STAtOR

periodiek van de VVS jaargang 9 nummer 1, maart 2008

STAtOR
Jaargang 9, nummer 1, maart 2008

STAtOR is een uitgave van de Vereniging voor Statistiek en Operationele Research (VVS). *STAtOR* wil leden, bedrijven en overige geïnteresseerden op de hoogte houden van ontwikkelingen en nieuws over toepassingen van statistiek en operationele research. Verschijnt 4 keer per jaar.

Redactie
Goos Kant (hoofdredacteur), Ana Isabel Barros, Mirjam Moerbeek, Gerrit Stemerdink (eindredacteur), Fred Steutel, Hilde Tobi, Marnix Zoutenbier.

Kopij en reacties richten aan
Prof. dr. G. Kant (hoofdredacteur), Faculteit der Economische Wetenschappen van de Universiteit van Tilburg, Postbus 90153, 5000 LE Tilburg, telefoon 013 - 4668234, mobiel 06-11045089, <G.Kant@uvt.nl>.

Bestuur van de VVS
Voorzitter: prof. dr. R. Gill <voorzitter@vvs-or.nl>
Secretaris: dr. C.G.H. Diks <c.g.h.diks@uva.nl>
Penningmeester: prof. dr. ir. C.A.G.M. van Montfort <kvmontfort@feweb.vu.nl>
Statistische dag: prof. dr. A.W. van der Vaart <aad@cs.vu.nl>
Namens de Bedrijfssectie (BDS):
prof. dr. R.J.M.M. Does <R.J.M.M.Does@uva.nl>
Namens de Biometrische Sectie (BMS):
prof. dr. A.H. Zwinderman <a.h.zwinderman@amc.uva.nl>
Namens de Economische Sectie (ECS):
dr. P.H.F.M. van Casteren <casteren@fee.uva.nl>
Namens het Ned. Genootschap voor Besliskunde (NGB):
prof. dr. J.J. van de Klundert <j.vandeklundert@math.unimaas.nl>
Namens de Sectie Mathematische Statistiek (SMS):
dr. P.J.C. Spreij <spreij@science. uva.nl>
Namens de Sociaal Wetenschappelijke Sectie (SWS):
prof. dr. J.K. Vermunt <j.k.vermunt@uvt.nl>

Leden- en abonnementenadministratie van de VVS
VVS, Postbus 2095, 2990 DB Barendrecht, telefoon 0180 - 623796, fax 0180 - 623670, e-mail <admin@vvs-or.nl>. Raadpleeg onze website over hoe u lid kunt worden van de VVS of een abonnement kunt nemen op *STAtOR* of op een van de andere periodieken.

VVS-website
http://www.vvs-or.nl

Advertenties
Monique van Hootegem, Moeflonstraat 5, 6531 JS Nijmegen, e-mail <hootegem@xs4all.nl>. *STAtOR* verschijnt in maart, juni, september en december.

Ontwerp en opmaak
Pharos / M. van Hootegem, Nijmegen

Uitgever

ISSN 1567-3383

Inhoud

Groei en verandering

Groei en verandering zijn processen waar wij dagelijks mee te maken hebben. Data die door dit soort processen gegenereerd wordt levert vaak een schat aan informatie en een uitdaging voor statistici op het gebied van modellering. Een proces waarmee wij allemaal mee te maken hebben gehad is de groei van ons lichaam tijdens onze kinderjaren. Het consultatiebureau gebruikt groeidiagrammen om de groei van de Nederlandse bevolking te documenteren. Over de constructie van deze groeidiagrammen gaat de bijdrage van Stef van Buuren.

De voorkeur van een persoon voor een bepaalde politieke partij kan over de tijd veranderen. De bijdrage van Mark de Rooij presenteert een classificatiemodel waarmee het stemgedrag voorspeld kan worden uit stemgedrag in het verleden en bepaalde achtergrondsvariabelen. De verandering in het stemgedrag ten tijde van de moord op Pim Fortuyn wordt hierbij als illustratie gebruikt.

Een gekozen regering zal zich moeten buigen over het begrotingsbeleid. De afgelopen vijftien jaar werd hiervoor gebruik gemaakt van scenario's voor de economische groei. In de voorspelling van deze groei wordt een voorzichtigheidsmarge gebruikt. Maar hoe groot moet deze marge eigenlijk zijn? Dat valt te lezen in de bijdrage van Henk Don en Freek van den Eijnden.

De huidige regering stelt zich trouwens als doel om het percentage vrouwen in de top van het bedrijfsleven en de universitaire wereld te laten groeien. Fred Steutel laat ons zien welke vrouwen er in de top van de mathematische statistiek te vinden zijn.

Dat er verschillen tussen mannen en vrouwen zijn is trouwens al langer bekend. Deze verschillen kunnen onder ander toegeschreven worden aan een verschillend genoom. Maar er valt nog veel meer informatie te halen uit de studie van ons genoom. Helene Thygesen en Koos Zwinderman laten zien hoe Poisson modellen toegepast kunnen worden in genomics-onderzoek.

Bij het schrijven van een artikel over cijfers moet je een aantal valkuilen vermijden, zoals te lezen is in de bijdrage van Ivo Gorissen. Hij geeft ons een aantal tips voor het schrijven van een artikel. Wellicht kunt u deze tips gebruiken om uw eigen schrijfstijl te optimaliseren. En met dit woord zijn we weer terug bij de bindende factor in dit nummer: groei en verandering.

Veel leesplezier!
De redactie

DE PARLEMENTAIRE VERKIEZINGEN 2002-2003

verandering nader bekeken door middel van een ideaalpunt classificatie model

Mark de Rooij

Het begin van de eenentwintigste eeuw kenmerkte zich door een sterke beroering in de Nederlandse politiek. Pim Fortuyn zorgde ervoor dat de traditionele balans in de Tweede Kamer zwaar geweld aangedaan werd, en de grote politieke partijen wisten zich geen raad. In mei 2002, vlak voor de parlementaire verkiezingen, werd Fortuyn vermoord, wat tot een ontwrichting van de gehele Nederlandse samenleving leidde. De parlementaire verkiezingen gingen toch door, aan de democratie wordt niet getornd, waarna het eerste kabinet Balkenende aan de macht kwam (CDA/LPF/VVD). Dit kabinet hield echter niet lang stand waardoor een herhaalde gang naar de stembus in 2003 nodig was, welke het kabinet Balkenende 2 opleverde, een coalitie van CDA, VVD en D66.

Stemintentie en het werkelijke stemgedrag

De Stichting Kiezersonderzoek die rond elke nationale gang naar de stembus de meningen peilt, verzamelt daarbij vele achtergrondvariabelen. Deze gegevens zijn vrij verkrijgbaar via het Steinmetz archief (Irwin, Van Holsteyn & Den Ridder, 2003). Een selectie van deze data zal hier geanalyseerd worden om naderhand de veranderingspatronen in deze roerige periode nog eens onder de loep te nemen. Van een steekproef van Nederlanders zijn gegevens beschikbaar op drie tijdstippen: een stemintentie vóór de verkiezingen van 2002, de werkelijke stem in 2002 en de stem in 2003.

We beperken ons tot de acht grootste partijen (PvdA, CDA, VVD, D66, GL, CU, LPF en SP). De steekproefgrootte is n=872, met volledige gegevens op ieder van de drie tijdstippen. Een histogram van de verdeling van de stemmen op ieder van de drie tijdstippen is gegeven in figuur 1, waar we zien dat er aanzienlijke veranderingen plaatsvonden. De LPF lijdt een flink verlies in 2003, het CDA wint geleidelijk over de tijd en de PvdA wint fors in 2003. We zijn geïnteresseerd in de veranderingsvraag: welke overgangen van partij naar partij worden vaak gemaakt en welke minder. Naast

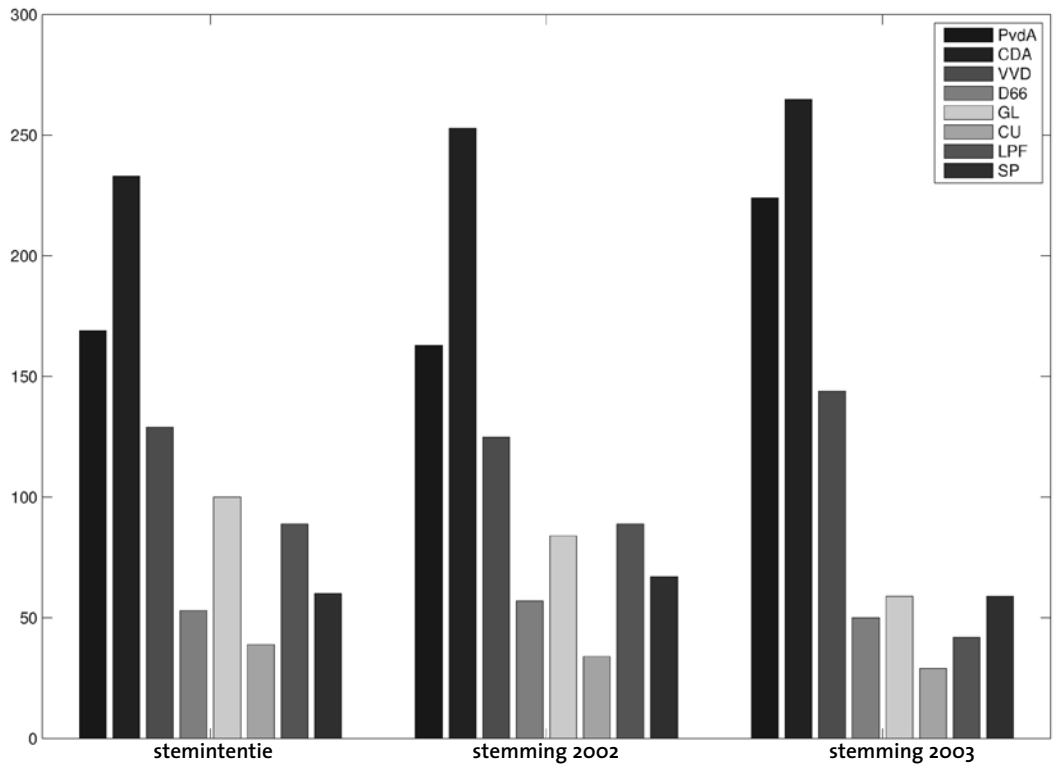
de keuzes op de drie tijdstippen hebben we twee achtergrondvariabelen, te weten sociale klasse en mate van verstedelijking. Sociale klasse is gemeten op een vijfpuntsschaal (*-2* arbeidersklasse; *-1* hogere arbeidersklasse; *0* middenklasse; *1* hogere middenklasse; *2* topklasse) en is een zelfinschatting van de persoon in één van deze categorieën. Mate van verstedelijking heeft de volgende categorieën: *-2* 1-499 adressen/km^2; *-1* 500-999 adressen/km^2; *0* 1000-1499 adressen/km^2; *1* 1500-2499 adressen/km^2; *2* >2500 adressen/km^2.

In de gegeven tijdsperiode lag er veel nadruk op de vraag hoe om te gaan met immigratie en allochtonen. Het kan verwacht worden dat mensen in steden – die dagelijks allochtonen zien in hun omgeving – een andere opinie hebben over deze omgang dan mensen die op het platteland wonen en daardoor ander stemgedrag vertonen. Voor mensen van een lagere sociale klasse kan het verwacht worden dat allochtonen hun baan inpikken, terwijl dit voor de hogere klassen minder waarschijnlijk is, wederom kan dit leiden tot verschillend stemgedrag over de tijd.

Het Model

Een ideaalpunt model

We verwachten dat de keuze van een persoon voor een politieke partij afhangt van de afstand tussen een ideaalbeeld van een politieke partij en het punt van de politieke partij zelf. Neem als voorbeeld dat een persoon vindt dat er geen buitenlanders meer toegelaten mogen worden: hij of zij is dan meer geneigd voor een partij te stemmen die als standpunt heeft `de grenzen gaan dicht' dan voor een partij die daar een meer genuanceerd beeld over heeft. Iedere persoon heeft zo een mening betreffende enkele zaken terwijl ieder van de politieke partijen dat ook heeft. De persoon zal dus kiezen voor een partij die dicht bij zijn standpunten ligt. Dit ideaalpunt van een



Figuur 1: Stemverdeling van de 872 personen over de acht politieke partijen op de drie tijdstippen.

persoon geven we weer met het symbool $\mathbf{y}_{it}$, het ideaalpunt van persoon i, (i=1,...,n) op tijdstip t. Iedere politieke partij heeft ook een positie, weergegeven door $\mathbf{z}_j$ (j=1, ..., G) waarin j de verschillende politieke partijen weergeeft. Het ideaalpunt wordt bepaald als een lineaire combinatie van de voorspellende variabelen. In ons geval zijn deze de vorige stem (intentie), sociale klasse, mate van verstedelijking, en eventueel een interactieterm van deze variabelen. Geef deze verklarende variabelen weer als $\mathbf{x}_{it}$ dan is $\mathbf{y}_{it} = \mathbf{B}^{\mathrm{T}}\mathbf{x}_{it}$. Nu we de ingrediënten van ons model hebben kunnen we de kans opschrijven dat een persoon i op tijdstip t voor partij j kiest, die is

$$\pi_{jt}(\mathbf{x}_{it}) = \frac{\exp\left(-d^2\left(\mathbf{y}_{it}, \mathbf{z}_j\right)\right)}{\sum_l \exp\left(-d^2\left(\mathbf{y}_{it}, \mathbf{z}_l\right)\right)}$$

Hierin staat d^2 voor de gekwadrateerde Euclidische ruimte van dimensionaliteit M. Een belangrijk hulpmiddel voor de interpretatie is de 'odds' van het kiezen voor de ene partij ten opzichte van een andere. De odds in het bovengegeven afstandsmodel wordt bepaald door de afstanden van een persoon tot twee politieke partijen. De odds is in het voordeel van de dichtstbijzijnde partij.

Factorizatie van de likelihood

We geven de keuzes van de persoon op de drie tijdstippen weer door $\mathbf{G}_i = (G_{i1}, G_{i2}, G_{i3})^{\mathrm{T}}$. Voor een overgangsmodel kan de gezamenlijke verdeling van de drie uitkomsten worden opgedeeld

$$f(G_{i1}, G_{i2}, G_{i3} \mid x_i) =$$
$$f(G_{i1} \mid x_i) f(G_{i2} \mid G_{i1}, x_i) f(G_{i3} \mid G_{i1}, G_{i2}, x_i)$$

Overgangsmodellen maken gebruik van deze opdeling en Bonney (1987) heeft laten zien dat standaardsoftware gebruikt kan worden voor de schatting van dit soort modellen door de matrix met voorspellende variabelen op een juiste manier te coderen, dat wil zeggen door $\mathbf{X}$ adequaat te definiëren. We nemen een multinomiale verdeling aan voor de uitkomsten waarin de kansen zijn gedefinieerd door middel van het afstandsmodel. De te schatten parameters zijn de regressiegewichten $\mathbf{B}$ en de posities van de politieke partijen $\mathbf{z}_j$.

