Leven in onzekerheid en eenvoud

Kiers, H.A.L.

Published in:
De Psycholoog

Publication date:
2001

Zeer gewaardeerde aanwezigen,

Leven in Onzekerheid en Eenvoud

Ik zal maar met de deur in huis vallen en meteen vertellen waarom ik heb gekozen voor deze ietwat sentimentele titel van mijn rede. *Leven in onzekerheid* slaat op de wetenschapper die zich als geen ander geconfronteerd weet met onzekerheden - en daarom misschien maar beter de "nooit-helemaal-zeker-weter" zou moeten heten. *Leven in eenvoud* doelt niet op het in materiële zin eenvoudige bestaan van de wetenschapper, maar slaat op de kennis van de wetenschapper, die namelijk alleen maar toepasbaar is op vergaand vereenvoudigde situaties. Desalniettemin geloof ik dat wetenschappers met hun sterk overgesimplificeerde en onzekere weergave van de werkelijkheid de samenleving van dienst kunnen zijn. Simplificatie leidt tot inzicht, en onzekerheid is een existentieel gegeven. Geen van beide kunnen in de wetenschap worden genegeerd.
Waarom kies ik er voor om mijn oratie aan dergelijke algemene begrippen op te hangen? Eén reden is dat juist het al dan niet verkrijgen van inzicht en zekerheid sturend is geweest in mijn eigen wetenschappelijke ontwikkeling. Op de middelbare school wist ik al dat ik wetenschapper wilde worden. Ik wilde weten hoe de natuur in elkaar zit, van melkwegstelsels tot quarks, met overigens vrij weinig interesse voor alles wat daar tussen zit. Kennis was toen voor mij onfeilbare kennis: zolang iets onzeker is, is het geen kennis. Ik had dan ook de neiging om wiskunde te gaan studeren (wat levert er immers zekerder kennis dan de wiskunde), maar mijn wiskunde-docent raadde me dat af en suggereerde om iets te kiezen wat dichter bij de werkelijkheid stond. Het is toen sterrenkunde geworden.
De basis van de sterrenkunde-studie bestond uit veel wis- en natuurkunde en daarmee werd mijn behoefte aan zekere kennis afdoende bevredigd. Maar ook merkte ik dat de relatie met de alledaagse werkelijkheid ver te zoeken was. Toen mij bleek dat psychologie óók een wetenschap was, en dat die over zoiets alledaags als het menselijk gedrag ging, was de overstap gauw gemaakt. Vol enthousiasme stortte ik me op de functieleer en vervolgens de sociale psychologie, maar mijn enthousiasme raakte gaandeweg getemperd toen ik leerde dat het in de psychologie niet bepaald gemakkelijk is om iets over het menselijk gedrag aan de weet te komen. De mens kan als onderzoeksobject immers de onderzoeksuitkomsten over diens gedrag beïnvloeden en doet dat ook. Waar natuurkunde zekere kennis bood over relatief oninteressante dingen als het opwekken van een elektrisch spanningsveld met behulp van een magneet, ging psychologie over veel interessanter zaken als interpersoonlijke aantrekking, maar leerde het me niet met zekerheid hoe ik een vonk kon doen overslaan. Iets wat overigens onder psychologiestudenten ook zonder die kennis wel lukte.
Mijn behoefte aan zekere kennis zal er zeker toe hebben bijgedragen dat ik me binnen psychologie toch weer ging richten op de meest wiskundige vakken, die waren bedoeld als ondergrond voor de statistiek. Zo was ik na de nodige omzwervingen alsnog in de pure wiskunde beland, en vond ik het een heel spannende uitdaging bezig te gaan met het bewijzen van een ingewikkelde wiskundige stelling die ook de echte wiskundigen nog niet hadden bewezen.
De zekerheid van wiskundige uitspraken is een groot genot, maar de keerzijde van wiskundig onderzoek is dat je, wanneer je er aan begint niet *zeker* weet of je dat genot later ook mag smaken: Als het je niet lukt om wiskundige bewijzen te leveren is het een schrale troost dat wiskunde, in geval van succes, altijd zekere kennis levert. Juist vanwege mijn behoefte aan zekerheid, zat ik dus in de wiskunde toch ook niet geheel op m'n plek. Weliswaar was, dankzij de samenwerking met en stimulans van Jos ten Berge succes in veel gevallen zo goed als zeker, maar het benodigde geduld ervoor kreeg ik zelf toch steeds minder. Terwijl kort na mijn afstuderen mijn onderzoek zich vooral richtte op de wiskundige fundamenten van

technieken voor gegevensverwerking, werd mijn onderzoek gaandeweg steeds meer gestuurd in de richting van de ontwikkeling van methoden voor gegevensverwerking met nadruk op de interpretatie van uitkomsten. Deze ontwikkeling werd gestimuleerd door adviesvragen van studenten en medewerkers over het analyseren van hun gegevens. Ik wilde hen kunnen uitleggen wanneer je welke techniek toepast, hoe je dat doet, en vooral ook, hoe je de resultaten ervan interpreteert. En daarmee kwam ik, ironisch genoeg, alsnog in aanraking met de onzekerheden waar psychologisch onderzoek bol van staat.

Waarom vereenvoudiging?

