Elementary Statistics on Trial (the case of Lucia de Berk)

Richard D. Gill, Piet Groeneboom and Peter de Jong

The trial of the Dutch nurse Lucia de Berk, suspected of several murders and attempts of murder was a very high profile case in the Netherlands. The initial suspicion rested mainly on quasi-statistical considerations, which produced (based partly on incorrect calculations) extremely small probabilities. Since the outcomes proved controversial, the court claimed to have dropped the statistical calculations from the verdict. But the verdict still rested on intuitive notions as "very improbable". So statistics were at center stage.

In the conviction of Lucia de Berk an important role was played by a simple (so-called) hypergeometric model, used by the law psychologist (H. Elffers) consulted as statistician by the court, which produced very small probabilities of occurrences of certain numbers of incidents.

In this article we want to draw attention to the fact that, if we take into account the variation among nurses in incidents they experience during their shifts, these probabilities become considerably larger. This points to the danger of using an oversimplified discrete probability model in these circumstances.

The outcomes of applying our alternative model to this case are in striking contrast with those of the first calculations which led to the initial suspicions and were instrumental in determining the atmosphere surrounding the trial and subsequent hysteria. The main result is that under the assumption of heterogeneity, the probability of experiencing a number of incidents (14) that led to Lucia's conviction is about 0.0206161 or *one in* 49 if the calculations are based on the same data as used by the law psychologist of the court. In his calculation, however, this probability was equal to one in 342 million.

Figure 1: Lucia de Berk before her imprisonment

The data

We use data from the unpublished reports of Elffers [1] and [2]. But before going into this, we want to make some general remarks on the data collection.

One of the key features of the data was the flawed data collection. Here different disciplines came into conflict: criminal investigation and scientific data gathering are very different. Their objectives, methods and results are not compatible. Criminal investigation is started when there is (suspicion of) a crime, hence one is looking for or hunting down a suspect. If there is need for meaningful statistics another methodology is needed, guaranteeing clear definitions and uniformity of the data collection. In the case of Lucia de Berk this clash of cultures proved disastrous. Incidents outside shifts of Lucia were discarded and some initially reported incidents were later relabeled without clear reasons. Extra shifts without incidents and incidents outside shifts of Lucia were subsequently brought to light. Moreover, the data collection rested for a large part on memory.

Clearly, the context of a criminal investigation produces a specific mindset: on the one hand the witnesses know what is looked for (and some of them may already be convinced of the guilt of the suspect), on the other hand fear of implicating one's self and friends can considerably distort memory. The data on shifts and incidents for the period which was singled out in Elffers' reports are given in the following table (our Table 1 corrects an error in [6], where the number of shift in the ward RCH-41 was erroneously given by 336 instead of 366 in their table on top of p. 235).

Table 1: Data on shifts and incidents

Hospital name (and ward number)	JCH	RCH-41	RCH-42	Total
Total number of shifts	1029	366	339	1734
Lucia's number of shifts	142	1	58	201
Total number of incidents	8	5	14	27
Number of incidents during Lucia's shifts	8	1	5	14

JCH and RCH denote the "Juliana Children's Hospital" and "Red Cross Hospital", respectively, and 41 and 42 were different ward numbers of the Red Cross hospital.

Elffers' method

We first discuss the analysis of the law psychologist H. Elffers, the statistician consulted by the court. This analysis was based on Table 1. As was noticed later, Lucia de Berk had actually done three shifts in RCH-41 instead of just one, but we will argue from the data used by H. Elffers. As explained in [6], Elffers argued by *conditioning* on part of the data and used two fundamental assumptions:

1. There is a fixed probability p for the occurrence of an incident during a shift (for example, p does not depend on whether the shift is a day shift or a night shift or on the nurse involved, etc.),

