclude that μ and $P(N(t) \geq 14)$ are estimated pretty good as well, and the algorithm seems to perform worse for the heterogeneity parameter ρ. Still, note that the estimation of the parameters partly depends on the simulated data set

Quantity of interest	**True value**	**Estimated value**	**Standard deviation**
$P(N(t) \geq 14)$	0.02061608	0.02296840	0.012096153
ρ	1.00000000	0.77979289	0.168898934
μ	0.01557093	0.01413278	0.002073995

Table 11.1: Overview of the true values of $P(N(t) \geq 14)$, ρ and μ with their estimated values and standard deviations using the Gibbs sampling algorithm.

of $n = 81$ observations. For example, this data set contains more nurses that experienced zero or only one incident(s) than one expects, due to it being a random draw from the negative binomial distribution.

Of course, it is possible to estimate $P(N(t) \geq 14)$ by plugging in the estimated values of ρ and μ in (8.1) as well. We obtain

$$P(N(t) \geq 14) \approx 0.02104627,$$

which is an even better approximation if we compare it with the ground truth.

Efficiency of the Gibbs sampler

In general, the convergence of the Gibbs sampling algorithm depends on the size n of the data set. The more information is available, the better the performance of the algorithm. However, having information about many nurses, which was what de Vos (2004) assumed, is not always realistic. Furthermore, if one looks closely to (11.2), the size of n and the range of ρ do have a lot of influence on how big or small (11.2) becomes. Some test runs of the algorithm show that if one wants to consider information about even more nurses, one must decrease the range of ρ and vice versa. If we do not take this into account, we obtain values at zero or infinity (not taking the proportionality constant into account). We already mentioned that it is a consequence of, e.g., the gamma function in the denominator of (11.2). Hence it can prevent us from using the Gibbs sampling algorithm the way we implemented it. Moreover, in theory, the parameter ρ can take values from zero to infinity and is therefore not restricted to some finite range. By making a choice for the maximum, we cannot explore values of ρ above 1.5, which is a pity since it is possible that these values of ρ can explain the data better.

In addition, we found that the algorithm does not necessarily perform better when one runs it for more iterations. Of course, more iterations lead to more computation time. E.g., the computation time is approximately 70 seconds for $T = 10000$ iterations (iMac 24-inch, M1, 2021, 16 GB unified memory). Therefore, it does not matter to run the algorithm for, e.g, only $T = 100000$ iterations with a burn in period of $m = 30000$ instead of the smaller number of iterations that we used before. One should let the algorithm perform for enough iterations for representative results.

Sensitivity to outliers: Gibbs sampler

To wrap up this subsection about the Gibbs sampling algorithm, we look at the sensitivity to outliers of this method. Suppose that we replace one of the data points in our set by an erroneous value that is much bigger than the rest. Then we would like to know what influence it has on the performance of the algorithm, and if the method is resistant to flawed data points.

The maximum number of incidents in the original data set of 81 nurses, given in Appendix E.2.1, was 15 incidents. In this set, we replaced the first data point $x_1 = 10$ by the erroneous values $x_1 = 100$ and $x_1 = 50$ sequentially. Then we ran our Gibbs sampler again. The results are presented in Table 11.2 together with the results of the original data set. Appendix C.4.1, Figures C.9 and C.10 show the corresponding traceplots based on the erroneous data sets. From Table 11.2 we derive that the estimation of $P(N(t) \geq 14)$ and the parameter ρ are the most affected by the errors. It was to be expected that the deviations from the original estimations are the greatest when we have a very large data flaw. We see that the standard deviations grow with the size of the flawed data point, except the ones of ρ. In this case, they decrease. Perhaps, it is caused by the restricted range of ρ. Only a smaller span of values can explain the outliers.

We conclude that μ can still be estimated reasonably and is therefore quite robust. However the estimated value of ρ is rather off, and it affects the estimation of $P(N(t) \geq 14)$ as well. Erroneous data leads to more optimistic probabilities, in favour of the suspect.

Quantity of interest	**Estimated value** ($x_1 = 100$)	**Standard deviation** ($x_1 = 100$)	**Estimated value** ($x_1 = 50$)	**Standard deviation** ($x_1 = 50$)	**Estimated value (Original)**	**Standard deviation (Original)**
$P(N(t) \geq 14)$	0.07365118	0.024090346	0.04624563	0.019001934	0.02296840	0.012096153
ρ	0.48180046	0.087184353	0.59067847	0.118639055	0.77979289	0.168898934
μ	0.01995040	0.003579974	0.01674288	0.002792654	0.01413278	0.002073995

Table 11.2: Overview of the estimated values and standard deviations of $P(N(t) \geq 14), \rho$ and μ for the flawed data sets using the Gibbs sampling algorithm.

11.2.2 Method of moments

We can also estimate the parameters of the mixed Poisson model in other ways. Most of the time, maximum likelihood estimators (MLE) are the preferred method to estimate parameters, since they are asymptotically efficient and, under regularity and identifiability conditions, consistent for the true parameters (van der Vaart, 1998, Section 3.3). However, moment-based estimators can be as efficient as maximum likelihood estimators for the negative binomial distribution (Savani and Zhigljavsky, 2006). Furthermore, the negative binomial distribution that we are dealing with is quite complicated notation wise, because of the specific form of the gamma distribution. So, finding the maximum likelihood estimators analytically will be difficult (even for a fixed time period t). Therefore, we want to have a look at what moments estimators can mean for us.

From Subsection 8.2.1 we know that the number of incidents that a nurse experiences during t shifts is $NB(\rho, (1 + t\mu/\rho)^{-1})$ distributed. Our goal is to estimate ρ and μ based on the data of a group of n nurses that works in a hospital ward. The information that we assume is available, contains the number of shifts t_i that each nurse worked during a predetermined period and the number of incidents that the nurse experienced x_i during the t_i worked shifts, for $i = 1, \ldots, n$. Note that the number of incidents is time-dependent and therefore can also be denoted by $x_i(t_i)$.

Before deriving the moment estimators for ρ and μ, we give a small explanation how moment estimation works. Let $X_1, \ldots, X_n$ be a sample from a distribution that depends on a k-dimensional parameter θ. The method of moments, see for example (van der Vaart, 1998, Section 3.3), proposes to estimate the parameter vector θ by the solution of a system of equations,

$$\frac{1}{n}\sum_{i=1}^{n} f_j(X_i) = E_\theta f_j(X), \qquad j = 1, \ldots, k,$$

for given functions $f_1, \ldots, f_k$. Hence the parameter is chosen such that the sample moments match the theoretical moments. In the context of hospital incidents, we want to match the first two moments of $N(t)$, the number of incidents that a nurse witnesses during time period t, to the sample moments. In Subsection 8.2.1, we already found that

$$\begin{aligned} E(N(t)) &= t\mu, \\ Var(N(t)) &= t\mu + \frac{(t\mu)^2}{\rho}. \end{aligned}$$

Thus the second moment of $N(t)$ equals

$$\begin{aligned} E(N^2(t)) &= Var(N(t)) + [E(N(t))]^2 \\ &= t\mu + \frac{(t\mu)^2}{\rho} + (t\mu)^2 \\ &= t\mu + (t\mu)^2\left(1 + \frac{1}{\rho}\right). \end{aligned}$$

Hence it follows that the moments are time dependent. This time dependency can be a problem if the number of shifts of the nurses are different, which is most often the case. We can in some way avoid this difficulty, by looking at the sum of first and second moments for each number of shifts t_i. In the upcoming part, we use the following notation for the first two sample moments, where $y_1, \ldots, y_n$ are some observations:

$$\begin{aligned} \overline{y} &:= \frac{1}{n}\sum_{i=1}^{n} y_i \\ \overline{y^2} &:= \frac{1}{n}\sum_{i=1}^{n} y_i^2 \end{aligned}$$

Finally, we denote the moment estimators of the parameters ρ and μ by $\hat{\rho}_{MOM}$ and $\hat{\mu}_{MOM}$, respectively.

For $i = 1, \ldots, n$, it holds true that

$$\begin{aligned} E(N(t_i)) &= t_i\mu, \\ E(N^2(t_i)) &= t_i\mu + (t_i\mu)^2\left(1 + \frac{1}{\rho}\right). \end{aligned}$$

The expectation of $N(t_i)$ and $N^2(t_i)$ can each be estimated by the number of incidents x_i that a nurse witnessed, and the squared number of incidents x_i^2, respectively. Since ρ and μ will be the same for all combinations of incidents and shifts, we can estimate the parameters by looking at the sum of the expectations of $N(t_i)$ and $N^2(t_i)$, for $i = 1, \ldots, n$, as we have stated before.

Then it holds true by linearity of expectations that

$$\begin{aligned} E\left(\sum_{i=1}^{n} N(t_i)\right) &= \sum_{i=1}^{n} E(N(t_i)) \\ &= \mu \sum_{i=1}^{n} t_i \\ &\approx \sum_{i=1}^{n} x_i, \end{aligned}$$

where we can perform the final approximation since we know how many incidents each nurse experienced during their t_i shifts. Hence we estimate $E(N(t_i))$ by x_i, for $i = 1, \ldots, n$. Therefore, we find that the moment estimator for μ equals

$$\hat{\mu}_{MOM} = \frac{\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} t_i} = \frac{n\overline{x}}{n\overline{t}} = \frac{\overline{x}}{\overline{t}}. \tag{11.3}$$

The quantity $\hat{\mu}_{MOM}$ is the average number of incidents per shift. This estimate is consistent with the one of Gill et al. (2018), which was that μ is the overall probability of an incident per shift. They estimated it as the ratio of total number of incidents to total number of shifts, which is exactly $\hat{\mu}_{MOM}$.

Now, we also want to find the moment estimator of ρ. Therefore, we will look at the sum of the second moments. It holds true that

$$\begin{aligned} E\left(\sum_{i=1}^{n} N^2(t_i)\right) &= \sum_{i=1}^{n} E(N^2(t_i)) \\ &= \mu \sum_{i=1}^{n} t_i + \mu^2 \left(1 + \frac{1}{\rho}\right) \sum_{i=1}^{n} t_i^2 \\ &\approx \sum_{i=1}^{n} x_i^2. \end{aligned}$$

In a similar fashion as for the first moments, we can approximate the sum of second moments by the sum of squared number of incidents for each nurse. Thus $E(N^2(t_i))$ is estimated by x_i^2, for $i = 1, \ldots, n$. Solving the above equation for ρ and plugging in the moment estimator for μ, we conclude that

$$\hat{\rho}_{MOM} = \frac{\left(\frac{\overline{x}}{\overline{t}}\right)^2}{\frac{\overline{x^2}}{\overline{t^2}} - \left(\frac{\overline{x}}{\overline{t}}\right)^2 - \frac{\overline{x}}{\overline{t^2}}}. \tag{11.4}$$

The full derivation of $\hat{\rho}_{MOM}$ can be found in Appendix A.6, Lemma A.2.

To summarise, we found that the moment estimator for (ρ, μ) is

$$(\hat{\rho}_{MOM}, \hat{\mu}_{MOM}) = \left(\frac{\left(\frac{\overline{x}}{\overline{t}}\right)^2}{\frac{\overline{x^2}}{\overline{t^2}} - \left(\frac{\overline{x}}{\overline{t}}\right)^2 - \frac{\overline{x}}{\overline{t^2}}}, \frac{\overline{x}}{\overline{t}} \right). \tag{11.5}$$

The above shows that one needs the distribution of the number of incidents and shifts to be able to estimate ρ and μ using moment estimation. Gill et al. (2018) only had access to information about the total number of incidents and shifts, and information about one particular nurse. Therefore, they could estimate the parameter μ but not the parameter ρ.

Suppose that we do not know the specific number of shifts that each nurse worked, but we do know the number of incidents that any of them experienced. We still look at a fixed time period, which is translated to the total number of shifts available during the period of interest. Then we can still estimate ρ and μ as follows. Keep in mind that in this case x_i is the number of incidents that a nurse witnessed during the whole time period t. Then it holds true that

$$\hat{\mu}_{MOM} = \frac{\overline{x}}{t} \tag{11.6}$$

and

$$\begin{aligned}\hat{\rho}_{MOM} &= \frac{\left(\frac{\overline{x}}{t}\right)^2}{\frac{\overline{x^2}}{t^2} - \left(\frac{\overline{x}}{t}\right)^2 - \frac{\overline{x}}{t^2}} \\ &= \frac{\overline{x}^2}{\overline{x^2} - \overline{x}^2 - \overline{x}} \\ &= \frac{\overline{x}^2}{s^2 - \overline{x}},\end{aligned} \tag{11.7}$$

where

$$s^2 := \overline{x^2} - \overline{x}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

is the sample variance. By using both moment estimators in our mixed Poisson model, we can calculate the probability that a nurse witnessed a certain number of incidents during some time period. The moment estimators may be less precise, since we cannot correct for the number of shifts that each nurse worked. However, we think that it is a good alternative when less information is available.

Furthermore, if we have a fixed time period t, then one can check that the moment estimator of μ coincides with the maximum likelihood estimator of μ.[5] It is hard to compare the moment estimator of ρ with the one that we obtain with maximum likelihood estimation, since the latter can only be found numerically using, e.g., Newton's method.[6] Regarding our first method using the number of shifts and incidents of each nurse to estimate the parameters, it becomes even more involved to determine the maximum likelihood estimators for ρ and μ algebraically. Therefore, we choose not to dive further into maximum likelihood estimation.

Application: method of moments

To study how well the method of moments works, we simulate the incidents that a group of $n = 1000$ nurses experienced, each during $t = 201$ shifts. We sample them from a $NB(\rho_0, (1 + t\mu_0/\rho_0)^{-1})$ distribution with parameters $\rho_0 = 1$ and $\mu_0 = 27/1734$, like we also did for the Gibbs sampler in the previous subsection. The data set can be found in Appendix E.2.2, Table E.4. We look at the estimated values of ρ and μ based on the information of a group of n nurses, with n ranging from 1 to 1000. The results are shown in Appendix C.4.2, Figures C.12 and C.13 for ρ and μ, respectively. Figures C.12a and C.13a shows the courses of ρ and μ, respectively, for up to 1000 nurses. Figures C.12b and C.13b gives a closeup for 300 nurses.

From Figure C.12a it follows that for our particular simulated data set, $\hat{\rho}_{MOM}$ converges to the true value $\rho_0 = 1$ based on the number of incidents of approximately 200 nurses. Figure C.13a implies that we need information about the incidents of about 350 nurses for μ to be close to the true value $\mu_0 = 27/1734$. Hence it follows that with enough information moment estimation works sufficiently. The result also implies that the moment estimators will likely be asymptotically consistent, which we will prove in the upcoming part. In addition, we will look at possible unbiasedness and asymptotic normality of the estimators. We end this subsection with a discussion about the sensitivity to outliers of the moment estimators $\hat{\rho}_{MOM}$ and $\hat{\mu}_{MOM}$.

Efficiency of the method of moments

First, we show that $\hat{\mu}_{MOM}$ given in (11.3) is an unbiased estimator for μ. It holds true that

$$\begin{aligned}E(\hat{\mu}_{MOM}) &= E\left[\frac{\sum_{i=1}^n N(t_i)}{\sum_{i=1}^n t_i}\right] \\ &= \frac{\sum_{i=1}^n E(N(t_i))}{\sum_{i=1}^n t_i} \\ &= \frac{\sum_{i=1}^n t_i\mu}{\sum_{i=1}^n t_i} \\ &= \mu.\end{aligned}$$

[5]See Appendix A.6, Lemma A.3.
[6]See Appendix A.6, Lemma A.4.

Now, suppose that μ is known. Then $1/\hat{\rho}_{MOM}$ is an unbiased estimator for $1/\rho$. By Appendix A.6, Lemma A.2, we know that

$$\frac{1}{\hat{\rho}_{MOM}} = \frac{1}{\mu^2}\frac{\overline{x^2}}{\overline{t^2}} - \frac{1}{\mu}\frac{\overline{t}}{\overline{t^2}} - 1.$$

Then it follows that

$$\begin{aligned}
E\left[\frac{1}{\hat{\rho}_{MOM}}\right] &= \frac{1}{\mu^2}E\left[\frac{\sum_{i=1}^n N^2(t_i)}{\sum_{i=1}^n t_i^2}\right] - \frac{1}{\mu}\frac{\sum_{i=1}^n t_i}{\sum_{i=1}^n t_i^2} - 1\\
&= \frac{1}{\mu^2}\frac{\sum_{i=1}^n E(N^2(t_i))}{\sum_{i=1}^n t_i^2} - \frac{1}{\mu}\frac{\sum_{i=1}^n t_i}{\sum_{i=1}^n t_i^2} - 1\\
&= \frac{1}{\mu^2}\frac{\sum_{i=1}^n t_i\mu}{\sum_{i=1}^n t_i^2} + \frac{1}{\mu^2}\frac{\sum_{i=1}^n (t_i\mu)^2\left(1+\frac{1}{\rho}\right)}{\sum_{i=1}^n t_i^2} - \frac{1}{\mu}\frac{\sum_{i=1}^n t_i}{\sum_{i=1}^n t_i^2} - 1\\
&= \frac{1}{\mu}\frac{\sum_{i=1}^n t_i}{\sum_{i=1}^n t_i^2} + 1 + \frac{1}{\rho} - \frac{1}{\mu}\frac{\sum_{i=1}^n t_i}{\sum_{i=1}^n t_i^2} - 1\\
&= \frac{1}{\rho}.
\end{aligned}$$

Hence we conclude that $\hat{\mu}_{MOM}$ is an unbiased estimator and $1/\hat{\rho}_{MOM}$ is an unbiased estimator, if μ is known.

Furthermore we show that, for fixed t, the moment estimators $\hat{\rho}_{MOM}$ and $\hat{\mu}_{MOM}$ are asymptotically consistent and asymptotically normal. By the weak law of large numbers (van der Vaart, 1998, Example 1.2), it holds that $\overline{x} \xrightarrow{\text{P}} E(N(t))$ and $\overline{x^2} \xrightarrow{\text{P}} E(N^2(t))$. Then by the continuous mapping theorem (van der Vaart, 1998, Theorem 1.7), it follows that

$$\hat{\mu}_{MOM} = \frac{\overline{x}}{t} \xrightarrow{\text{P}} \frac{E(N(t))}{t} = \frac{t\mu}{t} = \mu.$$

Again by the continuous mapping theorem (van der Vaart, 1998, Theorem 1.7), we have that $\overline{x}^2 \xrightarrow{\text{P}} [E(N(t))]^2$ and $s^2 \xrightarrow{\text{P}} Var(N(t))$. Then it holds true by similar reasoning that

$$\hat{\rho}_{MOM} = \frac{\overline{x}^2}{s^2 - \overline{x}} \xrightarrow{\text{P}} \frac{[E(N(t))]^2}{Var(N(t)) - E(N(t))} = \frac{(t\mu)^2}{t\mu + \frac{(t\mu)^2}{\rho} - t\mu} = \rho.$$

Finally by (van der Vaart, 1998, Theorem 3.7), under some regularity conditions, the moment estimators $\hat{\rho}_{MOM}$ and $\hat{\mu}_{MOM}$ are asymptotically normally distributed for fixed t. More details are given in Appendix A.6, Theorem A.6.

Sensitivity to outliers: method of moments

Like we did for the Gibbs sampling algorithm, we also look here at the sensitivity to outliers of the method of moments. We still use the same data set of a maximum of $n = 1000$ nurses as before, given in Appendix E.2.2. The maximum number of incidents in this set was 24. Similarly, we replace the first data point $x_1 = 4$ by the erroneous values $x_1 = 100$ and $x_1 = 50$ sequentially and compute the moment estimators $\hat{\rho}_{MOM}$ and $\hat{\mu}_{MOM}$. The results are presented in Appendix C.4.2, Figures C.14 and C.15 respectively.

The traceplots show that estimation of μ approaches the correct value quite fast. Using the number of incidents of about 200 nurses, the moment estimator $\hat{\mu}_{MOM}$ is already close to the true value $\mu_0 = 27/1734$. Hence we conclude that the moment estimator $\hat{\mu}_{MOM}$ is robust. This makes sense since it is also the maximum likelihood estimator which are often robust estimators (Weisstein, n.d.c). However, the moment estimator $\hat{\rho}_{MOM}$ of ρ is far off in both cases. For the case with the first data point equal to $x_1 = 100$ and using the incidents of 1000 nurses to estimate the parameter, the moment estimator $\hat{\rho}_{MOM}$ is still not close enough to say it is resistant to errors. It becomes better for the case $x_1 = 50$ but is still not “perfect” like in the original estimation. Since ρ depends on the sample variance of the data, it has major impact on the denominator of the moment estimator. Hence the sample variance becomes larger when there is a great difference between data points, and it follows that ρ grows much smaller.

We like to finish this subsection, by saying that the large underestimation of the parameter ρ and slight overestimation of the parameter μ lead to overestimation of the probability $P(N(t) \geq 14)$.

11.2.3 Hamiltonian Monte Carlo

Finally, we discuss a method that shows some resemblance with the Gibbs sampling algorithm. Instead of implementing a numerical method ourselves like we did before, we can also use some already programmed tools that we can apply to make the estimation of the parameters of a hierarchical model easier. One of those tools is the PyMC3 package for Python, see for example Salvatier et al. (2016). PyMC3 is a probabilistic programming package that allows us to fit a Bayesian model using different numerical methods, such as Markov Chain Monte Carlo and variational inference.

To perform Markov Chain Monte Carlo methods, the PyMC package offers different step methods. By default, it uses for example the Metropolis-Hastings algorithm for discrete random variables and the No-U-turn sampler for continuous random variables. Both are discussed in Subsection 3.4.3. Our motivation to dive into the numerical methods presented in the package besides our self-implemented Gibbs sampling algorithm, is that we can adapt our model in PyMC3 much easier. We do not necessarily have to make conjugacy assumptions for the prior distributions since we do not have to determine the full conditional distributions of the parameters of interest. Furthermore, using PyMC3 we can use Hamiltonian Monte Carlo methods such as the No-U-Turn sampler, which ensure much faster convergence to target distributions than Gibbs sampling methods (Hoffman and Gelman, 2014), without having to implement it ourselves.

Our original model, proposed by Gill et al. (2010), is given by (11.1). Since we are interested in estimating the parameters ρ and μ of the gamma distribution, we need to add another layer of priors for the parameters to our hierarchical model. In the implementation of the Gibbs sampling algorithm in Subsection 11.2.1 we assumed that the prior distributions of ρ and μ are independent, and that the prior of ρ is $Exp(1)$ distributed and the prior of μ is $Beta(0.01, 1)$ distributed. We will also use these assumptions for our implementation in PyMC3. Hence, we obtain the final hierarchical model:

$$\begin{aligned} N(t) \mid \Lambda = \lambda &\sim Pois(t\lambda) \\ \Lambda \mid (R, M) = (\rho, \mu) &\sim Gam(\rho, \rho/\mu) \\ R &\sim Exp(1) \\ M &\sim Beta(0.01, 1) \end{aligned} \tag{11.8}$$

This structure is given in Figure 11.1 as well.

One can find the implementation in Python in Appendix D.1.2, Listing D.2. As input, we need a vector of the incidents that the group nurses of interest experienced and a corresponding vector of the worked shifts of each nurse. We choose as the step method the No-U-Turn algorithm, hoping it will make our algorithm faster than the Gibbs sampling algorithm. Finally, we need for every nurse a corresponding parameter λ. Hence the dimension of a vector λ equals the number of nurses and is controlled by the shape parameter in PyMC3. As before, the parameters ρ and μ are estimated by their posterior means.

Application: No-U-Turn sampler

For application, we use the same simulated data set as for the Gibbs sampling algorithm, given in Appendix E.2.1, Table E.3, sampled from a $NB(\rho_0, (1 + r\mu_0/\rho_0)^{-1})$ distribution with parameters $\rho_0 = 1, \mu_0 = 27/1734$ and $r = 201$. The data set consists of the number of incidents that $n = 81$ nurses experienced during 201 shifts.

We choose to run the algorithm using four chains where each one consists of 1000 iterations. Appendix C.4.3, Figure C.18 shows the posterior density and the traceplots of the parameters ρ and μ. Forest plots of both parameters are given in Figure C.19 with 94% highest posterior density intervals. Figure C.20 shows the traceplot of $P(N(t) \geq 14)$.

In the same way as in the Gibbs sampling algorithm, we estimate the values of ρ and μ by the average of the simulated values, hence posterior mean, as well as the estimated value of $P(N(t) \geq 14)$. The results can be found in Table 11.3. From this table we can conclude that the heterogeneity parameter ρ is underestimated, which may be due to the randomly generated data set. The parameter μ and the probability $P(N(t) \geq 14)$ are estimated nicely. One might already notice that these results are close to the ones that we found using the Gibbs sampling algorithm.

Again, we can also compute $P(N(t) \geq 14)$ by plugging in the estimators for ρ and μ instead of the sample average. We obtain

$$P(N(t) \geq 14) \approx 0.020747,$$

Quantity of interest	True value	Estimated value	Standard deviation
$P(N(t) \geq 14)$	0.020616	0.022766	0.012361
ρ	1.000000	0.789139	0.175761
μ	0.015571	0.014140	0.002102

Table 11.3: Overview of the true values of $P(N(t) \geq 14)$, ρ and μ with their estimated values and standard deviations using the No-U-Turn sampler.

Quantity of interest	Estimated value ($x_1 = 100$)	Standard deviation ($x_1 = 100$)	Estimated value ($x_1 = 50$)	Standard deviation ($x_1 = 50$)	Estimated value (Original)	Standard deviation (Original)
$P(N(t) \geq 14)$	0.073724	0.024082	0.045435	0.019029	0.022766	0.012361
ρ	0.479113	0.089312	0.601485	0.121457	0.789139	0.175761
μ	0.019938	0.003577	0.016713	0.002750	0.014140	0.002102

Table 11.4: Overview of the estimated values and standard deviations of $P(N(t) \geq 14), \rho$ and μ for the flawed data sets using the No-U-Turn sampler.

which is almost the same as the ground truth of $P(N(t) \geq 14) \approx 0.0206$. Hence for this data set we may prefer this method of estimating the desired probability.

Efficiency of the No-U-Turn sampler

In contrast to the Gibbs sampling algorithm that we implemented before; this model programmed in PyMC3 has no limited range for the parameter ρ. There is also no trade-off between the number of data points n and the range of ρ. Hence, we can use as much data as we want.

The computation time for four chains with 1000 iterations each is approximately 8 seconds (iMac 24-inch, M1, 2021, 16 GB unified memory), which is quite fast. The computation time increases if, e.g., we increase the number of data points or the number of iterations of each chain.

If we create a summary of statistics, we see that the Gelman-Rubin diagnostic $\hat{R}$ are for all parameters ρ, μ and the vector λ close to one. This quantity tests for lack of convergence by comparing the variance between multiple chains to the variance within each chain (Salvatier et al., 2016). It holds that the between-chain and within-chain variance should be identical. If $\hat{R}$ is greater than one, it means that one or more of the chains have not yet converged. Hence it is a very good sign that for our model they are close to one.

Sensitivity to outliers: No-U-Turn sampler

Lastly, we look at the sensitivity to outliers. In the data set of 81 nurses given in Appendix E.2.1, we replaced the first data point $x_1 = 10$ by the erroneous values $x_1 = 100$ and $x_1 = 50$ sequentially. This is the same as we did for the Gibbs sampling algorithm and the method of moments. The results are presented in Table 11.4 together with the original results. Appendix C.4.3, Figures C.21 and C.23 show the traceplots and posterior densities of the parameters ρ and μ based on the erroneous data sets. Figures C.22 and C.24 show the forest plots with 94% highest posterior density intervals.

The estimated values and standard deviations almost coincide with the ones that we obtained using the Gibbs sampling algorithm. Hence, we can draw the same conclusion. The parameter ρ is mostly affected by the errors, whereas μ only deviates a little bit from the original estimation. Hence μ is a quite robust estimator, where ρ is not.

11.3 Comparison

Now, we would like to compare the three methods discussed before and look at which one we prefer. Therefore, we will first look at a situation that we discussed before. We consider the data set of 81 nurses used in Subsection 11.2.1, given in Appendix E.2.1. We estimate the parameters ρ and μ applying all three methods, and then use them to estimate $P(N(t) \geq 14)$ using the algorithms. To remind us, we assumed that each nurse works the same number of shifts, $t = 201$, and the true values of the parameters from which we sample are $\rho_0 = 1$ and $\mu_0 = 27/1734$.

Method	ρ	μ	$P(N(t) \geq 14)$
Gibbs sampler	0.783405	0.014126	0.022864
Method of moments	0.79274	0.013881	0.019436
No-U-Turn sampler	0.789139	0.014153	0.022714

Table 11.5: Comparison of estimators for the parameters ρ and μ and the probability $P(N(t) \geq 14)$ using different methods.

Method	ρ	μ	$P(N(t) \geq 14)$
Gibbs sampler	0.480542	0.019932	0.073602
Method of moments	0.12324	0.019409	0.085598
No-U-Turn sampler	0.479113	0.019938	0.073724

Table 11.6: Comparison of estimators for the parameters ρ and μ and the probability $P(N(t) \geq 14)$ using different methods. Here we replaced the first data point of the original data set with an erroneous value equal to $x_1 = 100$.

Method	ρ	μ	$P(N(t) \geq 14)$
Gibbs sampler	0.599253	0.016689	0.045381
Method of moments	0.29993	0.016338	0.063959
No-U-Turn sampler	0.601485	0.016713	0.045435

Table 11.7: Comparison of estimators for $P(N(t) \geq 14)$ and the parameters ρ and μ using different methods. Here we replaced the first data point of the original data set with an erroneous value equal to $x_1 = 50$.

Table 11.5 presents the results. We see that the estimations of the parameters ρ and μ and the probability $P(N(t) \geq 14)$ are very close, almost identical. Based on these results for a simulated data set, it thus does not really matter which method to apply for the estimation of the parameters of the mixed Poisson model. They all give the same outcome.

Therefore, we will also compare the robustness of the different estimators. Tables 11.6 and 11.7 show the estimations for the flawed data sets with $x_1 = 100$ and $x_1 = 50$, respectively. Again, the parameter μ is estimated almost identically for the Gibbs sampling algorithm, the method of moments and the No-U-Turn sampler. However, we see a remarkable difference for the estimation of the parameter ρ. The numerical methods estimate ρ closer to the truth than the method of moments. Maybe this result indicates that we should prefer the Gibbs sampling algorithm or the No-U-Turn sampler in this case.

Besides sensitivity to outliers, there are other things to consider when choosing a method for parameter estimation. For example, the method of moments is very simple and costs almost no computation time compared to the Gibbs sampling algorithm or the No-U-Turn sampler. Nonetheless, the No-U-Turn sampler is a lot faster than the Gibbs sampler (approximately ten seconds versus a little more than one minute). Furthermore, we already discussed that the correlation between the range of ρ and the number of data points that is used in the Gibbs sampling algorithm makes it somewhat restricted. We do not have this problem for our implemented No-U-Turn sampler.

11.3.1 Simulation study

In Appendix B we consider some more test scenarios that can be applied to the three methods that we discussed. In summary, we examined the outcomes of the three estimation methods using simulated data while varying the true value of ρ, the number of data points n and the group composition of the nurses. In the last case, we discriminated between nurses that all work the same number of shifts, and a group of nurses that work different numbers of shifts.

Using Analysis of Variance and Tukey's test we found that for the parameter μ the outcomes of the method of moments differ significantly from those of the Gibbs sampling algorithm and the No-U-Turn sampler (for a significance level of $\alpha = 0.05$). However, all methods seem to work fine if we compare them using scatter plots given in Appendix C.5, Figures C.29 through C.46, and perform better when more data is available.[7] For ρ there seems to be no significant difference when nurses work the same number of shifts, but in the other scenario we see that the method of moments and the Gibbs sampler differ.

From this simulation study, we also discovered that it is possible for ρ to be negative when using the method of moments. It happens when we correct for the number of worked shifts of each nurse. When all nurses work the same number of shifts, this correction does not take place as one can derive from (11.7). We think that it only happens when the data set is small. We also have a lot more outliers for ρ compared to the Gibbs sampler and the No-U-Turn sampler. Finally, the simulation study again confirms there is an upper bound for the number of data points that the Gibbs sampling algorithm can handle.

To give an example, where we see that the method of moments "corrects itself" for a larger data set, we consider a simulated data set of $n = 1000$ nurses sampled from a $NB(\rho_0, (1 + t\mu_0/\rho_0)^{-1})$ distributed with parameters $\rho_0 = 1$ and $\mu_0 = 27/1734$. Now, we assume that for 5% of the nurses works $t = 15$ shifts, 80% works $t = 150$ shifts and the final 15% works $t = 200$ shifts. (It is the same distribution of the shifts that we applied in our simulation study in Appendix B.) We randomly mix this group of nurses and estimate the values of ρ and μ based on the information of a group of n nurses, with n ranging from 1 to 1000. Then the developments of the estimators over the number of data points are shown in Appendix C.4.2, Figure C.16a and C.17 for ρ and μ, respectively. A closeup of the first ten data points for ρ is given in Figure C.16b. In this figure we see that for the first few data points the estimated value of ρ is negative, but then becomes and stays positive valued. If more data is available, both estimations for ρ and μ are very good. Hence, we conclude that with too little data, the moment estimation of ρ is not very reliable but for large data sets it is.

We compare the standard deviations of ρ and μ of the Gibbs sampler and the No-U-Turn sampler in the original examples from Subsection 11.2.1 and 11.2.3 with the asymptotic standard deviations of the method of moments.[8] For the Gibbs sampler, we found that $\rho \approx 0.779793$ with a standard deviation of 0.168899 and $\mu \approx 0.014133$ with a standard deviation of 0.002074. We assumed that $t = 201$ shifts. If we compute the asymptotic standard deviations of the method of moments for ρ and μ using the covariance matrix of (A.7), then we find that the asymptotic standard deviations of ρ and μ are 2.767126 and 0.018068, respectively. Hence the theoretical standard deviations and the ones we obtain using the algorithm differ approximately a factor ten. For the No-U-Turn sampler, we had that $\rho \approx 0.789139$ with a standard deviation of 0.175761 and $\mu \approx 0.014140$ with a standard deviation of 0.002102. We find that the asymptotic standard deviations are 2.799454 and 0.017992 for ρ and μ, respectively. These again differ about a factor ten with the standard deviations we obtained using the No-U-Turn sampler, which was to be expected since the results of the Gibbs sampler and the No-U-Turn sampler are very close to each other. Thus, the standard deviations obtained by the numerical methods are smaller. This seems to be a good thing, since they are more precise than in the asymptotic normal case.

11.3.2 Preference

Based on the above arguments, we advise to apply the No-U-Turn sampler since it is fast, gives good results based on the simulation tests and is quite robust compared to for example the method of moments. All methods seem to work fine to estimate μ: the overall probability of an incident per shift. The main difference is in the estimation of the amount of heterogeneity modelled by ρ.

The Gibbs sampling algorithm can also be used when the data set of nurses is small and does not have too many outliers since it shows more signs of robustness than the method of moments. The estimated value of ρ should not be too close to the chosen upper limit, since it may indicate that the "true" value of ρ is larger than this value. Hence, we advise to make this upper bound as big as possible when using the Gibbs sampler. If it is still too close to the upper bound, we advise to divert to the method of moments to estimate the parameters of the mixed Poisson model. This method also seems to perform well when we have a large data set. The "negative value" problem does not occur for large data sets. We expect the outcome of the method of moments to be the same as the No-U-Turn sampler in this case.

Finally, we like to emphasise that the method that Gill et al. (2018) used to estimate the parameter μ in the mixed Poisson model in their article agrees with the moment estimator that we derived in this thesis and hence will give the same answer. We saw that in most cases the Gibbs sampling algorithm and the No-U-Turn sampler also give approximately the same outcome for μ. However, we think that using either the method of moments (for large data sets), the Gibbs sampling algorithm or the No-U-Turn

[7] We do not expect the reader to thoroughly study these figures. The purpose is to illustrate the general outcomes.
[8] We take the results from Tables 11.1 and 11.3.

sampler to estimate the parameter ρ is a better option than the original idea of just assuming that $\rho = 1$ for convenience of calculations.

Note that the hierarchical models used in the Gibbs sampling algorithm and the No-U-Turn sampler can itself be applied as a model to simulate the number of incidents $N(t)$ that a nurse experience during t shifts. The distribution of $N(t)$ can be sampled from the posterior predictive distribution in PyMC3.