Bij wetenschappelijk onderzoek is een eerste stap altijd het vereenvoudigen van begrippen en resultaten. Onderzoekt men bij voorbeeld het verschil in stressbestendigheid van mannen en vrouwen dan zal allereerst het begrip stressbestendigheid moeten worden vereenvoudigd. Stressbestendigheid zelf kan bij voorbeeld beschreven worden door hoe snel iemand hoofdpijn of rugpijn krijgt, maar ook door hoe lichtgeraakt iemand is, hoe gauw uit het veld geslagen, of geneigd tot snoepen. Wanneer men een eenduidig antwoord wil op de vraag of mannen stressbestendiger zijn dan vrouwen, dan zal die veelheid van facetten van stressbestendigheid moeten worden vereenvoudigd tot één begrip, bij voorbeeld gedefinieerd als de gemiddelde score op schalen die meten in hoeverre men last heeft van alle bovengenoemde zaken (van hoofdpijn t/m neiging tot snoepen). Na deze eerste vereenvoudiging kan men weliswaar stressbestendigheid bij mannen en vrouwen meten, maar, zelfs al zouden we dat bij alle mannen en vrouwen ter wereld doen, dan nog hebben we geen antwoord op de vraag of mannen stressbestendiger zijn dan vrouwen. Ten eerste moeten we de vraag nader specificeren: willen we weten of alle mannen stressbestendiger zijn dan alle vrouwen, of willen we weten of mannen gemiddeld genomen stressbestendiger zijn dan vrouwen. De eerste vraag is misschien wel interessant, maar de uitkomst zal dat niet zijn: Er hoeft immers maar één vrouw gevonden te worden die stressbestendiger is dan één man, en de bewering gaat al niet meer op. Daarom wordt heel vaak met de vereenvoudiging gewerkt die ligt besloten in de vraag "Zijn mannen *gemiddeld* genomen stressbestendiger dan vrouwen?"[1]

Waarom onzekerheid?

Zelfs als men de stressbestendigheid van alle mannen en vrouwen ter wereld heeft onderzocht, is het resultaat toch maar zeer beperkt houdbaar: de wereldbevolking verandert binnen een seconde, en, erger nog, zelfs de stressbestendigheid van een individu zal op het ene moment anders uitvallen dan op het andere. En, in de praktijk werkt men natuurlijk ook niet met alle wereldburgers, maar in het gunstigste geval met een representatieve steekproef. Dergelijke onzekerheden in onze kennis zijn onvermijdelijk, maar worden desalnietemin, vooral in de populair-wetenschappelijke pers genegeerd.

[1] Er zijn overigens ook geheel andere vereenvoudigingen mogelijk. In plaats van het bepalen van de gemiddelde stressbestendigheid kan men bij voorbeeld ook kijken wat de "middelste" stressbestendigheidswaarde bij mannen is, en idem bij vrouwen. Een heel andere aanpak, die een heel gemakkelijk interpreteerbare maat voor vergelijking van stressbestendigheid van mannen en vrouwen levert, verkrijgt men door van alle mogelijke man-vrouw combinaties aan te geven in hoeveel procent van de gevallen de man stressbestendiger is dan de vrouw. Deze maat is als een soort kans op te vatten, die dan leidt tot de uitspraak dat men schat dat, bij voorbeeld in 60% van de gevallen dat men zowel een man als een vrouw voor zich heeft, de man het stressbestendigst van de twee zal zijn. Een vergelijkbare maat is Mann-Whitney's U grootheid, maar die wordt niet vaak als zodanig gerapporteerd, en is ook minder inzichtelijk.

Dat de pers niet van nuances houdt blijkt ook uit de fraaie krantenkop “Bij stress wordt het een meisje”[2]. De titel wekt de indruk dat men het geslacht van een te verwekken kind kan kiezen door zich in die tijd uitdrukkelijk aan stress bloot te stellen (als men een meisje wil) of juist zo ongestressed mogelijk door het leven te gaan (als men een jongetje prefereert). Bij nadere lezing van het artikel bleek het wat minder duidelijk in elkaar te zitten. Er was gevonden dat het doorstaan van zeer stressvolle gebeurtenissen zoals een hartaanval of een kankerdiagnose bij de partner de kans op een meisje met 2% verhoogde van 49% naar 51%, zodat de bedoelde stressbelasting toch wel wat veel gevraagd is voor deze marginale verandering. In de publieke media mag men dan niet houden van nuanceringen en gegevens over onzekerheden, in de wetenschap dienen die wel te worden geëxpliciteerd.

De rol van de statistiek

Maar al te vaak verwacht men van de statistiek dat daarmee uitkomsten kunnen worden bewezen, maar, zoals een collega het afgelopen zomer nog verwoordde, “Statistics is not a confirmatory science” (R. Rocci, personal communication, juli 2000), en dat geldt zelfs voor de zogenaamde confirmatieve methoden. De statisticus is een bescheidener rol toebedeeld, namelijk om de onderzoeker te adviseren in het vereenvoudigen van onderzoeksgegevens (data-vereenvoudiging), en in het aangeven van de onzekerheidsmarges rond de resultaten (onnauwkeurigheidsanalyse). Het proces van data-vereenvoudiging valt veelal samen met de zogenaamde beschrijvende statistiek, terwijl onnauwkeurigheidsanalyse vooral bedreven wordt in het kader van de inferentiële statistiek. De inferentiële statistiek beantwoordt de vraag “Wat zou er gebeuren als ik het overnieuw doe?”, zoals Ivo Molenaar dit al sinds jaar en dag formuleert. Beschrijvende statistiek geeft antwoord op de vraag die daar aan vooraf gaat: “Wat zit er eigenlijk in mijn data?” Derhalve zijn ook juist de beschrijvende technieken, die dus bedoeld zijn voor data-vereenvoudiging van groot belang in de data analyse[3].

