Illustratie: scusi

SEXY DATA SCIENCE

Daniel Oberski*

Laatst deed ik iets wat vast wel meer statistici af en toe doen: ik maakte me hardop zorgen over de replicatiecrisis. P-hacking, HARKing, data dredging, researcher degrees of freedom[1], … nu we er goed over na beginnen te denken is het eigenlijk nog een wonder dat er soms wél iets gerepliceerd wordt, jammerde ik. Een data scientist die mij had aangehoord vroeg toen: 'Waarom gebruiken die sociale wetenschappers niet gewoon een *holdout sample*?' Ja, waarom eigenlijk niet. Dat zou best wel eens kunnen werken! Simpel, maar een beetje als moderne kunst: je had er zelf aan kunnen denken maar dat heb je niet gedaan.

Een omgekeerd voorbeeld. Eén van de beroemdste data scientists ter wereld, Sandy Pentland van MIT, publiceerde in 2015 een artikel in *Science* waarin hij en zijn co-auteurs lieten zien dat 'vier spatio-temporele datapunten genoeg zijn om 90% van de individuen te herleiden' *ondanks anonimisering vooraf* (de Montjoye e.a., 2015). Niet alleen *Science*, maar ook *Nature* en zelfs de *New York Times* kopten geschokt: *With a few bits of data, researchers re-identify 'anonymous' data*.[2] In alle commotie was echter toch ook een groep onderzoekers minder ondersteboven van de resultaten. Een groep statistici uit Tarragona, gespecialiseerd in *statistical disclosure control*.[3] Zij publiceerden een commentaar in *Science* waarin ze lieten zien dat anonimisering wél tot adequate bescherming kan leiden. Tenminste, als je gebruik maakt van de over de laatste 40 jaar vergaarde statistische kennis over dit onderwerp (Sánchez e.a., 2016).[4]

Wat deze voorbeelden illustreren is het ontstaan van een synthese – goedschiks of kwaadschiks – tussen de 'twee culturen' die Leo Breiman vijftien jaar geleden al omschreef (Breiman, 2001). *In one corner*: Data Science[5] – kampioen pragmatisch aanpakken en ad-hoc scriptjes schrijven die iets nuttigs doen, kenner van algoritmes, presentator van business cases, koning van Facebook tot Twitter, enz. enz. met helden als Tukey, Breiman, Andrew Ng, en, laten we zeggen, Hemelrijk.[6] *In the other corner:* Statistiek – kampioen dingen tot op de bodem uitzoeken, koning van kans en consistentie, baron van bias tot steekproefvariantie, markies van modellering, randomizeerder van experimenten en kiezer van priors, met helden als Tukey, Breiman, Laplace, Bayes, Pearson, Fisher, Neyman, Rubin, en, in de vergelijking blijvend, Van Dantzig en Van Zwet.[7]

Twee werelden

Vanzelfsprekend is deze tegenstelling gechargeerd en is er al van oudsher enige overlap. Amazon gebruikt bijvoorbeeld *matrix factorization* om nieuwe producten aan te raden, en daar deed Gifi ook al aan (Van Der Heijden & Van Buuren, 2016). Tukey is een held in beide culturen; met name zijn artikel 'The future of data analysis' wordt beschouwd als onderdeel van de geschiedenis van beide velden. En biostatistici zijn al langer bezig met 'grote' datasets, vooral met meer kolommen dan rijen. De statisticus Tibshirani stelde om dat probleem op te lossen de lasso voor en die techniek vindt weer gretig aftrek in *machine learning* (Hastie e.a., 2015).[8] Later hebben statistici als Van de Geer zo beide velden kunnen verbeteren.