One might have also noticed that we did not round of the estimates for the parameters ρ, μ and the probability $P(N(t) \geq 14)$. For real applications, it needs to be done properly. We advise to take into consideration the standard deviation of each of the estimators. If we have a look at Table 11.1, we would recommend to round the estimate of the parameter ρ to two decimals, and the estimate of the parameter μ and the probability $P(N(t) \geq 14)$ up to three decimals.

11.4 Application: ship damage incidents

Lastly, we want to apply these estimation methods to real instead of simulated data. Unfortunately, as we saw in the Lucia de Berk case, we often do not have access to the distribution of the number of incidents among all nurses and their respectively working hours. In the article of Lucy and Aitken (2004), we came across an incomplete data set of nurses from some intensive care unit in England which included the hours worked per nurse and the number of events on duty. Unfortunately, after an email exchange with one of the authors C. Aitken (personal communication, January 25, 2022) it came to our knowledge that the full data set was not available anymore.

Therefore, we divert to other kinds of data that combines the principle of experiencing incidents and the total time that one is exposed to the risk of being involved in one. The data that we choose here is from a study about wave damage to cargo ships, and taken from (McCullagh and Nelder, 1989, Subsection 6.3.2). It is also given in Appendix E.4, Table E.23. To give some justification why we can apply the mixed Poisson model, it holds that damage incidents do not occur very often which may indicate that it can be modelled as a Poisson process. Furthermore, from the table we can derive that we deal with different types of ships. That is why it is justifiable to model the incident rate as a random process. Also, in the original analysis the authors used a Poisson regression to model the risk of damage caused by the three factors: ship type, year of construction and period of operation. Hence we think that it is valid to use a mixed Poisson model in this case.

We are interested in the columns “aggregate months service” and “number of damage incidents”. The aggregate months service can be interpreted as the total period at risk. Hence, we use one month of service as the time unit. We have a total of $n = 40$ data points. Note that from Table E.23 it follows that for type B, the ships' aggregate months of service is much larger than for the other types. Perhaps it indicates that each row contains the data of more than one ship. Unfortunately, it is not specified in (McCullagh and Nelder, 1989, Subsection 6.3.2).

Before we can estimate the parameters using the method of moments, Gibbs sampler or No-U-Turn sampler, we must establish some factors for the numerical methods. For the Gibbs sampler and No-U-Turn, we have to choose a value α for the first parameter of our beta prior of μ. Based on a look at the data, we choose $\alpha = 0.001$. We keep the same $Exp(1)$ prior as before for the parameter ρ. As the initial values for the parameters, we set $\rho_0 = 1$ and $\mu_0 = 0.001$. Finally, we must set an upper bound for ρ in the Gibbs sampling algorithm. Since we have no clue about the true value of ρ, we need to set it as big as possible. Therefore, we choose to let $R = (R_j)$ in (11.2) range from 0 to 6, with step size 0.005.

We use $T = 10000$ iterations, with a burn in period of $m = 3000$ iterations for the Gibbs sampling algorithm. For the No-U-Turn sampler implemented in PyMC3, we take four chains with each 1000 iterations. The results for each method are presented in Table 11.8. The respective traceplots and posterior density plots for the Gibbs sampler can be found in Appendix C.4.4, Figures C.25 and C.26,

Method	**Estimated value ρ**	**Estimated value μ**	**Standard deviation ρ**	**Standard deviation μ**
Gibbs sampler	2.331532	0.003370	0.824565	0.000543
Method of moments	-2.716572	0.002176	-	-
No-U-Turn sampler	2.300418	0.003375	0.788205	0.000524

Table 11.8: Comparison of estimators for the parameters ρ and μ applied to the ships data set taken from (McCullagh and Nelder, 1989, Subsection 6.3.2).

respectively. The posterior densities and traceplots of the No-U-Turn sampler are presented in Figure C.27, and the forest plots associated with this sampler are illustrated in Figure C.28.

From Table 11.8, we derive that we obtain a negative value because of the time correction for the moment estimator of ρ. Hence it cannot be correct, and it confirms that the method of moments is not suitable for estimating the amount of heterogeneity between ships in this data set. The data set of ships only contains 40 data points; hence it is quite small, and thus may be the cause of the negative estimation. We conclude that the Gibbs sampler and No-U-Turn sampler give almost the same approximation and standard deviation for both the parameters ρ and μ. The method of moments estimates overall probability of an incident per aggregate month service smaller.

From these observations and our insights from before about the performance of all the methods in different situations, we conclude that we can model the number of damage incidents of a ships during a time period of t aggregate months service as a $NB(\rho, (1+t\mu/\rho)^{-1})$ distribution with estimated parameters $\hat{\rho} \approx 2.3$ and $\hat{\mu} \approx 0.0034$.

We conclude that this "real" data set shows similar behaviour to the simulated data sets from before regarding the different estimation methods. Therefore, our advice to apply the No-U-Turn sampler to estimate the parameters ρ and μ in the mixed Poisson model remains the same.

Finally, we need to check whether the mixed Poisson model with our estimated parameters is a good fit to the original data about the damage incidents. Therefore, we will determine the *mean squared error* (MSE). The MSE measures the amount of error in statistical models. When the model perfectly predicts the data, we should have a MSE equal to zero. It is defined as (Frost, n.d.)

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{X}_i)^2, \tag{11.9}$$

where n is the number of data points, X_i the observed values, and $\hat{X}_i$ the predicted values for $i = 1, \ldots, n$. In our case, it holds true that $n = 40$ (see Table E.23). The predicted values are defined as $\hat{X}_i = E(N(t_i)) = t_i\mu$, where t_i equals the aggregate months of service. The observed values X_i are, of course, given by the damage incidents shown in Table E.23. Then it follows that for $\hat{\mu} \approx 0.0034$,

$$\text{MSE} \approx 399.95. \tag{11.10}$$

For sake of comparison, if we plug in $\hat{\mu} \approx 0.0022$ like the method of moments suggested, we find that

$$\text{MSE} \approx 109.26.$$

So according to the MSE, we should prefer $\hat{\mu} \approx 0.0022$ over $\hat{\mu} \approx 0.0034$. In Figure 11.2 the observations of the number of damage incidents are plotted against the aggregate months of service, where we added two lines that model the predicted values for each of the estimated values of μ. Indeed, it shows that for $\hat{\mu} \approx 0.0022$ the observed damage incidents are closer to the predicted values than for $\hat{\mu} \approx 0.0034$. However, the predicted values defined by $\hat{X}_i = t_i\mu$ do not depend on the parameter ρ that models the heterogeneity.

Therefore, we will have a look at another criterion to compare the models. We determine the *Bayesian information criterion* (BIC). The BIC criterion is often used to compare two or more models and we generally prefer models with a lower BIC criterion. The BIC criterion is defined as (Szabó and van der Vaart, 2021, Section 2.6)

$$\text{BIC} = -2\left[\sup_{\theta\in\Theta}\log\prod_{i=1}^{n}p_\theta(X_i) - \frac{d}{2}\log n\right]. \tag{11.11}$$

Here, $p_\theta(\cdot)$ is the likelihood function, d the number estimated parameters, n the number of data points and X_i the observed values for $i = 1, \ldots, n$.

In our case the likelihood function is given by the probability mass function of the $NB(\rho, (1+t\mu/\rho)^{-1})$ distribution. Furthermore, it holds true that $d = 2$ and $n = 40$. The data points are given by the pairs of damage incidents and aggregate months of service given in Table E.23. If we plug the estimated parameters $\hat{\rho} \approx 2.3$ and $\hat{\mu} \approx 0.0034$ into the likelihood function, we find that the BIC criterion for this model is given by

$$\text{BIC} \approx 177.12.$$

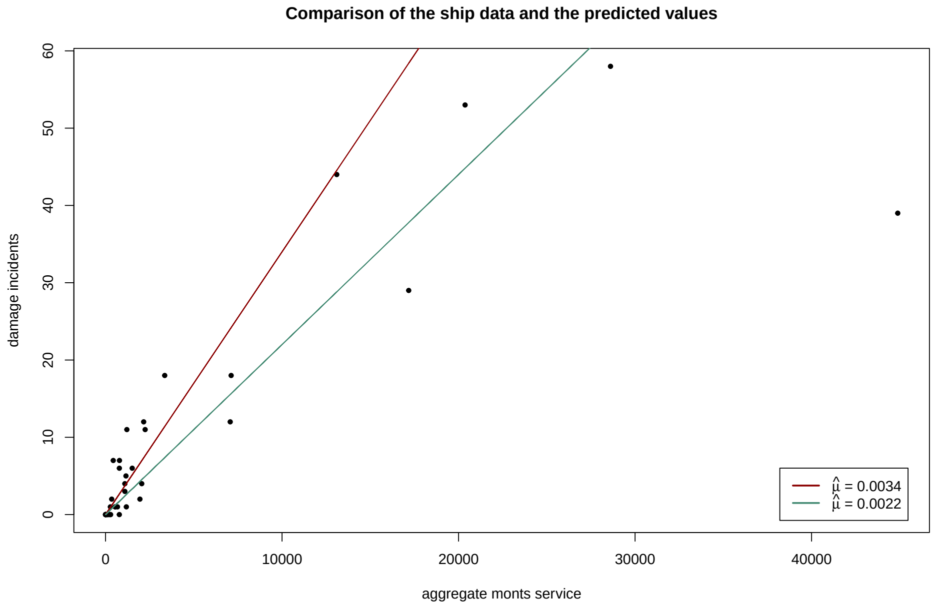


Figure 11.2: Comparison of observed values and predicted values of the mixed Poisson model for two values of μ for the ships data set taken from (McCullagh and Nelder, 1989, Subsection 6.3.2).

Now, if we plug in $\hat{\mu} \approx 0.0022$ and still use that $\hat{\rho} \approx 2.3$, we find that the BIC criterion is

$$\text{BIC} \approx 184.73.$$

Hence according to the BIC criterion, we should prefer $\hat{\mu} \approx 0.0034$ over $\hat{\mu} \approx 0.0022$ given that $\hat{\rho} \approx 2.3$ like we already suggested.

12
Linear Bayes (revised)

In Chapter 9 we already discussed how we disagree with the method of Alkemade (2015) to derive the likelihood ratio of experiencing fifteen arson incidents in some town given the hypotheses that there is no or (at least) one serial arsonist operating in that town. We present our analysis on the problem below.

12.1 Motivation

Since Alkemade (2015) made some interpretation errors in his original approach, we think it is important to give another perspective on the problem that he described in the report. In addition, we think the implicit assumption of him that residents of the Netherlands are indistinguishable is debatable and want to take this into account while determining the probability of having fifteen fires given there was no serial arsonist operating in town X.

We disapproved of the step that Alkemade (2015) took from the probability of having fifteen fires under the hypothesis of chance to the likelihood under the eight different hypotheses as well. Therefore, we want to raise awareness to the problem of obtaining the probability of a series of fires given the hypothesis of there being a serial arsonist. It is similar to the problem of a serial killer nurse. There is much attention in the literature about modelling a series of events given the scenario that all events happened coincidentally. Unfortunately, information about how to specify the probability that a series emerged while there was intent is very difficult to find.

In Chapter 10 we modelled the accidental fires in the arson case as a homogeneous Poisson process with a household-dependent incident rate. We want to use a similar method (to the one of Gill et al. (2018)) to determine the probability that in six months fifteen fires happened in town X given the hypothesis that there was no serial arsonist operating. Furthermore, we want to offer our ideas how one can obtain an order of magnitude for the same probability under the hypothesis that there is a serial arsonist present.

Note that the situation at hand differs a bit from the arson case discussed in Chapters 7 and 10. In that example, we looked at fires happening in one household. Here we have fires happening at different households in the same town. Therefore, one must account for the size of the town. We will do this using the number of households of the town, instead of, e.g., the actual size of the town. We think that this aspect is important, since the fires occurred nearby the homes of the residents of town X. Therefore, it is more appropriate to look at fires happening at random households instead of random locations. We are not interested in fires that took place randomly in town X, for example in some park or in a meadow. If one wants to use the information about the size of town, one can also look at spatial (Poisson) point processes. See for example Baddeley (2007).

12.1.1 Assumptions

Assume that no serial arsonist is operating in town X. Then instead of assuming that each resident of the Netherlands has the same probability of experiencing an arson incident, suppose that

- the arson incidents in town X can be modelled as a homogeneous Poisson process on the positive half line, where the positive half line coincides with time; and
- the intensity of this Poisson process, that is the intensity of an arson incident occurring in town X, is dependent of the number of households of town X and furthermore household-dependent. It is assumed that the household-dependent intensity is $Gam(\rho, \rho/\mu)$ distributed.

From the above, it follows that the households of town X each have some random intensity to experience an arson incident. The total intensity of the Poisson process then equals this household-dependent intensity times the number of households of town X. Hence, in some way this intensity is site-specific since it depends on the number of households.

We can also interpret the defined model as follows. The incidents of each household of town X are modelled as a homogeneous Poisson process with an household-dependent intensity Λ, that is $Gam(\rho, \rho/\mu)$ distributed. Then the arson incidents in town X are modelled as the sum of those n_h independent Poisson processes. It may seem controversial to assume that these Poisson processes are indeed independent since fires are often pooled in neighbourhoods. However, assuming independence between inhabitants of a town is even more debatable since fires will be clustered in households. The sum of two independent Poisson processes, with intensities λ_1 and λ_2, is again a Poisson process with intensity $\lambda_1 + \lambda_2$ (Adan and Resing, 2015, Section 2.6). Therefore, it holds true that the arson incidents in town X can be modelled as a homogeneous Poisson process with intensity $n_h\Lambda$.

We apply the same notation as in Chapters 8 and 9. To sum up:

- $N(t)$: the number of arson incidents that occurred in town X in time period t (Poisson process)
- t: time period of interest
- Λ: household-dependent intensity of the Poisson process $\{N(t) : t \geq 0\}$
- N_h: the number of households of the Netherlands
- n_h: the number of households of town X
- K: the number of fires in the Netherlands caused by arson

Similar to Chapter 9, we adapt a time unit of one year.

Remark. We discussed in Subsection 9.3.1 that we prefer to look at a level of households instead of individuals when we talk about, e.g., the number of arson incidents per year. From previous chapters, we know the number of households in the Netherlands, but we have no information about how many households lived in town X in 2013. Though the ratio of number of inhabitants of town X to number of residents in the Netherlands will likely not differ much from the ratio of number of households in town X to number of households in the Netherlands. Therefore, we will still use the information about the residents of town X and the Netherlands. One will see that only the ratio is important and therefore we assume that both ratios are equal.

Hence it holds true that

$$\frac{n_h}{N_h} = \frac{2400}{17000000} \approx 0.000141.$$

In Subsection 9.3.1 we explained that Alkemade (2015) made some mistakes during his data extractions from CBS (2014). He claimed that based on these fire statistics, it followed that $K \approx 3000$. However, using the same assumptions as Alkemade (2015) about the "Town X fires" and the same reference, we found that $K \approx 700$ instead.[1] Since we want to give a comparable probability to the one obtained by him, we will use the same data. Therefore, we assume that $K \approx 3000$.

Now, we have to determine the parameters of the mixed Poisson model. We assumed that Λ is $Gam(\rho, \rho/\mu)$ distributed. For the parameters of Λ it holds true that μ can be interpreted as the overall probability of experiencing a fire per year and ρ models the amount of heterogeneity between individuals. Similarly as in Chapter 10, we can estimate μ as the ratio of total number of fires caused by arson to total number of residents of the Netherlands. Hence $\mu = K/N_h$.

Now, the total intensity of the Poisson process is $n_h\Lambda$. If $X \sim Gam(\alpha, \beta)$ then it holds true for any constant $c > 0$ that $cX \sim Gam(\alpha, \beta/c)$ (Thompson, n.d., Section 10.4(ii)). So it follows that $n_h\Lambda$ is

[1] Alkemade (2015) argued that the arson incidents that took place in town X needed to be placed into the category of indoor fires.

$Gam(\rho, \rho/n_h\mu)$ distribution. Thus, we can also interpret the parameter $n_h\mu$ as the overall probability of having a fire in town X per year.

To determine ρ, we would need information about the distribution of fires caused by arson over the households of the Netherlands which is also explained in Section 10.3 and Chapter 11. However, since this information is not available to us, we set $\rho = 1$ for now.[2] Lastly, the time period of interest equals $t = 0.5$ years.

12.2 Model adaption

In the upcoming two subsections, we will first discuss how to obtain the probability of experiencing fifteen fires in approximately six months in town X given there is no serial arsonist operating in town X. Then we will look at the same probability under the remaining hypotheses and give our vision how to find these as well.

12.2.1 Probability of experiencing 15 fires

Again, assume that no serial arsonist is operating in town X. Hence H_1 is true. Then using the above formulated model, we have that

$$P(N(t) = j|n_h\Lambda = n_h\lambda) = \frac{(tn_h\lambda)^j}{j!}e^{-tn_h\lambda}, \qquad j = 0, 1, 2, \ldots,$$

and the probability density function of Λ equals

$$\pi(\lambda) = \frac{(\rho/\mu)^\rho}{\Gamma(\rho)}\lambda^{\rho-1}e^{-(\rho/\mu)\lambda}, \qquad \lambda \geq 0.$$

Using similar reasoning as in the proof of Theorem 8.1, it follows that

$$P(N(t) = j) = \binom{j+\rho-1}{j}\left(\frac{1}{1+tn_h\mu/\rho}\right)^\rho\left(\frac{tn_h\mu/\rho}{1+tn_h\mu/\rho}\right)^j. \tag{12.1}$$

Plugging in $\rho = 1$ and $\mu = K/N_h$, we find

$$P(N(t) = j) = \frac{1}{1+tn_hK/N_h}\left(\frac{tn_hK/N_h}{1+tn_hK/N_h}\right)^j. \tag{12.2}$$

We assumed that $n_h/N_h \approx 0.000141, K \approx 3000$ and $t = 0.5$. Hence, we obtain the following probability of experiencing fifteen fires in approximately six months in town X under H_1:

$$P(N(t) = 15) \approx 3.574 \times 10^{-12}$$

We want to stress that if we use $K \approx 700$ instead of $K \approx 3000$, we obtain an even smaller probability of about 10^{-20}.

Since it is debatable whether $\rho = 1$ is a good choice to model the amount of heterogeneity between households of experiencing an arson incidents, we added Figure 12.1 to show the influence of ρ on the desired probability. We can conclude that if we have more heterogeneity, hence households vary more in their risk of experiencing arson, then the probability of observing fifteen fires increases. Hence it is less exceptional to observe a series of fifteen arson incidents since there are more outliers in the population.

12.2.2 Likelihood ratio of experiencing 15 fires

To remind us, Alkemade (2015) formulated the following eight scenarios:

- H_1: There was no serial arsonist operating in town X.
- H_2: At least one unknown person was operating as a serial arsonist.
- H_3: Only suspect A was operating as a serial arsonist.

[2] We will later check how the value of ρ affects the probability of interest.

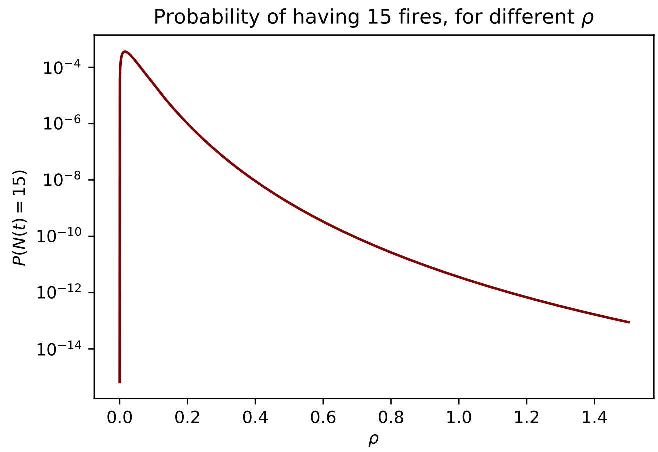


Figure 12.1: The probability of having fifteen arson incidents in approximately six months in town X, for different values of ρ. Observe that for the y-axis we used a logarithmic scale.

- H_4: Suspect A and at least one unknown person were operating as serial arsonists.
- H_5: Only suspect B was operating as a serial arsonist.
- H_6: Suspect B and at least one unknown person were operating as serial arsonists.
- H_7: Only suspect A and suspect B were operating as serial arsonists.
- H_8: Suspect A, suspect B and at least one unknown person were operating as serial arsonists.

To obtain the evidential value of having exactly fifteen fires in six months in town X given each of these hypotheses, we need to compute the remaining seven probabilities $P(E|H_i)$ for $i = 2, \ldots, 8$ besides the first one that we already calculated in the previous subsection. However, to model the occurrence of incidents while a serial arsonist is operating in town X is quite difficult. Therefore, we need to come up with some other method to get an idea how big or small these probabilities are.

To give some indication about the order of magnitude of these probabilities we need information about arsonists in general. First, we need to know how many arsonists are operating in the Netherlands every year. Then we can ask the question how many commit arson only once and what number is indeed a serial arsonist. Note that these numbers influence the prior probabilities of each of the hypotheses H_1 through H_8. Alkemade (2015) indeed uses this information to determine the prior probabilities for each of the scenarios, which can be read in his report. The definition of a serial arsonist is open to dispute, and often in literature one just talks about arsonists in general (Schoenmakers and van Ham, 2013, Section 4.2). Hence it is critical to be aware which definition is used while looking at data.

It holds true that based on media reports, in 2007 there were at least 40 individual cases of serial arsonism in the Netherlands (Schoenmakers and van Ham, 2013, Section 4.3). A Japanese study showed that from the total number of arsonists during the period of 1982 to 2005 (11652), there was 6% (708) convicted to serial arsonism. And in another study, it appeared that 18% (32) of the researched arson incidents (175) in the United Kingdom can be attributed to serial arsonists. Based on this information and (CBS et al., 2014, Table 4.7, pp.359-360), Alkemade (2015) concluded there were about 30 to 40 serial arsonists operating in 2013. He rounded it up to 50. Alkemade (2015) used this number to compute the probability that a serial arsonist lived in town X and committed arson there.

If we know how many serial arsonists operate in the Netherlands, we want to know whether we can put them into different categories. In our situation we are focused on the kind of arson that they commit, e.g., do they commit arson to properties or street objects. Hence it may be possible that the target influences how many fires the arsonist starts. In the reasoning of Alkemade (2015) about the class to which the "Town X fires" belong, it is also shown to be important. From (Schoenmakers and van Ham, 2013, Section 4.1) it follows that, in research, arsonists are also put into categories based on their underlying reason of committing arson. For example, there are individuals with psychological problems or that seek revenge and therefore commit arson. However, this type of categorisation is of no particular interest to us.

When we zoomed in on a specific type of serial arsonist, one might wonder what the distribution of the number of fires is over the serial arsonists. E.g., what is the mean number of objects or properties that an arsonist sets on fire? Or what is the maximum number that one has set on fire? From Dutch case studies, it is known by the police that serial arsonists caused dozens to in a few cases hundreds of fires (Schoenmakers and van Ham, 2013, Section 4.3). However, it is unknown how many of the total number of arson incidents in the Netherlands were caused by serials. In one study it was concluded that a serial arsonist caused an average of five arson incidents. In another study it was about 31 on average. This difference may be ascribed to various definitions and research populations. Alkemade (2015) said in his report that assuming about fifteen fires per serial arsonist coincides with the before mentioned data about the number of serial arsonists.

Using all of this information, we hope that an expert can give us an indication what the probability is that we observe a certain number of arson incidents in a predetermined time period, given that a serial arsonist or multiple serial arsonists are operating is some town. It is most important to have some idea of the order of magnitude of the probability since the probability under each scenario will be small. We want to specifically know the ratio between them and therefore we think that the order of magnitude is enough.

Combining these probabilities with the probability of the evidence given the hypothesis that no serial arsonist is operating in town X, we can find the desired likelihood ratio of seeing fifteen fires in town X in approximately six months.

12.3 Comparison

As we calculated in Subsection 12.2.1, the order of magnitude of the probability given that no serial arsonist is operating in town X is about 10^{-12} assuming a site-specific intensity rate of incidents. This probability is smaller than Alkemade's uncorrected probability of 10^{-10}. If we also have some idea for the probabilities given the other seven scenarios using an expert's opinion, we think that we have a clearer solution to the problem than Alkemade (2015) proposed in his report.

The actual probabilities are difficult to compare, since the one of Alkemade (2015) was formulated differently and both probabilities are based on unreliable data. We showed that one can obtain a quite different answer when different assumptions are made.

Part V

Recommendation

13
Overview

As seen in the previous chapters, throughout the literature different formats for presenting statistical evidence of incident series were used. The following four passed by in this research thesis: likelihood ratios for a case of credit card fraud; posterior probabilities for arson in one household; a p-value under the hypothesis of chance for a roster case in a hospital; and an exact probability for arson at several inhabitants of a town. Formally, following the Bayesian approach, an expert only reports a likelihood ratio (Sjerps, 2004). The prior odds must be determined by the judge, since the expert only has expert knowledge about the forensic evidence. Consequently, the expert cannot comment on the posterior odds.

In this chapter we present an overview about the different ways one can interpret statistical evidence for different situations involving an incident series following the literature. We summarise the results from Chapters 6 through 12, in which we considered and revised models that are used to provide statistical evidence for the above mentioned situations. In addition, we will shortly discuss our ideas for other circumstances that include an incident series and end this chapter with a conclusion. In the next chapter we will write our recommendation to the Netherlands Forensic Institute and therefore discuss what in our opinion the best approaches are to present the statistical evidence of an incident series.

13.1 Summary

The summary given below will answer our research question about the ways statistical evidence can be interpreted for different situations according to the literature and our own ideas.

13.1.1 Credit card fraud

The credit card fraud case that happened in England in the early part of 2000, is an example of a situation where we have a closed population of suspects, and we are almost sure that a crime took place. Therefore, it may seem like a simple situation to analyse but one must take into account some subtleties.

We looked at the approach of Lucy and Aitken (2004) how they interpreted the statistical evidence in their unpublished draft. The question of interest was: What is the value of evidence of roster data? It was stated in the article that the data consisted of information about the customers, the dates they visited the retail outlet and the assistants that worked at the retail outlet. The aim was to compute likelihood ratios for all possible combinations of assistants given the hypotheses of a specific group of assistants being responsible for the fraud versus there being no causal dependence between the presence of the assistants and the unauthorised transactions.

However, it turned out that the vaguely formulated evidence needed to be interpreted in a specific way. The evidence should have been defined as the event that one, or more than one, of the assistants under suspicion served the customers when they visited the retail outlet, and as a result of these visits the unauthorised transactions took place. However, the last part of this phrase is exactly what one wants to find out and therefore cannot be included in the evidence. If we assume that the evidence equals the event that one, or more than one, of the assistants under suspicion served the customers when they visited the retail outlet, then one cannot derive the likelihood ratios in the way that was done in the article. One needs attendance data of all the assistants that worked at the retail outlet.

The article shows a promising way of solving the problem. However, some of the assumptions are unreasonable and there are a few misconceptions. We advise others to have a look at this situation of credit card fraud again. If information about the attendance is available, then one could apply our suggested improvement for the model to hopefully obtain the value of evidence of the roster data. For complex situations we suggest implementing the method in a programming language of choice instead of calculating the likelihood ratios for different subsets of suspects by hand. We did not dive further into it but encourage the reader to give it a shot.

We conclude that for situations where there is a closed population of suspects and one is almost sure that a crime took place, the method described in Lucy and Aitken (2004) can give a first insight in how to determine in the evidential value of roster data. However, we have to be very precise on how we formulate the evidence. Therefore, at the moment we recommend investigating the proposed model further before we advise how to adapt them in real cases.

13.1.2 Arson

In Chapter 7 we gave a situation description about a series of arson incidents that took place in one household. The Netherlands Forensic Institute (NFI, 2019) came up with a model to determine a lower bound for the probability that some of these fires had a systematic cause, given that the suspect was involved in at least four fires in two years. It is a posterior probability, since the evidence was defined as the number of fires that the suspect witnessed. The role of an expert, following the Bayesian approach, is that they only provide for the likelihood ratio of the forensic evidence given the desired hypotheses, and leave the decision for the prior odds to the judge. Hence, they have no say in the posterior odds.

However, the Netherlands Forensic Institute remarked in their report that if we assume that the household of the suspect is a random household in the Netherlands, then the probabilities are applicable to the situation of the suspect. This assumption, that the suspect is indeed a random household in the Netherlands, influences the prior probabilities. Hence if we have more information about the situation of the suspect, the prior probabilities can be adapted and the probabilities for the suspect change. Hence by this remark, the Netherlands Forensic Institute tried to leave the ultimate decision to the court. Furthermore, as we already highlighted in Subsection 7.3.3, establishing the likelihood ratio instead of the posterior odds is practically impossible in this situation.

It was assumed that without any further information, households are indistinguishable. We performed a sensitivity analysis on the parameter N, that models the number of households in the Netherlands and assumed that only half of the households in the Netherlands is at risk. It resulted in a decrease of the posterior probability that i of the fires that the suspect experienced had a systematic cause, for $i = 1, 2, 3$. Therefore, we looked in Chapter 10 to the mixed Poisson model proposed by Gill et al. (2018) to model the accidental fires. We concluded that adding a modest amount of heterogeneity between households, by setting $\rho = 1$, had minor effect on the posterior probabilities. Information about the distribution of the number of fires in the Netherlands can help to set the value of ρ more accurate.

We left the situation of the arson incidents in one household with an unsolved problem. We could not determine the probability that at least four of the fires that the suspect experienced had a systematic cause, given there were at least four fires in two years. It has probably something to do with the upper and lower bounds that were used throughout the calculations, which need to be made more precise to obtain this specific probability.

The method proposed by the Netherlands Forensic Institute seems to be most suitable for situations where one wants to find out whether it is likely that an underlying systematic cause resulted in the events or that they happened coincidentally. This model helps to find an answer to the question how many of the incidents potentially had a systematic cause. However, it does not point out which of the incidents were induced by this systematic cause. This is sometimes even more important to know. Therefore, we will further discuss this issue in Section 15.1.

13.1.3 Roster cases

Our results regarding roster cases that were considered in Chapters 8 and 11 revolve around the probability that a nurse coincidentally experienced at least as many incidents as the Dutch nurse Lucia de Berk, which is a p-value. In Subsection 8.3.3 we already discussed that a p-value only gives us information that something extraordinary had happened and does not tell us what the cause was of the event. We believe that p-values can give us a first indication whether we should further investigate a case but should not be used as evidence in court. Most of the time in roster cases and incident series in general, a person is suspected of a crime because he or she was present at a lot of incidents. Then we want to first ask the question how rare this event is, before we look into how to prove that the suspect indeed did something wrong. We see p-values fit for this purpose in combination with the size of the population of interest.

We will elaborate by means of the example case of Lucia de Berk. In (8.3), we found out that using the model of Gill et al. (2018) the probability of a nurse coincidentally experiencing at least fourteen incidents during the same number of shifts as De Berk was approximately 1 in 49. Now, it is important to put this number into context. Lucia de Berk was present at this number of incidents during 201 shifts in three hospital wards. In total there happened 27 incidents during 1734 shifts, which was the period that De Berk worked at the Juliana Children's Hospital and Red Cross Hospital. If we know how many nurses worked at the three hospital wards, we can conclude whether it is realistic that 1 in 49 nurses experienced (at least) fourteen incidents. If the total number of nurses is much smaller than 49, this occurrence is suspicious, and we should investigate the case further. On the other hand, if this number is much larger, the event becomes less impressive and we can potentially decide to not dive deeper into it.

Now, one can also interpret the above phenomenon as looking at a random variable that describes the number of individuals in a population that experienced at least n incidents during time period t. It is distributed as a binomial random variable with size parameter N and "success" probability $P(N(t) \geq n)$, where the time period of interest t and the population size N are established in advance. The expected number of people in the population that experienced at least n incidents, which equals N times $P(N(t) \geq n)$, gives an indication of the rarity of the seen occurrence.

In the legal case of Lucia de Berk, law professor Henk Elffers was asked whether it may have been a coincidence that De Berk was present at so many resuscitations during her shifts (Elffers, 2002). He gave his answer in the form of a p-value equal to 1 in 342 million, which was not strange regarding the question that was asked. The problem was that the court interpreted it wrongly (prosecutor's fallacy) and at some point used it as indirect evidence against Lucia de Berk (Meester et al., 2006). However, Elffers' model was criticised on its assumptions that did not represent reality very much. We think that the assumptions that Gill et al. (2018) proposed reflect the situation better and therefore give the preferable answer to the original question. Perhaps it would not have led to the bad outcome for Lucia de Berk. Therefore, we conclude that realistic assumptions must be made to be able to use p-values as a tool for deciding whether to further investigate the series of incidents.

One needs to be very careful that p-values are not used as evidence against the suspect. We want to emphasise that they can be applied giving a first indication if something peculiar happened. Therefore, we need information about the population size of the group to which the suspect belongs.

In Chapter 11 we presented three methods how to estimate the two parameters ρ and μ used in the mixed Poisson model of Gill et al. (2018). We already concluded that, for small sample sizes, numerical methods such as the Gibbs sampling algorithm or the No-U-Turn sampler give good results using simulated data sets compared to the method of moments. We advise to use the No-U-Turn sampler in most situations, since it is fast, robust and seems to handle more complex situations (like small sample sizes and small values of ρ) well. The method of moments is suited for large data sets, which may be more common in arson cases instead of roster cases. Using one of these methods will give from our point of view an improvement to the mixed Poisson model, in comparison with the original choice of Gill et al. (2018) to set $\rho = 1$.

13.1.4 Linear Bayes

In Chapter 9 we focused on arson incidents happening at different locations in a town. We looked at a report by dr. Frans J.M. Alkemade and discussed his method to determine the evidential value of a series of fifteen fires happening in six months in town X. He distinguished between eight scenarios of there being none, one or multiple serial arsonists operating in town X. Before the value of evidence could be determined, Alkemade derived a rough estimate of the upper limit for the probability of having fifteen fires given there was no serial arsonist. This upper limit was sufficient for the linear Bayesian analysis of Alkemade. But as we explained in Subsection 9.3.1 there are more precise ways to calculate this probability. Furthermore, his data extraction was not accurate and the data itself was unreliable. In a sensitively analysis we discovered that his choices had quite some impact on the probability that Alkemade derived. Also, Alkemade corrected his probability for selection effects which in our opinion seems unnecessary.

Alkemade claimed there was no reason to discriminate between the scenarios of there being a serial arsonist, but this statement was not sufficiently supported. It followed from his argumentation that $P(E|H_1) = 10^{-6}$ and $P(E|H_i) = 1$ for $i = 2, \ldots, 8$ which is misleading, since the probability of having exactly fifteen fires is small given each of the hypotheses. Not just under H_1. Thus, what we miss in his line of reasoning, is how to obtain the probability of having a series of arson incidents given there is (at least) one serial arsonist active in town X. One must balance the probability under the hypothesis of chance to another probability given there is a systematic cause. Otherwise, the evidence is confusing for the judge.

Chapter 12 presents our analysis on the problem. We used a version of the method of Gill et al. (2018) to find the probability $P(E|H_1)$, where we assumed that the arson incidents in town X can be modelled as a homogeneous Poisson process with a random intensity. The intensity depended on the number of households in town X and the risk of each of the households that lived in town X to experience arson. This method led to a probability of 10^{-12}. We questioned whether we could compare this probability to Alkemade's uncorrected probability of 10^{-10}, since both of them relied on untrustworthy data. We concluded that different assumptions lead to different answers.

We specified the information that is needed for an expert to find the probability of having fifteen fires given that at least one serial arsonist is operating in town X. In summary, we think that details about how many serial arsonists there are for the prior and data on the distribution of fires over serial arsonists can already give an idea about the order of magnitude of this probability. Hence it will give a better answer to the problem, than Alkemade's original solution.

13.1.5 Other situations

Note that there are, of course, many more situations where we have a series of possible criminal events like we saw in Chapter 4. From our analysis, we can conclude that the mixed Poisson model proposed by Gill et al. (2018) for the roster case of Lucia de Berk is applicable in many situations involving a series of rare events. For example, traffic accidents or other cases of insurance fraud like the one we already brought up in Example 4.6 about the electric bike. It is primarily convenient for modelling incidents under the assumption of coincidence, and the model is also often used in case of overdispersion.