2. Incidents occur independently of one another.

On the basis of these assumptions, one can compute the probability that L incidents occur during Lucia's shifts, given the total number I of incidents and the total number N of shifts considered in the period of study. This is a *hypergeometric probability* given by

$$\frac{\binom{S}{L}\binom{N-S}{I-L}}{\binom{N}{I}} \qquad (1)$$

where S is the number of shifts of Lucia and I is the total number of incidents, and where $\binom{S}{L}$, etc. denote binomial coefficients. If we

just take all the data of Table 1 together, we have a total number of $N = 1734$ shifts, Lucia had $S = 201$ shifts, there was a total number $I = 27$ of incidents, and $L = 14$ incidents during shifts of Lucia. If we evaluate (1) with these values for N, S, I and L, we get the very small probability of about one in 4.2 million.

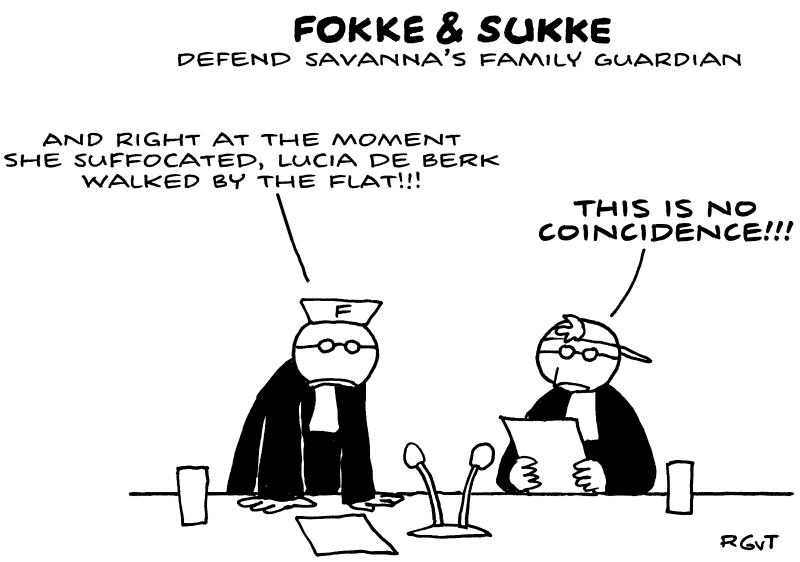

Figure 2: A Fokke & Sukke cartoon from 10-30-2007 in the Dutch newspaper NRC Next. The text was kindly translated into English for us by the creators of the cartoon: Reid, Geleijnse and Van Tol. Lucia de Berk was still in prison at that time. The canary and the duck are defending a family guardian, accused of being responsible for the death of the girl Savanna, who died by suffocation. The accused woman was in fact acquitted (by another defense!). What counselor Sukke is saying corresponds to what the law psychologist H. Elffers told the court: "Honored court, this is no coincidence. The rest is up to you.".

If we want to compute the probability (p-value) that a nurse is present with 14 *or more* incidents in Elffers' method of testing a null hypothesis of no systematic effects on these combined data (but he actually did not test it in this way on the combined data, see below), we have to sum the probabilities for $L = 14, 15, \ldots, 27$, and then we get the probability of about one in 3.8 million. This is a very small probability, although still about 100 times larger than the probability Elffers arrived at as described below. For the model we introduce in the next section, however, we get, using the same data, a probability of *one in* 49.

However, Elffers proceeded somewhat differently, not combining the data of the different hospitals. The details of what he actually did are described in [6]. The most important mistake he made in his calculation was to take the three hospitals separately, and multiplying the probabilities he got for these separately. This has the absurd consequence that a nurse working in several different hospitals gets a

higher chance of being accused of inexplicably being present at incidences than a nurse working in just one hospital. In this way he arrived at his estimate that the probability that Lucia de Berk was present at the given numbers of incidents at the Juliana Children's hospital and the Red Cross hospital was equal to one in 342 million. We refer the interested reader to [6] and to Chapter 7 "Math error number 7: the incredible coincidence" of the book [7].