De uitkomst

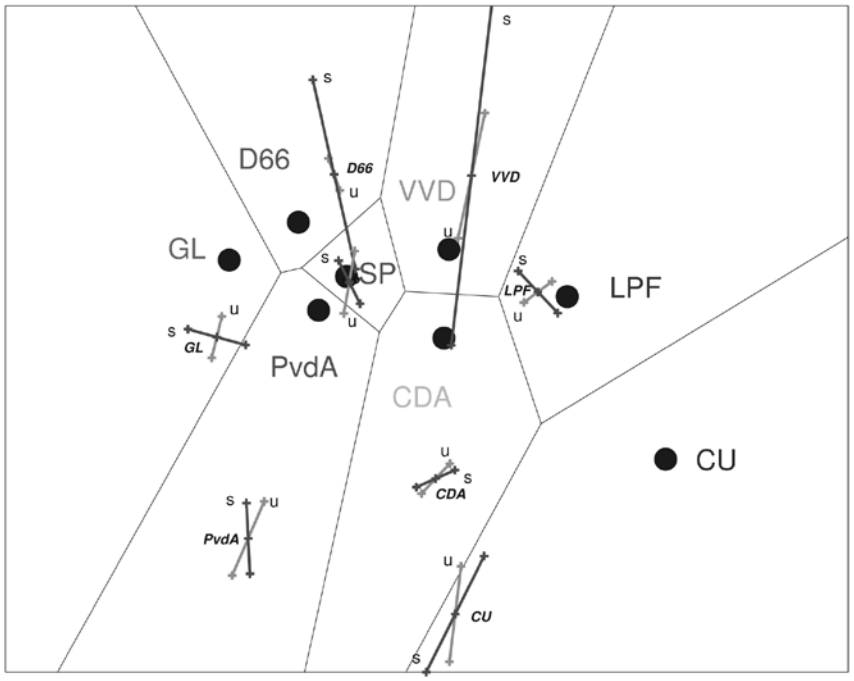
De likelihood ratio statistieken wijzen uit dat zowel de hoofdeffecten van de vorige keuze en de twee achtergrondvariabelen als een interactie tussen de vorige keuze en sociale klasse en mate van verstedelijking een rol spelen in de verklaring van de huidige stem. De oplossing is te zien in figuur 2. De grote stippen geven de posities van de politieke partijen weer ($\mathbf{z}_j$). De politieke partijen liggen zoals men kan verwachten, behalve de Socialistische Partij. Het lijkt erop dat deze partij het midden van het politieke spectrum heeft veroverd. We laten in figuur 2 niet alleen de posities van de partijen zien, maar ook de grenzen tussen twee partijen waar de odds in balans zijn.

De vorige stem wordt weergegeven door het centrum van de 'kruizen'. Over het algemeen zien we dat deze dicht bij de actuele stem ligt, wat betekent dat de meerderheid niet verandert. Sommige posities liggen dicht tegen de grenzen met een ander gebied aan, zie bijvoorbeeld de posities van LPF, CU en GL. Voor alle partijen geldt dat de vorige keuze in het gebied ligt waar dezelfde stem de hoogste kans heeft.

Het ideaalpunt ($\mathbf{y}_{it}$) wordt bepaald door de combinatie van vorige stem en de variabelen sociale klasse en mate van verstedelijking. Het feit dat er een interactie is tussen de sociale klasse of mate van verstedelijking enerzijds en vorige keuze anderzijds wil zeggen dat de richtingen en sterkte van de eerste twee variabelen voor iedere

vorige keuze anders is. Zo zien we dat voor CDA kiezers de twee variabelen nauwelijks verschil maken (de lijnen zijn heel klein) terwijl voor de VVD kiezers sociale klasse een grote rol speelt. De variabelen zijn weergegeven met de markers voor -2 en +2 waarbij het label van de variabele bij het positieve einde staat. De marker voor -1 (niet weergegeven) ligt dus in het midden van het nulpunt en de marker -2. Om het ideaalpunt van een specifieke persoon te bepalen nemen we zijn vorige keuze en vanaf deze positie tellen we de vectoren vanaf het nulpunt tot de markers behorende bij de persoon op ieder van de variabelen op. We kunnen. in andere woorden, een raster om ieder van de vorige keuzes leggen dat bepaald is door de twee variabelen. Dit raster geeft de ideaalpunten van de personen. Gebruik makend van deze informatie kunnen de volgende resultaten uit de grafiek worden gehaald:

- Personen die GL gestemd hebben en in de werkende klasse zitten hebben een verhoogde kans PvdA te stemmen;
- Personen die D66 gestemd hebben en in de werkende klasse zitten hebben een grote kans om SP te stemmen;
- Personen die in een grote stad wonen en eerder CU stemden neigen tot een CDA stem.



Figuur 2: Twee dimensionale oplossing. De grote stippen geven de categorieën van de uitkomstvariabele aan, de kleinere stippen de vorige keuze. De variabelen zijn gelabeld met een `s' voor sociale klasse en een `u' voor mate van verstedelijking (urbanization). De labels staan bij de positieve uitkomsten (+2).

Conclusies

We hebben een techniek laten zien waarmee we zeer specifieke hypotheses kunnen toetsen en ook de resultaten op een grafische manier kunnen presenteren. Wanneer we niet een tweedimensionale maar een (*G*-1)-dimensionale ruimte te gebruiken is het afstandsmodel gelijk aan het multinomiale logit model. Voor lagere dimensionaliteit schatten we minder parameters en hebben dus meer power voor het vinden van effecten. Daarnaast hebben we de mogelijkheid de resultaten door een specifieke biplot weer te geven. Een dergelijke grafische weergave leert ons meer dan alleen de regressiegewichten van het multinomiale logit model waarin de interactie-effecten moeilijk samen te vatten zijn.

Literatuur

Bonney, G.E. (1987). Logistic regression for dependent binary observations. *Biometrics*, 43, 951–973.

Irwin, G.A., Van Holsteyn, J.J.M & Den Ridder, J.M. (2003). *Dutch parliamentary election study 2002-2003: an enterprise of the foundation for electoral research in the netherlands (SKON)*, (computerfile) Steinmetz Archives, Amsterdam (P1628).

Mark de Rooij is universitair docent aan het Instituut voor Psychologisch Onderzoek van de Leidse Universiteit. Zijn onderzoek richt zich op statistische methoden voor longitudinale categorische data door middel van meerdimensionale schaaltechnieken. Hiervoor kreeg hij in 2006 van NWO een VIDI subsidie. Dit onderzoek werd uitgevoerd terwijl de auteur werd gesponsord door de Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO), vernieuwingsimpuls subsidie no. 452-06-002.
E-mail: <rooijm@fsw.leidenuniv.nl>.

Ambitie en voorzichtigheid in het begrotingsbeleid

Wat te doen bij meerdere uitkomsten per studie?

Van 1994 tot 2007 werd het begrotingsbeleid in Nederland gebaseerd op voorzichtige scenario's voor de economische groei. Tegenvallers zijn erger dan meevallers; het gebruik van een voorzichtig scenario verkleint de kans op tegenvallers. Maar hoe groot moet de voorzichtigheidsmarge zijn?

Henk Don en Freek van den Eijnden

Het trendmatige begrotingsbeleid dat de Nederlandse overheid sinds 1994 voert, heeft als centraal element een vooraf vastgelegde norm voor de ontwikkeling van de reële collectieve uitgaven, beter bekend als het uitgavenkader. Daarnaast wordt ook vooraf het saldo van lastenverlichtende en lastenverzwarende maatregelen vastgelegd. Het uitgavenkader en het lastensaldo worden aan het begin van een kabinetsperiode zó gekozen dat de doelstelling voor het begrotings-

saldo wordt bereikt bij een voorzichtige raming van de economische groei op middellange termijn. Beide paarse kabinetten (1994-1998 en 1998-2002) werkten met behoedzame scenario's waarin de economische groei jaarlijks ruim 0,6% lager lag dan de beste voorspelling tijdens de formatie. Vanaf 2002 is de voorzichtigheidsmarge verkleind naar 0,25%; begin dit jaar is hij verder teruggebracht tot nul. Wat is de achtergrond van de voorzichtigheidsmarge? Wat bepaalt zijn omvang? Welke marge is redelijk?

Tegenvallers zijn erger dan meevallers

In het advies dat ten grondslag lag aan het trendmatige begrotingsbeleid pleitte de Studiegroep Begrotingsruimte met nadruk voor het gebruik van voorzichtige uitgangspunten bij de raming van de economische groei, dit om de kans op budgettaire tegenvallers op voorhand zeer klein te maken (Studiegroep Begrotingsruimte, 1993). Dit was nodig om een herhaling van de jaren zeventig te voorkomen, toen opeenvolgende tegenvallers in de trendmatige economische groei het failliet betekenden van het structurele begrotingsbeleid dat door Jelle Zijlstra was ontwikkeld. De afkeer van tegenvallers betekent een asymmetrische doelfunctie voor het begrotingssaldo: tegenvallers zijn erger dan (even grote) meevallers.

Rond de eeuwwisseling adviseerden de Sociaal-Economische Raad (2000) en de Studiegroep (2001) om over te stappen naar een scenario waarin de groeiveronderstelling dichter bij de trendmatige groeiverwachting zou komen te liggen. Als reden werd gegeven dat de overheidsfinanciën de gevarenzone leken te hebben verlaten. En in 1999-2000 was gebleken dat grote meevallers de budgettaire spelregels onder druk zetten en zo hun eigen nadelen met zich meebrengen. De doelfunctie was dus minder asymmetrisch geworden, zodat de kans op tegenvallers wat groter mocht zijn. De kwantitatieve bepaling van de gewenste mate van voorzichtigheid bleef echter in nevelen gehuld.

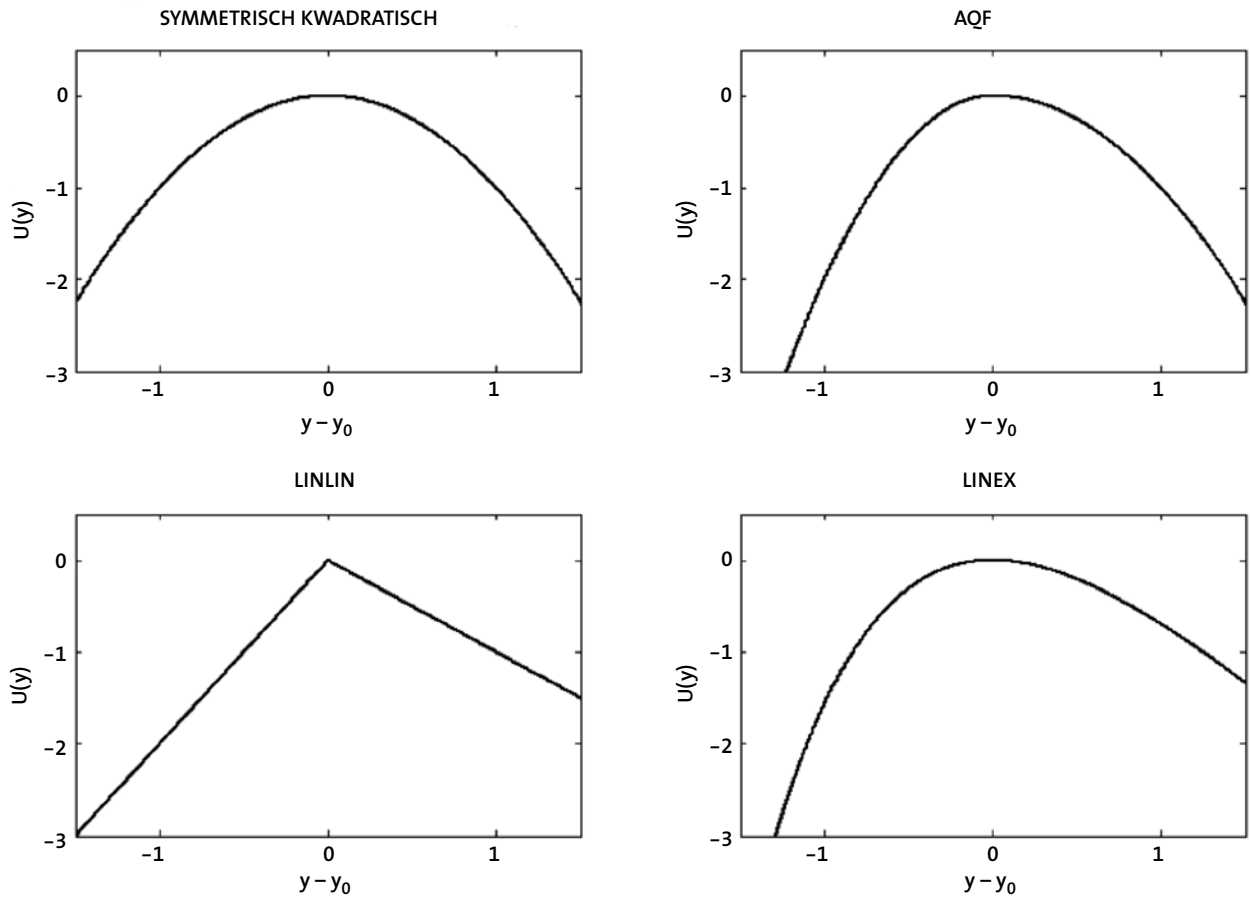
Bepaling voorzichtigheidsmarge

De optimale omvang van de voorzichtigheidsmarge kan worden afgeleid uit de mate van onzekerheid over het begrotingssaldo en de karakteristieken van de asymmetrische doelfunctie. Bij een gewone kwadratische doelfunctie geldt het beginsel van zekerheidsequivalentie: de verwachte waarde van de doelfunctie kan gemaximeerd worden door de doelfunctie te maximeren voor het scenario waarin de onzekere grootheid op haar verwachte waarde wordt gezet (Theil, 1958). Door de asymmetrie treedt een verschuiving op: de onzekerheid kan worden ingeruild voor een scenario waarin de onzekere grootheid een waarde aanneemt die iets verschoven is ten opzichte van haar verwachting. Deze verschuiving wordt aangeduid als de voorzichtigheidspremie (Kimball, 1990). Ook is het hanteren van een voorzichtig scenario equivalent met het kiezen van een hogere ambitiewaarde voor het begrotingssaldo in het basisscenario; de optimale verschuiving in ambitie kan worden afgeleid uit de voorzichtigheidspremie (Don, 2007a).

Om deze theorie te kunnen toepassen moeten we de onzekerheid over het begrotingssaldo kwantificeren en een geschikte doelfunctie voor het begrotingssaldo specificeren.

De mate van onzekerheid

Het trendmatige begrotingsbeleid richt zich op het structurele begrotingssaldo, zodat conjuncturele onzekerheden buiten beschouwing blijven.