Toch zijn statistiek en data science vaak nog twee werelden. Maar wel werelden die elkaar steeds vaker vinden. Dat is logisch, want beide velden groeien als kool op de vele nieuw ontgonnen datavelden en we zijn allemaal bezig met dezelfde problemen: data verzamelen, datakwaliteit beoordelen, data beschrijven, voorspellingen doen, oorzaken en gevolgen uit elkaar halen, bedrijven, overheden, en personen adviseren, resultaten communiceren aan een breed publiek, etc. Of je dat nu 'statistiek' of 'data science' wilt noemen, het ligt voor de hand dat je iets kan leren van een ander die er (ook) goed in is. Zeker als die ander nét een andere blik heeft, zoals bovenstaande voorbeelden wilden illustreren.

Statistiek en data science kunnen dan ook veel voor elkaar betekenen. Enerzijds kennen statistici veel algemene principes die nuttig kunnen zijn in het bredere gebied van de data science. Neem bijvoorbeeld het gebruik van Twitter bij het bestuderen van rampen. Klinkt mooi, maar zijn tweets tijdens een overstroming missing at random? Of neem de zorgen van een bedrijf als Booking.com over welke 'A/B-tests' ze moeten doen en hoe ze er honderden kunnen vergelijken. Kunnen ze misschien veel geld besparen met *fractional factorial designs*? Wat te doen als je een beslisboom wilt schatten op herhaalde metingen? Weet iemand daar een oplossing voor? En hoe kun je je klanten eigenlijk het beste vragen naar hun mening? Is vragenlijstontwerp arbitrair of heeft iemand daar als eens onderzoek naar gedaan?

Anderzijds heeft data science de statistiek ook veel te bieden. Er zijn veel categorieën maar ik noem er drie. Ten eerste een aanpak: die van data analyseerders als Tukey en Jan de Leeuw; het pragmatisme van goed genoeg en op tijd boven 'correct'. Ten tweede expertise in het aanboren van nieuwe databronnen, het omgaan met APIs, database management systemen, andere programmeertalen dan R en het creëren van kwalitatief hoogwaardige software met behulp van *unit testing*, *version control*, *staging*, goede documentatie en andere goede gewoonten. Ten derde maken technieken uit de wereld van *machine learning* en *data mining*: *regularization*, neurale netwerken, *support vector machines*, en het concept van *embeddings*, om maar wat voorbeelden te noemen, nog geen deel uit van het standaardarsenaal van veel statistici, maar kunnen vaak een nuttig gereedschap bieden voor hun taken.

Broodnodige kruisbestuiving

In Nederland zijn er al veel voorbeelden van samenwerking en, laten we eerlijk zijn, *rebranding*. Wie vroeger

statistisch consult gaf noemt zich nu vaak ook ‘data scientist’. Studenten van onze universiteiten vinden werk als data scientist bij overheden, bedrijven, en onderzoeksinstituten. Master-opleidingen zijn inhoudelijk aangepast en bevatten steeds meer data science. Daarbij zijn statistici, waaronder leden van de VvS+OR, nauw betrokken bij de data science-initiatieven van Nederlandse universiteiten, zoals die van Amsterdam, Leiden, Eindhoven, Tilburg, en (sinds kort) Utrecht. Dankzij dit alles is, in tegenstelling tot de situatie op andere plaatsen, de statistiek in Nederland vrij goed vertegenwoordigd in data science.

Toch kan er mijns inziens nog veel meer gebeuren, en dan vooral op het gebied van onderzoek. De eerste stap, *rebranding* en samenwerking, is gezet. De tweede, veel moeilijker, stap is het daadwerkelijk beter maken van beide vakgebieden door kruisbestuiving. En op dat gebied wil de VvS+OR een belangrijke bijdrage leveren.