Post-hoc testing

A reviewer of this paper has asked us to comment on the issue of the danger of post-hoc testing: testing a hypothesis using the same data which suggested that hypothesis. Elffers actually tried to take account of this problem in the following way. He started from the assumption that the number of incidents in the data from JCH was much larger than expected, and that the purpose of his analysis was to discover whether there was an association with any of the nurses who worked on the ward. He multiplied his p-value for the association with Lucia's shifts by 27, the number of nurses in that period who worked on the same ward. By the time he came to look at the data from RCH, Lucia was a prime suspect and he judged that no further Bonferroni type correction was required. Finally, he proposed to take a very small probability for the significance level of his test.

In fact, his starting assumption was false: in the previous year there had been no incidents in the ward, but the year before that, an even larger number. The hospital director had not revealed the information from two years ago to the investigators since the ward previously had had a different name (he had changed it).

One could try to use a Bayesian approach to deal with the post-hoc problem. There would be good arguments for a rather low prior probability of an arbitrary nurse being a serial killer. The difficult task for the Bayesian would be determining a reasonable model for number of incidents if Lucia is a murderer, since one has to take into account that some proportion of the incidents are not murders at all. Heterogeneity would also remain an issue for a Bayesian analysis. Explaining the methodology in a court of law could well be the biggest barrier.

Alternative model

We can model the incidents that a nurse experiences by a so-called Poisson process, with a nurse-dependent intensity A, where we use A for "accident proneness". A Poisson process is used to model incoming phone calls during non-busy hours, fires in a big city, etc. Since we believe the incidents to be rare, a Poisson process is an obvious choice for modeling the incidents that a nurse experiences.

This approach models two separate phenomena. Firstly, the intensity of nurses seeing or reporting incidents is modeled by introducing the random variable A. We assume that A has an exponential distribution, but other choices are also possible.

Note that we move away here from a simple discrete model, as used by Elffers, but use instead a *continuous* distribution for the "accident proneness" A of the nurse. Statistical models with continuously varying random variables are perhaps more difficult to explain to the judges, but are often much more realistic, which should be the only important consideration here.

Secondly, the number of incidents happening to a nurse on duty depends on A and the time interval she is working, and follows (conditionally on A) a Poisson distribution. The time interval is measured by the number of shifts the nurse has had.

Assuming that A is exponentially distributed implies, among other things, that it can easily happen that one nurse has twice the incident rate of another nurse. The probability of this event is $2/3$; in fact the probability of an incidence rate of a factor k times that of another nurse is $2/(k+1)$.

The statistical problem boils down to the estimation of the parameter, characterizing the mixture of Poisson processes for the different nurses. Combining the Juliana Children's Hospital and

the two wards of the Red Cross Hospital, Lucia had 201 shifts and 14 incidents.

A major flaw in the investigation is that the data collection is irreproducible and lacks rigorous methods and definitions. It crucially depended on the memory of people who knew what was sought after. But we will argue from the data in Table 1 above, which also was used in the computations of Elffers.

We'll take the overall probability of an incident per shift to be the ratio of total number of incidents to total number of shifts, $\mu = 27/1734$. If we take a shift to be our unit time interval, then this would be a so-called *moment estimate* of the mean intensity of incidents.

This means, that, conditionally on the time interval $T = 201$, the number of incidents follows a mixture of Poisson random variables with parameter $201A$, where the intensity A has an exponential distribution with first moment μ. Thus on average, an innocent Lucia would experience $201 \cdot \mu = 201 \cdot 27/1734 \approx 3.12976$ incidents. A picture of the probabilities that the number of incidents is greater or equal to $k = 1, 2, \ldots$ is shown in Figure 3, which is based on the calculations given at the end of this section.

