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

Column Piet takes his chance

Two cultures

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

Cultural differences

C. P. Snow wrote the essay *The Two Cultures*. I learn from Wikipedia (to which I also contribute from time to time) that it first existed in the form of a talk delivered the 7th of May 1959 in the Senate House, Cambridge, and that it was subsequently published as *The Two Cultures and the Scientific Revolution*.

This column is about the cultures of mathematical statistics and of medical statistics. It actually is my impression that people from the first category study papers written by people from the second category, but that the converse does not necessarily hold. But medical statisticians might profit from the newer insights of the mathematical statisticians.

There is always a danger that the phenomenon of 'looking down on' or the suggestion of it enters the picture. Pure mathematicians, looking down on applied mathematicians and vice versa (for different reasons). Mathematicians looking down on other sciences ("physicists do not know what a proof is", "social science is no science"). We could even look down on our earlier selves at a time where we did not understand things which we now cannot imagine not understanding.

It is true that if a pure mathematician says something about statistics, I do not expect it to be very sensible, but the pure mathematician would not expect me to say something very sensible about his/her particular subject. There is just all this specialized knowledge one needs to understand what the current problems in the field are. The whole 'looking down on' is a kind of herd mentality (perhaps apart from the looking down on oneself).

Unfortunately, often oracles in the fields are cited to get verdicts on matters these oracles (and the persons who cite them) know nothing about. All this is to explain that it is not my goal to bash medical statistics, only to suggest that other methods (from mathematical statistics) could be used (being close enough to the field).

I was primarily interested in how the so-called reproduction number was estimated. But since I could not find concrete information on how the RIVM (the Centre for Infectious Disease Control and Prevention of the National Institute for Public Health and the Environment) did this, I first focused on the estimation of the distribution of the incubation time about which (after some effort) I could find some more information.

Most of the literature on the estimation of the distribution of the incubation time can be found in journals of medical (not mathematical) statistics or just medical journals. These papers, unlike papers in mathematics, most of the time have many authors, and are often structured in the following way. First there is a summary of what colleagues in medical statistics or epidemiology have been doing. Then there is a summary of how data were collected, accompanied by multicolored pictures, meant as visualizations of the data. Next there might be some report on the use of simple parametric distributions (Weibull, gamma, log-normal, Erlang) in the estimation procedure. And then there might be some 'ancillary material', which has to be downloaded separately, and which might contain some formulas which are not included in the main body of the text. The ancillary material may even contain the real data.

If the estimates of the Weibull, gamma, log-normal and Erlang distributions are roughly the same, this fact is (mis)used to argue that these estimates are sound. But what if they are all far off the mark? And are these distributions really so different?

Let's take a look now at a paper of the above type with nine authors: [7]. Its objective is (citing directly from the paper): "To estimate the length of the incubation period of COVID-19 and describe its public health implications." It is indeed of the structure, described above. There is a multicolored picture on p. 3 (Figure 1) for the data on 181 cases. There are no formulas in the paper. There are references to what their direct colleagues did, for example the reference to [1] (which I discussed in my first column in this magazine). And, after some searching, I discovered the ancillary material in the 'Reproducible Research Statement' after the acknowledgements: "Statistical code and data set: available at [8]." This sounded promising: data on 88 subjects for the estimation of the distribution of the incubation time in [1], but data on 181 subjects would be available in [9]! Finding the right data set was again

(as with [1]) not entirely trivial, but anyway, in the end I could analyze these data with my own methods. Actually, to save the reader all this time-consuming searching, I put the data analyzed in [1] (and again analyzed by myself in [3]) into the main body of my text in [3].

Although the authors of [7] refer to [1], the model is different. The difference is that it is not assumed that we have exact data on when a person becomes symptomatic, but only have an interval in time for when this happened. In fact, in [1] there is also an interval, but this is only the interval of one day, whereas in [7] the interval can be 81 days. We have the following model.