Below we will discuss a situation that we found very interesting, but unfortunately could not analyse in detail in this research thesis. However, we want to point out a few remarks on the method used to interpret the statistical evidence.

Neighbourhood

Example 4.5 sketched a situation that was discussed in an article by Bolviken and Egeland (1995) about a fireman being seen in the neighbourhood of 24 out of 37 cases of forest fire prior to their onset. In the article, the probability of the fireman being seen near the scene of the fire is calculated assuming that the presence of the defendant at the various scenes of fire is purely coincidental. It is presented as a p-value, since according to Bolviken and Egeland (1995) it *'(...) is the most usual way of assessing the outcome of empirical tests of hypotheses'*. The main three modelling issues are deciding the number of stops made by the defendant during the critical time period, formalising the meaning of the defendant "being in proximity of a fire", and putting a value on the probability that the defendant is in the vicinity of the fire each time he stops.

Our primary question is whether the model is not too simplified, because of the assumptions that are made throughout the calculations. For example, the upper limit of the p-value solely depends on the distribution of the number of stops for a certain fire through its mean. The model is even more simplified by using the mean of the probabilities of proximity per fire and hence the mean road length adjacent to a fire instead of the individual defined probabilities. Bolviken and Egeland (1995) explain that it may be worthwhile to sacrifice some accuracy. According to them, it is equally important that a basis is created which allows the court to pass judgement on the assumptions underlying the analysis and on the values to be assigned to input parameters. A bridge between mathematics and law is required, which suggests that the assumptions and input should be kept simple.

Furthermore, Bolviken and Egeland (1995) choose to phrase their conclusions in terms of probabilities, instead of following the approach of likelihood ratios. In the discussion, they come back to this choice and show that the result from their analysis in their opinion can be applied using the Bayesian perspective. They state that the probability of the defendant being seen prior to at least c fires given that he is guilty, is equal to one since the defendant appears at every fire scene if he is the arsonist. However, this reasoning is not true. There might be some combination between coincidence and intent in this case, since some of these c fires might have happened without the interference of the fireman. It is the same problem that Gill et al. (2018) described in their article about the legal case of Lucia de Berk. They reasoned that it is a difficult task determining a reasonable model for the number of incidents given that De Berk is a murderer, since one must take into account that some of the incidents are not murders at all.

13.2 Conclusion

The main question is: What do we have to report when an expert is asked to analyse a series of similar, possibly criminal, events? As we saw in the above in most of the cases, it is possible to derive the probability of having an incident series given the hypothesis of coincidence. However, the problem is finding out the probability given the scenario of there being an underlying systematic cause. In the literature, there is hardly any attention to this scenario. Hence, we hope that this thesis comes to the attention of fellow researchers and in the future there will be written more about this problem. Furthermore, we would like to note that most literature that we found is focused on roster cases like the one of Lucia de Berk and less on other situations like fire incidents and traffic accidents. Therefore, we like to encourage others to do further research on these other phenomena as well.

We conclude that p-values, e.g., the ones reported by Elffers (2002) and Gill et al. (2018) in the Lucia de Berk case, and the one reported by Bolviken and Egeland (1995) for the fireman situation discussed above, can only be used as an indication whether something peculiar is going on. They can serve as a sign whether to start an investigation but may not be used as evidence! Using p-values as statistical evidence is dangerous since they are often interpreted wrongly. They only point out the situation given the coincidence scenario and do not balance it with the probability of seeing the event given there being a systematic cause. It is also true for the probability of having an exact number of incidents. The probability of having a cluster of events does not show that the suspect has done something wrong, e.g., he or she might have a higher risk of experiencing them. Therefore, it is important to also look for other forensic evidence. Though, in many of these cases there is almost no hard evidence against the accused (Fenton et al., 2022). Thus, one-sided probabilities may not be reported without context.

For this balancing probability we need experts in the field of specific types of incident series, for example car accidents, serial killer nurses or fire accidents, that can give us some idea of the order of magnitude of the opposite chance. This way a likelihood ratio can be reported to the court, which is desired according to the Bayesian approach.

For using a method with posterior odds instead of a likelihood ratio, like we saw in the arson case in Chapter 7, it is important to specify on what scale the problem is analysed. Hence the judge needs to know how the reported posterior odds can be changed when more information about the suspect is available. Only then, we think it is allowed to divert from reporting likelihood ratios to posterior odds.

Given the method of Gill et al. (2018) to model incidents, it is also possible to present the incident rate of the suspect to the judge and let them decide whether the series of events that the suspect experienced is suspicious based on this statistical evidence. This rate can be interpreted as the suspect's risk to witness such incidents. Using the gamma distribution that can be derived through the mixed Poisson model, we have an overview of the risks of the total population. Hence, we can derive how high the risk of the suspect is compared to others, on which the judge may base his verdict. If asked by the judge, a probability of experiencing the incident series, given the incident rate of the suspect, can also be included. Though, we need to make clear to the judge that this probability is misleading since it is not balanced with the probability given another cause than coincidence.

In most situations, the method to interpret statistical evidence is very dependable on the available information. It is critical that we are transparent about how data is collected, and incidents are defined, to avoid situations like the one of Lucia de Berk. By giving an overview of the various situations involving a series of incidents, methods how one can interpret the statistical evidence and stating what details are needed to evaluate it, we hope to contribute to research on how to deal with such situations.

14
Recommendation

Our final research question, which we formulated at the beginning of this thesis, was: What is the best statistical approach to provide for statistical evidence in each situation, and why? In this chapter we will present an answer to this question which forms our recommendation to the Netherlands Forensic Institute.

We think that the method proposed by Gill et al. (2018) is best suited for modelling accidental events. They propose a mixed Poisson model, which is often used to model rare events and considers heterogeneity in the population. For the discussion of this model, we refer to Chapter 8 for the original ideas and to Chapter 11 for the improvements regarding estimation of the parameters. Using this model a p-value can be derived under the hypothesis of chance, like Gill et al. (2018) did in their article. It may help to find the question whether to start an investigation. However, they may not be used as evidence in court.

Unfortunately, modelling rare events given the scenario of there being a systematic cause is problematic, because of the combination of coincidence and intent, and therefore may need the interference of an expert. It needs further investigation. In Subsection 12.2.2 we already presented our view on how certain information about how many incidents roughly take place given that the suspect had criminal intentions can be helpful to find the contrasting probability in case of serial arsonists. Combining the probabilities under this scenario and the one of coincidence, we should be able to report a likelihood ratio to the court as desired.

When one wants to find out whether it is likely that an underlying systematic cause resulted in the events or that they happened coincidentally, we advise to look at the model proposed by the Netherlands Forensic Institute (NFI, 2019) discussed in Chapter 7. This model helps to find an answer to the question how many of the incidents had a potential systematic cause but does not point out which of them where indeed induced by this systematic cause. In Chapter 10 we presented our improvements using the method proposed by Gill et al. (2018) to model the accidental events. The results are presented as posterior odds, which is different from the Bayesian approach of reporting likelihood ratios. Therefore, it needs to be clearly stated how the reported posterior odds can be changed when more information about the suspect is available.

Finally, for situations where we are sure that criminal activity took place and there is a closed population of suspects, it can be insightful to look at the ideas of Lucy and Aitken (2004) discussed in Chapter 6. Since there are a few misconceptions regarding the model and defined evidence, we recommend to first investigate the proposed model further before we advise how to adapt them in real cases.

15 Future research

As one might have noticed during this research, some questions were left unanswered. Therefore, in this chapter we will give suggestions for future research regarding the issue of incident series.

Before we talk about a few specific topics for future research, we like to discuss the following. As one might have noticed, not much information is publicly available about cases involving incident series. Therefore, for roster cases we simulated our own data in Chapter 11 and Appendix B to apply the discussed methods for estimating parameters of the mixed Poisson model. Furthermore, we used data of the desired format about wage damage to cargo ships to still include a realistic example. For other cases like to arson incidents, we had to assume some simplifications of the model since the distribution of incidents over the population was not accessible. We want to stress that in future cases, it might be possible to retrieve the desired information from parties such as the police or insurance companies. From our research, we now know what exact data is needed, which is helpful for future studies.

15.1 Causal questions

The goal of statistical evidence of the incident series is to give insight on whether something happened given an underlying “systematic” cause. But as soon as the expert presents his statistical analysis to the court, there is always a follow up question. What was the exact cause of the incidents? There are number of possibilities. The suspect might have had criminal intent, but he or she could have a higher risk of experiencing the events as well. The actions of the criminal could also have been just one of the causes leading to the incidents.

This last option becomes clearer when using an example. Consider the situation of Example 4.1 about medical staff injecting patients with insulin to potentially murder them. It is possible that some of the patients were already weak, and the insulin was just the last push they needed to eventually die. There are so called “angels of mercy” that think that a patient is better off if they no longer suffer from their illness and therefore decide to kill them (Wikipedia, 2022). Hence here we may ask whether the patient died of the insulin, or it just sped things up.

Furthermore, sometimes we are more interested in whether a specific incident had a systematic cause because, e.g., someone died during the event. Think about the fire incidents or traffic accidents. Our research does not answer these kinds of causal questions, but these are still important to reflect on. During literature research, it came to our notice that Fienberg and Kaye (1991) wrote an article about the effect of evidence of similar events on the posterior probability of guilt in an idealised situation. This article might be insightful to tackle these kinds of problems.

15.2 Literature

As highlighted in the introduction of this research thesis, it came to our knowledge that the Statistics and the Law Section Committee of the Royal Statistical Society is currently preparing a report about approaches to interpreting clusters of events (C. Aitken, January 25, personal communication, 2022). The authors of this hopefully soon published report are among others Richard Gill and Peter Green. Peter Green (personal communication, February 3, 2022) wrote that it will be a policy paper rather than a research paper. Nevertheless, it might be interesting for others to dive into it once it is released.

Furthermore, as we already explained in Subsection 13.1.5, Bolviken and Egeland (1995) wrote an article about a fireman being seen in the neighbourhood of many cases of forest fires prior to their onset. We encourage the reader to look at their methods in more detail, since it might be applicable to some situations where someone is often seen in the neighbourhood of a crime. Be aware of the remarks about the (over)simplification of the model and their reasoning about likelihood ratios.

15.3 Intent and coincidence

As already noted by Gill et al. (2018) in their article, to apply a Bayesian approach to a roster case like to one of Lucia de Berk, one needs to determine a reasonable model for the number of incidents given that the suspect is a murderer. However, it needs to be considered that some of the incidents are not murders at all. Hence there is a combination of coincidence and intent.

The solution to this problem would mean that besides determining the probability of seeing a series of incidents given the hypothesis of chance, we can also calculate the balancing probability. Thus, we can determine the likelihood ratio and the problem of misleading one sided probabilities is solved. The idea can be applied to fire incidents and traffic incidents as well. One solution that we thought of was interpreting the number of incidents as the sum of the accidental incidents and the ones that were intentional, like the Netherlands Forensic institute (NFI, 2019) did for the arson case, explained in Chapter 7. We encourage others to look into this idea.

15.4 Distribution of intensity rate

In Chapter 8, we saw that Gill et al. (2010) choose to model the intensity of a nurse seeing or reporting an incident with a gamma distribution. In Subsections 8.3.1 and 8.3.3, we said that it might be possible to replace the gamma distribution with another distribution. All that is needed to include heterogeneity between nurses is an extra parameter in the Poisson process. An obvious choice for a replacement is the Weibull distribution or the exponentiated exponential distribution, since both their density functions are quite similar to the one of the gamma distribution and simplify to the exponential distribution when one of the parameters is set equal to one (Kundu, 2004). One can look how it affects the results and whether one of the distributions is a better choice to model the nurse-dependent intensities. Maybe using non-parametric methods like kernel density estimation will also be an option.

Using other distributions rather than the gamma distribution, will not lead to a nice closed from expression like we had before. Therefore, it will be needed to divert to numerical methods to determine the posterior and unconditional distributions of the mixed Poisson model at hand. Presumably this implementation can be done using the PyMC3 package, see Salvatier et al. (2016), in Python that we also used in Chapter 11.

15.5 Spatial Poisson point process

Finally, in Chapter 12 we shortly addressed the existence of spatial Poisson point processes in comparison with "normal" Poisson processes. We argued that in the arson incidents situation, the aspect of fires happening at random households instead random locations was more important to us, and therefore we chose to apply a version of the mixed Poisson model to model the incidents happening in town X.

In our view, spatial (Poisson) point processes can be helpful in forensic settings. Using information from maps, one can predict how many incidents might happen in a region and maybe where they most likely will occur. Therefore, it is worth looking into when incidents tend to take place randomly in a certain area. A good starting point to learn more about spatial point processes is the lecture notes by Baddeley (2007), which are online available.

Part VI

Appendix

A
Proofs

A.1 Confidence intervals

The proof of Theorem A.1 is based on (Serra, 2020, Subsection 3.1.2), and the proof of Theorem A.2 is based on Windley (n.d.).

Theorem A.1 (Wald)**.** *Wald's (asymptotic)* $100(1-\alpha)\%$ *confidence interval for* p *is given by*

$$\left[\hat{p}_n - z_\alpha\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}}, \quad \hat{p}_n + z_\alpha\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}}\right]. \tag{A.1}$$

Proof. As we said in Subsection 3.2.1, it holds that

$$P\left(-z_{\alpha/2} \le \sqrt{n}\frac{\hat{p}_n - p}{\sqrt{\hat{p}_n(1-\hat{p}_n)}} \le z_{\alpha/2}\right) \approx 1-\alpha.$$

If we rewrite the above inequalities, we obtain the following:

$$\begin{array}{rcl} -z_{\alpha/2} \le & \sqrt{n}\dfrac{\hat{p}_n - p}{\sqrt{\hat{p}_n(1-\hat{p}_n)}} & \le z_{\alpha/2} \\ -\dfrac{z_{\alpha/2}}{\sqrt{n}} \le & \dfrac{\hat{p}_n - p}{\sqrt{\hat{p}_n(1-\hat{p}_n)}} & \le \dfrac{z_{\alpha/2}}{\sqrt{n}} \\ -z_{\alpha/2}\sqrt{\dfrac{\hat{p}_n(1-\hat{p}_n)}{n}} \le & \hat{p}_n - p & \le z_{\alpha/2}\sqrt{\dfrac{\hat{p}_n(1-\hat{p}_n)}{n}} \\ \hat{p}_n - z_{\alpha/2}\sqrt{\dfrac{\hat{p}_n(1-\hat{p}_n)}{n}} \le & p & \le \hat{p}_n + z_{\alpha/2}\sqrt{\dfrac{\hat{p}_n(1-\hat{p}_n)}{n}} \end{array}$$

Hence it gives us the desired (asymptotic) $100(1-\alpha)\%$ confidence interval (A.1) for p. ■

Theorem A.2 (Wilson)**.** *Wilson's (asymptotic)* $100(1-\alpha)\%$ *confidence interval for* p *is given by*

$$\left[\frac{\hat{p}_n + \frac{z_{\alpha/2}^2}{2n} - z_{\alpha/2}\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n} + \frac{z_{\alpha/2}^2}{4n^2}}}{1+\frac{z_{\alpha/2}^2}{n}}, \quad \frac{\hat{p}_n + \frac{z_{\alpha/2}^2}{2n} + z_{\alpha/2}\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n} + \frac{z_{\alpha/2}^2}{4n^2}}}{1+\frac{z_{\alpha/2}^2}{n}}\right]. \tag{A.2}$$

Proof. Wilson's (asymptotic) confidence interval for p can be proven in a different way than Wald's confidence interval. Let $z_{\alpha/2}$ be the desired upper $\alpha/2$-quantile of the standard normal distribution. Then by (3.8) it holds true that

$$z_{\alpha/2} \approx \sqrt{n}\frac{\hat{p}_n - p}{\sqrt{p(1-p)}}.$$

If we square both sides, and rewrite it a bit, we obtain

$$(\hat{p}_n - p)^2 = z_{\alpha/2}^2 \frac{p(1-p)}{n}.$$

Then it follows that

$$p^2 - 2\hat{p}_n p + \hat{p}_n^2 = \frac{z_{\alpha/2}^2 p}{n} - \frac{z_{\alpha/2}^2 p^2}{n},$$

and so we find the equality

$$\left(1 + \frac{z_{\alpha/2}^2}{n}\right) p^2 - \left(2\hat{p}_n + \frac{z_{\alpha/2}^2}{n}\right) p + \hat{p}_n^2 = 0.$$

The roots of this equality are given by

$$\begin{aligned}
p_{1,2} &= \frac{2\hat{p}_n + \frac{z_{\alpha/2}^2}{n} \pm \sqrt{\left(2\hat{p}_n + \frac{z_{\alpha/2}^2}{n}\right)^2 - 4\left(1 + \frac{z_{\alpha/2}^2}{n}\right)\hat{p}_n^2}}{2\left(1 + \frac{z_{\alpha/2}^2}{n}\right)} \\
&= \frac{2\hat{p}_n + \frac{z_{\alpha/2}^2}{n} \pm \sqrt{4\hat{p}_n^2 + \frac{4z_{\alpha/2}^2\hat{p}_n}{n} + \frac{z_{\alpha/2}^4}{n^2} - 4\hat{p}_n^2 - \frac{4z_{\alpha/2}^2\hat{p}_n^2}{n}}}{2\left(1 + \frac{z_{\alpha/2}^2}{n}\right)} \\
&= \frac{2\hat{p}_n + \frac{z_{\alpha/2}^2}{n} \pm \sqrt{\frac{4z_{\alpha/2}^2\hat{p}_n}{n} + \frac{z_{\alpha/2}^4}{n^2} - \frac{4z_{\alpha/2}^2\hat{p}_n^2}{n}}}{2\left(1 + \frac{z_{\alpha/2}^2}{n}\right)} \\
&= \frac{\hat{p}_n + \frac{z_{\alpha/2}^2}{2n} \pm \sqrt{\frac{z_{\alpha/2}^2\hat{p}_n}{n} + \frac{z_{\alpha/2}^4}{4n^2} - \frac{z_{\alpha/2}^2\hat{p}_n^2}{n}}}{1 + \frac{z_{\alpha/2}^2}{n}} \\
&= \frac{\hat{p}_n + \frac{z_{\alpha/2}^2}{2n} \pm \sqrt{z_{\alpha/2}^2\left(\frac{\hat{p}_n(1-\hat{p}_n)}{n} + \frac{z_{\alpha/2}^2}{4n^2}\right)}}{1 + \frac{z_{\alpha/2}^2}{n}} \\
&= \frac{\hat{p}_n + \frac{z_{\alpha/2}^2}{2n} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n} + \frac{z_{\alpha/2}^2}{4n^2}}}{1 + \frac{z_{\alpha/2}^2}{n}},
\end{aligned}$$

which are exactly the endpoints of the interval (A.2). ∎

A.2 Exponential distribution

The proof of Theorem A.3 is based on Dirksen (2019b).

Theorem A.3. *Let X_1, X_2 be two independent $Exp(1/\mu)$ distributed random variables. The probability of one of the random variables being at least a factor k times the other random variable equals $2/(k+1)$.*

Proof. The probability of one of the random variables being a factor k times the other random variable is the probability of the union of the events $\{X_1 \geq kX_2\}$ and $\{X_2 \geq kX_1\}$. It holds true that

$$
\begin{aligned}
P(X_1 \geq kX_2) &= \int\int_{\{(x_1,x_2):x_1 \geq kx_2\}} \frac{1}{\mu}e^{-x_1/\mu}\frac{1}{\mu}e^{-x_2/\mu}\,dx_1\,dx_2 \\
&= \int_0^\infty \int_{kx_2}^\infty \frac{1}{\mu^2}e^{-(x_1+x_2)/\mu}\,dx_1\,dx_2 \\
&= \int_0^\infty \frac{1}{\mu}e^{-x_2/\mu} \int_{kx_2}^\infty \frac{1}{\mu}e^{-x_1/\mu}\,dx_1\,dx_2 \\
&= \lim_{R\to\infty} \int_0^\infty \frac{1}{\mu}e^{-x_2/\mu}\left[-e^{-x_1/\mu}\right]_{x_1=kx_2}^{x_1=R} dx_2 \\
&= \int_0^\infty \frac{1}{\mu}e^{-x_2/\mu}e^{-kx_2}\,dx_2 \\
&= \int_0^\infty \frac{1}{\mu}e^{-(k+1)x_2/\mu}\,dx_2.
\end{aligned}
$$

Let $u = (k+1)x_2$. Then it follows that

$$
\begin{aligned}
P(X_1 \geq kX_2) &= \int_0^\infty \frac{1}{\mu}e^{-u/\mu}\frac{1}{k+1}\,du. \\
&= \frac{1}{k+1}\lim_{R\to\infty}\left[-e^{-u}\right]_{u=0}^{u=R} \\
&= \frac{1}{k+1}.
\end{aligned}
$$

By symmetry, we also have that $P(X_2 \geq kX_1) = 1/(k+1)$. Therefore, we conclude that the statement of the theorem holds. ■

A.3 Negative binomial distribution

The proof of Theorem A.4 is based on Ma (2011).

Theorem A.4. *The negative binomial distribution with parameters r and p can be viewed as a Pois(μ) distribution, where μ is itself a gamma distributed random variable with parameters $\alpha = r$ and $\beta = p/(1-p)$.*

Proof. We are going to show that the unconditional probability mass function of the defined hierarchical model is indeed the one of a negative binomial distribution. Let X be a Poisson random variable with parameter $\mu > 0$, where μ should be treated as a random variable M. Then it holds true that

$$P(X = k|M = \mu) = \frac{\mu^k}{k!}e^{-\mu}, \qquad k = 0, 1, 2, \ldots.$$

Suppose M is gamma distributed with parameters $\alpha > 0$ and $\beta > 0$. Then its probability density function π is given by

$$\pi(\mu) = \frac{\beta^\alpha}{\Gamma(\alpha)}\mu^{\alpha-1}e^{-\beta\mu}, \qquad \mu \geq 0,$$

where $\Gamma(\alpha)$ is the gamma function defined as in (2.6). We know that $\Gamma(\alpha+1) = \alpha\Gamma(\alpha)$ for $\alpha > 0$ and $\Gamma(n) = (n-1)!$ for $n = 1, 2, \ldots$ (Dekking et al., 2010).

Now, the unconditional distribution of X is obtained using the law of total probability:

$$\begin{aligned}
P(X = k) &= \int_0^\infty P(X = k|M = \mu)\pi(\mu)\,d\mu \\
&= \int_0^\infty \frac{\mu^k}{k!}e^{-\mu}\frac{\beta^\alpha}{\Gamma(\alpha)}\mu^{\alpha-1}e^{-\beta\mu}\,d\mu \\
&= \frac{1}{k!}\frac{\beta^\alpha}{\Gamma(\alpha)}\int_0^\infty \mu^{k+\alpha-1}e^{-(\beta+1)\mu}\,d\mu \\
&= \frac{1}{k!}\frac{\beta^\alpha}{\Gamma(\alpha)}\frac{\Gamma(k+\alpha)}{(\beta+1)^{k+\alpha}}\int_0^\infty \frac{(\beta+1)^{k+\alpha}}{\Gamma(k+\alpha)}\mu^{k+\alpha-1}e^{-(\beta+1)\mu}\,d\mu \\
&= \frac{1}{k!}\frac{\beta^\alpha}{\Gamma(\alpha)}\frac{\Gamma(k+\alpha)}{(\beta+1)^{k+\alpha}} \\
&= \frac{\Gamma(k+\alpha)}{k!\Gamma(\alpha)}\left(\frac{\beta}{\beta+1}\right)^\alpha\left(\frac{1}{\beta+1}\right)^k \\
&= \frac{\Gamma(k+\alpha)}{\Gamma(k+1)\Gamma(\alpha)}\left(\frac{\beta}{\beta+1}\right)^\alpha\left(\frac{1}{\beta+1}\right)^k \\
&= \binom{k+\alpha-1}{k}\left(\frac{\beta}{\beta+1}\right)^\alpha\left(\frac{1}{\beta+1}\right)^k
\end{aligned}$$

In the fourth step we used that the integral equals one, since the function inside the integral is the probability density function of a $Gam(k+\alpha, \beta+1)$ distributed random variable.

If we plug in $\alpha = r$ and $\beta = p/(1-p)$ we find that

$$\begin{aligned}
P(X = k) &= \binom{k+r-1}{k}\left(\frac{p/(1-p)}{p/(1-p)+1}\right)^\alpha\left(\frac{1}{p/(1-p)+1}\right)^k \\
&= \binom{k+r-1}{k}p^r(1-p)^k.
\end{aligned}$$

Because probability mass functions are uniquely determined, we conclude that the statement of the theorem is indeed true. ■

A.4 Law of rare events

The proof of Theorem A.5 is based on (Dekking et al., 2010, Section 12.2).

Theorem A.5 (Law of rare events)**.** *Let X be a $Bin(n,p)$ random variable. As $n \to \infty$ and $p \to 0$ such that $np \to \lambda$, for $\lambda > 0$, then X is asymptotically $Pois(\lambda)$ distributed. In other words, when the size of the population n is very large and the occurrence of an event rare, hence the probability p is very small, then we can approximate a binomial distributed random variable X by a Poisson distributed random variable.*

Proof. Let X be a $Bin(n,p)$ random variable with the property that $n \to \infty$ and $p \to 0$ such that $np \to \lambda$, for $\lambda > 0$. Then the probability mass function of X equals

$$P(X=k) = \binom{n}{k} p^k (1-p)^{n-k}, \qquad k = 0, 1, \ldots, n.$$

Now, we look at its limit as n becomes arbitrary large. It holds true that

$$\begin{aligned}
\lim_{n\to\infty} P(X=k) &= \lim_{n\to\infty} \binom{n}{k} p^k (1-p)^{n-k} \\
&= \lim_{n\to\infty} \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k} \\
&= \lim_{n\to\infty} \frac{n!}{k!(n-k)!} \left(\frac{np}{n}\right)^k \left(1 - \frac{np}{n}\right)^{n-k} \\
&= \lim_{n\to\infty} \frac{n!}{k!(n-k)!} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k} \\
&= \frac{\lambda^k}{k!} \lim_{n\to\infty} \frac{n!}{(n-k)!} \frac{1}{n^k} \left(1 - \frac{\lambda}{n}\right)^n \left(1 - \frac{\lambda}{n}\right)^{-k}.
\end{aligned}$$

We handle the above limit in three separate parts. First, notice that we can write

$$\begin{aligned}
\frac{n!}{(n-k)!} \frac{1}{n^k} &= \frac{n(n-1)(n-2)\ldots(n-k+1)}{n^k} \\
&= \frac{n}{n} \frac{n-1}{n} \frac{n-2}{n} \cdots \frac{n-k+1}{n}.
\end{aligned}$$

Each fraction in the above expression goes to 1 as n becomes arbitrarily large. Therefore, we find that

$$\lim_{n\to\infty} \frac{n!}{(n-k)!} \frac{1}{n^k} = 1.$$

For the second part, it holds that[1]

$$\lim_{n\to\infty} \left(1 - \frac{\lambda}{n}\right)^n = e^{-\lambda}.$$

Finally, since k is fixed we see that

$$\lim_{n\to\infty} \left(1 - \frac{\lambda}{n}\right)^{-k} = 1.$$

We conclude that

$$\lim_{n\to\infty} P(X=k) = \frac{\lambda^k}{k!} e^{-\lambda},$$

which is exactly the probability mass function of a Poisson distributed random variable with parameter $\lambda > 0$. ■

[1]It is well known that $\lim_{n\to\infty} \left(1 + \frac{x}{n}\right)^n = e^x$ (Adams and Essex, 2014, Chapter 3, Theorem 6).

A.5 Example credit card fraud

Lemma A.1. *In Example 6.1, it holds true that the probabilities of the evidence given the defence hypothesis for all allowed combinations of assistants are given by*

$$\begin{aligned}
P(E|\overline{G}, C, A_{c_1}, \dots, A_{c_{10}}, A_{r_1}) &= \frac{1}{216}, & P(E|\overline{G}, C, A_{c_1}, \dots, A_{c_{10}}, A_{r_5}) &= \frac{1}{6},\\
P(E|\overline{G}, C, A_{c_1}, \dots, A_{c_{10}}, A_{r_2}) &= \frac{2}{27}, & P(E|\overline{G}, C, A_{c_1}, \dots, A_{c_{10}}, A_{r_6}) &= \frac{4}{9},\\
P(E|\overline{G}, C, A_{c_1}, \dots, A_{c_{10}}, A_{r_3}) &= \frac{1}{54}, & P(E|\overline{G}, C, A_{c_1}, \dots, A_{c_{10}}, A_{r_7}) &= \frac{4}{27},\\
P(E|\overline{G}, C, A_{c_1}, \dots, A_{c_{10}}, A_{r_4}) &= \frac{1}{54}, & P(E|\overline{G}, C, A_{c_1}, \dots, A_{c_{10}}, A_{r_8}) &= 1.
\end{aligned}$$

Proof. To compute the above probabilities, we need the information of the subsets $A_{r_1}, \dots, A_{r_8}$ and $A_{c_1}, \dots, A_{c_{10}}$ and equation (6.4). It holds true that

$$\left[P(E|\overline{G}, c_q, A_{c_q}, A_{r_i})\right]_{iq} = \begin{pmatrix}
\frac{1}{2} & 1 & \frac{1}{2} & \frac{1}{3} & 1 & \frac{1}{3} & 1 & \frac{1}{3} & \frac{1}{2} & 1\\
1 & 1 & \frac{1}{2} & \frac{2}{3} & 1 & \frac{2}{3} & 1 & \frac{2}{3} & \frac{1}{2} & 1\\
\frac{1}{2} & 1 & \frac{1}{2} & \frac{2}{3} & 1 & \frac{2}{3} & 1 & \frac{1}{3} & \frac{1}{2} & 1\\
\frac{1}{2} & 1 & \frac{1}{2} & \frac{1}{3} & 1 & \frac{1}{3} & 1 & \frac{2}{3} & 1 & 1\\
1 & 1 & \frac{1}{2} & 1 & 1 & 1 & 1 & \frac{2}{3} & \frac{1}{2} & 1\\
1 & 1 & 1 & \frac{2}{3} & 1 & \frac{2}{3} & 1 & 1 & 1 & 1\\
\frac{1}{2} & 1 & 1 & \frac{2}{3} & 1 & \frac{2}{3} & 1 & \frac{2}{3} & 1 & 1\\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1
\end{pmatrix},$$

for $i = 1, \dots, 8$ (rows) and $q = 1, \dots, 10$ (columns). Then we find by (6.5) that

$$P(E|\overline{G}, C, A_{c_1}, \dots, A_{c_{10}}, A_{r_i}) = \begin{cases}
\frac{1}{2} \cdot 1 \cdot \frac{1}{2} \cdot \frac{1}{3} \cdot 1 \cdot \frac{1}{3} \cdot 1 \cdot \frac{1}{3} \cdot \frac{1}{2} \cdot 1 = \frac{1}{216} & \text{for } i = 1,\\
1 \cdot 1 \cdot \frac{1}{2} \cdot \frac{2}{3} \cdot 1 \cdot \frac{2}{3} \cdot 1 \cdot \frac{2}{3} \cdot \frac{1}{2} \cdot 1 = \frac{2}{27} & \text{for } i = 2,\\
\frac{1}{2} \cdot 1 \cdot \frac{1}{2} \cdot \frac{2}{3} \cdot 1 \cdot \frac{2}{3} \cdot 1 \cdot \frac{1}{3} \cdot \frac{1}{2} \cdot 1 = \frac{1}{54} & \text{for } i = 3,\\
\frac{1}{2} \cdot 1 \cdot \frac{1}{2} \cdot \frac{1}{3} \cdot 1 \cdot \frac{1}{3} \cdot 1 \cdot \frac{2}{3} \cdot 1 \cdot 1 = \frac{1}{54} & \text{for } i = 4,\\
1 \cdot 1 \cdot \frac{1}{2} \cdot 1 \cdot 1 \cdot 1 \cdot 1 \cdot \frac{2}{3} \cdot \frac{1}{2} \cdot 1 = \frac{1}{6} & \text{for } i = 5,\\
1 \cdot 1 \cdot 1 \cdot \frac{2}{3} \cdot 1 \cdot \frac{2}{3} \cdot 1 \cdot 1 \cdot 1 \cdot 1 = \frac{4}{9} & \text{for } i = 6,\\
\frac{1}{2} \cdot 1 \cdot 1 \cdot \frac{2}{3} \cdot 1 \cdot \frac{2}{3} \cdot 1 \cdot \frac{2}{3} \cdot 1 \cdot 1 = \frac{4}{27} & \text{for } i = 7,\\
1 \cdot 1 \cdot 1 \cdot 1 \cdot 1 \cdot 1 \cdot 1 \cdot 1 \cdot 1 \cdot 1 = 1 & \text{for } i = 8,
\end{cases}$$

as desired. ■

A.6 Parameter estimation

Lemma A.2. *The moment estimator of the parameter ρ in the mixed Poisson model, discussed in Subsection 11.2.2, is*

$$\hat{\rho}_{MOM} = \frac{\left(\frac{\overline{x}}{\overline{t}}\right)^2}{\frac{\overline{x^2}}{\overline{t^2}} - \left(\frac{\overline{x}}{\overline{t}}\right)^2 - \frac{\overline{x}}{\overline{t^2}}}. \tag{A.3}$$

Proof. From Subsection 11.2.2, we already have the following approximation:

$$\mu \sum_{i=1}^{n} t_i + \mu^2 \left(1 + \frac{1}{\rho}\right) \sum_{i=1}^{n} t_i^2 \approx \sum_{i=1}^{n} x_i^2$$

We are going to solve this equation for ρ. It holds true that

$$\mu^2 \left(1 + \frac{1}{\rho}\right) \sum_{i=1}^{n} t_i^2 = \sum_{i=1}^{n} x_i^2 - \mu \sum_{i=1}^{n} t_i,$$

and hence

$$\begin{aligned} 1 + \frac{1}{\rho} &= \frac{1}{\mu^2} \frac{\sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n} t_i^2} - \frac{1}{\mu} \frac{\sum_{i=1}^{n} t_i}{\sum_{i=1}^{n} t_i^2} \\ &= \frac{1}{\mu^2} \frac{n\overline{x^2}}{n\overline{t^2}} - \frac{1}{\mu} \frac{n\overline{t}}{n\overline{t^2}} \\ &= \frac{1}{\mu^2} \frac{\overline{x^2}}{\overline{t^2}} - \frac{1}{\mu} \frac{\overline{t}}{\overline{t^2}}. \end{aligned}$$

Consequently, we find that

$$\begin{aligned} \frac{1}{\rho} &= \frac{1}{\mu^2} \frac{\overline{x^2}}{\overline{t^2}} - \frac{1}{\mu} \frac{\overline{t}}{\overline{t^2}} - 1 \\ &= \frac{\frac{\overline{x^2}}{\overline{t^2}} - \frac{\overline{t}}{\overline{t^2}}\mu - \mu^2}{\mu^2}. \end{aligned}$$

By (11.3), the moment estimator for μ equals

$$\hat{\mu}_{MOM} = \frac{\overline{x}}{\overline{t}}.$$

Hence if we plug this equation into our expression for $1/\rho$, we find

$$\begin{aligned} \frac{1}{\rho} &= \frac{\frac{\overline{x^2}}{\overline{t^2}} - \frac{\overline{t}}{\overline{t^2}} \frac{\overline{x}}{\overline{t}} - \left(\frac{\overline{x}}{\overline{t}}\right)^2}{\left(\frac{\overline{x}}{\overline{t}}\right)^2} \\ &= \frac{\frac{\overline{x^2}}{\overline{t^2}} - \frac{\overline{x}}{\overline{t^2}} - \left(\frac{\overline{x}}{\overline{t}}\right)^2}{\left(\frac{\overline{x}}{\overline{t}}\right)^2}. \end{aligned}$$

