STAtOR

periodiek van de VVS jaargang 12 nummer 2, juli 2011

Hoe de p-waarde te gebruiken bij beslissingen

Eerste Kamerverkiezingen

Hoe optimalisatie in een aantal jaren in het DNA van TNT Express is gekomen

Karl Pearson tussen statistiek en filosofie

Rechtstreeks verwachtingen evalueren of de nul-hypothese toetsen?

Generalized reliability in industriële user studies

STAtOR
Jaargang 12, nummer 2, juli 2011

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
Joaquim Gromicho (hoofdredacteur), Ana Isabel Barros, Johan van Leeuwaarden, Mirjam Moerbeek, Gerrit Stemerdink (eindredacteur), Hilde Tobi. Vaste medewerker: Fred Steutel

Kopij en reacties richten aan
Prof. dr. J.A.S. Gromicho (hoofdredacteur), Faculteit der Economische Wetenschappen en Bedrijfskunde, afdeling Econometrie, Vrije Universiteit, De Boelelaan 1105, 1081 HV Amsterdam, telefoon 020-5986010, mobiel 06-55886747, <j.a.dossantos.gromicho@vu.nl>.

Bestuur van de VVS
Voorzitter:
prof. dr. Jacqueline Meulman <president@vvs-or.nl>
Secretaris:
dr. Irene Klugkist <bestuur@vvs-or.nl>
Penningmeester:
dr. Ad Ridder <penningmeester@vvs-or.nl>
Studentlid:
Maarten Kampert (Bsc) <student@vvs-or.nl>
Overige bestuursleden:
prof. dr. Fred van Eeuwijk (BMS), prof. dr. ir. Stan van Hoesel & dr. John Poppelaars (NGB), dr. Eric Cator (SMS), dr. Michel van de Velden (ECS), dr. Andries van der Ark (SWS),

Leden- en abonnementenadministratie van de VVS
VVS, Postbus 244, 6700 AE Wageningen, telefoon 0317 - 419572, fax 0317 - 421364, <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
www.vvs-or.nl

Advertentieacquisitie
René Luijk, p/a Leids Universitair Medisch Centrum, afdeling Medische Statistiek en Bio-informatica, Postzone S5-P, Postbus 9600, 2300 RC Leiden, 06 244 32 892, <adverteren.stator@vvs-or.nl>. STAtOR verschijnt in maart, juni, september en december.

Ontwerp en opmaak
Pharos | M. van Hootegem, Nijmegen

Druk
Drukkerij Zoeteweij, Yerseke

Uitgever

ISSN 1567-3383

Inhoud

Vakantie, een tijd vol keuzes

Mensen zijn er in soorten. Zo schijnen sommigen zich te kunnen ontspannen tijdens vakanties en ondervinden anderen juist stress vanwege de vele extra beslissingen die ze dan moeten nemen. Men zou, overdreven gesteld, kunnen zeggen dat de vrijheid van de vakantie onlosmakelijk de dwang van de keuze met zich brengt.

Statistici zijn net mensen: ook zij zijn er in soorten en ook zij zitten voortdurend met keuzes. Een keuze die vaak voorkomt is of men vóór of tegen Bayes is. Aan dit probleem worden in dit nummer van Stator maar liefst twee artikelen gewijd. Aart de Vos laat zien dat ‘klassiek’ en Bayes elkaar niet hoeven te bijten, maar heel goed als elkaars aanvulling beschouwd kunnen worden. En Rens van de Schoot gaat óók in in de complementaire functie van beide methoden. Voor wie toevallig in deze vakantie naar Denemarken gaat: in de slotkapel van het kasteel Frederiksborg in Hillerød hangen de wapenschilden van de leden van de Orde van de Olifant, de hoogste Deense onderscheiding. Een van die schilden is van de beroemde Deense natuurkundige Niels Bohr en zijn wapenspreuk luidt ‘Contraria sunt Complementa’. Dit slaat op Bohr’s interpretatie van de quantummechanica, maar na het lezen van de artikelen van De Vos en Van de Schoot kan men het ook op de statistiek toepassen.

Trouwens, over kiezen gesproken: de manier waarop de afgelopen maanden de Eerste Kamer werd gekozen leek voor de gewone kiezer wel hogere wiskunde. De redactie is verheugd u een zeer actueel en verhelderend overzicht van dit probleem te kunnen bieden. Jacob Jan Paulus en Jaap Praagman tekenen daarvoor.

Een heel ander artikel is geschreven door Hein Fleuren, Marco Hendriks en Annelies Woutersen. Zij beschrijven hoe TNT Express stapje voor stapje wordt veroverd door toepassingen van de OR.

De filosoof Richard Starmans vertelt over de belangrijke rol van Karl Pearson in de ontwikkeling van de statistiek. De redactie kijkt nu al uit naar enkele andere artikelen die hij heeft toegezegd te zullen schrijven over andere grootheden die aan de wording van ons vakgebied hebben bijgedragen. Zéér aanbevolen.

Dit is nog niet alles, u kunt er ook voor kiezen om als eerste het artikel van Jan Engel te lezen. Dan zult u zien dat de bekende Cronbachs alpha methode uit de sociale wetenschappen uitstekend kan worden gegeneraliseerd naar industriele studies.

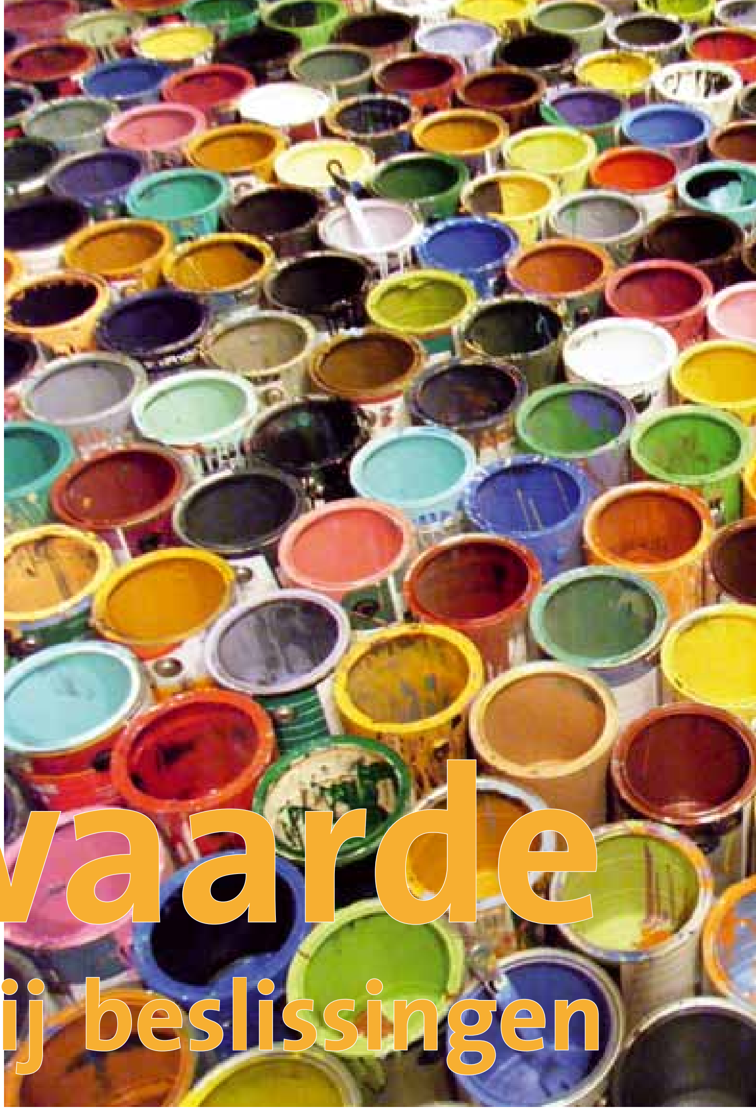
En natuurlijk hebben ook onze vaste columnisten rekening gehouden met de tijd van het jaar: hun bijdragen zijn luchtig en verstrooiend.

Ten slotte een In Memoriam voor Joop Kemperman. Hij was een van de vier promovendi van David van Dantzig en is in Nederland wat minder bekend geworden doordat hij al snel daarna naar Amerika verhuisde. Daar was hij een productief wetenschapper, met maar liefst 24 promovendi. Fred Steutel herdenkt hem.

Kortom, keus genoeg, maar laat het u niet tot stress voeren. Gooi desnoods een muntje op als u geen keus kunt maken, doet u in de vakantie toch nog aan kansrekening.

De redactie wenst u veel leesplezier.

Gerrit Stemerdink



Hoe de p-waarde te gebruiken bij beslissingen

ER ZIT SIGNIFICANT TE WEINIG VERF IN MIJN POTJES! MOET IK IETS DOEN?

Aart F. de Vos

Geen statistisch onderwerp is zo controversieel als de *p*-waarde. Bayesianen drijven er de spot mee, maar het is een basiselement in iedere standaardcursus statistiek. De *p*-waarde is de kans dat als de nulhypothese (H_0) waar is, de data zo extreem zijn als ze zijn, of nóg extremer. Dit krijgt betekenis in vergelijking met een norm, het significantie niveau α, dat meestal 0,05 genomen wordt. Als $p<\alpha$ dan is iets significant.

Centraal in de verwarring over de betekenis hiervan staat de '*p*-value fallacy': de *p*-waarde aanzien voor de kans dat de nulhypothese waar is. Dat is natuurlijk fout; die kans bestaat niet eens in de klassieke visie. Maar deze is wel nodig voor het nemen van beslissingen onder onzekerheid. Wat moet een bedrijfseconoom bij een steekproefuitkomst 'er zit significant te weinig verf in mijn potjes' anders dan denken dat je *iets moet doen* omdat de kans groot is dat er wat mis is? Dat er in het statistiekonderwijs geen koppeling wordt gelegd tussen toetsen en beslissen vind ik dramatisch. De statistiekopleiding wordt hierdoor vrijwel zinloos, soms zelfs gevaarlijk: de '*p*-value fallacy' is een variant van de 'prosecutors fallacy'. Het verloop van het proces Lucia de Berk is een pregnant voorbeeld – al ging daar nog veel

meer mis. ‘Als verdachte onschuldig is *(H_0)* is het gebeurde uiterst onwaarschijnlijk, dus zal verdachte wel schuldig zijn.’

Ook studenten die uitgebreid statistiek hebben gehad zijn verbijsterd als ik het volgende eenvoudige voorbeeld voorleg. Een student doet een multiple choice-tentamen: 50 vragen met steeds drie mogelijke antwoorden. Er zijn 27 goede antwoorden. Gegokt? Als de kans per vraag 1/3 is, is de kans op 27 *of meer* goed 0,002. Zeer significant. Dus niet gegokt? Als het een redelijk goede student is die soms leert en soms feest, is 65% kans dat hij gegokt heeft redelijk plausibel. 27 antwoorden goed is nog onwaarschijnlijker als hij geleerd heeft (en zijn kans per vraag 0,75 is) dan als hij gegokt heeft. Het gaat om de aannemelijkheidsverhouding van de data (27 goed dus *meer* doet niet ter zake). *Posterior odds is prior odds times likelihood ratio*.

Klassieke toets

Dus weg met de *p*-waarde? Nee. Ik denk een oplossing te hebben gevonden (op zijn minst voor bedrijfseconomen) in de vorm van een economische onderbouwing van de klassieke toets. Mijn stelling is dat er eenvoudige Bayesiaanse modellen zijn waarin het format $p<\alpha$ optimaal is voor het nemen van beslissingen. Die modellen zijn zinvol in een sequentiële procedure die een brug vormt tussen de frequentistische en de Bayesiaanse aanpak. De aanpak bestaat uit drie stappen.

1. Kies een α die zinnig is in de context van een beslissing (ik kom hierop terug).
2. Bereken de p-waarde en bepaal of $p<\alpha$; en zo ja, dan is iets significant.

Gevolgd door een beslissing die geparafraseerd kan worden met de slagzin:

3. Als iets niet significant is: niet verder over nadenken. Als wel: huur een Bayesiaan.

Net zoals in de gezondheidszorg. Een huisarts heeft twee mogelijke diagnoses: pluis of niet pluis. In het laatste geval verwijst hij naar de specialist. Met ‘pluis’ bedoelt hij dat de kans dat er iets aan de hand is te klein is om de kosten van nader onderzoek te rechtvaardigen.

Met goede software kan de student worden opgeleid tot ‘huiseconoom’ (stap 1 en 2). De Bayesiaan is dan de specialist. In vrijwel alle situaties is een volwaardige Bayesiaanse analyse lastig en dus kostbaar. Het vergt het opstellen van (deels subjectieve) priors en verliesfuncties en het in kaart brengen van diverse beslissingen.

De klassieke toets is typisch geschikt voor situaties met een hoge $P(H_0)$, de prior kans dat er niets de hand is. En functioneert als een soort zelfbescherming. Wij worden voortdurend bestormd door 'significante' feiten die iets zouden kunnen betekenen. Er over nadenken kost veel tijd en is daarom kostbaar. Meestal is het loos alarm. Dat is het geval in een fractie $P(H_0)(1-\alpha)$ van alle gevallen, als je de de regel 'denk na als de *p*-waarde kleiner is dan α hanteert.

De klassieke toets stelt 'verwerp H_0 als $p<\alpha$'. Met $\alpha = 0{,}05$. Of soms 0,01 om vaag omschreven redenen. Studenten leren braaf hoe je een *test statistic (S)* kan specificeren waarvan je de verdeling kent en dus de overschrijdingskans kunt opzoeken van de 'gerealiseerde *S*'. De *p*-waarde dus.

Nu kan bewezen worden dat vrijwel alle in een standaardboek *Statistics for Business Students* vermelde eenzijdige toetsen een Bayesiaanse interpretatie hebben die voorschrijft: doe iets als $S>k$ (of $S<k$, afhankelijk van de situatie). En dat dit ook geschreven kan worden als: 'doe iets (op grond van een kosten-baten afweging) als $p<\alpha_{im}$'. Met *p* de *p*-waarde en α_{im} de impliciete α die afhangt van priors en verliesfuncties.