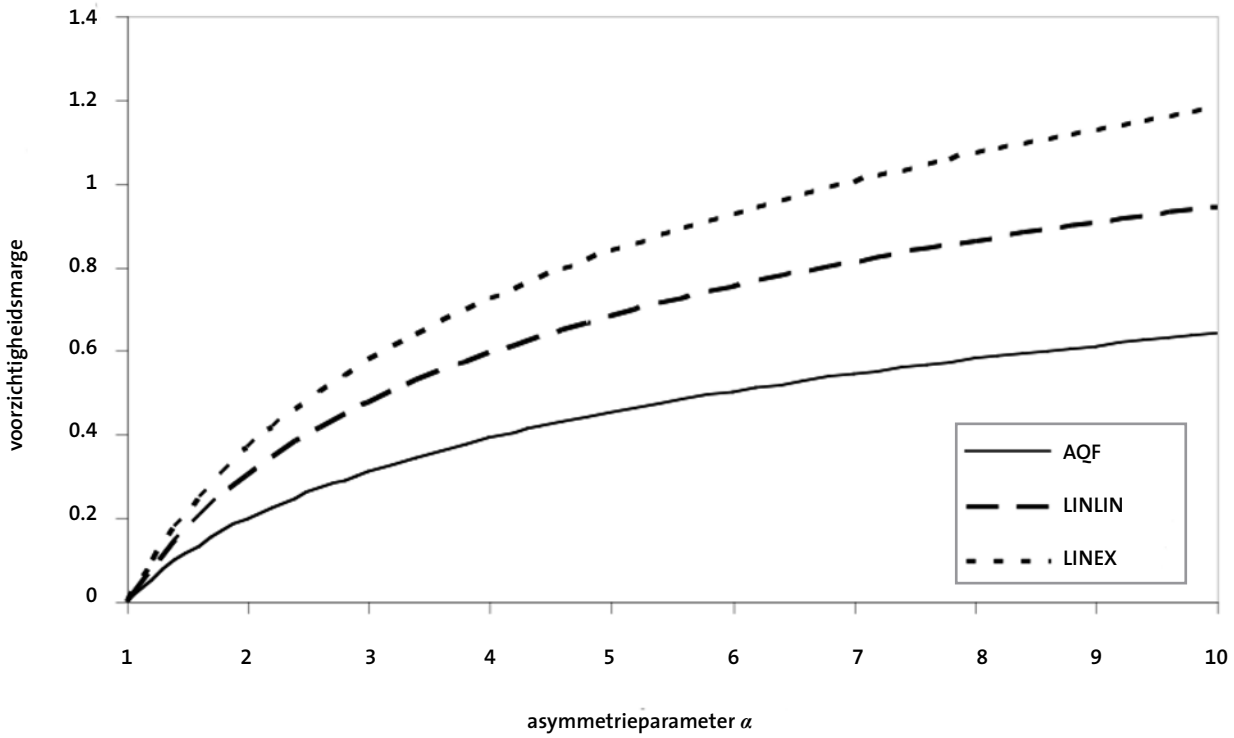
Figuur 1: Vier doelfuncties

De relevante onzekerheden in het structurele saldo hebben betrekking op het structurele groeitempo van de nationale economie en op niet aan de groei gerelateerde fluctuaties in het structurele saldo. Bij een horizon van vier jaar (de normale horizon van een kabinetsperiode) kan de totale relevante onzekerheid in het structurele begrotingssaldo worden geschat op een standaardfout van ongeveer 0,85%-punt van het bruto binnenlands product (BBP), zie Don en Van den Eijnden (2007). In de navolgende berekeningen wordt aangenomen dat deze onzekerheid zich gedraagt volgens een normale verdeling.

Een geschikte doelfunctie

Figuur 1 toont de drie asymmetrische doelfuncties die in de literatuur het meest gangbaar zijn: LINLIN, AQF en LINEX. Ter vergelijking is ook de gewone (symmetrische) kwadratische functie afgebeeld. LINLIN en AQF zijn aan weerszijden van het maximum bij y_0 lineair resp. kwadratisch, met verschillende coëfficiënten; LINEX neigt aan de ene zijde van het maximum naar een lineaire functie en aan de andere zijde naar een exponentiële functie.

Als asymmetrische variant van de gewone



Figuur 2: De voorzichtigheidsmarge in de groeiraming (in % jaarlijkse groei) als functie van de karakteristieken van de doelfunctie. Bron: Don (2007b)

kwadratische functie is AQF een voor de hand liggende keuze.We zullen nagaan hoe gevoelig de resultaten zijn voor een andere keuze. Voor LINLIN en AQF wordt de mate van asymmetrie bepaald door de waarde van een parameter a, $a > 0$. Een tegenvaller ($y < y_0$) is a maal erger dan een even grote meevaller ($y > y_0$). LINEX kent een parameter θ; tegenvallers zijn erger dan (even grote) meevallers als $\theta > 0$; de verhouding is echter niet constant. We kiezen θ steeds zo dat een tegenvaller van eenmaal de standaardfout a maal erger is dan een even grote meevaller. In de grafieken van figuur 1 geldt $a = 2$, een waarde die ons in eerste instantie redelijk lijkt. Ook op dit punt zullen we de gevoeligheid van de resultaten nagaan.

Als de doelfunctie U driemaal differentieerbaar is en de tweede afgeleide is strict negatief, dan is de voorzichtigheidspremie bij benadering gelijk aan $\frac{1}{2}\ \alpha.\sigma^2$, waarin $\alpha = -U'''/U''$ een maat is voor de voorzichtigheid die in U besloten ligt en α de standaardfout die de onzekerheid in de doelvariabele karakteriseert (Kimball, 1990). Voor een LINEX-doelfunctie is α constant en gelijk aan $-\theta$; ook geldt de formule voor de voorzichtigheidspremie hier exact bij een normaal verdeelde onzekerheid. Voor LINLIN en AQF geldt de formule zelfs niet bij benadering, omdat hun derde afgeleide in y_0 niet bestaat. Maar bij een normaal verdeelde onzekerheid heeft de voorzichtigheidspremie voor beide functies de vorm $\kappa.\alpha$, waarbij κ afgeleid kan worden uit de asymmetrieparameter α (Granger, 1969; Cain, 1994).

Resultaten

Figuur 2 laat zien hoe de gewenste voorzichtigheidsmarge in de jaarlijkse groei afhangt van de karakteristieken (aard en parameter) van de asymmetrische doelfunctie. De doelfunctie AQF met asymmetrieparameter 2 levert een voorzichtig scenario dat jaarlijks 0,2% onder de verwachte trendmatige groei ligt. De LINLIN functie levert bij dezelfde waarde van de asymmetrieparameter een voorzichtigheidsmarge van 0,3% jaarlijkse groei. LINEX geeft een nog iets grotere marge, van 0,37% jaarlijkse groei. We concluderen dat de marge van 0,25% jaarlijkse groei die sinds 2001 wordt gehanteerd, goed past bij redelijke veronderstellingen over de doelfunctie.

De behoedzame scenario's uit 1993 en 1997 kenden een marge van ruim 0,6% jaarlijkse groei. Deze waarde duidt op een aanzienlijk grotere asymmetrie in de doelfunctie. Bij AQF wordt deze marge pas bereikt bij $a = 9{,}5$, bij LINLIN vinden we deze marge als $a = 4{,}3$. Zulke verhoudingen zijn wellicht een goede weerspiegeling van de situatie in 1993-94, toen het begrotingstekort nog boven de 3% van het BBP lag en de overheidsschuld rond de 80%. Maar anno 2000 was er een overschot op de begroting en kwam de overheidsschuld onder de 60% van het BBP. Daarbij paste een minder zware straf op tegenvallers.

Het nieuwe regeerakkoord

Hoe moeten we het begrotingsbeleid van het nieuwe regeerakkoord nu beoordelen?

Het voorzichtige scenario is verlaten. Vergeleken met (het gewogen gemiddelde van) de verkiezingsprogramma's van de drie coalitiepartijen ging de overstap naar het trendmatige scenario gepaard met een verhoging van de streefwaarde voor het begrotingssaldo met 0,1% BBP (Don, 2007b). Maar een voorzichtigheidsmarge van 0,25% jaarlijkse groei is equivalent met een ambitieverschuiving voor het begrotingssaldo met 0,3% BBP, dus beduidend meer dan de feitelijke verschuiving. Kennelijk was de prijs die betaald moest worden voor een hogere ambitie bij het begrotingssaldo te hoog: voor drie partijen is de pijn van tekortbeperkende maatregelen bij andere beleidsdoelstellingen al gauw groter dan voor één partij. Ten opzichte van het gewogen gemiddelde van de drie verkiezingsprogramma's is het schrappen van de voorzichtigheidsmarge in het nieuwe regeerakkoord deels gebruikt voor hogere ambities op andere terreinen, zoals veiligheid en milieu.

Literatuur

Cain, M. (1994). Prediction Adjustments for Asymmetric Quadratic Loss with a Gaussian Process, *Journal of the Operational Research Society, 45 (10)*, 1179-1184.

Don, F. J. H. (2007a). *Ambitie en voorzichtigheid in het economisch beleid*, oratie Erasmus Universiteit Rotterdam.

Don, Henk (2007b). Ambitie begrotingsbeleid valt tegen. *Economisch Statistische Berichten, 4507*, 196-199.

Don, F. J. H. and F. O. van den Eijnden (2007). Prudence and ambition in fiscal policy. *Medium Econometrische Toepassingen, 15 (2)*, 2-9.

Granger, C. W. J. (1969). Prediction with a Generalized Cost of Error Function. *OR 20 (2)*, 199-207.

Kimball, M. S. (1990). Precautionary Saving in the Small and in the Large, *Econometrica*, 58 (1), 53-73.

SER (2000). *Sociaal-economisch beleid 2000-2004*, Advies 00/08. Den Haag: SER.

Studiegroep Begrotingsruimte (1993), *Naar een trendmatig begrotingsbeleid*, Negende rapport. Den Haag: Ministerie van Financiën.

Studiegroep Begrotingsruimte (2001). *Stabiel en duurzaam begroten*, Elfde rapport. Den Haag: Ministerie van Financiën.

Theil, H. (1958). *Economic Forecasts and Policy*.Amsterdam: North Holland.

Henk Don is bijzonder hoogleraar Econometrie en Economisch Beleid aan de Erasmus Universiteit Rotterdam; van 1994 tot 2006 was hij directeur van het Centraal Planbureau. E-mail: <don@few.eur.nl>.

Freek van den Eijnden studeert voor zijn MSc in Econometrics and Management Science aan de Erasmus Universiteit Rotterdam en werkte in 2006 2007 als student-assistent bij het Econometrisch Instituut.

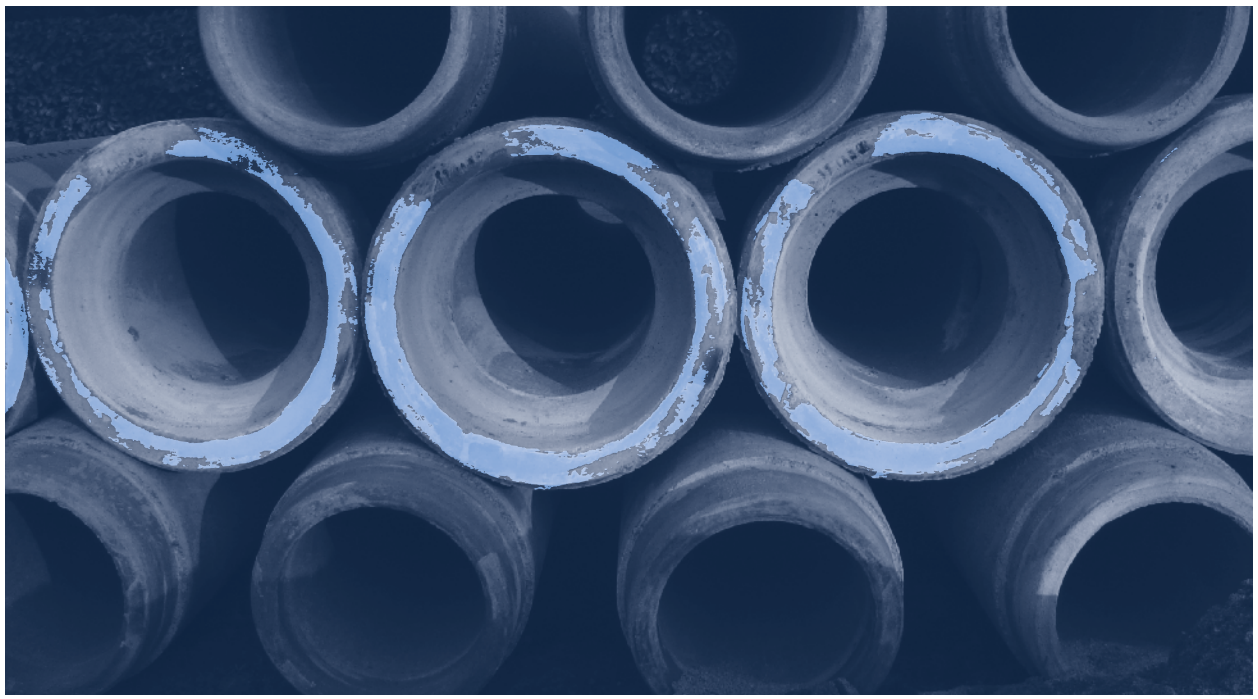


Foto: Pieter Bosch

HIËRARCHISCHE POISSON-MODELLEN

schatters en toepassingen in genomics-onderzoek

Helene H. Thygesen en A.H. Zwinderman

Voor veel telprocessen is een hiërarchisch Poisson-model een natuurlijke benadering: het aantal gebeurtenissen y_i in experiment *i* volgt een Poissonverdeling met intensiteit λ_{ji}, en over een groot aantal experimenten heen volgt de (latente) statistische grootheid $\{\lambda_j, i = 1..N\}$ weer bijvoorbeeld een gamma-verdeling of lognormaal-verdeling. De gamma-verdeling is een populaire keuze vanwege haar mooie analytische eigenschappen: de marginale verdeling van y_i wordt een negatieve binomiaalverdeling, en de posterior verdeling van $\lambda_i|y_i$ wordt een gamma-verdeling. Toch is een lognormaal-Poisson-model soms een natuurlijker model. In dit artikel beschrijven we een schatter voor het multivariate lognormaal-Poissonmodel en een toepassing daarvan in genomics-onderzoek, meer specifiek voor het interpreteren van sequencing data en SAGE (Serial Analysis of Gene Expression)-data.

Beschouw een experiment waarin een complex biochemisch proces wordt bestudeerd aan de hand van het aantal kopieën y van een bepaalde molecuul in een klein monster van het eindproduct. Een natuurlijk model voor dit experiment is dat er een groot aantal kopieën M zijn en dat de gebeurtenissen dat kopie i in het monster terecht komt, i=1..M onafhankelijk van elkaar zijn en kleine kansen p_i=**p** hebben. Dan volgt *y* een Poissonverdeling met als parameter λ=**pM**.

Maar waar **N** tellingen y_i, *i*=1..N beschikbaar zijn (voor verschillende soorten moleculen en/of verschillende experimenten), wordt *y overdispersed* omdat de scheikundige concentraties M_i, *i*=1..**N** (en dus de λ's) verschillen. Een model voor scheikundige concentratie is vaak iets als

$$M \sim A_1^{q1}\ A_2^{q2}\ A_3^{q3}\ \ldots$$

waar de factoren A_1^{q1} , A_2^{q2} etc. bijvoorbeeld concentraties van enzymen en reactanten kunnen zijn. De multiplicatieve vorm van het model betekent dat het model op log-schaal additief wordt, en dus zorgt de centrale limietstelling er voor dat log(M) normaal verdeeld wordt als er veel factoren zijn die min of meer onafhankelijk van elkaar zijn. Misschien zijn er een paar van de factoren, bijvoorbeeld A_1 en A_2, die een grote rol spelen en gerelateerd zijn aan covariaten in onze experiment. In dat geval krijgen we een standaard lineair model

$$\log(\lambda) = \log(p) + q_1 \log(A_1) + q_2 \log(A_2) + \varepsilon$$
$$\varepsilon = q_3 \log(A_3) + q_4 \log(A_4) + \ldots \text{ (normaal verdeeld)}$$

We hebben dit model toegepast op SAGE-data (Serial Analysis of Gene Expression). In SAGE wordt de concentratie van de transcriptieproducten (mRNA) van genen gemeten aan de hand van het aantal moleculen in een klein monster. Als je honderd monsters (patiënten) hebt en 30.000 soorten mRNA (genen), krijg je dus een 30.000 keer 100 tabel van (meestal kleine) tellingen. De covariaten zijn meestal gerelateerd aan het soort patiëntmateriaal (bloed of spier, kanker of controle, behandeld of onbehandeld) maar kunnen ook samenhangen met het soort gen (eiwitcoderend of niet, welk chromosoom, welke biologische functie etc.)

