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

Column Piet grijpt zijn kans

Nederland in tijden van corona

Piet Groeneboom, emeritus hoogleraar statistiek aan de TU Delft, schrijft dit keer in plaats van Casper Albers een column over een actueel statistisch onderwerp.

Op mijn avondwandeling werd ik aangeroepen door een buurvrouw, die in de tuin aan het werk was. Zij riep: "Het reproductiegetal staat op $1{,}4$, had je dat al gezien?" Nee, dat had ik niet gezien; ik was nog blijven steken op $1{,}29$. Op de televisie had ik iemand zien uitleggen dat dit betekende dat op dat moment 100 mensen gemiddeld 129 andere mensen infecteerden. De buurvrouw en ik hadden al eerder over het reproductiegetal gepraat. Mijn collega Richard Gill heeft het in dit verband over "A lie for children". Niet zo maar een leugen dus, een leugen *voor kinderen*.

Aan de andere kant had een oude vriend die nu in Duitsland woont, waar de natuurkundige Angela Merkel zelf aan het Duitse volk de betekenis van het reproductiegetal had uitgelegd, mij geschreven dat het voor hem niet duidelijk was waarom het reproductiegetal geen goede basis voor politieke beslissingen zou zijn, zolang het onmogelijk is de individuele besmettingsketens te reconstrueren.

Ik zelf dacht, toen ik Jaap van Dissel voor de eerste keer over het reproductiegetal zag praten op de televisie, ineens: "Hé, wiskundigen zouden hier iets nuttigs kunnen doen." Het was op een moment dat ik al verwikkeld was in een discussie met een groep kansrekenaars en statistici, waarvan één ons een stukje dat hij aan de krant wilde sturen had voorgelegd. Via deze discussie was ik beland op het COVID-19-mod-forum, waar ik onmiddellijk een 'channel' gestart ben, getiteld R_0 (R_0 om verwarring met de momenteel veel gebruikte taal `R` voor statistische software te vermijden, R_0 zou eigenlijk het reproductiegetal aan het begin van de pandemie zijn, het zogenaamde basis-reproductiegetal).

Merkwaardig is dat we steeds op de televisie mensen zien die de mededelingen over de hoogte van het reproductiegetal van het RIVM voor zoete koek lijken te slikken, terwijl we eigenlijk geen idee hebben hoe het RIVM dit heeft uitgerekend. Op de site van het RIVM [8] zien we onder het kopje 'reproductiegetal R' drie dingen:

1. 'Berekening waarde reproductiegetal R'. Dit is een link naar een artikel uit 2007 [7].
2. 'Code beschikbaar'. Dit is een link naar het `R`-package `EpiEstim` van Anne Cori en anderen.
3. '`R`-package `R` beschikbaar'. Dit is een link naar een ander `R`-package, het `R`-package `R0`.

Weten we nu hoe het RIVM R heeft berekend? Absoluut niet! We weten zelfs niet welke van de twee genoemde `R`-packages (en welke methoden hieruit) ze hebben gebruikt! We zullen daarom nu een ander onderwerp bekijken dat op dezelfde site gegeven wordt, maar waar meer informatie lijkt te worden verstrekt.

Figuur 1 Screenshot van de RIVM-webpagina 'Rekenmodellen openbaar en toegankelijk'.

De verdeling van de incubatietijd

Op de pagina 'Rekenmodellen openbaar en toegankelijk' op de RIVM-site [8] staat bovenaan 'Onderzoek naar incubatietijd van COVID-19' (zie Figuur 1 voor een screenshot). Dit bevat een link naar [1] en de twee andere verwijzingen ('Code beschikbaar' en 'Data beschikbaar') zijn allebei naar hetzelfde 'supplementary material' bij dit artikel, dat een databestand van 88 'travelers from Wuhan' en een R-script bevat.

Blij dat ik eindelijk een concrete databehandeling van het RIVM in handen kreeg! Helaas waren er wat moeilijkheden bij het runnen van het R-script, iets dat overigens in het algemeen vaak voorkomt. Waar gaat het om? De gegevens zijn van 88 personen die aan het eind van vorig jaar en/of het begin van dit jaar in Wuhan zijn geweest en daar waarschijnlijk het COVID-19-virus hebben opgelopen. Ook is het tijdstip bekend waarop zij symptomatisch zijn geworden. Het laatste is enigszins ongewoon, omdat dit tijdstip vaak ook op een of andere manier gecensureerd is, net als het tijdstip van de infectie dat hier *interval gecensureerd* is, zoals dat heet. Maar ik neem deze data nu maar even 'at face value'.

