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 attempted murders, was a 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." That put statistics at center stage.

In the conviction of de Berk, a simple (so-called) hypergeometric model played an important role. The law psychologist Henk Elffers, consulted as a statistician by the court, used it, and it produced very small probabilities of occurrences of certain numbers of incidents. 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 over-simplified discrete probability model in such circumstances.

The outcomes of applying our alternative model to this case are in striking contrast with those of the first calculations that 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 de Berk's conviction is about 0.021864 or one in 46 if the calculations are based on the same data as Elffers used. In his calculation, however, this probability was equal to 1 in 342 million.

The Data

The data for this discussion are from Elffers's unpublished reports. Before going into this, some general remarks about the data collection are needed.

One of the key features of the data was the flawed 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 begins when there is suspicion of a crime; one is looking for or hunting down a suspect. If there is a need for meaningful statistics, an approach must guarantee clear definitions and uniformity of the data collection.

In the case of Lucia de Berk, this clash of cultures proved disastrous. Incidents outside de Berk's shifts were discarded, and some initially reported incidents were later relabeled without clear reasons. Extra shifts without incidents and incidents outside shifts that de Berk worked were subsequently brought to light. Moreover, the data collection rested, for a large part, on memory.

Lucia de Berk

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, a fear of implicating oneself and friends can considerably distort memory. The data on shifts and incidents for the period that were singled out in Elffers's reports are shown in Table 1 (see also Meester, Collins, Gill, and van Lambalgen).

Elffers's Method

Analysis by Elffers, the statistician consulted by the court, was based on Table 1. As was noticed later, de Berk had actually worked three shifts in RCH-41 instead of just one, but this argument is based on

Table 1—Data on Shifts and Incidents

Hospital name (and ward number)	JCH	RCH-41	RCH-42	Total
Total number of shifts	1,029	336	339	1,704
Lucia's number of shifts	142	1	58	201
Total number of incidents	8	5	14	27
Number of incidents during de Berk'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.

the data used by Elffers. 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 de Berk'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

$$\frac{\binom{S}{L}\binom{N-S}{I-L}}{\binom{N}{I}}$$

where S is the number of shifts de Berk worked and I is the total number of incidents, and where $\binom{S}{L}$, etc., denote binomial co-efficients. Taking all the data of Table 1 together gives a total number of $N = 1{,}704$ shifts. Where de Berk had $S = 201$ shifts, there was a total number $I = 27$ of incidents, and $L = 14$ incidents during de Berk's shifts. If we evaluate this equation with these values for N, S, I, and L, we get the very small probability of about 1 in 3.4 million.

Computing the probability (p-value) that a nurse is present with 14 or more incidents in Elffers's 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) requires summing the probabilities for $L = 14, 15, \ldots, 27$, for the probability of about 1 in 3 million. For the model introduced in the next section, however, which used the same data, there is a probability of 1 in 46.

Elffers proceeded somewhat differently—not combining the data of the different hospitals (the details of what he actually did are described in Meester, Collins, Gill, and van Lambalgen). The most-important mistake he made in his calculation was to take the three hospitals separately, and multiply 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 incidents 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 number of incidents at the Juliana Children's Hospital and the Red Cross Hospital was equal to 1 in 342 million. (Meester, Collins, Gill, and van Lambalgen, "Math error number 7: the incredible coincidence" [Schneps and Colmez]).

Post-hoc Testing

There is a danger of post-hoc testing: testing a hypothesis using the same data that suggested that hypothesis. Elffers actually tried to take this problem into account by starting 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 de Berk's shifts by 26—the number of nurses in that period who worked on the same ward.

By the time he looked at the data from RCH, de Berk 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 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 approach would be determining a reasonable model for number of incidents if de Berk 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.

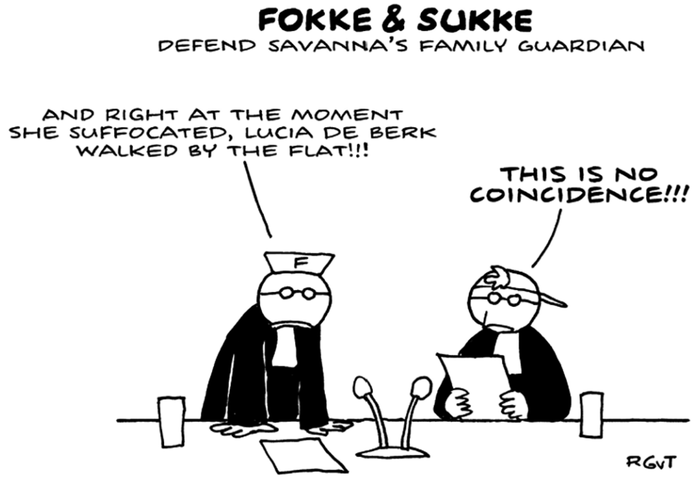
FOKKE & SUKKE
DEFEND SAVANNA'S FAMILY GUARDIAN
AND RIGHT AT THE MOMENT SHE SUFFOCATED, LUCIA DE BERK WALKED BY THE FLAT!!!
THIS IS NO COINCIDENCE!!!
F
RGvT

Figure 1. A Fokke & Sukke cartoon from October 30, 2007, in the Dutch newspaper *NRC Next*. (Translated into English by the creators of the cartoon, Reid, Geleijnse, and Van Tol). Lucia de Berk was still in prison at that time. The two ducks 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 Elffers told the court: "Honored court, this is no coincidence. The rest is up to you."

