Piet Groeneboom
Delft Institute of Applied Mathematics
TU Delft
p.groeneboom@tudelft.nl

Column Piet takes his chance

The Statistics Scene

Piet Groeneboom regularly writes a column on everyday statistical topics in this magazine.

As a starting columnist I am suddenly confronted with a lot of choices to make. For example, in telling a story, giving the reader a feel of the statistics scene, should I mention names or should I avoid that? I also have to avoid 'name dropping', which means that an older scientist tries to impress younger scientists with all the famous people he/she is acquainted with and whose names might mean nothing to the persons he/she addresses: "Peter A. told me..., and yes he had a point... On the other hand, Jerry said...

The additional danger of using names is that people are getting afraid of telling you something. "OK, I will tell you about our investigation of the confidence intervals in our COVID-19 research, but only if you promise not to make a mean joke about it in your column!"

Concluding this initial digression: I will use names, unless it would put the person in a negative light. And I will say "the statistician/probabilist..." to identify them for the reader. Some of the people in my story are on the picture below.

The Dutch statistics scene used to be dominated by professor Willem van Zwet, who passed away last year. Nowadays Aad van der Vaart (for a photo, see [7]) has perhaps taken over this torch. There were festivities in Leiden on the occasion of his 60th birthday where he became Knight in the Order of the Dutch Lion. On Wednesday that week there was a very pleasant dinner party at the Hortus in Leiden, where I was also invited and shared a table with a couple of prominent statisticians and Aad's wife and son, see the picture.

Moulinath Banerjee, a statistician on the opposite of the table from me, asked a waiter to take a picture of us. On the left of the table behind our table we see the 'rock star in statistics' Bradley Efron, 'the inventor of the bootstrap'. I'll come to speak about the bootstrap later. The organizers called him 'rockstar' (following a Dutch habit of combining separate words to one word, see [8] — at the time of this writing), but I read on internet: "rockstar is an energy drink, a video-game company, and a really terrible song by Nickelback which no one would ever want to be compared to, associated with, or forced to listen to." So I use two words, because I do not want Brad to be associated with these things. I was told that a famous saying of Bradley Efron is that he needed a whole year

Picture taken at the dinner in Hortus Leiden in 2019. Counterclockwise from the left: Piet Groeneboom, Maryse Loranger (wife of Aad van der Vaart), Pascal van der Vaart (son of Aad), Vera Wellner (wife of Jon Wellner), Moulinath (Mouli) Banerjee, Jon Wellner, Susan Murphy, Marloes Maathuis. On the left at the next table: Brad Efron, the 'rock star in statistics'.

to 'unlearn' mathematics, coming out of college (Caltech), meaning that only after a year of unlearning he was able to work on the really relevant things in statistics.

A pressing question (often) is: we have probability theory and statistics, what is the difference? At the mathematics departments of Dutch universities there used to be — apart from other chairs — two chairs, a chair for probability theory and a chair for mathematical statistics. In the eighties of the preceding century, when I became professor of mathematical statistics at the University of Amsterdam, there were big cuts in budget, which actually continued to affect me (us) during my career at the university. As a consequence I had meetings where I had to explain (unsuccessfully) to the pure mathematicians the difference between probability theory and mathematical statistics. They wanted to reduce the two chairs to one chair. I was even told that I got the position in Amsterdam because they thought I could do both chairs.

It may therefore be amusing to recall events taking place at the Mathematical Sciences Research Institute (MSRI) in Berkeley in 1991, where this issue was discussed at the beginning of a lecture I had to deliver there. As a matter of fact, I had prepared a speech on these events in Berkeley for Aad's dinner party (see above), but did not deliver my speech at the appropriate moment, although I was very much encouraged to do so by the statistician Marloes Maathuis — sitting to the left of me at the picture of the dinner table ("Piet, this is your chance!") — and probabilist/statistician Eric Cator (not on the photo). On my way to the Hortus from the Leiden train station I had even practiced my speech sitting down on a Leiden cafe terrace so that I could deliver it from memory.

One of our great novelists, Simon Vestdijk, has a passage in a book where he says: "My decisive error was not to drink (alcohol) before the event" (which therefore did not develop as hoped for). I guess I made the same mistake at the dinner party, staying completely sober.

But at a similar event (retirement dinner of the Dutch/British statistician Richard Gill in Leiden) I actually was not (completely) sober and I was at that time motioned to sit down during my speech by a prominent female member of the present Leiden statistics scene. This time, being completely sober, I was very reluctant to see the same scene of being motioned to sit down develop again. I abandoned therefore my plan to speak.

