

Een sprong naar de maan



Prof. dr. S. van Buuren

Een sprong naar de maan

Rede uitgesproken bij de aanvaarding
van het ambt van bijzonder hoogleraar
in de Toegepaste statistiek van preventie onderzoek
aan de Universiteit Utrecht
vanwege de Stichting Lorentz-van Iterson Fonds TNO
op 10 februari 2006 door

Prof. dr. S. van Buuren

*Mijnbeer de Rector Magnificus,
Zeer geachte toeboorders,*

Het zal u niet ontgaan zijn dat vanavond de Olympische Spelen worden geopend. Het is dan ook een mooi moment om terug te kijken op een gebeurtenis bij een eerdere Olympiade. Op 18 oktober 1968 maakte Bob Beamon in Mexico een sprong van 8 meter en 90 centimeter. Dat betekende destijds een spectaculaire verbetering van het wereldrecord verspringen met ruim 55 centimeter. Ik herinner me nog dat ik de afstand heb afgemeten met de huishoudcentimeter uit de naaidoos van mijn moeder op het schoolplein in de buurt. De standaard huishoudcentimeter heeft een lengte van 1 meter en 50 centimeter. De sprong van Beamon is bijna zes maal de volledige lengte van de centimeter. Legt u maar eens zes huishoudcentimeters achter elkaar. Ik heb er met grote bewondering naar staan kijken. Ook Beamon zelf kon achteraf zijn prestatie nauwelijks geloven.

Een jaar later, om precies te zijn op 20 juli 1969, stond Neil Armstrong op de onderste trede van de maanlander de Eagle, en sprak de woorden: “Een kleine stap voor een mens, een reusachtige sprong voor de mensheid”. Vervolgens nam hij het laatste stapje, en was daarmee de eerste mens op de maan. Op *zich* stelde dat stapje niet veel voor. In het Smithsonian Aviation and Space Museum in Washington kunt u een heuse maanlander aanschouwen. De afstand tussen de onderste trede van de Eagle en het maanoppervlakte bedraagt 75 centimeter, oftewel ongeveer de helft van de lengte van de rolcentimeter. Ik denk dat ieder van de aanwezigen in deze zaal zo’n stapje had kunnen zetten. De betekenis van het stapje van Armstrong is echter veel groter dan Beamon’s reuzensprong.

Wetenschappelijk vooruitgang gebeurt in stapjes, stappen en sprongen. In deze rede zal ik een lans breken voor de kleine stapjes. De beste manier om vooruit te komen is het zetten van meerdere, kleine stapjes. De stap van Armstrong is er slechts eentje in een lange keten. Dit stapje staat symbool voor de gehele keten. In de wetenschap staan kleine stapjes gelijk aan het achtereenvolgens oplossen van meerdere, schijnbaar eenvoudige problemen. Vele kleine stapjes maken één grote. Net als Bob Beamon kijken we dan achteraf misschien in verbazing achterom naar de afgelegde afstand.

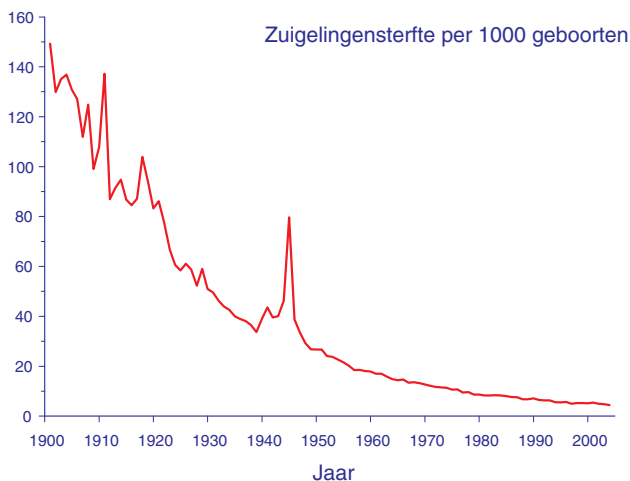
Per 1 januari 2005 financiert de Stichting LIFT TNO de bijzondere leerstoel Toegepaste statistiek van preventie onderzoek aan de discipline-groep Methoden & Technieken van de faculteit Sociale Wetenschappen van de Universiteit Utrecht. Het doel van deze leerstoel het uitbouwen, bestendigen en verankeren van de methodologische en statistische expertise van TNO Kwaliteit van Leven. Dit resulteert in een versterking van de kennis- en marktpositie van TNO, en in een verbreding van de mogelijkheden voor verwerving van onderzoeksgelden binnen en buiten Nederland. Tijdens deze bijeenkomst vertel ik u wat mijn leerstoel inhoudt, en welke plannen ik heb.

Ik zal trachten u een indruk te geven van wat preventie onderzoek inhoudt, en wat daarin de rol van statistiek is. Daarna leg ik u uit dat er twee routes naar vernieuwing zijn. Deze routes kunnen het beste met kleine stapjes afgelegd worden. Mijn conclusie is dat er een nieuwe benadering in de statistiek nodig is, modulaire statistiek.

Toegepaste statistiek van preventie onderzoek

Er bestaan vele soorten preventie, bijvoorbeeld: brandpreventie, misdaadpreventie, zwangerschapspreventie, inbraakpreventie en ziektepreventie. In mijn leerstoel gaat het om preventie van ziekten op zo vroeg mogelijke leeftijd.

Sinds het begin van de vorige eeuw heeft de preventieve geneeskunde grote successen geboekt. In de afgelopen 100 jaar is de kindersterfte in Nederland spectaculair gedaald. In het jaar 1900 overleden van elke 1000 kinderen er 150 voor de eerste verjaardag. Tegenwoordig zijn dat er 4 à 5 per 1000. Een deel van de daling kan op het conto geschreven worden van een sterk verbeterde hygiëne. Via open riolen kunnen ziekten zich gemakkelijk verspreiden. Nederland kende in het verleden ook open riolen. Tegenwoordig niet meer, en dat heeft bijgedragen tot een sterke afname van de kindersterfte.



Figuur 1 Afname in zuigelingensterfte in Nederlands sinds 1900 (bron: CBS).

In Nederland bestaat een uniek systeem van jeugdgezondheidszorg. U moet dan denken aan consultatiebureaus, periodiek geneeskundig onderzoeken op scholen, of, vroeger, het Groene Kruis. Het opsporen en voorkomen van ziekten tijdens de kindertijd draagt sterk bij aan de daling van de kindersterfte en chronische ziekten.

