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

Column Piet grijpt zijn kans

Danse macabre

Piet Groeneboom schrijft op regelmatige basis in dit blad een column, deze keer over een dans met 'dead subjects'.

Een niet zo geslaagde lezing

In 2011 werd ik uitgenodigd om te spreken op de zogenaamde 'Statistische Dag' in de Jaarbeurs in Utrecht. Voor zo'n gebeurtenis moest je een abstract schrijven die verscheen in het blaadje van de Vereniging voor Statistiek. Het leek mij toen wel leuk om een abstract te schrijven die de 'Danse macabre' tot onderwerp had. De Danse macabre is een bekend stuk van Camille Saint-Saëns dat ik voor het eerst op school bij muziekles had gehoord. Ik was toen erg 'gepakt' en eigenlijk pakt het me nog steeds als ik het weer hoor. Ik durf het bijna niet te zeggen, zeker niet nadat ik ergens heb gelezen dat Ravel (een van mijn favoriete componisten) heeft gezegd dat Saint-Saëns maar beter helemaal niet had kunnen bestaan.

Maar Saint-Saëns heeft volgens mij wel een heel mooie vioolsonate geschreven, die (naast bijvoorbeeld de sonates van César Franck en Gabriel Pierné — de zusjes Milstein gokken op Pierné, zie [5]) kandidaat is voor de 'Sonate de Vinteuil' in *À la recherche du temps perdu* van Marcel Proust. De vraag hierbij is steeds wat dan de 'petite phrase' is in deze sonates. Maar dit terzijde.

De duivel wekt in de Danse macabre de doden uit hun graven, die met hem (tot het kraaien van de haan) een dans op het kerkhof uit gaan voeren. Hij doet dit door de twee bovenste snaren van zijn viool in het interval A-Es in plaats van A-E te stemmen. Het interval A-Es is een verminderde kwint en wordt wel de 'diabolus in musica' genoemd. Heel passend dus dat de duivel dit interval gebruikt.

Waarom dacht ik aan de 'Danse macabre' voor mijn lezing? Aan het begin van mijn loopbaan als statisticus had ik een gesprek gehad met een bekende Nederlandse statisticus, omdat ik zijn steun probeerde te verwerven voor een project dat ik voor een student had aangevraagd. Ik zou daar zelf het isotone regressie-deel van voor mijn rekening nemen. Isotone regressie is regressieanalyse onder een orde-restrictie op de schatter, zoals monotonie, convexiteit, enzovoort. Ik schrijf nog steeds artikelen op dit gebied. Verder wijdde het Amerikaanse tijdschrift *Statistical Science* in november 2018 nog een speciaal nummer geheel aan isotone regressie. Maar de bekende statisticus zei: "The subject is dead!" Hij was wel Nederlands, maar gebruikte graag Engelse uitdrukkingen.

Twintig jaar later sprak ik met een bekende Amerikaanse statisticus in Stanford over de bootstrap en hij zei hetzelfde (maar nu over de bootstrap): "The subject is dead!" Dus schreef ik in mijn abstract dat ik een Danse macabre zou gaan uitvoeren met deze twee 'dead subjects'. In feite was mijn column 'Chernoff's distribution and the bootstrap' in dit tijdschrift ook zo'n dans.

Tot zover was er nog niet zo veel aan de hand, maar toen het tijdstip van de lezing naderde, vond ik dat het mijn plicht was om uit te leggen waarom de duivel aan het begin van de Danse macabre zijn viool in het diabolus in musica-interval stemde en

Afbeelding: fragment van object 450661, Wellcome Collection

Een historische afbeelding van een 'Danse macabre'

dat ook te laten horen en te wijzen op de verminderde kwint (of overmatige kwart wat eigenlijk hetzelfde is). Ik schafte daartoe op internet een uitvoering van Kyung Wha Chung (vroeger een van mijn favoriete violistes) en Charles Dutoit (met een mijns inziens wat te zwaar aangezet orkest) aan, die ik inderdaad in de Jaarbeurs heb laten horen.

Vooraf lijkt zoiets allemaal heel leuk, maar, zoals vaak gebeurt, op de dag zelf ging alles fout. De Jaarbeurs is op zichzelf al zo'n gebouw dat ik liever niet betreed, maar het heeft wel het voordeel dat het vanuit het station Utrecht gemakkelijk te bereiken is. Ik was dus keurig op tijd, maar de voorzitter van de sectie die mij had uitgenodigd was er niet. Bovendien had mijn abstract kennelijk heel veel toehoorders getrokken. Dat is in het algemeen moeilijk van tevoren te voorspellen, soms sta je voor een enorme zaal met anderhalve man en een paardenkop, maar in dit geval waren er niet genoeg stoelen. De Nederlandse statisticus die had verklaard dat isotone regressie een 'dead subject' was, was er ook, samen met een bekende toegepaste statisticus uit Stanford. Zij moesten stoelen uit een belendende zaal gaan halen.

