

On the (ab)use of statistics in the legal case against the nurse Lucia de B.

Ronald Meester*, Michiel van Lambalgen†, Marieke Collins‡, Richard Gill§

May 3, 2006

Abstract

We discuss the statistics involved in the legal case of the nurse Lucia de B. in The Netherlands, 2003-2004. Lucia de B. witnessed an unusual high number of incidents during her shifts, and the question arose as to whether this could be contributed to chance. We discuss and criticise the statistical analysis of Henk Elffers, a statistician who was asked by the prosecutor to write a statistical report on the issue. We discuss several other possibilities for statistical analysis. Our main point is that several statistical models exist, possibly leading to very different predictions, or perhaps different answers to different questions. There is no such thing as a ‘best’ statistical analysis.

1 Introduction; the case

In The Hague (The Netherlands), on March 24, 2003 the nurse Lucia de B. (hereafter called either ‘Lucia’ or ‘the suspect’) was sentenced to life imprisonment for allegedly killing or attempting to kill a number of patients in two hospitals where she had worked in the recent past: the *Juliana Kinderziekenhuis* (JKZ) and the *Rode Kruis Ziekenhuis* (RKZ). At the RKZ, she worked in two different wards, numbered 41 and 42 respectively. At the JKZ, an unusually high proportion of incidents occurred during her shifts,¹ and the

*Vrije Universiteit Amsterdam

†Universiteit van Amsterdam

‡Universiteit Utrecht

§Universiteit Utrecht

¹The precise technical definition of ‘incident’ is not important here; suffice it to say that an incident refers to the necessity for reanimation, regardless of the outcome of the reanimation.

question arose as to whether Lucia's presence at so many 'incidents' could have been accidental.

A statistical analysis was given by statistician Henk Elffers, who had looked into the matter at the request of the public prosecutor. In broadest terms, his conclusion was this: assuming only (as he says) that

1. the probability that the suspect experiences an incident during a shift is the same as the corresponding probability for any other nurse,
2. the occurrences of incidents are independent for different shifts,

then the probability that the suspect has experienced as many incidents as she in fact has, is less than 1 over 342 million. According to Elffers, this probability is so small that standard statistical methodology sanctions rejection of the nullhypothesis of chance. He did take care to note that in itself this does not mean the suspect is guilty.

Why do we write this article? Two of us (MvL and RM) became involved in the case as expert witnesses of the defense. We studied the method and conclusion of Elffers, and came to the conclusion that his numbers did not mean very much, if anything at all. Elffers (and the court, for that matter) completely overlooked the subjective element in the choice of a probabilistic model, and therefore the possibility of there being several models with very different predictions, or perhaps different answers to different questions!

The question as to how to use statistics in a case like this is not a question with a well-defined answer. Borrowing a phrase of Thomas Kuhn, we deal here with a *problem* rather than with a *puzzle*. There are many ways of doing statistics. One can argue whether to use a (subjective) Bayesian approach, or a classical frequentist approach. There is even a school called the likelihood approach which says that you should compute and report likelihood ratios, full stop. Within each school there can be many solutions to what appears to be the same problem. Moreover there is the question of the range of the model.

Hence, many different approaches are possible, using very different models, and with many different levels of sophistication. One can choose a very simple model, as Elffers did, giving precise results, albeit in a very limited context. One can also choose a much broader perspective, like a Bayesian point of view, which involves much more data, but whose conclusions are very much imprecise. There simply is no unique best way of dealing with the problem, and in this paper we want to elaborate this point significantly. In court, the judges continued to ask us: "So if you reject Elffers' numbers, why don't you give us better numbers", implicitly assuming that there exist

something like best numbers. One of the points of the present article is to argue against this.

This article is structured as follows. We will first present the relevant data and the statistical methodology used by Elffers. We discuss and criticise this methodology on several levels: not only do we offer a critical discussion of his overall approach, but we also think that within his paradigm, Elffers made several important mistakes. We also briefly discuss the way the court interpreted Elffers' report. Then we show how the method of Elffers could have been used in a way we believe is correct within his chosen paradigm, leading to a somewhat different conclusion. After that, we discuss a Bayesian point of view, as advocated by the Dutch econometrist De Vos, and then we move on to the so called epidemiological models, inspired by recent work of Lucy and Aitken. In the final section we try to formulate some conclusions.

2 The data and Elffers' method

Elffers tried to base his model entirely on data pertaining to shifts of Lucia and the other nurses, and the incidents occurring in those shifts. The data on shifts and incidents for the period which was singled out in Elffers' report are given in the following table:

hospital name (and ward number)	JKZ	RKZ-41	RKZ-42
total number of shifts	1029	336	339
Lucia's number of shifts	142	1	58
total number of incidents	8	5	14
number of incidents during Lucia's shifts	8	1	5

Later it was discovered that Lucia actually had done 3 shifts in RKZ-41 instead of just 1, and in our own computations later in this article, we will use this correct number.

When trying to put the situation sketched into a statistical model, one's first choice might be to build a model on the basis of epidemiological data concerning the probability of incidents during various types of shifts; this would allow one to calculate the probability that the suspect would be present accidentally at as many incidents as she in fact witnessed.

However, the trouble with this approach is that for the most part the requisite data are lacking. And even if the data were available, their use would be a subject of debate between prosecutor and defense; see Section 7.