There is an interval $[E_L, E_R]$ for the infection time and an interval $[S_L, S_R]$ for the time of becoming symptomatic. In the models used one can, just as in [3], shift the data in such a way that $E_L = 0$, which leaves us with three numbers: the time E ('Exit time' in the case of the data in [1], telling us when the person left Wuhan) and the times S_L and S_R, adapted for the shifting of E_L to zero. Denoting the time of becoming symptomatic by S, we have that S is the sum of the infection time I and the incubation time U. We assume, conditionally on the exit time E, that I and U are independent and also that the time of becoming infected is uniformly distributed on the interval $[0, E]$.

Now S can in fact lie in the interval $[0, E]$ (the person becomes symptomatic before the exit time E). In this case there is overlap between the intervals $[0, E]$ and $[S_L, S_R]$. Or we have that $E < S_L$ and then the incubation time U bridges the distance between the point $I \in [0, E]$ (infection time) and the point $S \in [S_L, S_R]$ (time of becoming symptomatic). It is very interesting that, in spite of the fact that we do not have a direct observation of I or U (we only know that their sum $S = I + U$ — time of becoming symptomatic — lies in some interval and that $I \leq E$), we can nevertheless estimate the density of the incubation time pretty well, *at least, if we use the right method!*

On the topic of interval censored data (and that is what we have here) there is a large literature in mathematical statistics, but this literature is usually completely ignored by (or unknown to?) the medical statisticians.

The model used by [7] is called the model for *doubly censored data*. I must add that this is how it is called by medical statisticians. In the world of mathematical statistics there is a completely different model which is also called the model for doubly censored data, on which there is the magnificent paper [6], but I'll stick to the use of this terminology in the world of medical statistics.

The analysis in [7] is based on [9] and uses the `R` package `coarseDataTools`, which is also created (among others) by Nicholas Reich and Justin Lessler, two of the nine authors of [7].

Now I would like to comment on the medical statistics paper [9]. This paper discusses both the model used in [1] and the model used in [7]. They call the model in [1] *single interval censored*. The log likelihood of a sample in the single interval censored case is of the form

$$\sum_{i=1}^{n} \log\{G(S_i) - G(S_i - E_i)\},$$

where n is the sample size, S_i the time of becoming symptomatic, E_i the exit time for the ith person, and G the (unknown) distribution function of the incubation time we want to estimate; $G(u) = 0$ if $u \leq 0$ (see also [4]).

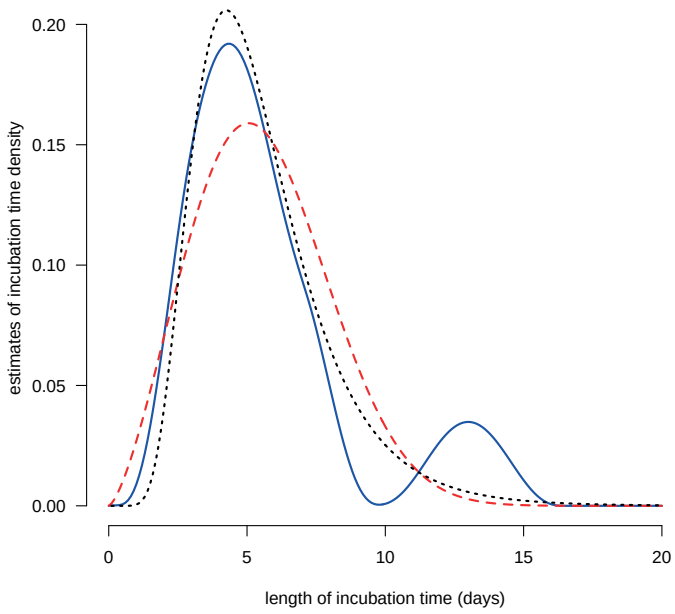

Figure 1 Estimates of the density of the incubation time for the data in [7], using bandwidth $h = 3.5$ for the nonparametric estimate. The red dashed curve is the Weibull estimate of the density for the data in [7], the black dotted curve the log-normal estimate of the density and the blue solid curve is the nonparametrics estimate. Sample size is $n = 181$.