Thus the moment estimator for ρ is

$$\hat{\rho}_{MOM} = \frac{\left(\frac{\overline{x}}{\overline{t}}\right)^2}{\frac{\overline{x^2}}{\overline{t^2}} - \frac{\overline{x}}{\overline{t^2}} - \left(\frac{\overline{x}}{\overline{t}}\right)^2},$$

as desired. ■

Lemma A.3. *Assume that the number of shifts t is fixed and the same for each nurse. The maximum likelihood estimator of the parameter μ in the mixed Poisson model, discussed in Subsection 11.2.2, is*

$$\hat{\mu}_{MLE} = \frac{\overline{x}}{t}. \tag{A.4}$$

Proof. Suppose that we have n nurses. Let x_i be the number of incidents that a nurse witnessed during t shifts, for $i = 1, \ldots, n$. Then the likelihood function equals

$$\begin{aligned}
\mathcal{L}(x_1, \ldots, x_n; \rho, \mu) &= \prod_{i=1}^{n} P(N(t) = x_i) \\
&= \prod_{i=1}^{n} \frac{\Gamma(x_i + \rho)}{x_i!\Gamma(\rho)} \left(\frac{1}{1 + t\mu/\rho}\right)^{\rho} \left(\frac{t\mu/\rho}{1 + t\mu/\rho}\right)^{x_i} \\
&= \left(\frac{1}{1 + t\mu/\rho}\right)^{n\rho} \left(\frac{t\mu/\rho}{1 + t\mu/\rho}\right)^{\sum_{i=1}^{n} x_i} \prod_{i=1}^{n} \frac{\Gamma(x_i + \rho)}{x_i!\Gamma(\rho)}.
\end{aligned}$$

It follows that the log likelihood function is

$$\begin{aligned}
\ell(x_1, \ldots, x_n; \rho, \mu) &= -n\rho \log(1 + t\mu/\rho) + \sum_{i=1}^{n} x_i \left[\log(t\mu/\rho) - \log(1 + t\mu/\rho)\right] \\
&\quad + \sum_{i=1}^{n} \left[\log \Gamma(x_i + \rho) - \log x_i! - \log \Gamma(\rho)\right] \\
&= -n\rho \log(1 + t\mu/\rho) + n\overline{x} \log(t\mu/\rho) - n\overline{x} \log(1 + t\mu/\rho) \\
&\quad + \sum_{i=1}^{n} \left[\log \Gamma(x_i + \rho) - \log x_i!\right] - n \log \Gamma(\rho). \qquad \text{(A.5)}
\end{aligned}$$

We compute the partial derivative with respect to μ of the log likelihood function and equal it to zero. It holds true that

$$\begin{aligned}
\frac{\partial \ell(x_1, \ldots, x_n; \rho, \mu)}{\partial \mu} &= -\frac{nt}{1 + t\mu/\rho} + \frac{n\overline{x}}{\mu} - \frac{nt\overline{x}/\rho}{1 + t\mu/\rho} \\
&= -\frac{nt(1 + \overline{x}/\rho)}{1 + t\mu/\rho} + \frac{n\overline{x}}{\mu} \\
&= 0.
\end{aligned}$$

If we solve this equation for μ, we find:

$$\begin{aligned}
\frac{nt(1 + \overline{x}/\rho)}{1 + t\mu/\rho} &= \frac{n\overline{x}}{\mu} \\
t\mu(1 + \overline{x}/\rho) &= \overline{x}(1 + t\mu/\rho) \\
t\mu &= \overline{x}
\end{aligned}$$

Hence

$$\hat{\mu}_{MLE} = \frac{\overline{x}}{t},$$

as desired. It is indeed a maximum likelihood estimator by the second derivative test. ■

Lemma A.4. *Assume that the number of shifts t is fixed and the same for each nurse. The maximum likelihood estimator $\hat{\rho}_{MLE}$ of the parameter ρ in the mixed Poisson model, discussed in Subsection 11.2.2, is the solution to the equation*

$$n \log(1 + \overline{x}/\rho) = \sum_{i=1}^{n} \sum_{j=0}^{x_i - 1} \frac{1}{\rho + j}. \qquad \text{(A.6)}$$

Proof. Again, we assume that we have n nurses and let x_i be the number of incidents that a nurse witnessed during t shifts, for $i = 1, \ldots, n$. Now, we take the partial derivative with respect to ρ of the log likelihood function given in (A.5) and equal it to zero. Define

$$\psi(x) := \frac{\Gamma'(x)}{\Gamma(x)},$$

which is called the *digamma function* (Weisstein, n.d.a). Then we find that

$$\begin{aligned}\frac{\partial \ell(x_1,\ldots,x_n;\rho,\mu)}{\partial \rho} &= -n\log(1+t\mu/\rho)+\frac{nt\mu/\rho}{1+t\mu/\rho}-\frac{n\overline{x}}{\rho}-\frac{n\overline{x}/\rho^2}{1+t\mu/\rho}-n\psi(\rho)+\sum_{i=1}^{n}\psi(x_i+\rho)\\ &= -n\log(1+t\mu/\rho)+\frac{1}{\rho}\frac{nt\mu(1+\overline{x}/\rho)}{1+t\mu/\rho}-\frac{n\overline{x}}{\rho}-n\psi(\rho)+\sum_{i=1}^{n}\left[\psi(\rho)+\sum_{j=0}^{x_i-1}\frac{1}{\rho+j}\right]\\ &= -n\log(1+t\mu/\rho)+\frac{n(t\mu-\overline{x})/\rho}{1+t\mu/\rho}+\sum_{i=1}^{n}\sum_{j=0}^{x_i-1}\frac{1}{\rho+j}\\ &= 0,\end{aligned}$$

where we used the digamma function satisfies the equation $\psi(x+1)=\psi(x)+1/x$ (Weisstein, n.d.a). By Lemma A.3, we know that

$$\hat{\mu}_{MLE}=\frac{\overline{x}}{t}.$$

So it follows that the maximum likelihood estimator of ρ satisfies the following equation:

$$n\log(1+\overline{x}/\rho)=\sum_{i=1}^{n}\sum_{j=0}^{x_i-1}\frac{1}{\rho+j}$$

■

Theorem A.6. *The moment estimators $\hat{\rho}_{MOM}$ and $\hat{\mu}_{MOM}$, discussed in Subsection 11.2.2, exist with probability tending to one and are asymptotically normal.*

Proof. Let ρ_0 and μ_0 be the true parameters of the mixed Poisson model. Define

$$e(\rho,\mu):=\begin{pmatrix}E(N(t))\\ E(N^2(t))\end{pmatrix}=\begin{pmatrix}t\mu\\ t\mu+(t\mu)^2\left(1+\frac{1}{\rho}\right)\end{pmatrix}.$$

To prove that the moment estimators $\hat{\rho}_{MOM}$ and $\hat{\mu}_{MOM}$ are asymptotically normal, we have to first check the regularity conditions of (van der Vaart, 1998, Theorem 3.7). They state that $e(\rho,\mu)$ has to be one-to-one on an open parameter set in $\mathbb{R}^2$, and continuously differentiable at (ρ_0,μ_0) with a nonsingular derivative e'_{ρ_0,μ_0}. Moreover, it has to hold that $E_{\rho_0,\mu_0}(N^2(t)+N^4(t))<\infty$.

It follows that

$$e'_{\rho,\mu}=\begin{pmatrix}\frac{\partial e_1(\rho,\mu)}{\partial\rho} & \frac{\partial e_1(\rho,\mu)}{\partial\mu}\\ \frac{\partial e_2(\rho,\mu)}{\partial\rho} & \frac{\partial e_2(\rho,\mu)}{\partial\mu}\end{pmatrix}=\begin{pmatrix}0 & t\\ -\frac{(t\mu)^2}{\rho^2} & t+t^2\mu\left(1+\frac{1}{\rho}\right)\end{pmatrix}.$$

Hence $e(\rho,\mu)$ is continuously differentiable at (ρ_0,μ_0), one-to-one and e'_{ρ_0,μ_0} is non-singular for $t\neq 0$. It holds that

$$\begin{aligned}E(N^2(t)+N^4(t)) &= E(N^2(t))+E(N^4(t))\\ &= \left[t\mu+(t\mu)^2\left(1+\frac{1}{\rho}\right)\right]+\left[t\mu+(t\mu)^2\left(7+\frac{7}{\rho}\right)+(t\mu)^3\left(6+\frac{18}{\rho}+\frac{12}{\rho^2}\right)\right.\\ &\left.+(t\mu)^4\left(1+\frac{6}{\rho}+\frac{11}{\rho^2}+\frac{6}{\rho^3}\right)\right]\\ &= 2t\mu+(t\mu)^2\left(8+\frac{8}{\rho}\right)+(t\mu)^3\left(6+\frac{18}{\rho}+\frac{12}{\rho^2}\right)+(t\mu)^4\left(1+\frac{6}{\rho}+\frac{11}{\rho^2}+\frac{6}{\rho^3}\right).\end{aligned}$$

One can check the above moments of $N(t)$ by using the moment generation function of the negative binomial distribution. Hence we have that $E_{\rho_0,\mu_0}(N^2(t)+N^4(t))<\infty$, as well.

Thus, the regularity conditions of (van der Vaart, 1998, Theorem 3.7) are met. We conclude that the moment estimators $\hat{\rho}_{MOM}$ and $\hat{\mu}_{MOM}$ exist with probability tending to one and satisfy

$$\sqrt{n}\left(\begin{pmatrix}\hat{\rho}_{MOM}\\ \hat{\mu}_{MOM}\end{pmatrix}-\begin{pmatrix}\rho_0\\ \mu_0\end{pmatrix}\right)\xrightarrow{d}N\left(0,e'^{-1}_{\rho_0,\mu_0}\Sigma_{\rho_0,\mu_0}(e'^{-1}_{\rho_0,\mu_0})^T\right), \tag{A.7}$$

where Σ_{ρ_0,μ_0} is the covariance matrix of the vector $(N(t), N^2(t))$ under (ρ_0, μ_0). The covariance matrix equals

$$\Sigma_{\rho,\mu} = \begin{pmatrix} cov(N(t), N(t)) & cov(N(t), N^2(t)) \\ cov(N^2(t), N(t)) & cov(N^2(t), N^2(t)) \end{pmatrix},$$

where one can check that

$$\begin{aligned} cov(N(t), N(t)) &= E(N^2(t)) - [E(N(t))]^2 \\ &= t\mu + \frac{(t\mu)^2}{\rho}, \end{aligned}$$

and

$$\begin{aligned} cov(N(t), N^2(t)) &= cov(N^2(t), N(t)) \\ &= E(N^3(t)) - E(N(t))E(N^2(t)) \\ &= t\mu + (t\mu)^2 \left(2 + \frac{3}{\rho}\right) + \frac{(t\mu)^3}{\rho} \left(2 + \frac{2}{\rho}\right), \end{aligned}$$

and

$$\begin{aligned} cov(N^2(t), N^2(t)) &= E(N^4(t)) - [E(N^2(t))]^2 \\ &= t\mu + (t\mu)^2 \left(6 + \frac{7}{\rho}\right) + (t\mu)^3 \left(4 + \frac{16}{\rho} + \frac{12}{\rho^2}\right) + \frac{(t\mu)^4}{\rho} \left(4 + \frac{10}{\rho} + \frac{6}{\rho^2}\right). \end{aligned}$$

■

B
Simulation study

In Chapter 11 we discussed three different methods to estimate the parameters ρ and μ of the gamma distribution that models the different incident rates of nurses in roster cases: a Gibbs sampling algorithm, the method of moments and a No-U-Turn sampler. Now, we want to perform some more simulation tests on these methods. Our goal is to look at the behaviour of the three models and discover if they differ for some invented scenarios. Based on this analysis, we want to give an advice which method to prefer.

B.1 Scenarios

We want to vary the following components:

- Values of ρ: $\rho = 0.25$, $\rho = 1$ and $\rho = 2.5$.
- Number of data points: $n = 25$, $n = 75$ and $n = 150$.
- Group composition of the nurses: (i) all nurses work $t = 150$ shifts, or (ii) 5% works $t = 15$ shifts, 80% works $t = 150$ shifts and 15% works $t = 200$ shifts.

We explain the thoughts behind these different scenarios. First, we already saw in Subsection 8.2.1 that the variance of the number of incidents during t shifts equals

$$Var(N(t)) = t\mu + \frac{(t\mu)^2}{\rho}.$$

If $\rho < 1$ the variance becomes larger than $t\mu + (t\mu)^2$, and smaller than $t\mu + (t\mu)^2$ if $\rho > 1$. Hence it is interesting to look at the behaviour of our estimators for different values of ρ since they influence the variance of the number incidents during a fixed time period. Secondly, the methods are dependent on the number of data points, as we already discussed before. We like to see if we still get similar outcomes if we vary the data sets. Finally, we assumed in our previous study that all nurses worked the same number of shifts. However, we think it is more likely that some nurses work more shifts than others. Therefore, we choose for an alternative scenario that in a group of nurses, 5% of the nurses works 15 shifts, 80% works 150 shifts and 15% works 200 shifts. Hence the first group portrays stand-in nurses, the second group represents the average nurses, and the final group works overtime.

The data sets are simulated in R. We assume that the data consisting of n nurses is sampled from a $NB(\rho, (1 + t\mu_0/\rho)^{-1})$ distribution, where $\mu_0 = 27/1734$ as before. In this way, we can compare the outcomes of the three models with the ground truths. Notice that for the mixed group, we vary the number of shifts t while drawing the data from the distribution. For each combination of the three components, we will simulate a new group of nurses and apply the three methods to the generated data set. Hence it is very well possible that some estimates will deviate more from the "true" parameters than others. It can be explained by a sampling variability of the randomly generated data. Therefore, we choose to perform the simulation for each scenario ten times and perform an Analysis of Variance to look whether the methods differ and specifically which ones.

Before we begin our simulation study, we want to point out that our Gibbs sampling algorithm is dependent on the chosen upper limit for the value of ρ. In our original simulations described in Subsection 11.2.1 we set this upper limit equal to 1.5, with a step size of 0.005. For the generated data that is sampled from the negative binomial distribution with parameter $\rho = 0.25$ or $\rho = 1$, it will not be a problem for the Gibbs sampling algorithm to estimate the parameter ρ close to their true values. This is because these values of ρ lie inside the interval from 0 to 1.5. However, for $\rho = 2.5$ it is not possible for the algorithm to estimate ρ correctly in advance. Hence, we will adapt the upper limit to 3, but keep the same step size of 0.005. Note that this choice will affect the computation time of the Gibbs sampling algorithm, since it has to compute a larger vector of probabilities given by (11.2).

We let the Gibbs sampling algorithm run for $T = 10000$ iterations with a burn-in period of $m = 3000$ iterations. For the No-U-Turn sampler, we take four chains with each 2500 iterations so that the number of possible divergences hopefully will not be too big compared to the total number of iterations.

The simulated data sets can be found in Appendix E.3 together with the obtained estimators for each method. In Appendix C.5, Figures C.29 through C.46, one can find scatter plots of the estimates of the parameters ρ and μ by each method.[1]

B.2 Analysis of Variance

As we said earlier, we want to perform an Analysis of Variance (ANOVA) to compare the different parameter estimation methods for various scenarios. Since for each scenario we simulate a new data set, and we have three methods that we can apply to estimate the two parameters ρ and μ, we choose to perform an Analysis of Variance using the randomised complete block design. This principle can be found in (Montgomery, 2019, Chapter 4). We use the following effects model to describe our observations for each of the parameters ρ and μ:

$$\hat{\rho}_{ij} = \rho_0 + \tau_i + \beta_j + \varepsilon_{ij}, \qquad \begin{cases} i = 1,2,3 \\ j = 1,\dots,10 \end{cases} \tag{B.1}$$

and

$$\hat{\mu}_{ij} = \mu_0 + \tau'_i + \beta'_j + \varepsilon'_{ij}, \qquad \begin{cases} i = 1,2,3 \\ j = 1,\dots,10 \end{cases} \tag{B.2}$$

where $\hat{\rho}_{ij}$ and $\hat{\mu}_{ij}$ are the observed estimators of method i of sample j for ρ and μ, respectively. The values ρ_0 and μ_0 represent the true values of the parameters. Furthermore, we have that τ_i and τ'_i are the effects of the ith method (treatment), and β_j and β'_j are the effects of the jth sample (block). Finally, ε_{ij} and ε'_{ij} are the random error terms.

We assume that effects of the three methods are fixed, the effects of the generated samples are random and hence $N(0, \sigma^2_\beta)$ distributed for some variance $\sigma^2_\beta > 0$. The random error terms are $N(0, \sigma^2_\varepsilon)$ distributed with $\sigma^2_\varepsilon > 0$, as usual.

Using this design, and thus creating blocks, we consider that some simulated data sets tend to deviate more from the ground truth than others. We are interested in testing the equality of the treatment means. The hypotheses of interest are

$$H_1 : \tau_1 = \tau_2 = \tau_3 = 0.$$
$$H_2 : \tau_i \neq 0 \,\text{at least one}\, i.$$

We use the test statistic

$$F_0 = \frac{MS_{\text{Treatment}}}{MS_E}, \tag{B.3}$$

where $MS_{\text{Treatment}}$ and MS_E are mean squares. We refer to (Montgomery, 2019, Chapter 4) for more details. We would reject H_1 if

$$F_0 > F_{\alpha,a-1,(a-1)(b-1)},$$

[1]We do not expect the reader to thoroughly study these figures. The purpose is to illustrate the general outcomes.

where $F_{\alpha,n,m}$ is the upper α-quantile of the F distribution with degrees of freedom n and m. The quantity α represents the predetermined significance level, $a = 3$ is the number of treatments and $b = 10$ the number of blocks.

When we do reject the null hypothesis, we perform a post-hoc test to figure out which treatment means differ. Hence which estimation method gives different outcomes. We will use Tukey's test. We test all pairwise mean comparisons. For equal sample sizes, the test declares that two means are significantly different if the absolute value of their sample differences exceeds

$$T_\alpha = q_\alpha(a, f)\sqrt{\frac{MS_E}{b}}, \tag{B.4}$$

where $q_\alpha(p, f)$ is the upper α percentage points of the Studentized Range Statistic. The quantity p is the size of group of sample means and f is the number of degrees of freedom associated with MS_E. In our case it holds true that $f = (a-1)(b-1)$.

The results are presented using ANOVA tables for a randomised complete block design. These tables are determined using Excel and IBM SPSS Statistics, and given in Appendix E.3.1, Tables E.5 through E.22. For some of the scenarios, we could not use the Gibbs sampling algorithm to estimate the parameters. Therefore, in these situations we will only compare the outcomes of the method of moments and the No-U-Turn sampler. We have that i only takes two values instead of three in this case. Also, for one sample the method of moments gave a negative estimated value for the parameter ρ, which is not possible. Therefore, we choose to ignore this observation and perform the analysis with nine blocks instead of ten. The discussed situations are marked in Appendix E.3.1.

B.2.1 Results

We apply a significance level of $\alpha = 0.05$. Since we also included the p-values in each of the tables, one can derive for each situation whether we reject the null hypothesis with respect to this significance level. If we do reject H_1 for one or both the parameters, we perform Tukey's test for the same significance level.

Now, we give a summary of the results. In most situations, for the parameter μ the null hypothesis is rejected ($p < 0.05$). Using Tukey's test, we found that in most of these cases the mean of the method of moments differs from the one of the Gibbs sampling algorithm and the No-U-Turn algorithm. However, while looking at the scatter plots in Appendix C.5, Figures C.29 through C.46, one sees that the spread of the estimators for each method are virtually the same. All estimates lie close to the true value, which was $\mu_0 = 27/1735 \approx 0.015571$. The spread becomes smaller when we increase the number of data points n. Hence the estimator becomes more accurate.

When looking at the results for ρ, we can conclude that for the scenario where nurses work the same number of shifts there is no difference between the treatment means. For the situation where nurses work different number of shifts, there is in general some difference between the results of the method of moments and those of the Gibbs sampling algorithm.

Furthermore, the Gibbs sampling algorithm overall does not work when we have many data points. Often it does not work in combination with a small value of ρ that we sampled from. From Appendix C.5, Figures C.29 through C.46, one can also infer that the other two methods estimate the value of ρ higher than the Gibbs sampling algorithm. In particular, it holds for $\rho_0 = 1$ and $\rho_0 = 2.5$. It may be due to the self-chosen upper bound in the algorithm.

For the method of moments, we notice that there are frequently outliers for ρ. It often happens when we have a small number of data points. For the scenario where nurses work different numbers of shifts, and hence correct in the moment estimator of ρ for the number of worked shifts, there is a possibility that ρ becomes negative. It is of course a disadvantage of this method. However, it is very well possible that this only happens for small data sets. Hence, we need to investigate this further.

For the No-U-Turn sampler, everything seems to work fine. Only for the combination of a small data set and a small value of ρ, we have some divergences. Hence it is important to have enough data available, but the rest works out nicely. So, this is a plus.

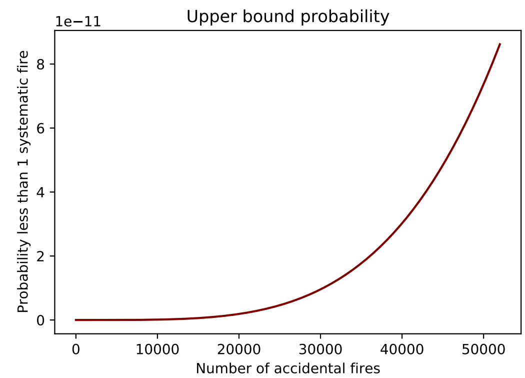
The advantage for the method of moments and No-U-Turn sampler above the Gibbs sampling algorithm, is that they (almost) always give us an estimate for both parameters. They also work a lot faster.

B.3 Conclusion

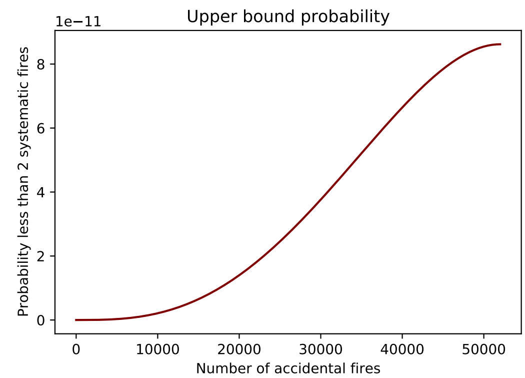
Ultimately, we prefer to use the No-U-Turn sampler. We refer to Section 11.3 for our motivation why, and the conclusion that we can draw from our simulation study and previously done experiments.

C
Figures

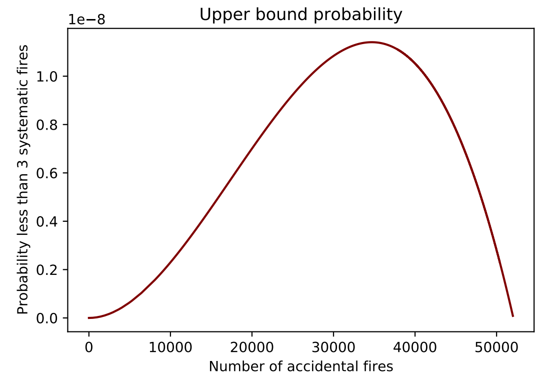
C.1 Arson



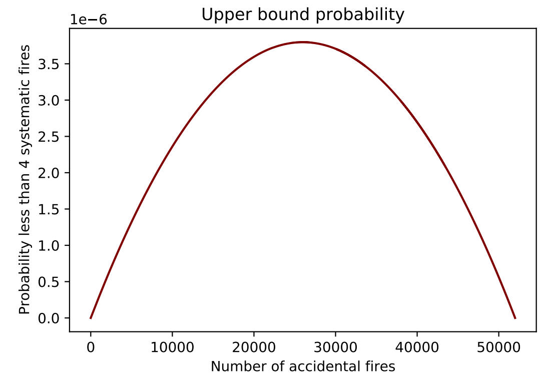
(a) $i = 1$



(b) $i = 2$



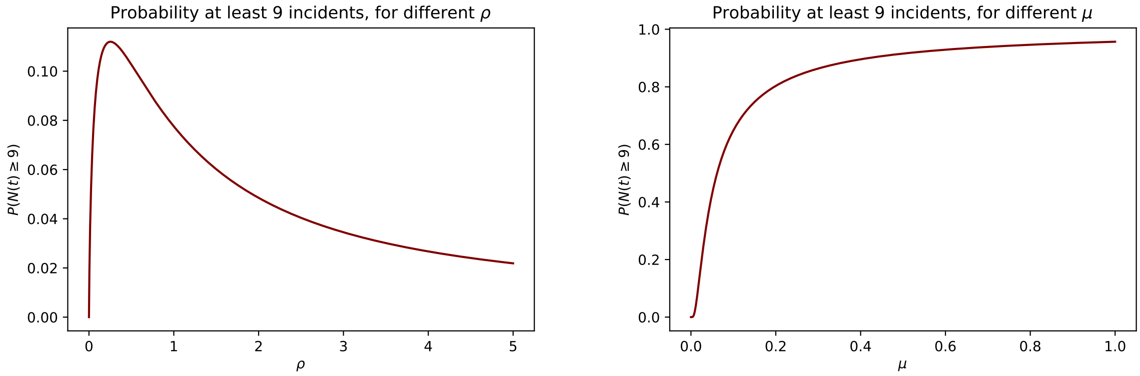
(c) $i = 3$



(d) $i = 4$

Figure C.1: Upper bound of the probability of having less than i systematic fires, using the binomial distribution for the number of accidental fires.

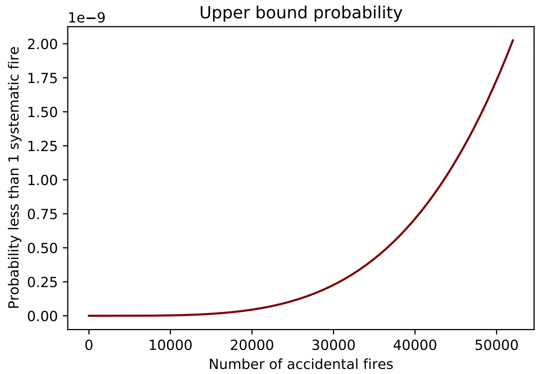
C.2 Roster cases



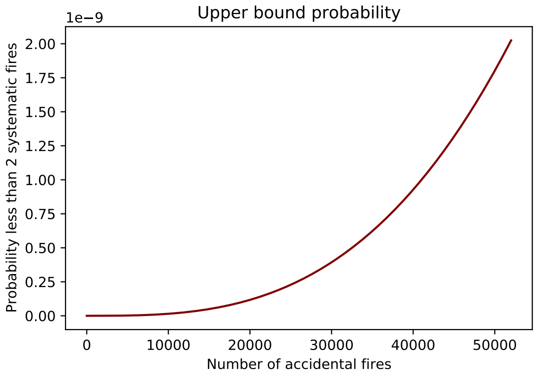
(a) $\rho \in (0, 5]$ (b) $\mu \in [0, 1]$

Figure C.2: The probability that the number of incidents in 203 shifts for one nurse is at least nine, for different values of ρ and μ. We set $t = 203$. In addition, we set $\mu = 26/1734$ in (a), and $\rho = 1$ in (b).

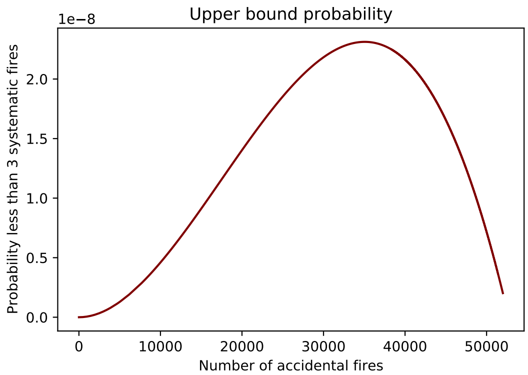
C.3 Arson (revised)



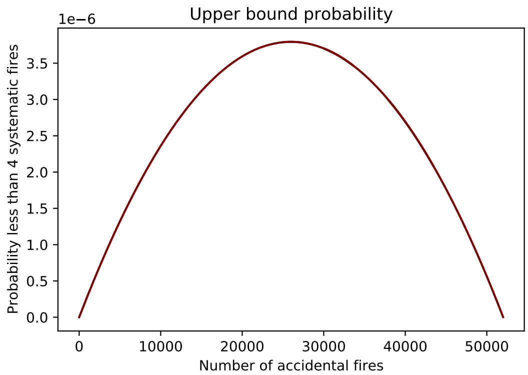
(a) $i = 1$



(b) $i = 2$



(c) $i = 3$



(d) $i = 4$

Figure C.3: Upper bound of the probability of having less than i systematic fires, using the negative binomial distribution for the number of accidental fires.

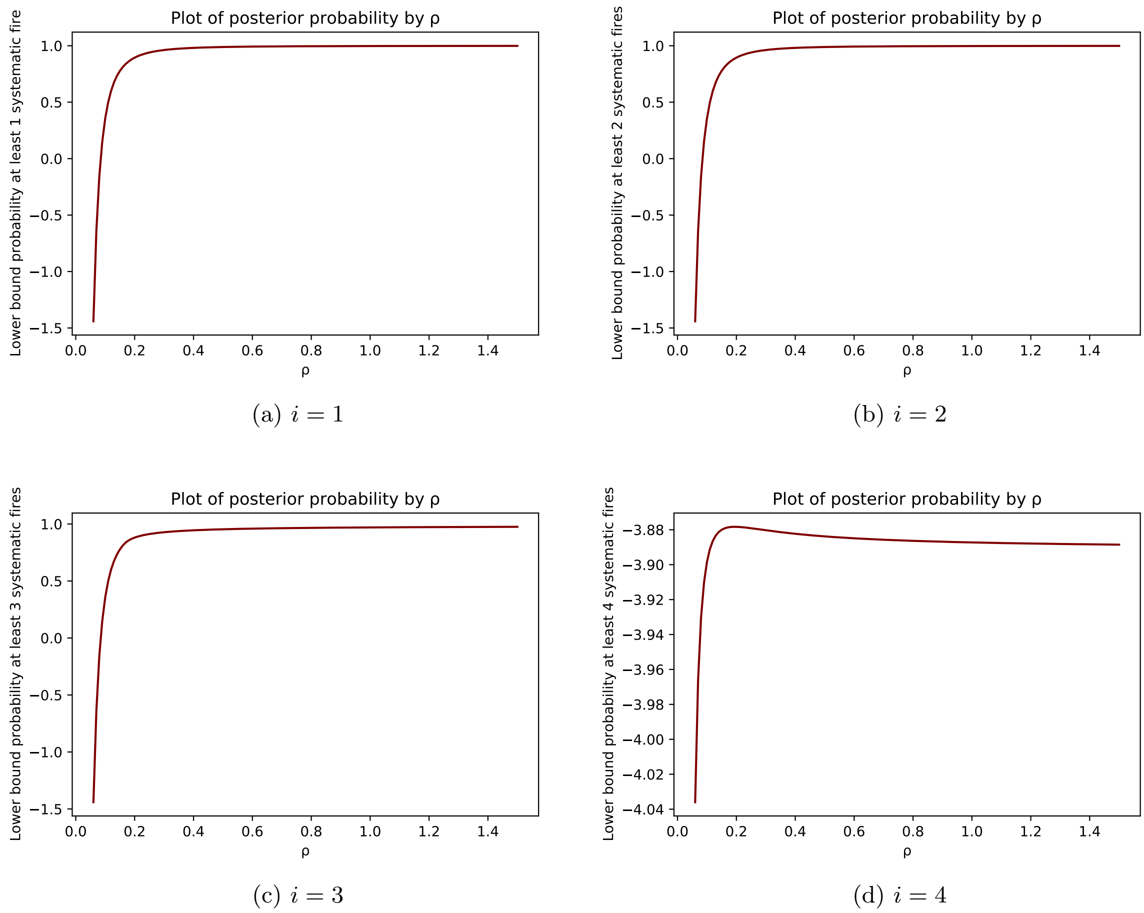


Figure C.4: Lower bound of the posterior probability of having at least i systematic fires for different values of ρ. Negative values for the posterior probability indicate that either q_1 or q_2 is negative valued, and hence have no meaning.

C.4 Roster cases (revised)

C.4.1 Gibbs sampling algorithm

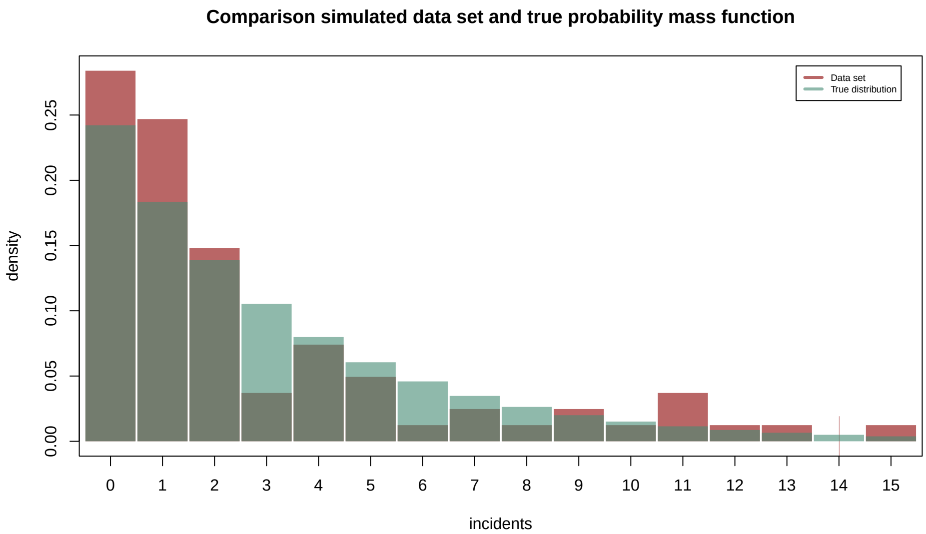
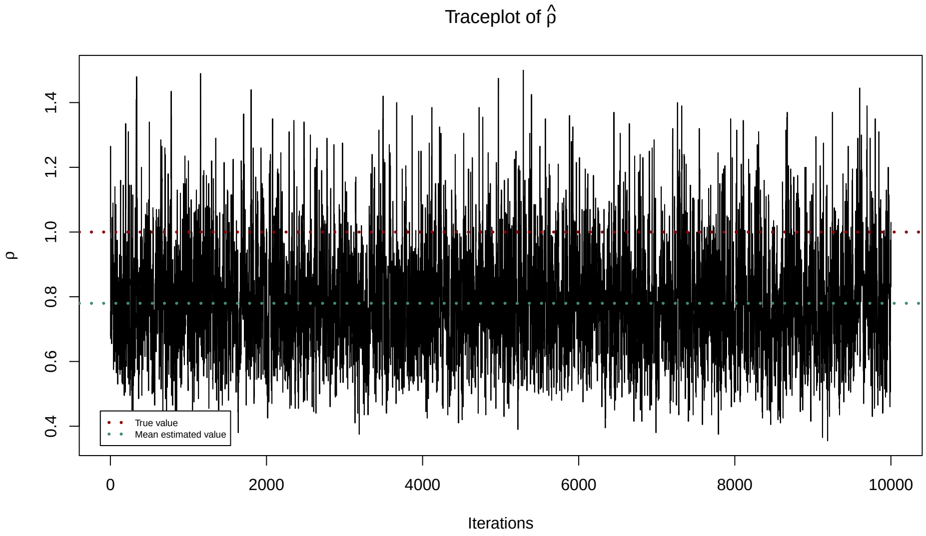
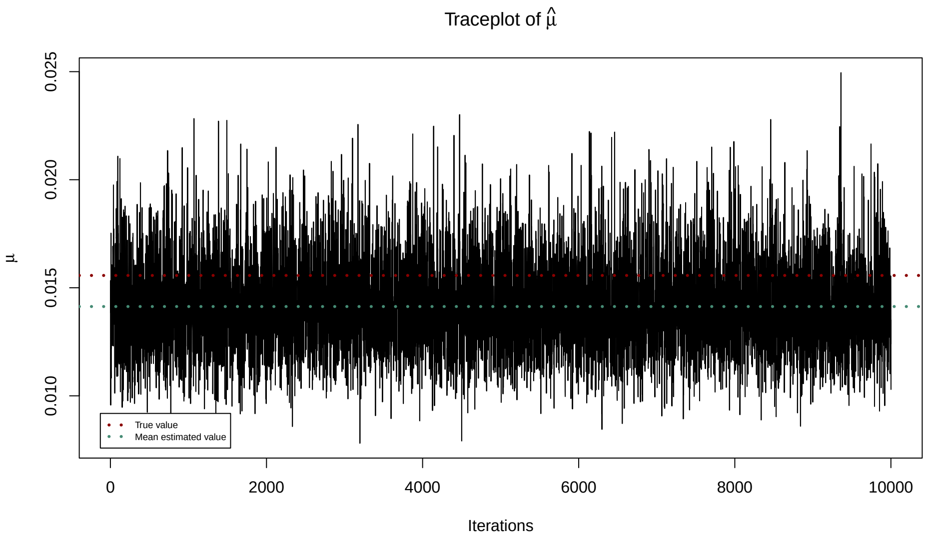


Figure C.5: Comparison of probabilities of simulated incidents of 81 nurses with true probability distribution for Gibbs sampling algorithm.