Een bekend voorbeeld is de hielprik. De hielprik bestaat uit het afnemen van een kleine hoeveelheid bloed bij pasgeborenen, uitgevoerd op de vierde dag na geboorte. Met de hielprik worden per jaar ongeveer 80 zuigelingen opgespoord met aangeboren stofwisselingsziekten zoals PKU of CHT. Indien dat niet zou gebeuren, dan lopen deze kinderen een blijvende achterstand op in groei en ontwikkeling, vaak leidend tot zwakzinnigheid. Bij tijdige behandeling ontwikkelen deze kinderen zich normaal. Preventie van deze ziekten voorkomt dus veel leed en kosten.

Preventie is bijzonder omdat het wordt aangeboden aan zich gezond voelende mensen die er meestal niet zelf om vragen. De meeste deelnemers doen 'voor niets' mee, d.w.z., zonder de preventie hadden ze de ziekte *ook* niet gekregen. De effecten van preventie bestaan uit het

behoud van gezondheid, en het *niet* optreden van ziekte. Effecten zijn pas zichtbaar na verloop van tijd, soms na vele jaren. (Schaapveld & Hirasing, 1997).

Goede preventie komt niet zomaar uit de lucht vallen. Vaak is jarenlang onderzoek nodig om antwoorden te vinden op belangrijke vragen zoals:

- Hoe groot is het gezondheidsprobleem?
- Wat kunnen we eraan doen?
- Wat zijn schadelijke gevolgen van preventie?
- Wat kost het?
- Hoe is het organisatorisch in te bedden?

De beantwoording van dit soort vragen vereist een goede methodologie. Met mijn leerstoel Toegepaste statistiek van preventie onderzoek wil ik een bijdrage leveren aan de vernieuwing en de verbetering van deze methoden en technieken. Met nieuwe methoden is beter onderzoek naar preventie, en daarmee betere preventie mogelijk.

Twee wegen naar innovatie

Vernieuwen is dus belangrijk. De missie van TNO luidt niet voor niets: “TNO maakt wetenschappelijke kennis toepasbaar om het innovatief vermogen van bedrijfsleven en overheid te versterken.” In de missie van de Universiteit Utrecht staat “Het doen van grensverleggend onderzoek”. Het kernbegrippen zijn hier ‘innovatie’ en ‘grensverleggend’. Ik heb me dan ook de vraag gesteld:

Op welke wijze draagt mijn leerstoel bij aan innovatie en aan grensverleggend onderzoek?

Om deze vraag te beantwoorden wil ik u graag uitleggen hoe vernieuwingen in de methodenleer tot stand komen. Ik zie twee wegen naar innovatie: de weg van de toepassing en de weg van de instrumentontwikkeling. Beide wegen zal ik uitleggen.

In mijn werk maak ik onderscheid tussen drie soorten activiteiten: *advise-ring*, *toepassing* en *instrumentontwikkeling*. Elke activiteit kan in principe los van de andere twee worden uitgevoerd, maar de kracht zit juist in de combinatie. Instrumentontwikkeling zonder toepassing is steriel, en leidt vaak niet tot een bruikbaar resultaat. Advisering zonder instrumentontwikkeling blijft beperkt tot het opdiepen van standaard recepten voor standaard problemen. Echte vernieuwing vereist zowel toepassing als instrumentontwikkeling. Hoe dit precies zit laat de volgende tabel zien.

	Bestaande taak	Nieuwe taak
Bestaande methode	Advisering	Toepassing
Nieuwe methode	Instrumentontwikkeling	Innovatie

De tabel beschrijft de relatie tussen advisering, toepassing en instrumentontwikkeling. In de bovenste twee cellen staan activiteiten waarin sprake is van een bestaande methode. In de onderste twee cellen is sprake van een nieuwe methode. In de twee linkercellen gaat het om bestaande taken. In de twee rechtercellen staan nieuwe taken.

De eerste situatie, de cel linksboven, bevat de combinatie *bestaande methode* en *bestaande taak*. De activiteiten in deze cel noem ik *advisering*. Het doel is het optimaal inzetten van een bekende methode op een bekende wijze. Zo toetsen we bijvoorbeeld het verschil tussen twee gemiddelden met een *t*-toets. Voor het voorspellen van uitkomsten gebruiken we regressie analyse. Standaard oplossingen voor standaard problemen.

De tweede situatie, de cel rechtsboven, bevat de combinatie *bestaande methode* en *nieuwe taak*. Hier gebruiken we een bekende methode op een nieuwe wijze. Op TNO doen we dat vaak. Voorbeelden zijn: het gebruik van psychometrische modellen voor het computerondersteund meten van fysieke activiteit, het gebruik van microsimulatie voor het schatten van de kosten-effectiviteit van bloeddrukscreening, en het gebruik van ROC-curves voor het doorrekenen van optimale verwijsregels voor kleine lengte. De methoden zelf zijn niet nieuw, maar binnen het toepassinggebied zijn ze dat wel. Ze vormen zo een toevoeging op de inhoudelijke literatuur.

De derde mogelijkheid is instrumentontwikkeling, de combinatie *nieuwe methode* voor een *bestaande taak*. Hieronder valt het verbeteren, uitbreiden en vernieuwen van een bestaande methode. Bijvoorbeeld, door koperleidingen te vervangen door glasvezelkabel kunnen we *sneller* gegevens verzenden. Of, door koperleidingen te vervangen door een GSM netwerk kunnen we op *meerdere plaatsen* gegevens verzenden.

Instrumentontwikkeling binnen het preventie onderzoek bestaat bijvoorbeeld uit verbeteren van instrumenten voor het meten van kwaliteit van leven. Ook valt hieronder: het verbeteren van statistische methoden voor het omgaan met ontbrekende gegevens, en het verbeteren van methoden voor het fitten van groeicurven. Iets wat al kon, kan nu beter.

