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

Column Piet takes his chance

Chernoff's distribution and the bootstrap

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

Chernoff's distribution

In 1964 the remarkable paper [2] appeared, where Herman Chernoff derived the limit distribution of an estimator of the mode of a density. This distribution has become the 'normal distribution of non-standard asymptotics' in the sense that many estimators in the realm of (what is now called) non-standard asymptotics have this distribution as limit distribution. To the list of estimators with this limit distribution can recently be added the nonparametric maximum likelihood estimator of the distribution function of the incubation time of Covid-19, as discussed in [6].

The journal in which Herman Chernoff published his paper does not rank very highly in the list of journals in mathematical statistics, and I asked Herman Chernoff why he had sent his paper to this journal. He answered: "Well, they asked me to write a paper for them, and this is what I got." Which shows again that it is the author that counts, and not the prestige of the journal in which the author publishes (think of Grigori Perelman who published his results on arXiv).

Herman Chernoff

I looked for information on Chernoff's distribution in Wikipedia [11], but was shocked to see my own name three times in the references but no mention of Chernoff's paper [2]! And no, it wasn't me who wrote that entry! My only contributions to Wikipedia are about the Supreme Court during the German occupation, where my father played a role (see [12]) and about the statistical arguments in the Lucia de Berk court case (see [13]).

Now, to quote the Wikipedia article [11]: Chernoff's distribution, named after Herman Chernoff, is the probability distribution of the random variable

$$Z = \operatorname*{argmax}_{x\in\mathbb{R}}\{W(x) - x^2\}, \tag{1}$$

where W is a two-sided Wiener process (or two-sided Brownian motion), satisfying $W(0) = 0$. One of the pleasant properties of Brownian motion is that we can always use lots of symmetries; in this case one can immediately see that the random variable Z in (1) has the same distribution as the random variable.

$$U = \operatorname*{argmin}_{x\in\mathbb{R}}\{W(x) + x^2\}, \tag{2}$$

Moreover, further properties of Brownian motion imply that the location of the maximum (argmax) and the location of the minimum (argmin) are almost surely unique, and we therefore have (almost surely) well-defined random variables.

As mentioned in [6], for the model introduced there, the nonparametric maximum likelihood estimator $\hat{F}_n$ of the distribution function of the incubation time distribution function maximizes the log likelihood over all (cumulative) distribution functions F such that $F(x) = 0$ for $x < 0$:

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

for a sample of size n, where S_i is the time of becoming symptomatic and E_i is the length of the exposure time of the ith person. It is proved in [7] that, if F_0 is the real distribution function of the incubation time and t is an interior point of the support of its density f_0, we have, under some conditions on the underlying distributions, denoting the convergence in distribution by $\xrightarrow{d}$, the following result for the maximum likelihood estimator (MLE) $\hat{F}_n$:

$$\lim_{n\to\infty} n^{1/3}\{\hat{F}_n(t) - F_0(t)\} \xrightarrow{d} cU, \tag{3}$$

where U is given by (2), for a constant $c > 0$, depending on the incubation time distribution and the distribution of the exposure times E_i. This means that, after rescaling, the limit distribution of

the maximum likelihood estimator is given by Chernoff's distribution (Theorem 4 in [7], where the dependence of the constant c on the underlying distributions is explicitly given).

A similar result was proved for a somewhat related model (the *interval censoring* model) 25 years ago in [3], and for the proof of (3) I had to go through similar steps again. The proof is quite complicated and I wish someone would come up with an easier one. The latest news on the analytical characterization of Chernoff's distribution is given in [8], where also references to earlier work can be found.

Bootstrap confidence intervals

In [6] an estimate of the (cumulative) distribution function of the incubation time of Covid-19 on the basis of data of 88 travelers from Wuhan was given by the MLE, but also by the smoothed maximum likelihood estimator (SMLE), given by:

$$\hat{F}_{nh}(t) = \int \mathbb{K}((t-y)/h)d\hat{F}_n(y), \tag{4}$$

where $h>0$ is the bandwidth, $\hat{F}_n$ the MLE of the distribution function, and $\mathbb{K}$ the integrated kernel

$$\mathbb{K}(x) = \int_{-\infty}^{x} K(u)\,du, \quad x\in\mathbb{R},$$

see (3) in [6]. Here K is a smooth symmetric kernel, symmetric around zero, with support $[-1,1]$, and h a bandwidth (which we choose here to be 4, on a scale of days).