Er was dus een chaotische beginsituatie, met allemaal mensen die stoelen uit belendende zalen gingen halen en een ontbrekende voorzitter. Maar uiteindelijk arriveerde hij toch een kwartier te laat. Je zou denken dat hij me dan wat extra tijd zou geven, maar nee dat wilde hij niet (op dat soort dingen moet je ook voorbereid zijn). Dus daar begon ik met mijn uitleg over de diabolus in musica en het fragment met Kyung Wha Chung en Charles Dutoit. Toen was ik al bijna door de mij toegestane tijd heen.

Later heb ik nog via via begrepen dat alles wat ik had gezegd onbegrijpelijk was, zowel wat ik over de diabolus in musica had gezegd als wat ik nog over de 'two dead subjects' te melden had in de geringe resterende tijd. Gelukkig kon ik dus in deze column in ieder geval nog even die diabolus in musica opnieuw uitleggen.

Een zeer geslaagde lezing

Maar het kan ook anders. In 1980 was ik in Seattle bij een lezing van David Freedman (die hoogleraar was in Berkeley). David Freedman kwam daar een lezing over de bootstrap (een van de twee 'dead subjects', zie boven) houden. Eerlijk gezegd dacht ik voorafgaand aan die lezing dat de bootstrap flauwekul was, gewoon een variant van simuleren, waar veel ophef over gemaakt werd en waar een krachtige reclamecampagne achter zat.

Maar na die lezing van David Freedman was ik helemaal 'om'. Wat was dat een goede lezing, die bovendien op tijd begon! Er hoefden ook geen stoelen uit belendende zalen te worden gehaald. Alles wat David Freedman zei was volledig begrijpelijk en het was ook nog heel interessant. Het was in de begintijd van de artikelen over de bootstrap, niet lang nadat Bradley Efron zijn eerste artikelen hierover geschreven had (zie [2] en [3]).

Het prettige (voor mij) van het feit dat een echte wiskundige zoals David Freedman, die eveneens bijzondere boeken over de theorie van de Brownse beweging heeft geschreven, zich met dit onderwerp ging bezighouden was dat ik daardoor ook ineens begreep waar het eigenlijk om ging.

Hij schreef in die tijd een artikel met Peter Bickel (zie [1]), waarin bijvoorbeeld Theorem 1 hieronder stond. Bij de aanvankelijke vorm van de bootstrap (de 'klassieke bootstrap') trekken we met teruglegging uit onze oorspronkelijke steekproef $X_1,\ldots,X_n$. Deze trekkingen worden meestal aangeduid met X_i^*. We kunnen een bootstrap steekproef van dezelfde omvang nemen: $X_1^*,\ldots,X_n^*$ of een steekproef van andere omvang $X_1^*,\ldots,X_m^*$, waarbij m zowel groter als kleiner dan n kan zijn. Over het trekken van steekproeven van kleinere, verdwijnende omvang $m=o(n)$, waarbij desondanks $m\to\infty$ als $n\to\infty$, is een heel boek geschreven: zie [4].

In de stelling hieronder is $\bar{X}_n$ het gemiddelde van de oorspronkelijke steekproef, $\bar{X}_m^*$ het gemiddelde van de bootstrap steekproef, S_m^* is gedefinieerd door

$$S_m^*=\left\{m^{-1}\sum_{i=1}^m\left(X_i^*-\bar{X}_m^*\right)^2\right\}^{1/2}$$

(de positieve wortel) en $N(0,\sigma^2)$ duidt een normale verdeling met verwachting 0 en variantie σ^2 aan.