Door wat verder te studeren in het R-script ontdekte ik dat het eigenlijke databestand dat voor de analyse gebruikt was een selectie uit en bewerking van het op de site verstrekte databestand 'BAKER_Supplementary material S1_data.tsv' was en 'data.input' heette in het R-script; beide databestanden zijn te vinden op mijn GitHub-repository [4].

Er zijn wat eigenaardigheden. De eerste persoon heeft als vertrekdatum uit Wuhan 4-1-2020, maar 3-1-2020 als datum voor het krijgen van symptomen. In data.input is hier in beide gevallen 3-1-2020 van gemaakt. Algemeen is het eind van de blootstellingsperiode gelijkgesteld aan het tijdstip van het krijgen van symptomen als dit nog in Wuhan gebeurde. 'Blootstellingsperiode' is de vertaling van het Engelse 'exposure time', een uitdrukking die ik eigenlijk liever zou gebruiken, maar ik ben nu eenmaal in het Nederlands begonnen te schrijven. Verder: Wuhan-reiziger nummer 67 is 0 dagen in Wuhan geweest ('connecting flight'). Daar heb ik maar één dag van gemaakt. Dat is overigens de enige verandering die ik in het 'data.input'-bestand heb aangebracht.

Ten slotte zijn er een groot aantal NA's ('not available') voor de aanvang van de blootstellingsperiode in Wuhan. De NA's zijn allemaal op -18 gezet in het getransformeerde databestand, wat (kennelijk) betekent: 18 dagen voor oudjaar, het oudjaar zelf is op 0 gezet (heb ik gededuceerd).

Om te kijken of ik er hetzelfde uitkreeg als het RIVM als ik hun type analyse hierop los zou laten, evenwel zonder hun (mijns inziens niet relevante) Bayesiaanse terminologie erbij te halen, heb ik een C++-programmaatje geschreven dat ook te vinden is op [4]. Tussen haakjes: C++ is een zich nog steeds ontwikkelende taal die om verschillende redenen veruit superieur is aan R. Het is veel sneller, er zijn veel meer 'debug'-faciliteiten, C++11 heeft heel goede random number-generatoren (Mersenne-twister, enzovoort), de geheugenbehandeling is beter en ten slotte is de taal veel meer 'solide' dan R (C++ heeft zich natuurlijk ook al heel lang ontwikkeld vanuit C).

Als je eenmaal een C++-programma hebt geschreven kun je dit meestal gemakkelijk doorsluizen naar een R-script via het R-package Rcpp (zie [2]) en dat is ook wat ik in [4] heb gedaan. De conclusie van mijn exercitie is: ja, ik krijg er ongeveer hetzelfde uit als zij als ik het in hun R-code gegenereerde bestand data.input gebruik en *parametrische* maximum likelihood gebruik. Maar daarmee is de kous nog niet af, zie hieronder.

Niet-parametrische maximum likelihood

Het heeft er een beetje de schijn van dat veel epidemiologen hoogstens bekend zijn met het begrip 'parametrische maximum likelihood' en dat bij hen het begrip 'niet-parametrische maximum likelihood' niet bekend is. Al meer dan zestig jaar is echter bekend dat je ook iets heel anders kunt doen, namelijk *niet-parametrische maximum likelihood*. Een voorbeeld van zo'n niet-parametrische maximum likelihood-schatter is de schatter van een monotoon dalende dichtheid in een artikel uit 1956 [3].

We bespreken hier nu kort de parametrische en niet-parametrische maximum likelihood-methode in het onderhavige geval. Hierbij centreren we voor de eenvoud van notatie, en zonder verlies van algemeenheid, de blootstellingsperiode zo dat het linkereindpunt 0 is, dat wil zeggen, we verminderen deze met het begin van de blootstellingsperiode. De log-likelihood (ik gebruik maar de heel erg ingeburgerde Engels-Amerikaanse terminologie en probeer niet Nederlands-puristisch te zijn) voor de i-de waarneming die ons ter beschikking staat zou zijn:

$$\log \int_{t \in [0,E_i]} g(S_i - t)\, dF_i(t),$$

dat wil zeggen de logaritme van de dichtheid van één waarneming. Hier is g de onbekende dichtheid van de incubatietijd waarin we geïnteresseerd zijn, S_i het tijdstip waarop persoon i symptomatisch is geworden, verminderd met het beginpunt van zijn/haar blootstellingsperiode, E_i is het eind van de blootstellingsperiode, verminderd met het begin van de blootstellingsperiode en F_i is de verdelingsfunctie van het (verschoven) infectietijdstip op het interval $[0, E_i]$. We hebben hier in feite te maken met een *deconvolutie* probleem. Om de dichtheid g van de incubatietijd-verdeling (waar het ons dus eigenlijk om gaat) identificeerbaar te maken, moeten we een aanname doen met betrekking tot de verdelingsfunctie F_i van de blootstellingsperiode. We maken hiervoor de meest voor de hand liggende aanname, die inderdaad meestal gemaakt wordt, namelijk dat deze verdeling uniform is op het interval, zie bijvoorbeeld [6]. Ook in [1] wordt deze aanname gemaakt, waarin dit echter, mijns inziens ten onrechte, in Bayesiaans terminologie, een 'prior distribution' wordt genoemd.

Als we deze aanname maken, wordt de log-likelihood van alle (88) waarnemingen, onder de aanname dat deze onafhankelijk

zijn, de logaritme van de dichtheid van alle waarnemingen:

$$\sum_{i=1}^{n} \log\Big\{\int_{t\in[0,E_i]} g(S_i-t)\,dt/E_i\Big\} \tag{1}$$

waarbij $n=88$.

De scheiding der geesten treedt nu echter op in de manier waarop g wordt geschat. Het RIVM, samen met de bijna volledige epidemiologische gemeenschap, doet alsof er alleen maar een aantal keuzes hiervoor mogelijk zijn: de Weibull-verdeling, de gamma-verdeling en de log-normale verdeling. Het doet er nu even niet toe wat hiervoor de formules zijn, die men trouwens op internet kan opzoeken. Het gaat om het principe dat men een verdeling (die overigens duidelijk begrensde support heeft) wil schatten met een aantal bekende verdelingen met oneindige support die door een klein aantal parameters gekarakteriseerd worden. Omdat er geen dwingende inhoudelijke reden is om voor één van deze verdelingen te kiezen, zien we steeds artikelen waar ze maar allemaal worden geprobeerd, terwijl ze er eigenlijk waarschijnlijk allemaal naast zitten als we wat nauwkeuriger naar de data kijken.

Het maximum likelihood-principe vertelt ons om g te schatten door (1) te maximaliseren naar g. Dat wil zeggen, we maximaliseren de logaritme van de dichtheid van de waarnemingen als functie van de onbekende verdeling.

Omdat de noemer E_i bij de integraal in (1) geen rol speelt in de maximalisatie van (1), komt dit neer op het maximaliseren van

$$\sum_{i=1}^{n} \log\{G(S_i)-G(S_i-E_i)\}, \tag{2}$$

waarbij G de (cumulatieve) verdelingsfunctie is horend bij de dichtheid g (de geïntegreerde dichtheid dus). Van de Weibull-enzovoortgemeenschap mogen we hier dus alleen een Weibull-enzovoortverdelingsfunctie kiezen. Maar waarom eigenlijk? Wat gebeurt er als we (2) gewoon maximaliseren over *alle* verdelingsfuncties? We krijgen dan een zogenaamde discrete verdelingsfunctie als oplossing, corresponderend met discrete kansen op een eindig aantal (niet vooraf gegeven) punten: de *niet-parametrische maximum likelihood-schatter van de verdeling!* Deze verdelingsfunctie is stuksgewijs constant en wordt getoond in Figuur 2 links, samen met de verdelingsfunctie die uit de maximum likelihood-procedure voor de Weibull-verdeling komt.

Het probleem (2) te maximaliseren over alle verdelingsfuncties is niet geheel triviaal. Er bestaan verschillende (iteratieve) methoden die besproken worden in [5]. We kunnen bijvoorbeeld het zogenaamde Expectation Maximization-algoritme (EM-algoritme) gebruiken en ook het veel snellere iteratieve convexe minorantalgoritme (zie [5] en [4]).

We kunnen deze schatting van de verdelingsfunctie van de incubatietijd ook als basis nemen voor een gladdere schatting van de verdelingsfunctie, de zogenaamde 'Smoothed Maximum Likelihood Estimate' (SMLE) en zelfs van de (absoluut continue) dichtheid (als we aannemen dat deze bestaat). Als in [5] (zie bijvoorbeeld sectie 1.2), kunnen we de SMLE berekenen, evenals een schatting van de dichtheid, in beide gevallen gebruikmakend van de niet-parametrische MLE. De SMLE wordt gedefinieerd door

$$\tilde{F}_{nh}(t) = \int \mathrm{IK}((t-y)/h)\,d\hat{F}_n(y), \quad t\in\mathbb{R}, \tag{3}$$

waarbij $h>0$, $\hat{F}_n$ de MLE van de verdelingsfunctie, en IK een geïntegreerde kern