Data-vereenvoudiging

Ik heb zo straks gezegd dat het rapporteren van gemiddelden een vorm van data-vereenvoudiging is. Dit is echter de simpelste situatie. Als men onderzoek doet meet men namelijk zelden alleen maar één eigenschap. Men wil bij voorbeeld stressbestendigheid relateren aan de bloeddruk, of aan de intelligentie, of aan de persoonlijkheid, en moet dan dus bij een aantal mensen naast de stressbestendigheid ook de bloeddruk meten, de intelligentie en de persoonlijkheid. Bloeddruk meten gaat relatief gemakkelijk. Intelligentie is een complexer begrip, en kan bij voorbeeld worden uitgesplitst naar verbale, numerieke en ruimtelijke intelligentie. Domweg melden dat iemands IQ 120 is is dus niet zo informatief. Waar men bij IQ tenminste dan nog kan spreken over een soort algemene intelligentie, ligt dat bij persoonlijkheid wat moeilijker[4], en worden er aan persoonlijkheid tegenwoordig zo’n vijf dimensies toegekend, die niet zonder gevoel voor drama de Big Five zijn genoemd (Goldberg,

[2] *Intermediair*, 9-9-1999.

[3] Nu doet zich het merkwaardige fenomeen voor dat onderzoekers vaak zo gericht zijn op de inferentiële statistiek dat ze de technieken die daarvoor gebruikt worden ook gebruiken voor de beschrijvende analyse. Zo zijn sommige onderzoekers zo gewend om hun data te analyseren met variantie-analyse of t-toets dat ze bij voorbaat de scores op een variabele gaan onderverdelen in categorieën als hoog versus laag, terwijl men beschikt over scores in een aanzienlijk bredere range.

[4] Echter, Hofstee (1996, zie ook Hofstee, ten Berge & Hendriks, 1998) merkt op dat een dergelijke algemene persoonlijkheidsfactor wel degelijk zinvol te definieren is en *sociale wenselijkheid* genoemd zou kunnen worden, mits deze dan niet als (zelf)beoordelaarsartefact wordt opgevat, maar als persoonsbeschrijvende eigenschap.

1981). Misschien zijn die Big Five echter wel *te Big* want het idee dat iemands persoonlijkheid, oftewel, hoe iemand zich gedraagt, in termen van 5 hoofdkenmerken kan worden beschreven lijkt toch wel een wat al te sterke vereenvoudiging. De persoonlijkheid is hier ingeperkt tot algemene gedragskenmerken van personen. Daardoor wordt er heel wat, met name situatie-gebonden gedrag buiten beschouwing gelaten. Het zal immers zeker ook van de situatie afhangen hoe iemand zich gedraagt. Zelfs een doorgaans zeer beheerst persoon kan uitbarsten in situaties waar anderen de schouders over ophalen, en een enthousiaste bungee-jumper kan in de sociale omgang best heel angstig zijn. Al ettelijke decennia geleden is bedacht dat het nuttig zou zijn om patronen te ontdekken in dergelijke situatie-gebonden gedragingen. Het mooie is dat tegenwoordig daarvoor ook de technieken beschikbaar zijn. Men kan dan namelijk zogenaamde drieweg-gegevens verzamelen en deze analyseren met drieweg-technieken (Tucker, 1966, Kroonenberg & de Leeuw, 1980). Drieweg-gegevens zijn dan in dit geval scores van een aantal personen (de eerste weg) op een aantal gedragsbeschrijvende items (de tweede weg), met betrekking tot verschillende situaties (de derde weg). In een onderzoek waarvoor ik met Iven van Mechelen de gegevens heb geanalyseerd (van Mechelen & Kiers, 1999, Kiers & van Mechelen, 2001), gingen de items over uitingen van spanning of angst: Zo moesten proefpersonen op een schaal van 1 tot 5 aangeven in hoeverre hun hart sneller gaat kloppen, ze misselijk worden, etc., in situaties als, een nacht alleen in een bos verkeren, een bezoek brengen aan een psycholoog, of een toespraak houden voor een grote zaal. In die laatste situatie kan ik mij op dit moment vrij goed inleven, en ik zou nu dus heel goed kunnen aangeven in hoeverre mijn hart sneller gaat kloppen, ik misselijk wordt, etc., maar als proefpersoon zou ik dit dus moeten doen voor alle situaties die de onderzoeker heeft bedacht.

Hier in onze vakgroep is Maaike ten Berge een wat grootschaliger aanpak begonnen, waarin 64 proefpersonen moesten aangeven in hoeverre elk van de 71 mogelijke reacties van toepassing waren in 142 verschillende situaties, hetgeen in totaal dus meer dan een half miljoen gegevens opleverde. Het doel is nu om patronen te ontdekken in dergelijke grote drieweg-data sets. Mijn onderzoek naar de analyse van drie- en meerweggegevens richt zich dan ook vooral op de vraag hoe we uit dergelijke grote onoverzichtelijke hoeveelheden data patronen kunnen destilleren. Daartoe is het handig om de gegevens te vereenvoudigen op zo'n manier dat personen, reacties en situaties allemaal in groepen worden ingedeeld. Vervolgens kunnen we dan aangeven in hoeverre een bepaalde typische angstreactie bij elk type persoon door een bepaald type situatie wordt uitgelokt. Zo'n patroon zou dan kunnen zijn dat een bepaald type persoon vooral *heftige lichamelijke reacties* vertoont in typische *beoordelingssituaties*, maar in *eenzame situaties* niet bijzonder angstig reageert, terwijl een ander type persoon vooral een positief spanningsgevoel krijgt in *eenzame situaties*, maar verder niet warm of koud wordt van *beoordelingssituaties*.[5]

Wanneer zo dan een algemeen beeld is verkregen over hoe mensen verschillen in hoe ze reageren op verschillende situaties, kan men vervolgens tests ontwikkelen om systematisch na te gaan, of zelfs te voorspellen, hoe een individueel persoon reageert op bepaalde situaties. Dergelijke situatiegebonden gedragsbeschrijvingen kunnen van groot belang zijn bij personeelsselectie, klinisch psychologische diagnostiek, etc.