Because of this, Elffers tried to set up a model which uses only the shift data given above. This he achieved by *conditioning* on part of the data. He assumed that

1. there is a fixed probability p for the occurrence of an incident during a shift (hence p does not depend on whether the shift is a day or a night shift, etc.),
2. incidents occur independently of each other.

It is now straightforward to compute the *conditional* probability of the event that (at the JKZ, say) all incidents occur during Lucia's shifts, *given* the total number of incidents and the total number of shifts *in the period under study*. Indeed, if the total number of shifts is n , and Lucia had r shifts, then the conditional probability that Lucia witnessed x incidents given that k incidents occurred, is

$$\frac{\binom{r}{x} p^x (1-p)^{r-x} \binom{n-r}{k-x} p^{k-x} (1-p)^{n-r-k+x}}{\binom{n}{k} p^k (1-p)^{n-k}} = \frac{\binom{r}{x} \binom{n-r}{k-x}}{\binom{n}{k}}. \quad (1)$$

Note that this quantity does not depend on the unknown parameter p . This distribution is known as the *hypergeometric* distribution. With this formula, one can easily compute the (conditional) probability that the suspect witnessed at least the number of incidents as she actually has, for each ward.

However, according to Elffers, this computation is not completely fair for the suspect. Indeed, the computation is carried out only because of the bare fact that there were so many incidents during her shifts. It would, therefore, be more reasonable (according to Elffers) not to compute the probability that Lucia has witnessed so many incidents, but instead the probability that *some* nurse witnessed so many incidents. At the JKZ, there were 27 nurses taking care of the shifts and therefore Elffers multiplies his outcome by 27; he calls this the *post hoc correction*. According to Elffers, this correction only needs to be done at the JKZ; at the RKZ this is no longer necessary since the suspect was already identified as being suspect on the basis of the JKZ data.

Elffers arrives at his final figure (the aforementioned 1 in 342 million) by simply multiplying the outcomes for the three wards (with post hoc correction at the JKZ, but without this correction at the RKZ).

3 Discussion of Elffers' method

There are a number of problems and points of concern with the method of Elffers. In the following, we list some of these.

3.1 Conditioning on part of the data

As we remarked already, conditioning on the number of incidents has a big advantage, namely that under the hypothesis of chance, the unknown parameter p cancels in the computations. It is the very conditioning that makes computations possible in Elffers' model.

The idea of conditioning at inference time on quantities that were not naturally fixed at data sampling time has some history. It seems that Fisher first proposed this idea for exact inference on a 2×2 contingency table [5]. In [8], some justification is offered for this technique. Conditioning is reasonable, according to Mehta and Patel, if "the margins [...] are representative of nuisance parameters whose values do not provide any information about the null hypothesis of interest." In the current case however, it is debatable whether or not the number of incidents does not provide information. Intuitively, it should make a big difference whether the number of incidents during the time span of interest is very high or very low, compared to other periods (provided the incident-rate is constant in time). In other words, it seems to us that although in itself sometimes being defensible, conditioning is only allowed if one realises that one possibly discards relevant information. In other words, the limitations of such an approach should be made abundantly clear. Elffers, however, is silent on this matter. We elaborate on this issue in Section 7 below, where we will back up our criticism with some computations.

3.2 Using data twice: the post hoc correction

One of the problems with this approach is the fact that the data of the JKZ is used twice. First to identify the suspect, and after that again in the computations of Elffers' probabilities. This procedure should raise eyebrows amongst statisticians: it is one of these problems that seem to arise all over the place: one sets up an hypothesis on the basis of certain data, and after that one uses the same data to test this hypothesis. It is clear that this raises problems, and it is equally clear that Elffers' method shares this problem. In a way, Elffers seems to be aware of this. After all, his post hoc correction was introduced for exactly this reason.

However, this post hoc correction is a striking instance of an unacknowledged subjective choice employed by Elffers. To see this, first note that Elffers restricts the statistical analysis to the wards at which the suspect worked. Why? Indeed, the question of the prosecutor had to do with the question whether or not Lucia actually killed her patients. In the formulation of this question, the word ‘ward’ does not appear, and this means that Elffers himself decided to just consider the level of wards. We do not claim or think that this decision was wrong; there are good arguments to defend it, the most important one probably being the simplicity of the resulting model. But one can also envision a statistical analysis of *all* wards in, say, The Netherlands, perhaps with different probabilities for incidents in different wards. When we now condition on the number of incidents in each ward, then again the number of incidents of the suspect has the same hypergeometric distribution as before. However, a posthoc correction in this hypothetical statistical analysis would have to incorporate *all* nurses in The Netherlands. In this hypothetical statistical analysis, the computations concerning the suspect would still only depend on the data of her own ward. Hence, multiplication by the number of nurses in the ward of Lucia, does not necessarily follow from the fact that we only use data from her own ward; the level of the posthoc correction is quite arbitrary.