Het verschil tussen frequentisten en Bayesianen is eigenlijk dat de laatsten beweren te weten hoe α moet worden bepaald. Dat doen ze overigens zelden. Bayesianen conditioneren op de data. De kansverdeling van data onder H_0 interesseert ze niet. Wat jammer is, omdat de representatie van een beslissingsprobleem in termen van *p* en α als voordeel heeft dat *p* alleen afhangt van H_0 en de data, en α van priors en verliesfucties. De coherente manier om over de situatie te denken is in termen van een gegeven *p* en een prior voor α. Waarbij het bij sequentieel beslissen dus een optie is om meer informatie over α te vergaren (een Bayesiaan in te huren).

De eerste uitdaging is nu om software te ontwikkelen voor stap 1. Een eenvoudig model dat de belangrijkste inputs en schattingen van kosten omzet in zinnige waarden voor α. De range van zinnige waarden zal over het algemeen vrij groot zijn: verliesfuncties in termen van α zijn vlak rond het optimum. Het is vooral essentieel om echt foute keuzes te vermijden. Interessant is het ook om na te gaan wanneer 0,05 echt fout is.

De formele kant in vogelvlucht

In het basisgeval is er één parameter θ van belang. H_0 is $\theta = \theta_0$, H_1 is $\theta > \theta_0$ (of $\theta < \theta_0$). De formule die aangeeft wanneer je moet beslissen iets te doen, omdat het verwachte nut groter is dan dat van niets doen is: doe iets als

$$BF(S) = \int \frac{f(S|\theta)}{f(S|\theta_0)}\, \pi(\theta|H_1)\, d\theta > \frac{\pi(H_0)}{\pi(H_1)} \frac{L(1,0)}{L(0,1)} = k \qquad (1)$$

(f = kansdichtheid; π = prior kans(dichtheid))

Aan de linkerkant staat de 'Bayes Factor' ($BF(S)$). (= $P(S|H_1) / P(S|H_0)$). Die is afhankelijk van $\pi(\theta|H_1)$, de subjectieve voorverdeling van θ onder H_1. Oftewel de lastige vraag: 'Wat zou er aan de hand kunnen zijn?' Aan de rechterkant de *prior odds* van de hypothesen, maal de ratio van de verwachte verliezen bij het nemen van de twee mogelijke verkeerde beslissingen: $L(1,0)$ iets doen terwijl er niets aan de hand is) en $L(0,1)$ (ten onrechte niets doen).

De impliciete α in (1) is

$$\alpha_{im} = P_{S|Ho}\,(BF(S) > k) \qquad (2)$$

een ongebruikelijk mengsel van frequentistisch ($P_{S|Ho}$) en Bayesiaans. Bayesianen kijken naar de ongelijkheid $BF(S) > k$. Dat heeft als bezwaar dat aan beide kanten priors voorkomen. Het format $p < \alpha_{im}$ heeft dit bezwaar niet. Het kan bovendien ook worden gebruikt wanneer $L(0,1)$ afhangt van

θ, wat normaliter het geval zal zijn. De formule wordt dan $\alpha_{im} = P_{S|Ho}\,(S>k)$ waarbij k met de Bayesiaanse machinerie berekend kan worden.

Een voorbeeld

Volledig Bayesiaans beslissen is lastig. Maar verschillende beslissingen onderling coherent maken is eenvoudiger. Stel θ is de verwachting van het gemiddelde tekort aan verf in mijn potjes en S het gemiddelde tekort in mijn steekproef. Als het echt te weinig is krijg ik een boete. Vergelijking (1) beschrijft de situatie.

Onderstaande tabel is gemaakt met (1) en (2). $S \sim N(\theta,1)$, $\pi(\theta|H_1) = \exp(-\theta)$. S is data, $S_0|H_0 \sim N(\theta_0,1) = N(0,1)$.

S	0,91	1,5	1,645	2	2,193	2,63
BF(S)	1,12	1,91	2,23	3,43	4,46	8,92
$P(S_0 > S\|H_0)$	0,181	0,067	0,050	0,023	0,014	0,006

De $\alpha = 0{,}05$ hoort bij $S = 1{,}645$, welbekend. Stel dat bekend is dat dit de optimale toetsdrempel is om nog een extra steekproef te trekken.

Een denkbare opgave voor het tentamen statistiek voor bedrijfskundigen in 2020?

a. laat zien dat als die extra steekproef 2 maal zo duur wordt men moet gaan werken met een α van 0,014.
b. laat zien dat als $\pi(H_1)$ stijgt van $1/(2n+1)$ naar $1/(n+1)$ $\alpha = 0{,}181$ genomen moet worden.
c. maak een voorbeeld van veranderingen die $\alpha = 0{,}006$ opleveren.

Tweezijdig toetsen

Als mijn steekproef doet vermoeden dat er te weinig verf in mijn potjes zit, met een p-waarde van 0,03 moet ik dat dan negeren omdat er ook te veel verf in mijn potjes had kunnen zitten? Economen leren dat ze moeten bedenken of ze eenzijdig of tweezijdig moeten toetsen en in het laatste geval de p-waarde met 2 moeten vermenigvuldigen. Tamelijk onzinnig. Verschillende waarden van α zijn relevant voor te veel of te weinig verf. De verliesfuncties verschillen. En beter nadenken ga je in dit geval alleen over de α die hoort bij te weinig. Meer in het algemeen geldt dat p-waarden hun magische betekenis verliezen als men beseft hoe onzeker men is over de relevante α.

Conclusie

De klassieke toets en Bayesiaans beslissen onder onzekerheid kunnen geïntegreerd worden tot iets zinvols. En een gewoon voorbeeld over verf in potjes biedt een ander perspectief dan abstracte pogingen tot verzoening van beide richtingen (Andrews, 1994; Berger, 2003; zie ook De Vos & Francke, 2008). De grootste uitdaging is om studenten te leren een redelijke α te kiezen, zodat het huren van Bayesianen beperkt kan blijven tot echt belangrijke gevallen.

LITERATUUR

Andrews, D.W.K. (1994). The large sample correspondence between classical hypothesis tests and Bayesian posterior odds tests. *Econometrica* 62, 1207-1232.
Berger, J.O. (2003). Could Fisher, Jeffreys and Neyman have agreed on testing? *Statistical Science* 18, 1-32.
de Vos, A.F. & Francke, M.K. (2008). Bayesian Unit Root Tests and Marginal Likelihood (www1.fee.uva.nl/pp/bin/1015fulltext.pdf)

AART DE VOS is econometrist en verbonden aan de faculteit der economische wetenschappen en bedrijfskunde van de Vrije Universiteit Amsterdam. E-mail: <a.f.de.vos@vu.nl>

Ter gelegenheid van zijn pensionering vindt op ***30 september 2011*** een ***afscheidssymposium*** *plaats over het onderwerp van dit artikel. Topspecialisten op dit gebied hebben hun medewerking toegezegd. Zie ook de website <http://personal.vu.nl/a.f.de.vos/>.*

EERSTE KAMERVERKIEZINGEN

Jacob Jan Paulus & Jaap Praagman

De Eerste Kamer wordt gekozen via een getrapte verkiezing, de burgers kiezen Provinciale Statenleden die vervolgens de Eerste Kamerleden kiezen. Begin maart 2011 waren de verkiezingen voor de Provinciale Staten en door de vraag of de gedoogcoalitie een meerderheid zal halen is er dit jaar veel aandacht voor de Eerste Kamerverkiezing. In de landelijke media is er veel gespeculeerd over strategisch stemmen [1,2]. Want door strategische samenwerking kunnen partijen meer Kamerzetels halen. Tijd voor een wiskundige analyse.

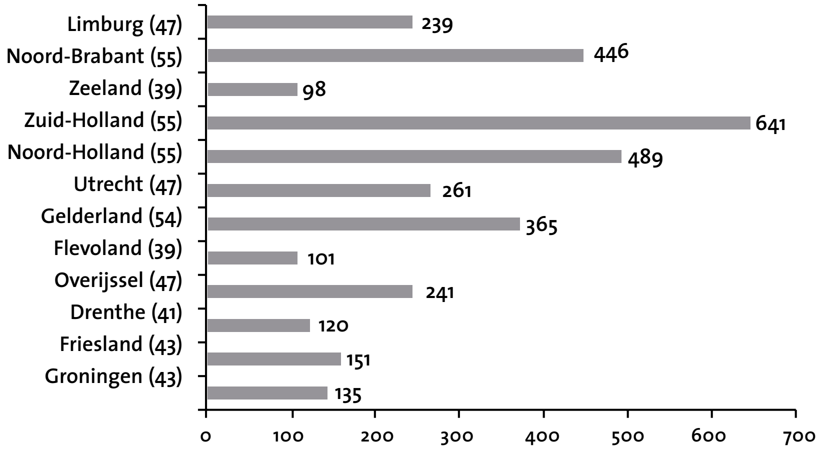
Spelregels

Bij de Eerste Kamer verkiezing op 23 mei 2011 mag elk Statenlid één stem uitbrengen. Omdat het inwoneraantal per provincie verschilt, worden de stemmen gewogen. Deze weging, de zogenaamde *stemwaarde*, wordt per provincie bepaald door het inwoner aantal te delen door het honderdvoud van het aantal Statenleden in die provincie. De stemwaarde en het aantal Statenleden per provincie staan in figuur 1 weergegeven.

De totale waarde van de stemmen uitgebracht op een partij is het behaalde *stemcijfer* van die partij. De stemcijfers bepalen de zetelverdeling in de Eerste Kamer en dat gaat als volgt [3,4]. Met de kiesdeler worden de zogenaamde kale zetels verdeeld. Het aantal kale zetels dat een partij krijgt is gelijk aan het naar beneden afgeronde quotiënt van stemcijfer en kiesdeler. Deze kiesdeler is de som van de stemcijfers gedeeld door het aantal Kamerzetels. In 2011 is de kiesdeler 166561/75. De resterende zetels worden vervolgens één voor één verdeeld zodanig dat de eerstvolgende restzetel wordt toegekend aan die partij waarvoor na toekenning van die zetel het quotiënt van stemcijfer en aantal zetels het grootste is.

Verkiezingen 2011

In tabel 1 staat per provincie de zetelverdeling van de Provinciale Staten na de verkiezing van maart 2011. Stemt ieder Statenlid bij de Eerste Kamer-



Figuur 1. Provincie (aantal statenleden) en stemwaarde

verkiezing op de eigen partij dan resulteert dat in de stemcijfers zoals vermeld in de voorlaatste kolom. Delen door de kiesdeler en naar beneden afronden levert het aantal kale zetels in de laatste kolom. (De lokale partijen zijn gebundeld in de Onafhankelijke Senaats Fractie, OSF.)

In totaal zijn daarmee 69 zetels verdeeld. De resterende 6 zetels worden één voor één verdeeld volgens het hiervoor beschreven principe van grootste gemiddelde stemcijfer per zetel, zie tabel 2.

	GR	FR	DR	OV	FL	GD	UT	NH	ZH	ZL	NB	LB	Stem-cijfer	Kale zetels
CDA	5	8	6	11	4	9	6	5	6	6	10	10	24238	10
VVD	6	6	9	8	9	11	11	13	12	7	12	8	34518	15
PVV	3	4	4	4	6	6	5	6	8	5	8	10	21064	9
PvdA	12	11	12	9	6	9	7	11	10	7	7	6	29639	13
GroenLinks	3	2	2	2	2	4	4	5	3	1	3	3	10656	4
D66	3	2	2	3	3	4	5	6	5	2	5	2	13781	6
SP	6	3	4	4	3	5	4	5	5	3	8	6	16825	7
ChristenUnie	3	3	2	3	3	3	2	1	2	2			5708	2
SGP				2	1	2	1		2	4			3248	1
PvdD	1				1	1	1	1	1		1		2438	1
50PLUS				1	1	1	1	1	1		1	2	3022	1
OSF	1	4						1		2			1424	0
													166561	69

Tabel 1. Zetelverdeling van de Provinciale Staten na de verkiezing van maart 2011

Rekenen

Interessant wordt het doordat de Statenleden in principe vrij zijn in hun keuze en dus ook op een andere partij mogen stemmen. Dat biedt de mogelijkheid om afspraken te maken tussen samenwerkende partijen en zo de zetelverdeling te beïnvloeden. Al begin april was er sprake van een mogelijke afspraak tussen 50PLUS en de OSF. Als bijvoorbeeld alle OSF stemmen terecht komen bij de 50PLUS partij, dan krijgt de 50PLUS partij een stemcijfer van 3022+1424 = 4446. Wat direct twee kale zetels oplevert in plaats van één. Dit ten koste van de SP, omdat er dan nog maar vijf restzetels te verdelen zijn. Maar met een kleine verhoging van het stemcijfer, bijvoorbeeld doordat een Zeeuws GroenLinks Statenlid zijn stem aan de SP geeft, kan de SP voorbij de PVV schuiven. Wat vervolgens voor de gedoogcoalitie CDA, VVD en PVV weer aanleiding kan zijn om een defensieve strategie te zoeken, et cetera. Dit vraagt om een wiskundige analyse naar de strategische stemmogelijkheden.