[5] Dergelijke resultaten komen niet zo maar uit een drieweg-analyse naar voren. Ten eerste blijkt het in de praktijk zelden zo te zijn dat men heel duidelijk groepen personen kan onderscheiden, simpelweg omdat er altijd wel weer aardig wat personen qua eigenschappen precies tussen groepen personen in blijken te zitten, zodat er een vloeiende overgang van de ene groep naar de andere groep ontstaat. Voorts, als men al wel in staat is reacties en situaties te vereenvoudigen in groepen reacties resp. situaties, dan zou het vervolgens ook nog wenselijk zijn als de patronen van soorten reacties op soorten situaties relatief eenvoudig zijn. Maar eenvoud aan de ene kant gaat niet altijd samen met eenvoud aan de andere kant. Een oplossing voor het vinden van bruikbare compromissen tussen deze vormen van eenvoud wordt beschreven in Kiers (1998).

Een fascinerend aspect aan de statistiek is dat het in heel verschillende vakgebieden wordt gebruikt, en dat de toepassingen qua gegevensanalyse soms heel sterke overeenkomsten hebben. Zo werk ik al een aantal jaren succesvol samen met Age Smilde en zijn groep chemometrici, die zich vooral richten op de analyse van chemische gegevens. Daarvoor worden bij voorbeeld mengsels van onderling reagerende chemische stoffen waarvan men de samenstelling niet kent doorgelicht door er licht van verschillende golflengtes doorheen te zenden en op een serie achtereenvolgende tijdstippen te meten hoeveel van het licht van elke golflengte door zo'n mengsel wordt geabsorbeerd. Ook hier is sprake van drieweg-gegevens, waarbij de wegen dan zijn (1) de verschillende mengsels, (2) de verschillende golflengtes en (3) de verschillende tijdstippen. Het doel is hier de samenstelling van de mengsels te achterhalen. Hoewel de materie sterk verschilt, is de doelstelling uiteindelijk sterk vergelijkbaar met die in de psychologie. In de chemie probeert men de ingrediënten te ontdekken die verantwoordelijk zijn voor het gedrag van bepaalde stoffen, terwijl we in de psychologie ingrediënten proberen te ontdekken die verantwoordelijk zijn voor het gedrag van mensen.
Drieweg-gegevens komen ook voor in de economie, en dan met name bij het marktkundig onderzoek[6], waarbij men dan bij voorbeeld gegevens wenst te analyseren die door uzelf en mij worden aangedragen voor zover u meedoet aan het sparen van airmiles en dergelijke. Doordat van elke klant precies wordt vastgelegd hoeveel deze wanneer van welk product, in welk filiaal koopt krijgt men al gauw grote complexe data sets. Zo kan men bij voorbeeld de aantallen aankopen van een groot aantal verschillende producten, door een groot aantal klanten, in een groot aantal weken analyseren, en is men alweer bij de drieweg-analyse beland. Het doel van de analyse is dan onder meer categorieën van klanten met soortgelijk koopgedrag te ontdekken. Dergelijke marktsegmenten kunnen dan van op maat verzorgde reclameaanbiedingen worden voorzien, wat geheel en al in het belang is van de klant (die dan bij voorbeeld niet langer gepijnigd hoeft te worden met aanbiedingen van veel te dure producten die hij toch niet kan kopen), en wellicht nog veel meer van het bedrijf zelf.
Een laatste toepassing waar ik me mee bezig houd [7] ligt op het terrein van de neurologie, en dan wel met name het onderzoek wat gebruik maakt van fMRI-hersenscans. Terwijl één hersenscan al een groot databestand oplevert van vele pixelintensiteiten, is het voor onderzoek naar welke hersengebieden geactiveerd worden tijdens welk soort mentale activiteit gebruikelijk om een hele serie scans onder verschillende condities te nemen, en dat dan van verschillende proefpersonen. Dergelijke enorme data sets kunnen, dankzij een comprimeringstruc (zie Kiers & Krijnen, 1991, Kiers, Kroonenberg & ten Berge, 1992, Kiers & Harshman, 1997) waar ik een paar jaar geleden aan heb gewerkt, toch heel gemakkelijk en snel met drieweg-analyse worden geanalyseerd. Met dergelijke analyses kan men dan trachten op te sporen welke activiteiten in de hersenen veelvuldig samengaan, of zelfs gekoppeld zijn, en hoe dergelijke basale vormen van hersenactiviteit gerelateerd kunnen worden aan het al dan niet optreden van ziekten zoals Huntington's disease en Alzheimer. Aldus kan men meer inzicht krijgen in de neurologische correlaten van menselijk (of dierlijk) gedrag.

Kansrekening

Als men aan statistiek denkt, denkt men vaak ook aan kansrekening. Toch is kansrekening binnen de beschrijvende statistiek, oftewel, de zojuist behandelde data-vereenvoudiging, lang niet altijd aan de orde. Sommige van onze vakgenoten vinden de data-vereenvoudiging, waaraan met name in Leiden en Groningen veel gedaan wordt, een minderwaardige tak van

[6] zie bij voorbeeld Cooper, Klapper & Inoue (1996).
[7] in samenwerking met dr. Paul Maguire, Neuroimaging center, AZG, Groningen.

sport binnen de statistiek. Dit blijkt uit het feit dat ze ons, vanwege het feit dat we kansrekening weinig gebruiken, de "kansarmen" in ons vakgebied noemen. In plaats van ons tegen deze ietwat denigrerende benaming te verzetten, is het misschien beter om de term als geuzennaam te gaan voeren, al was het maar omdat de betekenis van geuze, "bedelaar", wel zeer dicht tegen kansarm aanligt, en desondanks de geuzen belangrijke pioniers in de vaderlandse geschiedenis werden.