Interessanter is indien het instrument óók nog iets kan dat niet eerder kon. We krijgen dan innovatie, een *nieuwe methode* voor een *nieuwe taak*. In de 17^e eeuw bouwde Antonie van Leeuwenhoek microscopen die tot 200 maal vergrootten, een voor die tijd ongekend hoge scherpte. Zijn beroemde tijdgenoot Robert Hooke kwam niet verder dan 20 tot 30 maal. Van Leeuwenhoek was een zeer kundig lenzenmaker, en gebruikte slechts één lens. Zijn instrumenten waren eenvoudiger dan die van Hooke, die twee lenzen gebruikten en die meer op hedendaagse microscopen lijken. De technische prestatie van Van Leeuwenhoek was op zich al bewonderenswaardig, maar de werkelijke vernieuwing zat in wat hij ermee deed. Van Leeuwenhoek zag als eerste algen, bacteriën, bloedcellen, en bracht zo een nieuwe wereld van microscopisch leven onder de aandacht van de wetenschap.

Ook in de statistiek kennen we innovatie. Omstreeks 1990 voltrok zich een revolutie door de ontwikkeling van nieuwe Bayesiaanse rekenmethoden. Deze methoden lossen bestaande problemen op een nieuwe wijze op, en belangrijker nog, bieden ook een aanpak voor allerlei onopgeloste problemen. Inmiddels zijn deze methoden vrij gangbaar, en heeft de benadering al tot veel innovatieve toepassingen geleid. Vernieuwing is dus iets nieuws doen op een nieuwe manier.

Er zijn twee wegen naar innovatie, via toepassing, en via instrumentontwikkeling.

De eerste weg naar innovatie loopt via toepassingen. Vaak moet een bestaande methode worden aangepast voor een optimaal resultaat. In feite doen we dan aan een vorm van instrumentontwikkeling. Soms zijn er zoveel aanpassingen nodig dat we kunnen spreken van een nieuw instrument met een nieuwe toepassing. Karakteristiek is dat de gebruikte methodologie vaak een combinatie is van meerdere, relatief eenvoudige technieken. Ook de reis naar de maan is een combinatie van vele relatief eenvoudige stapjes. Het mooie van deze route is dat ze door de praktijk gestuurd wordt. Het gevaar is dat de benodigde aanpassingen tot foute conclusies kunnen leiden. Of, in het geval van de maan, een fatale landing. Een goede wetenschappelijke inbedding is daarom vereist.

De tweede weg naar innovatie loopt via instrumentontwikkeling. Soms kan het nieuwe instrument iets dat niet met een bestaand instrument kan. Het mooie van deze route zijn de nieuwe mogelijkheden die het instrument kan bieden, leidend tot nieuwe inzichten zoals Van Leeuwenhoek's microscoop. Het gevaar van deze weg is dat instrumentontwikkeling een doel op zich kan worden, zonder minimaal één duidelijke toepassing. Opgepast moet worden voor het Zwitserse zakmes, een instrument dat van alles en nog wat kan, maar niet één ding echt goed. Instrumenten die pogen alles in één sprong op te lossen zijn vaak ingewikkeld.

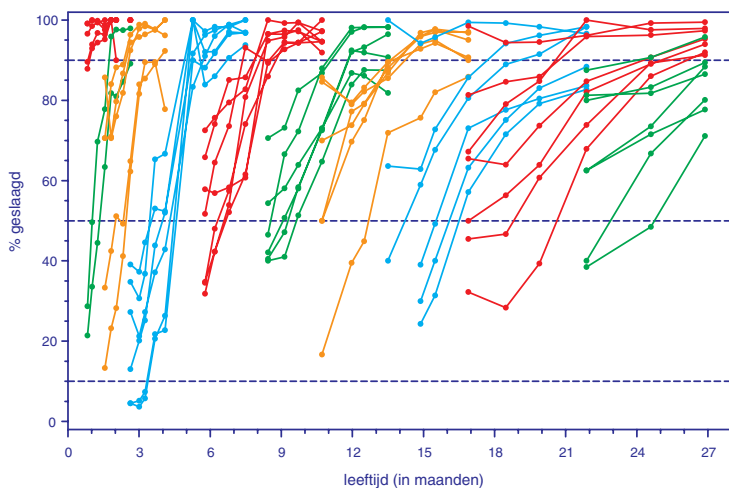
Bij TNO verloopt innovatie op het gebied van de statistiek vooral via de weg van de toepassing. Op de universiteit ligt de nadruk op de weg van de instrumentontwikkeling. Beide zijn waardevol, en versterken elkaar. Ik zie voor mijzelf een taak weggelegd om de sterke punten van beide benaderingen te combineren.

In de rest van dit verhaal ga ik in op twee voorbeelden uit mijn praktijk. In het eerste voorbeeld loopt de weg naar innovatie via de toepassing. In het tweede via instrumentontwikkeling.

De weg van de toepassing

Een statistische toepassing is het gebruik van *bestaand* statistisch gereedschap voor een *nieuwe* taak. Ik zal nu een voorbeeld geven van het gebruik van een bestaande methode, het Rasch model, op een nieuw gebied, het meten van ontwikkeling bij kinderen.

De Zwolse kinderarts Van Wiechen ontwikkelde in de 60-er jaren een methode voor het volgen van ontwikkeling van kinderen. Deze methode staat bekend als het Van Wiechen schema, en heeft breed ingang gevonden in het consultatiebureau. Het Van Wiechen schema bevat 59 indicatoren, vaak eenvoudige vaardigheidstestjes. Bijvoorbeeld, op de leeftijd van 2 maanden kijkt de verpleegkundige of de baby een voorwerp met ogen en hoofd volgt. Op 9 maanden moet het kind een blokje van de ene in de andere hand kunnen overgeven. Het Van Wiechen schema heeft zichzelf ruimschoots bewezen in de praktijk. Toch kan het iets niet dat in de praktijk wel handig zou zijn. Het meet geen ontwikkeling. Ontwikkeling is een toename van vaardigheden over de tijd. Het vaststellen van ontwikkeling vereist een score op elk tijdstip, en daarin voorziet het schema niet.