An Alternative Model

The incidents that a nurse experiences can be modeled by a so-called Poisson process, with a nurse-dependent intensity A, using A for "accident proneness." A Poisson process is used to model incoming phone calls during non-busy hours, residence 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 might be 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, 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 a 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, de Berk 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. We will argue from the data in Table 1, however, which was also used in Elffers's computations.

If the overall probability of an incident per shift is the ratio of total number of incidents to total number of shifts, μ = 27 / 1,704. If a shift is the 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 de Berk would experience $201 \cdot \mu = 201 \cdot 27 / 1{,}704 \approx 3.18486$ incidents. A picture of

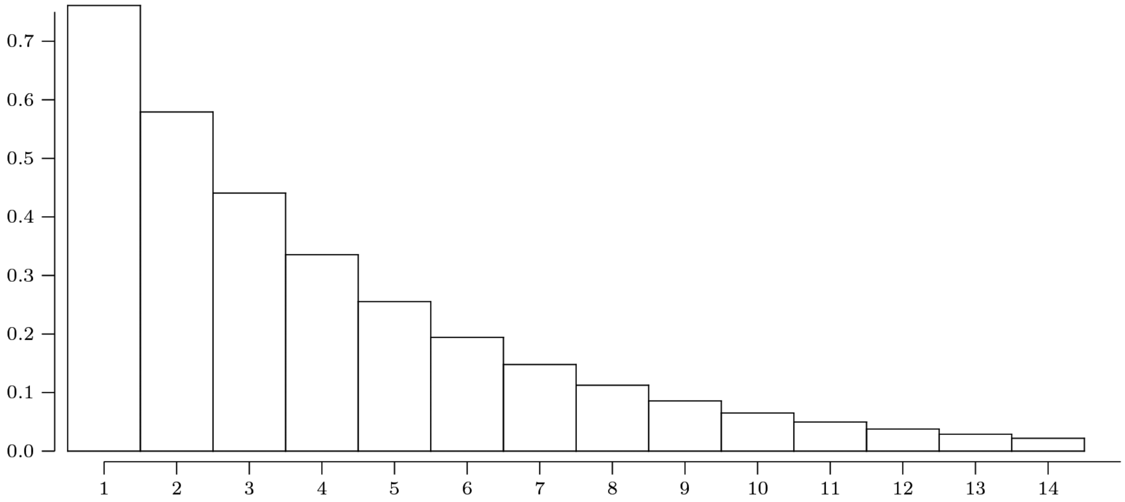

Figure 2. Probabilities (in the Poisson model) that the number of incidents in 201 shifts for one nurse is at least 1, 2, 3, … , if $\mu = 27 / 1{,}704$. The probabilities are given by the heights of the columns above—1, 2, 3…, respectively.

the probabilities that the number of incidents is bigger than $k = 1, 2, \dots$ is shown in Figure 2, which is based on the calculations 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)

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

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

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. If some nurses experience more or less incidents than others, the end result in all cases is over dispersion 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):

$$\mathrm{var}(N) = (1 + t\mu)^2 - (1 + t\mu)$$
$$= t\mu + (t\mu)^2;$$

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 is 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 on 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 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, $k = 1, 2, \dots$, is given by an integral, namely

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

(see, e.g., Feller, 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 / 1,704) 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 is the accident proneness, we have to integrate with respect to the density of the exponential distribution with expectation μ, taking $\lambda = ta$. For the probability that a nurse experiences more than k incidents, that results in:

$$\int_0^\infty \mathbb{P}\{I \geq k | A = a,$$

$$T = t\}\,\frac{e^{-a/\mu}}{\mu}\,da =$$

$$\int_0^\infty \left\{\frac{1}{(k-1)!}\int_0^{ta} e^{-x}x^{k-1}\,dx\right\}$$

$$\frac{e^{-a/\mu}}{\mu}\,da = \left(\frac{t\mu}{1+t\mu}\right)^k.$$

This is the geometric distribution with parameter $1/(1 + t\mu)$. With $k = 14$ and $t\mu = 3.18486$, this yields 0.0218641 or about 1 in 46.