Schatter, univariaat geval

Je hebt SAGE-data van één patiënt en wilt de parameters μ en σ in de verdeling van log(λ) schatten. Helaas is de marginale verdeling van *y,* gegeven μ en σ, niet in gesloten vorm beschikbaar, en de ML-schatter ook niet. Wel kunnen we de verwachte momenten (gemiddelde en variantie) van *y* theoretisch berekenen. De moment-methode, waarmee μ en σ worden geschat zodat de theoretische momenten overeenkomen met de empirische momenten, leidt echter tot een zeer grote standaardfout, en dat is ook goed te begrijpen: het histogram van *y* is scheef en dus heeft de rechterstaart onevenredig veel invloed op de variantie. Je zou dan kunnen voorstellen om de momenten van log(*y*) te gebruiken, net zoals je zou doen als je de λ's in plaats van de *y* kon waarnemen. Helaas, dan krijgen de lage tellingen onevenredig veel invloed (als er maar één *y*=0 voorkomt wordt E(log(*y*))= -∞ en VAR(log(*y*))= ∞).

De ML-schatter voor het univariate geval is feitelijk niet moeilijk te berekenen met moderne computers. Je schrijft $L(y|\mu,\sigma) = \int \text{Poisson}(y|\lambda)\ \text{Normaldens}(\log(\lambda)|\mu,\sigma)\ d\log(\lambda)$, dat geëvalueerd kan worden met numerieke integratie voor al de kleine y's, bijvoorbeeld y<10. Voor de grotere y's kun je ook (als je geen numerieke integratie wilt doen), Bulmer's approximatie gebruiken. (Google op Bulmer lognormal Poisson voor details). Als we L kunnen evalueren, kunnen we het ook maxima-

liseren, bijvoorbeeld met Newton's methode.

Toch willen wij hier een andere methode noemen, omdat wij die nodig hebben in meer complexe gevallen waarin wij de likelihood-functie niet kunnen evalueren. Het idee is dat wij de moment-methode kunnen verbeteren door de data te transformeren zodat de variantie stabiel wordt. Als *y* Poisson-verdeeld is met parameter λ, dan is de variantie van $\sqrt{y}$ approximatief stabiel. De momenten van $\sqrt{y}$ zijn daarom approximatief sufficiënt voor het schatten van μ en σ. Dus er moet een redelijk goede schatter voor μ en σ zijn die op de momenten van $\sqrt{y}$ gebaseerd is. Er is geen gesloten-vorm van zo'n schatter maar wat we kunnen doen is het verband tussen de $\sqrt{y}$-momenten en (μ,σ)-waardes vast leggen door een regressieanalyse uit te voeren op gesimuleerde data.

Schatter, bivariaat geval

De volgende stap is de schatting van een bivariaat normale verdeling voor een dataset van twee patiënten. Dus voor elk gen *i* volgen $y_{i,1}$ en $y_{i,2}$ Poisson-verdelingen met parameters $\lambda_{i,1}$ en $\lambda_{i,2}$, en over de genen heen volgt $(\log(\lambda_{.,1}), \log(\lambda_{.,2}))$ een bivariaat normaalverdeling. Om zo'n bivariate verdeling te schatten hebben wij de parameters in de marginale verdelingen voor $(\log(\lambda_{.,1}), \log(\lambda_{.,2}))$ nodig, plus de correlatie tussen de twee. De parameters in de marginale verdelingen worden geschat met de ML-methode. Vervolgens passen we het simuleren- en regressie-trucje toe: voor een aantal waardes van de correlatie, lopend van -1 tot 1, simuleren we datasets met de zojuist gevonden marginale-verdelings-parameters, en dan berekenen we de wortel-kwadraatsom $\sum\sqrt{(y_{i,1\ i,2})}$. Tenslotte fitten we het verband tussen de correlatie en de wortel-kwadraatsom (bijvoorbeeld met een spline) en interpoleren de ware correlatie van de empirische wortel-kwadraatsom.

Hogere-orde inferentie

Nu we de parameters in een multivariaat normale verdeling hebben geschat, is de weg plaveid voor alle leuke dingen die je met normaal verdeelde data kunt doen. Principale-componenten-analyse, bijvoorbeeld.

Waar je uiteindelijk naar toe wilt is het schatten van de invloed van patiënt-gerelateerde covariaten op de expressie van individuele genen. Voor elk gen *i* krijg je een model van het type

$$\log(\lambda_{i,j}) = b_i + \text{sum}\,(\,a_{k,i}\,x_{k,j}\,) + \varepsilon_{i,j}$$

waar $x_{k,j}$ de waarde is van de k'de covariaat voor de *j*'de patiënt. Je kunt dit model rechtstreeks schatten op basis van de tellingen $y_{i,j}$, $j=1...\#_{\text{patiënten}}$. (Methode van Vencio, zie referentie). Dat zou echter zonde zijn omdat het aantal genen veel groter is dan het aantal patiënten, zodat je veel informatie verliest door de genen één voor één te analyseren. Beter is een empirisch-Bayesiaanse benadering waar je eerst de verdeling (over de genen heen) van de coëfficiënten *a* en *b* afleidt uit de covariantie-matrix en vervolgens per gen deze verdeling als prior gebruikt in een Bayesiaans model voor de tellingen van dat gen. In principe zou je ook een volledig Bayesiaanse model kunnen fitten voor de hele dataset maar dat kost vooralsnog te veel rekenkracht.

Waarom niet gewoon de gamma-verdeling?

Voordat we de schatter voor het multivariate log-normaal Poisson-model ontwikkelden, probeerden we het eerst met een gamma Poisson-model, waar de schatters in gesloten vorm geschreven kunnen worden. Het grootste nadeel van de gamma-verdeling is dat hij domweg niet een goede fit oplevert voor SAGE-data. De staart van de verdeling is te dun en moest aangevuld worden met een

non-parametrische component. In het multivariate geval werd het nog erger omdat de gamma-verdeling slechts approximatief gesloten is onder multiplicatie. We kozen voor een additieve variant van de bivariate gamma-verdeling (Google op *Trivariate reduction* gamma voor details) :

$$\lambda_{.,j} = u_{.,j} + w_{.,j,k}\,; \quad \lambda_{.,k} = u_{.,k} + w_{.,j,k}$$

maar die veronderstelt dat de shape-parameter hetzelfde is voor de gedeelde component *w* en de onafhankelijke componenten *u* (wat niet altijd klopt) en bovendien is een additief model biologisch niet plausibel. Tenslotte staat trivariate reductie geen negatieve correlaties toe.

Het gamma-model blijft natuurlijk aantrekkelijk omdat alles in gesloten vorm gaat en je dus niet onze simulatie- en regressie-truc nodig hebt. Vooral bij de Bayesiaanse inferentie per gen is het een groot voordeel dat je de empirische prior analytisch kunt updaten, terwijl je met een lognormaal-prior een MCMC moet draaien of iets dergelijks.

Naast het gamma-Poisson-model is er een andere pragmatische benadering die je kunt gebruiken als je snel resultaten wilt zonder specifieke software: gewoon de wortel van de tellingen nemen en dan in een lineair model stoppen, alsof ze multivariaat normaal verdeeld zijn! Het probleem met die benadering is dat het niet biologisch plausibel is dat de covariaten lineair werken op wortelschaal, dus de coëfficiënten krijgen geen goede biologische interpretatie. Toch werkt het in de praktijk vaak goed als het gaat om, bijvoorbeeld, het clusteren van patiënten in de PCA-ruimte of het identificeren van genen waarvan de expressie met bepaalde patiënt-covariaten correleert. Het is in ieder geval veel beter dan het analyseren van de data op lineaire schaal (de hoge tellingen krijgen dan te veel invloed) of log-schaal (de lage tellingen krijgen dan te veel invloed) of het vergelijken van groepen met een Fisher- of chi-kwadraat-toets (dan wordt de binnen-groepsvariatie genegeerd).

Toepassingen van het model

Bijna alle literatuur-referenties naar het lognormaal-Poisson-model zijn ecologische toepassingen: het aantal door veldbiologen gevonden nesten van vogelsoort *i* in district *j* in maand *k* is Poisson verdeeld met parameter λ_{ijk}, en men probeert dan een lineair model voor λ_{ijk} te schatten. Toch is het ook een heel natuurlijk model voor teldata uit andere disciplines waar grote hoeveelheden teldata worden verzameld, bijvoorbeeld in de archeologie, sociale wetenschappen, sterrenkunde en (internet)-verkeer. Het schatten en de inferentie is een beetje omslachtig en dat is waarschijnlijk de reden waarom het model moeilijk naar andere disciplines migreert.

De SAGE techniek wordt niet meer zoveel gebruikt, maar er is een aantal andere technologieën voor genomisch onderzoek in opkomst. Het sequencen van eiwitten en DNA wordt steeds goedkoper en vindt steeds meer toepassingen. Hier leent het lognormaal-Poisson-model zich waarschijnlijk goed voor, net als bij SAGE.

Bij sequencing-data, waaronder SAGE, is er echter een complicerende factor: een aanzienlijk deel van de getelde moleculen is niet goed geclassificeerd, en dus wordt het waargenomen aantal $y_{i,j}$ de som van het ware aantal plus het aantal ten onrechte als *i* geclassificeerde moleculen. Er bestaat een aantal methoden om hiermee rekening te houden:

1. Het negeren van de gen-tags (RNA-sequensen), die niet in bekende genen voorkomen. Voor goed bestudeerde organismen zoals mensen, muizen en fruitvliegen is de lijst van genen en de bijhorende RNA-sequensen tegenwoordig vrij compleet. (zie <bioinfo.amc.uva.nl/HTMseq>)
2. Ook kan je de de kans dat een telling van

een bepaalde sequens op een meetfout berust berekenen aan de hand van het aantal gemeten sequensen die erop lijken: Als bijvoorbeeld de sequens AAAAAAAAAG slechts een keer is geteld en de sequens AAAAAAAAAT honderd keer, is de kans groot dat de op G eindigende meting eigenlijk op T had moeten eindigen.

3. Tenslotte kan je de histogrammen (danwel kruistabellen, in het bivariate geval) fitten met een mengmodel met twee componenten, te weten de autentieke metingen en de meetfouten.

Idealiter zouden we deze drie methodes willen integreren in een Bayesiaans raamwerk. Zo ver zijn we nog niet. In de praktijk gebruiken we methode 1, maar schatten we tevens het aantal meetfouten aan de hand van methode 2 en 3 om te verifiëren dat ons model voor de meetfouten correct is. Het blijkt aardig te kloppen, maar niet perfect: een van de problemen is dat de kopieerfouten die tijdens de evolutie tot de diversiteit van het genoom hebben geleid, op de meetfouten lijken. Daarom worden statistisch geschatte meetfouten confounded door de oververtegenwoordiging van op elkaar lijkende sequensen in de natuur. Het is moeilijk om dat in het model in te bouwen.

Literatuur

Thygesen, H.H. and A.H. Zwinderman (2006). Modeling SAGE data with a truncated gamma-Poisson model. *BMC Bioinformatics*, 7, 157.

Vencio R.Z.N. et al. (2004). Bayesian model accounting for within-class biological variability in serial analysis of gene expressions (sage). *BMC Bioinformatics,* 5, 119.

Helene H. Thygesen is verbonden aan het Medical and Pharmaceutical Statistics Research Unit, Lancaster University. E-mail: <helene.h.thygesen@gmail.com>. A.H. Zwinderman is hoogleraar Biostatistiek bij de Afdeling Klinische Epidemiologie en Biostatistiek van het Academisch Medisch Centrum van de Universiteit van Amsterdam. Email: <a.h.zwinderman@amc.uva.nl>.

column

LOF DER GECIJFERDHEID

Fred Steutel

Statistiek is mooi, zelfs statistieken kunnen mooi zijn, maar je moet er wel voor kunnen rekenen. Het is opmerkelijk hoeveel je met eenvoudig rekenwerk kunt achterhalen. Het schijnt dat een mathematisch fysicus aan de hand van een kleurenopname van een ontploffende atoombom allerlei gegevens kon afleiden over de werking van dit *device*, die als 'zeer geheim' waren geclassificeerd. Omgekeerd denken sommige mensen dat je de afstand van de aarde tot de zon kunt 'uitrekenen', als je maar heel knap kunt rekenen; dat daar meetgegevens voor nodig zijn komt niet bij ze op.

Meetgegevens, daar gaat het om in de statistiek. Maar als je wat hebt, dan kun je ook wat. Als je niet kunt rekenen of geen gevoel voor getallen hebt, kom je natuurlijk nergens, zoals doorgaans in de dagbladen. Zo stond onlangs stond in *de Volkskrant* een diagram waarin de aantallen

sigaretten werden weergegeven die per jaar in Nederland waren verkocht – in miljoenen. Voor 2006 kwam daar 14 uit: 14 miljoen sigaretten in een jaar. Het is duidelijk dat dat niet kan kloppen: nog niet één sigaret per inwoner per jaar: er is een factor duizend zoek. Immers, als een kwart van de Nederlanders dagelijks tien sigaretten rookt, kom je op vier miljoen maal tien maal 365, dat is 14 miljard. Het aan accijns opgehaalde bedrag stond ook in miljoenen euro's: twee. Dat moest dus ook twee miljard euro zijn. Per sigaret is dat 14 eurocent en voor de gemiddelde roker dus zo'n vijfhonderd euro per jaar. Hij zal er niet door weerhouden worden.

Een voorbeeld van extreme ongecijferdheid. In een interview in *NRC Handelsblad* van Leijendekker met Paul Schnabel stond 'Tussen het jaar duizend en achthonderd jaar later was de gemiddelde groei nog geen een procent. Daarom is er in die tijd maar weinig veranderd.' Weet iemand wat een groei van een procent per jaar doet in achthonderd jaar? Een groei met een factor 2865 of 286400 procent! Groei met 0,6 procent per jaar ('nog geen één procent') geeft een factor 120. Er stond trouwens niet in het verhaal wàt er groeide en hoe dat werd gemeten.

Bij sommige getallen heb je het gevoel dat er instanties zijn die de zaken zorgvuldig schimmig houden. Over het tot stand komen van aardolieprijzen is maar weinig bekend. Sommige dingen mogen we weten. De totale hoeveelheid opgepompte ruwe olie per jaar: 35 miljard vaten van 159 liter, de prijs van een vat (159 liter) ruwe olie op de *spotmarkt*: 95 dollar, de benzineprijs aan de pomp: 1,50 euro. Maar zelfs hoe die 1,50 euro is opgebouwd weten we niet precies. En wat we helemaal niet weten is wat er aan de Saudische sjeiks wordt betaald voor een vat olie, of wie er profiteren van prijsstijgingen op de *spotmarkt* die geen materiële, maar een politieke of psychologische oorzaken hebben. Wat we wel weten is dat de winsten van de oliemaatschappijen stijgen met een stijgende olieprijs. Hoe die winsten tot stand komen blijft schimmig. Het zou goed zijn als iemand eens precies uitzocht waar al het oliegeld blijft. Iets voor een parlementaire enquête?

Een tweede voorbeeld is de bierprijs. Ook daar is vrij veel schimmigheid. Wat zijn de productiekosten van een vat bier en wat zou een redelijke prijs voor een pilsje zijn? In de jaarcijfers van Heineken lees ik: 'vaste' productiekosten van vier miljard euro, exclusief marketing en verkoopkosten, voor 120 miljoen hectoliter bier. Dat is 0,33 euro per liter. Doe er wat bij voor verkoop en marketing: 50 cent per liter. De horeca betaalt per liter meestal tussen de 1.00 en 1,80 euro, maar er zijn enorme verschillen, en kroegbazen worden door bierproducenten soms in een ijzeren greep gehouden. Supermarkten betalen soms maar één euro per liter. Heineken zou dus bij productiekosten van 0,50 een winst van 0,50 euro per liter maken. Uit de winstcijfers blijkt echter dat de (netto) winst ongeveer 0,05 euro per lier is. Blijft over de vraag: waarom 0,25 liter bier in een café bijna twee euro kost, en een 0,50 liter bier het dubbele. Dat laatste is onredelijk omdat de prijs van een glaasje bier maar voor een heel klein deel door de prijs van het bier bepaald wordt: 0,25 liter kost de uitbater nog geen 0,40 cent.