Heterogeneity of any kind increases the variation in the number of incidents experienced by a randomly chosen nurse over a given period of time (given number of shifts). From the well-known relations for conditional expectation (E) and variances (var)

$$\mathrm{E}(X) = \mathrm{E}(\mathrm{E}(X|Y)),$$

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

it follows that whereas for a Poisson distributed random variable variance and mean are equal, for a mixture of Poisson's (with different conditional means), the variance is larger than the mean. So if some nurses experience more or less incidents than other, in all cases the end-result is *overdispersion* caused by *heterogeneity*.

Applied to the current model which is geometric with parameter $(1 + t\mu)^{-1}$ (see the computation at the end of this section):

$$\begin{aligned}\mathrm{var}(N) &= (1 + t\mu)^2 - (1 + t\mu) \\ &= t\mu + (t\mu)^2,\end{aligned}$$

where the latter term neatly splits over the expected variance of the Poisson process plus the variance of the conditional parameter of the Poisson process which we assumed to be exponential.

The fact that a modest amount of heterogeneity turns an almost impossible occurrence into something merely mildly unusual, is strong support for further empirical research whether and if so in what forms heterogeneity plays a role in healthcare. It can have major implications in different areas, such as medical research (representing an extra source of variation) and training of medical staff.

Computation of the probabilities in the mixed Poisson model

If N is a Poisson random variable with parameter λ, the probability that the number of incidents is bigger than k, $k = 1, 2, \ldots$, is given by an integral, namely

$$\frac{1}{(k-1)!} \int_0^\lambda e^{-x} x^{k-1}\, dx, \tag{2}$$

see, e.g., [3], Exercise 46, p. 173. This means that if we assume that the "accident proneness" of the nurses has an exponential distribution with expectation μ (in our case estimated by $27/1734$) and the parameter of the Poisson distribution for the nurse is given by ta, where t is the time interval (in our case $t = 201$) and a the accident proneness, we have to integrate (2) with respect to the density of the exponential distribution with expectation μ, taking $\lambda = ta$ So we get for the probability that a nurse

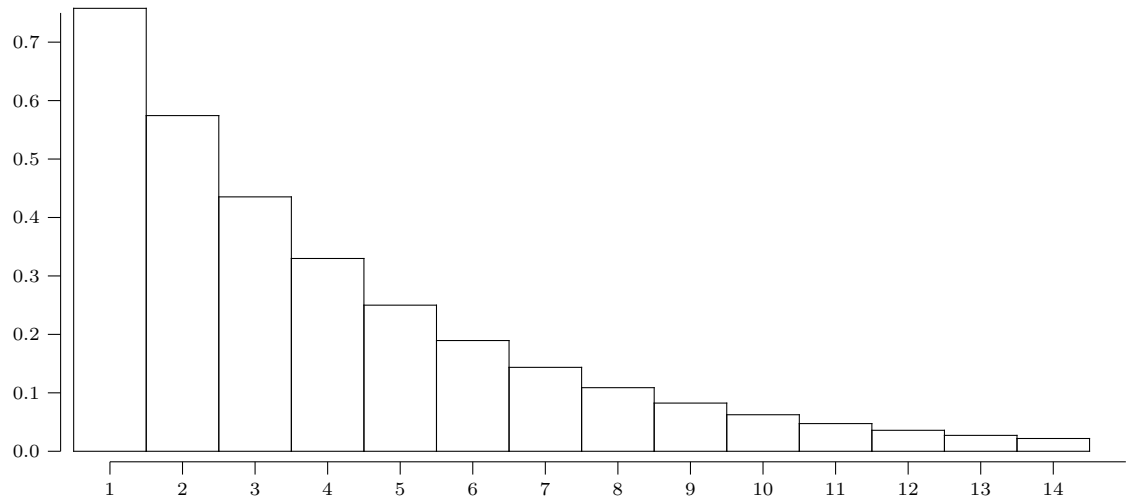

Figure 3: Probabilities (in the Poisson model) that the number of incidents in 201 shifts for one nurse is at least 1,2,3,..., if $\mu = 27/1704$. The probabilities are given by the heights of the columns above $1, 2, 3, \ldots$, respectively.