Geen vaste kiesdeler

Vanuit wiskundig perspectief is het werken met de kiesdeler en vervolgens één voor één verdelen van de restzetels onnodig ingewikkeld. Waar het in wezen om gaat is het stemcijfer per verdeelde zetel. In het eerder genoemde voorbeeld, waar iedere partij op zichzelf stemt, krijgt de SP de laatste restzetel met een stemcijfer van 2103,13 per zetel. Iedere toegewezen zetel representeert dus een stemcijfer van 2103,13 of meer. Omgekeerd geldt dat iedere partij ook voor elke 2103,13 aan stemcijfer een zetel toegewezen heeft gekregen. Kortom de uiteindelijke zetelverdeling correspondeert met de zetel-

	Stemcijfer/ (zetels +1)	Zetels	Stemcijfer/ (zetels +1)	Zetels	Stemcijfer/ (zetels +1)	Zetels	Stemcijfer/ (zetels +1)	Zetels	Stemcijfer/ (zetels +1)	Zetels	Stemcijfer/ (zetels +1)	Zetels
CDA	2203,45	11	2019,83	11	2019,83	11	2019,83	11	2019,83	11	2019,83	11
VVD	2157,38	15	2157,38	16	2030,47	16	2030,47	16	2030,47	16	2030,47	16
PVV	2106,40	9	2106,40	9	2106,40	9	2106,40	9	2106,40	10	1914,91	10
PvdA	2117,07	13	2117,07	13	2117,07	13	2117,07	14	1975,93	14	1975,93	14
GroenLinks	2131,20	4	2131,20	4	2131,20	5	1776,00	5	1776,00	5	1776,00	5
D66	1968,71	6	1968,71	6	1968,71	6	1968,71	6	1968,71	6	1968,71	6
SP	2103,13	7	2103,13	7	2103,13	7	2103,13	7	2103,13	7	2103,13	8
ChristenUnie	1902,67	2	1902,67	2	1902,67	2	1902,67	2	1902,67	2	1902,67	2
SGP	1624,00	1	1624,00	1	1624,00	1	1624,00	1	1624,00	1	1624,00	1
PvdD	1219,00	1	1219,00	1	1219,00	1	1219,00	1	1219,00	1	1219,00	1
50PLUS	1511,00	1	1511,00	1	1511,00	1	1511,00	1	1511,00	1	1511,00	1
OSF	1424,00	0	1424,00	0	1424,00	0	1424,00	0	1424,00	0	1424,00	0
		70		71		72		73		74		75

Tabel 2. Verdeling van de restzetels in de Eerste Kamer na de Provinciale Statenverkiezing van maart 2011

verdeling die we krijgen als we 2103,13 als kiesdeler hanteren. Het hele proces van berekenen van kale zetels en toekennen van restzetels kunnen we dus vervangen door het kiezen van de deler waarbij precies 75 zetels worden verdeeld. Probleem is echter dat deze feitelijk resulterende deler van tevoren onbekend is en afhankelijk van het stemgedrag.

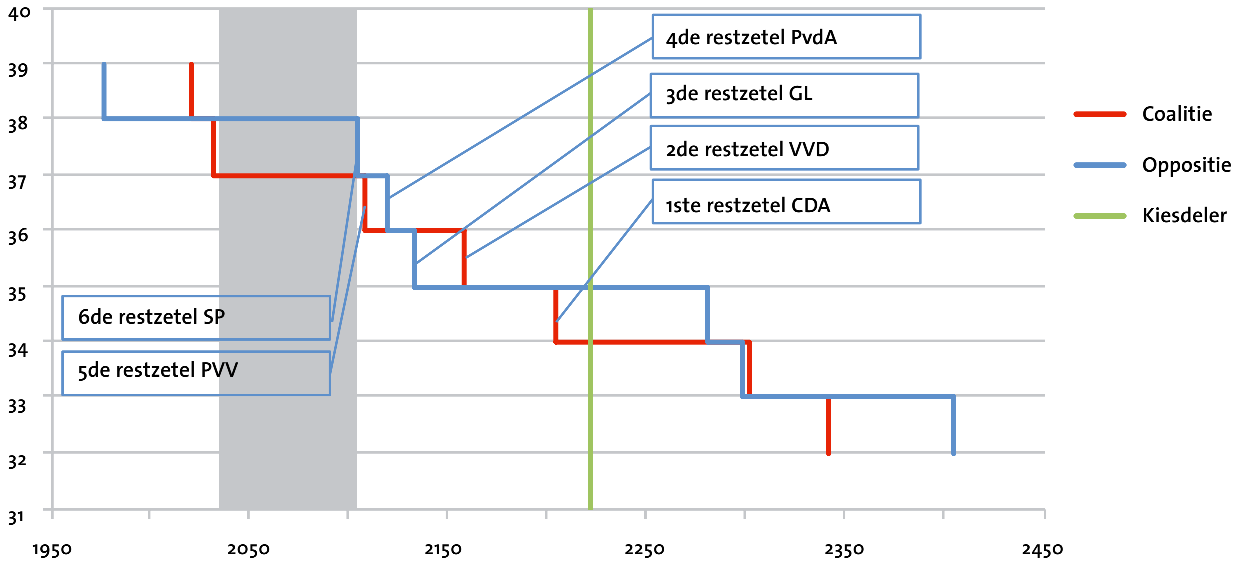
In figuur 2 lichten we dit op een iets andere manier toe, waar het aantal aan een partij toe te wijzen zetels is uitgezet tegen de deler. Weer voor de situatie dat alle Statenleden op hun eigen partij stemmen. De partijen zijn gegroepeerd als coalitie (CDA, VVD en PVV) en oppositie (alle andere partijen). Bij de groene lijn, de officiële kiesdeler, zijn 34 zetels aan de coalitie toegekend en 35 aan de oppositie. Door de deler lager te kiezen dan de kiesdeler, worden er meer zetels verdeeld. Het proces van restzetelverdeling komt dus overeen met het naar links verschuiven van de deler totdat er precies 75 zetels zijn verdeeld.

Door het kiezen van een andere stemstrategie kan een groep partijen de eigen curve beïnvloeden. Als bijvoorbeeld het CDA een stem, en daarmee stemwaarde, aan de PVV geeft zal de trede *5de restzetel PVV* naar rechts verschuiven, maar de trede *1ste restzetel CDA* schuift juist naar links. De kans op een restzetel voor de PVV zou hiermee mogelijk verbeteren omdat het stemcijfer per zetel wordt verhoogd. Echter, de kans op een restzetel voor het CDA zal verslechteren.

Wiskundig model

Om een optimale strategie te bepalen voor een groep partijen G modelleren we het probleem als volgt. Laat variabele X_{pij} het aantal stemmen zijn dat in provincie p door partij i uitgebracht is op partij j, en laat ω_p de stemwaarde zijn in provincie p. Dan is het verkregen stemcijfer van partij j gelijk aan $\Sigma_{p,i}\omega_p X_{pij}$. Laat variabele D de deler zijn en de variabele Z_j het behaalde aantal zetels van partij j. Met α_{pi} geven we het aantal statenleden aan dat partij i in provincie p heeft. De randvoorwaarden zijn als volgt. Er zijn 75 zetels te verdelen; $\Sigma_j Z_j = 75$. Het aantal stemmen uitgebracht komt overeen met het aantal statenleden; $\Sigma_j X_{pij} = \alpha_{pi}$. Het stemcijfer geeft een boven- en ondergrens op het aantal zetels van een partij;

$$Z_j \leq \frac{\Sigma_{p,i}\omega_p X_{pij}}{D} < Z_j + 1.$$



Figuur 2. Het aantal aan een partij toe te wijzen zetels uitgezet tegen de deler

Nu we dit model geformuleerd hebben, lopen we tegen twee dingen aan. Wat te doen met het wiskundig gezien moeilijk oplosbare kwadratisch probleem en wat is de strategie van de partijen die niet in groep *G* zitten? De sleutel voor beide problemen zit in het tijdelijk vastzetten van de deler.

Door het vastzetten van de deler wordt het model lineair en eenvoudiger om op te lossen. We laten daarmee tijdelijk de eis vallen dat 75 zetels verdeeld moeten worden. Worden er minder dan 75 verdeeld, dan is de deler klaarblijkelijk te hoog gekozen. En als er meer dan 75 worden verdeeld, dan is de deler te laag gekozen. Een *binary search* op de deler leidt snel tot een goede keuze.

Voor een gegeven deler kan een groep partijen nu haar beste strategie bepalen, onafhankelijk van de strategie van de anderen. Of bij die deler precies 75 zetels worden verdeeld, is echter wel afhankelijk van de strategie van de andere partijen.

Laat de parameter β_{ij} gelijk zijn aan 1 als partijen *i* en *j* tot dezelfde groep en 0 als dat niet zo is. En als we de variabele voor de deler *D* vervangen door een vast getal δ, wordt het ILP model als volgt:

$$\text{maximize} \sum_j z_j$$

$$\text{s.t.} \sum_j X_{pij} = \alpha_{pi} \ \forall p,i,$$

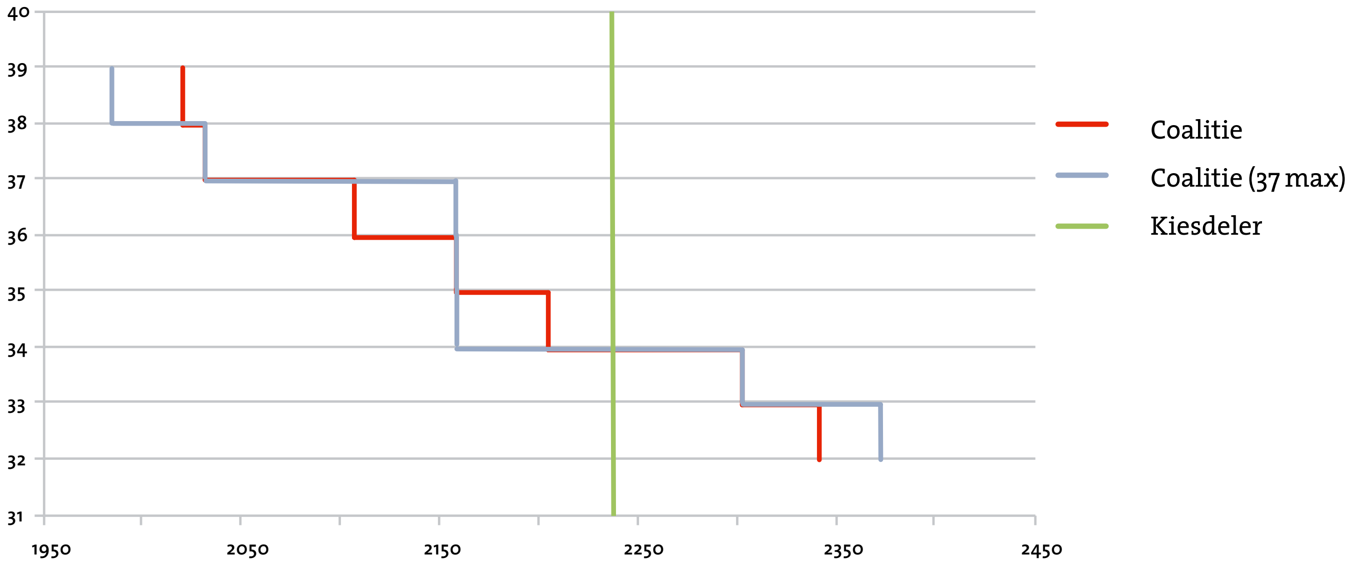
$$X_{pij} \le \alpha_{pi}\,\beta_{ij} \ \forall p,i,j,$$

$$Z_j \le \frac{\Sigma_{p,i}\,\omega_p\, X_{pij}}{\delta} \quad \forall j$$

$$X_{pij}, Z_j \ \text{integer.}$$

Uitkomsten

Zoals eerder geschetst kan de gedoogcoalitie haar kans op een 37ste zetel (de restzetel van de PVV) verstevigen door een deel van het stemcijfer van het CDA naar de PVV te verschuiven. Zoeken naar de grootste waarde van δ waarvoor de gedoogcoalitie 37 zetels kan halen, levert de strategie op die staat afgebeeld in figuur 3. In deze strategie gaat een stemwaarde van 506 van het CDA naar de PVV, door in de provincie Zuid-Holland één CDA-er te laten stemmen op de PVV en omgekeerd in de provincie Groningen één PVV-er op het CDA.



Figuur 3. Strategie voor het vinden van de grootste waarde van δ waarvoor de gedoogcoalitie 37 zetels kan halen

Voor elk zetelaantal kan gezocht worden naar een strategie die bij een zo groot mogelijke waarde van δ dat zetelaantal behaalt. Ofwel, elk zetelaantal heeft een sterkste strategie voor het halen van dat aantal zetels. Het resultaat staat weergegeven in figuur 4 als *Coalitie (X max)*; het maximaal haalbare zetelaantal gegeven een deler. In de vorige grafieken correspondeerde een lijn met één strategie, maar deze lijn correspondeert met het maximaal haalbare over alle mogelijke strategieën. De strategie verschilt dus per trede in de grafiek.

Voor de oppositie kan een vergelijkbare analyse gemaakt worden. Het lijkt politiek gezien niet realistisch om de oppositie als één grote samenwerkende groep te beschouwen. Ter illustratie kiezen we hier voor vier subgroepen; Groep 1 bestaat uit PvdA, SP, D66 en GroenLinks, groep 2 uit ChristenUnie en SGP, groep 3 uit alleen PvdD en groep 4 uit 50PLUS en OSF. Binnen deze beperking is het maximaal haalbare voor de oppositie afgebeeld in figuur 4, naast het eerder bepaalde maximaal haalbare voor de coalitie.

Als alle partijen, gegeven deze groepsindeling, de beste strategie volgen zal de coalitie 36 zetels krijgen en de oppositie 39 zetels. Het valt op dat bij optimale strategieën de 74ste en 75ste zetel worden verdeeld bij delers die heel dicht bij de officiële kiesdeler liggen.