Statistiek is een schone zaak, maar sommige belangrijke en minder belangrijke zaken worden er niet door opgehelderd.

Fred Steutel *is emeritus hoogleraar kansrekening aan de TU Eindhoven. Hij is redacteur van STAtOR.*
E-mail: <f.w.steutel@tue.nl>

SCHRIJVEN OVER CIJFERS

Bij het schrijven van een artikel over cijfers moet je een aantal valkuilen vermijden, zoals te technisch schrijven, te veel details laten zien, te veel facetten willen bespreken, te snel aan de slag willen, je te zeer vereenzelvigen met je artikel. Uitgangspunt is dat je geen artikel moet schrijven om te laten zien wat je allemaal weet, maar om de lezer een helder verhaal te vertellen. Nadat het artikel helemaal af is, moet je wat tijd inruimen om te leren hoe je de volgende keer sneller en beter te werk kunt gaan.

Ivo Gorissen

Je hebt je onderzoek afgesloten en je wil of moet publiceren. Als dat zonder problemen tot het gewenste resultaat leidt, lees dan niet verder. Je verdoet je tijd en kunt beter een van de andere artikelen uit dit tijdschrift gaan lezen.

Voor de anderen gaan we verder. Voor je begint met schrijven moet je je kritisch afvragen of je een goede en gedegen analyse hebt uitgevoerd en of je niet nog ondergedompeld bent in de cijfers, maar er daadwerkelijk 'boven' staat. Of, anders gezegd of je de cijfers niet alleen kent, maar ook het verhaal dat ze vertellen. Pas dan kun je beginnen met het schrijfproces. En... neem daar de tijd voor! Want het is zonde als een onderzoek waarin je veel tijd en energie hebt gestopt niet goed uit de verf komt omdat de presentatie onvoldoende is.

Maak een plan

Veel onderzoekers maken bij het schrijven van een artikel de fout te beginnen met schrijven. Dat lijkt logisch, maar is het niet. Als je een onderzoek gaat doen dan begin je toch ook niet met onderzoeken. Nee, je verkent de boel eerst een beetje, gaat een plan smeden, toetst dat links en rechts, werkt je plan uit, herschrijft het een paar keer, zorgt voor groen licht en gaat dan pas over tot het doen van onderzoek. Waarom doe je dat met schrijven dan niet zo?

Voor wie is het artikel bedoeld?

Je schrijft niet voor jezelf. Dat weet je natuurlijk zelf ook wel, toch neem je jezelf vaak ongemerkt als referentie. Je schrijft op wat je zelf zou willen

SCHRIJFTIPS

Vertel een verhaal
Zorg dat je een verhaal hebt, dat is iets anders dan een schier eindeloze aaneenrijging van feiten. Het is het antwoord op de vraag: Wat is er aan de hand? Aanknopingspunten daarbij zijn: wat er op dit moment speelt in de media, beleidsagenda's, een herkenbare categorie (ouderen, voedsel, kerosineverbruik), et cetera.

Formuleer zo bondig mogelijk
Als een artikel vlot loopt is de kans groter dat je lezer door gaat met lezen. Een groot deel van de stijltips hieronder geven concrete handvatten voor bondig formuleren.

Begin met de conclusies en werk deze later pas uit
Dan weet je als lezer direct of je het interessant genoeg vindt om door te lezen of beter iets anders kunt gaan doen. Journalisten werken ook zo. Bovendien trek je direct de aandacht van de lezer. En ook al leest iemand het artikel maar voor een deel, je belangrijkste mededelingen heeft hij al opgepikt.

Zorg voor een pakkend begin
In veel gevallen gebeurt dat in de vorm van een lead. Houd deze kort en geef daarin al direct je hoofdconclusie weer. Vermijd hier het gebruik van cijfers.

Stijl
Wat betreft schrijfstijl is er een aantal zaken waar je alert op moet zijn bij het schrijven van een artikel.

- Behandel één hoofdonderwerp per paragraaf.
- Begin elke alinea met een kernzin.
- Schrijf begrijpelijk. Laat je woordkeuze afhangen van je lezerspubliek.
  TIP Leg het voor aan iemand uit je potentiële lezerspubliek.
- Wissel korte en lange zinnen af.
- Maak je zinnen niet te ingewikkeld. Zorg dat woorden die bij elkaar horen dicht bij elkaar staan, ofwel vermijd tangconstructies. Schrijf bijvoorbeeld niet: DANS (Data Archiving and Networked Services), een initiatief van KNAW en NWO, zorgt in samenwerking met onderzoekers voor de dataopslag en blijvende toegankelijkheid van onderzoeksgegevens in de alfa- en gammawetenschappen. Maar wel: DANS (Data Archiving and Networked Services) zorgt voor de dataopslag en blijvende toegankelijkheid van onderzoeksgegevens in de alfa- en gammawetenschappen. DANS is een initiatief van KNAW en NWO en werkt samen met onderzoekers.
  TIP Bekijk alle zinnen die drie of meer regels nodig hebben met de nodige scepsis en hak ze zonodig in tweeën of in drieën.
- Vermijd omhaal van woorden.
  TIP Zoek alle 'Het...dat'- en 'Er....dat'-zinnen en herschrijf ze.
- Schrijf zoveel mogelijk in de actieve vorm.
  TIP Kijk bij elke 'wordt' en 'worden' of je er niet zonder kunt.
- Vermijd overmatig gebruik van cijfers.
  TIP Streep alle cijfers door en bekijk welke je echt nodig hebt.
- Ga niet de data uit de tabel beschrijven.
- Beperk beeldspraak en gebruik geen al te populair taalgebruik.
- Humor is een moeilijk stijlmiddel, dat je maar beter kunt vermijden.

Het gebruik van tabellen, grafieken en andere figuren
Een tabel, grafiek of ander figuur kan duizend woorden waard zijn, maar ook voor complete verwarring zorgen. Houd de presentatie zo helder mogelijk. Je gebruikt grafieken als je de lezer in één oogopslag wil laten zien wat er aan de hand is en de exacte waarden niet het belangrijkste zijn, maar wel: of iets sneller stijgt dan iets anders, of iets groter is dan iets anders. Je gebruikt tabellen als de exacte waarden van de cijfers wel belangrijk zijn. Tabellen kunnen je een grote dienst bewijzen omdat ze je de kans geven om in de tekst nog maar weinig cijfers te gebruiken. Denk ook eens na over andere vormen van presentatie, zoals geografische kaarten, stroomschema's et cetera. Alles waarmee je je verhaal beter kunt vertellen is de moeite meer dan waard.

lezen. Niet doen! Maak een studie van je publiek. Je artikel zal gepubliceerd worden in een tijdschrift of op een website. Dat biedt je dus al je eerste aanknopingspunt. Zo kun je aan de binnenkant van de omslag van deze *STAtOR* zien dat dit een tijdschrift van de Vereniging voor Statistiek en Operationeel Research is dat haar leden, bedrijven en overige geïnteresseerden op de hoogte wil houden van ontwikkelingen en nieuws over toepassingen van statistiek en operationele research. Soms staat op de site van een tijdschrift nog wat extra informatie over de doelgroep en de invalshoek van het tijdschrift. Gebruik deze informatie, want het voorkomt dat je je artikel in een later stadium flink moet aanpassen om alsnog bij de doelstelling van het tijdschrift aan te sluiten. Dat is trouwens ook wel zo fijn voor de tijdschrift- of webredactie. Als het goed is heb je nu een beeld van de mensen voor wie je schrijft: opleidingsniveau, voorkennis en vakgebied. Je kunt je nu bezig gaan houden met de vraag: wat interesseert die mensen?

Waarover ga je schrijven?

Misschien heb je een heel duidelijke boodschap die je wil verkondigen, of is er heel veel informatie uit je onderzoek naar voren gekomen die allemaal de moeite waard is. Als je een duidelijke boodschap hebt, moet je je serieus afvragen of je potentiële lezers hierop zitten te wachten en welke invalshoek je moet kiezen om hun aandacht te krijgen. Als je een grote berg interessante informatie hebt, moet je kiezen. Want zoals Goethe al zei:'In der Beschränkung zeigt sich erst der Meister'. Verplaats je in de positie van je beoogde lezers: Wat houdt hen bezig, wat wordt er op dit moment gepubliceerd op het betrokken vakgebied, welke discussies spelen er en wat kan jouw onderzoek toevoegen? Vervolgens moet je weer kiezen. En wel voor de invalshoek. Vaak is hier ook keuze te over: (sub)categorieën van je onderwerp, niveaus of ontwikkelingen en allerlei indelingen zoals regio, geslacht, schoolsoort, burgerlijke staat et cetera. Allemaal erg interessant, maar ingewikkeld om het allemaal tegelijk te vertellen. Bedenk waar je lezer het meest in geïnteresseerd zal zijn en laat de rest gewoon zitten. Kiezen doet pijn, maar als je een zo groot mogelijk effect wil sorteren met je artikel, dan is dat onvermijdelijk.

Hoe structureer je je artikel?

Als je weet voor wie je gaat schrijven en waarover dan moet je nog nadenken over hoe je de informatie structureert. De eerste aanwijzing daarvoor krijg je door het medium (tijdschrift, internetsite) te bekijken. Je vindt hier informatie over de omvang van het artikel, het gebruik van grafieken en tabellen en andere illustraties, de opbouw van artikelen, het gebruik van een lead, de wijze van indeling in paragrafen en eventuele subparagrafen, het gebruik van noten, noem maar op. Bepaal eerst de hoofdstructuur en ga dit vervolgens verder uitwerken. Schrijven over de uitkomsten van wetenschappelijk onderzoek kan grofweg op twee manieren. Je begint te vertellen hoe je onderzocht hebt en komt uiteindelijk toe aan het presenteren van de resultaten of je begint met het presenteren van je belangrijkste uitkomsten en legt niet of terloops uit hoe je onderzocht hebt. Hierbij is ook weer de vraag waarop je lezer gespitst is. Een belangrijke aanwijzing krijg je ook nu weer van het medium waarin je wil publiceren. Het ene tijdschrift is gericht op het delen van academische kennis met vakgenoten terwijl het andere juist wetenschappelijke informatie wil aanbieden voor het maken van beleidskeuzen. In het eerste geval begin je vaak met een uitgebreide verhandeling over je onderzoek en neem je de resultaten uitgebreid op in een bijlage, terwijl je in het andere geval hoogtepunten uit de uitkomsten presenteert en de onderzoeksmethode (hooguit) in bijvoorbeeld een bijlage vermeldt.

Schrijven

Eindelijk ben je nu zover dat je echt kunt gaan schrijven. Dat schrijfproces kun je op verschillende manieren inrichten. Je kunt gewoon gaan schrijven, zien waar het schip strandt en het eerste concept vervolgens gedegen uitwerken. Je kunt het verhaal ook tegen iemand vertellen, om feedback vragen en dan gaan schrijven. Je kunt er naar streven om het allemaal direct zo goed mogelijk te doen, zodat je weinig nakijktijd kwijt bent. Zoek vooral de manier die voor jou het beste werkt. Maar blijf uitgaan van je oorspronkelijke plan. Het is echter niet de bedoeling dat je daar onnodig star aan vast blijft houden. Blijkt bij de uitwerking dat er een paragraaf bij of weg moet, verander dat. Maar pas wel je plan aan. Dat voorkomt dat je op een bepaald moment op drift raakt en het oorspronkelijke doel uit het vizier verliest.

Als je je eerste concept af hebt volgt een periode van herlezen en herschrijven. Hierbij moet je op een aantal zaken letten, zoals spelfouten, logische opbouw, korte en lange zinnen afwisselen, consequent werkwoordtijdengebruik, niet te passief schrijven, stijlgebruik, te veel omhaal van woorden en nog veel meer. Het is handig als je weet wat jouw zwakke punten zijn.

Als je het concept dan een aantal malen bekeken hebt en je bent tevreden, juich dan niet te snel. Leg het weg en geef het eventueel ook aan een referent te lezen. Die heeft een frisse blik op het artikel. Ga lekker een of twee weken met iets anders aan de slag en pak het vervolgens weer op. Wedden dat je ontdekt dat het nog beter kan. Herhaal dit proces zonodig nog een of twee keer.

En dan is je artikel klaar voor verzending. Vaak breekt er dan een nieuw proces aan: er komt commentaar van de redactie die het onder ogen krijgt. Belangrijk hierbij is dat je je open opstelt, want dan maak je optimaal gebruik van de expertise van de redacteur. Die heeft namelijk in de loop der tijd de nodige kennis en ervaring opgedaan met de eisen van een goed stuk in zijn medium. Geef de onmogelijkheden aan, maar wees vooral creatief in het benutten van de mogelijkheden.

Evalueren

Als het goed is ben je dan waar je wezen wil: Je artikel wordt gepubliceerd! Voor velen stopt het schrijfproces hier. Je doet er echter goed aan om het hele proces nog eens goed op een rijtje te zetten. Meestal blijft het namelijk niet bij het publiceren van dit ene artikel. Tijdens het schrijven heb je heel veel feedback gekregen. Maak daar gebruik van! Noteer wat je volgende keer beter wil doen. Dit is ook het moment om je zwakke punten op taalgebied op te sporen (passief, veel omhaal van woorden et cetera) zodat je die er een volgende keer vroegtijdig uit kunt halen. Daarmee voorkom je tijd- en energieverspilling voor de correctie van oude - en dus vermijdbare - fouten. Het noteren van leerpunten kun je het best al tijdens het proces doen, maar je zou dat in elk geval aan het einde van de rit moeten doen.

Zoals je gezien hebt kost het schrijven van een goed artikel tijd, veel tijd. Maar ik ben er van overtuigd dat het een investering is die zich ruim terugverdiend. Succes.

Met dank aan collega André Mares voor zijn kritische commentaar op een eerdere versie van dit artikel.

Literatuur

Centraal Bureau voor de Statistiek. *Schrijfwijzer*. Voorburg/Heerlen: interne publicatie.
Renkema, J. (2002). *Schrijfwijzer*. Den Haag: Sdu Uitgevers.
United Nations Economic Commission for Europe (2006). *Making Data Meaningful, A guide to writing stories about number*. New York and Genova.

Ivo Gorissen is als wetenschappelijk redacteur werkzaam bij het Centraal Bureau voor de Statistiek. E-mail:<igsn@cbs.nl>.