So the SMLE is the convolution of the kernel $\mathbb{K}_h(\cdot) \stackrel{\text{def}}{=} \mathbb{K}(\cdot/h)$ with a discrete measure, given by the MLE $\hat{F}_n$.

Now, does the result (3) help us to derive the properties of (4)? One could say: a little bit. We do not really need to know that the limit distribution is Chernoff's distribution, because this 'fine behavior' of the MLE is washed away in the convolution with the kernel. On the other hand, once we know that this is the limit behaviour of the MLE, we know how to look for upper bounds we'll need in developing theory for the SMLE (4), in particular for constructing confidence intervals.

Today the standard method to achieve this is called the bootstrap. The bootstrap simulates the model we are using from the estimates themselves, in this case the MLE $\hat{F}_n$. The name 'bootstrap' comes from a version (American?) of the Baron von Münchhausen tales, where the baron pulls himself out of the swamp by his 'bootstrap' (in the copy I read as a child he pulls himself out of the mud by his hair).

However, one always has to prove that this will 'work', which means that one has to show that this procedure really reproduces the random behaviour one wants to simulate. In the present case it has for example been proved in [10] that the method *does not work* for the MLE in the analogous model of interval censoring if one tries to use the standard method for doing this, which is to resample with replacement from the data and to compute the estimates for these samples. In fact, the bootstrap has been proved to be *inconsistent* in this case. There is little doubt that the usual bootstrap will fail similarly for the MLE in the present model.

But one can in fact reproduce the convergence to Chernoff's distribution in (3) by using a different kind of bootstrap, as is suggested in [10]. In this version of the bootstrap one resamples from a smoothed version $\tilde{F}_n$ of the MLE, where $\tilde{F}_n$ has the property that $\forall u\in\mathbb{R}$:

$$\lim_{n\to\infty} n^{1/3}\{\tilde{F}_n(t+n^{-1/3}u) - \tilde{F}_n(t) - f_0(t)\,n^{-1/3}u\} = 0, \tag{5}$$

almost surely at points t in the interior of the support of the distribution, where f_0 is the density of the underlying incubation distribution (assumed to exist). In fact, the smoothed maximum likelihood estimator (SMLE) $\tilde{F}_{nh}$ has the desired property.

Although it is remarkable that we can indeed reproduce the 'Chernoffian behavior' by using this smooth bootstrap, one could (rightly) object to this method that if we assume that there exists an estimator having the property (5), we can do better and in fact use the SMLE $\tilde{F}_{n,h_n}$ as our estimator of the distribution function instead of the MLE. We then can prove:

$$n^{2/5}\{\tilde{F}_{n,h_n}(t) - F_0(t)\} \xrightarrow{d} N(\mu,\sigma^2),$$

if $h_n \sim kn^{-1/5}$, for $k>0$. That is, we have convergence to a normal distribution $N(\mu,\sigma^2)$, with mean μ and variance σ^2 specified in [7], at a faster rate than achieved by the nonparametric MLE in (3).

The construction of the 95% bootstrap confidence intervals in Figures 1 and 2 proceeded in the following way.

The (original) sample is $(E_1,S_1),\ldots,(E_n,S_n)$.
Sample $I_1^*,\ldots,I_n^*$ uniformly from $[0,E_1]$, …,$[0,E_n]$, respectively.
Sample $U_1^*,\ldots,U_n^*$ from the SMLE $\tilde{F}_{nh}$ of the incubation distribution function.
Let $S_i^* = I_i^* + U_i^*$. Then our bootstrap sample is:

$$(E_1,S_1^*),\ldots,(E_n,S_n^*).$$

Note that we keep the E_i fixed, relieving us from the duty of estimating its distribution.
Compute for each bootstrap sample either the MLE $\hat{F}_n^*$or the SMLE $\tilde{F}_{nh}^*$.

For Figure 1 all 1000 values $\hat{F}_n^*(t) - \int \mathbb{K}_h(t-y)\,d\hat{F}_n(y)$ were ordered and the percentiles $P_{0.025}(t)$ and $P_{0.975}(t)$ were determined. This gives the bootstrap intervals:

$$\big[\hat{F}_n(t) - P_{0.975}(t), \hat{F}_n(t) - P_{0.025}(t)\big].$$

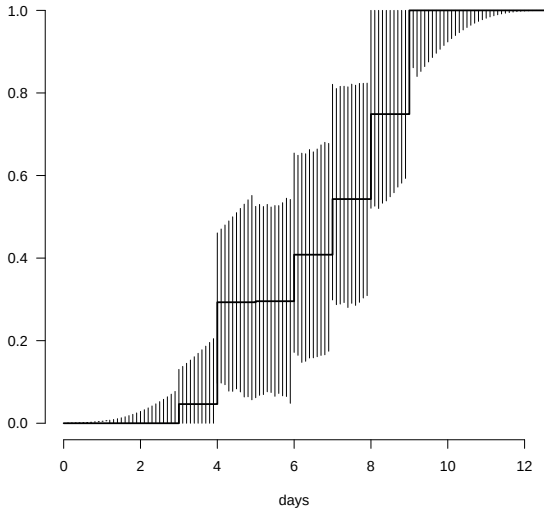

Figure 1 Estimate of the distribution function of the incubation time for the data in [1] by the MLE, with pointwise 95% bootstrap confidence intervals.

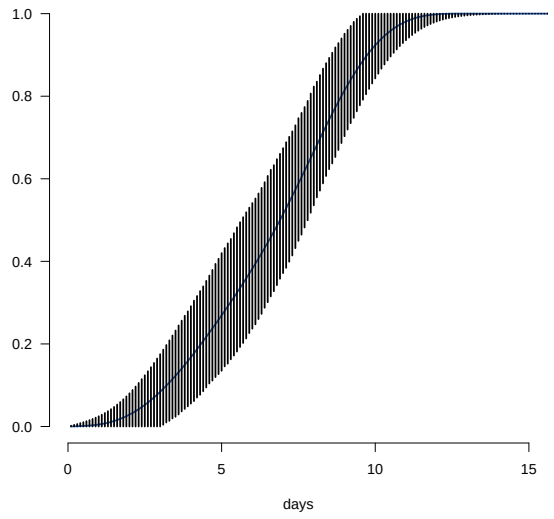

Figure 2 Estimate of the distribution function of the incubation time for the data in [1] by the SMLE, with pointwise 95% bootstrap confidence intervals.

For Figure 2 all 1000 values $\tilde{F}^*_{nh}(t) - \int \mathbb{K}_h(t-y)\,d_{nh}(y)$ were ordered and the percentiles $\tilde{P}_{0.025}(t)$ and $\tilde{P}_{0.975}(t)$ were determined. Note that we do not subtract $\tilde{F}_{nh}(t)$ but instead subtract the convolution of the kernel $\mathbb{K}_h$ with $d\tilde{F}_{nh}$. This gives the bootstrap intervals:

$$[\tilde{F}_{nh}(t) - \tilde{P}_{0.975}(t), \tilde{F}_{nh}(t) - \tilde{P}_{0.025}(t)].$$

A similar procedure for the density estimates yields the confidence intervals in Figure 3.

A comparison

We now try to throw more light on the difference between the nonparametric approach and the approach using distributions like the Weibul, log-normal, et cetera. for the incubation time distribution. We investigated, in a follow-up of a study presented by the medical statistician Ronald Geskus at a session of the Joint Statistical Meeting (JSM) August this year, how these methods behave in the estimation of the 95th percentile of the distribution. To this end we generated 1000 samples of size $n = 500$, using a Weibull distribution to generate the incubation time distribution.

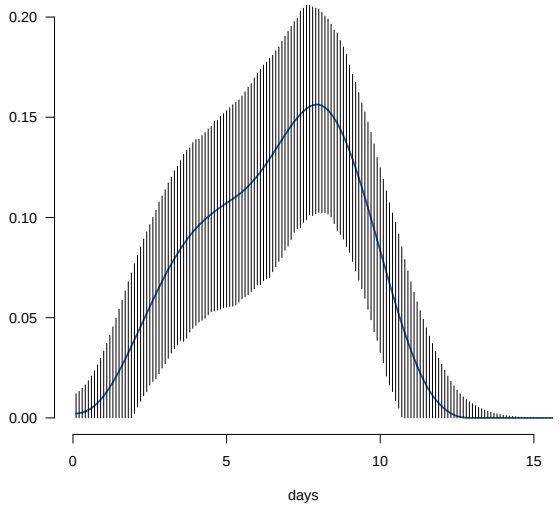

Figure 4 Estimate of the density of the incubation time for the data in [1], with pointwise 95% bootstrap confidence intervals.