Om die bijdrage te coördineren en data science en statistiek inhoudelijk bij elkaar te brengen hebben we daarom de spiksplinternieuwe Sectie Data Science opgericht. Onze doelstellingen zijn:

- Samenbrengen van mensen die met data science bezig zijn uit *alle* onderzoeksvelden, inclusief statistiek, OR, informatica, en toegepaste wetenschappen, met als doel het *wederzijds* uitwisselen van kennis en verbeteren van methoden;
- Verbreden VvS+OR door nieuwe aanwas uit andere vakgebieden;
- Profileren statistiek in Nederland via data science;
- Coördineren en organiseren van concrete activiteiten, zoals:
  - Hackathons (een eerste voorproefje zal plaatsvinden op de Dag voor Statistiek en Besliskunde op 23 maart);
  - Lezingen en webinars informatici, statistici, data scientists voor gemengd publiek (aftrap wordt gegeven door de lezing van Max Welling op 23 maart);
  - Andere activiteiten die de lezer in wil brengen!

De Sectie Data Science is actief betrokken bij de organisatie van de Dag voor Statistiek en Besliskunde. Het bestuur van de sectie bestaat uit Daniel Oberski (voorzitter), Katrijn Van Deun, Alessandro Di Bucchianico, Arend Oosterhoorn, en Maarten Joosen met ondersteuning van Erik-Jan van Kesteren. Zie ook de website http://sectiedatascience.nl/.

* Met dank aan Katrijn van Deun en Alessandro Di Bucchianico

Noten

1. https://fivethirtyeight.com/features/science-isnt-broken
2. https://bits.blogs.nytimes.com/2015/01/29/with-a-few-bits-of-data-researchers-identify-anonymous-people/
3. Voor lezers die niet bekend zijn met dat veld: het zijn de technieken die bijvoorbeeld het CBS al sinds jaar en dag gebruikt om vast te stellen of statistieken die zij maken niet herleid kunnen worden tot personen. Zie bijvoorbeeld Hundepool e.a. (2012).
4. Overigens kreeg de groep van Pentland het laatste woord (de Montjoye & Pentland, 2016) dat ze aanwendden om zich te beklagen over het gebrekkig begrip van de statistici van de *big*-heid van hun data. De discussie is dus verre van afgelopen!
5. Op de vraag of data science nu eigenlijk statistiek is of andersom of misschien wel geen van beide ga ik hier niet in. Ook zal ik sommige termen niet uit het Engels vertalen als ik denk dat dat verwarring op zou leveren.
6. Als Hemelrijk meer data had gehad. Dank aan Peter Grünwald voor deze suggestie.
7. Mijn oprechte excuses als uw voorkeur er niet tussen staat, of uw afkeur wel. Hopelijk geeft het nochtans een idee van de tegenstelling die door Breiman werd geschetst. (Ik laat me overigens graag corrigeren).
8. Kleine kanttekening hierbij: Breiman stelde als eerste een methode voor variabelenselectie voor, quasi gelijktijdig (één jaar eerder) met de lasso en dit is de nonnegative garrote.

Literatuur

Breiman, L. (2001). Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). *Statistical Science: A Review Journal of the Institute of Mathematical Statistics*, *16*(3), 199–231.

De Montjoye, Y.-A., & Pentland, A.S. (2016). Response to Comment on ‘Unique in the shopping mall: On the reidentifiability of credit card metadata’. *Science*, *351*(6279), 1274.

De Montjoye, Y.-A., Radaelli, L., Singh, V.K., & Pentland, A.S. (2015). Unique in the shopping mall: On the reidentifiability of credit card metadata. *Science*, *347*(6221), 536–539.

Hastie, T., Tibshirani, R., & Wainwright, M. (2015). *Statistical Learning with Sparsity*. CRC Press.

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Nordholt, E.S., Spicer, K., & De Wolf, P.-P. (2012). *Statistical Disclosure Control*. John Wiley & Sons.

Sánchez, D., Martínez, S., & Domingo-Ferrer, J. (2016). Comment on ‘Unique in the shopping mall: On the reidentifiability of credit card metadata’. *Science*, *351*(6279), 1274–1274.

Van der Heijden, P.G.M., & Van Buuren, S. (2016). Looking back at the Gifi system of nonlinear multivariate analysis. *Journal of Statistical Software*, *73*(4).