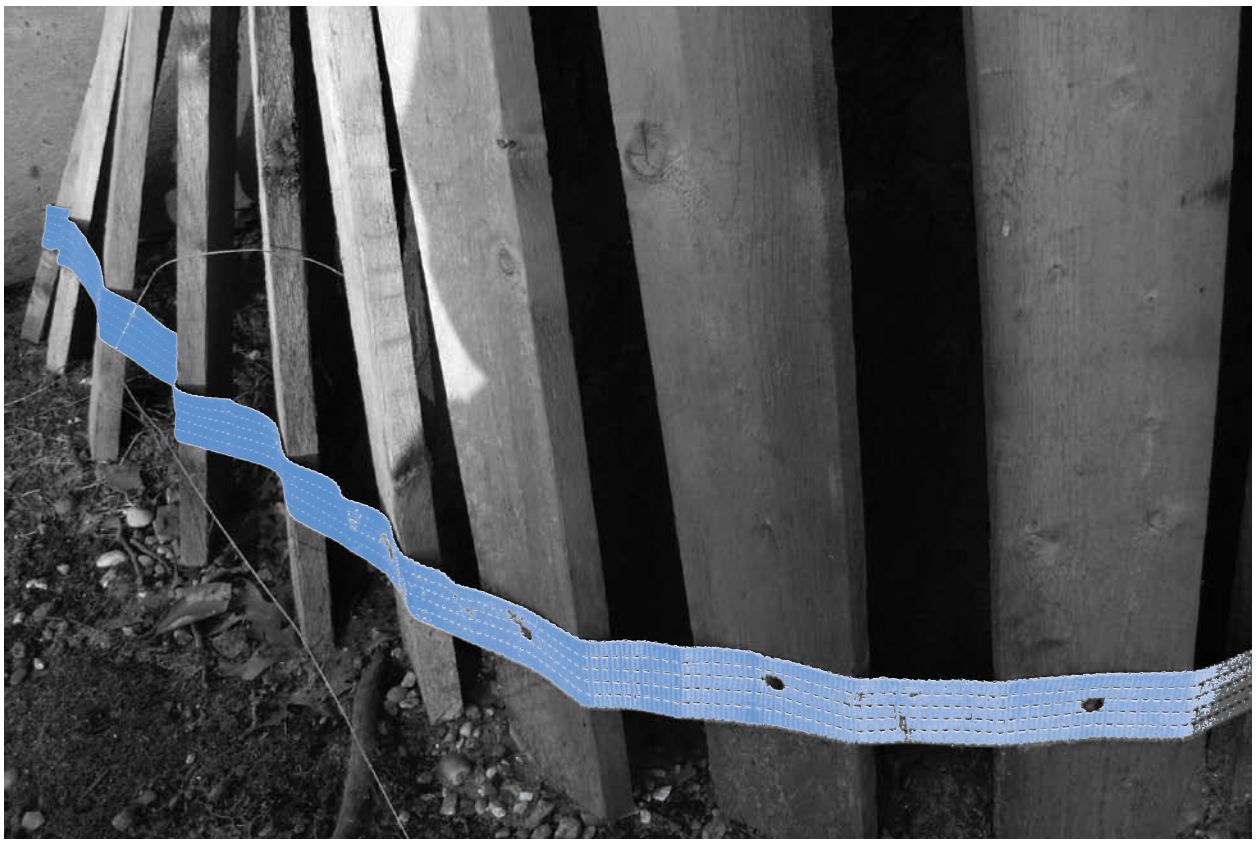


Foto: Pieter Bosch

GROEIDIAGRAMMEN

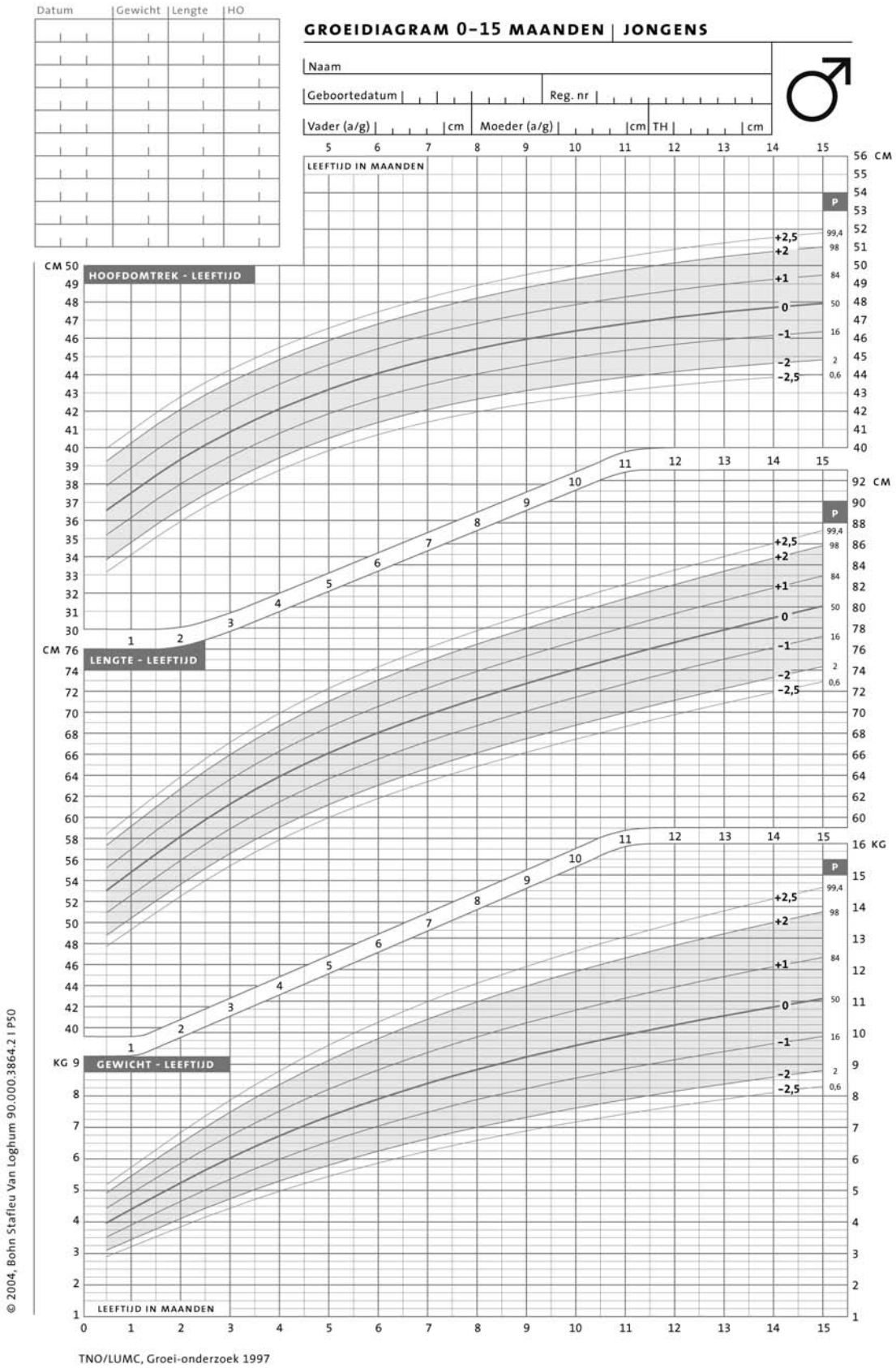
Stef van Buuren

Vrijwel iedere Nederlandse zuigeling bezoekt het consultatiebureau (CB). Het CB heeft tot taak de groei en ontwikkeling van het kind te volgen, en waar nodig door te verwijzen naar de huisarts. De verpleegkundige van het CB meet regelmatig de lengte, het gewicht en de hoofdomtrek van het kind, en tekent vervolgens de meetwaarden in op het groeidiagram. Mocht de groei van het kind afwijken van hetgeen normaal is (te kort, te lang, te licht, te zwaar), dan vindt nader onderzoek plaats naar de mogelijke oorzaken. In Nederland worden jaarlijks ongeveer 200.000 kinderen geboren. Het groeidiagram is daarmee de meest gebruikte grafiek van Nederland.

Figuur 1 op de volgende pagina is het groeidiagram voor jongetjes tot een leeftijd van 15 maanden. Soortgelijke diagrammen zijn beschikbaar voor jongens/meisjes, andere leeftijden, etnische groepen, en andere uitkomsten. In dit stuk ga ik in op de constructie en het gebruik van groeidiagrammen.

Landelijke Groeistudies

In Nederland worden de groeidiagrammen berekend op basis van gegevens uit landelijke groei-



Figuur 1: Groeidiagram voor jongens van 0 - 15 maanden

studies. Dit zijn studies die eenmaal per 10 à 15 jaar worden gehouden, en waarin een grote groep kinderen wordt gemeten en gewogen. De meest recente studie, de Vierde Landelijke Groeistudie, werd in 1997 door TNO en LUMC uitgevoerd. (Fredriks et al, 2000) Eerdere groeistudies stammen uit de jaren 1955, 1965 en 1980. De steekproefomvang van de studies varieert tussen grofweg 20.000 en 50.000 kinderen. Nederland is één van de weinige landen ter wereld waar met een zekere regelmaat een groeistudie plaatsvindt. De groei van Nederlandse kinderen behoort dan ook tot de best gedocumenteerde ter wereld.

De lezer zou zich kunnen afvragen of het wel nodig is elke 10 à 15 jaar een groeistudie uit te voeren. Het blijkt echter dat er aanzienlijke verschillen bestaan in de lengte van opeenvolgende generaties. In het jaar 1860 was de gemiddelde lengte van een 20-jarige jongen gelijk aan 164 cm. In 1997 was dat 184 cm, een verschil van 20 cm in het gemiddelde. De Vierde Landelijke Groeistudie toonde ook aan dat in 1997 kinderen aanzienlijk dikker waren dan in 1980. Dergelijke trends zijn alleen met periodieke studies te vinden.

Figuur 1 bevat een grafische weergave van de verdeling van hoofdomtrek, lengte en gewicht naar leeftijd van gezonde kinderen. De verdelingen zijn niet representatief voor alle kinderen in Nederland, maar vormen een (normatieve) beschrijving van een ideaalpopulatie. Kinderen in de steekproef van de groeistudie met een afwijkende lengte of hoofdomtrek (bijvoorbeeld door ziekte, zeer laag geboortegewicht, etc.) worden buiten de referentiepopulatie gelaten. De omvang van de groep uitvallers is overigens gering, rond de 1 procent. De groeidiagrammen zijn berekend op basis van de referentiepopulatie van gezonde kinderen.

Schatten van centielen

Het lengte-diagram laat zeven SD-lijnen zien, beschreven met labels tussen -2.5 SD en +2.5 SD. De interpretatie van deze labels komt overeen met de bekende *Z*-score. Dus bijvoorbeeld onder de mediane lijn, de '0 SD lijn', valt 50% van de populatie. Enkele jaren geleden is de weergave van percentiellijnen vervangen door SD lijnen. Deze zijn wellicht wat minder intuïtief, maar *Z*-scores vergemakkelijken het werken met extremen en zijn tegenwoordig gangbaar.

Het belangrijkste probleem bij de berekening van de diagrammen is dat de referentieverdeling moet worden gladgestreken in twee richtingen tegelijkertijd: binnen de leeftijd en over de leeftijd. Twintig jaar geleden omschreef Tim Cole het schatten van centielen als 'something of a black art'. Sindsdien zijn er ruim 30 verschillende statistische methoden voor dit probleem ontwikkeld, en geldt 'centile estimation' als een apart subspecialisme. De methoden verschillen in een aantal opzichten:

- sommige methoden schatten elk centiel apart, terwijl andere alle centielen tegelijk schatten, meestal onder de aanname van een verdeling;
- sommige methoden behandelen leeftijd als continu, terwijl anderen verwachten dat leeftijd in groepen is geclassificeerd;
- de gebruikte smoothing methode varieert: splines, kernel regression, polynome regressie, etc.

Borghi et al (2006) hebben criteria voorgesteld om te kiezen tussen de methoden. Deze zijn:

- extreme centielen moeten zo precies mogelijk worden geschat;
- centiellijnen mogen elkaar niet kruisen;
- centielen en *Z*-score moeten met een formule kunnen worden berekend;
- leeftijd moet als continu worden geanalyseerd;

• waar nodig moet de verdeling scheefheid en kurtosis modelleren.

Toepassing van deze criteria leidde tot een shortlist van vijf methoden. De eenvoudigste daarvan is de LMS-methode (Cole & Green, 1992). Uitgezonderd de mogelijkheid voor kurtosis voldoet de LMS-methode aan alle vereisten. De LMS-methode is gebruikt voor het maken van de diagrammen van de Vierde Landelijke Groeistudie.

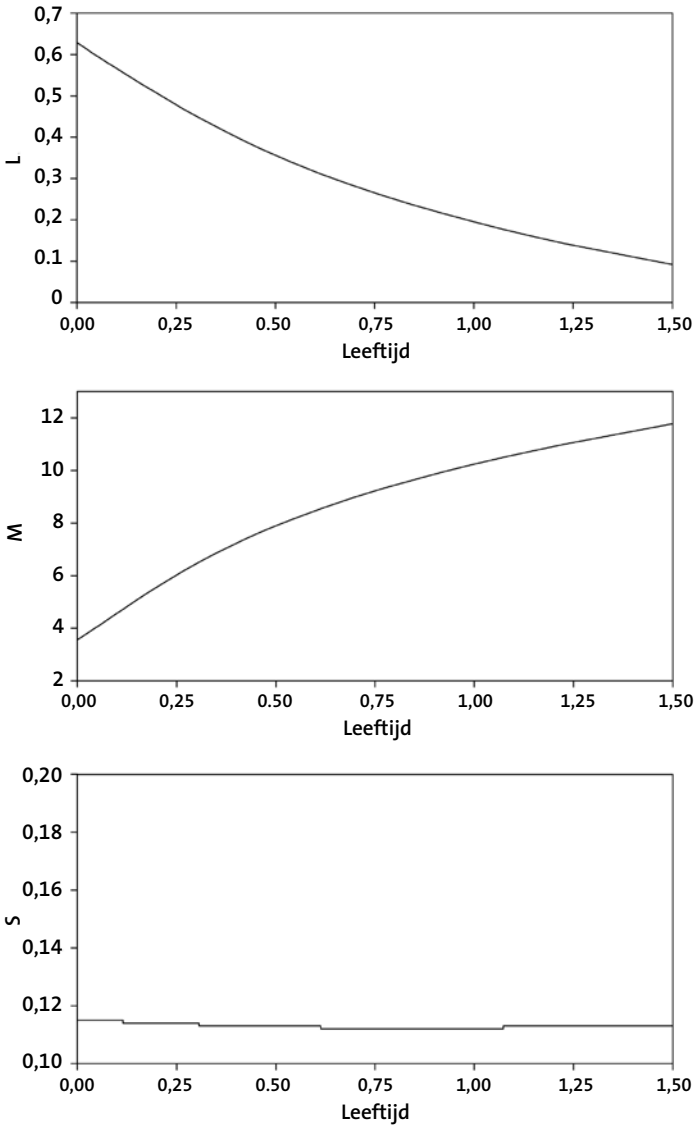
De LMS-methode veronderstelt dat de uitkomstmaat op iedere leeftijd normaal verdeeld is na een Box-Cox transformatie, waarbij de exacte vorm van de transformatie afhangt van leeftijd. De Box-Cox verdeling heeft drie parameters, aangeduid als *L*, *M* en *S*. De leeftijds-afhankelijke *L*-curve modelleert de scheefheid, de *M*-curve beschrijft hoe de mediaan van de uitkomstmaat met leeftijd varieert, en de *S*-curve geeft weer hoe de coefficient of variation met leeftijd samenhangt. De *L*-, *M*- en *S*-curve worden geschat met behulp van penalized maximum likelihood.

Figuur 2 laat zien hoe de *L*-, *M*- en *S*-curves van gewicht met de leeftijd variëren. De *L*-curve loopt van 0.6 (rondom geboorte) naar 0.1 (rond 1.5 jaar). Dit geeft aan dat de verdeling verandert van vrijwel normaal (rondom geboorte) naar vrijwel lognormaal (rondom 1.5 jaar). De *L*-, *M*- en *S*-curven worden op vaste leeftijden getabuleerd, de referentiestandaard.