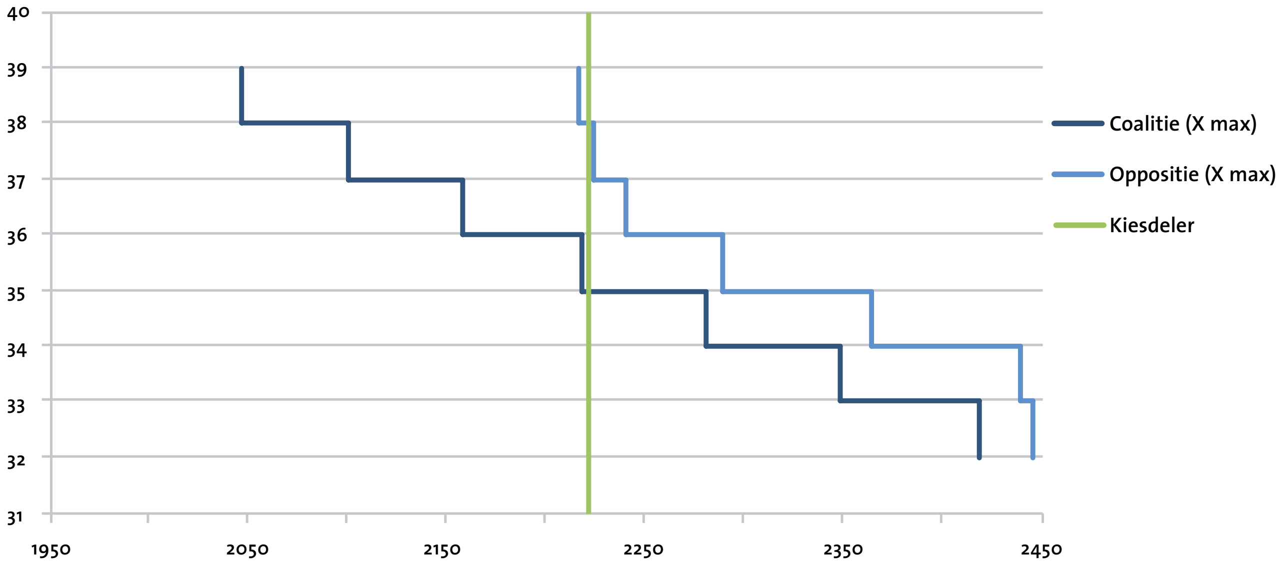
Open eindjes

De wiskundige kant van de verkiezing is hiermee belicht, er zijn echter nog wat losse eindjes. Een bewijs van NP-compleetheid en het bestaan van een puur Nash evenwicht zijn waarschijnlijk niet moeilijk te geven. Echter, groepsindelingen en stemafspraken zijn politieke kwesties, die ongetwijfeld veel moeilijker zijn dan de beschreven wiskunde.

LITERATUUR

1. NOS dossier *Strijd om de Eerste Kamer*: http://nos.nl/dossier/210939-eerste-kamerverkiezingen-2011/thema/serie-strijd-om-de-eerste-kamer/
2. Eén Vandaag uitzending *Verkiezingen een strijd op landelijk niveau,* 17 februari 2011: http://beta.uitzendinggemist.nl/afleveringen/1062892
3. Website Eerste Kamer: www.eerstekamer.nl/
4. Website Kiesraad: www.kiesraad.nl/

Jacob Jan Paulus en Jaap Praagman zijn beiden werkzaam bij Consultants in Quantitative Methods (CQM). E-mail: <paulus@cqm.nl> & <praagman@cqm.nl>



Figuur 4. Het maximaal haalbare zetelaantal gegeven een deler

Hoe optimalisatie in een aantal jaren in het DNA van TNT Express is gekomen

HEIN FLEUREN, MARCO HENDRIKS & ANNELIES WOUTERSEN

We schrijven 2004, een jaar waarin stevige groei voor TNT Express een soort natuurconstante was. Op een namiddag heeft op het CentER for Applied Research een aantal mensen een kort overzichtje gemaakt wat de grote 4 Express spelers (UPS, Fedex, DHL en TNT Express) aan optimalisatie deden. Voor UPS en Fedex is redelijk wat in de literatuur (business en wetenschappelijk) te vinden. Voor DHL ook nog wel wat maar voor TNT... Met deze boodschap naar TNT getogen, alwaar we (achteraf) blijkbaar op het juiste moment bij de juiste man aan tafel kwamen: Marco Hendriks, toendertijd verantwoordelijk voor *business improvement* in Express' netwerken. In dit artikel belichten we wat er zoal bij komt kijken om Operations Research succesvol toe te passen in een groot bedrijf als TNT.

Wat is een Express netwerk?

Grote Express spelers vervoeren dagelijks miljoenen pakjes wereldwijd van zakelijke adressen naar zakelijke adressen, en in toenemende mate ook consumentenadressen (e-commerce). Deze verzending kent specifieke tijdsrestricties, zowel in het aantal transit dagen als de tijd waarvoor het moet worden afgeleverd (voor 9:00 uur, voor 12:00 uur of gehele dag). De algemene infrastructuur die gekozen wordt is er een van vele depots van waaruit busjes rondrijden om zendingen op te halen en die naar het depot te brengen. Vandaaruit worden ze met grote trucks naar een hub, zeg maar een groot sorteercentrum, gebracht waar ze 's nachts verder worden gesorteerd naar, óf de volgende hub, óf het einddepot. De trucks rijden volgens een vooraf uitgedacht schema en vertrekken op tijd, of de truck nu vol is of niet. Eenmaal aangekomen op het einddepot worden de pakjes weer met busjes vervoerd naar de afleveradressen. Ophalen en wegbrengen duiden we aan met PuD (Pickup & Delivery) en het verkeer tussen depots en hubs met Linehaul. PuD-vans en linehaul-trucks worden over het algemeen allemaal ingehuurd bij transportbedrijven die rijden onder TNT-vlag.

Eerste project: Netwerk Italië

Het idee waarmee we aankwamen bij TNT was leuk maar moest eerst in het eigen jargon vertaald worden en, TNT is commercieel, er kwam gewoon een tender waarop 5 bedrijven hebben gereageerd. Na selectie bleef de combinatie van de Universiteit van Tilburg en ORTEC over. We mochten aan de slag met het linehaul-netwerk van Italië om onze bewering waar te maken dat met Operations Research voordelen te behalen waren.

De verleiding is groot, zeker voor OR'ers, om direct in modellen te gaan denken maar we hebben dit, ook met hulp van TNT, kunnen weerstaan. We zijn begonnen met een goede data-analyse. Met behulp van AIMMS hebben we diverse bestanden van zendingen, linehaul-schema's, voertuigen en geografische gegevens aan elkaar gekoppeld. Dit bleek, na de altijd weer nodige *cleaning* van gegevens, al snel een goudmijn te zijn om allerlei snelle verbeteringen te vinden. Voorbeelden zijn voertuigen die halfvol relatief dicht achter elkaar in dezelfde richting reden, voertuigen die een te lage belading hadden maar waarvan de lading makkelijk via een andere route kon gaan. Dit lijkt triviaal maar is het zeker niet als men bedenkt dat er in dit netwerk zo'n 500 tot 600 trucks elke nacht rondrijden met dagelijks redelijk wisselende volumes. De netwerken met al hun aansluitingen die gehaald moeten worden laten zich goed vergelijken met een spoorboekje zoals dat bij de spoorwegen gehanteerd wordt.

Na de data-analyse was het veel helderder waar de grote knelpunten zaten en konden we gericht gaan optimaliseren. In dit netwerk is dit met Shortrec (routeringspakket van ORTEC) gebeurd, waar uiteraard de TNT-specifieke kenmerken moesten worden ingebracht. In de optimalisatie hebben we uiteindelijk gekozen om een lange lijst van locale verbeteringen aan het management voor te leggen en niet het schema radicaal om te gooien. Deze lange lijst van verbeteringen is door de transport-managers van Italië geheel doorgenomen op haalbaarheid. Verschillende aangedragen verbeteringen sneuvelden vanwege diverse praktijkredenen, maar uiteindelijk bleef een groter dan verwachtte besparing over: zo'n 5 procent op een netwerk van vele tientallen miljoenen kosten per jaar. Eind 2005 was de eerste mutatie in het DNA van TNT een feit.

Volgende projecten

De interesse was gewekt en volgende netwerk linehaul-projecten dienden zich aan in landen als Frankrijk, Duitsland, Slovenië en Spanje. Ook deze projecten hebben we gedaan met de ontwikkelde tools en methodologieën, maar daarnaast ontstond er een sterke behoefte aan het delen van kennis op optimalisatiegebied maar ook op implementatiegebied. Met het laatste gaat het er met name om hoe een aangetoonde besparing ook werkelijk gerealiseerd kan worden en hoe de overgang zo soepel mogelijk kan verlopen. Dit resulteerde in de oprichting van een zogenaamde COP (Community of Practice) waar diverse deelnemers uit de landen maar ook kennisorganisaties en zelfs het World Food Program bij aanschoof. Drie keer per jaar komen we voor twee dagen bij elkaar om een specifiek onderwerp van netwerk optimalisatie en alles wat daarmee samenhangt te bespreken. Belangrijk hierbij is dat een dergelijke kennisclub niet een theekransje wordt maar dat er ook echte beslissingen worden voorbereid die door het management kunnen worden bekrachtigd (of verworpen). Dit vereist een behoorlijk sterke aansturing aangezien specialisten nogal eens de neiging hebben uit te wijden en erg veel details te willen meenemen.

Belangrijk binnen elk optimalisatieproject van TNT is de driehoek: People & Organisation, Processes & Procedures en Tools & Technology. Aan alle deze drie aspecten moet ruim aandacht worden gegeven, het moet qua uitwerking goed op elkaar aansluiten en als geheel worden afgetekend door het management. Een paar voorbeelden: (optimalisatie)-tools kunnen nog zo goed zijn, als ze niet ingebed zijn in de (eventueel veranderde) processen dan werken ze niet. Als bij prachtige oplossingen de mensen niet meewerken, of de organisatie heeft andere performance indicatoren, ook dan zal er weinig effect in de praktijk zijn.

Andere gebieden

Na twee jaar veel in de netwerkhoek te hebben uitgebouwd, kwamen ook langzamerhand de vragen of ook in de andere *supply chain*-onderdelen van TNT, namelijk PuD, de luchtvrachtdivisie en de depots & hubs geoptimaliseerd kon worden. Op een zelfde wijze en met hetzelfde accent op de hierboven beschreven driehoek zijn hier projecten opgestart en COP's opgezet.

Nu wordt er met *vehicle routing*-software driftig geoptimaliseerd in het ophalen en afleveren van zendingen. De inzet van personeel in de depots en hubs wordt geoptimaliseerd. Er zijn tools die ondersteunen bij het inrichten en opzetten van nieuwe hubs en depots. Complete TNT-specifieke oplossingen zijn uitgewerkt voor de netwerken en zware MIP-modellen gebouwd voor de optimalisatie van de Express luchtvloot die ruim 40 vliegtuigen omvat.

Om een indruk te geven: vorig jaar zijn in al deze gebieden samen ruim 100 kleinere en grotere optimalisatieprojecten uitgevoerd!

Suboptimalisatie?

Van het begin af aan hebben we ons gerealiseerd dat de complete *supply chain* geoptimaliseerd moet worden. Maar als er nog veel opgebouwd moet worden is het toch verstandig om met de *supply chain* onderdelen te beginnen. *Supply chain* denken is lastig in een groot bedrijf waar één van de onderdelen al een wereld op zich is met een grote complexiteit. Het zou bijvoorbeeld kunnen betekenen dat we moeten investeren in mensen in depots en hubs (of sorteermachines) waardoor deze duurder worden terwijl we de voordelen in de netwerken en PuD behalen. Maar wat als de depot/hub-manager onder vrij zware targets staat?

Enigszins geholpen door de crisis eind 2008 zijn we de eerste complete *supply chain* modellen gaan ontwikkelen. De eerste modellen zijn simpele 'doorrekenmodellen' om de complete keten in door TNT zelf aangedragen alternatieven door te rekenen. Maar schijn bedriegt hier: het doorrekenmodel van de keten bevat als onderdelen weer zo'n vier optimalisatiemodellen en heel veel berekeningen. Een gemiddelde run laadt 60 miljoen data-elementen en heeft een uurtje rekentijd op een zware PC nodig.

Verdere kennisopbouw

Gedurende de afgelopen jaren ontdekten we dat het heel lastig is om met medewerkers in het veld te werken die onvoldoende de tools en de achterliggende ideeën begrijpen. Dit heeft geleid dat we, tezamen met TiasNimbas, de GO-academy (GO van Global Optimisation) hebben opgezet. De GO-academy kent 5 modules van elk vier dagen, welke na 2 jaar tijd wordt afgerond met een afstudeeropdracht van een vijftal maanden in de praktijk van de TNT-landen. In april 2011 hebben we de tweede lichting van veertig Supply Chain Masters, letterlijk vanuit de hele wereld, hun diploma kunnen overhandigen. De inhoud van de GO-academy wordt ook constant aangepast aan veranderingen in de strategie van Express. Op dit moment zijn we bezig met de voorbereiding van alweer het vijfde cohort. Dit betekent dat we in enkele jaren tijd honderden mensen op hetzelfde kennisniveau brengen die daarnaast ook nog eens eenzelfde 'taal' spreken.

En nu verder?

Wellicht is de indruk ontstaan dat we wel bijna klaar moeten zijn. Niets is minder waar. Er valt voor OR-ers nog heel veel te doen, vooral in het toepassen (*deployment*) van alle kennis. Maar ook infrastructuuroptimalisatie, integratie van *lean*-concepten en optimalisatie, *supply chain* denken samen met de klant en geoptimaliseerde depots & hubs staan nog maar aan het begin. Dit, ondanks dat TNT in een hoog tempo gespecialiseerde afdelingen hiervoor heeft opgezet. Het lijstje waarmee we vijf jaar geleden naar TNT gingen is allang niet leeg meer maar helaas/gelukkig zit de concurrentie ook niet stil...