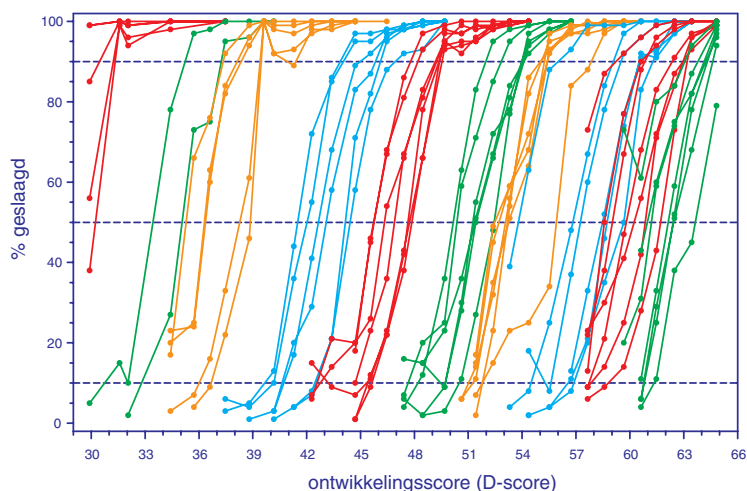
Figuur 2 Slaagkans naar leeftijd voor 59 indicatoren in het Van Wiechen schema.

Onlangs heb ik samen met twee collega's, Gert Jacobusse en Paul Verkerk, een oplossing hiervoor gevonden. (Jacobusse *et al*, 2006). Figuur 2 geeft per indicator het percentage geslaagden weer bij verschillende leeftijden. De data komen uit het SMOCK onderzoek. (Herngreen *et al*, 1992). De kleuren geven de verschillende bezoeken aan. Oudere kinderen kunnen meer, en daarom lopen alle curven omhoog. Merk op dat de curven aan het begin veel sneller stijgen dan aan het eind. Ontwikkeling op zeer jonge leeftijd gaat dus sneller.

Op deze gegevens hebben we een Rasch model gefit. Het Rasch model is een model uit de psychometrie, en wordt veel gebruikt voor het maken van onderwijstoetsen, zoals de CITO toets. Het model veronderstelt het bestaan van een onderliggende continue schaal. In ons geval kunnen we deze onderliggende schaal interpreteren als een continue ontwikkelingsscore voor kinderen tussen 0 en 2 jaar. Het model beschrijft hoe de slaagkans per indicator toeneemt met de ontwikkelingsscore.

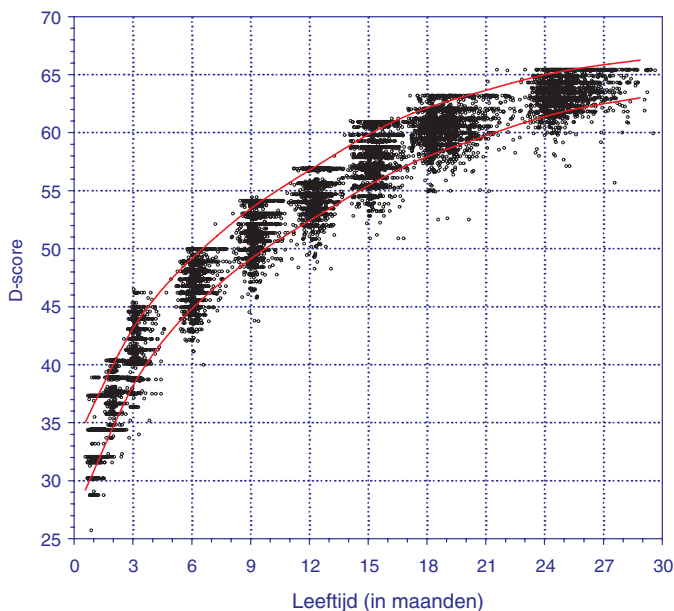
Indien het Rasch model past, dan kunnen we op eenvoudige wijze de z.g. D-score uitrekenen aan de hand van het aantal goedgeemaakte indicatoren. Het verschil tussen twee D-scores op verschillende tijdstippen meet ontwikkeling, net zoals bij groei van lengte of gewicht.

Een belangrijke veronderstelling van het Rasch model is dat alle curven evenwijdig aan elkaar lopen. In de figuur is dat duidelijk niet het geval. Op jongere leeftijd loopt de slaagkans veel sneller op dan op latere leeftijd. Om dit aspect te kunnen beschrijven zouden we eigenlijk een ingewikkelder model met een tweede parameter nodig hebben, namelijk een parameter die de steilheid van elke curve aangeeft. Tot onze verassing bleek het simpele Rasch model juist heel goed te passen.



Figuur 3 Slaagkans van 59 indicatoren in het Van Wiechen schema naar ontwikkelscore (D-score).

Figuur 3 geeft het resultaat van het model. Op de horizontale as staat de D-score. In de vorige figuur stond hier de leeftijd. Op de verticale as staat, zoals voorheen, de slaagkans. De curven binnen de figuur geven, per indicator, het percentage geslaagden aan bij een bepaalde ontwikkelscore. We zien dat alle curven, zoals verwacht, stijgen met de D-score. In tegenstelling tot de eerdere figuur lopen de curven min of meer evenwijdig aan elkaar, zoals dat volgens het Rasch model hoort.



Figuur 4 Verband van de D-score met leeftijd, aangevuld met percentiellijnen P10 en P90.

Waarom past dit model zo goed? Figuur 4 bevat de verklaring. We zien het verband tussen D-score en leeftijd. Elk stipje correspondeert met een bezoek. Zoals verwacht stijgt de D-score van de kinderen met de leeftijd, maar bij de jongste kinderen is de stijging veel sterker. Leeftijd en D-score hangen volgens een kromlijinig verband met elkaar samen. Dit lijkt veel op bijvoorbeeld het verband tussen lengte en leeftijd. We hebben dat nergens opgegeven, en het komt vanzelf uit de schaalanalyse. Het Rasch model heeft gezorgd voor een ontwikkelschaal waarin elk stapje even moeilijk is. Omdat de ontwikkeling van zeer jonge kinderen sneller is, stijgt de curve sneller voor de jongste kinderen.

Hoe draagt deze toepassing bij aan innovatie?

We hebben een kleine, maar belangrijke, stap gezet. We kunnen ontwikkeling van kinderen tussen 0 en 2 jaar kwantificeren op een interval-schaal. Dit opent een heel scala aan nieuwe mogelijkheden:

- we kunnen verschijscores berekenen en vergelijkingen over de tijd maken;
- we kunnen referentiediagrammen van ontwikkeling maken;
- we kunnen de ontwikkeling van een kind vergelijken met die van leeftijdsgenoten;
- we kunnen expliciete verwijscriteria maken;
- we kunnen met de computer het aantal benodigde vragen minimaliseren;
- we kunnen de D-score als voorspellende of als uitkomstmaat in analyses opnemen.