Theorem 1. *Suppose $X_1,X_2,\ldots$ are independent, identically distributed, and have finite positive variance σ^2. Along almost all sample sequences $X_1,X_2,\ldots$, given $(X_1,\ldots,X_n)$, as n and m tend to ∞:*

a. *The conditional distribution of $\sqrt{m}(\bar{X}_m^*-\bar{X}_n)$ converges in distribution to $N(0,\sigma^2)$.*

b. *$S_m^*\to\sigma$ in conditional probability, that is, for $\epsilon>0$,*

$$P\left\{|\,S_m^*-\sigma\,|>\epsilon\mid X_1,\ldots,X_n\right\}\to 0,\ \textit{almost surely}.$$

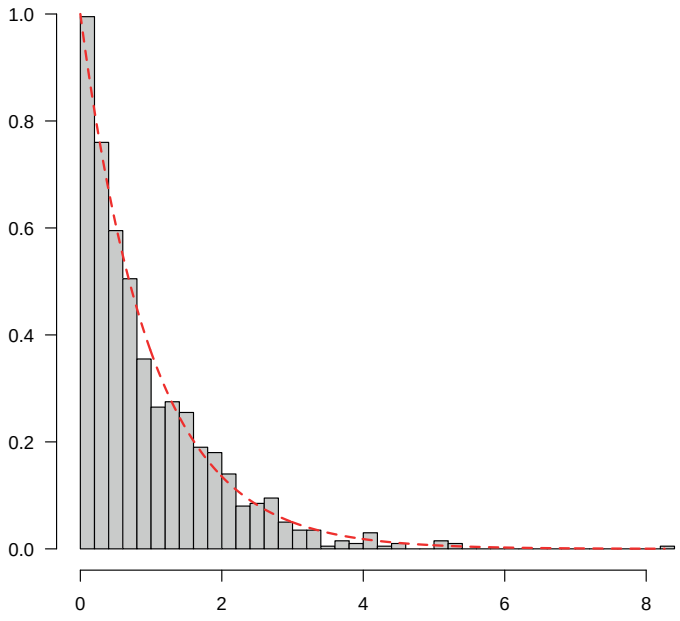
Wat de bootstrap onder andere onderscheidt van gewone simulatie is dat we (voorwaardelijk) bijna zekere convergentie *langs bijna alle rijen* $X_1,X_2,\ldots$ krijgen. We zouden het ongelukkig kunnen treffen en met een heel onwaarschijnlijke rij en daardoor met erratisch gedrag van de bootstrap steekproeven te maken kunnen hebben. Maar we kunnen nu eenmaal met slechte steekproeven te maken krijgen en dat is 'all in the game'.

Ook is het zo dat de bovenstaande stelling gaat over een situatie waarin we de bootstrap eigenlijk niet nodig hebben en waar de centrale limietstelling ons al geeft wat we nodig hebben. Maar Bickel en Freedman merken terecht op dat het interessant is om te zien dat de bootstrap het doet in een situatie die we ook analytisch kunnen behandelen.

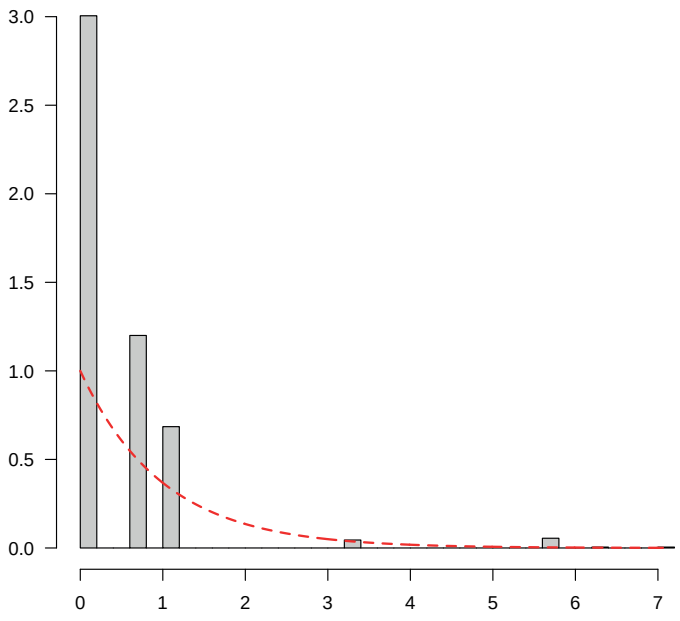
Iemand in het gehoor vroeg bij de lezing van Freedman: "Is het niet zo dat de bootstrap bij de mediaan niet 'werkt'?" De (steekproef) mediaan is de (een) 'middelste' waarneming bij de geordende waarnemingen als de steekproefgrootte oneven is en het gemiddelde van de twee middelste waarnemingen als de steekproefgrootte even is. Freedman antwoordde "That's a lie!" Dat vond ik toen heel grappig. Dus niet gewoon "Nee, het werkt ook voor de mediaan", maar "That's a lie!"