HEIN FLEUREN is hoogleraar Operations Research aan de Universiteit van Tilburg en is gespecialiseerd in de praktische toepassing van Operations Research.
E-mail: <heinfleuren@orcoach.nl>

MARCO HENDRIKS is als Director Optimisation & Engineering werkzaam bij TNT Express.
E-mail: <marco.hendriks@tnt.com>

ANNELIES WOUTERSEN is als Supply Chain Consultant werkzaam bij ORTEC.
E-mail: <annelies.woutersen@ortec.com>

AGENDA

21 tot en met 26 augustus 2011

Het **58ste World Statistics Congress** van het International Statistical Institute (ISI) vindt in Dublin van 21 – 26 augustus 2011 plaats. Het wetenschappelijk programma omvat diverse, interessante onderwerpen. Op 24 augustus is er een speciale themadag 'Water, quality and quantity'; statistische kwesties met betrekking tot water en de waterkwaliteit staan dan centraal. Het tweejaarlijks congres wordt bezocht door statistici die op verschillende terreinen werkzaam zijn, zoals in de financiële sector en het bedrijfsleven, bij de overheid en wetenschappelijk onderzoeksinstituten of in het onderwijs. Het congres wil een ontmoetingsplek zijn om ervaringen en expertise te delen en om tot nieuwe inzichten te komen. Zie voor het hele programma, aanmelding en indienen van papers de website <www.isi2011.ie>.

17 tot en met 19 augustus 2011

De tweejaarlijks conferentie van de **International Association for Official Statistics (IAOS)** die van 17 – 19 augustus in Belfast wordt gehouden, kan gezien worden als een 'satellite' van het ISI-congres in Dublin. Het thema van de conferentie is 'The Demography of Ageing and Official Statistics'. Centraal staan de omvang en kenmerken van de vergrijzing en de impact op het sociale en economische beleid. IAOS wil de kennis over overheidsstatistieken bevorderen en het gebruik van effectieve en efficiënte statistische diensten wereldwijd stimuleren. De IAOS-conferentie is georganiseerd door het Northern Ireland Statistics and Research Agency (NISRA), in samenwerking met het Centre for Ageing Research and Development in Ireland (CARDI) en de Queen's University in Belfast. Zie voor meer informatie de website </www.nisra.gov.uk/IAOS2011.html>.

Karl Pearson

tussen Statistiek en Filosofie

Richard Starmans

De Britse statisticus Karl Pearson (1857-1936) neemt in de geschiedenis van de statistiek een onwrikbare plaats in. Hij geldt met Francis Galton en Ronald Fisher als de wegbereider van de moderne statistiek en was een belangrijk gangmaker van de probabilistische revolutie die zich eind negentiende eeuw in diverse disciplines voltrok. Dat gebeurde in aloude, gevestigde wetenschappen als de fysica (statistische mechanica van Boltzmann, Maxwell en Gibbs), maar ook in nieuwe, opkomende disciplines, zoals de sociale wetenschappen (Quetelet en later Durkheim), de psychologie en vooral de biologie (evolutietheorie, genetica, zoölogie), die eerst in nauwe wisselwerking met de statistiek tot bloei kwam. De biometricus Pearson en (later) de mendeliaan Fisher namen daarbij het voortouw. Ook nu is de naam van Pearson nog steeds verbonden aan vele statistische inzichten en methoden: correlatie- en regresssie-analyse, de Chi-kwadraat goodness-of-fitness toets, aan termen als standaardeviatie en kurtosis en uiteraard aan zijn fundamentele werk op het gebied van (families van) scheve verdelingen.

Filosofie en statistiek

Pearsons intellectuele erfenis overstijgt de statistiek in engere zin. Hij was evenzeer een veelzijdig en vernieuwend filosoof, die blijk gaf van een diepgaand inzicht in de wijsgerige kernproblemen die het tijdsgewricht waarin hij leefde bepaalden. Hij hield zich actief bezig met de grondslagen van wetenschapsgebieden als wiskunde, natuurkunde en evolutietheorie, maar bewoog zich eveneens moeiteloos op het terrein van politieke filosofie en letterkunde. Vooral zijn lijvige *The Grammar of Science* (GoS), dat hij op 35-jarige leeftijd in 1892 publiceerde, biedt een rijk palet aan wetenschapsfilosofische thema's; het inductieprobleem en causaliteit, de relativiteit van ruimte en tijd, de ontologische status van de werkelijkheid, evolutietheorie en ethiek. Het werk beleefde vele herdrukken en kende zowel befaamde bewonderaars (Einstein) als tegenstanders (de filosoof C.S. Peirce).

Filosofie en statistiek vormen bij Pearson bovendien twee zijden van de zelfde medaille. Kennis van de aard van deze verbondenheid is dan ook onontbeerlijk voor een juiste waardering van zijn nalatenschap. Vooral GoS geldt als de sleutel tot zijn statistische werk. Kennistheoretisch betoont Pearson zich in GoS een uitgesproken empirist. Hij schaart zich daarmee in een aloude Britse, anti-metafysische traditie (Bacon, Locke, Hume), die ook in de negentiende eeuw vele aanhangers kende en waarvan John Stuart Mill en Herbert Spencer de invloedrijkste waren.

Anti-causalist

Maar Pearson ging verder dan zijn illustere landgenoten en bekende zich tot de fenomenalistische filosofie van zijn vroegere leermeester Ernst Mach, volgens welke doctrine de werkelijkheid bestaat uit elementen (kleuren, trillingen, tijden), die zich manifesteren in een stroom van *sense data* of directe gewaarwordingen. Deze zijn neutraal (noch geestelijk, noch materieel) en vanuit deze gewaarwordingen kunnen zowel materiële objecten als bewustzijnsinhouden worden geconstrueerd. De wetenschap moet geen verklaringen nastreven of causale mechanismen postuleren, maar heeft primair een denkeconomische functie; de stroom van primaire sensaties moet zo goed mogelijk in kaart worden gebracht en samengevat in een overzichtelijke taal. Empirische generalisaties, zoals Boyle's gaswet, waren nog wel toegestaan, de in die tijd net opgekomen kinetische gastheorie uiteraard niet. Ook Pearson stond huiverachtig tegenover het toekennen van een ontologische status aan begrippen die niet direct gekoppeld waren aan zintuiglijke waarnemingen, zoals niet-verifieerbare theoretische constructen en hij verwierp het zoeken naar een diepere of hogere werkelijkheid die schuilgaat achter de verschijnselen.

Na of naast Bertrand Russell werd hij de belangrijkste anti-causalist van zijn tijd; ook zijn bijdragen aan de correlatieanalyse moeten in dit licht worden bezien. Daarnaast legt Pearson in GoS een sterk positivistische, om niet te zeggen sciëntistische visie aan de dag. Hij stelt een bijkans onbegrensd vertrouwen in de moderne wetenschap als regulerend mechanisme voor de samenleving, dat van mensen volwaardige burgers moest maken en de staat op wetenschappelijke leest zou schoeien. De verwantschap met de grondlegger van het positivisme, August Comte is in deze onmiskenbaar. Pearson combineert dit alles tot slot met een idealistische, dat wil zeggen antimaterialistische filosofie over de aard van de (kenbare) werkelijkheid. Deze bestaat niet onafhankelijk van de menselijke geest en Pearson beschouwt wetenschap vooral als een classificatie en analyse van de inhoud van de geest. Haar domein wordt daarmee veeleer het bewustzijn dan een onafhankelijk hiervan gegeven externe wereld.

Het belang van de Pearsoniaanse revolutie voor de ideeëngeschiedenis ligt deels in de wijze waarop hij deze wijsgerige inzichten uitwerkt in een constructieve statistische methodologie. Minstens zo belangrijk is daarbij dat zijn werk het sluitstuk vormt van een nieuw wereldbeeld, dat tot stand kwam aan het eind van de negentiende eeuw. Filosofisch gezien tekende zich in die eeuw een historisering van de werkelijkheid af, kwam tweeduizend jaar substantiedenken onder druk te staan en vond volgens de filosoof Ian Hacking een ‘erosie van het determinisme’ plaats.

Daaraan vooraf ging echter een ruim tweeduizend jaar durend emancipatieproces van de verwante en weerbarstige begrippen *variatie* en *verandering*, waarmee filosofen, wetenschappers en wiskundigen van oudsher hebben geworsteld. Dat begon al bij de Griekse natuurfilosofen, die een verklaring zochten voor de verschillende aspecten van verandering: verandering van plaats (beweging), ontstaan en vergaan, verandering van hoedanigheid en hoeveelheid. Het bestaan ervan werd dikwijls ontkend, onmogelijk geacht of gereduceerd tot non-verandering. Variatie werd beschouwd als afwijking van een regel of norm, die in het gunstigste geval moest worden verklaard. De noties stonden voor onvolmaaktheid en onvoorspelbaarheid en werden soms zelf incoherent geacht op logische en metafysische gronden (Parmenides). Met name de grilligheid van de (levende) natuur en de veelvuldigheid van haar verschijningsvormen vormden een obstakel voor een deterministisch wereldbeeld en stonden een wiskundige benadering van de verschijnselen in de weg. Zelfs pioniers van de statistiek als Laplace (‘errorfuncties’) en Quetelet (‘l’homme moyen’) konden er slechts enige greep op krijgen door de gegevens in het keurslijf van de normale verdeling te dwingen.

Inferentiële statistiek

Mede door de evolutietheorie van Wallace en Darwin kreeg de opvatting dat variatie inherent is aan de natuur en de aarde een lange geschiedenis van verandering kent, vaste grond onder de voeten. Galton, onvermoeibaar pleitbezorger voor mathematisering van de wetenschap, overtuigde zijn vriend Pearson dat variatie zich wel degelijk voor een wiskundige behandeling leende, zonder pejoratieve duidingen van afwijkingen, errorfuncties, etcetera. Pearson deed daaraan recht door de variatie niet te identificeren in ‘errors’, maar in de verschijnselen zelf (gecodeerd in data), en terug te voeren tot verschillende (klassen van) kansverdelingen.

Een cruciale stap zette hij met zijn vroege werk over scheve verdelingen. Op basis van grote aantallen door Galton verzamelde (biometrische) data zag hij in dat vele verschijnselen niet normaal, doch scheef waren verdeeld en met behulp van vier parameters (gemiddelde, standaardafwijking, scheefheid en welving) konden worden beschreven en geclassificeerd. Volgens Pearson kon zelfs de eenheid van de wetenschappen worden gevonden in deze constructieve, zij het zeer arbeidsintensieve methode. De variabiliteit in de natuur manifesteerde zich in een puntenwolk van metingen en Pearson zocht naar het ‘best fitting’ model, de functie die het beste paste bij de data; veeleer een zuinige beschrijving, dan een causaal, onderliggend mechanisme waardoor deze data waren gegenereerd. Als eerste gaf hij daarmee kansverdelingen een volwaardige plaats in de wetenschap en opende hij de deur naar inferentiële statistiek.

Pearson zag de wereld op een niveau van abstractie waarbij data, variatie in data, datagenererende mechanismen en parameters van de verdelingen de werkelijkheid veeleer coderen en opbouwen en niet zozeer een (vermeende) fysische werkelijkheid representeren of afbeelden. Gesteund door zijn macheaanse en idealistische filosofie interpreteerde hij zijn kansverdelingen op een welhaast Pythagoreïsche manier, al werd Pearsons universum niet gerealiseerd door getallen, getalsverhoudingen en getallentypen, maar door kansverdelingen met bijbehorende parameters en vormden de puntenwolken de objecten van de wetenschap. Het wereldbeeld verloor weliswaar veel van zijn directe aanschouwelijkheid, maar paradoxaal genoeg was de werkelijkheid als statistische verdeling in macheaanse zin 'waarneembaar', dichtbij de gegevens, kenbaar; een beschrijving van de feitelijke data, een grote, maar eindige deelverzameling van de verzameling van alle mogelijke metingen, alleen beschikbaar in een 'ideale' situatie. Zolang deze deelverzameling groot genoeg was zouden de berekende 'parameters' dezelfde zijn als die van de gehele verzameling.

Voor zover Pearson een concept van statistische inferentie kende, was dit vooral gebaseerd op het idee van de *goodness-of-fit*. Van daadwerkelijk *significance testing* of *maximum likelihood estimation*, of zelfs van een volwassen statistisch modelbegrip is dan uiteraard nog geen sprake. Daartoe dienden andere wegen te worden ingeslagen. Dat deed bij voorbeeld Pearsons jongere tegenstrever Fisher, die niet werd beperkt door macheaanse doctrines en dieper in het Pythagoreïsche universum schouwde. Voor hem was de werkelijkheid een abstracte wiskundige verdeling, niet waarneembaar of samenvallend met de gegevens, welke slechts een (aselecte) steekproef vormden, met een eigen verdeling. Waarheid was een vaste, maar onbekende, populatiegrootheid en kon slechts worden geschat.

Eerste filosoof van de statistiek

Het is opmerkelijk dat Pearsons werk in de moderne epistemologie allerminst tot de canon behoort. Allereerst was Pearson een wegbereider van het logisch positivisme dat in de jaren twintig van de vorige eeuw in Wenen en Berlijn opkwam. Bovendien domineert in de contemporaine wetenschapsfilosofie de 'probabilistische invalshoek'. Waarschijnlijkheids-modellen spelen een cruciale rol in moderne benaderingen van causaliteit, in het Wetenschappelijk Realisme Debat, en uiteraard in de (Bayesiaanse) confirmatietheorie, die al decennia lang een toonaangevend paradigma vormt binnen de kennisleer. Klassieke filosofische vragen worden in probabilistische termen geduid. Kernvragen als: Wat is de werkelijkheid? Bestaat deze onafhankelijk van de geest? Hebben we er toegang toe? Zo, ja, hoe? Kunnen we er ware uitspraken over doen? Zo ja, wat is waarheid en hoe is deze verbonden met de werkelijkheid?