In the *doubly censored* case, though, the log likelihood is of the form

$$\sum_{i=1}^{n} \log[\mathbb{G}(S_{R,i}) - \mathbb{G}(S_{R,i} - E_i) - \mathbb{G}(S_{L,i}) + \mathbb{G}(S_{L,i} - E_i)],$$

where $\mathbb{G}(u) = \int_0^u G(x)dx$, $u > 0$, is the *integrated* distribution function G, where $\mathbb{G}(u) = 0$ if $u \leq 0$, and where $[S_{L,i}, S_{R,i}]$ is the interval for the time of becoming symptomatic for the ith person; E_i is again the exit time. So we maximize over a *convex* function $\mathbb{G}$ instead of a monotone function G.

It is clear that these maximization problems are totally different. Nevertheless, [9] has an appendix with the title 'Proof of non-equivalence in interval-censored likelihoods' to show that the likelihoods are not of the same type. May I say that this strikes me as a little bit odd?

I now take a look at the results of the parametric methods in [7] and compare these with the results of the nonparametric method. The estimates of the density (with respect to Lebesgue measure) of the incubation time distribution, based on the data in [7] are shown in Figure 1. I used the `R` package `coarseDataTools` to compute the parametric estimates based on the Weibull and log-normal distributions. The nonparametric estimate is bimodal, where the second mode is between 10 and 15. This could point to the existence of a subpopulation with longer incubation times of 10 to 15 days. The rigid parametric models do not allow this bimodality.

Note that we cannot make histograms of direct data here, and that we first have to perform the 'deconvolution' (see the concluding remarks below) of the time of becoming symptomatic into infection time and incubation time before we can produce the estimate of the incubation time. Moreover, even the time of becoming symptomatic is not directly observed, but only known to belong to a certain interval.

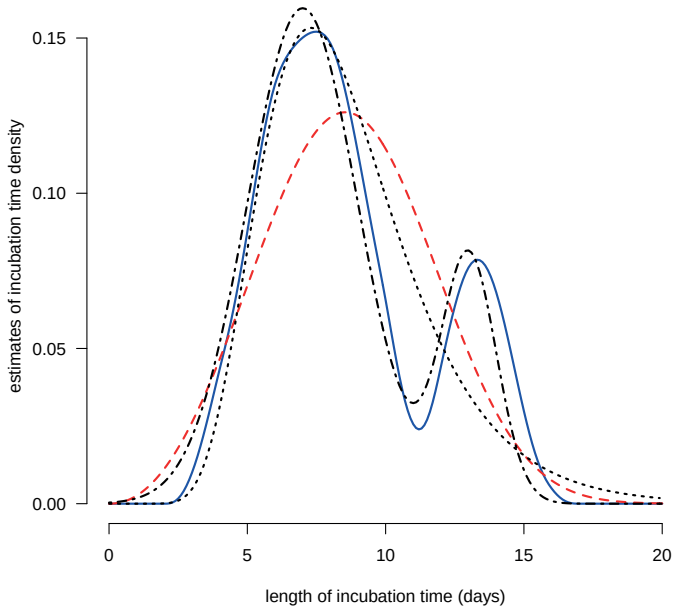

Figure 2 Estimates of the incubation time density for simulated data, using bandwidth $h = 3$ for the nonparametric estimate. The red dashed curve is the Weibull estimate of the density, the black dotted curve the log-normal estimate, the blue curve the nonparametric estimate and the black dashed-dotted curve the real density. Sample size is $n = 500$.

To see whether this bimodality could be real and yet completely missed by the parametric methods, I generated the incubation times from a mixture of normals on $[0,20]$ by rejection sampling. I used a sample of $n = 500$ elements. The exit times E_i were generated from a uniform distribution on $[0,30]$ and the infection times from a uniform distribution on $[0,E_i]$, conditionally on E_i. This yields sums S_i of the infection times and incubation times, upon which I dropped intervals $[S_{L,i},S_{R,i}]$ by drawing uniform random variables on $[-3,-1]$ and $[1,3]$. The R script for this can be found on [5] and was written by Slavik Koval.