experiences k or more incidents:

$$\int_0^\infty \mathbb{P}\{I \geq k | A = a, T = t\} \frac{e^{-a/\mu}}{\mu}\, da = \int_0^\infty \left\{ \frac{1}{(k-1)!} \int_0^{ta} e^{-x} x^{k-1}\, dx \right\} \frac{e^{-a/\mu}}{\mu}\, da$$
$$= \left(\frac{t\mu}{1+t\mu} \right)^k.$$

This is the geometric distribution with parameter $1/(1 + t\mu)$. With $k = 14$ and $t\mu = 3.12976$ this yields 0.0206161 or about *one in* 49.

An early version of this paper used a revision of Elffers' dataset proposed by Professor Ton Derksen, philosopher of science, who together with his sister, medical doctor Metta de Noo, was the first to actively contest the court's reasoning in the case of Lucia de Berk.

Our model then led us to a right tail probability of *one in nine*. We later noticed that Derksen had also removed all incidents which the court finally decided not to count as provenly caused by Lucia; he used the legal argument that Elffers had previously been instructed by the judges to do the same for the data from the Juliana Children's Hospital. This does not make any statistical sense.

Going back to original medical records, Derksen and de Noo also found inconsistencies in the classification and timing of several incidents, which underlines the unreliability of the data. Correcting the data for apparent errors would also improve the results from the defence point of view.

We decided in the present paper to stick with Elffers' numbers in order to focus on our main point concerning the impact of heterogeneity.

Extended discussion of heterogeneity

We showed that a modest amount of heterogeneity leads to very different orders of magnitude in the outcomes of crucial calculations. Here we address some of the underlying mechanisms which may lead to the postulated heterogeneity.

Clearly, the data in this case show heterogeneity. The data stem from two hospitals with very different patients, young children in the JCH and elder adult patients in the RCH. The data come from three wards and the rates of incidents per shift vary considerably for each ward.

We describe two general mechanisms causing heterogeneity. The first one concerns properties of subjects directly related to the intensity of the rate of incidents. The other mechanism is more indirect and results from "spurious correlations", in which properties not related to the underlying intensity influence the measurement via unexpected dependencies and systematic variations in variables assumed to be independent and uniform.

Related to this is another aspect of the data: the degree to which a specific model or null-hypothesis is susceptible to small variations in the data. We will show this to be the case in the original calculations. Although our example is tuned to this very specific case, it refers to a much more general *caveat*. It should be established how stable certain models are under small perturbations of the data.

Are nurses interchangeable?

According to medical specialists we have spoken to, nurses are completely interchangeable with respect to the occurrence of medical emergencies among their patients. However, according to nursing staff we have consulted, this is not the case at all. Different nurses have different styles and different personalities, and this can and does have a medical impact on the state of their patients. Especially regarding care of the dying, it is folk knowledge that terminally ill persons tend to die preferentially on the shifts of those nurses with whom they feel more comfortable. As far as we know there has been no statistical research on this phenomenon.

There is another obvious way in which the intensity of incidents depends on characteristics that vary over the population. Any event that can turn out to be an "incident" starts with the call of a doctor. And in all cases it is the nurse who decides to call a doctor. This decision is influenced by professional and personal attitude, past experience and personality traits like self-confidence. It seems obvious to us that these characteristics vary greatly in any population. Hence we assume that the intensity of experiencing incidents varies accordingly.

Inadequacy of the hypergeometric distribution as a model and spurious correlations

The model underlying the null-hypothesis (which led to the hypergeometric distribution) depends on two assumptions: Both the incidents and the nurses are assigned to shifts uniformly and independently of each other.

Above we have established two ways in which characteristics of individual subjects may lead to variation in the intensity of experiencing an incident. This variation is in contrast with one of the assumptions underlying the hypergeometric distribution: uniformity.