This means that for each sample of size $n = 500$ we generated random variables $V_1, \ldots, V_n$, where each V_i had the Weibull distribution function

$$x \mapsto F_{\alpha,\beta}(x) \stackrel{\text{def}}{=} 1 - \exp\{-\beta x^{-\alpha}\}, \tag{6}$$

for $x \geq 0$ (zero otherwise), where $\alpha, \beta > 0$. In the simulations we took $\alpha = 3.03514$ and $\beta = 0.002619$ (these were values that came out of the study of travelers from Wuhan, discussed in [6]) for generating the samples. We also generated random variables $E_1, \ldots, E_n$ (exposure times), uniform on $[1,30]$, and for each i a variable $U_i \in [0, E_i]$ (infection time), uniform on $[0, E_i]$. This means that the sample on the basis of which the estimators of the incubation time distribution are computed consists of the pairs

$$(E_i, S_i) = (E_i, U_i + V_i), \quad i = 1, \ldots, n,$$

where S_i is the time of becoming symptomatic for the ith person, letting the exposure time have origin zero.

In the Weibull approach to the problem, we maximize for $\alpha, \beta > 0$:

$$\sum_{i=1}^{n} \log\{F_{\alpha,\beta}(S_i) - F_{\alpha,\beta}(S_i - E_i)\}, \tag{7}$$

where $F_{\alpha,\beta}$ is defined by (6). This gives a maximum likelihood estimate $F_{\hat{\alpha},\hat{\beta}}$ of the distribution function, where $(\hat{\alpha}, \hat{\beta})$ maximizes (7) over (α, β). The estimate of the 95th percentile is then defined by $F^{-1}_{\hat{\alpha},\hat{\beta}}(0.95)$, where $F^{-1}_{\hat{\alpha},\hat{\beta}}$ denotes the inverse function.

In the log-normal approach to the problem, we maximize for $\alpha \in \mathbb{R}$ and $\beta > 0$:

$$\sum_{i=1}^{n} \log\{G_{\alpha,\beta}(S_i) - G_{\alpha,\beta}(S_i - E_i)\}, \tag{8}$$

where $G_{\alpha,\beta}$ is defined by

$$G_{\alpha,\beta}(x) = \Phi((\log x - \alpha)/\beta), \tag{9}$$

for $x > 0$ (zero otherwise), where $\beta > 0$ and Φ is the standard normal distribution function. The estimate of the percentile is then given by $G^{-1}_{\hat{\alpha},\hat{\beta}}(0.95)$, where $(\hat{\alpha}, \hat{\beta})$ maximizes (8) over (α, β).

In the *nonparametric* maximum likelihood approach we simply maximize

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

over *all* distribution functions F. This give the nonparametric MLE $\hat{F}_n$, from which we compute the SMLE $\tilde{F}_{nh}(t) = \int \mathbb{K}_h(t-y)\,d\hat{F}_n(y)$ and the estimate of the 95th percentile $\tilde{F}^{-1}_{nh}(0.95)$. The bandwidth h was chosen to be $h = 2$ here.

The results of this simulation for 1000 samples of size $n = 500$ are shown in the box plot Figure 4. The box plot is an invention of John Tukey, who started his mathematics career as a topologist: Tukey's lemma ("There is a maximal member of each non-void family of finite character") is one of the statements equivalent to the axiom of choice and the well-ordering principle listed in the famous book on topology [9] He is also known for his work on the Fast Fourier Transform (Cooley–Tukey FFT algorithm) and many other contributions to science. The box plot gives an image of the variability of the data, which in this case consist of the estimates of the 95th percentile of the incubation time distribution. The grey colored box ('interquartile box') is between the first and third quartile

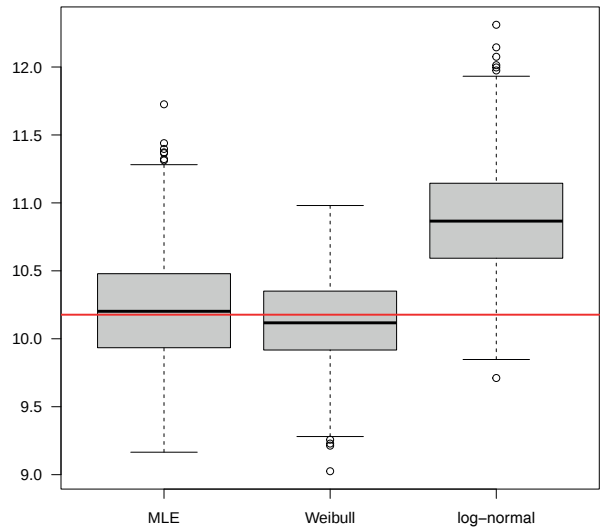

Figure 4 Box plot of 95th percentile estimates for the nonparametric, Weibull and log-normal maximum likelihood estimators for 1000 samples of size $n = 500$. The incubation time data are generated from a Weibull distribution.

of these 1000 estimates. The 'whiskers' are at a distance equal to 1.5 times the interquartile range from the boundaries of the interquartile box if there are outliers and otherwise at the smallest or largest observation. The outliers are further away than 1.5 times the interquartile range from the boundaries of the interquartile box. The black line segments in the boxes are at the position of the median. Finally, the red line denotes the value of $F^{-1}_{\alpha,\beta}(0.95) \approx 10.17716$ where $(\alpha, \beta) = (3.03514, 0.002619)$ (the values used to generate the Weibull incubation time distribution).