Met behulp van de referentietabel kan voor iedere gewichtmeting *Y* de *Z*-score worden berekend als volgt:

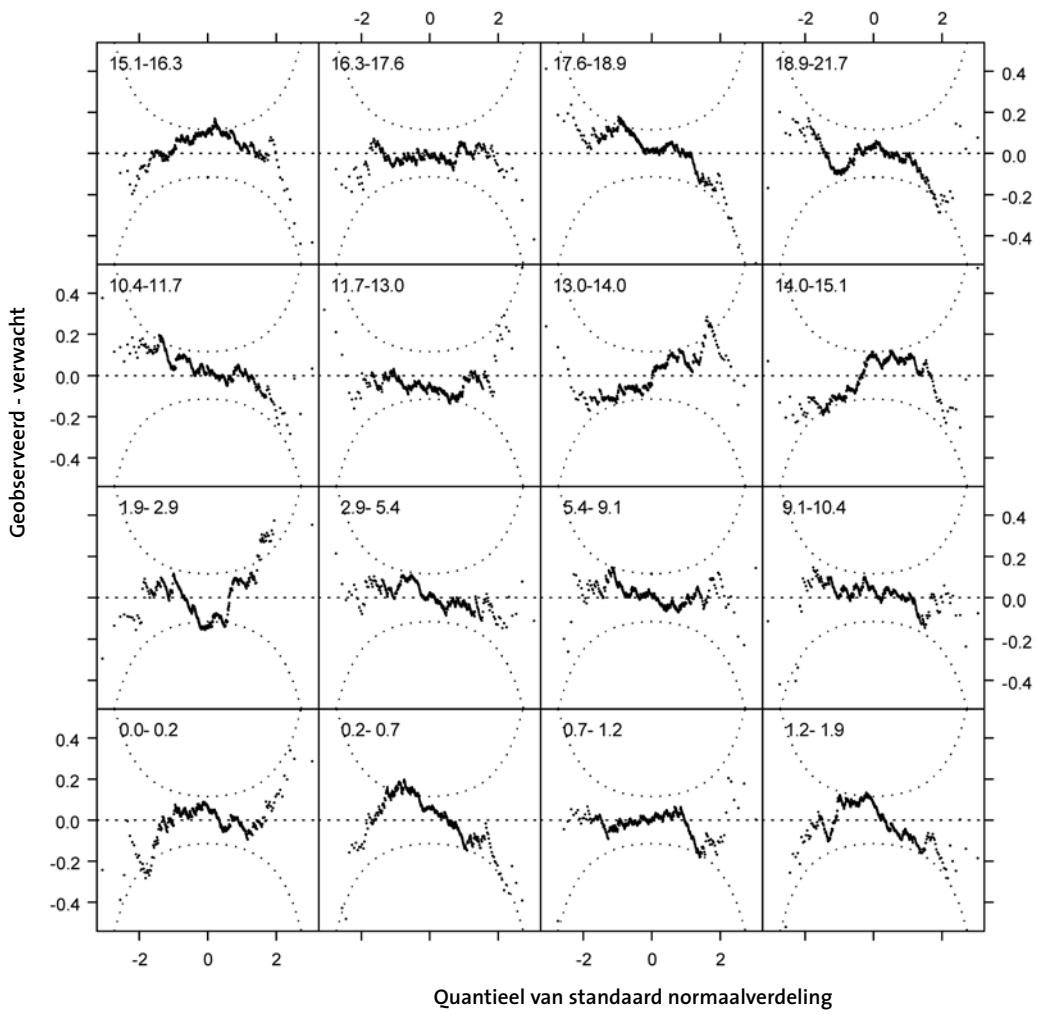
$$Z = \frac{(Y / M_t)^{L_t} - 1}{L_t S_t} \quad \text{indien } L_t \neq 0$$

$$Z = \frac{\ln(Y / M_t)}{S_t} \quad \text{indien } L_t = 0$$



Figuur 2: *L*-, *M*- en *S*-curve van het LMS model voor gewicht (kg) naar leeftijd van Nederlandse jongens 0-1½ jaar.

waarbij L_t , M_t en S_t de referentiewaarden zijn op leeftijd *t*. Voorbeeld: Stel een jongentje van 16 weken oud weegt 8 kg. Invullen levert op *Z* = ((8/6.51)^0.448-1)/(0.448*0.113) = 1.91. Met een *Z*-score van +1.91 is dit jongetje voor zijn leeftijd aan de zware kant. Aan de hand van de *Z*-tabel kunnen we afleiden dat in de referentiepopulatie slechts 2.8% van de leeftijdsgenoten zwaarder is dan 8 kg. Indien de leeftijd van het kind tussen



Figuur 3: Worm plot voor lengte van Nederlandse jongens (0-21 jaar) voor een goed passend LMS model (model 0,10,6R). Dit model is gebruikt om het groeidiagram te tekenen.

de tabelleeftijden in valt, dan worden eerst de *L*-, *M*- en *S*-waarden lineair geïnterpoleerd vanuit de omliggende tabelleeftijden.

Worm plot

Eén van de aspecten die nog niet aan bod is gekomen is de wijze waarop het definitieve LMS-model wordt bepaald. Het LMS-model bevat drie *smoothing* parameters, één voor ieder van van *L*-, *M*-, en *S*-curves. De *smoothing* parameter wordt aangeduid met het aantal vrijheidsgraden. De keuze van het juiste aantal vrijheidsgraden is van groot belang voor het uiteindelijke diagram. De worm plot (Van Buuren & Fredriks 2001) is een hulpmiddel voor het vinden van de juiste waarden.

Het idee achter de worm plot is eenvoudig. Voor een passend model is de *Z*-score van de kinderen in de referentiepopulatie op alle leeftijden normaal verdeeld met gemiddelde 0 en variantie 1.

Figuur 3 is de worm plot van een passend model voor lengte van jongens 0-21 jaar.

De worm plot is een aangepaste versie van de Quantile-Quantile plot, de QQ-plot. Op de *Y*-as van de worm plot staat het verschil tussen empirische en theoretische quantiel (d.w.z. detrended QQ-plot). Verder is uitgesplitst naar 16 leeftijdsgroepen (zie hoek linksboven), en is het 95% betrouwbaarheidsinterval getekend. De naam *worm plot* is ontleend aan de vorm en gedrag van de datapunten. Wanneer het model past, dan zullen de wormen zich grofweg binnen het betrouwbaarheidsinterval bewegen. Specifieke vormen van misfit zijn: de worm gaat niet door de oorsprong, de worm staat scheef op de horizontale as, de vorm geeft een kwadratisch patroon, etc. Afhankelijk van de vorm van de worm kan het op specifieke plaatsen LMS-model worden bijgesteld. Op deze wijze is iteratief een goed passend LMS-model te vinden. Modellen fitten is wormen temmen.

Conclusie

Bij het maken van groeidiagrammen komt de nodige statistiek kijken. Groeidata zijn voor de statisticus een waar paradijs. Het aantal kinderen is groot, het aantal maten is klein, de maten hangen sterk samen, iedereen begrijpt wat lengte en gewicht is, en de resultaten worden in de praktijk intensief gebruikt.

Tijdige preventie van ziekten vereist goede instrumenten. Het groeidiagram is een eerste en belangrijke stap naar een evidence-based benadering voor een tijdige opsporing van risicofactoren en aandoeningen. Onlangs is de Vijfde Landelijke Groeistudie gestart. Deze wordt uitgevoerd door TNO, VUMC en LUMC. De studie zal eind 2009 nieuwe referentiewaarden opleveren. Sinds enkele jaren werkt TNO aan richtlijnen en protocollen voor verwijzingen op basis van groei. Hierbij wordt onder meer het longitudinale karakter van het groeipatroon benut. Dit onderzoek is momenteel vol in ontwikkeling, en het is dan ook te vroeg om er hier dieper op in te gaan. Maar wellicht kan ik het lezerspubliek van *STAtOR* over een aantal jaren met een update verblijden.

Relevant materiaal is te vinden op de website <www.stefvan buuren.nl>.

Literatuur

Borghi, E., de Onis, M., Garza, C., van den Broeck, J., Frongillo, E. A., Grummer-Strawn, L., van Buuren, S., Pan, H., Molinari, L., Martorell, R., Onyango, A., Martines, J. (2006). Methods for constructing the WHO child growth references: Recommendations of a statistical advisory group. *Statistics in Medicine,* 25, 247-265.
Cole, T. J., Green, P. J. (1992). Smoothing reference centile curves: the LMS method and penalised likelihood. *Statistics in Medicine*, 11, 1305-1319.
Fredriks, A. M., van Buuren, S., Burgmeijer, R. J., Meulmeester, J. F., Beuker, R.J., Brugman, E., Roede, M. J., Verloove-Vanhorick, S. P., Wit, J. M. (2000). Continuing positive secular growth change in The Netherlands 1955-1997. *Pediatric Research*, 47, 316-323.
Van Buuren, S., Fredriks, A. M. (2001). Worm plot: A simple diagnostic device for modeling growth reference curves. *Statistics in Medicine*, 20, 1259-1277.

Stef van Buuren is statisticus bij TNO Kwaliteit van Leven te Leiden en hoogleraar Toegepaste statistiek van preventie onderzoek aan de faculteit Sociale Wetenschappen van de Universiteit Utrecht.
E-mail: <stef.vanbuuren@tno.nl>.

Wachtenden bij Ikea in Yokohama.

WACHTEN OP LEVEN EN DOOD

Onno Boxma

Wachtrijtheorie is (niet) dood

'*Queueing theory is dead*.' Dat was de openingszin van Walter Smith, tijdens zijn hoofdvoordracht bij een in 1972 door IBM georganiseerd congres over puntprocessen. Smith, overigens een meester in het maken van prikkelende opmerkingen, stelde dat de meeste artikelen over wachtrijen uit die jaren onbelangrijk waren. Het vakgebied zat inderdaad toentertijd in een dip.

Cohen's standaardwerk *The Single Server Queue* uit 1969 gaf goed weer hoe ver men wel kon gaan met het analyseren van prestatiematen voor het klassieke wachtrijmodel van één rij klanten van één type, te bedienen door één bediende.

Veel van die prestatiematen werden uitgedrukt in genererende functies en Laplace transformaties, wat in toenemende mate tot een gevoel van onbehagen leidde; afgezien van de bepaling van momenten uit getransformeerden via differentiatie stuitte men op een *Laplacian curtain*, want men wist niet goed hoe kansverdelingen numeriek uit gecompliceerde getransformeerden konden worden bepaald. En zodra het onderwerp van studie complexer werd – meerdere bedienden, meerdere typen klanten, of meerdere wachtrijen in een netwerk – waren sterk vereenvoudigende aannamen vereist, zoals exponentialiteit van alle betrokken verdelingen.

Smith had dus wel een punt, en wachtrijtheorie zou bepaald niet het eerste wetenschapsgebied zijn geweest dat een zachte dood is gestorven. In veel disciplines worden bij tijd en wijle soortgelijke meningen geventileerd. Zo menen sommigen dat de mathematische statistiek haar beste tijd gehad heeft, en de theorie

Een kat met negen levens

Vanaf 1972 heeft de wachtrijtheorie, op dat moment een VUT-ter (zie het eind van deze column) een sterke bloeiperiode doorgemaakt. Een der succesfactoren is een stevige inbedding in de toegepaste kansrekening, met rechtstreekse toegang tot de beste kansrekenaars en hun modernste gereedschap. Nog belangrijker echter lijkt de hechte band met de computer-communicatiegemeenschap, wat overigens nog geen garantie voor het eeuwige leven is. Hoe dan ook, de wachtrijtheorie lijkt zich momenteel elke tien jaar drastisch te vernieuwen, waarbij de impulsen bijna steeds komen uit de hoek van de computer-communicatienetwerken (veel minder uit produktiesystemen, en voorlopig nog niet uit bijvoorbeeld het wegverkeer of de gezondheidszorg). Midden jaren tachtig richtte de aandacht van het vakgebied zich op vloeistofmodellen, waarbij een vloeistofstroom een grote stroom datapakketjes representeerde. In het vorige decennium gaven metingen aan Internetverkeer aanleiding tot een geheel nieuwe kijk op verkeersmodellen, met veel aandacht voor zwaarstaartige verdelingen en long-range dependence. Ook werden modellen en analysemethoden ontwikkeld voor draadloze communicatienetwerken. Zulke modellen leken dikwijls nauwelijks meer op de klassieke wachtrijmodellen uit 1972.

De laatste jaren krijgt de wachtrijtheorie nieuwe impulsen uit nieuwe diensten en communicatiemogelijkheden als YouTube en BitTorrent.Metingen suggereren dat BitTorrent (letterlijk: stortvloed van bits) momenteel al verantwoordelijk is voor 30-40 procent van het Internet verkeer, en dus wordt de prestatie-analyse ervan heel belangrijk.

BitTorrent is een protocol voor de uitwisseling van data; denk aan het downloaden van films. Het protocol zorgt voor coordinatie van de downloads. De download zelf gebeurt decentraal, en bestaat uit het uitwisselen van stukken bestanden tussen alle gebruikers die op dat moment meedoen aan het up- en downloaden. Het feit dat een gebruiker tegelijk zelf aan het downloaden is en al verworven stukken onder andere gebruikers (zijn {\em peers}) distribueert geeft een enorme winst in vergelijking met de klassieke {\em single server} situatie: in theorie kunnen N kopieën van een file in orde log(N) tijd worden gedistribueerd, in plaats van in orde N. De wachtrijtheoreticus smult bij de gedachte dat klanten hier tegelijk zelf ook bedienden zijn. Momenteel worden diverse modellen voor zulke {\em peer-to-peer} netwerken ontwikkeld en bestudeerd; onder meer in het Europese Euro-FGI netwerk waartoe onze groep en een CWI-groep behoren. Een wachtrijkenner die er twintig jaar uit is geweest en deze modellen ziet, zou het vak niet terugkennen.

van Markovbeslissingsprocessen bloeit lang niet meer zo als dertig jaar geleden. Niet dat dit geen gebieden zijn die overduidelijk van enorm belang zijn, goed ingebed in onderwijsprogramma's en volop in gebruik in de praktijk. Het probleem is eerder dat er te weinig wetenschappelijke *grand challenges* liggen, of dat men geen flauw benul heeft hoe die kunnen worden aangepakt. Denk maar aan het probleem van de *state space explosion* bij Markovbeslissingsprocessen. In zo'n situatie keren veel van de beste onderzoekers zich af van hun vakgebied, om uitdagingen in verwante gebieden op te zoeken.

Het al dan niet optreden van wetenschappelijke doorbraken die een vakgebied een belangrijke nieuwe impuls geven, of die zelfs een nieuw vakgebied vestigen, is doorgaans sterk aan het toeval onderhevig. Doorbraken zijn bijna per definitie onverwacht en onvoorspelbaar. Zij die de houdbaarheidsdatum van hun vakgebied verdedigen scheppen er bijvoorbeeld genoegen in te wijzen op de bij velen aan het begin van de twintigste eeuw heersende opvatting, dat de fysica op sterven na dood was. Een flagrant verkeerde inschatting, maar wie had de doorbraken van Einstein, Planck c.s. kunnen voorzien?

Rond de tijd van Smith' uitspraak '*Queueing theory is dead*' begon ik me, als student, in de wachtrijtheorie te verdiepen – in zalige onwetendheid van zijn doodsverklaring, en nog minder van de ophanden zijnde wederopstanding.

Wat Smith bij bovengenoemd IBM congres niet kon voorzien was de enorme impuls die de prestatie-analyse van computernetwerken zou gaan geven aan de wachtrijtheorie. Het begon niet direct met wetenschappelijke doorbraken, maar met spannende nieuwe vragen, voortvloeiend uit de ongebreidelde groei van computersystemen.