Next we discuss sources of correlation which correspond to indirect rather than direct causation: we speak then of spurious-correlation, correlation which can be explained by confounding factors, by common causes.

There are serious reasons to doubt the uniformity of incidents over shifts. There may occur periodical differences. The population of a hospital ward may vary over the seasons. The patients may differ in character and severity of illness due to seasonal influences. There are differences between day and night shifts and between weekend shifts and shifts on weekdays. An extensive study of Dutch Intensive Care Units admissions shows a marked increase in deaths when the admission falls outside "office hours"[5]. Recall that there have to be nurses on

duty throughout the night and throughout the weekends, while the medical specialists tend to have "normal working hours". Finally there is the periodical cycle of the circadian rhythm, influencing the condition of the patients and the attention of the medical staff [4].

Notice that circadian variation in e.g. mortality and the resulting variation of incident rate between different shifts over the day interacts with the variation in the number of nurses on a shift, with more personnel on the day shifts. This can result in a higher number of nurses with an incident on their shift if the incident rate is higher during day time shifts and conversely, a lower number in the opposite case.

There may be other, non-periodical variations that affect the uniformity of incidents. In the case of the Juliana Children's hospital there has been a rather sensitive matter of policy: whether very ill children, who are not going to live for very long, should die at home or in the hospital wards. We understand that this policy did change at least once at the JCH in the period of interest. Presumably a change in policy concerning where the hospital wants children to die, will have impact on the rate of incidents. Further, incidents may be clustered, since one patient can give rise to several incidents.

On the other hand the way nurses are assigned to shifts is certainly not uniform and 'random'. Nurses take shifts in patterns, for example several night shift on a row, alternated by rows of evening or day shifts. Nurses are assigned to shifts according to skills, qualification and other characteristics. Maybe some nurses take relatively more weekend shifts than others, because of personal circumstances.

Although both the assignment of nurses to shifts and the assignment of incidents to shifts are not uniform processes, one could hope that there might be some 'mixing' condition that makes the ultimate result indistinguishable from the postulated independence and uniformity. Certainly one may hope, but this magical mechanism should at least be made plausible.

Taken together, even if we consider both the shifts of a given nurse as a random process, and the incidents on a ward as a random process, and even if we consider the two processes as stochastically independent of one another, the assumption of constant intensities of either is a guess, not based on any evidence or argument. There may be patterns in the risk of incidents and there are certainly patterns in the shifts of nurses. These patterns may be correlated, through the process by which shifts are shared over the different nurses according to their different personal situations, their different wishes for particular kinds of shifts, their different qualifications, and the changing situation on the ward.

How stable are the hypergeometric probabilities under small changes in the data?

Consider the data of the ward at JCH. These numbers and their interpretation are at the root of what turned out to be one of the gravest miscarriages of justice of the Dutch Juridical system. Under the assumption of the hypergeometric distribution the probability of this configuration is very small, less than one in nine million. The configuration is in some respects extreme: eight out of eight incidents occur in the shift of one nurse. However the data are in another respect also conspicuous: no incidents occur in the 887 shifts where this nurse was not present (see Table 1). The data collection had been far from flawless, with no formal definition of incident, no or incomplete documentation, and rested at least in part on recollection of witnesses who were aware of which facts were looked for.

Assuming the possibility of tiny flaws in the process of data acquisition, it is legitimate to investigate the effect of $1, 2, \ldots, 8$ incidents that could have been forgotten, or overlooked. This amounts to allowing a maximal error of less than one percent. The results are quite remarkable; in table 2 we give the probabilities..

Table 2: The effect of perturbations on the probabilities

Shifts with incidents outside Lucia's (postulated)	0	1	2	3	4
Probability	1/9043864	1/1137586	1/257538	1/79497	1/29989
Shifts (continued)	5	6	7	8	
Probability	1/13051	1/6329	1/3341	1/1889	

The very small numbers vanish easily. Six or more incidents not remembered, not reported, or just defined away make the difference between astronomically small on the one hand and very unusual on the other. This shows that the probabilities are very sensitive to small errors in the data.