Clearly the bimodality is indeed not noticed by the Weibull and log-normal estimates. On the other hand, the nonparametric estimate is remarkably close to the real density and completely reproduces the bimodality. The log-normal estimate seems a bit closer to the real density on the interval $[0,7]$. Both parametric estimates compensate for the missing of the bimodality by a global shift to the right.

For the mathematical statisticians, there is still a lot of work to be done in this direction. For example, I gave in [3] a sketch of the proof that in the single interval censored model the optimal bandwidth in the estimation of the density will converge at rate $n^{-1/7}$ in a continuous version of the model, if n is the sample size, and that the rate of convergence of the estimate itself is $n^{-2/7}$. Why is this 'asymptopia', as it is called by some medical statisticians, of practical importance? It is important because it leads to an automatic method for choosing the bandwidth. We can only use this particular method if we know the rate of convergence. Apart from this, 'asymptopia' is implicit in the computation of the confidence intervals for the parametric methods, as computed by using the bootstrap in the R package `coarseDataTools`.

What these rates are for the doubly interval censored model is still unknown. It is clear, though, that if we want more accurate estimates of the density of the incubation time and other characteristics of the pandemic, we have to explore these matters further and have to abandon the classical parametric methods.

Conclusion

My goal was to get information on the length of the incubation time for COVID-19. This is of importance for quarantine measures, et cetera. But we do not have direct observations. What we can (almost) observe is the time of becoming symptomatic. This is the sum of the time to infection and the time from infection till being symptomatic (i.e., the incubation time). Under a conditional independence assumption the distribution of this sum is the *convolution* of the distribution of time till infection and incubation time. We want to pull out of this the information on the incubation time distribution: this is the *deconvolution* aspect. We can do this under some assumptions on the infection time distribution, for example that this distribution is uniform on the time interval (till leaving Wuhan, for example). On top of this there is still the difficulty that even the time till being symptomatic is not observed exactly: we only have an interval for it, sometimes a day, sometimes a lot of days. That is the *interval censoring* aspect.

I discussed the analysis of a new data set linked to [7], with the suggestion of bimodality ('two cultures' again), exhibited by the nonparametric estimate. To illustrate that the presence of possible bimodality will be missed by the parametric methods, as used in [1] and [7], I generated a bimodal incubation time distribution, and applied the different methods to this data set. Figure 2 clearly shows that the nonparametric estimate reproduces the bimodality remarkably well and that indeed the parametric estimates completely miss the bimodality and are inferior estimates in this situation. That the usual parametric methods do not pick up the bimodality is not so surprising, but it might be surprising that the nonparametric methods do so well here. The R scripts for producing the test data and the pictures in this column can be found in [5].

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

3 Piet Groeneboom, Estimation of the incubation time distribution for COVID-19, *Statistica Neerlandica* (2020).

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

5 Piet Groeneboom, Doubly censored data, (2021), https://github.com/pietg/doubly_censored_data.

6 M.G. Gu and C.-H. Zhang, Asymptotic properties of self-consistent estimators based on doubly censored data, *Ann. Statist.* 21(2) (1993), 611–624.

7 S.A. Lauer, K.H. Grantz, Q. Bi, F.K. Jones, Q. Zheng, H.R. Meredith, A.S. Azman, N.G. Reich and J. Lessler, The incubation period of coronavirus disease 2019 (covid-19) from publicly reported confirmed cases: Estimation and application, *Annals of Internal Medicine* 172 (2020), 577–582.

8 Steven A. Lauer, ncov_incubation (2020), https://github.com/HopkinsIDD/ncov_incubation.

9 Nicholas G. Reich, Justin Lessler, Derek A.T. Cummings and Ron Brookmeyer, Estimating incubation period distributions with coarse data, *Stat. Med.* 28(22) (2009), 2769–2784.