Lavenberg en Reiser in datzelfde IBM, en anderen, zouden spoedig komen met netwerkmodellen voor *window flow control* en voor computersystemen met een vaste multiprogrammeringsgraad, en met oplossingsmethoden in de vorm van produktvormresultaten en de *Mean Value Analysis* methode.

Kleinrock had ondertussen wachtrijmodellen ontwikkeld voor het ARPAnet, een voorloper van het Internet; ook schreef hij het baanbrekende boek *Queueing Systems*, waarbij vooral het in 1976 verschijnende Vol. 2: *Computer Applications* veel deed om het vakgebied toegankelijk te maken voor elektrotechnici en informatici.

100 years of queueing

Volgend jaar viert de wachtrijgemeenschap dat de Deense ingenieur A.K. Erlang honderd jaar geleden het eerste artikel over wachtrijen publiceerde: The theory of probabilities and telephone conversations (*Nyt Tidsskrift for Matematik* B, Vol. 20, 1909). Tijdens het congres '100 years of queueing. The Erlang centennial' dat in Kopenhagen gaat plaatsvinden van 1-3 april 2009 zullen diverse sprekers terugblikken, en vooral vooruitkijken hoe het vak zich ontwikkelt en waar de grootste uitdagingen liggen. Tien jaar later zal ik in *STAtOR* rapporteren wat ervan is gekomen – als *STAtOR* dan nog bestaat ...

Onno Boxma is hoogleraar Stochastische Besliskunde bij de Faculteit Wiskunde en Informatica van de TU Eindhoven en wetenschappelijk directeur van EURANDOM. Hij is hoofdredacteur van Queueing Systems. Email: <boxma@win.tue.nl>.

VROUWEN IN DE MATHEMATISCHE STATISTIEK

Fred Steutel

Tijdens een redactievergadering opperde ik – in een onbewaakt ogenblik – het idee om op zoek te gaan naar een artikel over 'Vrouwen in de Statistiek'. Ik verwachtte dat een van de Nederlandse vrouwelijke statistici bereid zou zijn deze leemte op te vullen. Toen ik twee, overigens heel vriendelijke, blauwtjes had gelopen ('dat moet een man maar doen'), besloot ik zelf een poging te wagen. Geholpen door het wereldwijde web kwam ik op het spoor van allerhande wetenswaardigheden over vrouwen in de statistiek, met inbegrip van de kansrekening – het onderscheid is tanende. Al schrijvende merkte ik dat dit stuk vrij veel gaat over mensen die ik persoonlijk ken. Aan de verleiding om boven dit verhaal 'Mijn vrouwen in de statistiek' te zetten, heb ik met succes weerstand geboden. Mijn eigen achtergrond in de wiskunde maakt dat dit verhaal vooral gaat over vrouwen in de wiskundige statistiek; ik zal het voorvoegsel 'wiskundige' of 'mathematische' niet steeds herhalen. Ik heb geen oordeel over vrouwen in de statistiek (of daarbuiten). Ik doe verslag van wat ik heb gehoord, gezien of gelezen. Ik heb niet gezocht naar statistische gegevens over vrouwen in de statistiek. Ik heb geen volledigheid nagestreefd.

Ervaring

Ik kwam niet helemaal onbeslagen ten ijs, want ik had een paar jaar geleden in het *Nieuw Archief voor Wiskunde* een stukje geschreven over de tamelijk beroemde kansrekenaar(ster) Alexandra

Bellow (née Bagdasar), echtgenote van de beroemde schrijver Saul Bellow, eerder echtgenote van de kansrekenaar Ionescu Tulcea en later echtgenote van de zeer bekende wiskundige Alberto Calderón. Zij is specialist op het terrein van limietstellingen. Voor de details verwijs ik naar [2] van de literatuurlijst. Bovendien had ik in verband met een column over 'zwarte statistici' gezien dat het internet een schat aan informatie bergt over vrouwelijke wiskundigen.

Verkenningen

De zoekterm '*women in statististics*' levert maar liefst 13400 hits; meer dan je in een eindige tijd kunt raadplegen. Maar ook hier geven de eerste tientallen antwoorden de meest relevante informatie. Het is opvallend dat vrouwelijke statistici, en vrouwelijke wiskundigen in het algemeen, vaak door hun vaders zijn aangemoedigd om te gaan studeren. Dat gold al voor de 'Egyptische' wiskundige Hypathia (370-415 AD), die in Alexandrië op gruwelijke wijze door 'christelijk' gepeupel werd vermoord, en voor een van de bekendere zwarte vrouwelijke wiskundigen, Kate Okikiolu, die in 1991 promoveerde; haar wiskundige stamboom gaat terug tot de grote Leibniz, die promoveerde in 1666. Zij volgde het voorbeeld van haar Nigeriaanse vader, een van de productiefste 'zwarte wiskundigen'. Het gold ook voor één van eerste vrouwelijke statistici, Florence Nightingale. Het leek haar vader goed '*to let her study mathematics instead of doing worsted work* [een soort borduurwerk] *and practicing quadrilles*'.

Dicht bij huis

Al in mijn tijd (1956-1964) op het Mathematisch Centrum (CWI) waren daar vrouwen die aan statistiek deden: Constance (Stan) van Eeden, over wie later meer, Doralien Wabeke, die met Hemelrijk een boekje publiceerde met 'Elementaire statische opgaven met uitgewerkte oplossingen', Emmy Bos-Levenbach, die met Ton Benard schreef over '*The plotting of observations on probability paper*', en Rina Korswagen, die na het kandidaatsexamen de studie staakte. Van deze vier is alleen Stan van Eeden in de statistiek bezig gebeleven, en met veel succes. Zij promoveerde bij Van Dantzig en vertrok niet lang daarna naar de Michigan. Zij publiceerde met haar echtgenoot C.H. Kraft het boek *A nonparametric introduction to statistics*. Zij was lange tijd hoogleraar in Montréal, Canada en is al jaren *professeure associée* in Montréal en *University Professor* in Vancouver. Haar belangstelling lag (en ligt) op de terreinen van parametervrije methoden en selectieprocedures. Zij zal in de rest van dit verhaal regelmatig opduiken. Aan Nederlandse universiteiten zijn twee vrouwelijke hoogleraren Statistiek: Jacqueline Meulman in Leiden, die heeft gezongen tijdens haar oratie, en Mathisca de Gunst aan de VU, specialist in biostatistiek. Een derde Nederlandse, Sara van de Geer, is hoogleraar statistiek in Zürich. Aan de TU-Eindhoven waren een tijdlang twee Belgische vrouwen in de statistiek: Irene Gijbels en Gerda Claeskens, die nu beiden hoogleraar zijn – in België. Er zijn hier nu weer twee vrouwelijke statistici: Sonja Kuhnt en Francesca Nardi; de laatste op een *tenure track position*.

ISI

Het International Institute of Statistics (ISI), sinds mensenheugenis gevestigd in de burelen van het CBS, heeft een belangrijke rol gespeeld in de verbreiding van de statistiek. Het is opgezet

als overkoepelend orgaan van nationale bureaus voor de statistiek, en werd later het ook een thuis voor mathematisch statistici en kansrekenaars. Een van de directeuren was professor Denise A. Lievesley (van 1989 tot 1991), daarna was zij directeur van het U.K. Data Archive at the University of Essex. Zij was president van de Royal Statistical Society en vice-president van het ISI. Zij is nu werkzaam bij Unesco en tevens, sinds kort, president van ISI. Een andere vrouwelijke statisticus die jarenlang aan het ISI verbonden was, is Stan van Eeden, tot voor kort hoofdredacteur van *Statistical Theory & Method Abstracts*. Zij was daar niet de enige vrouw in de statistiek. Toen ik als hulpje van Stan regelmatig naar Voorburg ging, trof ik daar de *editor's assistants* Ann Daniels en Tineke de Boer en de ISI-secretaresse Ank Lepping. Ik bewaar goede herinneringen aan de samenwerking en aan het gemeenschappelijk lunchen en koffiedrinken.

Promovendae

Eén van de methoden om inzicht te krijgen in aantallen vrouwelijke statistici (weer inclusief kansrekenaars) is het tellen van promovendi. Omdat veel promovendi 'afstammen' van Van Dantzig, is de *Scientific Family Tree* door Constance van Eeden (referentie [1]) een goede ingang. Helaas worden daar geen voornamen gegeven en is er – terecht misschien – geen geslachtsaanduiding. Dit betekent dat ik alleen de vrouwen die ik ken, als zodanig kan identificeren. Een tweede nadeel, waarover ik al eerder (ook tegenover Stan) mijn bedroefdheid uitsprak, is het uitsluiten van Runnenburg en diens 'nageslacht'; hij was wel leerling van Van Dantzig, maar promoveerde, kort na diens dood, formeel bij N.G. de Bruijn. Een andere ingang is de *Mathematics Genealogy Project* op het wereldwijde web; daar worden wel vaak voornamen gegeven, maar de data zijn nogal onvolledig. Het percentage vrouwen onder de promovendi is laag, hooguit 20 %. Na Stan van Eeden met vijf of zes vrouwen (allemaal in de VS of Canada) is van Zwet het meest 'vrouwvriendelijk': Sara van de Geer, Mathisca de Gunst en Marta Fiocco. Piet Groeneboom had twee vrouwelijke promovendi: Anoesjka Cabo (ook bekend als violiste) en Marloes Maathuis. Ook Richard Gill had twee promovendae: Catharina Oudshoorn en Jaqueline Lok. Door bovengenoemde oorzaken is het moeilijk om gegevens over vrouwelijke kansrekenaars te vinden. Marie Colette van Lieshout is ooit bij mij op sollicitatiebezoek geweest, maar gaf er – met reden - de voorkeur aan bij Adrian Baddeley te promoveren. Ronald Meester had twee promovendae: Lorna Booth en Cornelia Quant. Hun – en zijn – doctor-voorvaderen gaan via David Hilbert, Felix Klein terug tot Otto Mencke, die in 1666 promoveerde. Misschien dat De Haan ook een aantal vrouwen onder zijn promovendi had, maar in zijn cv worden geen voornamen van promovendi gegeven. Bij Chris Klaassen is de Surinaamse Shanti Venetiaan gepromoveerd.

Twee Florence Nightingales

De eerste vrouwelijke statisticus is volgens velen Florence Nightingale (1820 – 1910). Zij leerde wiskunde van de beroemde Sylvester en (misschien) statistiek van de beroemde Belg Quetelet. Een van haar uitspraken luidt: '*To understand God's thoughts we must study statistics, for these are the measure of His purpose*' – in ieder geval mooi gezegd. Zij leverde bijdragen op het terrein van het verzamelen, tabelleren en grafisch weergeven van gegevens, met name in verband met aantallen gewonden en gesneuvelden in de Krim-

oorlog. Kwade tongen beweren dat de onhygiënische werkmethoden uit die tijd de gewonden meer kwaad dan goed gedaan hebben.

F.N. David (1909 – 1993) is een van de bekendste vrouwelijke kansrekenaars, bekend van het boek Games, Gods and Gambling. Eén van haar artikelen, met D.E. Barton, draagt de provocerende titel *Four-letter words*; de gangbare interpretatie van die uitdrukking is de verzameling woorden die in het Nederlands met drie letters worden geschreven, maar dit is een artikel over het tellen van – abstracte – woorden in een – abstract – alfabet. Pas sinds vrij kort weet ik dat de voorletters F.N. staan voor Florence Nightingale.

Op de website van de American Statistical Association (ASA) vind ik onder het kopje *Women have made valuable contributions* de volgende statistici, actief in de tweede helft van de vorige eeuw. Gertrude Cox, een van de oprichters van de Interrnational Biometric Society en een van de presidenten van de ASA. Janet Norwood, ook president geweest van ASA, expert op het gebied van de officiele statistiek. Tiffany T. Sundelin, een *quality control engineer,* werkt(e) in de vliegtuigindustrie (ik kon verder weinig over haar vinden). Tischa Agnessi, specialist op gebied voorspellingmodellen in verband met *fraudulent customer behavior*. Amy Alice Derrow, expert op het gebied van computerprogramma's voor patiëntenstatistiek. Van bijna al deze vrouwen had ik nog nooit gehoord.

Prijzen

Voor de mensen die wel eens Berkeley bezochten, was Elizabeth L. Scott een goede (of mogelijk een kwade) bekende – ik heb haar een keer, geheel neutraal, ontmoet tijdens een van de laatste Berkeley Symposia. Zij leidde als rechterhand van Jerzy Neyman de afdeling statistiek. In een in memoriam noemt David haar *formidable*. Zij was zelf een statisticus van formaat, met speciale belangstelling voor toepassingen in de astronomie, waarin ze was afgestudeerd. Naar haar is de Elizabeth L. Scott Award genoemd. De eerste ontvanger van deze prijs was Florence Nightingale David. Deze prijs was niet alleen bedoeld voor vrouwen; Ingram Olkin ontving hem in 1998. Naar David is ook een prijs genoemd: Florence Nigthingale David Award, die wel alleen voor vrouwen bedoeld is. De eerste winnaar is Megan Kruse. Er is ook een Nederlander naar wie een prijs is genoemd: Henri Willem Methorst, algemeen secretaris van het ISI en later, van 1911 tot 1947, directeur van het *permanent office*. Vrouwelijke ontvangers van de Methorst Medaille zijn ondermeer de Finse Asta Manninen en ... Stan van Eeden.

Tot slot

De vraag of het verstandig is om vrouwelijke statistici in het voetlicht te plaatsen zal ik niet beantwoorden. Het was aardig, en tot op zekere hoogte leerzaam, om er 'onderzoek' naar te doen en in mijn eigen geheugen te graven. Al met al is het wellicht beter om ook in de statistiek de vrouwen hun eigen plaats te laten bepalen.

PS Bij het tot stand komen van deze STAtOR zijn minstens vier vrouwen betrokken.

Literatuur
1. Van Eeden, Constance. (2000). *The Scientific Family Tree of David van Dantzig*. Amsterdam: CWI.
2. Steutel, Fred W. (2005). Uit de schaduw vandaan. *Nieuw Archief voor Wiskunde* 5/6, 3, 225.

Fred Steutel *is emeritus hoogleraar kansrekening aan de TU Eindhoven. Hij is redacteur van STAtOR.*
E-mail: <f.w.steutel@tue.nl>

AGENDA

27 maart 2008

De Dag voor Statistiek en Besliskunde 2008 zal wederom op het CBS in Voorburg plaatsvinden, en wel op donderdag 27 maart. Sprekers en overige informatie staan binnenkort op de site <www.vvs-or.nl>. Noteer deze dag alvast in uw agenda.

26 - 28 juni 2008

We would like to invite you to participate in the **Workshop on Nonparametric Inference – WNI2008**, which will be held in Coimbra, Portugal, on June 26-28, 2008. For more information please see the website <www.mat.uc.pt/~wni2008>.

7 - 11 juli 2008

De 23e International Workshop on Statistical Modelling vindt van 7 t/m 11 juli 2008 plaats in Utrecht. Deadline voor het insturen van abstracts: 10 februari 2008. Zie: <www.fss.uu.nl/iwsm2008>.

17 - 21 augustus 2008

On behalf of ISCB and the LOC it is a great pleasure to invite you to attend the 29th annual conference in Copenhagen, 17-21 August 2008. Please visit <www.iscb2008.info> for further details.

27 - 28 augustus 2008

Oprichters van EURANDOM, voormalige postdocs en vooraanstaande sprekers vanuit industrie en wetenschap blikken terug op 10 jaar EURANDOM en kijken vooruit naar nieuwe ontwikkelingen. Voor inlichtingen, zie <www.eurandom.nl>.