Dit zijn belangrijke inhoudelijke vorderingen.

Is er ook sprake van een nieuwe methode? In feite is dit een recht-toe-recht-aan toepassing van het Rasch model op een nieuw terrein. Maar er is methodologisch gezien ook iets interessants aan de hand. Indien we het algemenere model met twee parameters per indicator hadden gefit, dan hadden we nooit de goede oplossing gevonden. Bij toeval stuitte ik op een lang vergeten artikel van Frederic Lord in Psychometrika uit 1975 waar dat het geval is. (Lord, 1975) Lord fit eerst een 2-parameter-model, en constateert dat er een interpretatieprobleem is omdat beide parameters samenhangen. Als oplossing past Lord achteraf een schaaltransformatie toe. Onze analyse suggereert dat Lord's methodeodeloos complex is. Het 1-parameter Rasch model levert namelijk automatisch de door Lord gewenste transformatie op. Lord maakt een sprong vooruit ter grootte van 2 parameters, en moet vervolgens een stap terugzetten. Het Rasch model zet een stap ter grootte van 1 parameter, en komt meteen op de juiste plek uit. Het inzicht dat het Rasch model de transformatie van Lord overbodig maakt is nieuw.

Via de weg van de toepassing is niet alleen vooruitgang te boeken op inhoudelijk gebied. De weg van de toepassing kan ook leiden tot nieuwe methodologische inzichten en tot nieuwe instrumenten. Binnen het

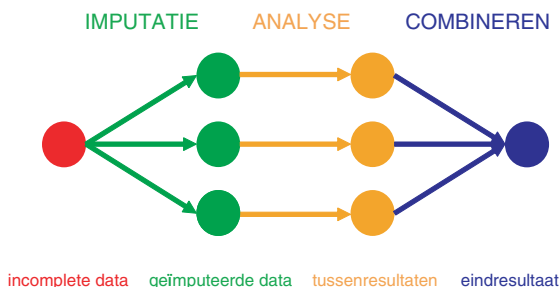
projectenwerk van TNO zijn de mogelijkheden beperkt om dit soort methodologische ideeën uit te werken. Mijn leerstoel biedt daartoe wel een mogelijkheid.

De weg van de instrumentontwikkeling

De tweede weg naar innovatie loopt via instrumentontwikkeling. Instrumentontwikkeling is het verbeteren van een methode om een bestaande taak beter te kunnen uitvoeren. Ik ga nu dieper in op een instrument voor de analyse van incomplete gegevensbestanden.

Ontbrekende gegevens vormen een belangrijk praktisch probleem. Er zijn veel oorzaken die kunnen leiden tot ontbrekende data. De respondent slaat de vraag over, stuurt het formulier niet terug, is misschien te ziek om te antwoorden, of verhuist uit het onderzoeksgebied. Ontbrekende data zijn vervelend. Met ontbrekende waarden kunnen we niet rekenen. Zelfs een eenvoudige maat als het gemiddelde is niet meer gedefinieerd als er ontbrekende waarden zijn. Vaak zoekt de onderzoeker, of de producent van de statistische software, een ad-hoc oplossing. Bijvoorbeeld, het gemiddelde wordt berekend over alleen de respondenten waarvoor wél een waarde bekend is. Het is dan natuurlijk de vraag of die respondenten corresponderen met de groep waarover de onderzoeker uitspraken wil doen. Er zijn boeken volgeschreven over de mogelijke valkuilen van dit soort ad-hoc oplossingen. Dergelijke methoden leiden vaak tot onderbenutting van kostbaar verzamelde data, en kunnen tot foute conclusies leiden. (Little & Rubin, 2002)

Gelukkig zijn er ook manieren om het beter te doen. Een goede en algemeen bruikbare aanpak is multiële imputatie. Multiële imputatie is ontwikkeld door Donald Rubin van Harvard University, één van meest creatieve hedendaagse statistici. Rubin introduceert een onderscheid tussen het missing data probleem en het analyse probleem. Het missing data probleem is het probleem dat wordt veroorzaakt doordat de informatie onvolledig is. Het analyseprobleem is het wetenschappelijk interessante probleem waarvoor de gegevens in eerste instantie verzameld zijn. In multiële imputatie worden beide problemen apart, en tot op zekere hoogte onafhankelijk van elkaar, opgelost.



Figuur 5 Schematische weergave van de werking van multiële imputatie.

Figuur 5 is een schematische voorstelling van de werking van imputatie. We beginnen aan de linkerkant met een incomplete gegevensset. In deze gegevensset zijn, om welke reden dan ook, gegevens onbekend. Imputatie betekent dat de ontbrekende waarden vervangen worden door surrogaatwaarden, oftewel imputaties. Vervolgens kunnen we de geïmputeerde data analyseren alsof zij compleet was. Dit heeft echter een belangrijk nadeel. De statistische analyse maakt geen onderscheid tussen de echte en de verzonnen gegevens. Het gevolg is dat de onzekerheidsmarges uit de analyse systematisch te klein zijn, en dus incorrect.

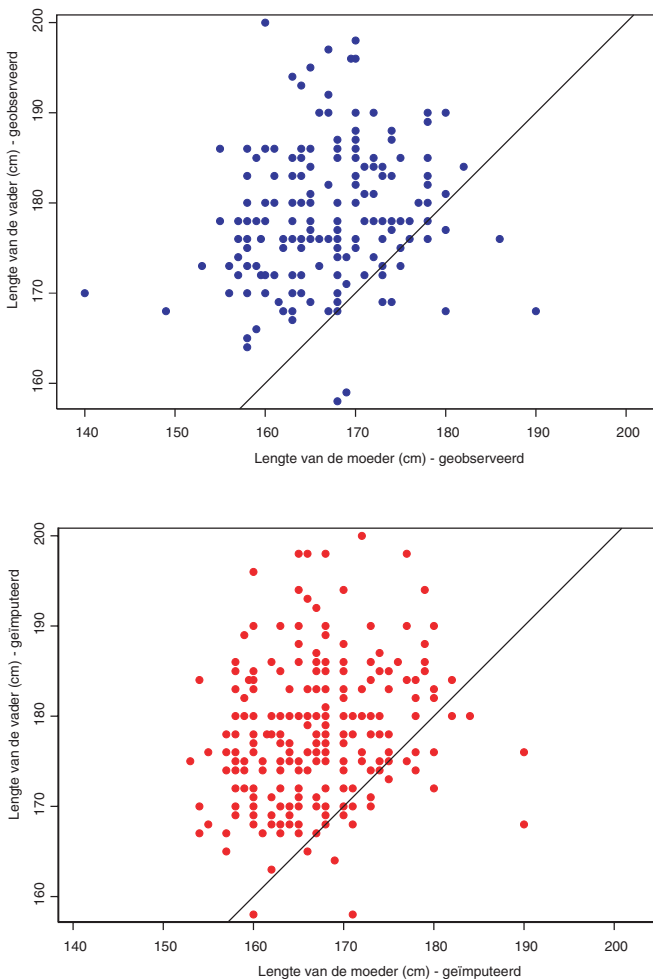
Het idee van Rubin is even geniaal als eenvoudig. Rubin stelt voor om niet één, maar meerdere, bijvoorbeeld drie, geïmputeerde datasets te maken, en die elk apart te analyseren. Dit leidt niet tot één resultaat, maar tot drie. Met behulp van eenvoudige combinatieregels kunnen hieruit de juiste schattingen worden berekend met de juiste onzekerheidsmarges. (Rubin, 1987)