Dat de bootstrap 'werkt' voor de mediaan betekent dat we hiervoor een soortgelijke stelling als Theorem 1 kunnen formuleren en dat de verdeling van de mediaan in het door de bootstrapsteekproeven gesimuleerde gedrag voorwaardelijk op de oorspronkelijke steekproeven bijna zeker naar dezelfde limietverdeling convergeert als de verdeling van de mediaan in de oorspronkelijke steekproef.

'Werkt' de bootstrap altijd? Nee, natuurlijk niet. In [1] wordt het voorbeeld van het maximum van een uniforme steekproef gegeven. Als $X_1,\ldots,X_n$ een steekproef uit een uniforme verdeling op het interval $[0,\theta]$ is, geldt voor het maximum, dat meestal met $X_{(n)}$ wordt aangeduid, dat $n(\theta-X_{(n)})/\theta$ in verdeling naar een standaard exponentiële verdeling convergeert. Als de klassieke



Figuur 1 Histogram van $n(X_{(n)}-X^*_{(n)})/X_{(n)}$, gebaseerd op 1000 parametrische bootstrap steekproeven $X^*_1,\ldots,X^*_n$, waarbij $X_{(n)}$ is gebaseerd op één oorspronkelijke steekproef $X_1,\ldots,X_n$, met $n=1000$, uit de uniforme verdeling op $[0,2]$. De rode gestreepte kromme is de standaard exponentiële dichtheid.



Figuur 2 Histogram van $n(X_{(n)}-X^*_{(n)})/X_{(n)}$, gebaseerd op 1000 niet-parametrische bootstrap steekproeven $X^*_1,\ldots,X^*_n$ met teruglegging uit dezelfde steekproef $X_1,\ldots,X_n$ als in Figuur 1. De rode gestreepte kromme is weer de standaard exponentiële dichtheid.

bootstrap zou werken voor het maximum, zou $n(X_{(n)}-X^*_{(n)})/X_{(n)}$, voorwaardelijk op $X_1,\ldots,X_n$, met kans 1 ook naar een standaard exponentiële verdeling moeten convergeren, Merk op dat hier de rol van θ door $X_{(n)}$ is overgenomen en de rol van $X_{(n)}$ door $X^*_{(n)}$ in de bootstrap-steekproeven.

Maar de voorwaardelijke verdeling van $n(X_{(n)}-X^*_{(n)})/X_{(n)}$ heeft in dit geval helemaal geen limiet! Wel kunnen we in dit geval de klassieke zogenaamde niet-parametrische bootstrap vervangen door de parametrische bootstrap die wel werkt. In dat geval genereren we *uniforme* bootstrap-steekproeven $X^*_1,\ldots,X^*_n$ uit $[0,X_{(n)}]$ en berekenen het maximum $X^*_{(n)}$ in deze bootstrap-steekproeven.

Dat de parametrische bootstrap *wel* werkt betekent dat de voorwaardelijke limietverdeling van $n(X_{(n)}-X^*_{(n)})/X_{(n)}$ wel bijna zeker dezelfde is als die van $n(\theta-X_{(n)})/\theta$, wat wordt geïllustreerd door Figuur 1, waarin het histogram als dichtheidsschatter wordt gebruikt (relatieve frequentie delen door wijdte van staaf).

Dat het helemaal misgaat voor de oorspronkelijke 'klassieke' bootstrap wordt getoond in Figuur 2. In feite kan worden aangetoond dat in dit geval de voorwaardelijke kans dat $n(X_{(n)}-X^*_{(n)})/X_{(n)}=0$ bijna zeker convergeert naar $1-e^{-1}\approx 0{,}63$. Hierdoor zal de hoogte van de linker staaf van het histogram bijna zeker naar oneindig convergeren als $n\to\infty$ en de staafwijdte naar 0 gaat.

Referenties

1 Peter J. Bickel en David A. Freedman, Some asymptotic theory for the bootstrap, *The Annals of Statistics* 9(6) (1981), 1196–1217.

2 Bradley Efron, Bootstrap methods: another look at the jackknife, *Ann. Statist.* 7(1) (1979), 1–26.

3 Bradley Efron, Computers and the theory of statistics: thinking the unthinkable, *SIAM Rev.* 21(4) (1979), 460–480.

4 Dimitris N. Politis, Joseph P. Romano en Michael Wolf, *Subsampling*, Springer Series in Statistics, Springer, 1999.

5 Gerard Scheltens, La Sonate de Vinteuil, 2017, https://www.opusklassiek.nl/cd-recensies/cd-gsch/gschdebussyo1.htm.