An analogy might clarify this point. Consider a lottery with tickets numbered 1 to 1.000.000. The jackpot falls on the ticket with number 223.478, and the ticket has been bought by John Smith. John Smith lives in the DaCostastraat in the city of Leiden. Given these facts we may compute the chance that John Smith wins the jackpot; a simple and uncontroversial model shows that this probability will be extremely small. Do we conclude from this that the lottery was not fair, since an event with very small probability has happened? Of course not. We can also compute the probability that someone in the DaCostastraat wins the jackpot, but it should be clear that the choice of the Da Costastraat as reference point is completely arbitrary. We might similarly compute the probability that someone in Leiden wins the jackpot, or someone living in Zuid-Holland (the state in which Leiden is situated). With these data-dependent hypotheses there simply is no uniquely defined scale of the model at which the problem must be studied.

The analogy with the case of Lucia will be clear: the winner of the jackpot represents the suspect being present at 8 out of 8 incidents, the street represents the ward. Elffers restricts his model to the ward in which something unusual has happened. With perhaps equal justification, another statistician might have considered the entire JKZ (Leiden, in the analogy) instead of the ward as basis for her computations – with vastly higher prob-

ability for the relevant event to happen somewhere. Still another statistician might have taken the Netherlands as the basis for the computation, which yields again a higher probability. The important point to note is that *subjective choices are unavoidable here*; and it is rather doubtful whether a court's judgement should be based on such choices. If one wants to avoid these kind of subjective choice, one should adopt an approach where the data is not used twice. In the next section we discuss such an approach.

One more remark on the issue: although Elffers' posthoc correction is not totally without rationale, its use is badly flawed since we know the nurses have widely varying numbers of shifts, many of them have rather small numbers of shifts. If we had done Elffers' posthoc correction using each nurses individual incident rate rather than raw number of incidents, we would have got a much less small probability than Elffers did. Perhaps the correct computation here is even a nice weapon for the defense.

3.3 Multiplication is not allowed

Elffers multiplies the three probabilities from the three wards. The multiplication means that he is assuming that under his null-hypothesis, incidents occur completely randomly in each of the three wards (as far as the allocation of shifts to nurses is concerned), independently over the wards, but with possibly different rates in each ward. If one accepts his earlier null-hypothesis as an interesting hypothesis to investigate, then this new hypothesis could also be of interest.

What is the meaning of the probability which Elffers finds? It is the probability, under this null-hypothesis of randomness, and conditional on the total number of incidents in each ward, that a nurse with as many shifts as Lucia in each ward separately, would experience as many (or more) incidents than she did, in all wards simultaneously. Is the fact that this probability is very small, good reason to discredit the null hypothesis?

First we should understand the rationale of Elffers' method when applied to one ward. He is interested to see if a certain null-hypothesis is tenable (whether his null-hypothesis is relevant to the case at hand, is another matter). He chooses in advance for whatever reason he likes, a statistic (a function of the data) such that large values of that statistic would tend to occur more easily if there actually is a, for him, interesting deviation from the null-hypothesis. Since his null-hypothesis completely fixes the distribution of his chosen statistic, he can compute the probability that the actually observed value could be equalled or exceeded under that hypothesis. The resulting probability is called the p -value of the statistical test. If the null-

hypothesis and the statistic had been chosen in advance of collecting the data, then it becomes hard to retain belief in the null-hypothesis if the p -value is very small. Elffers in fact follows the following procedure: he has selected (arbitrarily) a rather small threshold, $p = 0.001$. When a p -value is smaller than 0.001 he will declare that the null-hypothesis is not true. Following this procedure, and in those cases when actually the null-hypothesis was true, he will make a false declaration once in a thousand times.

If the null-hypothesis corresponds to a person being innocent of having committed a crime, then his procedure carries the guarantee that not more than one in a thousand innocent persons are falsely found guilty. (Presumably, society does accept some small rate of false convictions, since absolute certainty about guilt or innocence is presumably an impossibility. But perhaps one in a thousand is a bit too large a risk to take).

Now we return to Elffers' multiplication of three p -values, one for each ward. Does this result in a new p -value?

An easy argument shows that the probability as computed is in itself pretty meaningless. Suppose there are 100 wards and the null-hypothesis is true (including the independence over the wards). A nurse with the same number of shifts as Lucia in each ward has approximately a probability of a half to have as many incidents as Lucia, in each ward separately. Multiplying, the probability that she "beats" Lucia in all wards is approximately 1 in 2 to the power one hundred, or approximately one in a million million million million million. Yet we are assuming the complete randomness of incidents within each ward! Clearly we have to somehow discount the number of multiplications we are doing.

Is there something else that Elffers could have done, to combine the results of the three wards? Yes; and in fact, classical statistics offers many choices. For instance he could have compared the total number of incidents of Lucia over the three wards, to the probability of exceeding that number, given the totals per ward and the numbers of shifts, when in each ward separately incidents are assigned uniformly at random over all the shifts. In the language of statistical hypothesis testing, he should have chosen a single test-statistic based on the combined data for his combined null-hypothesis, and computed its p -value under that null-hypothesis. Perhaps it would be reasonable to weight the different wards in some way. Each choice gives a different test-statistic and a different result. The choice should be made in advance of looking at the data, and should be designed to react to the kind of deviation from the null-hypothesis which it is most important to detect. Such a choice can be scientifically motivated but it is in the last analysis subjective.