In the US it was and perhaps still is the habit to start a lecture with an awful joke or at least some appetizers, and so I started my lecture in Berkeley at MSRI by saying that many probabilists say that statistics is applied probability, in particular probabilists looking for a job, but that I did not agree with this point of view. For one thing because the motivation is very different for the members of the two groups. Just after having said this, Aad van der Vaart who was in the audience asked loudly: "When did you discover this, yesterday?" What could I say? I had my usual blackout, trying to think about the meaning of this interruption. So there was a short silence, after which Aad added for reasons again not clear to me: "Of course, I am ten years younger than you!" Actually, I found this rather amusing, because I knew that the distance was in fact larger than that. But it reminded me of an earlier event in Tashkent or rather Uzbekistan. I must confess that it was perhaps not very logical to think of that event, but neither was the succession of the two interruptions (I think), unless there is some higher point of view uniting the two interruptions.

The first World Congress of the Bernoulli Society was held in Tashkent in 1986 and there had (somewhat half-heartedly) been some social events scheduled, one of them being the non-supervised climbing of a mountain near Tashkent. This type of activity has some appeal to me, so I joined a small group, containing Aad van der Vaart and the specialist in the theory of extreme values Laurens de Haan, to climb this mountain.

For some reason I was the only person to reach the top of the mountain that day. In preparation of my speech in Hortus Leiden I checked my recollection of the event with Aad at the dinner party and he remembered it very clearly. In my climbing to the top I had released rocks (by which 'rock star' gets yet another meaning) and Aad had been forced to go downhill with his face to the top to keep an eye on the rocks I had set into motion in climbing to the top. I just learned this at the dinner party from Aad and intended to include it in my speech. Some people even interpreted this mountain event as my attempt to kill the competition! So Aad's second interruption at my MSRI lecture was possibly inspired by his anger about it. In our next climb of a mountain he would be first on the top!

After my lecture at MSRI Joel Zinn, an American probabilist who was in the audience, asked me whether Aad was my 'shill'. He did not use the word shill (which I learned later from the statistician Jon Wellner, also on the picture), but asked whether this was a prearranged act. This sounded to me like an attractive interpretation of what happened, and since that time I have been very prone to having 'shills' in the audience, willingly or unwillingly. For example, the probabilist/statistician Eric Cator was my shill in Seattle, in the so-called 'prelim' course, preparing students to the preliminary exam for being allowed to write a PhD (we have no such thing in the Netherlands). If your shill in the audience shouts "Nonsense!" just after you made some statement, the students immediately wake up and you have their undivided attention.

So, ..., I now described the gist of my intended speech at the dinner party, which would in particular have been interesting for Aad's wife and son, sitting to the right of me on the photo of the dinner party. I did not expect to be able to do that! So at least one good thing (for me) came out of becoming a columnist of NAW.

And finally: what is the difference between probability theory and mathematical statistics? The answer is very simple: probability theory works with *one* probability measure, statistics deals with a whole family (usually uncountable) of probability measures. This distinction leads to totally different methods one has to use. To illustrate this, I continue where I left off in my first column [6].

The distribution of the incubation time of COVID-19

The column [6] was about the estimation of the distribution of the incubation time of COVID-19. It was written in Dutch. This as a consequence of a misunderstanding. It was because my predecessor Casper Albers had written his last columns in Dutch, although he had started in English. A colleague in Rome wrote to me that he understood 80% using Google Translate. If I also want to reach people like him, it is better to write in English.

The Dutch/British statistician Richard Gill asked me to write a more technical paper on the same subject for the Dutch journal *Statistica Neerlandica*, which publishes in English. I actually did that, although another colleague told me that publishing in *Statistica Neerlandica* meant that my paper was lost for science (since *Statistica Neerlandica* is often not available in foreign libraries). But I realized that, by an agreement between Wiley and Delft University of Technology, my paper would have 'open access' on internet, and that therefore the 'lost for science' matter might

not be too serious. Richard is associate editor of the journal. *Statistica Neerlandica* has a publishing system run by Wiley, which means that they prefer manuscripts written in MS Word and handle manuscripts via a firm in India which does indeed not understand TeX, let alone BibTeX. Anyway, it has (after a rather difficult communication with the firm in India) been published now [5] and shows that the typical rate for the nonparametric density estimator is $n^{2/7}$ in a continuous version of the model, if n is the sample size. I think this result is new. As spelled out in the Appendix of [5], it hinges on the computation of an adjoint in an (infinite dimensional) Hilbert space and there is no explicit solution one can use. One has in fact to solve an integral equation numerically. I really wished that we could get to this result in a simpler way, but this is where we stand right now. It uses a local version of a theory on differentiable functionals, initiated by a paper of Aad van der Vaart [10].

Fortunately, the operators on the Hilbert space have an interpretation as conditional expectations, which facilitates their computation. Another interesting aspect is that the limit distribution of the full nonparametric maximum likelihood estimator in this model is still unknown, but that we can nevertheless derive the limit distribution of the density estimator based on it. I wonder whether I still will be alive when finally these insights will reach the community of epidemiologists and medical statisticians.