Wij kansarmen gebruiken dan de kansrekening misschien soms wat te weinig, omgekeerd denk ik dat, wat de "kanskapitalisten" doen, namelijk het zonder meer voorop zetten van het denken in kanstermen ook niet verstandig is. Allereerst wil ik aanvoeren dat er legio situaties zijn waarin kansrekening helemaal niet aan de orde is, omdat men, zoals bij voorbeeld in toegepast onderzoek, helemaal niet wil generaliseren naar andere dan de eigen setting[8]. Stel dat men echter wel wil generaliseren, bij voorbeeld wanneer men algemene uitspraken wil doen over het verschil in stressbestendigheid van mannen en vrouwen. Omdat men niet "de gemiddelde man" of "de gemiddelde vrouw" op kan sporen, trekt men een willekeurige steekproef[9] van mannen en vrouwen, en kan men vervolgens statistische uitspraken doen over het verschil in stressbestendigheid van mannen en vrouwen. Dan is hierbij een, in principe, bescheiden rol weggelegd voor kansuitspraken. In de praktijk worden die kansuitspraken veelal gedaan in het kader van een nulhypothesetoetsingsprocedure, en eenmaal in dat kader gebracht krijgen de kansuitspraken vaak opeens een hoofdrol toebedeeld. Naar mijn mening komt dat vooral doordat veel onderzoekers dergelijke kansuitspraken niet goed begrijpen en daardoor denken dat ze veel bruikbaarder zijn dan ze eigenlijk zijn.

Het denken in kanstermen *is* ook niet gemakkelijk, en sluit soms heel slecht aan bij de intuïtie[10]. Om dit te illustreren gebruik ik graag het volgende voorbeeld[11]: Hoe groot denkt u dat de kans is dat er in een groep van 30 personen twee individuen op dezelfde dag jarig zijn? Of, om een voorbeeld te geven wat beter past bij het aantal mensen wat hier in de zaal zit: Hoe groot is de kans dat in een groep van 140 mensen, twee mensen dezelfde 4-cijferige PIN-code hebben? Probeert u voor uzelf te bedenken hoe groot die kansen zouden zijn. Hoe groot zou dus, eerst maar eens, de kans zijn dat in een groep van 30 personen er twee op dezelfde dag jarig zijn.

Deze vraag legde ik laatst ook aan mijn klas van 30 psychologiestudenten voor. Van hen dachten er 4 dat die kans tussen de 20 en 40% lag, en de anderen dachten dat die kans kleiner was dan 20%. Niemand dacht dat die kans groter was dan 40%, terwijl het juiste antwoord is dat die kans maar liefst 71% is. Toen ik hierover voor het eerst las kon ik het, zelfs nadat ik het had nagerekend, niet meteen geloven. Pas nadat ik met de computer een flink aantal keren een steekproef van 30 willekeurige dagen uit het jaar had getrokken, en ik inderdaad zag dat er in heel wat steekproeven van 30 willekeurige dagen minstens twee zelfde dagen voorkwamen was ik overtuigd.[12]

---

[8] Te denken valt aan situaties waarin men "de hele populatie" onderzoekt (bij voorbeeld bij praktijkonderzoek waarin een compleet bedrijf wordt doorgelicht). Overigens blijft ook in zulke situaties veelal een belangrijke taak voor de (beschrijvende) statistiek weggelegd in het aanbrengen van overzicht in de complexiteit, zie bij voorbeeld Hand (1999).

[9] Aldus worden toevalsprocessen door onderzoekers zelf gecreëerd; inferentiële statistiek is vooral bedoeld om de grootte van de fluctuaties tussen willekeurige steekproeftrekkingen in te schatten.

[10] Zoals bij voorbeeld wordt ondersteund door het werk van de groep rond Tversky en Kahneman, zijn mensen niet bepaald goede *intuitive statisticians*.

[11] Over dit, inmiddels mede dankzij de NWO wetenschapsquiz 2000 vrij bekende voorbeeld, las ik voor het eerst in het boek *Het laatste raadsel van Fermat* van Simon Singh.

[12] Ook de studenten wilden dit niet meteen geloven. Op hun uitdrukkelijk verzoek heb ik toen in deze groep van dertig studenten alle verjaardagen opgevraagd en bleken er - tot mijn niet geringe vreugde, en tot hun verrassing – inderdaad twee studenten op dezelfde dag jarig te zijn. Aangezien het echter gaat om een kans van duidelijk minder dan 100% zal uiteraard niet elke docent het geluk hebben om deze uitkomst zo overtuigend te illustreren.

En hoe zit het dan met die pincodes. Nou, de kans dat in een groep van 140 mensen twee mensen dezelfde pincode hebben is 62%, bij een groep van 200 is dat 87%, en in een groep van 400, dat wil zeggen wat wij tegenwoordig ongeveer in onze collegezalen aantreffen, is die kans 99.97%. Kortom, in zo'n collegezaal zitten geheid twee studenten met dezelfde pincode, wat maar weer eens illustreert dat zelfs een personal identification number helemaal niet zo persoonlijk is.
We zien dus dat het werken met kansen niet altijd even gemakkelijk is. Nou zullen de kanskapitalisten dit opvatten als een pleidooi voor verdergaand en intensiever onderwijs in de kansrekening, maar ik zou hier op een andere kant willen wijzen. Ik denk dat het niet laat zien dat mensen niet goed kansen kunnen *uitrekenen*, want ook degenen die wel weten hoe je zo'n kans uitrekent schatten die kansen verkeerd in. Ik was zelf bij voorbeeld na al heel wat voorbeelden te hebben doorgerekend toch nog totaal verrast door de kans van 99.97% bij 400 pincodes. Ik denk dat het eerder laat zien dat het voor mensen lastig is zich een voorstelling te maken van de gebeurtenissen waar men het bij dergelijke kansuitspraken eigenlijk over heeft[13], en als dat te lastig wordt zal dat leiden tot een verkeerd gebruik van kansuitspraken. Zo zouden mensen nu kunnen denken dat er 71% kans is dat, als ze zelf in een groep van dertig zitten, er iemand op dezelfde dag jarig is als hij of zijzelf, of dat er 62% kans is dat een van de hier aanwezigen dezelfde pincode heeft als uzelf. Dit klopt niet. De kansuitspraken die die 71% en 62% opleverden waren net een stapje ingewikkelder.