An easy way to combine (under the null-hypothesis) independent p -values is a method due to Fisher (and can be found in his book [5]: multiply (as Elffers did) the three p -values for the separate tests (denoted by p_1, p_2 and p_3), and compare this with the probability distribution of the product of the same number of uniform random numbers between 0 and 1. A standard argument from probability theory reduces this to a comparison of $-2 \sum_i \log p_i$ with a chi-squared distribution with $2n$ degrees of freedom. What is in favour of this method is its simplicity. Choosing this one is just as much a subjective choice as any other.

3.4 The Quine-Duhem problem

In fact, talk of ‘the rejection of the nullhypothesis’ is somewhat imprecise. It was observed by the philosopher Quine, and before him by the historian of science Duhem, that the falsificationist picture of an hypothesis H logically implying a prediction P , which when falsified must lead to the abandonment of H , is too simplistic.

Consider the following example. Suppose our thermodynamic theory implies that water boils at 100C at sea level; and suppose furthermore that our observations show water to boil at 120C. Does this mean thermodynamics is false? Not necessarily, because there might be something wrong with the thermometer used. That is, the logical structure of the prediction is rather

‘Thermodynamics + Properties of thermometer’ imply ‘water boils at 100C at sea level’.

More formally, a prediction P from an hypothesis H always has the form $H \& A_1 \& \dots \& A_n \Rightarrow P$, where the $A_1 \& \dots \& A_n$ are the auxiliary hypotheses. If we find that P is false, we can conclude only not- $(H \& A_1 \& \dots \& A_n)$ from which something can be concluded about H only if we have independent corroboration of the $A_1 \& \dots \& A_n$. The same phenomenon occurs in statistics, and in particular in this case.

In order to be able to make calculations, we have to make several assumptions which go much beyond the nullhypothesis of interest, namely that the probability of an incident is independent of the presence of Lucia. This nullhypothesis is compatible with the following auxiliary hypotheses, all of which directly contradict the assumptions behind Elffers’ model:

- the probability of an incident during a night shift is larger than during a day shift (more people die during the night);

- the probability of an incident during a shift depends on the prevailing atmospheric conditions (which may have an effect on respiratory problems);
- the occurrence of an incident in shift $n + 1$ is not independent on the occurrence of an incident in shift n (a successful reanimation in shift n may be followed by death in shift $n + 1$);
- there is no a priori reason to assume that all nurses have equal probability to witness incidents; for instance, as our own informal inquiries in hospitals have shown, terminally ill patients often die in the presence of a nurse with whom they feel ‘comfortable’.

This is just a small sample of the different auxiliary hypotheses that could be envisaged. The main point is this: only if the auxiliary hypotheses used in setting up the model closely mirror reality can the occurrence of an improbable outcome be used to cast doubt on the nullhypothesis. In the absence of such independent verification of the auxiliary hypotheses, the occurrence of an improbable outcome might as well point to a silly choice of the model.

4 The court’s interpretation of Elffers’ numbers

In its judgement of March 24, 2003, the court glossed Elffers’ findings as follows (the numbering corresponds to the court’s report)²:

7. In his report of May 29, 2002, dr. H. Elffers concludes that the *probability* that a nurse *coincidentally* experiences as many incidents as the suspect is less than 1 over 342 million. (emphasis added)

8. In his report of May 29, 2002, dr. H. Elffers has further calculated the following component probabilities

a. The *probability* that one out of 27 nurses would *coincidentally* experience 8 incidents in 142 out of a total of 1029 shifts ... is less than 1 over 300.000.

b. The *probability* that the suspect has *coincidentally* ...

(emphasis added)

²The original Dutch version can be found at www.rechtspraak.nl

11. The court is of the opinion that the probabilistic calculations given by dr. H. Elffers in his report of May 29, 2002, entail that it must be considered *extremely improbable* that the suspect experienced all incidents mentioned in the indictment *coincidentally*. These calculations *consequently show* that it is *highly probable* that there is a *connection between the presence of the suspect and the occurrence of an incident*. (emphasis added)

We have cited these excerpts from the court’s judgement because the italicized phrases should raise eyebrows among statisticians. The judgement of the court is ambivalent, and it is unclear whether or not the court makes the famous mistake known as the *prosecutor’s fallacy*. Clearly, one *should* talk about the probability that something happened, under the assumption that everything was totally random. The judgement of the court could however also be interpreted as the probability that something accidentally happened. This is quite different, as is easily illustrated with the following formal translation into mathematical language.

Writing E for the observed event, and H_0 for the hypothesis of chance, Elffers calculated $P(E | H_0) < 342 \cdot 10^{-6}$, and the court (perhaps) concluded that $P(H_0 | E) < 342 \cdot 10^{-6}$. Writing

$$P(H_0 | E) = \frac{P(E | H_0) \cdot P(H_0)}{P(E)},$$

we see that prior information about $P(H_0)$ and $P(E)$ is required to come to any conclusion.

We would like to note that Elffers did not make this mistake himself, but during the testimony of two of the authors of this article (RM and MvL), the court of appeal certainly did.

5 Elffers’ method revised

There is an obvious way to revise Elffers’ method in a way that avoids the scale problems and the double use of data. Indeed, since the incidents at the JKZ were the ones that attracted attention to the suspect, we should ignore these incidents in the computations. Note that this does not necessarily lead to throwing away information: it is precisely due to the incidents at the JKZ that we perform computations at the RKZ at all.