Het vinden van goede imputaties is een kunst op zich. Voor elke onbekende waarneming hebben we een plausibele verdeling van surrogaatwaarden nodig waaruit we imputaties kunnen trekken. Deze verdeling geeft aan wat de waarde had kunnen zijn indien zij geobserveerd was geweest. Stel bijvoorbeeld dat de lengtemeting van een kind ontbreekt,

maar dat we wel weten dat het kind 10 jaar oud is. Uit de Vierde Landelijke Groeistudie weten we dat de lengte van een Nederlands kind met een leeftijd van 10 jaar normaal verdeeld is met een gemiddelde van 143 cm en een standaard deviatie van 6.5 cm. Uit deze verdeling kunnen we surrogaatwaarden trekken. Indien we nog meer van het kind weten, bijvoorbeeld of het een jongen of meisje is, de lengte van de ouders, of de kledingmaat van het kind, dan kunnen we ook die informatie inbrengen om de verdeling te specificeren. Extra informatie reduceert de onzekerheid in de imputaties, en draagt bij aan een kwalitatief betere schatting.

Een probleem ontstaat indien een persoon meerdere ontbrekende gegevens heeft, dus bijvoorbeeld lengte én gewicht. Lengte en gewicht hangen samen, en we moeten daarmee rekening houden bij het maken van de imputaties. Als we dat niet zouden doen, dan zouden we absurde verhoudingen tussen lengte en gewicht kunnen krijgen. Eén strategie is om de gezamenlijke verdeling van lengte en gewicht op te stellen, en daaruit de twee imputaties tegelijkertijd te trekken. In de praktijk is dat een doodlopende weg, met name omdat er geen theoretische multivariate verdelingen bestaan met voldoende flexibiliteit.

TNO werkt sinds 1990 met een veel eenvoudiger methode. Deze methode deelt het probleem op naar de variabelen. Hoe gaat dat precies in zijn werk? We beginnen met een dataset waarin we voor gewicht een willekeurige startwaarde invullen, en imputeren de lengte alsof we gewicht werkelijk geobserveerd hadden. Daarna draaien we de zaak om: we imputeren gewicht onder de aanname dat de zojuist gevonden lengte imputaties de werkelijk geobserveerde waarden waren. Daarna draaien we de rollen van lengte en gewicht weer om, etc. Dit proces stoppen we na, zeg, 10 keer.



Figuur 6 Relatie tussen vader- en moederlengte. Boven: echte data. Beneden: geïmputeerde data.

Dit is eenvoudig, maar hoe is deze methode? In enkele bijzondere gevallen is dit proces equivalent aan een Gibbs sampler, en daarvan zijn de

theoretische eigenschappen bekend. Uit simulaties blijkt dat de methode ook goed werkt voor tal van andere missing data problemen. Mist goed uitgevoerd levert de methode de juiste schattingen met de juiste betrouwbaarheidsintervallen. (Van Buuren *et al*, 2006) Figuur 6 illustreert dat de relatie tussen vader- en moederlengte behouden blijft. Boven staan de echte data, beneden een set van imputaties. Gegeven deze bemoedigende resultaten zou ik graag in het kader van mijn leerstoel verder werken aan een theoretische onderbouwing van de methode. Daarnaast is het nuttig om, samen met collega Hox, de methode geschikt te maken voor datasets met een multilevel structuur.

Tot dusver heb ik de methode beschreven als pure instrument-ontwikkeling, dus als een verbetering op het bestaande instrumentarium. Echte innovatie krijgen we pas indien de methode toegepast wordt, en bij voorkeur ook iets nieuws kan. Is dat hier het geval?

In de afgelopen jaren heb ik de methode op TNO toegepast op projecten van uiteenlopende aard. Voorbeelden zijn:

- imputatie van bloeddruk bij de analyse van overlevingsduren (Van Buuren *et al*, 1999);
- imputatie van ouderlengte bij de analyse van verwijzingsregels voor kleine lengte (Van Buuren *et al*, 2004, Van Dommelen *et al*, 2005); en
- imputatie van het aantal contacten met de huisarts bij analyse van de inkomensgevolgen van eigen bijdragen (Van Buuren *et al*, 1998).

In geen van deze toepassingen was het mogelijk het missing data probleem op een andere wijze op te lossen. De methode kan dus iets dat geen andere methode kan.

De methode biedt ook een springplank naar een nieuwe statistische techniek, die ik hier *tijdrasteren* zal noemen. In veel longitudinale studies lukt het niet om alle personen op exact de juiste tijdstippen te meten, en dat compliceert de statistische verwerking van de gegevens aanzienlijk. Tijdrasteren is een voorbewerking op onregelmatige observaties. Het doel is een nieuwe dataset te genereren waarin alle personen keurig op alle tijdstippen een meting krijgen. De daaropvolgende analyse kan dan eenvoudig van aard zijn. Met multiële imputatie kan de juiste mate van onzekerheid in de data worden meegenomen. De methode heeft poten-

tiel belangrijke toepassingen in longitudinaal onderzoek naar groei en ontwikkeling. Andere interessante uitbreidingen zijn het analyseren van randomised response data, en het genereren van multivariate edits.