The Dutch/South African probabilist Guus Balkema noticed that the bandwidths I was using for the SMLE (Smoothed Maximum Likelihood Estimator) and the density estimator, based on the nonparametric MLE (Maximum Likelihood Estimator), were 3 and 4, corresponding to the formulas (3) and (4) in my column, respectively. He said: "So you advocate the nonparametric MLE, because you do not need any parameters there, but do you choose the bandwidth parameters on the basis of the numbers of your formulas?" An astute remark, which launched me on the automatic bandwidth selection theory in this case. I find this issue rather fascinating and have some partial solutions in [5], but the matter is still not completely solved. I use the bootstrap for this (not Efron's original bootstrap), but I may say more on this in a column later this year.

I also got a very interesting referee report which I cannot resist citing from. The referee says: "This is a technical follow-up of an earlier contribution in Dutch by the same author. In that Dutch paper, the author makes several remarks that suggest a limited knowledge where mathematical epidemiology of infectious diseases is concerned. That is no problem for a column-like contribution in Dutch, but should be remedied in a formal publication like the one envisaged here. The author should study the literature and place his contribution in the proper context."

Ignorance is bliss... The good news is that the referee allows me to reveal my ignorance in my columns for NAW ("no problem"). Unfortunately, the referee does not say what all this relevant literature is that I should have cited. It would mean that perhaps someone has now derived the limit distribution of the MLE in the model above? Or that someone else derived the $n^{2/7}$ rate of the density estimate based on it and determined the (normal) limit distribution? All this is very relevant for the confidence intervals which certainly will be too narrow if one uses the parametric models, apart from the fact that the corresponding estimators of the density will be inconsistent.

Another thing of interest is that the associate editor did not know who this referee was; he originally thought that it was the person whom he had asked to do the refereeing, but since this person first declined, the referee was chosen by AI (Artificial Intelligence) and his/her name was hidden from him. There is of course the suspicion that there is a connection with the RIVM, but this is only speculation. It was indeed my goal to reach the RIVM (the Centre for Infectious Disease Control and Prevention of the National Institute for Public Health and the Environment, this full name was one of the benefits of the refereeing of my paper [5]), but this has been an utterly unsuccessful endeavour so far. It might not get better after my present column.

The corresponding author of [1] never answered my questions about the difficulties with the R scripts on the accompanying site. The (lack of) information on the computation of the reproduction number in [9] is still unchanged. At a request on the COVID-19-mod forum, I translated Mathematica files for an epidemiological model into C++ and from there to R, using Rcpp. This helped of course to brush up my "limited knowledge where mathematical epidemiology of infectious diseases is concerned". I also bought and read the nicely written book [3], but this is not really very up-to-date on the statistical treatment of these problems (if I may say so).

It surprised me that many epidemiologists seem to use Mathematica for simulations. Mathematica is great and I use it a lot, but not for simulations. In fact, the simulations in the Mathematica files became much faster in my C++ implementation. However, my translation of the Mathematica files is now sitting for seven months in a private repository of my GitHub site [4] because there were (if I understood correctly) plans to write a paper on it which did not materialize, in spite of the original enthusiasm about this ("Super!"). Hopefully there will be some movement in all this before the next pandemic starts raving around the earth.

References

1 Jantien A. Backer, Don Klinkenberg and Jacco Wallinga, Incubation period of 2019 novel coronavirus (2019-nCov) infections among travellers from Wuhan, China, 20–28 January 2020, *Euro Surveill.* 25 (2020).

2 Tom Britton and Gianpaolo Scalia Tomba, Estimation in emerging epidemics: bases and remedies, *J. R. Soc. Interface* 16 (2019).

3 Odo Diekmann, Hans Heesterbeek and Tom Britton, *Mathematical Tools for Understanding Infectious Disease Dynamics*, Princeton University Press, 2013.

4 Piet Groeneboom, Incubation time, https://github.com/pietg/incubationtime, 2020.

5 Piet Groeneboom, Estimation of the incubation time distribution for covid-19, *Statistica Neerlandica,* 2020.

6 Piet Groeneboom, Nederland in tijden van corona, *Nieuw Archief voor Wiskunde* 5/21 (2020), 181–184.

7 Leiden University, Symposium Aad van der Vaart, 2019, https://www.universiteitleiden.nl/en/news/2019/06/symposium-aad-van-der-vaart.

8 Leiden University, Statistical rockstar, 2019, https://www.universiteitleiden.nl/en/news/2019/06/statistical-rockstar-brad-efron-opens-leiden-statistics-centre.

9 RIVM, De zorg voor morgen begint vandaag. Rekenmodellen openbaar en toegankelijk, 2020, https://www.rivm.nl/coronavirus-covid-19/hoe-berekeningen-bijdragen-aan-bestrijding-van-virus/rekenmodellen.

10 A.W. van der Vaart, On differentiable functionals, *Ann. Statist.* 19 (1991), 178–204.