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

Our model led us to a right tail probability of 1 in 9. We later noticed that Derksen had also removed all incidents that the court finally decided not to count as provably caused by de Berk; 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 defense point of view.

We decided to stick with Elffers's numbers here to focus on our main point concerning the impact of heterogeneity.

Extended Discussion of Heterogeneity

As shown, a modest amount of heterogeneity leads to very different orders of magnitude in the outcomes of crucial calculations. Some underlying mechanisms 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 older adult patients in the RCH. The data come from three wards, and the rates of incidents per shift vary considerably for each ward.

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 that are 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. This was 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 people tend to die preferentially on the shifts of those nurses with whom they feel more comfortable, although 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 a call to 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 such as self-confidence. It seems obvious that these characteristics vary greatly in any population, so 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.

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.

Sources of correlation can correspond to indirect rather than direct causation; spurious correlation can be explained by confounding factors and common causes.

There are serious reasons to doubt the uniformity of incidents over shifts. Periodical differences may occur. 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" (see Hans, Kuijsten, Brinkman, Meynaar, et al.)—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 (see Kuhn).

Notice that circadian variation in areas such as 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 daytime shifts and, conversely, a lower number in the opposite case.

There may be other, nonperiodical 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 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 an 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 shifts in a row, alternated by rows of evening or day shifts. Nurses are assigned to shifts according to skills, qualifications, and other characteristics. Some nurses might take relatively more weekend shifts than others because of personal circumstances.

Although the assignment of both nurses to shifts and 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. This magical mechanism should at least be made plausible.

Taken together, even if we consider both the shifts of a given nurse and the incidents on a ward as random processes, and 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 With Small Changes in the Data?

Consider the data for 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 in the Dutch judicial system. Under the assumption of the hypergeometric distribution, the probability of this configuration is very small; less than 1 in 9 million.

The configuration is, in some respects, extreme: Eight out of eight incidents occur during 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. The data collection had been far from flawless, with no formal definition of incident, and 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, . . . , 8 incidents that could have been forgotten or overlooked. This amounts to allowing a maximal error of less than 1 percent. The results are quite remarkable; Table 2 shows the probabilities.

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 quite sensitive to small errors in the data.

A judgment 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 for 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.

Table 2—The Effect of Perturbations on the Probabilities

Shifts with incidents outside Lucia de Berk's (postulated)	0	1	2	3	4
Probability	1/9,043,864	1/1,137,586	1/257,538	1/79,497	1/29,989
Shifts (continued)	**5**	**6**	**7**	**8**	
Probability	1/13,051	1/6,329	1/3,341	1/1,889	

Concluding Remarks

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 beyond the case of Lucia de Berk. What remains is a very short description of how the case ended in acquittal.

de Berk was arrested in December 2001. The court (of appeals) stated that it did not include statistical considerations as a 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 when de Berk was working. In this way, statistical considerations were crucial, but the verdict was immunized against formal statistics. de Berk 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 de Berk case was fiercely debated in public and the statistical notions remained 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 judicial path had been followed until the end, a new "fact"—a so-called *novum*—had to be found.

In 2008, de Berk was allowed to wait for the end of the legal proceedings from outside prison, and two years later, she was finally acquitted of all murder accusations.

Further Reading

Elffers, H. May 8, 2002. 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. *http://www.math.leidenuniv.nl/~gill/Elffers1eng.pdf*.

Elffers, H. May 29, 2002. 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. *http://www.math.leidenuniv.nl/~gill/Elffers2eng.pdf*.

Feller, William. 1968. *An introduction to probability theory and its applications*, Vol. 1, Third edition. New York-London-Sydney: John Wiley & Sons, Inc.

Hans, A.J., Kuijsten, M., Brinkman, Sylvia, Meynaar, Iwan A., et al. 2010. Hospital mortality is associated with ICU admission time. *Intensive Care Medicine*, doi: *10.1007/s00134-010-1918-1*.

Kuhn, G. 2000. Circadian rhythm, shift work, and emergency medicine. *Annals of Emergency Medicine*, 37:88{98. doi: *10.1067/mem.2001.111571*.

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

Schneps, L., and Colmez, C. 2013. *Math on trial. How numbers get used and abused in the courtroom*. New York: Basic Books.

About the Authors

Richard Gill is emeritus professor of mathematical statistics at Leiden University. He is interested in statistics and the law, as well as statistics in quantum physics, survival analysis, and semiparametric models.

Piet Groeneboom is emeritus professor of statistics of Delft University of Technology. His present research focuses on shape-restricted inference and computer algorithms for non- and semiparametric statistical models.

After finishing his PhD thesis on central limit theorems with applications in Malliavin calculus, **Peter de Jong** started his company on internet publishing. He taught statistics for actuaries. He is also editorial advisor to the scientific publisher Ios Press on the role of statistics in their portfolio.