Het gevaar van moeilijk te begrijpen kansuitspraken

De gangbare statistische methode van de *nulhypothesetoetsing* berust op een ongeveer even complex soort kansuitspraken, en leidt dan ook tot veel misverstanden en verkeerd gebruik. Stel men vindt dat het verschil in gemiddelde stressbestendigheid bij mannen en vrouwen 1.2 is op een schaal van 5. Bij nulhypothesetoetsing rekent men dan de kans uit dat, *als in werkelijkheid de gemiddelde stressbestendigheid van mannen en vrouwen precies gelijk zou zijn*, men in de steekproef een verschil zou vinden van 1.2 of zelfs meer. Als die kans dan kleiner is dan 5% redeneert men vervolgens:

- Als er in werkelijkheid geen verschil was, zou de kans heel klein zijn dat ik toevallig zo'n groot verschil vond als ik nu vond.
- Ik vind zo'n groot verschil, dus lijkt het me niet aannemelijk dat mannen en vrouwen in werkelijkheid precies dezelfde gemiddelde stressbestendigheid hebben.
- Daarom lijkt het me aannemelijk dat mannen in werkelijkheid gemiddeld stressbestendiger zijn dan vrouwen.

Ik heb nu een kansuitspraak gebruikt, maar wat moet ik me nu concreet bij die kans die kleiner was dan 5% voorstellen? Dat is bijna net zo moeilijk als de kans op 2 gelijke verjaardagen in een groep van 30. In een onderzoek van Sietske Nicolai[14] aan onze eigen vakgroep bleken maar liefst 31 van de 43 deelnemers (stafleden en promovendi) die kans verkeerd te interpreteren. En zij vroeg de onderzoekers dan nog expliciet naar wat die kans betekent. De werkwijze in de praktijk suggereert een nog ernstiger misinterpretatie. Men hanteert namelijk al bijna 80 jaar de ongeschreven wet dat, wanneer de bovengenoemde kans kleiner is dan 5%, het onderzoeksresultaat "statistisch significant" wordt genoemd. Dit maakt het werken met dergelijke toetsingsprocedures een stuk eenvoudiger, maar ik ben bang dat

[13] Ik denk dat men deze kansuitspraak vaak te veel op zichzelf betrekt en dat men dus, zelfs als de formulering overduidelijk is, denkt dat het gaat om 'zoiets als' de kans dat iemand op dezelfde dag jarig is als men zelf, en dat men, alleen al uit ervaring, weet dat die kans vrij klein is. Die inschatting is ook juist want de kans hierop is slechts 8%.
[14] A.S. Nicolai (1999) *Regels en communicatie in wetenschappelijk onderzoek*. Verslag leeronderzoek, Vakgroep Psychologie, RUG, Groningen.

veel mensen die lezen dat het gevonden verschil in stressbestendigheid "statistisch significant" is, ten onrechte concluderen dat dan *bewezen* is dat mannen stressbestendiger zijn dan vrouwen[15]. Alsof alle onzekerheid omtrent het onderzoeksresultaat in één klap verdwenen is. Alsof de onzekerheid is afgekocht nu één keer die specifieke kans is uitgerekend.
Ik denk dat dit misverstand in belangrijke mate veroorzaakt is door de erg ongelukkig gekozen term "significantie". "Significant" in het dagelijks spraakgebruik wordt opgevat als zinvol, belangrijk, maar een "statistisch significant" resultaat hoeft nog helemaal niet inhoudelijk belangrijk te zijn. Bij grote steekproeven zijn zelfs heel kleine, inhoudelijk onbelangrijke verschillen al "statistisch significant". Evenmin is een zogenaamd "significant" resultaat een "zeker" resultaat. De onzekerheid in de uitkomst is door de nulhypothesetoetsing met de binaire uitkomst wel/niet significant prettig onder de mat geveegd, maar dit is natuurlijk wel domweg een vorm van misleiding.
Nu ligt het natuurlijk voor de hand om te stellen dat dit niet het probleem van de statisticus is, die immers keurig volgens de theorie handelt en drommels goed weet wat de statistische uitspraken betekenen. Als de praktijk-onderzoeker niet goed weet hoe de statistische redenaties precies in z'n werk gaan en daardoor redeneerfouten gaat maken, dient de praktijk-onderzoeker er toe aangezet te worden zich beter te bekwamen in de statistiek. Ik zou het daarmee eens zijn als nulhypothesetoetsing de enig zinvolle statistische procedure zou zijn. Gelukkig zijn er echter ook andere manieren om aan te geven in hoeverre een onderzoeksuitkomst last kan hebben van toevalsfluctuaties, en *hoeft* men dus helemaal niet te werken met de conceptueel lastige nulhypothesetoetsing en de misleidende significantie-terminologie.
Met name handig is het gebruik van betrouwbaarheidsintervallen, waarmee rond een gevonden waarde, bij voorbeeld het eerdergenoemde verschil in stressbestendigheid van 1.2, een interval, bij voorbeeld lopend van 0.8 tot 1.6 wordt geconstrueerd waarvan men mag verwachten dat dit het echte verschil in gemiddelden vast wel zal bevatten. Hier staat "vast wel" dan bij voorbeeld voor een kans van 95%. Een belangrijk voordeel van deze weergave boven die van de uitkomst van een significantietoetsing is dat elke suggestie van een onfeilbare uitkomst weggenomen is, en dat de onnauwkeurigheid zelfs juist is geëxpliciteerd, en de grootte ervan is aangegeven. Het gevaar van het misbruik van de nulhypothesetoetsing, namelijk dat *significantie* voor "wel effect" en *niet-significantie* voor "geen effect" wordt aangezien is hier geheel en al afwezig.
Wat verder bijzonder prettig is, is dat men helemaal niets verliest ten opzichte van de nulhypothesetoetsing: De uitslag van zo'n toets, als iemand die om wat voor reden dan ook toch zou willen doen, is onmiddellijk af te leiden uit zo'n betrouwbaarheidsinterval. We verliezen niets, maar we winnen inzicht in de onnauwkeurigheid van onze uitspraken wanneer we betrouwbaarheidsintervallen gebruiken[16]. Daarmee is er dus geen enkele reden meer om bij een onderzoek de uitslag van een nulhypothesetoetsing te vermelden, terwijl er wel een belangrijk bezwaar tegen is (namelijk het grote risico op misinterpretatie)[17]. Desondanks