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

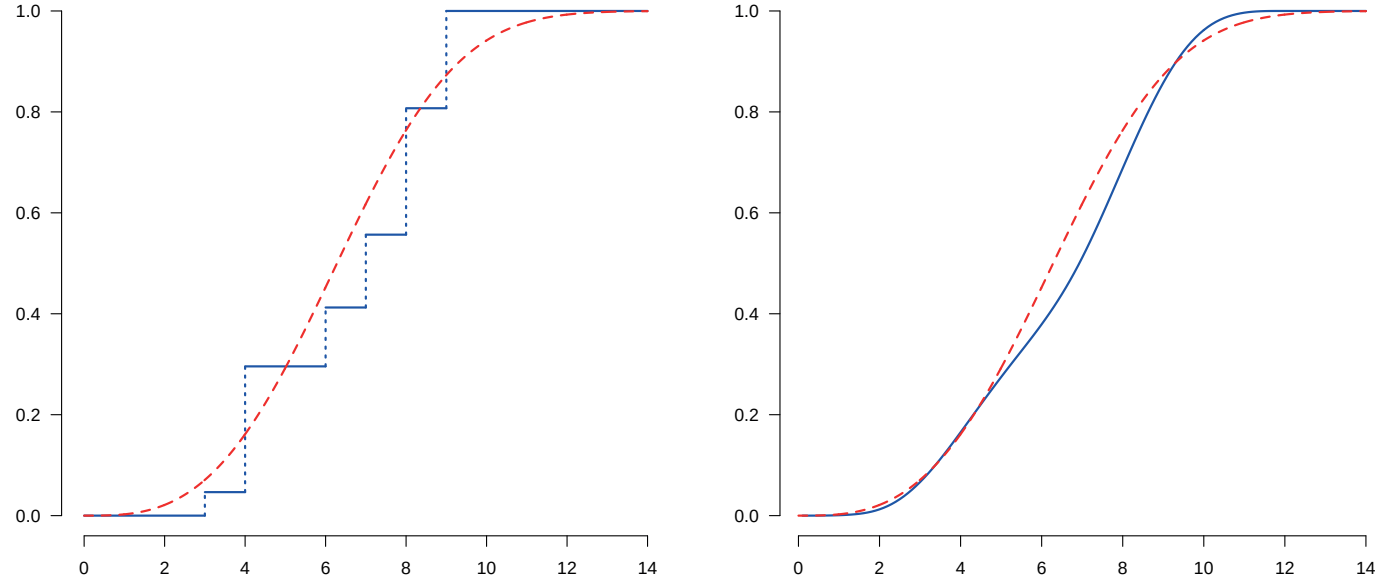
is. K is hier een symmetrische kern met support $[-1,1]$, bijvoorbeeld de zogenaamde triweight kern

$$K(u) = \frac{35}{32}(1-u^2)^3 1_{[-1,1]}(u), \quad u\in\mathbb{R}.$$

We schatten de dichtheid door:

$$\tilde{f}_{nh}(t) = h^{-1}\int K((t-y)/h)\,d\hat{F}_n(y), \quad t\in\mathbb{R}. \tag{4}$$

Voor de analyse hier namen we $h=3$ in (3) en $h=4$ in (4). Er is veel theorie ontwikkeld voor hoe groot we h moeten kiezen, maar we kunnen daarop hier niet ingaan. Het resultaat is te zien in de figuren, waar de schatters vergeleken worden met wat de Weibull maximum likelihood-schatters opleveren.



Figuur 2 Links: niet-parametrische Maximum Likelihood-schatter (MLE) van de verdelingsfunctie van de incubatietijd (blauw, sprongen zijn gestippeld) en Weibull MLE van de verdelingsfunctie (rood, gestreept). Rechts: Smoothed Maximum Likelihood Estimator (SMLE) van de verdelingsfunctie van de incubatietijd (blauw) en Weibull MLE (rood, gestreept).