(a) Traceplot ρ



(b) Traceplot μ

Figure C.6: Traceplots of the estimated parameters ρ and μ using the Gibbs sampler. We used the number of incidents of $n = 81$ nurses, for $t = 201$ shifts. The dashed dark red lines indicate the true values $\rho_0 = 1$ and $\mu_0 = 27/1734$ of the parameters. The dashed cyan lines indicate the mean of the estimated parameters post burn-in.

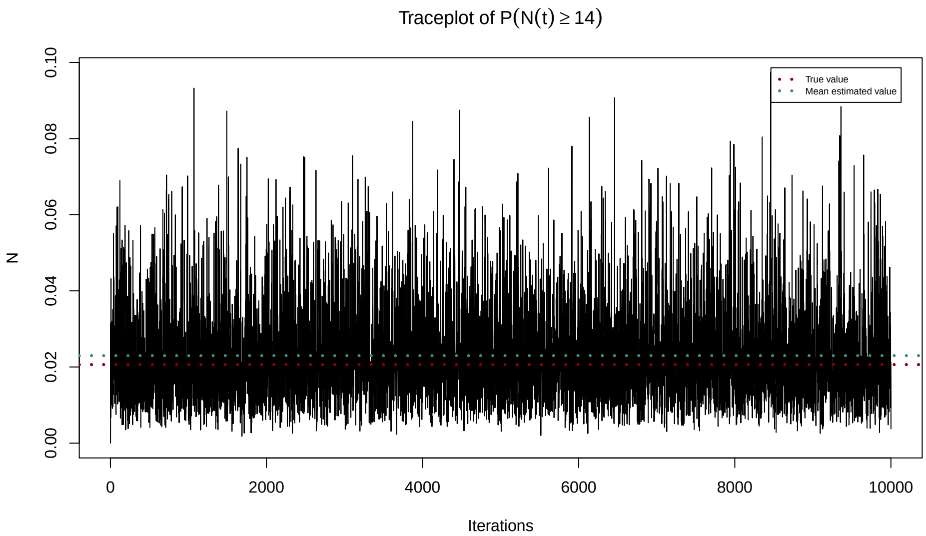
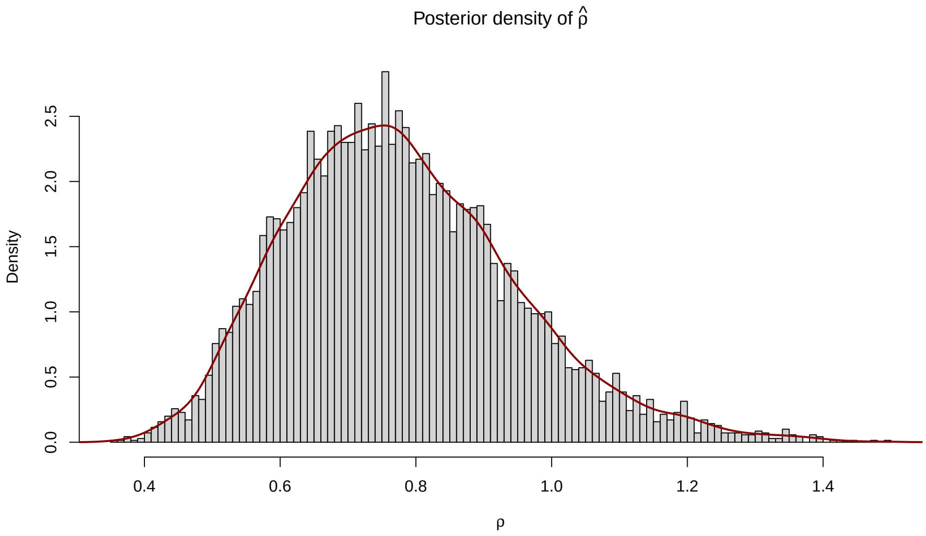
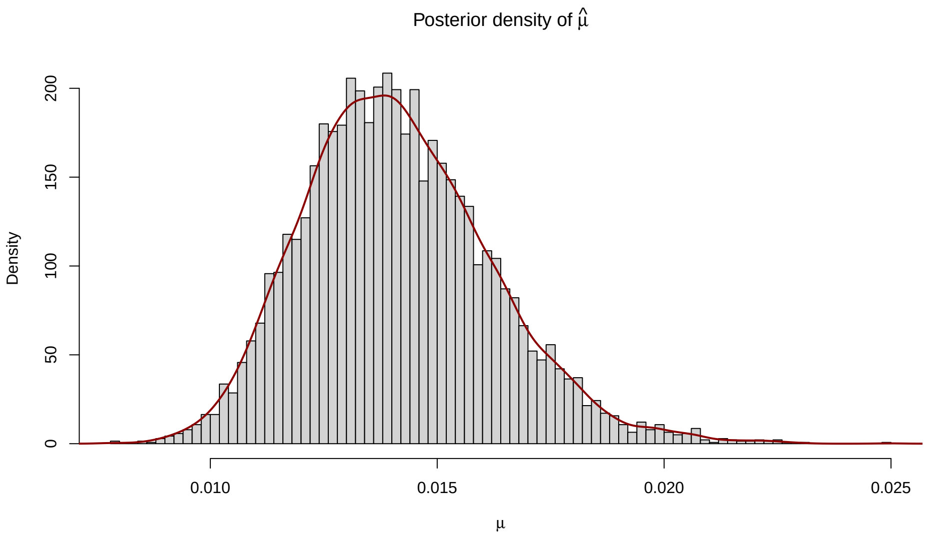


Figure C.7: Traceplot of the estimated probability $P(N(t) \geq 14)$ using the Gibbs sampler. We used the number of incidents of $n = 81$ nurses, for $t = 201$ shifts. The dashed dark red line indicates the true value of the probability $P(N(t) \geq 14)$ with the true parameters $\rho_0 = 1$ and $\mu_0 = 27/1734$ plugged in. The dashed cyan line indicate the mean of the estimated probability post burn-in.

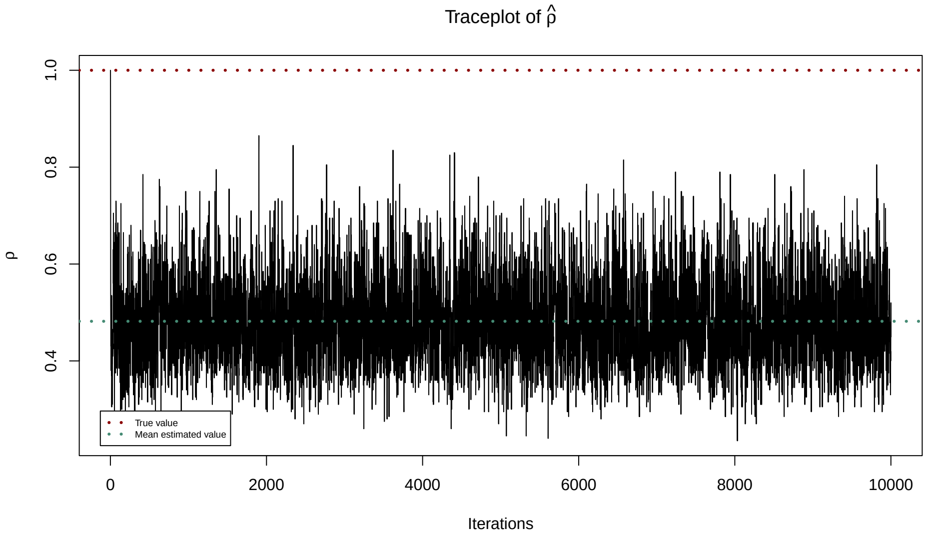


(a) Histogram ρ

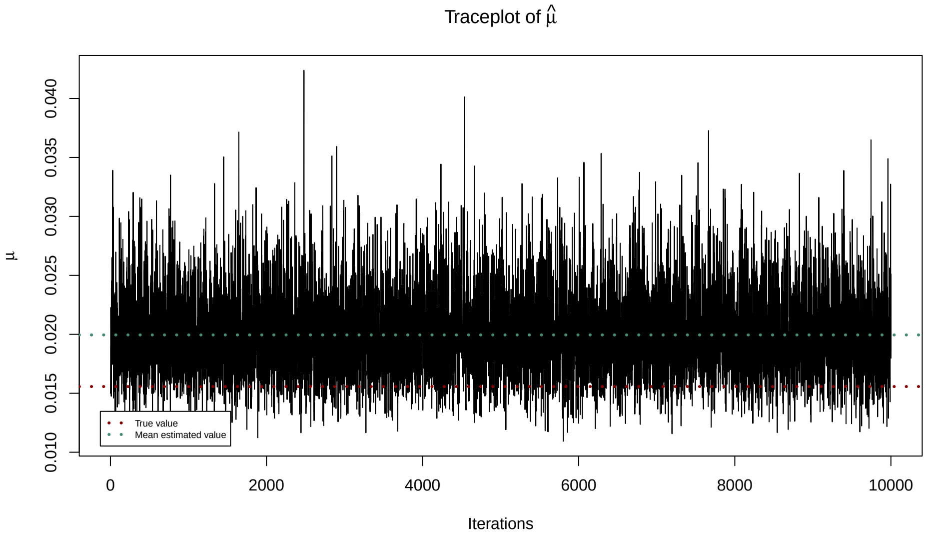


(b) Histogram μ

Figure C.8: Histograms of the posterior densities post burn-in of the parameters ρ and μ using the Gibbs sampler. We used the number of incidents of $n = 81$ nurses, for $t = 201$ shifts. The dark red lines indicate the corresponding density functions.

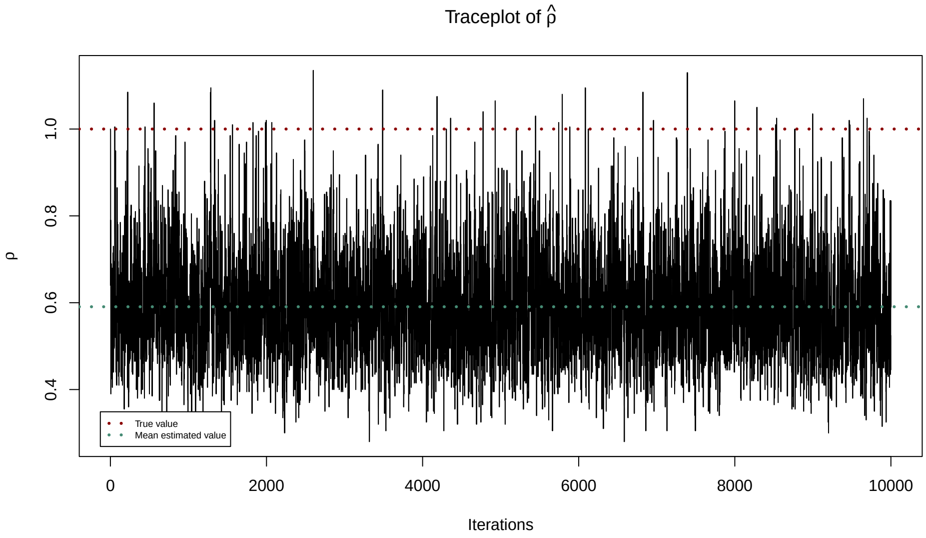


(a) Traceplot ρ

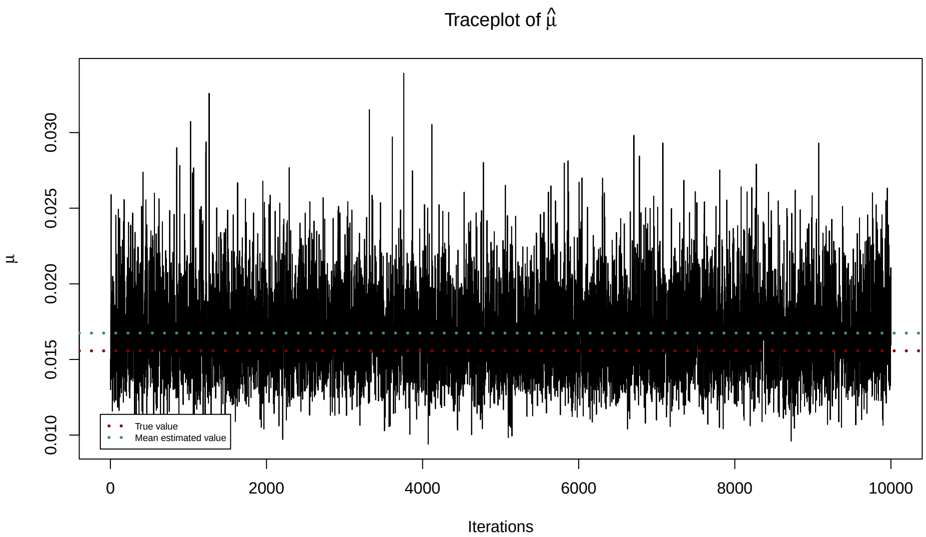


(b) Traceplot μ

Figure C.9: Traceplots of the estimated parameters ρ and μ using the Gibbs sampler. We used the number of incidents of $n = 81$ nurses, for $t = 201$ shifts, where we replaced the first data point with an erroneous value equal to $x_1 = 100$. The dashed dark red lines indicate the true values $\rho_0 = 1$ and $\mu_0 = 27/1734$ of the parameters. The dashed cyan lines indicate the mean of the estimated parameters post burn-in.



(a) Traceplot ρ



(b) Traceplot μ

Figure C.10: Traceplots of the estimated parameters ρ and μ using the Gibbs sampler. We used the number of incidents of $n = 81$ nurses, for $t = 201$ shifts, where we replaced the first data point with an erroneous value equal to $x_1 = 50$. The dashed dark red lines indicate the true values $\rho_0 = 1$ and $\mu_0 = 27/1734$ of the parameters. The dashed cyan lines indicate the mean of the estimated parameters post burn-in.

C.4.2 Method of moments

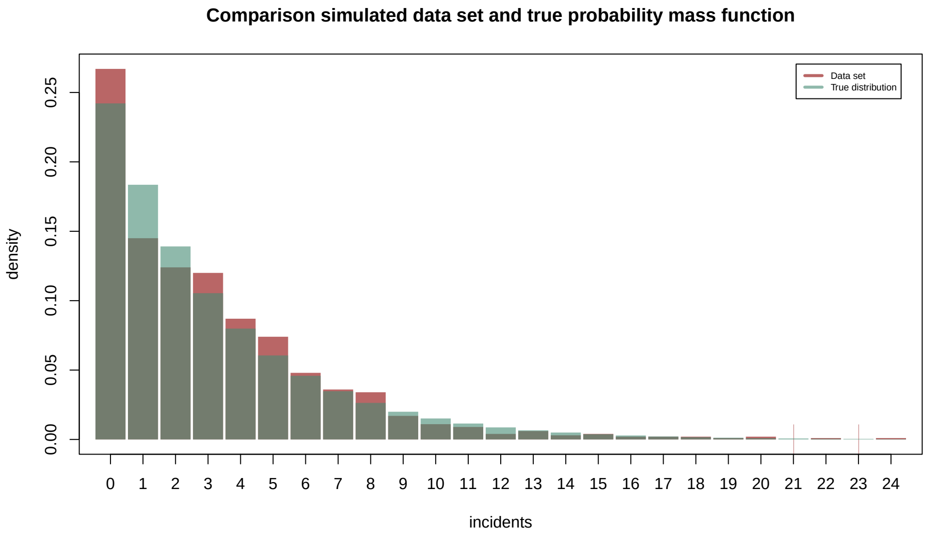
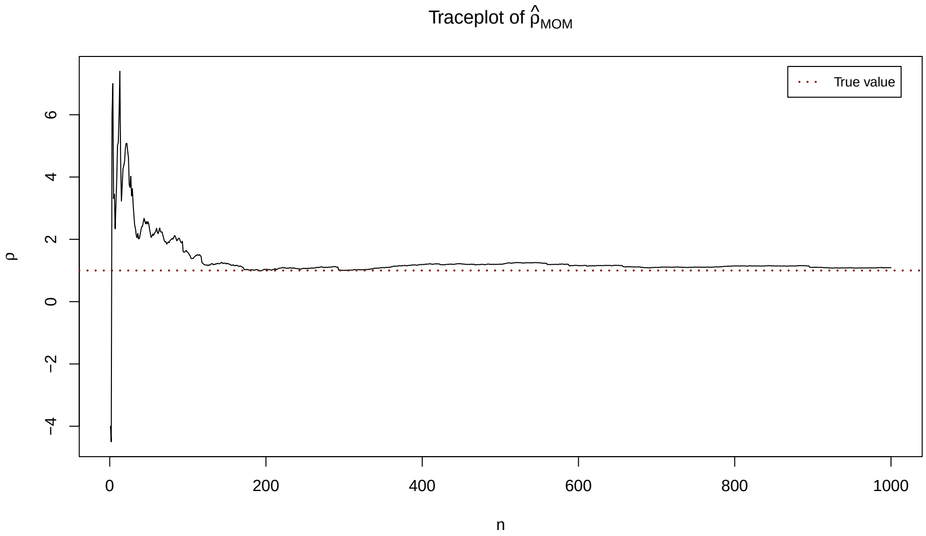
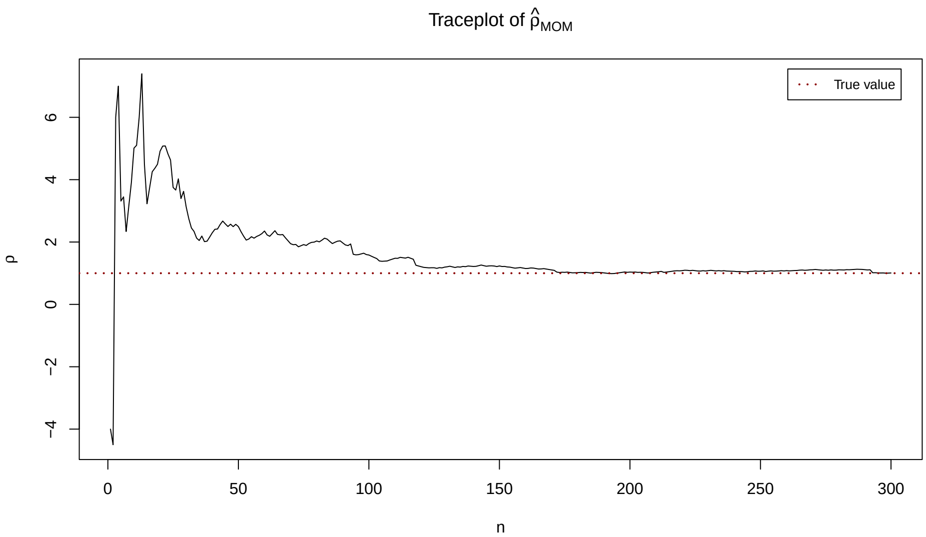


Figure C.11: Comparison of probabilities of simulated incidents of 1000 nurses with true probability distribution for the method of moments.

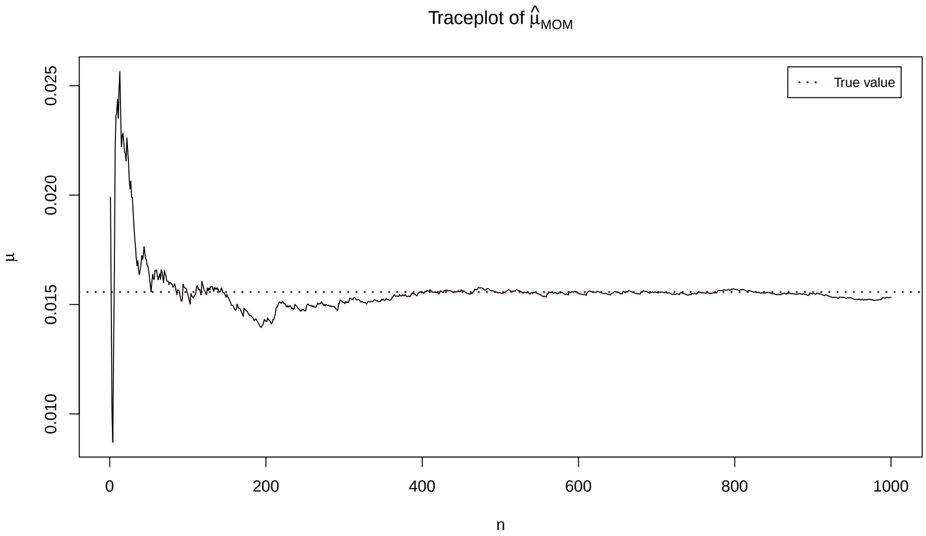


(a) Traceplot of maximum $n = 1000$ nurses

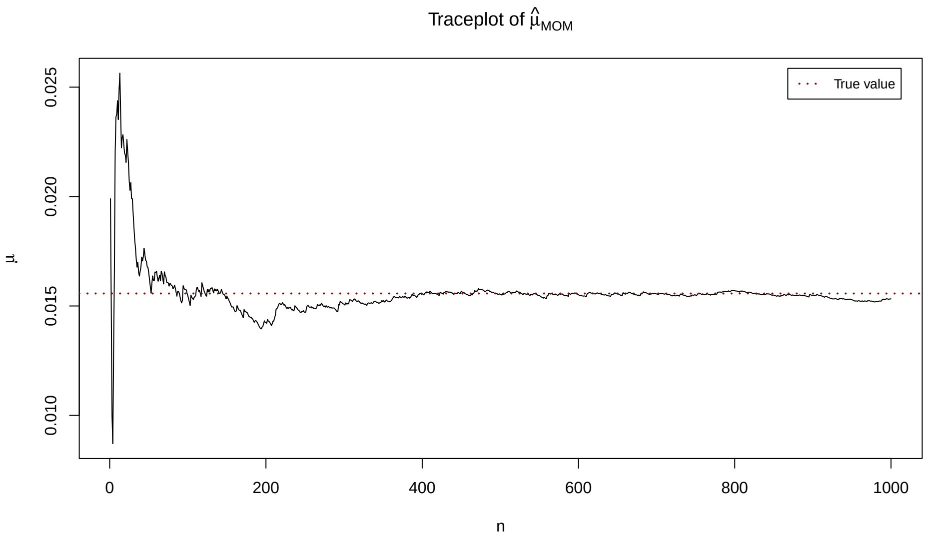


(b) Closeup of maximum $n = 300$ nurses

Figure C.12: Traceplot of the moment estimator $\hat{\rho}_{MOM}$ of the parameter ρ using the number of incidents of n nurses, for $t = 201$ shifts. The dashed dark red lines indicate the true value of the parameter ρ, which is $\rho_0 = 1$.

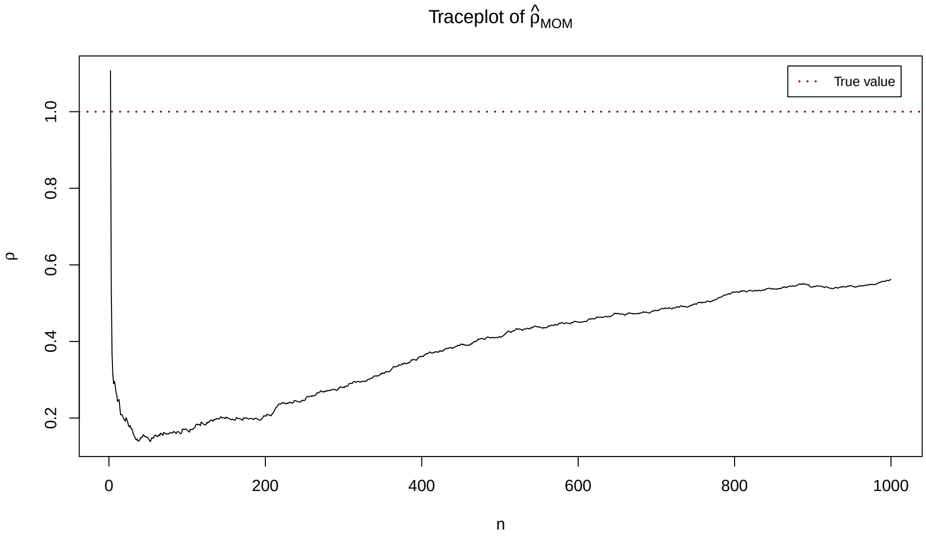


(a) Trace plot of maximum $n = 1000$ nurses

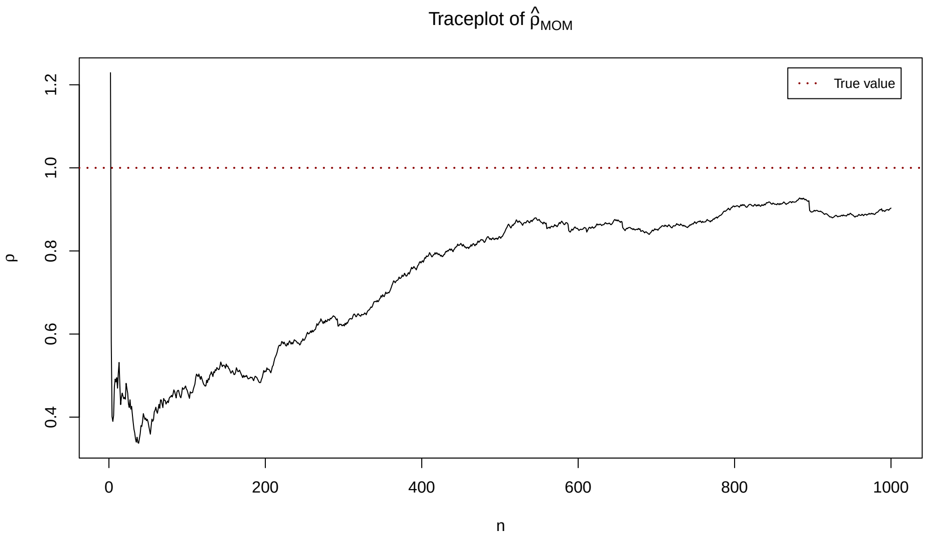


(b) Closeup of maximum $n = 300$ nurses

Figure C.13: Traceplot of the moment estimator $\hat{\mu}_{MOM}$ of the parameter μ using the number of incidents that n nurses experienced in $t = 201$ shifts. The dashed dark red lines indicate the true value of the parameter μ, which is $\mu_0 = 27/1734$.

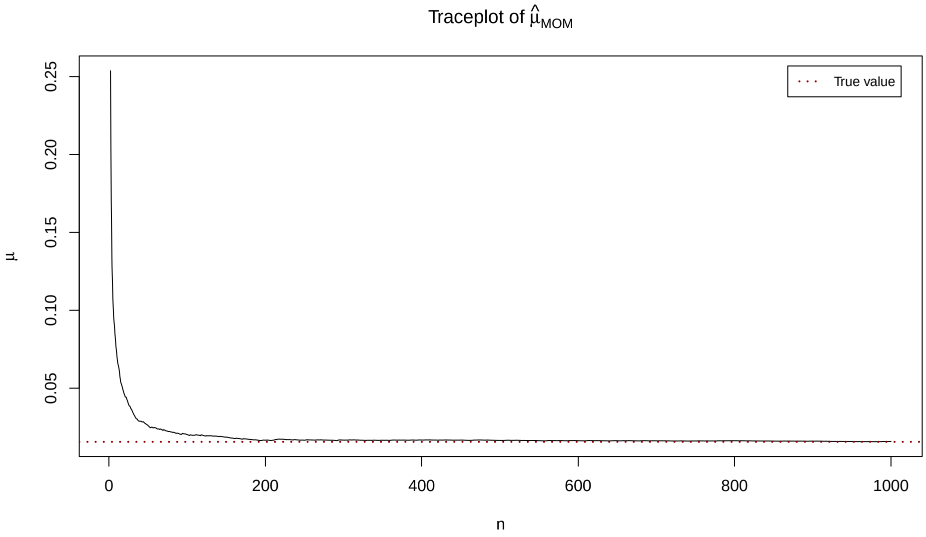


(a) $x_1 = 100$

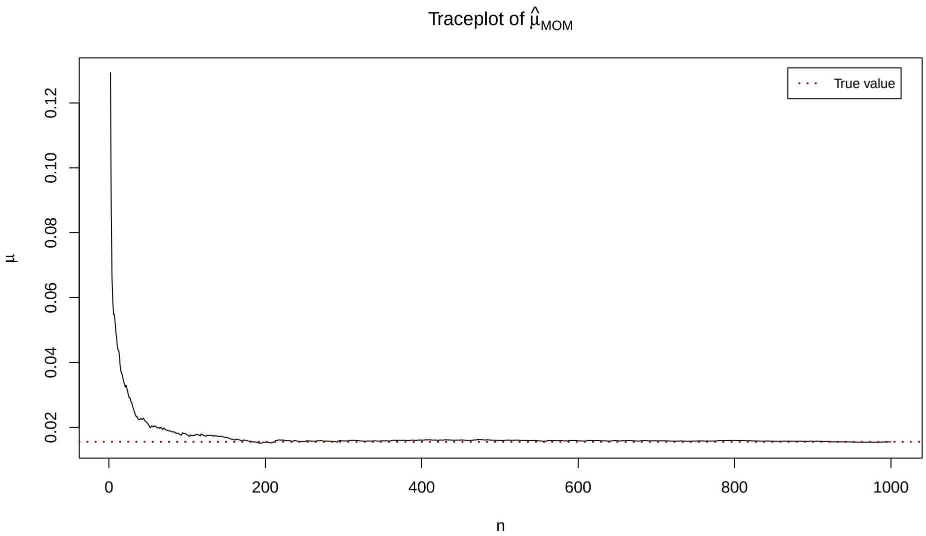


(b) $x_1 = 50$

Figure C.14: Traceplot of the moment estimator $\hat{\rho}_{MOM}$ of the parameter ρ using the number of incidents of n nurses, for $t = 201$ shifts, where we replaced the first data point with an erroneous value equal to (a) $x_1 = 100$ and (b) $x_1 = 50$. The dashed dark red lines indicate the true value of the parameter ρ, which is $\rho_0 = 1$.

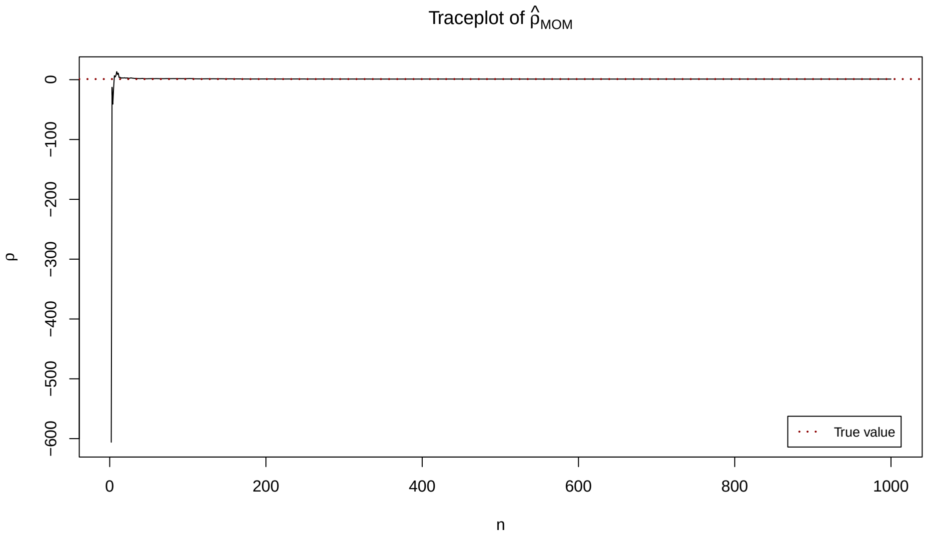


(a) $x_1 = 100$

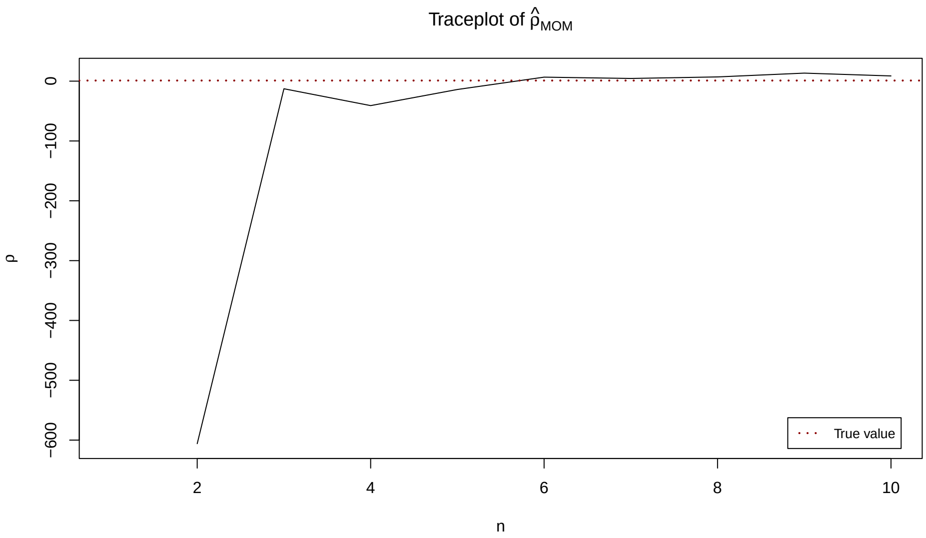


(b) $x_1 = 50$

Figure C.15: Traceplot of the moment estimator $\hat{\mu}_{MOM}$ of the parameter μ using the number of incidents of n nurses, for $t = 201$ shifts, where we replaced the first data point with an erroneous value equal to (a) $x_1 = 100$ and (b) $x_1 = 50$. The dashed dark red lines indicate the true value of the parameter μ, which is $\mu_0 = 27/1734$.



(a) Traceplot of maximum $n = 1000$ nurses



(b) Closeup of maximum $n = 10$ nurses

Figure C.16: Traceplot of the moment estimator $\hat{\rho}_{MOM}$ of the parameter ρ using the number of incidents of n nurses, where 5% of the nurses works $t = 15$ shifts, 80% works $t = 150$ shifts and 15% works $t = 200$ shifts. The dashed dark red lines indicate the true value of the parameter ρ, which is $\rho_0 = 1$.

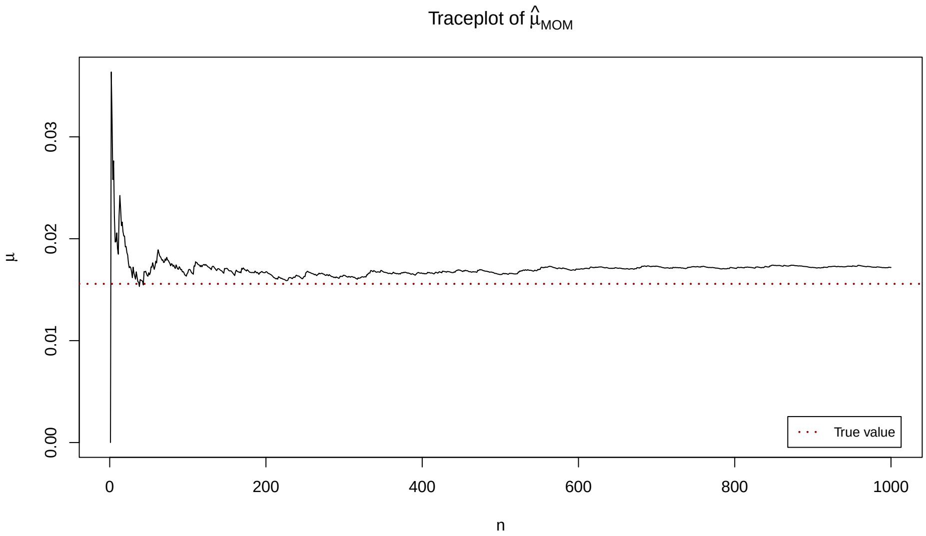


Figure C.17: Traceplot of the moment estimator $\hat{\mu}_{MOM}$ of the parameter μ using the number of incidents of n nurses, where 5% of the nurses works $t = 15$ shifts, 80% works $t = 150$ shifts and 15% works $t = 200$ shifts. The dashed dark red lines indicate the true value of the parameter μ, which is $\mu_0 = 27/1734$.