[15] Ook in statistische leerboeken spreekt men soms over "statistisch bewijzen" (zie bij voorbeeld Van Peet et al., 1997, p.11).
[16] Nog algemener is het werken met puntschatting en bijbehorende schatting van de standaardfout, op basis waarvan elk gewenst betrouwbaarheidsinterval kan worden opgesteld door de lezer.
[17] Dit geldt niet alleen voor eenvoudige situaties, zoals vergelijking van twee gemiddelden, maar ook voor de complexere van regressie-analyse en variantie-analyse. In feite worden hiermee het werken met oninzichtelijke toetsingsgrootheden als F-waarden vervangen door het werken met de beschrijvende parameters waarin men eigenlijk is geïnteresseerd, namelijk regressiegewichten, gemiddelden, hoofd- en interactie-effecten en verklaarde varianties, waar rondom dan weer betrouwbaarheidsintervallen kunnen worden aangegeven.

wordt in de praktijk de nulhypothesetoetsing nog veel gebezigd[18]. Een recente Task force on Statistical Inference (Wilkonson et al., 1999) heeft nu echter verscherpte richtlijnen voor de American Psychological Association voorgesteld, waarin het rapporteren van effect-groottes en interval-schattingen voor "primary outcomes" verplicht gesteld wordt (p.599).[19]

Politiek-maatschappelijke gevolgen

Verkeerd gebruik van statistiek, wat dus vooral in de hand gewerkt wordt door het gebruik van significantie-uitspraken, kan ook de politiek-maatschappelijke discussie bezoedelen. Zo rapporteerde de Sunday Telegraph in 1997: "Meeroken veroorzaakt geen kanker – nu officieel." Dit was volgens de krant gebaseerd op een rapport van de Wereld Gezondheids Organisatie. De krant had dit nieuws toegespeeld gekregen door de tabaksindustrie, die alleen de uitkomst "geen significant effect" uit het rapport aan de krant doorbriefde. Het rapport zelf vermeldde echter dat uit het onderzoek was gekomen dat er onder meerokende echtgenoten en meerokende werknemers een ongeveer 16% verhoogde kans is op het krijgen van longkanker. Maar, vermeldde het rapport ook, de aantallen kankerpatiënten waren te klein om dit percentage als statistisch significant te kunnen bestempelen. En dit laatste gebrek aan significantie werd dus door de tabaksindustrie aangegrepen als statistisch bewijs dat er geen effect was[20]. Het gebruik van nulhypothesetoetsing werkt dit soort misleiding in de hand. Dit had niet zo gemakkelijk plaats kunnen vinden wanneer zowel wetenschappers als "gebruikers" gewend zouden zijn om te denken in onderzoeksuitkomsten en onnauwkeurigheidsmarges.

Een algemene aanpak voor onnauwkeurigheidsanalyse

[18] Dit gebeurt ondanks het feit dat bezwaren tegen nulhypothesetoetsing al vele decennia door diverse auteurs worden verkondigd, zoals Berkson (1938, 1942), Rozeboom (1960), Meehl (1967), Carver (1978), Guttman (1985), Cohen (1994), zie ook Harlow, Mulaik & Steiger (1997). Waarom deze bezwaren nog maar nauwelijks gevolgen hebben gehad voor de werkwijze in de praktijk, laat zich alleen maar raden. Schmidt suggereert dat dit komt doordat de "typical researcher" wellicht de volgende (stuk voor stuk onjuiste) gedachten zou hebben (p.126): "Significance tests have been repeatedly criticized by methodological specialists, but I find them very useful in interpreting my research data, and I have no intention of giving them up. If my findings are not significant, then I know that they probably just occurred by chance and that the true difference is probably zero. If the result is significant, then I know I have a reliable finding. The *p* values from the significance tests tell me whether the relationships in my data are large enough to be important or not. I can also determine from the *p* value what the chances are that these findings would replicate if I conducted a new study. These are very valuable things for a researcher to know. I wish the critics of significance testing would recognize this fact."

[19] Niet alleen de significantie-uitspraken resulterend uit nulhypothesetoetsing worden vaak verkeerd begrepen. Ook het werken met het begrip *power* of *onderscheidingsvermogen* is een bron van verwarring, omdat het in feite gebaseerd is op twee geneste conditionele uitspraken. Namelijk de uitspraak "Als de nulhypothese waar zou zijn, dan zou de kans op het vinden van een uitkomst groter dan een bepaalde kritieke waarde .05 zijn, en die kritieke waarde is dan *x*", en de uitspraak "Als de werkelijke waarde μ zou zijn, dan is de power de kans op het vinden van een waarde groter dan de bij de nulhypothese vastgelegde kritieke waarde". Het begrip power is echter alleen maar van nut bij nulhypothesetoetsing, en dan met name om tegemoet te komen aan één van de bezwaren daarvan, namelijk, het bezwaar dat niet-significantie lang niet altijd betekent dat er geen of hooguit een heel klein effect is. Zodra echter de nulhypothesetoetsing vervangen wordt door betrouwbaarheidsintervallen is het begrip power overbodig geworden. Daarvoor in de plaats kan men gaan letten op de intervalbreedte, waarbij geldt dat hoe smaller het interval, des te nauwkeuriger het resultaat. Net als men door de keuze van de steekproefgrootte de power vooraf kan instellen, kan men ook door de keuze van de steekproefgrootte vooraf de intervalbreedte, en dus de nauwkeurigheid van het resultaat instellen. Alweer is het mooie dat we niets verliezen, maar veel winnen.