Pearson was de eerste statisticus die antwoorden voorstelde op al deze vragen, die hij combineerde met een constructieve statistische methodologie. Juist omdat hij zich bewoog op het snijvlak van statistiek, filosofie en methodologie werd hij feitelijk de eerste filosoof van de statistiek. Immers, de filosoof van de statistiek zal nieuwe statistische theorieën en paradigma's altijd trachten te verankeren in deze kernvragen en ze (mede) beschouwen en beoordelen als pogingen om deze te beantwoorden. Pearson stond niet alleen. Wellicht minder expliciet, maar niet minder doorwrocht gaf de generatie na hem (Fisher, Jerzy Neyman en Egon Pearson, Bayesianen als Savage en De Finetti) zich rekenschap van deze kwesties. De 'Science of Data' begon met de vraag naar de aard en status van data en vooral naar de relatie met de werkelijkheid, waaraan zij toch op de een of andere manier ontleend zijn, van waaruit zij zijn gegenereerd, die zij moeten representeren, benaderen, coderen, simuleren, vervangen of wellicht zelfs overbodig maken; zulks afhankelijk van de ingenomen filosofische positie.

Richard Starmans is verbonden aan de Faculteit Bètawetenschappen (Department of Information and Computing Sciences) van de Universiteit Utrecht. Hij doet onderzoek op het snijvlak van filosofie, statistiek en informatica.
E-mail: <starmans@cs.uu.nl>

RECHTSTREEKS VERWACHTINGEN EVALUEREN OF DE NUL HYPOTHESE TOETSEN?

Rens van de Schoot

In de wetenschappelijke literatuur wordt door vrijwel alle onderzoekers klassieke nul hypothese toetsing (NHT) gebruikt om antwoord te geven op de onderzoeksvraag. NHT hoeft echter niet per se de beste keuze te zijn om de onderzoeksvraag te beantwoorden. In dit artikel laat ik zien waarom en introduceer ik vervolgens een recent ontwikkelde methode die een veelbelovend alternatief biedt, namelijk Bayesiaanse Model Selectie (BMS).

Informatieve hypothesen

Onderzoekers hebben vaak bepaalde verwachtingen over hoe de werkelijkheid er uit ziet. Verwachtingen en hypothesen kunnen gebaseerd zijn op eerder (literatuur-) onderzoek, of wetenschappelijk debat. Het doel van veel onderzoekers is het evalueren van deze verwachtingen om te bepalen welke de beste is. Met andere woorden, welke verwachting de meeste steun krijgt van de verzamelde data. Verwachtingen zijn geformuleerd in termen van wat ik *informatieve* hypothesen noem. Dit omdat er *a priori*, dat is voordat er data zijn verzameld, *informatie* bestaat over bijvoorbeeld de ordening tussen twee (of meer) groepsgemiddelden gebaseerd op een eerder gepubliceerd artikel: $H_A : \mu_1 < \mu_2 < \mu_3 < \mu_4$, waarbij het teken '<' aangeeft dat het eerste gemiddelde (μ_1) lager is dan het tweede gemiddelde (μ_2), enz. Een andere theorie veronderstelt echter een andere ordening, bijvoorbeeld $H_B : \mu_1 < \mu_2 < \mu_3 = \mu_4$. De vraag is nu welke van deze twee informatieve hypothesen de beste is.

Onderzoekers *willen* deze verwachtingen wel evalueren, maar ze *kunnen* dit niet zo maar doen. Het is namelijk vrijwel onmogelijk om met NHT complexe informatieve hypothesen te evalueren. Als onderzoekers dit toch proberen omdat er geen alternatieven voor handen zijn, dan ontstaan er enkele problemen.

Wat gaat er mis?

Er is door de tijd heen veel literatuur verschenen met kritiek op het gebruik van NHT en het gebruik van p-waarden. Ik zal me nu voornamelijk richten op waar het mis gaat bij het evalueren van informatieve hypothesen door NHT.

Met NHT moet vaak een hele stapel output geëvalueerd worden om een onderzoeksvraag te beantwoorden: F-toets, post-hoc toetsen, groepsgemiddelden, etc. Deze stapel output kan makkelijk leiden tot verwarring en het geeft slechts indirect steun voor de a priori verwachtingen. Bij NHT is de hypothese die daadwerkelijk getoetst wordt immers de bekende nul hypothese *er is niks aan de hand* versus het alternatief *er gebeurt iets, maar we weten niet wat*. Merk op dat deze nul (H_0) en alternatieve hypothese (H_1) niet hetzelfde zijn als de informatieve hypothesen H_A, en H_B die de onderzoekers eigenlijk wilden evalueren. Als de nul hypothese en de alternatieve hypothese geen onderdeel zijn van de onderzoeksvraag, geven de NHT resultaten in dat geval geen direct antwoord op de onderzoeksvraag.

Als dan toch H_0 wordt getoetst, dan wordt de traditionele p-waarde gebruikt om deze te verwerpen of niet te verwerpen. Het omslagpunt van deze beslissing ligt meestal bij de welbekende waarde van $p < ,05$. Deze drempelwaarde van ,05 is niet alleen willekeurig gekozen, maar laat alleen ruimte voor de conclusie dat een nul hypothese wel of niet wordt verworpen met niks daar tussenin. Dit kan leiden tot vreemde beslissingen, bijvoorbeeld in het geval dat $p = ,051$ of $p = ,049$. Het mag duidelijk zijn dat beide situaties niet veel van elkaar verschillen. En toch wordt H_0 in het eerste geval niet verworpen en in het tweede geval wel.

Wanneer H_0 wordt verworpen, weten we eigenlijk nog steeds niks over de informatieve hypothesen H_A en H_B, aangezien de alternatieve hypothese geen informatie bevatte over de ordening tussen de gemiddelden. Hoe kan een onderzoeker dan toch uitspraken doen over de informatieve hypothesen?

Kan het ook anders?

Om erachter te komen welke van een set van informatieve hypothesen het meest waarschijnlijk is, gaan we deze *direct* tegen elkaar afzetten met behulp van Bayes Factors (BFs). BFs zijn een model selectie maat en houden niet alleen rekening met hoe goed de hypothese bij de data past, maar ook hoe complex de hypothese is. Dit resulteert in een directe vergelijking tussen de onderzochte hypotheses. Een BF van 10 betekent dat H_A tien keer zo veel steun van de data krijgt dan H_B en een BF van 1 betekent dat beide hypotheses even veel steun van de data krijgen.

Er is diverse software voor het uitvoeren van (M)AN(C)OVA's, multipele multivariate regressie analyse, kruistabel analyse, en latente klasse analyse met BMS (zie www.fss.uu.nl/ms/informativehypothesis voor de gratis te downloaden software). De gebruiker hoeft alleen de hypothesen te specificeren in termen van restricties tussen de statistische parameters, zoals $\mu_1 < \mu_2$, en de dataset aan te leveren, en de software levert de BFs automatisch uit. Laten we nu met een inhoudelijk voorbeeld bekijken hoe BMS werkt.

Voorbeeld

Merkwaardig genoeg is het rendement van door NWO toegekende promotiesubsidies in Nederland tot nu toe niet onderzocht. In een recent gepubliceerde artikel onderzochten Van de Schoot, Sonneveld en Lockhorst (2011) het al dan niet succesvolle verloop van de promotietrajecten waarvoor in de jaren 1999 tot en met 2001 bij Maatschappij- en Gedragswetenschappen een subsidieaanvraag werd ingediend die werd gehonoreerd. Succesvol staat hier voor: de besteding van de subsidie werd bekroond met de verdediging van het proefschrift voor september 2010. Er werd een onderscheid gemaakt tussen projecten waarin de promovendi in een kleiner programmatisch verband werkten (met collegae al dan niet op dezelfde werkplek) en individuele projecten die men als *stand alone* zou kunnen beschouwen.

	n	gepromoveerd	niet-gepromoveerd	rendement
Totaal programma-projecten	112	87	25	77,7%
Individueel project	107	96	11	89,7%

Tabel 1. Subsidierendement uitgesplitst naar ‘individueel project’/‘project deel uitmakend van het programma’

Er waren twee hypothesen geformuleerd:

> H_1: er zijn procentueel evenveel succesvol afgeronde projecten bij individuele subsidies als bij programmatische subsidies.

Versus

> H_2: er zijn relatief meer succesvol afgeronde projecten bij individuele subsidies dan bij programmatische subsidies.

Merk op dat de onderzoekers in dit geval wel degelijk waren geïnteresseerd in de nul hypothese, met als kanttekening dat BMS veronderstelt dat het verschil niet precies nul is, maar ongeveer nul.

Uit de databestanden van NWO is een lijst samengesteld van alle promotieprojecten waarvoor in de jaren 1999 tot en met 2001 een subsidieaanvraag werd gehonoreerd. Zie tabel 1 voor de percentages geslaagd dan wel gefaald op de peildatum. Er bestaat een ondervertegenwoordiging van programmatische subsidies bij het totaal aantal geslaagde projecten. Maar is dit verschil ook significant en/of relevant?

De klassieke methode om deze vraag te beantwoorden is om een chi-kwadraat toets uit te voeren. Dan blijkt dat er een significant verschil bestaat tussen geobserveerde en verwachte waardes, $X^2_{1,\,219} = 8{,}31$; $p = {,}004$. Aangezien het hieruit nog niet duidelijk wordt waar precies de verschillen zitten, hebben we ook BMS gebruikt om de hypotheses direct af te zetten tegen elkaar. Uit deze analyse blijkt dat de tweede hypothese 28 keer zoveel steun van de data krijgt als de eerste hypothese. Oftewel, er is veel steun in de data voor de hypothese dat er relatief meer succesvol afgeronde projecten zijn bij individuele subsidies dan bij programmatische subsidies.

Conclusie

De resultaten van nul hypothese toetsting geven vaak geen direct antwoord op de onderzoeksvraag. Informatieve hypothesen kunnen niet direct met elkaar vergeleken worden, iets dat met Bayesiaanse Model Selectie wel kan. Ook wanneer de ordening van de groepsgemiddelden niet geheel overeen komt met één van de hypothesen, geeft BMS een interpreteerbaar resultaat. Zelfs bij veel complexere onderzoeksvragen dan in dit artikel besproken, bijvoorbeeld met meerdere (on)afhankelijke variabelen, meerdere meetmomenten over de tijd heen, covariaten, meer groepen, enz., geeft BMS nog steeds een enkel interpreteerbaar getal per hypothese. BMS resulteert dus in makkelijker te interpreteren resultaten dan NHT en geeft bovendien een direct antwoord op de onderzoeksvraag. BMS is daardoor een veelbelovend alternatief voor NHT en zal steeds vaker opduiken in de wetenschappelijk literatuur.

Het artikel is grotendeels gebaseerd op het proefschrift van de auteur en de genoemde literatuur.

LITERATUUR

Hoijtink, H., Klugkist, I., & Boelen, P. A. (Eds.). (2008). *Bayesian evaluation of informative hypotheses in psychology*. New-York: Springer.

Van de Schoot, R., Hoijtink, H. & Doosje, S. (2009). Rechtstreeks Verwachtingen Evalueren of de Nul Hypothese Toetsen? Nul Hypothese Toetsing versus Bayesiaanse Model Selectie. *De Psycholoog 4,* 196-203.

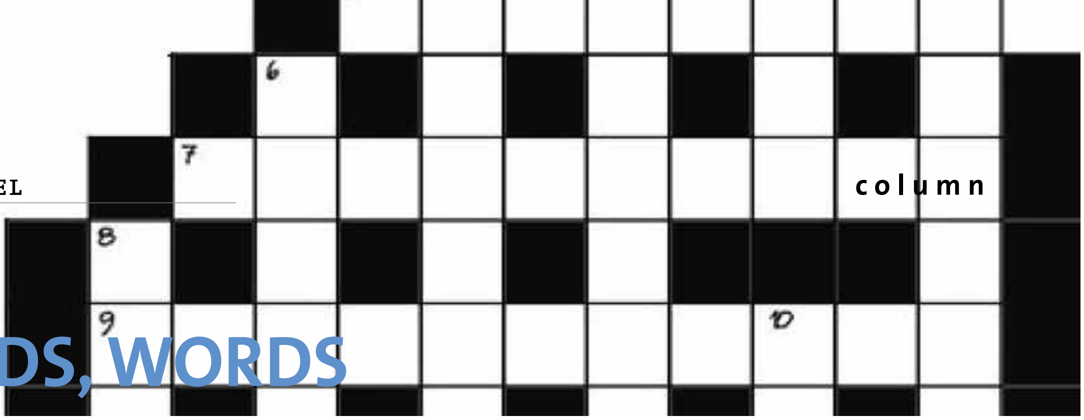
Van de Schoot, R., Hoijtink, H., Mulder, J., Van Aken, M. A. G., Orobio de Castro, B., Meeus, W. & Romeijn, J.-W. (2011). Evaluating Expectations about Negative Emotional States of Aggressive Boys using Bayesian Model Selection. *Developmental Psychology, 47,* 203-212

Van de Schoot, R., Hoijtink, H., & Romeijn, J-W (2011). Moving beyond traditional null hypothesis testing: Evaluating expectations directly. *Frontiers in Quantitative Psychology and Measurement, 2:24. doi: 10.3389/fpsyg.2011.00024*

Van de Schoot. R., Sonneveld, H., & Lockhorst, D. (2011). Het lot van promotieprojecten: Rendement van magw/nwo-subsidies. *De Psycholoog*, 4, 10-18.

RENS VAN DE SCHOOT is als universitair docent verbonden aan de Faculteit Sociale Wetenschappen (afdeling Methoden en Statistiek) van de Universiteit Utrecht
E-mail: <a.g.j.vandeschoot@uu.nl>

FRED STEUTEL



WORDS, WORDS

De combinatie statistiek en taal heeft een lange traditie. Het begon misschien met Markov, die het – naderhand – naar hem genoemde begrip Markov-keten ontwikkelde aan de hand van de opeenvolging van letters in een bekend Russisch gedicht. Wat later werd door statistici de frequentie van letters in teksten geturfd en geanalyseerd. De moderne techniek biedt nieuwe mogelijkheden.