Conclusie

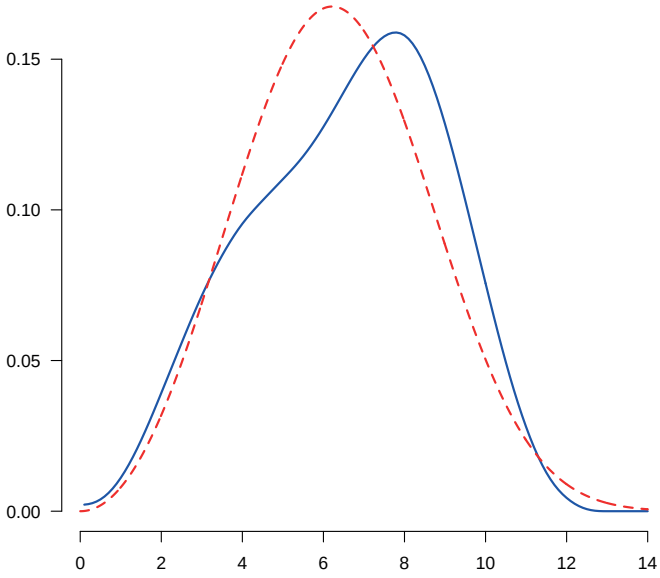
We hebben de site [8] 'Rekenmodellen openbaar en toegankelijk' van het RIVM bekeken. Helaas zijn we hier niets wijzer geworden over de manier waarop het reproductiegetal door het RIVM wordt berekend. Het ging wat beter met een schatting van de verdeling van de incubatietijd, waarvoor een R-script geleverd werd dat na herbenoemen van een bestand dat er niet was en installeren van extra packages gerund kon worden. Hierdoor kon ook achterhaald worden wat het getransformeerde databestand was dat men eigenlijk bewerkt had.

Voor het schatten van de verdeling van de incubatietijd gaf ik een andere methode, gebaseerd op niet-parametrische maximum likelihood, daarmee vermijdend dat we ons in een strijd tussen log-normale, gamma- of Weibull-verdeling hoeven te begeven. De methodes die ik gebruikt heb kunnen allemaal gecheckt worden op [4] ('de berekening is openbaar en toegankelijk').

Als we de verdelingsfuncties, verkregen via de Weibull maximum likelihood-methode, vergelijken met die verkregen via de niet-parametrische maximum likelihood-methode, zien we dat deze schattingen op elkaar lijken, vooral de SMLE ('Smoothed Maximum Likelihood Estimate') en de Weibull MLE ('Weibull Maximum Likelihood Estimate'), zie Figuur 2 rechts. De niet-parametrische aanpak leent zich er dus ook voor om eventuele keuze voor een parametrisch model te rechtvaardigen.

Bekijken we echter één niveau dieper de dichtheid, dan zien we in Figuur 3 een opmerkelijk verschil: de modus van de dichtheid (locatie van het maximum) is verschoven van $6{,}2$ naar $7{,}8$ dagen. Dat is dus $1{,}6$ dagen verschil! De betekenis van dit verschil en de waarde die we hieraan moeten hechten, zouden we graag verder willen onderzoeken. Berekeningen van dit type zijn van belang voor de quarantainetijd.

Deze studie is natuurlijk maar gebaseerd op 88 mensen, van wie de gegevens ook nogal onvolledig zijn vanwege de vele NA's ('not available') aan het begin van de blootstellingsperiode. Maar in ieder geval treft de niet-parametrische methode niet het verwijt dat we een verdeling gepakt hebben (Weibull, gamma, enzovoort) waarvoor we geen echt argument hebben en die alleen al door zijn rigide beperkingen een vertekend beeld kan geven van de data.



Figuur 3 Schatter van de dichtheid van de incubatietijd, gebaseerd op de niet-parametrische MLE van de verdelingsfunctie (blauw) en Weibull MLE van de dichtheid (rood, gestreept).

Voor de niet-parametrische schatters hebben mathematisch statistici de laatste zestig jaar heel veel theorie ontwikkeld en we kunnen er hetzelfde soort dingen mee doen als met de parametrische schatters, zoals (niet noodzakelijk symmetrische!) betrouwbaarheidsintervallen uitrekenen voor de schattingen van de momenten en de verdeling zelf. Er lijkt dus voor onderzoekers die zich met dit soort data bezighouden voldoende aanleiding om zich eens in deze niet-parametrische methoden te verdiepen. En voor mathematisch statistici om deze belangrijke epidemiologische onderwerpen te bestuderen.

Referenties

1 Jantien A. Backer, Don Klinkenberg en 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 Dirk Eddelbuettel, *Seamless R and C++ Integration with Rcpp,* Springer, New York, 2013,

3 Ulf Grenander. On the theory of mortality measurement. II, *Skand. Aktuarietidskr.* 39 (1956), 125–153.

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

5 Piet Groeneboom en Geurt Jongbloed, *Nonparametric Estimation under Shape Constraints*, Cambridge Univ. Press, 2014.

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

7 J. Wallinga and M. Lipsitch. How generation intervals shape the relationship between growth rates and reproductive numbers, *Proc. R. Soc. B* 274 (2007), 599–604.

8 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.