C.4.3 Hamiltonian Monte Carlo

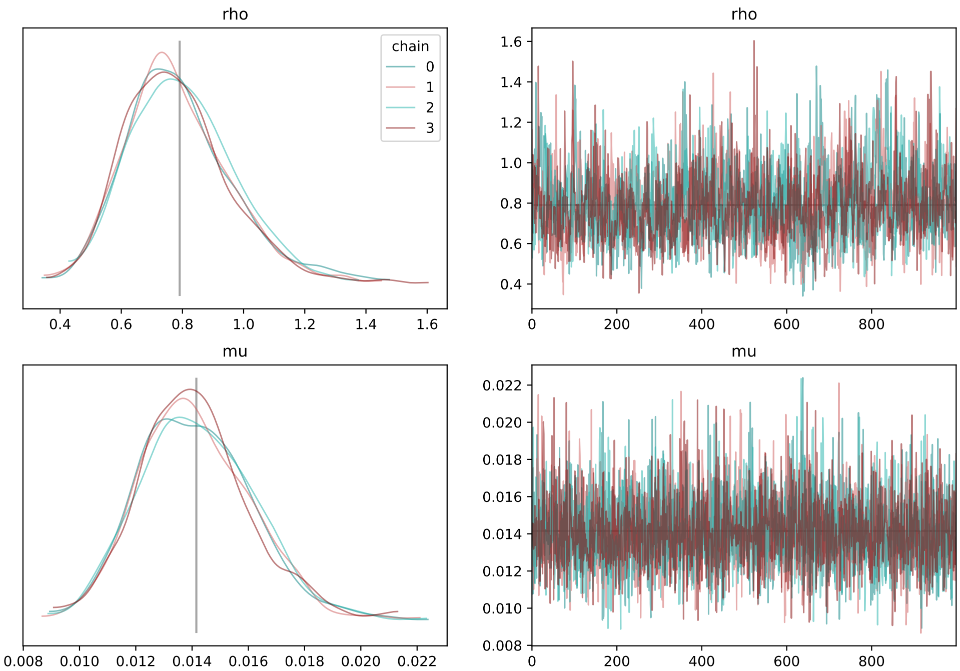
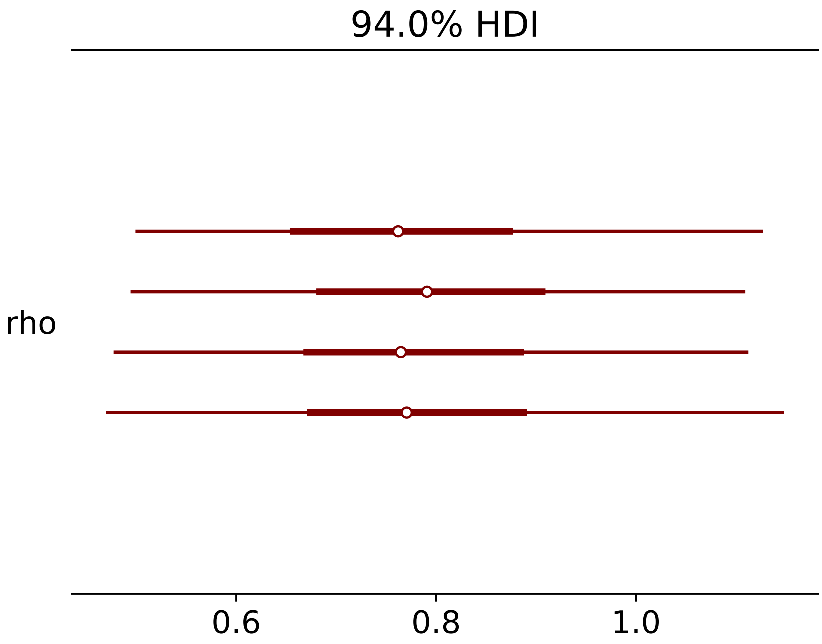
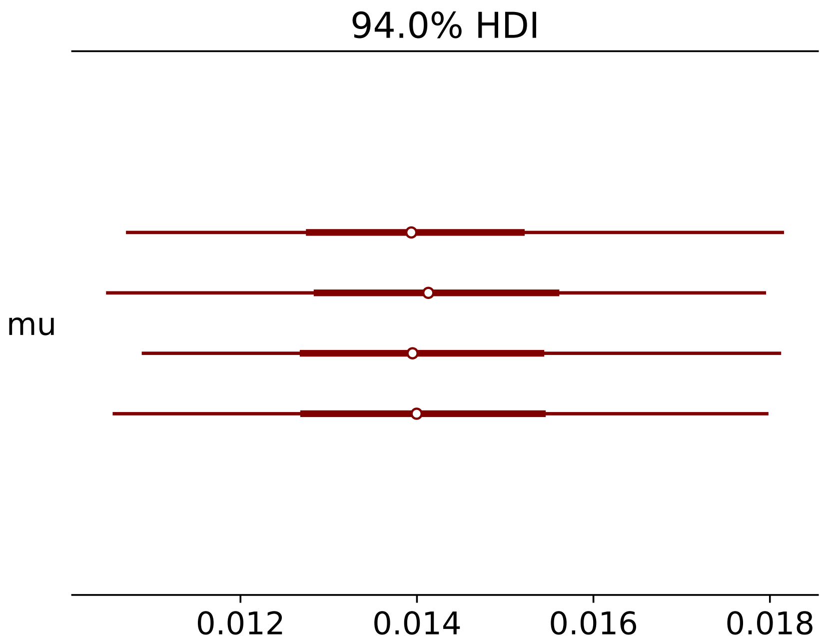


Figure C.18: Posterior densities and traceplots of the parameters ρ and μ using the No-U-Turn sampler in PyMC3. We used the number of incidents of $n = 81$ nurses, for $t = 201$ shifts. Each color represents one chain. The vertical and horizontal lines are the posterior means of the parameters. The true values of the parameters are $\rho_0 = 1$ and $\mu_0 = 27/1734$.



(a) Forest plot ρ



(b) Forest plot μ

Figure C.19: Forest plots of the estimation of the parameters ρ and μ using the No-U-Turn sampler in PyMC3. We used the number of incidents of $n = 81$ nurses, for $t = 201$ shifts. Each 94% highest posterior density interval (HDI) represents one chain. The true values of the parameters are $\rho_0 = 1$ and $\mu_0 = 27/1734$.

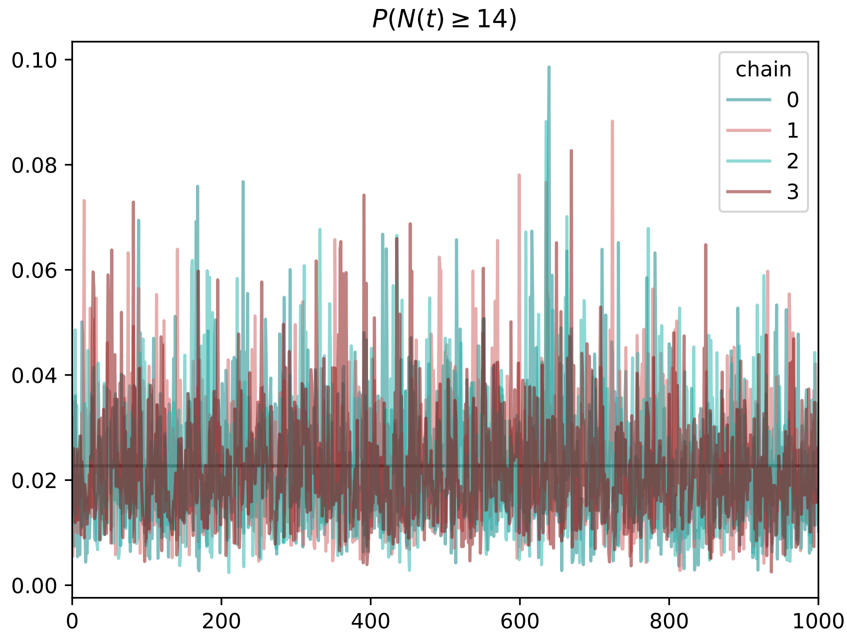


Figure C.20: Traceplot of the estimation of the probability $P(N(t) \geq 14)$ using the No-U-Turn sampler in PyMC3. We used the number of incidents of $n = 81$ nurses, for $t = 201$ shifts. Each colour represents one chain. The horizontal line is the posterior means of the probability.

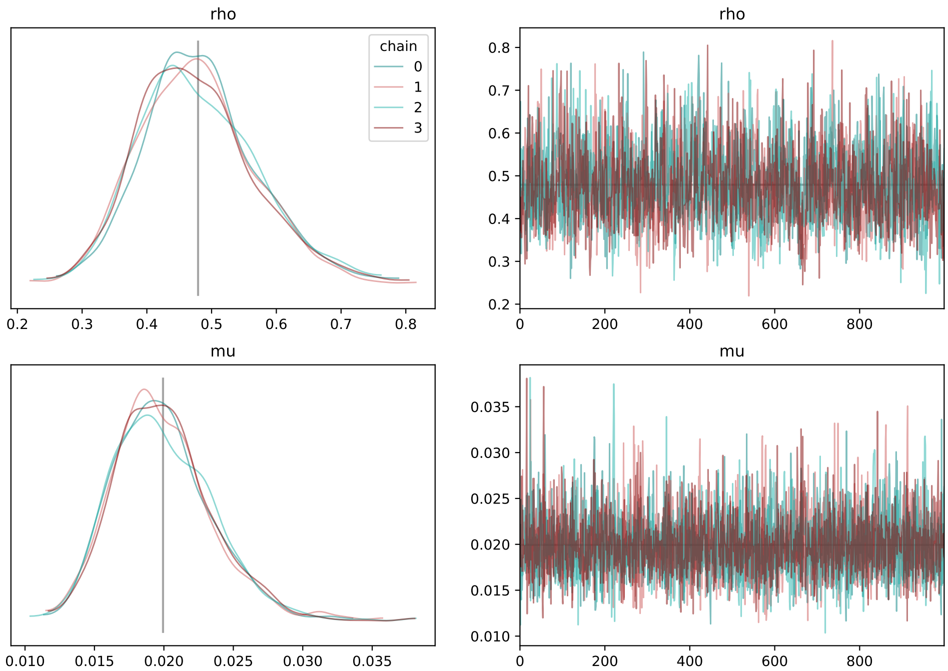
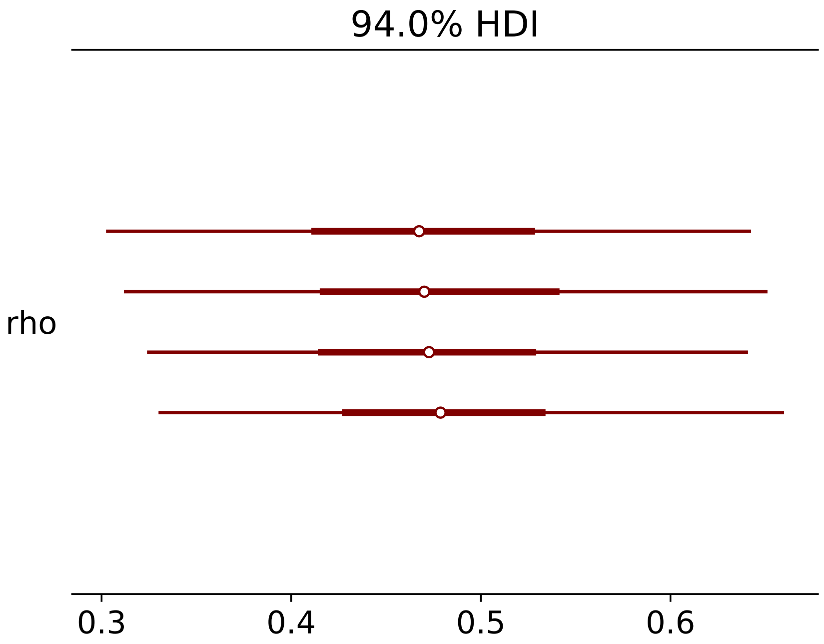
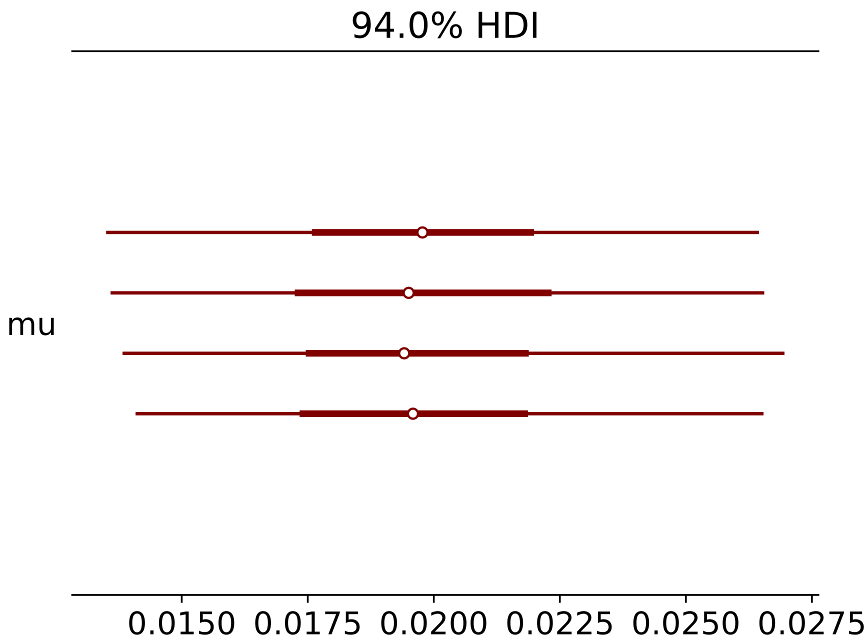


Figure C.21: Posterior densities and traceplots of the parameters ρ and μ using the No-U-Turn sampler in PyMC3. We used the number of incidents of $n = 81$ nurses, for $t = 201$ shifts, where we replaced the first data point with an erroneous value equal to $x_1 = 100$. Each colour represents one chain. The vertical and horizontal lines are the posterior means of the parameters. The true values of the parameters are $\rho_0 = 1$ and $\mu_0 = 27/1734$.



(a) Forest plot ρ



(b) Forest plot μ

Figure C.22: Forest plots of the estimation of the parameters ρ and μ using the No-U-Turn sampler in PyMC3. We used the number of incidents of $n = 81$ nurses, for $t = 201$ shifts, where we replaced the first data point with an erroneous value equal to $x_1 = 100$. Each 94% highest posterior density interval (HDI) represents one chain. The true values of the parameters are $\rho_0 = 1$ and $\mu_0 = 27/1734$.

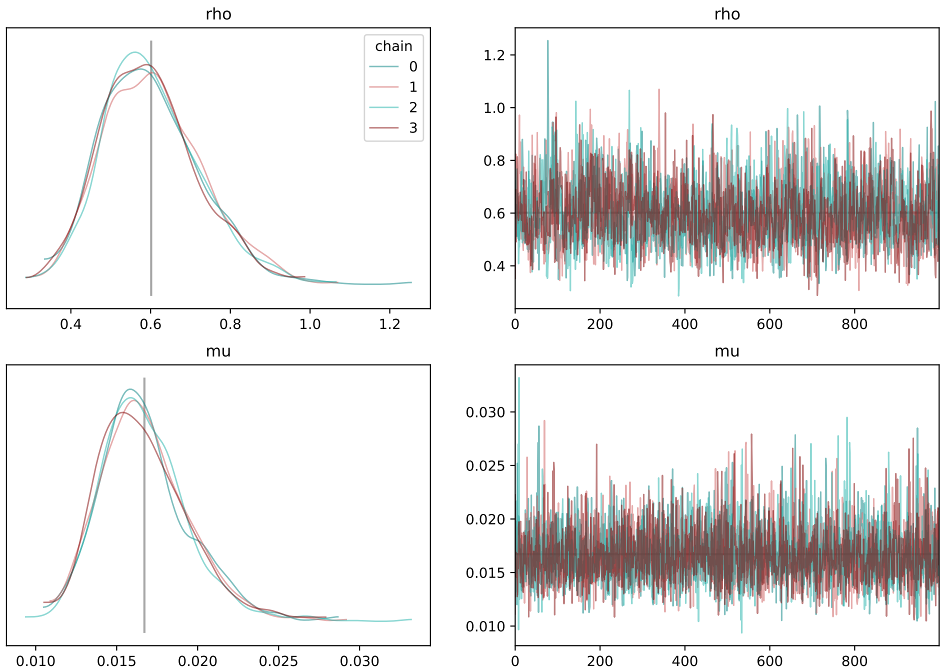
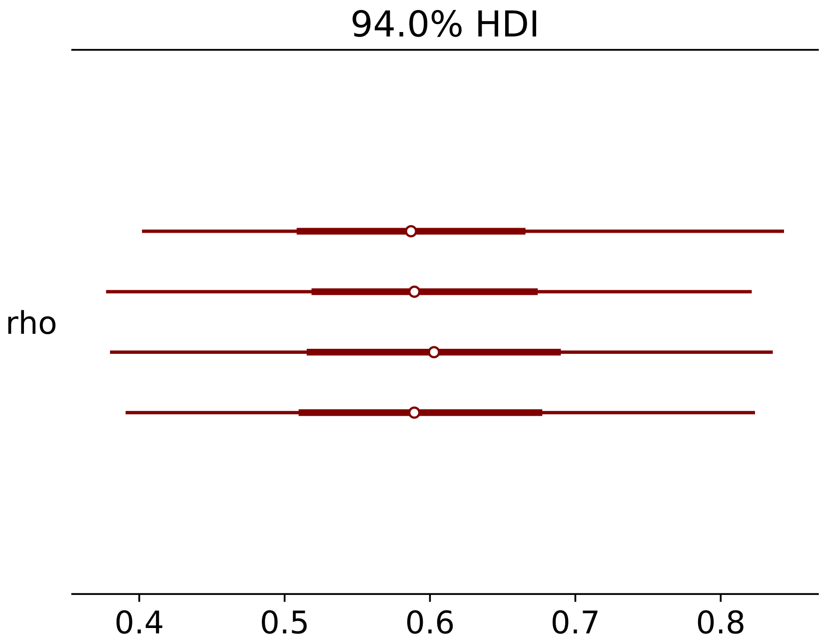
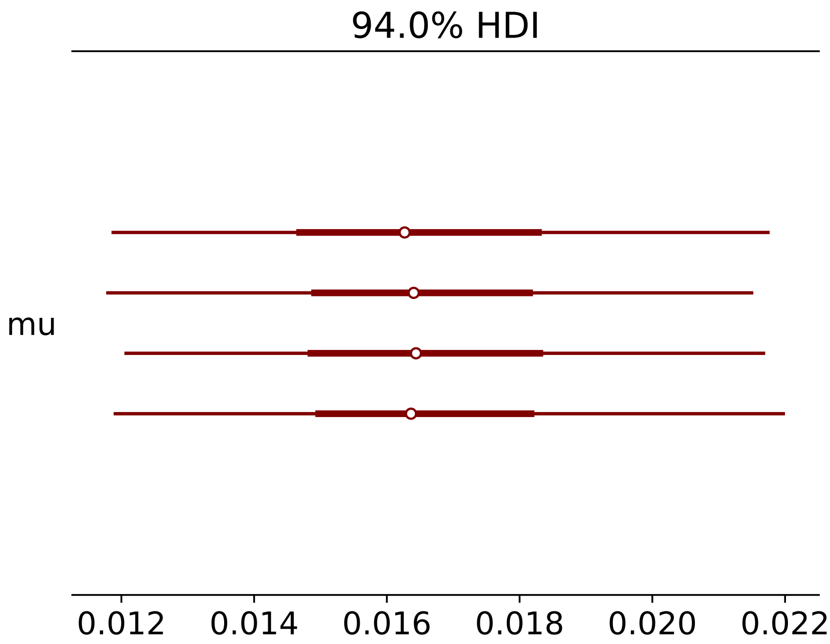


Figure C.23: Posterior densities and traceplots of the parameters ρ and μ using the No-U-Turn sampler in PyMC3. We used the number of incidents of $n = 81$ nurses, for $t = 201$ shifts, where we replaced the first data point with an erroneous value equal to $x_1 = 50$. Each colour represents one chain. The vertical and horizontal lines are the posterior means of the parameters. The true values of the parameters are $\rho_0 = 1$ and $\mu_0 = 27/1734$.



(a) Forest plot ρ

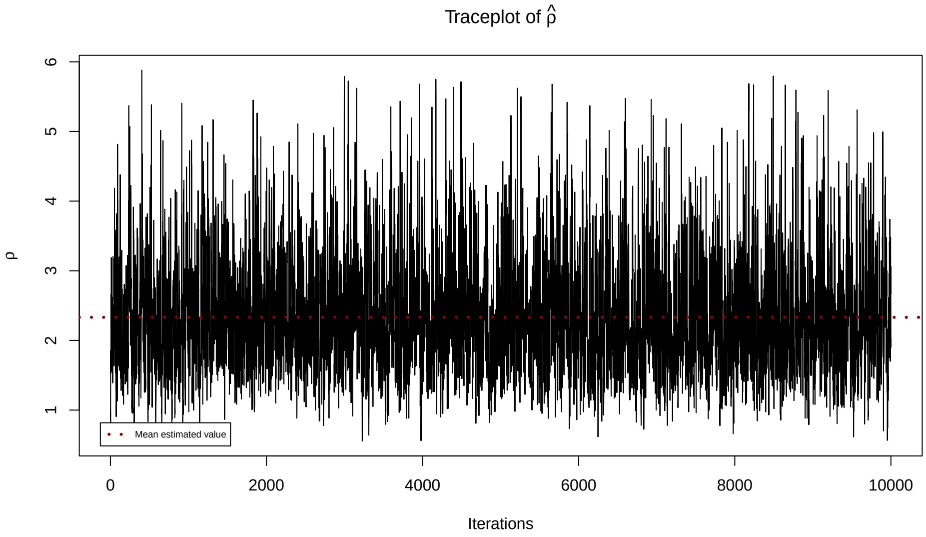


(b) Forest plot μ

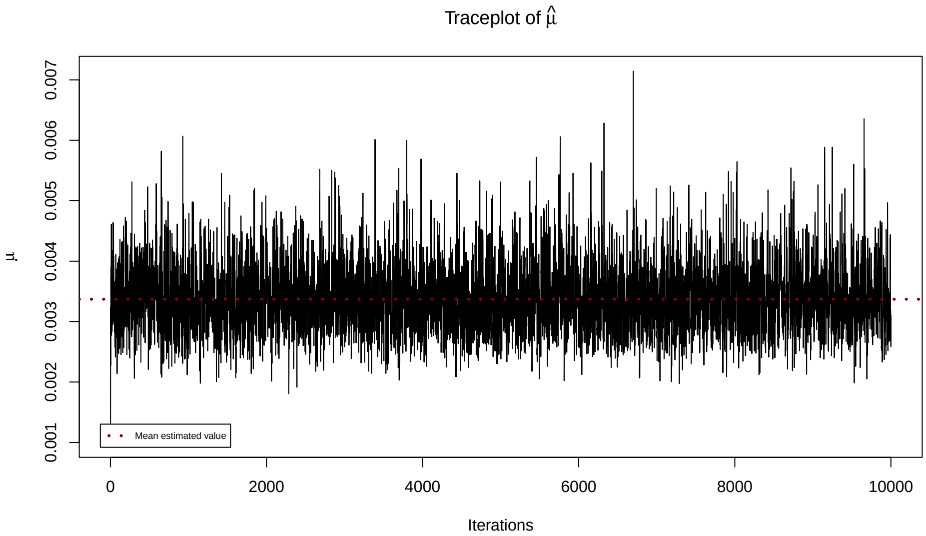
Figure C.24: Forest plots of the estimation of the parameters ρ and μ using the No-U-Turn sampler in PyMC3. We used the number of incidents of $n = 81$ nurses, for $t = 201$ shifts, where we replaced the first data point with an erroneous value equal to $x_1 = 50$. Each 94% highest posterior density interval (HDI) represents one chain. The true values of the parameters are $\rho_0 = 1$ and $\mu_0 = 27/1734$.

C.4.4 Application: ship damage incidents

Gibbs sampling algorithm

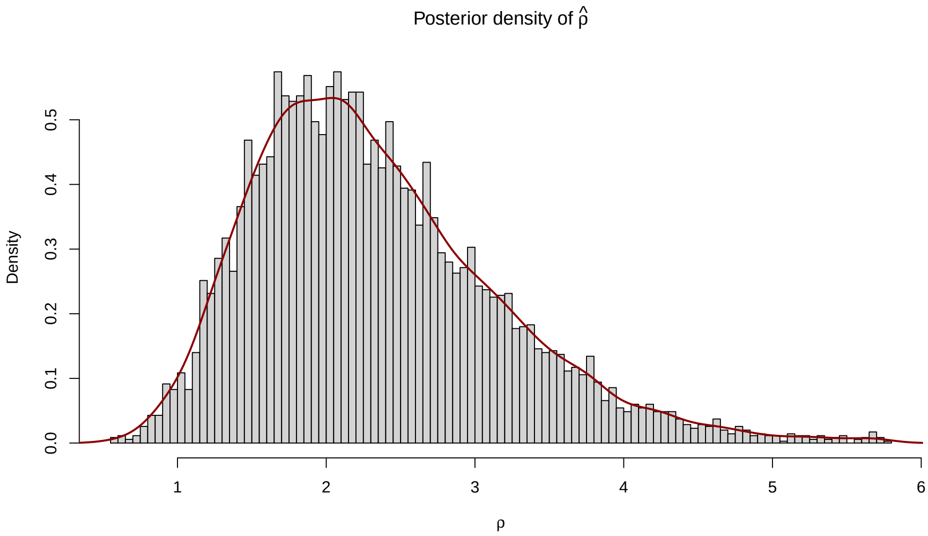


(a) Traceplot ρ

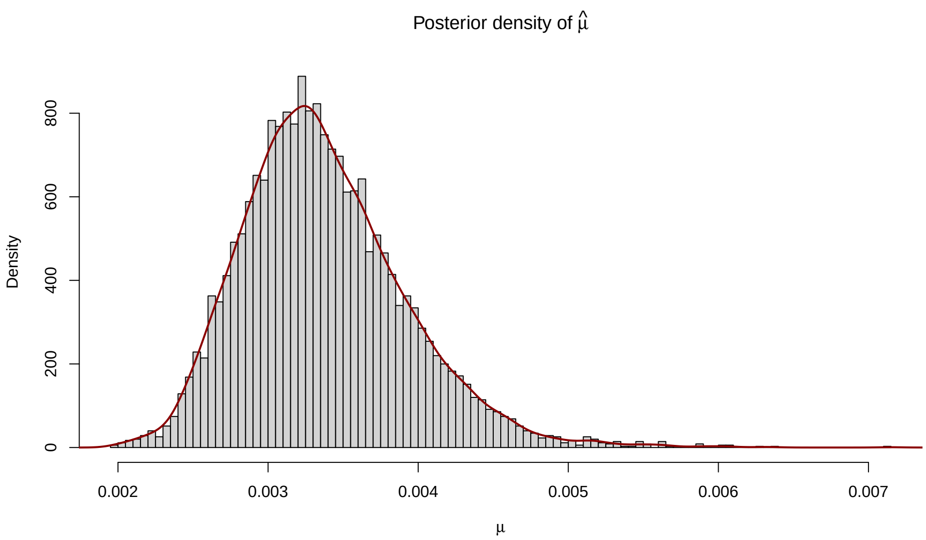


(b) Traceplot μ

Figure C.25: Traceplots of the estimated parameters ρ and μ using the Gibbs sampler. We used the data from a study about wave damage to cargo ships taken from (McCullagh and Nelder, 1989, Subsection 6.3.2). The dashed dark red lines indicate the mean of the estimated parameters post burn-in.



(a) Histogram ρ



(b) Histogram μ

Figure C.26: Histograms of the posterior densities post burn-in of the parameters ρ and μ using the Gibbs sampler. We used the data from a study about wave damage to cargo ships taken from (McCullagh and Nelder, 1989, Subsection 6.3.2). The dark red lines indicate the corresponding density functions.

Hamiltonian Monte Carlo

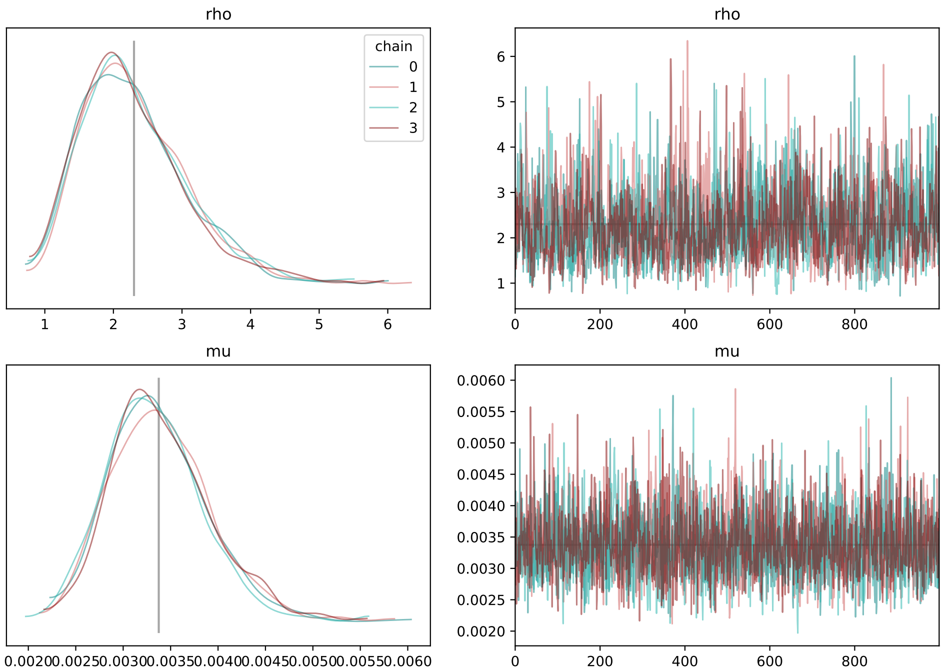
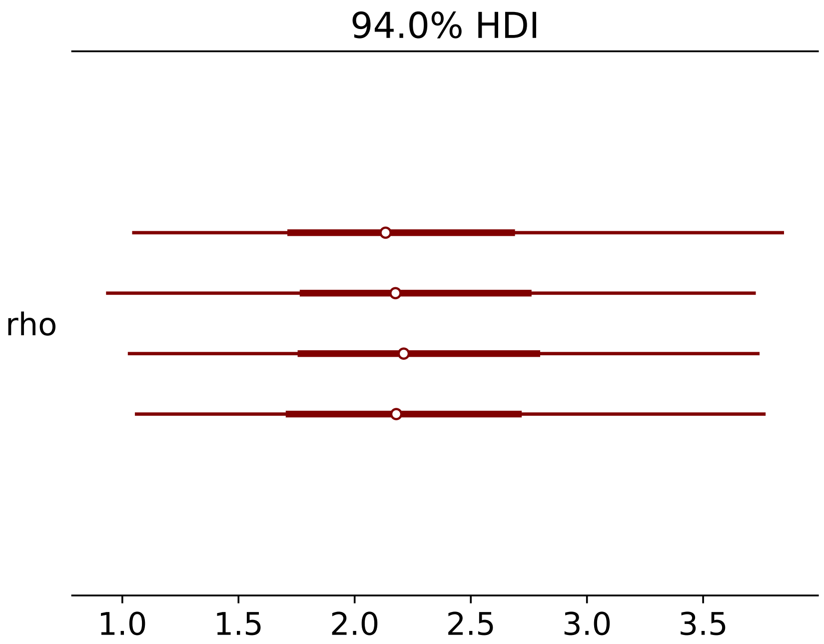
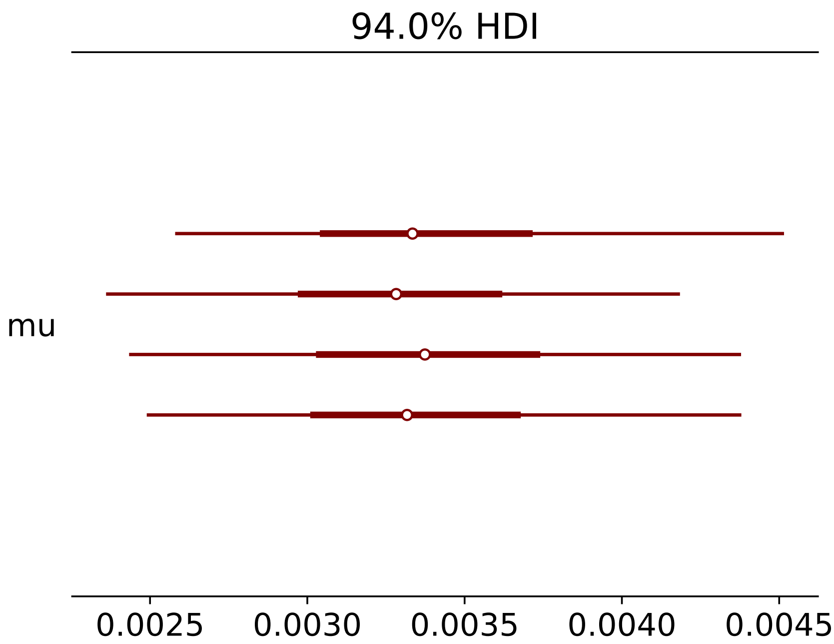


Figure C.27: Posterior densities and traceplots of the parameters ρ and μ using the No-U-Turn sampler in PyMC3. We used the data from a study about wave damage to cargo ships taken from (McCullagh and Nelder, 1989). Each colour represents one chain. The vertical and horizontal lines are the posterior means of the parameters.



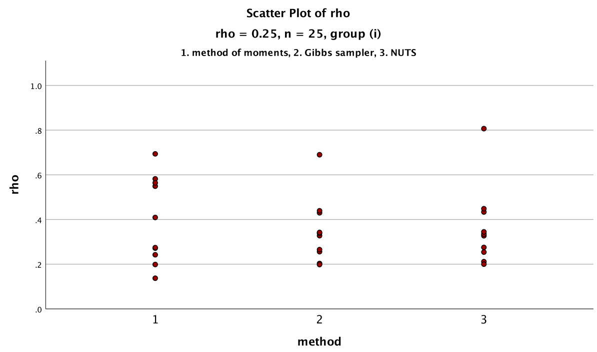
(a) Forest plot ρ



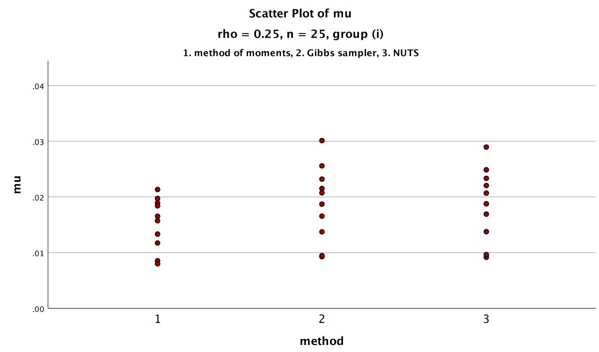
(b) Forest plot μ

Figure C.28: Forest plots of the estimation of the parameters ρ and μ using the No-U-Turn sampler in PyMC3. We used the data from a study about wave damage to cargo ships taken from (McCullagh and Nelder, 1989, Subsection 6.3.2). Each 94% highest posterior density interval (HDI) represents one chain.

C.5 Simulation study

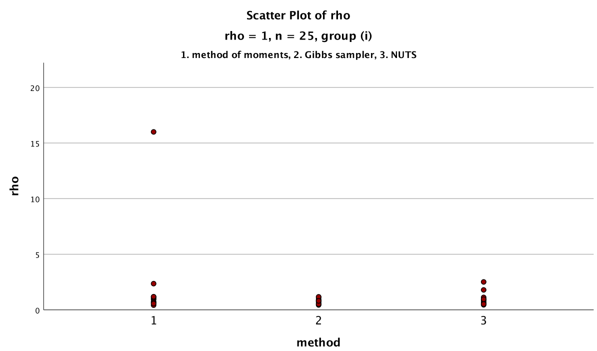


(a) Parameter ρ



(b) Parameter μ

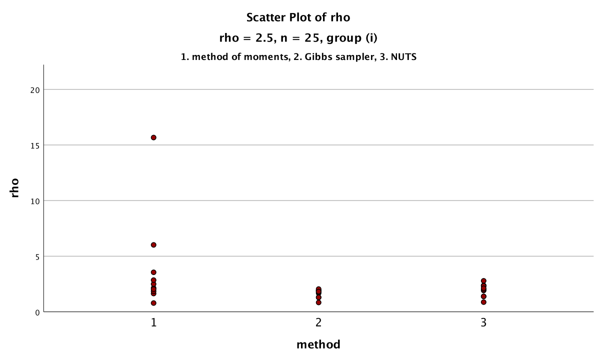
Figure C.29: Scatter plots of the estimated parameters ρ and μ of the mixed Poisson model by the different estimation methods, where $\rho_0 = 0.25, n = 25$, group composition (i).