Mijn vrouw en ik lossen iedere week het Scryptogram in *NRC Handelsblad* op. We gebruiken daarbij soms een woordenboek; om te kijken of een bedacht woord bestaat of om te zoeken naar woorden die, bij voorbeeld, met ‘boer’ beginnen.

In uiterste nood gebruiken wij een modern wapen: de Dikke Van Dale op dvd. Die heeft een optie ’zoeken op woord(vormen)’. Daar kun je een woord invullen, waarbij de nog ontbrekende letters door vraagtekens zijn vervangen. Voorbeeld: ‘?st???s?’. In dit geval is er maar één oplossing: ‘asterisk’. Als je al flink wat letters hebt, is het aantal mogelijkheden niet meer dan enkele tientallen; als je weinig letters hebt, zijn er soms honderden mogelijkheden. Een enkele keer, als je een fout gemaakt hebt of als de puzzelmaker een woord gebruikt dat Van Dale niet kent, krijg je de mededeling ‘Er zijn geen artikelen met de door u gezochte vorm’.

Als je nog veel vraagtekens hebt, zijn er heel veel manieren om een woord te maken. Dat opent de mogelijkheid – wiskundigen zoeken graag naar extreme situaties – om het aantal woorden te vinden met een gegeven aantal letters. Dat gaat als volgt. Je tikt zeven vraagtekens, drukt op ‘Enter’ en vindt 17.318 woorden van zeven letters, van ‘aaibaar’ tot en met ‘zygenen’. Dat laatste woord betekent ‘grote familie van vlinders, waartoe o.a. de sint-jansvlinder behoort (Zygaenidae)’.

Zo kunnen we de woorden van alle mogelijke lengten afwerken. Alleen, bij woorden van drie of minder letters bestaan veel van de resultaten uit afkortingen, en bij woorden met heel veel letters zijn de resultaten vaak samenstellingen van meer dan één woord. Tussen deze extremen in is het interessant. Het langste echte woord dat ik vond heeft 31 letters (hierbij geldt de ij als één letter): medeverantwoordelijkheidsheffing.

Eén van de vragen die je kunt stellen is: ‘Van welke lengte zijn er de meeste woorden?’ Van Dale geeft bij iedere lengte braaf het aantal woorden, en het blijkt dat woorden van tien letters het meest voorkomen; de lengten 9, 10 en 11 leveren achtereenvolgens 32789, 33549 en 29398 woorden. Zit hier nog een soort model achter? Je kunt je voorstellen dat het aantal letters dat een woord telt ongeveer een Poisson-verdeling heeft met een gemiddelde van 10. Omdat het totaal aantal woorden in de Dikke Van Dale ongeveer 240.000 is, komen de fracties woorden met respectievelijk 9, 10 en 11 letters uit op 0,1366; 0,1398 en 0,1225. Dat komt goed overeen met de kansen op 8, 9 en 10 in een Poisson-verdeling met gemiddelde 9,5. Het aantal letters in een woord is natuurlijk minstens een, en dat aantal komt dan goed overeen met ‘één plus een Poisson-verdeelde grootheid met gemiddelde 9,5’.

Maar, de taal laat zich gelukkig slechts zeer ten dele in getallen vangen.

FRED STEUTEL is emeritus hoogleraar kansrekening aan de TU Eindhoven. E-mail: <f.w.steutel@tue.nl>

Foto: BoH, Kenterschalm

Generalized reliability in industriële user studies

JAN ENGEL

> *Love is the total absence of fear. Love asks no questions. Its natural state is one of extension and expansion, not comparison and measurement. – Gerald Jampolsky*

Echter, in alle andere gevallen dan liefde, zullen we toch wel meten. En dat niet alleen: het trekken van conclusies omtrent de werkelijkheid gaat het beste als we ook precies genoeg kunnen meten. Maar wat is precies genoeg? In dit artikel gaan we er van uit dat we in een industriële user studie een vergelijking willen maken tussen verschillende condities, bijvoorbeeld tussen verschillende producten. Dit wordt gedaan door participanten in een meetpanel die hun beoordelingen vastleggen in een of meer meetschalen.

Precies genoeg is dan in dit geval: kunnen de participanten voldoende goed onderscheid maken tussen de condities? Doen ze dat consistent?

Dit artikel heeft de volgende opbouw. Na een introductie van het begrip ‘reliability’ uit de psychometrie zullen we dit generaliseren naar een versie die bruikbaar is voor het bepalen van meetprecisie in user studies, een *generalized reliability* die we G-reliability zullen noemen. Deze geeft antwoord op de vraag: hoe precies kan het panel, maar ook de individuele participant daarin, de verschillen tussen condities vaststellen? Vervolgens laten we aan de hand van een voorbeeld zien dat G-reliability een handige maat is voor de karakterisering van een meetpanel. G-reliability hangt nauw samen met het onderscheidingsvermogen van de toets op conditieverschillen en dit zullen we illustreren. Een discussie besluit dit artikel.

Meetprecisie van items

Bij het kwantificeren in empirisch onderzoek wordt veelal de vraag gesteld naar meetprecisie. In de psychometrie, met name in de klassieke test theorie (CTT), hanteert men daarvoor het begrip ‘reliability’. Hierbij wordt een eigenschap van de participanten in de studie gemeten op een eendimensionale meetschaal met K items. Daarbij wordt het volgende meetmodel verondersteld (Spector, 1992):

$$Y = P + E$$

Hierin is P de echte waarde van een participant met variantie σ_P^2, Y de gemeten versie ervan terwijl E het verschil aangeeft, de error met variantie σ_e^2.

Hoewel onderwerp van veel discussie, zie bijvoorbeeld Clarke and Watson (1995), is Cronbach α (Cronbach, 1947) een standaardmaat voor reliability. Onder bepaalde veronderstellingen in het meetmodel kan worden aangetoond dat Cronbach α een schatter is van de reliability

$$\frac{\sigma_P^2}{\sigma_P^2 + \sigma_e^2 / K}$$

De reliability is een relatieve maat voor meetprecisie. Deze beantwoordt de vraag: hoe goed kunnen we de verschillen in P-waarden van participanten, gemeten door σ_P^2, onderscheiden op grond van de gemiddelden van de meetwaarden Y aan de K items, met variantie $\sigma_P^2 + \sigma_e^2 / K$. Het bepalen van reliability gebeurt overigens veelal pas in de (observatie) studie zelf. In industriële user studies is het veelal mogelijk, alvorens een experiment uit te voeren, een beeld te krijgen van de precisie waarmee gegevens worden verkregen. We zullen nu reliability generaliseren naar de industriële setting.

Generalisatie van reliability

Bij het bepalen van reliability in CTT vraagt men zich in feite af: hoe goed kan ik verschillen onderscheiden tussen participanten, gegeven de meetvariatie van items. Die participanten zijn hier de meettarget. In een industriële user studie zijn participanten geen meettarget, maar onderdeel van de meetmethode. De meettarget wordt gevormd door de condities, de producten, die worden aangeboden aan de participanten. Het begrip reliability is dan nog steeds heel bruikbaar, maar we zullen het herformuleren. De vraag is dan: hoe precies kunnen we verschillen vaststellen tussen de meetcondities gegeven de variatie van items en van participanten? Jammer genoeg ziet men dat ook in dit geval nog steeds Cronbach α wordt berekend, maar dan nu over de hele dataset van participanten en condities. Deze geeft helaas geen duidelijk, en een soms zelfs misleidend, beeld van de reliability waar het echt om gaat: het meten van verschillen tussen condities. De verschillen tussen participanten zijn helemaal niet interessant. Dat dit mis kan lopen zullen we laten zien in een voorbeeld in de volgende paragraaf.

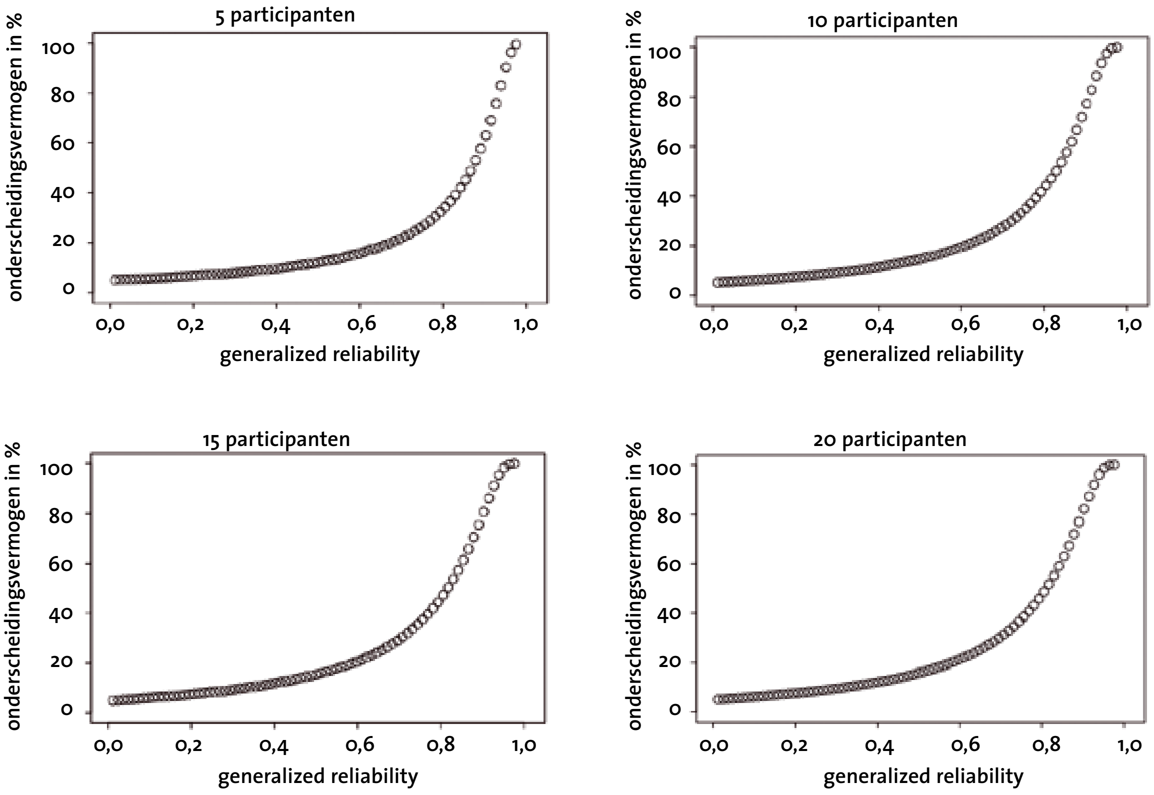
Om tot een generalized reliability te komen dienen we eerst een redelijk statistisch model te formuleren voor de gegevens van een user studie. We gaan nu uit van het geval dat we één factorcondities testen op elke participant in een Within Subject Design (WSD), een veel voorkomend type user studie. Dan zou een redelijk model als volgt kunnen zijn:

$$Y_{ijk} = \mu + P_i + C_j + PC_{ij} + \beta_k + e_{ijk}$$

In dit model vinden we de effecten terug van participanten P_i, i = 1,...,I; condities C_j, j = 1,...,J; de interactie van participanten en condities PC_{ij}, de test items β_k, k = 1,...,K en de residuele fout e_{ijk}. We veronderstellen voorts dat de P_i, PC_{ij} en e_{ijk} stochastisch zijn en onderling onafhankelijk, met varianties σ_P^2, σ_{PC}^2 en σ_e^2. Voorts hanteren we

$$\sigma_C^2 = \Sigma(C_j - C)^2 / (J-1)$$

als maat voor de verschillen tussen condities. We gaan nu de reliability op twee manieren generaliseren.



Figuur 1. Onderscheidingsvermogen van de F-toets op condities tegen generalized reliability, voor vier panelgrootten. Resultaten zijn voor twee condities (J = 2)

Geval 1: de G-reliability per participant,

$$\frac{\sigma_C^2}{\sigma_C^2 + \sigma_e^2 / K}$$

Deze vorm kennen we al uit de vorige paragraaf. Het enige verschil is dat we participant vervangen hebben door condities. Hiermee karakteriseren we de meetprecisie van een participant in het panel. Hoe goed kan deze condities onderscheiden?

Geval 2: de G-reliability over het meetpanel van participanten,

$$\frac{\sigma_C^2}{\sigma_C^2 + \sigma_{PC}^2 /I + \sigma_e^2/(I^*K)}$$

Deze expressie geeft het volgende weer: hoe goed onderscheiden we de verschillen tussen condities zoals gemeten door σ_C^2 door het gemiddelde per conditie van de data Y over items én participanten. Dit gemiddelde heeft variatie $\sigma_C^2 + \sigma_{PC}^2 / I + \sigma_e^2 / (I^*K)$. De aanpak is analoog aan het eerste geval, het resultaat is wat anders.

Aardig is nu dat een eenvoudige momentenschatter van G-reliability wordt gegeven door de grootheid (F-1)/F waarbij F de waarde is van de F-toets van Anova voor het toetsen van condities. In het eerste geval toetsen we het effect van condities voor een bepaalde participant. In het tweede geval doen we dat voor alle participanten samen. De F-toets wordt gebruikt als schatter voor G-reliability. De resultaten volgen uit de verwachtingswaarden van de Mean Squares waaruit de F-toets is opgebouwd, met andere woorden uit de EMS tabel van de variantieanalyse.

Er is een verband met een eerder, en tamelijk onbekend resultaat. Hoyt (1941) liet al zien dat

Cronbach α ook kan worden bepaald door (F-1)/F voor het geval dat in CTT wordt bestudeerd, zie de vorige paragraaf. Maar het geldt veel algemener! We kunnen dan eenvoudig G-reliability bepalen via de Anova F-toets, en standaard software is geschikt om dit uit te voeren.