Doing similar computations as Elffers, but now restricted to the RKZ and without any correction, we clearly obtain very different numbers. If we

first take the data of the two wards together, then we have a total number of 675 shifts, Lucia having 61 of them (note the correction of numbers). There were 19 incidents, 6 of which were during one of Lucia’s shifts. Under the same hypothesis as Elffers, a similar computation now leads to a probability of 0.0038, which of course is much and much larger than the number obtained by Elffers. In particular, Elffers himself used a significance level of 0.001, meaning that in this case the null hypothesis should *not* be rejected, in sharp contrast to Elffers’ conclusion.

However, it is perhaps more reasonable to make a distinction between the two wards. There are several ways of dealing with this. One could combine p -values as in Section 3.3, or, alternatively, treat both wards independently with the hypergeometric method as Elffers, and ask for the probability that the sum of two independent hypergeometric random variables (with their proper parameters) exceeds 6. A simple computation leads to the conclusion that this probability is equal to 0.022, still bigger than the previously found 0.0038.

It is clear that some of the aforementioned problems remain in this revised form of the method. Nevertheless, we believe that the revised form is an improvement, since we get rid of a post hoc correction and the double use of data. Within the approach chosen by Elffers, this seems to us the only way to avoid arbitrary and subjective corrections. Doing things right here is in fact a useful argument for the defense.

6 A Bayesian approach to the problem

During and after the trial, a public debate arose in The Netherlands about the way statistics was used in this case. Apart from Henk Elffers and two of the authors of this article, also Aart de Vos, an econometrist, entered the discussion. De Vos claimed that a Bayesian approach would solve all scale problems; see [9]-[10]. In a national newspaper, he came to the conclusion that Lucia was *not* guilty with probability at least 10%, a number in sharp contrast with Elffers’ outcomes. We summarise his method here, without going into details.

A Bayesian analysis works as follows. Let E denote the evidence at hand, H_d the null hypothesis (the hypothesis that L is innocent), and H_p denote the alternative hypothesis (the hypothesis that L is guilty).

A straightforward application of Bayes’ rule now gives

$$\frac{P(H_p|E)}{P(H_d|E)} = \frac{P(E|H_p)}{P(E|H_d)} \cdot \frac{P(H_p)}{P(H_d)}.$$

In (other) words,

$$\text{posterior odds} = LR \cdot \text{prior odds}.$$

We interpret $P(H_d|E)$ as the probability of H_d after evaluating the evidence E . The posterior odds are - at least in theory - nice to work with, because any new evidence (E_{new}) can be implemented to give new posterior odds. For example, suppose we first had

$$\text{“old” posterior odds} = \frac{P(E|H_p)}{P(E|H_d)} \cdot \frac{P(H_p)}{P(H_d)},$$

then, after this new evidence, we get new posterior odds:

$$\begin{aligned} \frac{P(H_p|E, E_{\text{new}})}{P(H_d|E, E_{\text{new}})} &= \frac{P(E_{\text{new}} \cap E|H_p)}{P(E_{\text{new}} \cap E|H_d)} \cdot \frac{P(H_p)}{P(H_d)} \\ &= \frac{P(E_{\text{new}}|H_p, E)}{P(E_{\text{new}}|H_d, E)} \cdot \text{“old” posterior odds}. \end{aligned}$$

This is all nice in theory, but the questions that arises once you try to use this in a law suit are obvious: can we make sense of $P(H_p)$ and $P(H_d)$? For what kind of evidence it is possible to compute $\frac{P(E|H_p)}{P(E|H_d)}$? And can we make sense of $\frac{P(E_{\text{new}}|H_p, E)}{P(E_{\text{new}}|H_d, E)}$? The latter question is particularly challenging, because it is difficult to see how the different pieces of evidence are related.

In the case at hand, the following facts were brought up by De Vos as relevant evidence. After each piece of evidence we write between parantheses the likelihood ratio for that piece of evidence as used by De Vos.

1. E_1 ; the fact that the suspect never confessed ($\frac{1}{2}$);
2. E_2 ; the fact that two of the patients had certain toxic substances in their blood (50);
3. E_3 ; the fact that 14 incidents were reported during Lucia’s shifts (7,000);
4. E_4 ; the fact that suspect had written in her diary that ‘she had given in to her compulsion’ (5).

It seems obvious to us that these facts are hardly, if at all, expressable as numbers. The numbers of De Vos can hardly be motivated. The prior probability $P(H_p)$ is taken to be 10^{-5} , and then finally, De Vos assumes independence between the various facts, and ending up with posterior odds equal to roughly 8,75. This means that suspect is guilty with probability close to 90%, certainly not enough to convict anybody.

6.1 Discussion

The numbers obtained by De Vos are in sharp contrast with Elffers' outcomes. However, we find it hard to believe that anybody would take any of these numbers seriously. It is clear from the analysis that his priors and likelihood ratios are very, very subjective. Any change in his priors would lead to very different answers. More generally, Bayesian approaches would require judges to give their priors in order to motivate their verdicts. It is unclear what the role of the defense would be in this situation: can they reasonably object to the judges' subjective priors? The attempt to use every piece of evidence in the statistics, has to lead to meaningless results.