Modulaire statistiek

Ik heb laten zien dat innovatie in de methodenleer via twee wegen kan plaatsvinden. Bij beide wegen kunnen we het probleem opdelen in relatief eenvoudige subproblemen. Dat is over het algemeen een goede strategie, en binnen de weg van de toepassing gebeurt dat vrij automatisch. Binnen de weg van de instrumentontwikkeling is opdelen minder gebruikelijk. In het laatste gedeelte van deze rede reik ik een meer stapsgewijs alternatief aan, dat ik ‘modulaire statistiek’ noem.

Onder invloed van de explosieve toename in rekenkracht hebben statistici de afgelopen jaren tal van nieuwe statistische methoden en technieken ontwikkeld. Vaak ontstaan nieuwe modellen door bestaande modellen te combineren. Ze worden daarmee weliswaar algemener, maar ook complexer en daardoor wellicht minder bruikbaar. De prijs van complexiteit is dat het model ons niets leert over de werkelijkheid.

Waarvoor dient een model? Een model is een hulpmiddel voor het brein. Op weg hier naar toe heeft u wellicht de plattegrond van de stad Utrecht bestudeerd. Een plattegrond uit het stratenboek is niets anders dan een model: het is een sterk vereenvoudigde weergave van het weggennet. Kaarten met te weinig, of juist teveel details -bijvoorbeeld ook alle fietspaden- zijn minder bruikbaar voor het beoogde doel, in uw geval het reizen naar het Academiegebouw. Het juiste niveau van detail is niet alleen bij plattegronden belangrijk, maar ook bij statistische modellen. Afhankelijk van het beoogde gebruik moeten we in statistische modellen spaarzaam zijn met het toevoegen van detail uit de data.

Geconfronteerd met een lastig analyseprobleem, kunnen we kiezen uit twee strategieën. In de eerste, integrale benadering, proberen we alle interessante fenomenen in één model te vangen. Dit heeft het voordeel dat de analyse optimaal gebruik kan maken van de aanwezige informatie. Het fitten van het model en de interpretatie van de uitkomsten kan ech-

ter lastig zijn. Een tweede benadering, de modulaire aanpak, tracht het probleem op te delen in subproblemen, die ieder opeenvolgend op optimale wijze worden aangepakt. Dit heeft als voordeel dat de verschillende stapjes redelijk eenvoudig zijn, maar bij het opdelen kan informatie verloren gaan.

Het is buitengewoon merkwaardig dat de statistische literatuur zich vrijwel exclusief bezighoudt met de integrale benadering. Nieuwe methoden en modellen veralgemeniseren de bestaande, vaak door het toevoegen van parameters. Ze maken dus grotere sprongen. Ik heb mij daar zelf ook schuldig aan gemaakt. Op blz. 128 van mijn proefschrift introduceerde ik een zeer algemene verliesfunctie, die ik destijds de nogal pretentieuze naam ‘omnibus loss function’ gaf (Van Buuren, 1990).

Hiermee kon men alle problemen uit het boek van Gifi (Gifi, 1990) oplossen, maar ook nog een heleboel meer. Ik kan u nu wel verklappen dat de ‘omnibus loss function’ nooit erg succesvol is geweest. Bij mijn weten heeft niemand er de afgelopen 15 jaar naar verwezen, inclusief ik zelf. De integrale benadering is een prima middel om te promoveren, en dwingt wellicht eenzelfde soort bewondering af als die voor de sprong van Beamon. Ze levert echter niet het meest bruikbare instrumentarium op.

Vanaf deze plaats wil ik pleiten voor *modulaire statistiek*, de andere benadering. In de modulaire statistiek deelt de gebruiker het analyseprobleem op in subproblemen, en lost elk van deze subproblemen apart op met kleine, overzichtelijke, en gespecialiseerde instrumenten. De totale analyse is te beschrijven als een aantal geschakelde modellen en technieken. De verbinding tussen deze stappen wordt gevormd door data. Elke stap heeft een data invoer en een data uitvoer. De uitvoer van de ene stap vormt de invoer voor de volgende. Na elke stap kan de gebruiker precies kijken of de stap doet waarvoor die bedoeld is. Door stappen verschillend te schakelen en te nesten kan met een beperkt aantal basismodellen toch complexe situaties worden geanalyseerd.

Een voorbeeld: Daniel Ho en collega's hebben onlangs een methode beschreven om paal en perk te stellen aan de ongebreidelde zoektochten naar statistisch significante resultaten die veel vakgebieden teisteren.

(Ho *et al*, 2005) De methode komt erop neer de data analyse te splitsen in twee stappen. Stap 1 is een pre-processing stap die door middel van matching verschillen in achtergrondvariabelen tussen groepen balanceert. De dataset die voortkomt uit deze stap kan vervolgens gebruikt worden als invoer voor de standaard statistische analyse die men toch al van plan was.

Een ander voorbeeld is multiële imputatie. Stap 1 lost het missing data probleem op, en levert niet één maar bijvoorbeeld drie datasets. Stap 2 is de analyse van elk van de datasets. Stap 3 is het samennemen van de drie analysesresultaten.

Een derde voorbeeld is response conversie, een methode die ik in opdracht van de Europese Commissie heb ontwikkeld voor het verbeteren van vergelijkbaarheid van data uit verschillende landen. (Van Buuren *et al*, 2005) Response conversie hercodeert de scores behaald op verschillende instrumenten naar een gezamenlijke schaal. Stap 1 levert een nieuwe dataset op waarin, voor zover mogelijk, voor verschillen tussen meetinstrumenten is gecorrigeerd. Stap 2 bestaat uit de gebruikelijke tabuleringen.

Een vierde voorbeeld is het kwantificeren van categorische data. Optimal scaling technieken kennen aan elke categorie één of meerdere schaalwaarden toe die optimaal zijn voor een gegeven lineair model. Het feit dat schaalwaarden kunnen verschillen tussen modellen compliceert de interpretatie van de schaalwaarden. In zijn presidential address uit 1988 suggereert Jan de Leeuw daarom een twee-staps strategie. (De Leeuw, 1988) Daarin wordt het vinden van schaalwaarden losgekoppeld van de data analyse in stap twee. Ook dit is een voorbeeld van een modulaire aanpak.