Daniel Oberski is Universitair Hoofddocent Data Science Methodology bij de vakgroep Methodologie en Statistiek van de Universiteit Utrecht.
E-mail: d.l.oberski@uu.nl. Twitter: @DanielOberski

Gerrit Stemerdink

column

M/V of V/M

Moeten mannen eerst worden genoemd en daarna vrouwen? Of is het andersom? Denk maar niet dat zoiets niet belangrijk is.

Ik neem geen standpunt in, al was het maar dat ik niet verwikkeld wil raken in welke feministische discussie dan ook. Ik neem slechts waar. Nog steeds is het in de westerse wereld gebruikelijk om mannen voorop te stellen. Zo heet mijn echtgenote J.C. Stemerdink-van de Wetering. Maar sinds enige tijd is de wet veranderd, ik vind dat volkomen terecht. Bij een huwelijk kan men nu kiezen hoe men wil heten. Als ik nu zou trouwen zou ik als man de naam Van de Wetering-Stemerdink mogen voeren.

In de wetenschap kan over de man/vrouw volgorde óók verschillend worden gedacht. Zelfs binnen een en hetzelfde vakgebied als de biologie! Daar heb ik een aardige ervaring mee gehad.

Een groot deel van de jaren tachtig van de vorige eeuw was ik werkzaam aan de Landbouwuniversiteit Wageningen. Het was een heerlijke tijd, als statisticus bij het Rekencentrum kreeg ik alle vrijheid me in nationaal en internationaal verband bezig te houden met vergelijking van software op aspecten als rekennauwkeurigheid en gebruikersgemak. Ook werd ik veel geraadpleegd door onderzoekers uit alle mogelijke vakgebieden, de LUW was een heel brede universiteit, ondanks de beperking die de naam suggereert.

Ooit promoveerde daar een statisticus op een onderzoek naar variantiecomponenten in kruisingsschema’s. Zoiets is belangrijk bij pogingen de eigenschappen van een ras te verbeteren, of dat nu melkkoeien of doperwtjes zijn. Gebruikelijk is dan dat er een stamboom wordt opgesteld en dat gekeken wordt in hoeverre de eigenschappen van de ouders in hun nakomelingen terug te vinden zijn. Er komen dan termen als Fragaria virginiana Mill. x Fragaria chiloensis L. aan te pas. Ik had het proefschrift met veel belangstelling gelezen, daarbij was me niets bijzonders opgevallen. Tenslotte had ik nooit enige aandacht besteed aan het verschijnsel man/vrouw-volgorde, ik vond dat eigenlijk niet relevant.

Maar Wageningen telde in die tijd een nogal activistische groep vrouwelijke wetenschappers. Dat ze zo activistisch waren zag ik als een logische reactie op een wereld die voor 80% door vaak bijzonder masculien denkende mannen werd bevolkt. Tijdens de promotie kwam een van deze dames met de opmerking dat het vreemd was dat in kruisingsschema’s bij planten éérst de vrouw genoemd werd en dan de man, maar dat het bij dieren andersom was. Zij wilde weten wat de promovendus daarvan vond, ook al omdat het verwarrend was beide varianten in één proefschrift aan te treffen. Ik heb nooit kunnen achterhalen of deze vraag voorgekookt was, zoals veel vragen bij promoties. Maar het antwoord was grandioos. De letterlijke tekst weet ik niet meer, maar het kwam neer op: ‘Ik vind het niet zo’n probleem, het is maar net wat men gewend is. We hebben hier te maken met twee separaat ontwikkelde tradities. Maar laat ik u gerust stellen, in tegenstelling tot de notatievolgorde staat tijdens de lijfelijke uitvoering van de kruising in de dierenwereld altijd het vrouwtje voor en het mannetje achter’. De aula barstte los in een geweldige lachbui, de rector magnificus had grote moeite het decorum van de plechtigheid te herstellen...

Gerrit Stemerdink is eindredacteur van *STAtOR*.
E-mail: gjstemerdink@hotmail.com