7 An epidemiological approach

In [2] and [3], Lucy and Aitken discuss a totally different way of modelling cases like this, and we include a discussion of their method here. This method does not rely on conditioning on the number of incidents, but instead on epidemiological data. Even though this data is not available in our case, it is of considerable interest to us. Indeed, we can investigate what would have happened, in case this data would have been available. Moreover, it is possible to make a reasonable guess for the missing data. As we will see, different assumptions on the missing data leads to radically different conclusions, hereby providing some evidence for the statement that Elffers should have informed the court that his method lacks this data, and should therefore have been viewed with serious scepticism.

The basic assumption of Lucy and Aitken is that the probability distribution of the number X of incidents witnessed by a certain nurse, is given by a Poisson distribution, hence

$$P(X = k) = e^{-\mu r} \frac{(\mu r)^k}{k!},$$

where r is the number of shifts of the nurse, and $\mu > 0$ is a parameter.

The argument why a Poisson distribution might be reasonable here, is that under the same assumptions as Elffers, the total number of incidents follows a binomial distribution with low 'succes' probability. It is well known that such a binomial distribution is well-approximated by a Poisson distribution.

We take care to note that this Poisson model is connected to Elffers' conditional approach. If we adapt the model of Lucy and Aitken and subsequently condition on the number of incidents, then we are led (under the

hypothesis of chance) to the hypergeometric distribution in Elffers' computation. In this sense, Elffers model is even more general than the model of Lucy and Aitken. We also remark that the current Poisson model also suffers from the problem mentioned in Section 3.4; also in this model, we cannot distinguish between for instance day and night shifts.

The hypothesis of chance can now be formulated as saying that every nurse, including the suspect, has the *same* parameter μ .

The hypothesis H_p of the prosecutor can have several forms. Perhaps the first hypothesis that comes to mind is $H_p : \mu_L > \mu$, where μ_L is the parameter corresponding to the suspect, and μ is the parameter corresponding to all other nurses. However, under this composite hypothesis, one cannot perform computations, and in the sequel we will consider several explicit values of μ_L instead.

To be more explicit, consider a situation with I nurses, and let k_i be the number of incidents witnessed by nurse i , $i = 1, \dots, I$. Denote by r_i the number of shifts of nurse i , and let E be the event that nurse i witnessed k_i incidents, for $i = 1, \dots, I$. This leads to

$$P(E|H_d) = \prod_{i=1}^I e^{-\mu r_i} \frac{(\mu r_i)^{k_i}}{k_i!},$$

and, assuming that the suspect is nurse j , to

$$P(E|H_p) = \frac{e^{-\mu_L r_j} (\mu_L r_j)^{k_j}}{k_j!} \prod_{i=1, i \neq j}^I e^{-\mu r_i} \frac{(\mu r_i)^{k_i}}{k_i!}.$$

A simple computation that shows that the likelihood ratio becomes

$$\text{LR} = \frac{P(E|H_p)}{P(E|H_d)} = e^{\mu r_j - \mu_L r_j} \left(\frac{\mu_L r_j}{\mu r_j} \right)^{k_j}. \quad (2)$$

In order to evaluate the outcome of any computation with this likelihood ratio, we may use the following scale for describing the height of a likelihood ratio, see [4]:

	evidence is
$LR = 1$	equally likely under H_p as under H_d
$1 < LR < 100$	slightly more likely under H_p than under H_d
$100 \leq LR < 1000$	more likely under H_p than under H_d
$1000 \leq LR < 10,000$	much more likely under H_p than under H_d
$LR > 10,000$	very much more likely under H_p than under H_d .

As was noted by Meester and Sjerps in [6] and [7], one should be extremely careful when using a table like this, since the likelihood ratio only becomes an absolute meaningful number in combination with the prior odds of the two competing hypotheses. For now, we do not worry too much about this point, and we shall try to make a few realistic assumptions for doing computations. We concentrate on the RKZ, for the same reasons as before. In the following computations, we take the data of the two wards at the RKZ together. One probably feels that we should allow different incident rates between different wards; in that case the numbers would come out even better for the suspect.

I: At first sight, a reasonable assumption for the prosecutor is to estimate μ by the incidents during shifts of all nurses, apart from the suspect. However, this choice is biased, since when we throw away the incidents and shifts of the nurse with the highest incident-rate, we might end up with a rate which is too low. This is unfavourable for the suspect. Nevertheless, this choice leads to

$$\mu = \frac{13}{614}.$$

Lucy and Aitken proceed by choosing μ_L in such a way that the expected number of incidents witnessed by the suspect is precisely k_j , that is, $\mu_L r_j = k_j$ hence

$$\mu_L = \frac{6}{61}.$$

These assumptions lead to a likelihood ratio of 90.7, and this is in the range where the evidence is only slightly more likely under H_d .

II: If we would estimate μ based on all incidents, we would get

$$\mu = \frac{19}{675}$$

and this would lead to a likelihood ratio of about 25 (keeping μ_L as above).

III: To see why it is important to realise that the lacking data could play a very important role, consider the hypothetical situation in which data outside the observed period would have led to an estimate of $\mu = 7/60$. This means that although in the observed time frame, Lucia had a remarkably high number of incidents relative to the other nurses, the number was actually quite normal when considering a larger time frame. Needless to say that this information should be helpful for the suspect. Indeed, doing the computation in this case leads to a likelihood ratio of only 1.1.