Het aantal voorbeelden is naar believen uit te breiden:

- cluster analyse gebeurt vaak in twee stappen;
- de bootstrap is een modulaire methode waarbinnen de hoofdanalyse wordt genest;
- LISREL modellen kan vaak men opvatten als een combinatie van factor analyse en regressie analyse;

- multilevel analyse kan men benaderen met een regressie analyse per niveau;
- het gebruik van standaard deviatie scores vereenvoudigt de daaropvolgende statistische analyse, etc.

Overigens is met al deze technieken voorzichtigheid geboden. Collega Hoijtink heeft laten zien dat het onoordeelkundige exporteren van factorscores naar een volgende analyse een bias kan introduceren, en heeft methoden voorgesteld die hiervoor corrigeren. (Hoijtink & Boomsma, 1996)

Er is weinig onderzoek gedaan naar de wijze waarop de overdracht van de data tussen de verschillende stappen geregeld kan worden. De modulaire statistiek is een groot, nog braakliggend terrein. Daar valt volop moois te bouwen.

Conclusie

Aan het eind van het verhaal gekomen wil ik het een en ander samenvatten. Statistiek speelt een centrale rol in het onderzoek naar preventie. Vernieuwing van het statistisch gereedschap blijft nodig om ook in de toekomst antwoord te kunnen geven op belangrijke vragen in het preventie onderzoek.

Ik heb twee wegen naar vernieuwing geschetst: de weg van de toepassing, en de weg van de instrumentontwikkeling. Ik hoop dat duidelijk is geworden dat toepassing niet zonder instrumentontwikkeling kan, en instrumentontwikkeling niet zonder toepassing. Welke weg men ook bewandelt, vele kleine stapjes overbruggen een grotere afstand dan één grote sprong.

Bob Beamon was een fantastische verspringer, maar zelfs hem is het nooit gelukt om naar de maan te springen. Toch kunnen we er komen. Met kleine stapjes.

Geachte aanwezigen,

De Stichting Lorentz-van Iterson Fonds TNO wil ik danken voor de waardering die blijkt uit het instellen van het bijzonder hoogleraarschap in de Toegepaste statistiek van preventie onderzoek.

Ook de Raad van Bestuur TNO, de directie TNO Kwaliteit van Leven, de universiteit en de faculteit wil ik danken voor het in mij gestelde vertrouwen. Ik zal mijn best doen het waar te maken.

Het feit dat ik hier nu voor u sta is de uitkomst van een jarenlang proces, waarin de ideeën langzamerhand zijn gerijpt en steeds vastere vormen hebben gekregen. Twee mensen hebben mij daarbij enorm geholpen. Aan TNO zijde is dat Erik Verrips, die altijd heeft gedaan alsof het de normaalste zaak van de wereld was dat ik professor zou worden. Aan de Utrechtse kant is dat Peter van der Heijden, die de concrete mogelijkheid bood en me met raad en daad terzijde stond. Deze leerstoel is ook jullie verdienste.

Ik voel me bevoorrecht dat ik de afgelopen jaren op dit belangrijke en boeiende terrein heb mogen werken, en ik hoop dat nog lange tijd te blijven doen. Uiteraard weer in goede samenwerking met de medewerkers van de afdelingen Jeugd, Preventie en Bewegen en Kwaliteit in de Zorg van TNO Kwaliteit van Leven in Leiden, en de medewerkers van disciplinegroep Methoden & Technieken van de Universiteit Utrecht.

Ik dank u allen zeer dat u de moeite hebt willen nemen naar mij te komen luisteren. U zult ondertussen uitkijken naar de receptie. Ik ook. U kunt mij volgen richting Senaatszaal. Uit mijn verhaal kunt u opmaken dat u dat het beste kunt doen met kleine stapjes.

Ik heb gezegd.

de Leeuw J (1988). Multivariate analysis with linearizable regressions. *Psychometrika*, 53, 437-454.

Gifi A (1990). *Nonlinear multivariate analysis*. New York: Wiley.

Herngreen WP, Reerink JD, Noord-Zaadstra BM van, Verloove-Vanhorick SP, Ruys JH (1992). SMOCC: design of a representative cohort-study of live-born infants in The Netherlands. *European Journal of Public Health*, 2, 117-122.

Ho DE, Imai K, King G, Stuart EA (2005). *Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference*. Assessed from <http://www.stat.columbia.edu/~gelman/blog/> on Dec 14, 2005.

Hooijink H, Boomsma A (1996). Statistical inference based on latent ability estimates. *Psychometrika*, 61, 313-330.

Jacobusse G, van Buuren S, Verkerk PH (2006). An interval scale for development of children 0-2 years. *Statistics in Medicine*, in press.

Little RJA, Rubin DB (2002). *Statistical analysis with missing data. Second edition*. New York: Wiley.

Lord FM (1975). The 'ability' scale in item characteristic curve theory. *Psychometrika*, 40, 205-217.

Rubin DB (1987). *Multiple imputation for nonresponse in surveys*. New York: Wiley.

Schaapveld K, Hirasing RA (1997). *Preventiegids. 2e herziene druk*. Assen: Van Gorcum.

van Buuren S (1990). *Optimal scaling of time series*. M&T series 16. Leiden: DSWO Press.

van Buuren S, Perenboom RJM, Rhodes MGH, Boshuizen HC (1998). *Gebandicapten Informatie DoorDenk Systeem. GIDS prototype eigen betalingen*. TNO Rapport, PG98.014.

van Buuren S, Boshuizen HC, Knook DL (1999). Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in Medicine*, 18, 681-694.

van Buuren S, van Dommelen P, Zandwijken GR, Grote FK, Wit JM, Verkerk PH (2004). Towards evidence based referral criteria for growth monitoring. *Archives of Diseases in Childhood*, 89, 336-341.

van Buuren S, Eyres S, Tennant A, Hopman-Rock M (2005). Improving comparability of existing data by Response Conversion. *Journal of Official Statistics*, 21, 53-72.

van Buuren S, Brand JPL, Groothuis-Oudshoorn CGM, Rubin DB (2006). Fully conditional specification in multivariate imputation. *Journal of Statistical Computation and Simulation*, in press.

van Dommelen P, van Buuren S, Zandwijken GRJ, Verkerk PH (2005). Individual growth curve models for assessing evidence-based referral criteria in growth monitoring. *Statistics in Medicine*, 24, 3663-3674.