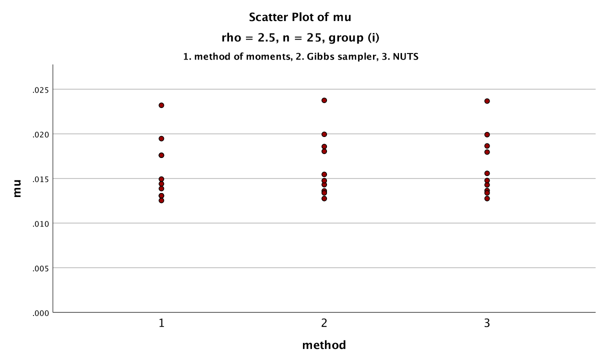
(a) Parameter ρ

(b) Parameter μ

Figure C.30: Scatter plots of the estimated parameters ρ and μ of the mixed Poisson model by the different estimation methods, where $\rho_0 = 1, n = 25$, group composition (i).

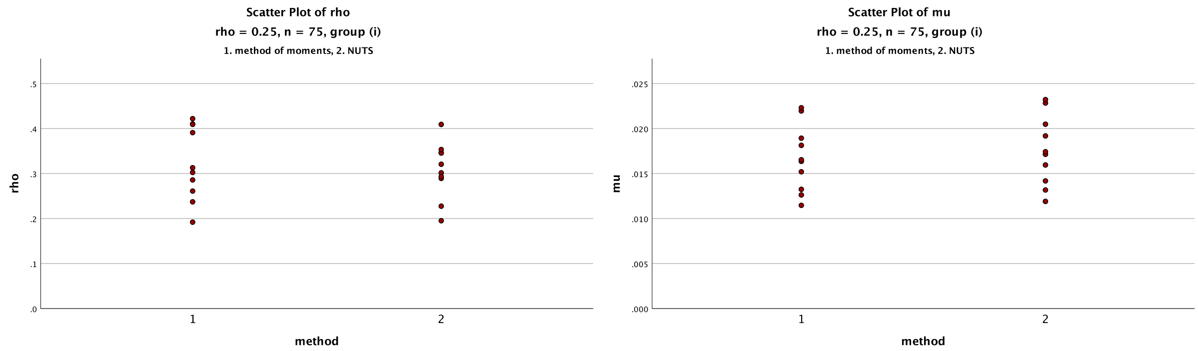


(a) Parameter ρ



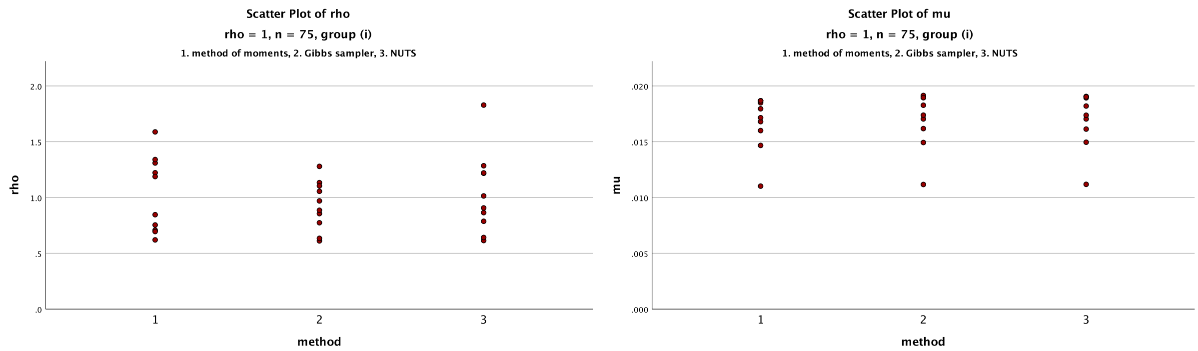
(b) Parameter μ

Figure C.31: Scatter plots of the estimated parameters ρ and μ of the mixed Poisson model by the different estimation methods, where $\rho_0 = 2.5, n = 25$, group composition (i).



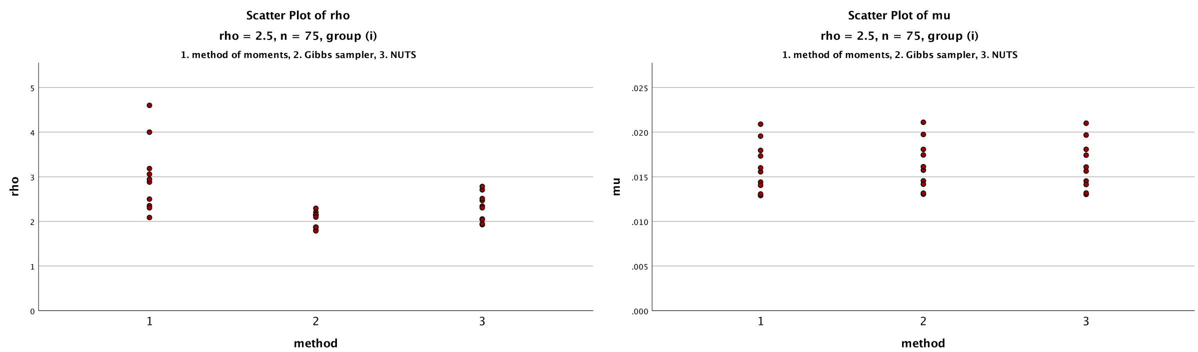
(a) Parameter ρ (b) Parameter μ

Figure C.32: Scatter plots of the estimated parameters ρ and μ of the mixed Poisson model by the different estimation methods, where $\rho_0 = 0.25, n = 75$, group composition (i).



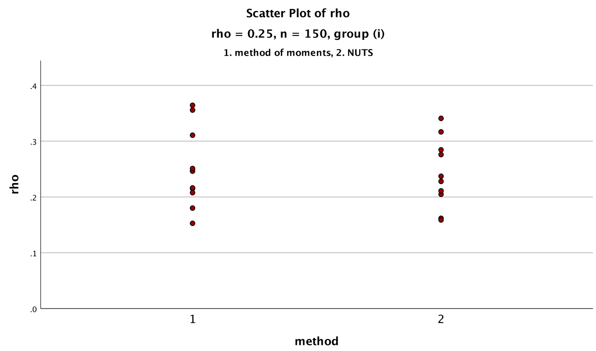
(a) Parameter ρ (b) Parameter μ

Figure C.33: Scatter plots of the estimated parameters ρ and μ of the mixed Poisson model by the different estimation methods, where $\rho_0 = 1, n = 75$, group composition (i).

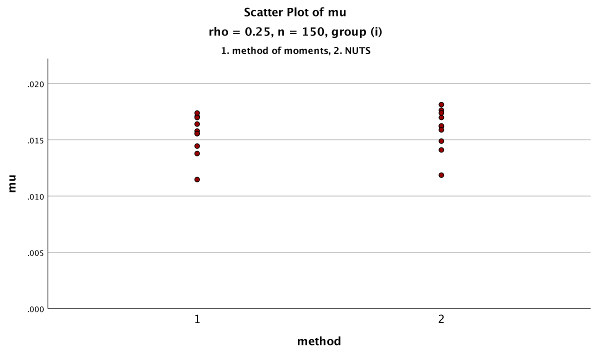


(a) Parameter ρ (b) Parameter μ

Figure C.34: Scatter plots of the estimated parameters ρ and μ of the mixed Poisson model by the different estimation methods, where $\rho_0 = 2.5, n = 75$, group composition (i).

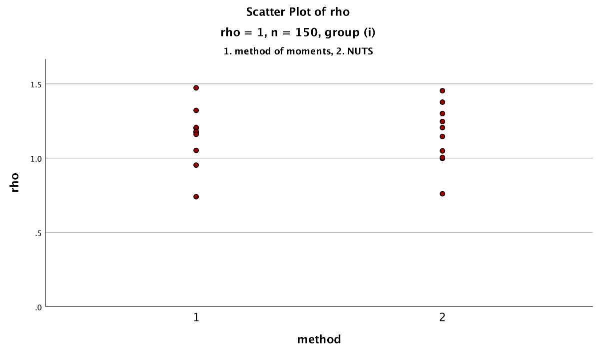


(a) Parameter ρ

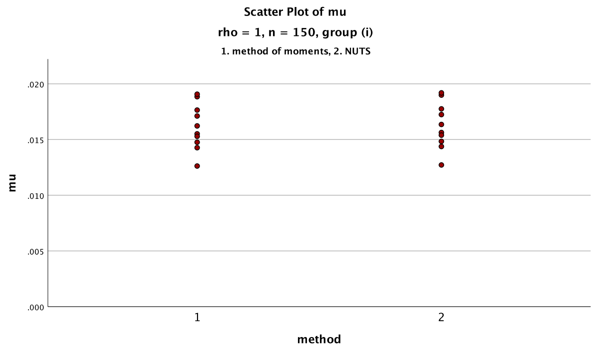


(b) Parameter μ

Figure C.35: Scatter plots of the estimated parameters ρ and μ of the mixed Poisson model by the different estimation methods, where $\rho_0 = 0.25, n = 150$, group composition (i).

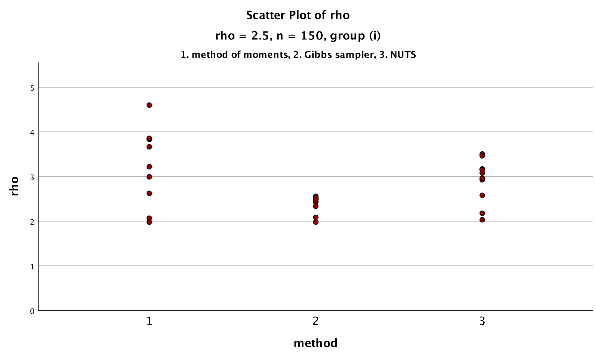


(a) Parameter ρ

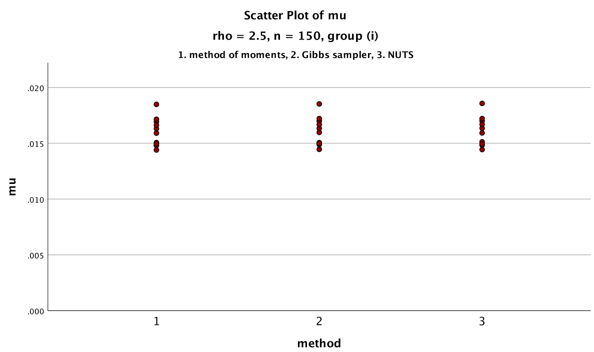


(b) Parameter μ

Figure C.36: Scatter plots of the estimated parameters ρ and μ of the mixed Poisson model by the different estimation methods, where $\rho_0 = 1, n = 150$, group composition (i).

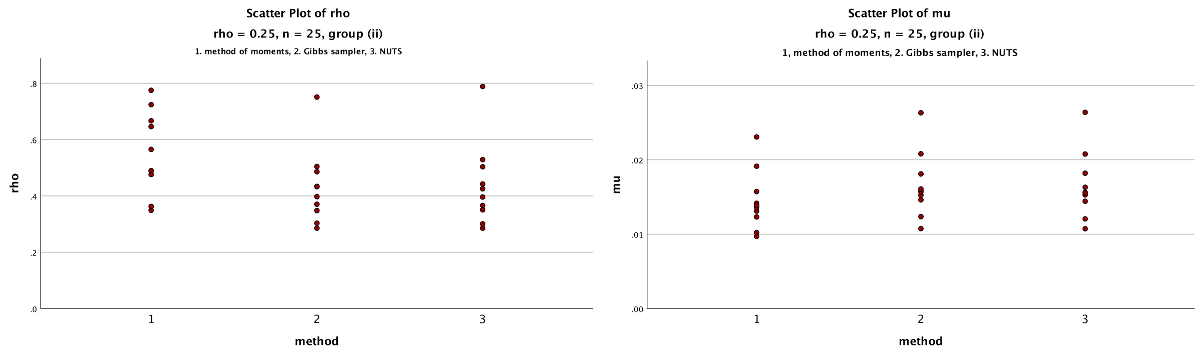


(a) Parameter ρ



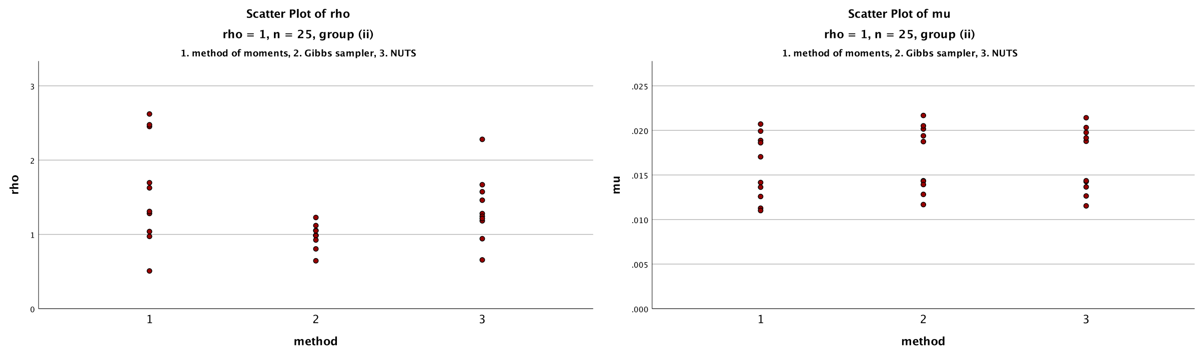
(b) Parameter μ

Figure C.37: Scatter plots of the estimated parameters ρ and μ of the mixed Poisson model by the different estimation methods, where $\rho_0 = 2.5, n = 150$, group composition (i).



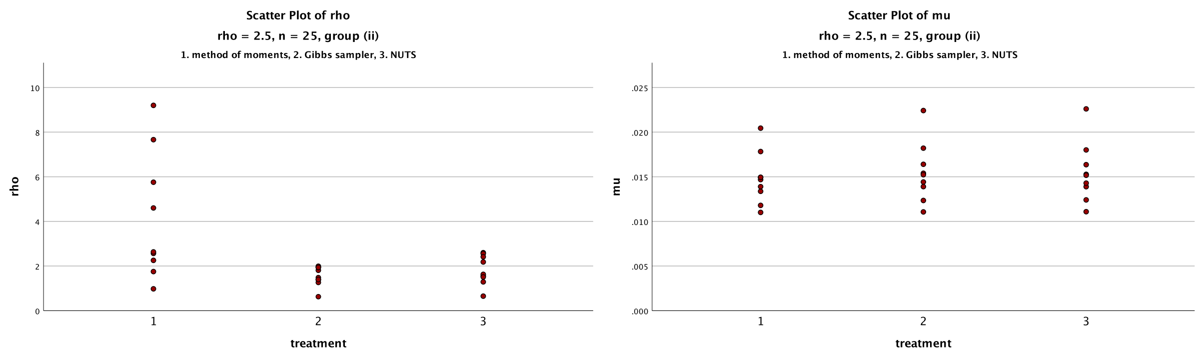
(a) Parameter ρ (b) Parameter μ

Figure C.38: Scatter plots of the estimated parameters ρ and μ of the mixed Poisson model by the different estimation methods, where $\rho_0 = 0.25, n = 25$, group composition (ii).



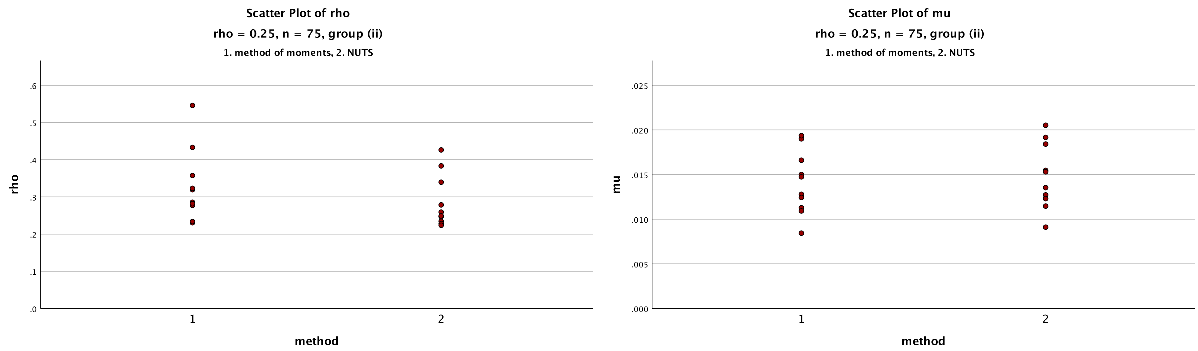
(a) Parameter ρ (b) Parameter μ

Figure C.39: Scatter plots of the estimated parameters ρ and μ of the mixed Poisson model by the different estimation methods, where $\rho_0 = 1, n = 25$, group composition (ii).



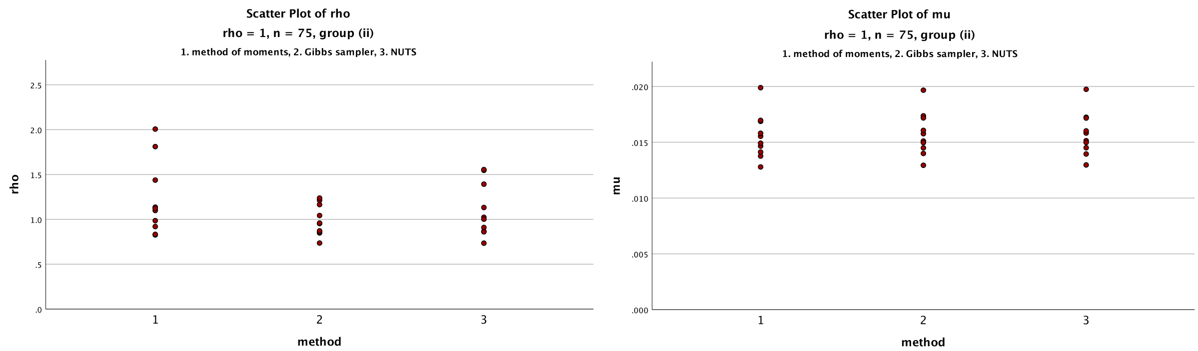
(a) Parameter ρ (b) Parameter μ

Figure C.40: Scatter plots of the estimated parameters ρ and μ of the mixed Poisson model by the different estimation methods, where $\rho_0 = 2.5, n = 25$, group composition (ii).



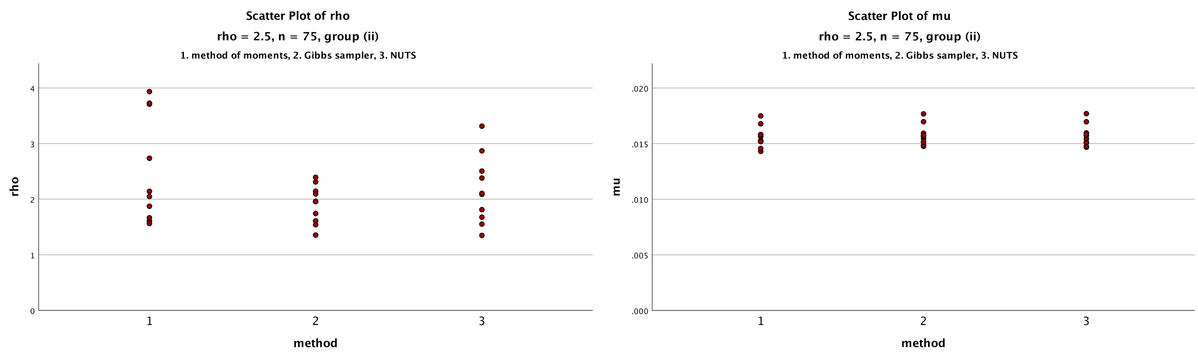
(a) Parameter ρ (b) Parameter μ

Figure C.41: Scatter plots of the estimated parameters ρ and μ of the mixed Poisson model by the different estimation methods, where $\rho_0 = 0.25, n = 75$, group composition (ii).



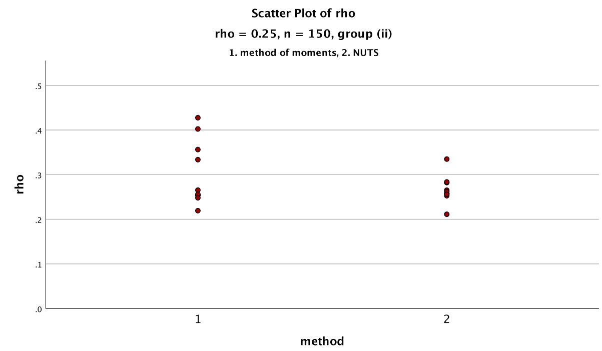
(a) Parameter ρ (b) Parameter μ

Figure C.42: Scatter plots of the estimated parameters ρ and μ of the mixed Poisson model by the different estimation methods, where $\rho_0 = 1, n = 75$, group composition (ii).

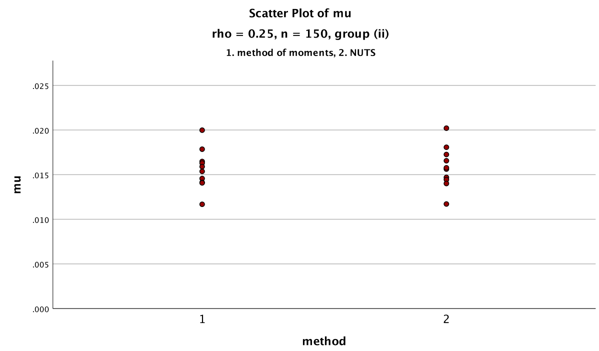


(a) Parameter ρ (b) Parameter μ

Figure C.43: Scatter plots of the estimated parameters ρ and μ of the mixed Poisson model by the different estimation methods, where $\rho_0 = 2.5, n = 75$, group composition (ii).

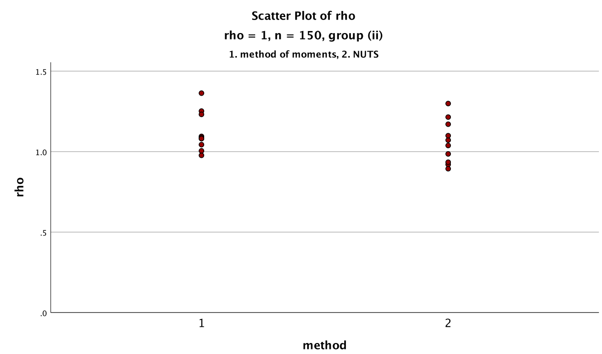


(a) Parameter ρ

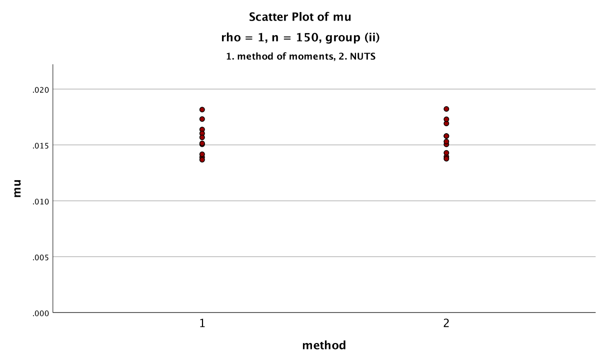


(b) Parameter μ

Figure C.44: Scatter plots of the estimated parameters ρ and μ of the mixed Poisson model by the different estimation methods, where $\rho_0 = 0.25, n = 150$, group composition (ii).

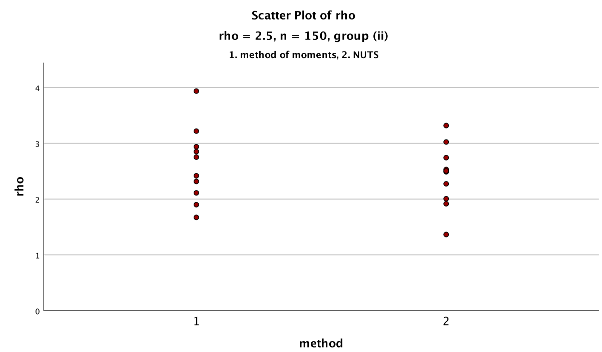


(a) Parameter ρ

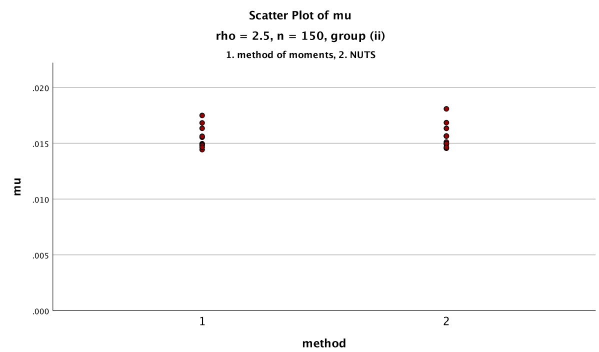


(b) Parameter μ

Figure C.45: Scatter plots of the estimated parameters ρ and μ of the mixed Poisson model by the different estimation methods, where $\rho_0 = 1, n = 150$, group composition (i).



(a) Parameter ρ



(b) Parameter μ

Figure C.46: Scatter plots of the estimated parameters ρ and μ of the mixed Poisson model by the different estimation methods, where $\rho_0 = 2.5, n = 150$, group composition (ii).

D Listings

The code of the Gibbs sampler and the No-U-Turn sampler from Chapter 11, that are also given below, can be accessed through Github: `https://github.com/PCMDEBRUIN/Incident-series.git`. See *GS_Gill_R_final.R* and *GS_Gill_pymc3_final.py*, respectively. In addition, one can find the Gibbs sampler programmed in Python instead of R (*gibbs.py*), and the Python code on how Figures C.1 and C.3 can be obtained (*plot_systematic.py* and *plot_systematic_poissongamma.py*, respectively).

D.1 Roster cases (revised)

D.1.1 Gibbs sampling algorithm

Listing D.1: Gibbs sampler discussed in Subsection 11.2.1, programmed in R.

```r
library('coda')
library('invgamma')
library('tidyverse')
library('writexl')

T = 10000 # size of Markov chain
m = 3000 # burn in period
n = 27 * 3 # number of nurses
r = 201 # number of shifts (Lucia)

# parameter choice of Gill et al.
mu0 <- 27/1734
rho0 <- 1
N0 <- 1 - pnbinom(13, size = rho0, prob = 1 / (1 + r * mu0 / rho0), log = FALSE)

# simulate data based on Gill's model
x <- rnbinom(n, size = rho0, prob = 1 / (1 + r * mu0 / rho0))
# number of incidents that a nurse experiences during r shifts
x_bar <- mean(x)
x2_bar <- mean(x ** 2)

# parameters of interest
rho <- rep(0, T + 1)
mu <- rep(0, T + 1)

# quantity of interest P(N(t) >= 14)
N <- rep(0, T + 1)

# initialization
rho[1] = 1
mu[1] = 0.015
N[1] = 0

```

```r
# Gibbs sampler
for(t in 1:T){
  # lambda
  lambda <- rep(0, n)
  for(i in 1:n){
    lambda[i] <- rgamma(1, shape = x[i] + rho[t], rate = r + rho[t] / mu[t])
  }

  # full conditional mu (inverse gamma)
  mu[t + 1] <- rinvgamma(1, n * rho[t] - 0.01, rho[t] * sum(lambda))

  # full conditional rho (discretized)
  prob <- rep(0, 300)
  for(j in 1:300){
    # p0 is Gamma(1,1) distributed
    prob[j] <- (dgamma(j * 0.005, shape = 1, rate = 1, log = FALSE)
    / (gamma(j * 0.005) ** n)) * exp(-(j * 0.005 / mu[t + 1])
    * sum(lambda) + n * j * 0.005 * log(j * 0.005 / mu[t + 1])
    + (j * 0.005 - 1) * sum(log(lambda)))
  }
  c <- sum(prob)
  rho[t + 1] <- sample((1:300) * 0.005, size = 1, replace = TRUE, prob = prob / c)

  # probability experience at least 14 incidents
  N[t + 1] <- 1 - pnbinom(13, size = rho[t + 1],
  prob = 1 / (1 + r * mu[t + 1] / rho[t + 1]), log = FALSE)
}

# save dataset
save(x, rho, mu, N, file = 'Gibbs_sampler_dataset.Rdata')
df <- data.frame('rho' = rho, 'mu' = mu, 'probability' = N)
dg <- data.frame('incidents' = x)
write_xlsx(dg, '~/Documents/Mathematical_Sciences/Research_Thesis/R/Plot/Gibbs_sampler/
simulated_dataset_incidents_Gibbs.xlsx')
write_xlsx(df, '~/Documents/Mathematical_Sciences/Research_Thesis/R/Plot/Gibbs_sampler/
simulated_dataset_outcome_Gibbs.xlsx')

setwd('/Users/patricia/Documents/Mathematical_Sciences/Research_Thesis/R/Plot/
Gibbs_sampler')

# traceplot rho
pdf('GS_rho.pdf', width = 10, height = 6)
traceplot(as.mcmc(rho), main = expression(paste('Traceplot of ', hat(rho))),
          ylab = expression(paste(rho)), xaxt = 'n')
axis(1, at = seq(0, T, by = T / 5))
abline(h = rho0, col = 'red4', lwd = 3, lty = 3)
abline(h = mean(rho[(m + 1):(T + 1)]), col = 'aquamarine4', lwd = 3, lty = 3)
legend('bottomleft', inset = 0.025, legend = c('True value', 'Mean estimated value'),
       col = c('red4', 'aquamarine4'), lty = 3:3, cex = 0.6, lwd = 3:3, bg = 'white')
dev.off()

# traceplot mu
pdf('GS_mu.pdf', width = 10, height = 6)
traceplot(as.mcmc(mu), main = expression(paste('Traceplot of ', hat(mu))),
          ylab = expression(paste(mu)), xaxt = 'n')
axis(1, at = seq(0, T, by = T / 5))
abline(h = mu0, col = 'red4', lwd = 3, lty = 3)
abline(h = mean(mu[(m + 1):(T + 1)]), col = 'aquamarine4', lwd = 3, lty = 3)
legend('bottomleft', inset = 0.025, legend = c('True value', 'Mean estimated value'),
       col = c('red4', 'aquamarine4'), lty = 3:3, cex = 0.6, lwd = 3:3, bg = 'white')
dev.off()

# traceplot P(N(t) >= 14)
pdf('GS_N.pdf', width = 10, height = 6)
traceplot(as.mcmc(N), main = expression(paste('Traceplot of ', P(N(t) >= 14))),
          ylab = 'N', xaxt = 'n')
axis(1, at = seq(0, T, by = T / 5))
abline(h = N0, col = 'red4', lwd = 3, lty = 3)
abline(h = mean(N[(m + 1):(T + 1)]), col = 'aquamarine4', lwd = 3, lty = 3)
legend('topright', inset = 0.025, legend = c('True value', 'Mean estimated value'),
       col = c('red4', 'aquamarine4'), lty = 3:3, cex = 0.6, lwd = 3:3, bg = 'white')
dev.off()
```

```
# density plot rho
pdf('GS_rho_plot.pdf', width = 10, height = 6)
hist(rho[(m + 1):(T + 1)], breaks = 100,
     main = expression(paste('Posterior density of ', hat(rho))),
     xlab = expression(paste(rho)),
     freq = FALSE)
lines(density(rho[(m + 1):(T + 1)]), col = 'red4', lwd = 2) # add corresponding density
dev.off()

# density plot mu
pdf('GS_mu_plot.pdf', width = 10, height = 6)
hist(mu[(m + 1):(T + 1)], breaks = 100,
     main = expression(paste('Posterior density of ', hat(mu))),
     xlab = expression(paste(mu)),
     freq = FALSE)
lines(density(mu[(m + 1):(T + 1)]), col = 'red4', lwd = 2) # add corresponding density
dev.off()

# summary true value, mean and sd of parameters and P(N(t) >= 14)
par <- data.frame(N[(m + 1):(T + 1)], rho[(m + 1):(T + 1)], mu[(m + 1):(T + 1)])
colnames(par) <- c('P(N(t) >= 14)', 'rho', 'mu')
trueval <- c(N0, rho0, mu0)
mean <- sapply(par, FUN = mean)
sd <- sapply(par, FUN = sd)
dh <- cbind(trueval, mean, sd)
dh

# save table with outcomes
save(dh, file = 'GS_outcomes_table.Rdata')
```

D.1.2 Hamiltonian Monte Carlo

Listing D.2: No-U-Turn sampler discussed in Subsection 11.2.3, programmed in Python (PyMC3).

```
import os
import time
import pyreadr

import arviz as az
import numpy as np
import pymc3 as pm
import matplotlib.pyplot as plt

from scipy.special import gamma

az.style.use(['default'])

# import data simulated data set from R
os.chdir('/Users/patricia/Documents/Mathematical_Sciences/Research_Thesis/Python/Plot')
dataset = pyreadr.read_r('/Users/patricia/Documents/Mathematical_Sciences/
Research_Thesis/R/Plot/Gibbs_sampler/Gibbs_sampler_dataset.Rdata')

x = dataset['x'].to_numpy() # number of incidents
x_data = np.squeeze(x)

n = 27 * 3 # number of nurses
r = 201 # number of shifts

r_data = np.empty(n)
r_data.fill(r) # vector of the number of shifts

# parameter choice of Gill et al.
# rho0 = 1
# mu0 = 27/1734

# simulate data set
# x_data = np.random.negative_binomial(n = rho0, p = 1 / (1 + r * mu0 / rho0), size = n)
# vector of the number of incidents

niter = 1000
chains = 4

with pm.Model() as mixed_poisson:
    # define priors
    rho = pm.Gamma('rho', alpha = 1, beta = 1)
    mu = pm.Beta('mu', alpha = 0.01, beta = 1)

    # model heterogeneity
    lamb = pm.Gamma('lamb', alpha = rho, beta = rho / mu, shape = n)

    # observations
    x_obs = pm.Poisson('x_obs', mu = r_data * lamb, observed = x_data)

    # model specifications
    step = pm.NUTS()

output_dir = os.path.abspath(os.path.join(os.path.dirname(__file__), '..', 'Output'))
time_str = time.strftime('%Y%m%d-%H%M%S')

with mixed_poisson:
    # draw posterior samples
    trace = pm.sample(niter, step = step, return_inferencedata = False,
    chains = chains, cores = 1)

    # specifications plots
    var_names = ['rho', 'mu']
    chain_prop = {'color': ['teal', 'indianred', 'lightseagreen', 'maroon'],
    'alpha': [0.5]}
    lines = (('rho', {}, np.mean(trace['rho'])), ('mu', {}, np.mean(trace['mu'])))

    # traceplot of the parameters
    pm.plot_trace(trace, figsize = (12,8), var_names = ['rho', 'mu'], lines = lines,
    chain_prop = chain_prop, compact = True, legend = True)
```

```
    plt.savefig(os.path.join(output_dir, f'Traceplot rho and mu {time_str}.pdf'),
    format = 'pdf', dpi = 600)

    # forest plots of the parameters
    pm.plot_forest(trace, var_names = 'rho', colors = 'maroon')
    plt.savefig(os.path.join(output_dir, f'Forestplot rho {time_str}.pdf'),
    format = 'pdf', dpi = 600)

    pm.plot_forest(trace, var_names = 'mu', colors = 'maroon')
    plt.savefig(os.path.join(output_dir, f'Forestplot mu {time_str}.pdf'),
    format = 'pdf', dpi = 600)
    plt.show()

# summary of parameters
with mixed_poisson:
    summary = pm.summary(trace, round_to = 10, kind = 'all')
    print(summary)

def poissongamma(n, rho, t, mu):
    """
    Parameters
    ----------
    n : value of p-value P(N(t) >= n)
    rho : parameter of the mixed Poisson model that models the heterogeneity
    t : time period (given in desired time unit)
    mu : parameter of the mixed Poisson model that models the overall probability
    of having an incident per time unit

    Returns
    -------
    p-value P(N(t) >= n)

    """
    p = 1/(1 + (t * mu) / rho)
    q = 0

    for k in range(0, n):
        q += gamma(k + rho) / (gamma(rho) * gamma(k + 1)) * np.power(p, rho)
        * np.power(1 - p, k)

    return 1 - q

# traceplot P(N(t) >= 14)
x = np.linspace(1, niter, niter, dtype = 'int')
N = poissongamma(14, rho = trace['rho'], t = r, mu = trace['mu'])
# N0 = poissongamma(14, rho = rho0, t = r, mu = mu0) # true value

plt.plot(x, N[0:niter], label = '0', color = 'teal', alpha = 0.5)
plt.plot(x, N[niter:2 * niter], label = '1', color = 'indianred', alpha = 0.5)
plt.axhline(y = np.mean(N), color = 'black', alpha = 0.4)
plt.plot(x, N[2 * niter:3 * niter], label = '2', color = 'lightseagreen', alpha = 0.5)
plt.plot(x, N[3 * niter:4 * niter], label = '3', color = 'maroon', alpha = 0.5)
plt.xlim(0, niter)
plt.legend(loc = 'upper right', title = 'chain')
plt.title('$P(N(t) \geq 14)$')
plt.savefig(os.path.join(output_dir, f'Traceplot N {time_str}.pdf'),
format = 'pdf', dpi = 600)
```

E
Data

Some of the simulated data sets and outcomes that we refer to below can be accessed through Github: `https://github.com/PCMDEBRUIN/Incident-series.git`.