7.1 Discussion

It is clear that the drawbacks of the conditional approach of Elffers become quite apparent here. Of course, when the necessary data is not available, or not very reliable one has to be careful with drawing any conclusion. However, even though it is unfavourable for the suspect to assume that the ‘ordinary’ incident rate is estimated by the incident rate of the other nurses in the relevant time frame, this choice already leads to conclusions that differ from those of Elffers. It seems however, that the results here are comparable to a correct use of Elffers’ method, see Section 5.

Comparing the epidemiological approach to the approach of Elffers, we notice one remarkable feature. In Elffers’ approach, the actual probability of an incident was irrelevant, his method only compares the number of incidents witnessed by the suspect, to the incidents witnessed by the other nurses. In the approach of Lucy and Aitkin, it is almost the opposite: it is irrelevant how many incidents the other nurses witnessed, at least if the estimate for μ is obtained from data outside the observed time frame. Intuitively, one feels that a combination of these aspect might be preferable. This combination is achieved in the notion of *relative risk*, the subject of the next paragraph.

8 Relative risk

In [2] and [3], Lucy and Aitken define the term *relative risk* as follows: the relative risk R_j of a nurse j is the fraction of her shifts during which an incident took place, divided by the fraction of the remaining shifts during which an incident took place. More formally,

$$R_j = \frac{k_j/r_j}{\sum_{i \neq j} k_i / \sum_{i \neq j} r_i}.$$

For example, the relative risk of Lucia for the RKZ for the two wards together is equal to

$$\frac{\frac{6}{61}}{\frac{13}{614}} \approx 4.65.$$

The fact that Lucia had the highest relative risk is clearly not enough to warrant any investigation; some nurse must have the highest relative risk. The more important question is how high a relative risk should be in order to be suspicious.

The distribution of the highest relative risk depends on many variables, like the number of nurses, the way the shifts are spread among the nurses,

the number of shifts, and of course also on the modelling assumptions themselves. In this section we again concentrate on the model of Lucy and Aitken of the previous section.

The numbers in the definition of the relative risk only depend on the considered time span, comparing the amount of incidents (s)he witnessed to the amount of incidents the other nurses witnessed. But when adding the parameter μ , which one should preferably base on the number of incidents divided by the number of shifts *outside* the considered time span, it also becomes important what the long-time average number of incidents is. This means we have combined the use of data from both Elffers' procedure and the one of Lucy and Aitken in the previous section.

It is now useful to do some numerical simulations to obtain some idea about the distribution of this highest relative risk.

8.1 Simulating relative risk

We have no data concerning the number of shifts of all other nurses, apart from Lucia. Therefore, we have simulated a situation where all nurses have done the same number of shifts. In order to stay as close to reality as possible, we note that in the RKZ we have a total of 675 shifts of which Lucia did 61 (note the remark after the table of data). Therefore we simulated a situation in which 11 nurses all did 61 shifts. Hence $r_i = r$ for all i and $I = n/r$. This leads to

$$R_j = \frac{k_j}{\sum_{i=1}^I k_i - k_j} (I - 1).$$

We are interested in the nurse with the highest relative risk for each group of I nurses. Since all nurses work the same number of shifts, this is simply the nurse with the most incidents.

We have run 1000 simulations in the case of Lucia for the data of the RKZ, first for both wards together, then for each ward separately. Here are the results. The values for μ in the first column are based on the frequency of incidents of all other nurses in the RKZ; the values of μ in the second column are based on the overall frequency of incidents, including Lucia, and the values in the third column are just rather arbitrary high numbers.

whole RKZ	$\mu = \frac{13}{614}$	$\mu = \frac{19}{675}$	$\mu = \frac{7}{60}$
Lucia's p -value	0.121	0.042	0
RKZ-41	$\mu = \frac{4}{333}$	$\mu = \frac{5}{336}$	$\mu = \frac{1}{4}$
Lucia's p -value	0.787	0.681	0
RKZ42	$\mu = \frac{9}{281}$	$\mu = \frac{14}{339}$	$\mu = \frac{1}{10}$
Lucia's p -value	0.383	0.286	0.031

8.2 Discussion

For $\mu = \frac{13}{614}$, L's relative risk of approximately 4.65 lies between the 879th and 880st of the 1000 highest relative risks. In other words, it is a high relative risk, but not extremely high. For $\mu = \frac{19}{675}$, L's relative risk lies between the 958th and the 959th highest relative risks and for μ even higher, for example $\mu = \frac{7}{60}$, L's relative risk is larger than the largest of the highest simulated relative risks.

From this, we may conclude that if data on the number of incidents outside the time span L worked at the RKZ would indicate μ to be large, L's relative risk would be extremely high and this could be used as evidence against her in court. This may seem strange, since in the likelihood ratio approach of the previous section, a larger μ implied a *lower* likelihood ratio, which is in favor of the defendant. The fact that a large μ does not work in favour of the defendant in the relative risk approach, is because if μ really is very large, then we do not expect much spread in the the relative risks of the nurses. If the total number of incidents is coincidentally very small, then the relative risks will be widely spread. So under the hypothesis that all the nurses are the same, whenever the total number of incidents is very small, we are more likely to see a very extreme relative risk and hence (incorrectly) reject the hypothesis. The defense could object strongly to the use of this procedure; all the more so since we don't really know the value of μ , and we don't really know if it is constant in time.