Onderscheidingsvermogen van de F-toets

G-reliability in geval 2 heeft een relatie met het onderscheidingsvermogen van de F-toets op condities. We kunnen dit als volgt inzien. We formuleren de nulhypothese H_0: $C_1 =\dots = C_J$, en het alternatief H_1: niet alle C_j zijn gelijk. Indien H_0 waar is heeft de F-toets een centrale F-verdeling, als H_0 onwaar is heeft de F-toets een niet-centrale F-verdeling, met niet-centraliteits (nc) parameter die we λ noemen. Maar die parameter λ kunnen we uitdrukken in de G-reliability , en is daarvan een monotoon stijgende functie:

$$\lambda = (J\text{-}1)\ (G\text{-}reliability\ /(1\text{-}\ G\text{-}reliability))$$

Hoe groter G-reliability, hoe groter het onderscheidingsvermogen; zie figuur 1. G-reliability heeft direct een betekenis voor de kwaliteit van toetsen.

Uit figuur 1 trekken we twee conclusies:

1. gegeven een waarde van de generalized reliability is het aantal participanten niet erg bepalend voor het onderscheidingsvermogen van de F-toets, en
2. voor een onderscheidingsvermogen van 80% is toch al snel een, relatief hoge, generalized reliability nodig van 0,90. Aan de Cronbach α worden vaak lagere eisen gesteld.

Voorbeeld: meetprecisie van een meetpanel

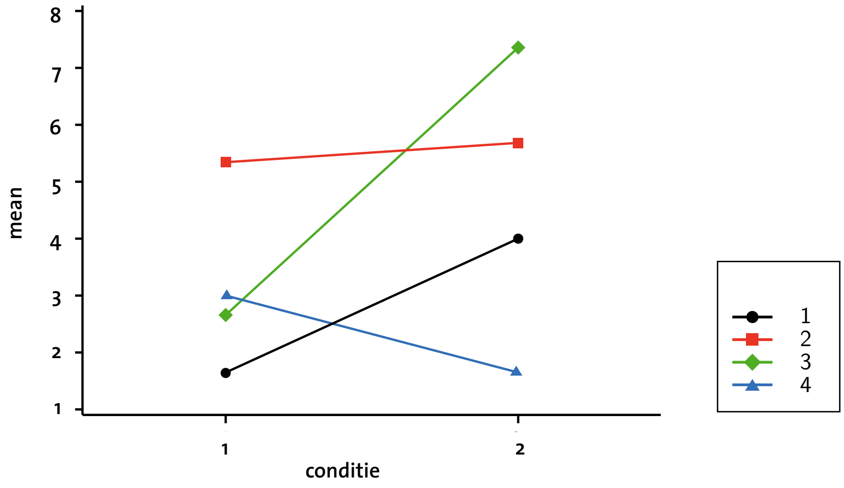
Het voorgaande geeft methoden om de reliability van een meetpanel na te gaan. In het volgende, eenvoudige, voorbeeld heeft elk van vier participanten, onder twee verschillende condities, een oordeel gegeven op een 9-punts meetschaal met drie items. Zie tabel 1.

Na berekening blijkt Cronbach α = 0,924. Er lijkt dus, op grond van Cronbach α, niets aan de hand, en onze meetschaal lijkt het doet te doen. Maar we meten de verkeerde reliability: die van de totale variatie, dus van participanten, condities, de interactie, en niet van condities alleen! Voor de F-toets op condities vinden we F = 1,343, en daarmee de G-reliability (F-1)/F = 0,26, en dat is direct een stuk minder optimistisch. Kijken we naar de individuele G-reliability waarden, gebaseerd op de F-toets uitkomsten per participant, dan vinden we 0,98, -12,00, 0,96 en 0,75. Daaruit blijkt dat de tweede participant het wel heel slecht doet. Laten we deze weg, dan vinden we echter een G-reliability van 0,14.

Een en nader wordt wellicht duidelijk aan de hand van de interactieplot van figuur 2. We zien hierin dat participant 2 inderdaad slecht discrimineert: een non-discriminator. Maar er is meer: de interactie is groot en daarmee de meetprecisie van het panel gering. Nog sterker, door die grote interactie is Cronbach α groot, maar voor het echte doel, het bepalen van verschil tussen condities, is die interactie juist funest! De waarde

Participant	Conditie	Item 1	Item 2	Item 3
1	1	2,00	1,00	2,00
1	2	4,00	3,00	5,00
2	1	5,00	6,00	5,00
2	2	6,00	4,00	7,00
3	1	3,00	2,00	3,00
3	2	6,00	7,00	9,00
4	1	3,00	4,00	2,00
4	2	1,00	2,00	2,00

Tabel 1. Meetgegevens op 9-punt schaal van vier participanten, twee condities en drie items



Figuur 2. Interactie plot van participanten en condities voor de gegevens van tabel 1

van de G-reliability geeft dit adequaat weer: die is heel laag. De grote variantie van de interactie vinden we terug in de noemer van G-reliability. Een verbetering in de situatie? Wanneer dit een reële situatie is zou men kritisch moeten kijken of de participanten wel geschikt zijn voor hun taak, en of selectie/opleiding niet duidelijk een noodzaak is. Zomaar meer participanten opnemen vergroot de waarde van I, en daarmee de G-reliability , maar is misschien niet de meest effectieve weg.

Discussie

De G-reliability geeft nuttige informatie om de meetprecisie vast te stellen in een industriële user studie. We hebben dit laten zien aan de hand van een WSD met één factorcondities. Overigens is dit resultaat natuurlijk eenvoudig te generaliseren voor andere designs, zoals een WSD met meerdere factoren, of voor het Between Subject Design. De karakterisering van afzonderlijke participanten ziet men ook terug in de sensometrie (Brockhoff en Skovgaard, 1994). Hierbij wordt ook de F-waarde van de individuele participant gehanteerd als maat voor kwaliteit, maar er is geen standaardisatie. Het aardige is dat nu de link wordt gelegd tussen deze waarde uit de sensometrie en de reliability uit de psychometrie. Ten slotte geldt dan ook een volgende spin-off: uit het betrouwbaarheids-interval voor de nc-parameter λ kunnen we een betrouwbaarheids-interval bepalen voor G-reliability. En nuttige resultaten hiervoor worden gegeven door bijvoorbeeld Steiger (2004).

LITERATUUR

Brockhoff, P.M. & Skovgaard, I.M. (1994). Modelling individual differences between assessors in sensory evaluations. *Food Quality and Preference*, 5, 215-224.

Clarke, L. A. & Watson, D. (1995). Constructing validity: basic issues in objective scale development. *Psychological Assessment*, 7, 309-319.

Cronbach, L.J. (1947). Test “reliability”: Its meaning and interpretation. *Psychometrika*, 12, 1-6.

Hoyt, C. (1941). Test reliability by analysis of variance. *Psychometrika* , 6, 135-160.

Spector, P.E. (1992). *Summated rating scale construction. An introduction*. Newbury Park: Sage publications.

Steiger, J.H. (2004). Beyond the F test: effect size confidence intervals and tests of close fit in the analysis of variance and contrast analysis. *Psychological Methods*, 9, 164-182.

JAN ENGEL is senior consultant bij CQM in Eindhoven.
E-mail: <Engel@cqm.nl>

IN MEMORIAM

JOOP KEMPERMAN (1924-2011)

J.H.B. (Joop) Kemperman werd op 16 juli 1924 in Amsterdam geboren. Hij studeerde wiskunde aan de Universiteit van Amsterdam, waar hij in 1948 doctoraal examen deed. Hij werkte daarna aan het Mathematisch Centrum (CWI) onder leiding van David van Dantzig, bij wie hij in 1950 promoveerde op een proefschrift getiteld 'The General One-Dimensional Random Walk with Absorbing Barriers, with Applications to Sequential Analysis'.

In 1951 emigreerde hij naar de V.S. Hij was achtereenvolgens hoogleraar aan de Purdue University, University of Rochester en de Rutgers University.

In 1970 ontmoete ik hem voor het eerst in Rochester tijdens mijn bezoek aan Julian Keilson. Ik kwam daar aan na een lange rondrit door de VS vanuit Austin, Texas. Hij had me zijn tweede auto willen lenen – mijn auto was al verkocht – ware het niet dat hij die kort daarvoor al had uitgeleend en defect had teruggekregen.

Dat tekende hem: hij was heel genereus, in alle opzichten, ook met zijn inzichten in de wiskunde. Hij had een opvallend algemeen advies voor wiskundigen: gebruik geen ongelijkheden. Bij het gebruik van ongelijkheden raak je altijd informatie kwijt. Laat dus altijd gelijkheden staan, ook al weet je van sommige grootheden dat ze klein zijn.

Ik kwam hem daarna regelmatig tegen op congressen en weer later, wat uitgebreider, bij mijn tweede bezoek aan Keilson in Rochester in 1972. Toen heb ik ook Joops vrouw Wilna ontmoet, een actieve, opgewekte vrouw. Zij stierf in 1995. De bijzetting van haar as in Alkmaar werd door veel collega's van Joop bijgewoond. Zij was een vrolijke vrouw, en door de vele herinneringen aan haar opgetogen leven kreeg de plechtigheid een uitgesproken opgewekt karakter.

Kemperman was een heel veelzijdige wiskundige. Hij werkte in de analyse, de discrete wiskunde, en vooral, in de kansrekening en statistiek. Verder was hij redacteur van verschillende vaktijdschriften, waaronder *The Annals of Mathematical Statistics*, *The Annals of Probability*, *Stochastic Processes and Applications* en, buiten de stochastiek, van *Aequationes Mathematicae*. Hij had veel promovendi. Het Mathematics Genealogy Project geeft 24 namen; de eerste promoveerde in 1954, de laatste in 1994. Hij was corresponderend lid van de Koninklijke Nederlandse Academie van Wetenschappen.

Zijn eerste artikel verscheen in 1949 in *Indagationes Mathematicae* en schreef hij samen met J.G. van der Corput. De titel luidde 'The second pearl of the theory of numbers'. Zijn laatste artikel schreef hij samen met Mark Brown en verscheen in 2009 in *Probability in Engineering and Information Sciences*: 'Sharp two-sided bounds for distributions under a hazard rate constraint'.

De laatste jaren leed hij aan de ziekte van Parkinson. Joop is op 12 juni in East Brunswick overleden. Op 2 juli is op de Sint Barbara begraafplaats in Alkmaar zijn as met die van zijn vrouw verenigd.

Fred Steutel

Johan van Leeuwaarden

BUURTFEEST

Terwijl ik met een bitterbal op de maat van de muziek mijn gehemelte kastijd zie ik uit mijn ooghoek iemand naderen.

‘Leopold, aangenaam, de overbuurman’, zegt de zelfbewuste vijftiger.

Bittergarnituur maakt je benaderbaar.

Leopold vervolgt: ‘Dus jij werkt aan de TU.’

‘Klopt’, zeg ik aarzelend, niet wetende wat deze bevestiging teweeg zal brengen.

‘Wat doe je dan precies?’

Tja, denk ik, daar gaan we weer. Maar het is feest, dus ik besluit het een kans te geven.

‘Nou, ik doe wiskunde, deels lesgeven, deels onderzoek.’

‘Aha, onderzoek. Dus je bent uitvinder?’

‘Niet echt. Ik probeer wel steeds iets nieuws te bedenken, maar geen uitvindingen.’

‘Maar ik dacht dat ze aan de TU echt dingen uitvonden, zoals MRI scans en elektrische auto’s?’

‘Klopt, dat is ook zo.’

‘Maar jij vindt geen dingen uit?’

‘Nou, misschien toch wel, maar zo heb ik het niet eerder bekeken.’

‘Noem dan eens wat van je uitvindingen.’

Voetbal was een beter onderwerp geweest, maar er is nu geen weg meer terug.

‘Kijk, ik werk eerder aan wiskundige formules, of modellen, die anderen dan weer kunnen gebruiken bij het ontwikkelen van producten. Die modellen en formules zijn nieuw, en in die zin een uitvinding.’

‘Maar wanneer anderen jouw formules gebruiken, levert jou dat dan geld op?’

‘Nee, want iedereen mag die formules gebruiken.’

‘En die modellen, leveren die dan iets op?’

‘Die modellen zijn abstracties van de werkelijkheid. Ze geven wel inzicht, maar leiden doorgaans niet tot een tastbaar product. Althans, mijn modellen niet.’

‘Dus dat onderzoek van jou levert niets op?’

‘Geen geld nee. Wel die formules dus.’

‘Ja, maar die formules geef je blijkbaar voor niets weg. Waarom houd je ze niet voor jezelf, en probeer je er iets mee te doen dat geld oplevert?’

‘Maar dan zou ik een eigen bedrijf moeten beginnen?’

‘Dat klopt ja, en daar lijkt me niets mis mee. De kranten staan vol met verhalen over de universiteiten. Er is een groot begrotingstekort, en het schijnt dat jullie je broek in de toekomst zelf moeten ophouden. Je kunt dus maar beter zorgen dat je iets nuttigs doet, iets wat geld oplevert.’

‘Ik heb het ook gelezen, maar…’

‘En terecht ook’, onderbreekt Leopold, ‘wij maar geld in de universiteiten pompen zonder dat we er ooit iets van terug zien.’

‘Ik denk dat je gelijk hebt’, zeg ik, terwijl ik besluit me te richten op het volgende buurtfeest.

‘Goed dan’, zegt Leopold, ‘zal ik nog een biertje halen?’

‘Doe mij maar iets sterkers’, zeg ik, terwijl ik weer overga tot mijn bitterbal.

Johan van Leeuwaarden is werkzaam in de groep Stochastische Besliskunde bij de faculteit Wiskunde en Informatica van de Technische Universiteit Eindhoven. Tevens is hij research fellow bij EURANDOM.
E-mail: <j.s.h.v.leeuwaarden@tue.nl>