E.1 Data Lucia de Berk

Hospital name (and ward number)	**JCH**	**RCH - 41**	**RCH - 42**	**Total**
Total number of shifts	1029	366	339	1734
Lucia de Berk's number of shifts	142	1	58	201
Total number of incidents	8	5	14	27
Number of incidents during Lucia de Berk's shifts	8	1	5	14

Table E.1: Original data on shifts and incidents at the Juliana Children's Hospital (JCH) and the Red Cross Hospital (RCH). Here, 41 and 42 denote the different ward numbers of the Red Cross Hospital (Gill et al., 2018).

Hospital name (and ward number)	**JCH**	**RCH - 41**	**RCH - 42**	**Total**
Total number of shifts	1029	366	339	1734
Lucia de Berk's number of shifts	142	3	58	203
Total number of incidents	11	5	10	26
Number of incidents during Lucia de Berk's shifts	7	1	1	9

Table E.2: Corrected data on shifts and incidents at the Juliana Children's Hospital (JCH) and the Red Cross Hospital (RCH). Here, 41 and 42 denote the different ward numbers of the Red Cross Hospital (Gill et al., 2010).

E.2 Simulated data roster cases

E.2.1 Gibbs sampler

We simulated the number of incidents that 81 nurses experienced during $t = 201$ shifts. The simulated data is saved in the Excel document *simulated_dataset_incidents_Gibbs.xlsx* in Github. The same data can also be accessed through the Rdata document *Gibbs_sampler_dataset.Rdata*. The frequency table of the incidents is given in Table E.3.

Number of incidents	**Frequency**
0	23
1	20
2	12
3	3
4	6
5	4
6	1
7	2
8	1
9	2
10	1
11	3
12	1
13	1
14	0
15	1

Table E.3: Frequency table of simulated incidents of 81 nurses for Gibbs sampling algorithm.

E.2.2 Method of moments

We simulated the number of incidents that 1000 nurses experienced during $t = 201$ shifts. The simulated data is saved in the Excel document *simulated_dataset_incidents_MM.xlsx* in Github. The frequency table of the incidents is given in Table E.4.

Number of incidents	**Frequency**
0	267
1	145
2	124
3	120
4	87
5	74
6	48
7	36
8	34
9	17
10	11
11	9
12	4
13	6
14	3
15	4
16	2
17	2
18	2
19	1
20	2
21	0
22	1
23	0
24	1

Table E.4: Frequency table of simulated incidents of 1000 nurses for the method of moments.

E.3 Data, estimators and ANOVA simulation study

We simulated the number of incidents of the nurses for each scenario described in Appendix B.1. The simulated data can be found in the Excel document *incidents_and_shifts_simulation_study.xlsx* in Github. The estimates of the parameters ρ and μ for each simulated data set and method are saved in the Excel document *simulation_outcomes.xlsx*.

The performed Analysis of Variance can be found in the Excel document *simulation_ANOVA.xlsx*. The corresponding Analysis of Variance tables are presented in the upcoming subsection. Furthermore, the results can also accessed through the IBM SPSS Statistics document *Output.spv*.

E.3.1 ANOVA tables

* Gibbs sampling algorithm disregarded. ** One block disregarded.

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0	P-value
Treatments (methods)	0.010	2	0.005	1.161	0.336
Blocks (samples)	0.731	9	0.81		
Error	0.075	18	0.04		
Total	0.816	29			

(a) Parameter ρ.

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0	P-value
Treatments (methods)	$8.775 \cdot 10^{-5}$	2	$4.388 \cdot 10^{-5}$	21.566	< 0.001
Blocks (samples)	0.001	9	$1.059 \cdot 10^{-4}$		
Error	$3.662 \cdot 10^{-5}$	18	$2.035 \cdot 10^{-6}$		
Total	0.001	29			

(b) Parameter μ.

Table E.5: Analysis of Variance for parameters ρ and μ, with $\rho_0 = 0.25, n = 25$, group composition (i).

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0	P-value
Treatments (methods)	17.115	2	8.557	1.302	0.296
Blocks (samples)	90.971	9	10.108		
Error	118.309	18	6.573		
Total	226.395	29			

(a) Parameter ρ.

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0	P-value
Treatments (methods)	$7.526 \cdot 10^{-6}$	2	$3.763 \cdot 10^{-6}$	38.395	< 0.001
Blocks (samples)	$4.415 \cdot 10^{-4}$	9	$4.905 \cdot 10^{-5}$		
Error	$1.764 \cdot 10^{-6}$	18	$9.801 \cdot 10^{-8}$		
Total	$4.507 \cdot 10^{-4}$	29			

(b) Parameter μ.

Table E.6: Analysis of Variance for parameters ρ and μ, with $\rho_0 = 1, n = 25$, group composition (i).

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0	P-value
Treatments (methods)	29.988	2	14.994	2.688	0.095
Blocks (samples)	75.627	9	8.403		
Error	100.412	18	5.578		
Total	206.027	29			

(a) Parameter ρ.

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0	P-value
Treatments (methods)	$1.570 \cdot 10^{-6}$	2	$7.850 \cdot 10^{-7}$	47.638	< 0.001
Blocks (samples)	$3.289 \cdot 10^{-4}$	9	$3.654 \cdot 10^{-5}$		
Error	$2.966 \cdot 10^{-7}$	18	$1.648 \cdot 10^{-8}$		
Total	$3.307 \cdot 10^{-4}$	29			

(b) Parameter μ.

Table E.7: Analysis of Variance for parameters ρ and μ, with $\rho_0 = 2.5, n = 25$, group composition (i).

Source of Variation	**Sum of Squares**	**Degrees of Freedom**	**Mean Square**	$\boldsymbol{F_0}$	$\boldsymbol{P}$**-value**
Treatments (methods)	0.001	1	0.001	0.452	0.518
Blocks (samples)	0.075	9	0.008		
Error	0.020	9	0.002		
Total	0.096	19			

(a) Parameter ρ.

Source of Variation	**Sum of Squares**	**Degrees of Freedom**	**Mean Square**	$\boldsymbol{F_0}$	$\boldsymbol{P}$**-value**
Treatments (methods)	$3.813 \cdot 10^{-6}$	1	$3.813 \cdot 10^{-6}$	84.363	< 0.001
Blocks (samples)	$2.605 \cdot 10^{-4}$	9	$2.895 \cdot 10^{-5}$		
Error	$4.068 \cdot 10^{-7}$	9	$4.519 \cdot 10^{-8}$		
Total	$2.648 \cdot 10^{-4}$	19			

(b) Parameter μ.

Table E.8: Analysis of Variance for parameters ρ and μ, with $\rho_0 = 0.25, n = 75$, group composition (i).*

Source of Variation	**Sum of Squares**	**Degrees of Freedom**	**Mean Square**	$\boldsymbol{F_0}$	$\boldsymbol{P}$**-value**
Treatments (methods)	0.070	2	0.035	2.902	0.081
Blocks (samples)	2.454	9	0.273		
Error	0.218	18	0.012		
Total	2.742	29			

(a) Parameter ρ.

Source of Variation	**Sum of Squares**	**Degrees of Freedom**	**Mean Square**	$\boldsymbol{F_0}$	$\boldsymbol{P}$**-value**
Treatments (methods)	$5.564 \cdot 10^{-6}$	2	$2.782 \cdot 10^{-7}$	65.782	< 0.001
Blocks (samples)	$1.666 \cdot 10^{-4}$	9	$1.851 \cdot 10^{-5}$		
Error	$7.612 \cdot 10^{-8}$	18	$4.229 \cdot 10^{-9}$		
Total	$1.673 \cdot 10^{-4}$	29			

(b) Parameter μ.

Table E.9: Analysis of Variance for parameters ρ and μ, with $\rho_0 = 1, n = 75$, group composition (i).

Source of Variation	**Sum of Squares**	**Degrees of Freedom**	**Mean Square**	$\boldsymbol{F_0}$	$\boldsymbol{P}$**-value**
Treatments (methods)	4.830	2	2.415	16.446	< 0.001
Blocks (samples)	4.127	9	0.459		
Error	2.643	18	0.147		
Total	11.601	29			

(a) Parameter ρ.

Source of Variation	**Sum of Squares**	**Degrees of Freedom**	**Mean Square**	$\boldsymbol{F_0}$	$\boldsymbol{P}$**-value**
Treatments (methods)	$1.150 \cdot 10^{-7}$	2	$5.750 \cdot 10^{-8}$	102.580	< 0.001
Blocks (samples)	$2.011 \cdot 10^{-4}$	9	$2.234 \cdot 10^{-4}$		
Error	$1.009 \cdot 10^{-8}$	18	$5.605 \cdot 10^{-10}$		
Total	$2.012 \cdot 10^{-4}$	29			

(b) Parameter μ.

Table E.10: Analysis of Variance for parameters ρ and μ, with $\rho_0 = 2.5, n = 75$, group composition (i).

Source of Variation	**Sum of Squares**	**Degrees of Freedom**	**Mean Square**	$\boldsymbol{F_0}$	$\boldsymbol{P}$**-value**
Treatments (methods)	$3.250 \cdot 10^{-4}$	1	$3.250 \cdot 10^{-4}$	0.271	0.615
Blocks (samples)	0.070	9	0.008		
Error	0.011	9	0.001		
Total	0.081	19			

(a) Parameter ρ.

Source of Variation	**Sum of Squares**	**Degrees of Freedom**	**Mean Square**	$\boldsymbol{F_0}$	$\boldsymbol{P}$**-value**
Treatments (methods)	$1.193 \cdot 10^{-6}$	1	$1.193 \cdot 10^{-6}$	109.805	< 0.001
Blocks (samples)	$6.142 \cdot 10^{-5}$	9	$6.825 \cdot 10^{-6}$		
Error	$9.779 \cdot 10^{-8}$	9	$1.087 \cdot 10^{-8}$		
Total	$6.271 \cdot 10^{-5}$	19			

(b) Parameter μ.

Table E.11: Analysis of Variance for parameters ρ and μ, with $\rho_0 = 0.25, n = 150$, group composition (i).*

Source of Variation	**Sum of Squares**	**Degrees of Freedom**	**Mean Square**	$\boldsymbol{F_0}$	$\boldsymbol{P}$**-value**
Treatments (methods)	$3.050 \cdot 10^{-4}$	1	$3.050 \cdot 10^{-4}$	0.093	0.768
Blocks (samples)	0.712	9	0.079		
Error	0.030	9	0.003		
Total	0.742	19			

(a) Parameter ρ.

Source of Variation	**Sum of Squares**	**Degrees of Freedom**	**Mean Square**	$\boldsymbol{F_0}$	$\boldsymbol{P}$**-value**
Treatments (methods)	$6.489 \cdot 10^{-8}$	1	$6.489 \cdot 10^{-8}$		< 0.001
Blocks (samples)	$7.650 \cdot 10^{-5}$	9	$8.5003 \cdot 10^{-6}$		
Error	$1.587 \cdot 10^{-9}$	9	$1.763 \cdot 10^{-10}$		
Total	$7.657 \cdot 10^{-5}$	19			

(b) Parameter μ.

Table E.12: Analysis of Variance for parameters ρ and μ, with $\rho_0 = 1, n = 150$, group composition (i).*

Source of Variation	**Sum of Squares**	**Degrees of Freedom**	**Mean Square**	$\boldsymbol{F_0}$	$\boldsymbol{P}$**-value**
Treatments (methods)	3.936	2	1.968	16.196	< 0.001
Blocks (samples)	6.919	9	0.769		
Error	2.187	18	0.122		
Total	13.042	29			

(a) Parameter ρ.

Source of Variation	**Sum of Squares**	**Degrees of Freedom**	**Mean Square**	$\boldsymbol{F_0}$	$\boldsymbol{P}$**-value**
Treatments (methods)	$1.607 \cdot 10^{-8}$	2	$8.037 \cdot 10^{-9}$	15.191	< 0.001
Blocks (samples)	$4.500 \cdot 10^{-5}$	9	$4.500 \cdot 10^{-6}$		
Error	$9.461 \cdot 10^{-9}$	18	$5.256 \cdot 10^{-10}$		
Total	$4.502 \cdot 10^{-5}$	29			

(b) Parameter μ.

Table E.13: Analysis of Variance for parameters ρ and μ, with $\rho_0 = 2.5, n = 150$, group composition (i).

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0	P-value
Treatments (methods)	0.0953	2	0.048	11.959	< 0.001
Blocks (samples)	0.473	9	0.053		
Error	0.072	18	0.004		
Total	0.640	29			

(a) Parameter ρ.

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0	P-value
Treatments (methods)	$2.812 \cdot 10^{-5}$	2	$1.406 \cdot 10^{-5}$	83.856	< 0.001
Blocks (samples)	$4.959 \cdot 10^{-4}$	9	$5.510 \cdot 10^{-5}$		
Error	$3.018 \cdot 10^{-6}$	18	$1.677 \cdot 10^{-7}$		
Total	$5.270 \cdot 10^{-4}$	29			

(b) Parameter μ.

Table E.14: Analysis of Variance for parameters ρ and μ, with $\rho_0 = 0.25, n = 25$, group composition (ii).

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0	P-value
Treatments (methods)	1.948	2	0.974	9.359	0.002
Blocks (samples)	4.733	9	0.526		
Error	1.874	18	0.104		
Total	8.556	29			

(a) Parameter ρ.

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0	P-value
Treatments (methods)	$5.443 \cdot 10^{-6}$	2	$2.721 \cdot 10^{-6}$	26.686	< 0.001
Blocks (samples)	$3.601 \cdot 10^{-4}$	9	$4.001 \cdot 10^{-5}$		
Error	$1.836 \cdot 10^{-6}$	18	$1.020 \cdot 10^{-7}$		
Total	$3.674 \cdot 10^{-4}$	29			

(b) Parameter μ.

Table E.15: Analysis of Variance for parameters ρ and μ, with $\rho_0 = 1, n = 25$, group composition (ii).

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0	P-value
Treatments (methods)	37.196	2	18.598	9.698	0.002
Blocks (samples)	39.197	8	4.900		
Error	30.683	16	1.918		
Total	107.076	26			

(a) Parameter ρ.

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0	P-value
Treatments (methods)	$2.973 \cdot 10^{-6}$	2	$1.486 \cdot 10^{-6}$	9.133	0.002
Blocks (samples)	$2.449 \cdot 10^{-4}$	8	$3.061 \cdot 10^{-5}$		
Error	$2.604 \cdot 10^{-6}$	16	$1.627 \cdot 10^{-7}$		
Total	$2.505 \cdot 10^{-4}$	26			

(b) Parameter μ.

Table E.16: Analysis of Variance for parameters ρ and μ, with $\rho_0 = 2.5, n = 25$, group composition (ii).**

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0	P-value
Treatments (methods)	0.009	1	0.009	4.886	0.054
Blocks (samples)	0.114	9	0.013		
Error	0.016	9	0.002		
Total	0.139	19			

(a) Parameter ρ.

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0	P-value
Treatments (methods)	$2.772 \cdot 10^{-6}$	1	$2.772 \cdot 10^{-6}$	16.291	0.003
Blocks (samples)	$2.336 \cdot 10^{-4}$	9	$2.596 \cdot 10^{-5}$		
Error	$1.531 \cdot 10^{-6}$	9	$1.701 \cdot 10^{-7}$		
Total	$2.379 \cdot 10^{-4}$	19			

(b) Parameter μ.

Table E.17: Analysis of Variance for parameters ρ and μ, with $\rho_0 = 0.25, n = 75$, group composition (ii).*

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0	P-value
Treatments (methods)	0.263	2	0.132	4.720	0.022
Blocks (samples)	2.049	9	0.228		
Error	0.502	18	0.028		
Total	2.813	29			

(a) Parameter ρ.

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0	P-value
Treatments (methods)	$3.310 \cdot 10^{-7}$	2	$1.655 \cdot 10^{-7}$	12.480	< 0.001
Blocks (samples)	$1.040 \cdot 10^{-4}$	9	$1.155 \cdot 10^{-5}$		
Error	$2.387 \cdot 10^{-7}$	18	$1.326 \cdot 10^{-8}$		
Total	$1.045 \cdot 10^{-4}$	29			

(b) Parameter μ.

Table E.18: Analysis of Variance for parameters ρ and μ, with $\rho_0 = 1, n = 75$, group composition (ii).

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0	P-value
Treatments (methods)	1.750	2	0.875	6.295	0.0008
Blocks (samples)	10.124	9	1.125		
Error	2.502	18	0.139		
Total	14.376	29			

(a) Parameter ρ.

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0	P-value
Treatments (methods)	$1.057 \cdot 10^{-7}$	2	$5.285 \cdot 10^{-8}$	4.314	0.029
Blocks (samples)	$2.429 \cdot 10^{-5}$	9	$12.699 \cdot 10^{-6}$		
Error	$2.205 \cdot 10^{-7}$	18	$1.225 \cdot 10^{-8}$		
Total	$2.462 \cdot 10^{-5}$	29			

(b) Parameter μ.

Table E.19: Analysis of Variance for parameters ρ and μ, with $\rho_0 = 2.5, n = 75$, group composition (ii).

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0	P-value
Treatments (methods)	0.006	1	0.006	3.362	0.100
Blocks (samples)	0.040	9	0.004		
Error	0.017	9	0.002		
Total	0.063	19			

(a) Parameter ρ.

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0	P-value
Treatments (methods)	$2.046 \cdot 10^{-7}$	1	$2.046 \cdot 10^{-7}$	6.358	0.037
Blocks (samples)	$9.659 \cdot 10^{-5}$	9	$1.073 \cdot 10^{-5}$		
Error	$2.896 \cdot 10^{-7}$	9	$3.218 \cdot 10^{-8}$		
Total	$9.708 \cdot 10^{-5}$	19			

(b) Parameter μ.

Table E.20: Analysis of Variance for parameters ρ and μ, with $\rho_0 = 0.25, n = 150$, group composition (ii).*

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0	P-value
Treatments (methods)	0.017	1	0.017	2.757	0.131
Blocks (samples)	0.240	9	0.027		
Error	0.057	9	0.006		
Total	0.315	19			

(a) Parameter ρ.

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0	P-value
Treatments (methods)	$3.887 \cdot 10^{-8}$	1	$3.887 \cdot 10^{-8}$	2.074	0.184
Blocks (samples)	$3.906 \cdot 10^{-5}$	9	$4.340 \cdot 10^{-6}$		
Error	$1.687 \cdot 10^{-7}$	9			
Total	$3.927 \cdot 10^{-5}$	19			

(b) Parameter μ.

Table E.21: Analysis of Variance for parameters ρ and μ, with $\rho_0 = 1, n = 150$, group composition (ii).*

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0	P-value
Treatments (methods)	0.189	1	0.189	5.651	0.041
Blocks (samples)	6.626	9	0.736		
Error	0.301	9	0.033		
Total	7.116	19			

(a) Parameter ρ.

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0	P-value
Treatments (methods)	$7.001 \cdot 10^{-8}$	1	$7.001 \cdot 10^{-8}$	4.456	0.064
Blocks (samples)	$2.065 \cdot 10^{-5}$	9	$2.294 \cdot 10^{-6}$		
Error	$1.414 \cdot 10^{-7}$	9	$1.571 \cdot 10^{-8}$		
Total	$2.086 \cdot 10^{-5}$	19			

(b) Parameter μ.

Table E.22: Analysis of Variance for parameters ρ and μ, with $\rho_0 = 2.5, n = 150$, group composition (ii).*

E.4 Ship damage incidents

Ship type	Year of construction	Period of operation	Aggregate months service	Number of damage incidents
A	1960-64	1960-74	127	0
A	1960-64	1975-79	63	0
A	1965-69	1960-74	1095	3
A	1965-69	1975-79	1095	4
A	1970-74	1960-74	1512	6
A	1970-74	1975-79	3353	18
A	1975-79	1960-74	0	0*
A	1975-79	1975-79	2244	11
B	1960-64	1960-74	44882	39
B	1960-64	1975-79	17176	29
B	1965-69	1960-74	28609	58
B	1965-69	1975-79	20370	53
B	1970-74	1960-74	7064	12
B	1970-74	1975-79	13099	44
B	1975-79	1960-74	0	0*
B	1975-79	1975-79	7117	18
C	1960-64	1960-74	1179	1
C	1960-64	1975-79	552	1
C	1965-69	1960-74	781	0
C	1965-69	1975-79	676	1
C	1970-74	1960-74	783	6
C	1970-74	1975-79	1948	2
C	1975-79	1960-74	0	0*
C	1975-79	1975-79	274	1
D	1960-64	1960-74	251	0
D	1960-64	1975-79	105	0
D	1965-69	1960-74	288	0
D	1965-69	1975-79	192	0
D	1970-74	1960-74	349	2
D	1970-74	1975-79	1208	11
D	1975-79	1960-74	0	0*
D	1975-79	1975-79	2051	4
E	1960-64	1960-74	45	0
E	1960-64	1975-79	0	0**
E	1965-69	1960-74	789	7
E	1965-69	1975-79	437	7
E	1970-74	1960-74	1157	5
E	1970-74	1975-79	2161	12
E	1975-79	1960-74	0	0*
E	1975-79	1975-79	542	1

Table E.23: Number of reported damage incidents and aggregate months service by ship type, year of construction and period of operation (McCullagh and Nelder, 1989, Subsection 6.3.2). Notes by author: * Necessarily empty cells. ** Accidentally empty cell.

Bibliography

R.A. Adams and C. Essex. *Calculus: A Complete Course*. Pearson, eighth edition, 2014. ISBN 978-0-321-78107-9.

I. Adan and J. Resing. Queueing systems. Distributed at Mastermath course *Queueing Theory*, March 26, 2015.

C. Aitken and F. Taroni. *Statistics and the Evaluations of Evidence for Forensic Scientists*. John Wiley & Sons, Ltd, second edition, November, 2004. ISBN 978-0-470-01122-5. doi: 10.1002/0470011238.

F. Alkemade. Een bayesiaanse visie op de zaak xxx. Technical report, Ministerie van Justitie en Veiligheid, October 1, 2015.

F.J.M. Alkemade. Dr. frans j. m. alkemade (home page), n.d. URL `https://nl.linkedin.com/in/dr-frans-j-m-alkemade-07a0878`. LinkedIn.

A. Baddeley. Spatial point processes and their applications, 2007. URL `https://link.springer.com/content/pdf/10.1007/978-3-540-38175-4_1.pdf`.

E. Bolviken and T. Egeland. Arson, statistics and the law: can the defendant's proximity to a large number of fires be explained by chance? *Science & Justice*, 35:97–104, 1995. doi: https://doi.org/10.1016/S1355-0306(95)72640-5.

N. Bou-Rabee and J.M. Sanz-Serna. Geometric integrators and the hamiltonian monte carlo method. *Acta Numerica*, 27:1–92, May, 2018. ISSN 1474-0508. doi: 10.1017/s0962492917000101. URL `http://dx.doi.org/10.1017/S0962492917000101`.

Brandweeracademie. Risicogroepen en brandveiligheid. Technical report, Instituut Fysieke Veiligheid, July 1, 2019. URL `https://www.ifv.nl/kennisplein/Documents/20190701-BA-Risicogroepen-en-brandveiligheid.pdf`.

N. Brouwer, P. de Bruin, O. Schavemaker, and C. Smans. The case of lucia de berk. Essay written for Utrecht University course *Orientation on Mathematical Research*, January 23, 2020.

R. Burks. Insulin as a murder weapon, August 6, 2021. URL `https://www.chemistryworld.com/opinion/insulin-as-a-murder-weapon/4014013.article`.

C. Caballero-Cárdenas. The nuts algorithm and its efficiency, March 26, 2021. URL `https://www.uma.es/media/files/The_NUTS_algorithm_and_its_efficiency.pdf`.

CBS. Brandweerstatistiek 2013. Technical report, Centraal Bureau voor de Statistiek, 2014. URL `https://www.cbs.nl/nl-nl/publicatie/2014/45/brandweerstatistiek-2013`.

CBS, WODC, and Raad voor de rechtspraak. Criminaliteit en rechtshandhaving 2013. Technical report, Centraal Bureau voor de Statistiek, Wetenschappelijk Onderzoek- en Documentatiecentrum, Raad voor de rechtspraak, 2014. URL `https://www.cbs.nl/nl-nl/publicatie/2014/45/criminaliteit-en-rechtshandhaving-2013`.

M. Collins. Statistics on trail. what statisticians haw to say in the case of lucia de b. Master's thesis, Utrecht University, December 22, 2005.

J.D. Cook. Inverse gamma distribution, October 3, 2008. URL `http://www.johndcook.com/inverse gamma.pdf`.

C. Dahlman. De-biasing legal fact-finders with bayesian thinking. *Topics in Cognitive Science*, 12:1115–1131, 2020. doi: 10.1111/tops.12419.

C. Dahlman and A. Pundik. The problem with naked statistical evidence. *Philosophical Foundations of Evidence Law*, January 2, 2021. doi: http://dx.doi.org/10.2139/ssrn.3758860.

M. David and D. Jemna. Modeling the frequency of auto insurance claims by means of poisson and negative binomial models. *Analele Stiintifice ale Universitatii Al I Cuza din Iasi - Sectiunea Stiinte Economice*, 68: 151–168, 2015. doi: 10.1515/aicue-2015-0011.

A.P. Dawid. Bayes's theorem and weighing evidence by juries. *Bayes's Theorem*, 113, August 3, 2001. doi: 10.5871/bacad/9780197263419.003.0004.

A.P. Dawid. Sally clark appeal. statement by professor a.p. dawid, n.d. URL `http://www.statslab.cam.ac.uk/~apd/SallyClark_report.doc`.

A.F. de Vos. *Rekenen aan de zaak Lucy de B.*, pages 163–174. Boom Juridische uitgevers, 2004. ISBN 9054545240.

F.M. Dekking, C. Kraaikamp, H.P. Lopuhaä, and L.E. Meester. *A Modern Introduction to Probability and Statistics. Understanding Why and How.* Springer, 2010. ISBN 9781849969529.

S. Dirksen. Lecture notes stochastic processes. Distributed at Utrecht University course *Stochastische Processen*, 2019a.

S. Dirksen. Stochastic processes - homework assignment 4. Distributed at Utrecht University course *Stochastische Processen*, June, 2019b.

H. Elffers. Verdeling reanimatie- en overlijdensgevallen in het juliana kinderziekenhuis en rode kruisziekenhuis. Technical report, Nederlands Studiecentrum Criminaliteit en Rechtshandhaving, May 8, 2002.

N. Fenton, R.D. Gill, D. Lagnado, and M. Neil. Statistical issues in serial killer nurse cases. *arXiv:2106.00758*, March 23, 2022.

S.E. Fienberg and D.H. Kaye. Legal and statistical aspects of some mysterious clusters. *Journal of the Royal Statistical Society*, 154, 1991.

J. Frost. Mean squared error (mse), n.d. URL `https://statisticsbyjim.com/regression/mean-squared-error-mse/`.

Gerechtshof 's-Hertogenbosch. Ecli:nl:ghshe:2016:5165, November 22, 2016. URL `https://uitspraken.rechtspraak.nl/inziendocument?id=ECLI:NL:GHSHE:2016:5165`.

R.D. Gill, P. Groeneboom, and P. de Jong. Elementary statistics on trail - the case of lucia de berk. *arXiv.org (stat.AP) 1009.0802*, August 8, 2010.

R.D. Gill, P. Groeneboom, and P. de Jong. Elementary statistics on trail - the case of lucia de berk. *Chance*, pages 9–15, December 3, 2018. doi: 10.1080/09332480.2018.1549809.

S. Glen. Coverage probability, February 6, 2022. URL `https://www.statisticshowto.com/coverage-probability/`.

L. Green. *Queueing Analysis in Health Care*, volume 91, pages 281–307. Springer, October, 2006. ISBN 978-0-387-33635-0. doi: 10.1007/978-0-387-33636-7_10.

M.D. Hoffman and A. Gelman. The no-u-turn sampler: Adaptively setting path lengths in hamiltonian monte carlo. *Journal of Machine Learning Research*, 15:1351–1381, 2014.

Z. Jianhua. Chapter 3: Methods for generating random variables, n.d. URL `https://www.ynufe.edu.cn/pub/tsxy/jhzhao/teach/CS/notes/lec3.pdf`.

A. Kim. Exponenital distribution - intuition, derivation, and applications, August 6, 2019. URL `https://towardsdatascience.com/what-is-exponential-distribution-7bdd08590e2a`.

W. Koehrsen. The poisson distribution and poisson process explained, August, 2019. URL `https://towardsdatascience.com/the-poisson-distribution-and-poisson-process-explained-4e2cb17d459`.

D. Kundu. Exponentiated exponential distribution. *Encyclopedia of Statistical Sciences*, 2004. URL `http://home.iitk.ac.in/~kundu/paper101.pdf`.

D. Lucy. Commentary on meester *et al.* 'on the (ab)use of statistics in the legal case against nurse lucia de b.'. *Law, Probability and Risk*, 5:251–254, 2006. doi: 10.1093/lpr/mgm004.

D. Lucy and C. Aitken. A review of the role of roster data and evidence of attendance in cases of suspected excess deaths in a medical context. *Law, Probability and Risk*, 1:141–160, December 1, 2002. doi: https://doi.org/10.1093/lpr/1.2.141.

D. Lucy and C. Aitken. The evidential value of roster and attendance data in cases where fraud or medical malpractice may be suspected. Preprint, 2004.

D. Ma. The negative binomial distribution, July 11, 2011. URL `https://probabilityandstats.wordpress.com/tag/poisson-gamma-mixture/`.

P. McCullagh and J.A. Nelder. *Generalized Linear Models*. Chapman and Hall, second edition, 1989.

R. Meester and K. Slooten. *Probability and Forensic Evidence: Theory, Philosophy, and Applications*. Cambridge University Press, April, 2021. ISBN 9781108428279. doi: 10.1017/9781108596176.

R. Meester, M. Collins, R. Gill, and M. van Lambalgen. On the (ab)use of statistics in the legal case against the nurse lucia de b. *Law, Probability and Risk*, 5:233–250, 2006. doi: 10.1093/lpr/mgm003.

M. Meredith. Point estimates: mean, median or mode?, April 21, 2021a. URL `https://mmeredith.net/blog/2019/Mean_median_mode.htm`.

M. Meredith. Mean vs median for bayesian posteriors, February 12, 2021b. URL `https://mmeredith.net/blog/2021/Posterior_mean_median.htm`.

D.C. Montgomery. *Design and Analysis of Experiments*. Wiley, tenth edition, June, 2019.

R.G. Newcombe. Two-sided confidence intervals for the single proportion: comparison of seven methods. *Statistics in Medicine*, 17:857–872, April 30, 1998. doi: 10.1002/(SICI)1097-0258(19980430)17:8⟨857::AID-SIM777⟩3.0.CO;2-E.

NFI. Statistisch onderzoek naar een serie branden in xxx. Technical report, Netherlands Forensic Institute, 2019.

A. Pundik. Rethinking the use of statistical evidence to prove causation in criminal cases: a tale of (im)probability and free will. *Law and Philosophy*, 40:97–128, April, 2021. doi: https://doi.org/10.1007/s10982-020-09389-0.

Rijksoverheid. Het nederlands forensisch instituut, February 2014. URL `https://www.forensischinstituut.nl/over-het-nfi/publicaties/publicaties/2017/10/18/algemene-informatie-over-het-nederlands-forensisch-instituut-nfi`.

S.M. Ross. *Introduction to Probability Models*. Academic Press, 2014.

J. Salvatier, T.V. Wiecki, and C. Fonnesbeck. Probabilitic programming in python using pymc3. *PeerJ Computer Science*, 2, 2016. doi: 10.7717/peerj-cs.55.

V. Savani and A.A. Zhigljavsky. Efficient estimation of parameters of the negative binomial distribution. *Communications in Statistics - Theory and Methods*, 35, June, 2006. doi: 10.1080/03610920500501346.

L. Schneps and C. Colmez. *Math on Trail. How Numbers Get Used and Abused in the Courtroom*. Basic Books, 2013.

Y. Schoenmakers and T. van Ham. *Seriebrandstichters*, pages 75–92. Boom Lemma uitgevers, 2013. ISBN 9789059319912.

P. Serra. Lecture notes nonparametric statistics. Distributed at Mastermath course *Nonparametric Statistics*, June 8, 2020.

M. Sjerps. Statistiek als bewijsmiddel in het strafrecht. *STAtOR*, 5:16–20, June, 2004.

M. Sjerps and R. Meester. Selection effects and database screening in forensic science. *Forensic Science International*, 192:56–61, 2009. doi: 10.1016/j.forsciint.2009.07.022. URL `https://www.sciencedirect.com/science/article/pii/S037907380900320X`.

P.J.C. Spreij. Measure theoretic probability. Distributed at Mastermath course *Measure Theoretic Probability*, May 16, 2021. URL `https://staff.fnwi.uva.nl/p.j.c.spreij/onderwijs/master/mtp.pdf`.

M. Stephens. Summarizing and interpreting the posterior (analytic), January 28, 2017. URL `https://stephens999.github.io/fiveMinuteStats/summarize_interpret_posterior.html`.

B. Szabó and A.W. van der Vaart. Bayesian statistics. Distributed at Mastermath course *Bayesian Statistics*, May 18, 2021.

E.A. Thompson. Gamma distributions: Lm 4.6, n.d. URL `https://sites.stat.washington.edu/thompson/S341_10/Notes/week4.pdf`.

A.W. van der Vaart. Mathematische statistiek. Distributed at Mastermath course *Asymptotic Statistics*, December, 1998.

Verbond Van Verzekeraars. Aantal brandclaims gedaald, wel zorgen om accu's en batterijen, October 17, 2019. URL `https://www.verzekeraars.nl/publicaties/actueel/aantal-brandclaims-gedaald-wel-zorgen-om-accu-s-en-batterijen`.

Volksgezondheidenzorg.info. Aantal instellingen voor medisch specialistische zorg, 2021. URL https://www.volksgezondheidenzorg.info/onderwerp/ziekenhuiszorg/cijfers-context/aanbod#node-aantal-instellingen-voor-medisch-specialistische-zorg.

E.W. Weisstein. Digamma function, n.d.a. URL https://mathworld.wolfram.com/DigammaFunction.html.

E.W. Weisstein. Negative binomial distribution, n.d.b. URL https://mathworld.wolfram.com/NegativeBinomialDistribution.html.

E.W. Weisstein. Robust estimation, n.d.c. URL https://mathworld.wolfram.com/RobustEstimation.html.

Wikipedia. Overdispersion, May 14, 2021. URL https://en.wikipedia.org/wiki/Overdispersion.

Wikipedia. Angel of mercy (criminology), March 7, 2022. URL https://en.wikipedia.org/wiki/Angel_of_mercy_(criminology).

J. Windley. Confidence intervals - proportions - wilson score interval, n.d. URL https://www.cpp.edu/~jcwindley/classes/sta2260/Confidnece%20Intervals%20-%20Proportions%20-%20Wilson.pdf.