A judgement on data quality is not only the concern of a statistician. Judges are used to inconsistent and incomplete data (statements), psychologists are very well aware of the possible fallacies of memory. Both groups have their own professional standards of how to deal with these phenomena. A statistician, however, should point out what the effects of these phenomena can be on the outcome of his models.

If this model is used to corroborate evidence this sensitivity should be made explicit, just as adverse workings of a medicine are mentioned explicitly for the users.

Concluding remarks

In the body of this paper we have shown the considerable effect that a modest amount of heterogeneity can have on tail probabilities. The broader impact of allowing heterogeneity in the analysis of (medical) research has interesting consequences outside the case of Lucia de Berk. What remains is a very short description of how the case ended in acquittal. Lucia was arrested in december 2001. As indicated in the introduction, the court (of appeal) stated that it did not include statistical considerations as basis for its verdict. This may be true for formal statistical considerations, but the essential step in the construction of the guilty verdict was that only one or two cases of murder had to be proven convincingly, the rest of the murders could be considered proven based on the "very improbable" occurrence of incidents during the shifts of Lucia. In this way statistical considerations were crucial, but the verdict was immunized against formal statistics. In this way Lucia was convicted in 2004 for seven murders and three attempts of murder. What followed was a long legal struggle where the emphasis was on the validity of the medical arguments and increasingly intricate juridical matters. The Lucia case was fiercely debated in public and the statistical notions remained here an important issue. Statisticians, now banned from the courtrooms, continued to play a role, for example by mobilizing the scientific community. Gradually the notion emerged that a gross miscarriage of justice had taken place. A complicating factor remained that, since the juridical path had been followed till the end, a new "fact", a so-called *novum* had to be found. In 2008, Lucia was allowed to wait for the end of the legal proceedings outside prison, and two years later she was finally acquitted of all murder accusations.

References

[1] H. Elffers. Distribution of incidents of resuscitation and death in the Juliana Kinderziekenhuis [Juliana Children's Hospital] and the Rode Kruisziekenhuis [Red Cross Hospital]. Unpublished report to the Court, May 29 2002. URL `http://www.math.leidenuniv.nl/~gill/Elffers2eng.pdf`.

[2] H. Elffers. Distribution of incidents of resuscitation and death in the Juliana Kinderziekenhuis [Juliana Children's Hospital] and the Rode Kruisziekenhuis [Red Cross Hospital]. Unpublished report to the Court, May 8 2002. URL `http://www.math.leidenuniv.nl/~gill/Elffers1eng.pdf`.

[3] William Feller. *An introduction to probability theory and its applications. Vol. I.* Third edition. John Wiley & Sons, Inc., New York-London-Sydney, 1968.

[4] Kuhn G. Circadian rhythm, shift work, and emergency medicine. *Ann Emerg Med*, 37:88–98, 2000. doi: 10.1067/mem.2001.111571.

[5] Hans A. J. M. Kuijsten, Sylvia Brinkman, Iwan A. Meynaar, Peter E. Spronk, Johan I. van der Spoel, Rob J. Bosman, Nicolette F. de Keizer, Ameen Abu-Hanna, and Dylan W. de Lange. Hospital mortality is associated with icu admission time. *Intensive Care Med*, online first, 2010. doi: 10.1007/s00134-010-1918-1.

[6] R. Meester, M. Collins, R.D. Gill, and M. van Lambalgen. On the (ab)use of statistics in the legal case against the nurse Lucia de B. *Probability and Risk*, 5:233–250, 2007. With discussion by David Lucy.

[7] Leila Schneps and Coralie Colmez. *Math on trial.* Basic Books, New York, 2013. ISBN 978-0-465-03292-1. How numbers get used and abused in the courtroom.