9 Conclusion

It is not easy to draw a clearcut conclusion from all this. For one thing, one can say that Elffers' methodology has a rationale in this case: had Elffers' computation led to the conclusion that in his model chance would have been a reasonable guess, then it seems that there would not have been a case against Lucia de B. at all.

In this case, however, Elffers' numbers did raise interest, and there should

have been reflection on what to do next. It is clear that had Elffers used his model correctly, that is, without double use of data, without some arbitrary posthoc correction, and without his unallowed multiplication, then the resulting numbers would have been different. In fact, the outcome would not have led to the rejection of the null hypothesis of chance, with significance level of 0.001 (Elffers' own choice), although the reported probabilities of 0.0038 and 0.022 (see Section 5) would have been uneasily low for sure.

Following an epidemiological approach does not really lead to a different conclusion. The likelihood ratios of 90.7 and 25 reported in Section 7 would in itself not lead to conviction, but are again somewhat uneasily high. Similar remarks apply to the relative risks in Section 8. It seems that whatever method one uses, the conclusions are more or less the same.

The weakness of Elffers' approach can also be seen as its strength. The model can be criticised (and should be criticised) but once we have the model, there are no further parameters to be worried about, and which could lead to disagreement between the prosecutor and the defense. This does not mean, clearly, that the model of Elffers should be seen as the final word on the issue.

The Poisson model of Lucy and Aitken suffers from the fact that any conclusion by either party can be questioned by the other on the basis of the choice of the parameter μ . If one of the parties can raise reasonable doubts about the validity or reasonableness of the parameter choice, then the numbers arising from that model can be questioned as well. This being said, an analysis as carried out in Section 7 does show that the ingredients missing in Elffers' model, could be very important, and Elffers should have been aware of this.

The more sophisticated a model becomes, the more possibilities for criticising it one has. This becomes abundantly clear in the Bayesian approach of De Vos in Section 6. De Vos tries to incorporate everything into his mathematical model. To us, this seems impossible, and the result of the computations of De Vos do not mean much, if anything at all.

The big question remains whether or not statistics can play a role in cases like this. It is clear that one should be extremely careful. As we mentioned in the introduction, we clearly have a problem rather than a puzzle here. Every statistician makes choices, has to make choices: it seems almost impossible to produce an uncontroversial number. In fact, perhaps the only uncontroversial number in this article are the numbers in Section 5. Indeed, these numbers do not suffer from double use of data or scaling problems, and do not involve any parameter choice either. Perhaps these numbers are - after all - the only possible contribution of statistics to the

present case. This statement might be surprising, given the available data. But it is one thing to say that a number is relevant and should be used. It is quite another thing to work out a reasonable way to use it. Not all numbers can or even should be used in a statistical fashion.

On June 18, 2004, the court of appeal in The Hague again found Lucia de B. guilty and sentenced her to life imprisonment plus detention in a psychiatric hospital in case she would ever be pardoned. This time the judgement made no mention at all of statistical arguments; evidence which had played a secondary role during the first trial now assumed primary importance. Hence for several reasons this was a Pyrrhic victory at most (at least for the authors). If the court of appeal had explicitly repudiated the form of statistical argument employed by the public prosecutor and the first court, future cases would have been able to use this jurisprudence.

However, incorporating the statistical argument in the second judgement would have required the court of appeal to take an explicit stand on all the issues raised above. In fact, careful writers on the foundations of statistics have pointed out that evaluating a statistical conclusion involves even more:

In applying a particular technique in a practical problem, it is vital to understand the philosophical and conceptual attitudes from which it derives if we are to be able to interpret (and appreciate the limitations of) any conclusions we draw. ([1], page 332)

Evidently the court of appeal was not willing to dig this deep; but the quote as well as the case of Lucia de B. may serve as a reminder to lawyers and judges that the interpretation of statistical arguments is by no means immune to disputation.

References

- [1] V. Barnett, *Comparative Statistical Inference*, Wiley (1999).
- [2] D. Lucy and C. Aitken, *A review of role of roster data and evidence of attendance in cases of suspected excess death in a medical context*, *Law Probability and Risk* **1**, 61-160 (2002).
- [3] D. Lucy and C. Aitken, *The evidential value of roster and attendance data in cases where fraud or medical malpractice may be suspected*, preprint.

- [4] I.W. Evett, G. Jackson, J.A. Lambert and S. McCrossan, *The impact of the principles of evidence interpretation on the structure and content of statements*, *Science & Justice* **40**, 233-239 (2000).
- [5] R.A. Fisher, *Statistical methods for research workers*, Oliver and Boyd, Edinburgh (1925).
- [6] R. Meester and M. Sjerps, *The evidential value in the DNA database controversy and the two-stain problem*, *Biometrics* **59**, 727-732 (2003).
- [7] R. Meester and M. Sjerps, *Why the effect of prior odds should accompany the likelihood ratio when reporting DNA evidence*, *Law, Probability and Risk* **3**, 51-62 (2004)
- [8] C.R. Mehta and N.R. Patel, *Exact inference for categorical data*, www.cytel.com/Papers/sxpaper.pdf (1997).
- [9] A.F. de Vos, *A primer in Bayesian Inference*, preprint.
- [10] A.F. de Vos, *Informative priors and Empirical Bayesian inference*, preprint.