It can be seen that, since the incubation time data were generated from a Weibull distribution, the estimates of the quantiles assuming this distribution have indeed the smallest variation, but are slightly biased. But the nonparametric estimates, not making the assumption that the distribution is of the Weibull type, are also pretty good, whereas the estimates, assuming a log-normal distribution are completely off (in fact, these estimates are inconsistent, i.e., will not converge to the right value if the sample size tends to ∞), and indeed the model is 'misspecified' for these estimates. But the term 'misspecified' does not apply to the computations with the SMLE, since that estimate adapts to the underlying distribution and provides consistent estimates.

Some remarks of the statistician Richard Gill on this: One sees that the uncertainty about what interest us, is not much larger when one only uses the nonparametric SMLE than when one assumes that the incubation time distribution is Weibull (which is the correct distribution in this simulation setting), but much larger when one assumes log-normal. While there is absolutely no scientific (medical) reason to 'believe' Weibull, or to 'believe' log-normal. They lead to completely different statistical inferences, hence could lead to completely different policy recommendations. How to choose? We don't have to. We statisticians have figured out how to do it better, not making these assumptions.

Conclusion

In [6] the nonparametric maximum likelihood estimator $\hat{F}_n$ of the distribution function of the incubation time for Covid-19 was introduced, as an alternative to the parametric estimates used in this case, see, e.g., [1]. At that time it was still unknown what the local limit distribution of $\hat{F}_n$ (at a fixed point t) was. In the mean time it has been proved in [7] that, after rescaling, the limit distribution is Chernoff's distribution ((3) above), which was not exactly unexpected, but rather hard to prove.

It is argued above (and also in [5]) that one can probably better use smoothed nonparametric maximum estimates, of which Figures 2 and 3 give examples. Moreover, we can produce confidence intervals in a fully automatic way, using the smooth bootstrap, generating bootstrap samples from the SMLE. The ordinary bootstrap fails in producing valid confidence intervals via the MLE itself, as can be deduced from results in [10].

The methods discussed here provide an alternative for the parametric models which are usually applied in this context, estimating the incubation time distribution by, e.g., Weibull, gamma or log-normal distributions. The latter methods will not be able to catch finer aspects of the data, in contrast with, for example, the nonparametric density estimates. Moreover, they will only perform well if the underlying distribution is of the assumed type (Weibull, log-normal, et cetera) and otherwise will be inconsistent. On the other hand, the nonparametric methods will be consistent, irrespective of the underlying distribution.

All programs for generating the pictures in this column were written in C++ and are available as R scripts on [4].

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 H. Chernoff, Estimation of the mode, *Ann. Inst. Statist.* Math. 16 (1964), 31–41.

3 P. Groeneboom, Lectures on inverse problems, in *Lectures on Probability Theory and Statistics (Saint-Flour, 1994)*, Lecture Notes in Math., Vol. 1648, Springer, 1996, pp. 67–164.

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

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 Piet Groeneboom, Nonparametric estimation of the incubation time distribution (2021), arXiv:2108.12606.

8 Piet Groeneboom, Steven Lalley and Nico Temme, Chernoff's distribution and differential equations of parabolic and Airy type, *J. Math. Anal. Appl.* 423(2) (2015), 1804–1824.

9 John L. Kelley, *General Topology*, D. Van Nostrand Company, 1955.

10 Bodhisattva Sen and Gongjun Xu, Model based bootstrap methods for interval censored data, *Comput. Statist. Data Anal.* 81 (2015), 121–129.

11 Wikipedia, Chernoff's distribution, https://en.wikipedia.org/wiki/Chernoff%27s_distribution, 2019.

12 Wikipedia, Supreme Court of The Netherlands in World War 2, https://en.wikipedia.org/wiki/Supreme_Court_of_the_Netherlands#Second_World_War, 2021.

13 Wikipedia, Lucia de Berk, https://en.wikipedia.org/wiki/Lucia_de_Berk, 2021.