[20] De hier vermelde gebeurtenissen vormen een vereenvoudigde weerslag van het artikel *Significant meeroken* van G. Feenstra in zijn column *Kantlijn* in *de Volkskrant*, 1997.

Statistiek begint met data-vereenvoudiging en vervolgt dan met onnauwkeurigheidsanalyse. Als die vereenvoudiging het berekenen van gemiddelden is, is het berekenen van een betrouwbaarheidsinterval een eenvoudig klassiek statistisch probleem. Maar wat te doen als onze vereenvoudiging niet simpelweg tot gemiddelden leidt. Stel dat we, in plaats van het gemiddelde de mediaan willen gebruiken, of dat onze vereenvoudiging uit de uitkomsten van een drieweg-analyse bestaat. Ook dan willen we inzicht in de onnauwkeurigheid van onze resultaten, want we willen antwoord op de vraag "Wat zou er uitkomen als ik het onderzoek overnieuw zou doen?", of, "als ik een net wat andere steekproef had". Om een antwoord op deze vraag te krijgen kunnen we *altijd* gebruik maken van de prachtige, en heel eenvoudig te begrijpen bootstrap-procedure (Efron, 1979, zie ook Efron & Tibshirani, 1993). Ruwweg komt deze erop neer dat we bekijken hoeveel onze resultaten fluctueren wanneer we de gegevens in onze steekproef willekeurig vervangen door andere gegevens uit onze steekproef. De fluctuaties die we dan aantreffen vormen een basis op grond waarvan we, rond wat voor uitkomsten dan ook maar, betrouwbaarheidsintervallen kunnen bepalen.
Naast de bootstrap-procedure zijn er nog andere methoden om inzicht te geven in de algemene bruikbaarheid van resultaten uit onderzoek. Te denken valt daarbij aan kruisvalidatie-methoden, die expliciet de vraag beantwoorden in hoeverre een *beschrijving* op grond van een bepaalde steekproef, ook algemener toepasbaar is, en dus ook op gegevens uit een andere steekproef. Hoe dit dan het best kan worden gedaan voor drieweg-analyse is een nog deels open liggende vraag (zie ook Kiers & van Mechelen, 2001)[21]

Vrijheid en Verantwoordelijkheid

Wetenschappers zijn gewend te discussiëren. U ziet dat ook aan de afbeelding hier achter mij, die u ter afleiding wellicht al uitvoerig hebt bestudeerd. Hier ziet u dat men met elkaar samenwerkt, maar ook onenigheid heeft, zoals hier rechts van mij de met stuurse blik van elkaar wegkijkende personen. Gek genoeg lijken veel onderzoekers echter niet te willen discussiëren over de keuze van statistische technieken, alsof ze die keuzevrijheid helemaal niet willen hebben. Het komt veelvuldig voor dat men vraagt of je een bepaalde procedure wel "mag" toepassen (alsof de statisticus dat voor het zeggen heeft), hoeveel proefpersonen je minstens "moet" hebben, of hoe je "moet" beslissen hoeveel factoren je bij een factor analyse gebruikt. Misschien heeft dit te grote geloof in de autoriteit van de statisticus te maken met het feit dat men statistiek ziet als een methode om de, vooral in de sociale wetenschappen niet bepaald harde resultaten, hard te maken (zie Porter, 1984, p.393), en daar dan ook harde regels van verwacht. Echter, ook in de statistiek geldt Sartre's (1965) paradoxale imperatief: "U bent vrij, kies!"[22]. Even Sartreaans is de opmerking dat men ook verantwoordelijk is voor de eigen keuzen. Vrijheid brengt verantwoordelijkheid met zich mee. Kortom, het goede nieuws wat ik hier wil uitdragen is dat men in de keuze van te gebruiken data-vereenvoudingingstechnieken geheel vrij is, mits deze gevolgd wordt door een degelijke verantwoording, waarin ook aandacht is voor een weergave van de onzekerheden in de uitkomsten. Het mooie is dat dit laatste ook *altijd* kan, want mocht er voor de onnauwkeurigheidsanalyse rond de uitkomsten geen standaardprocedure beschikbaar zijn, dan kan men altijd nog van bootstrap-procedures gebruik maken.
En wat betekent dit alles voor het onderwijs?

[21] In de context van componenten-analyse is kruisvalidatie eerder voorgesteld door Ten Berge (1986).
[22] "…vous êtes libre, chosissez, c'est-à-dire inventez. Aucune morale générale ne peut vous indiquer ce qu'il y a à faire; il n'y a pas de signe dans le monde. Les catholiques répondront: mail il y a des signes. Admettons-le; c'est moi-même en tout cas qui choisis le sens qu'ils ont." (p.46-47)

Toen ik bij een eerdere gelegenheid in onze vakgroep een soortgelijk betoog hield kreeg ik de vraag of dit dan ook geen consequenties voor het onderwijs zou moeten hebben. Op dat moment had ik daar geen beter antwoord op dan, tja, dat zou eigenlijk wel moeten, maar ik zou niet weten hoe. Inmiddels heb ik daar andere ideeën over, omdat me nog duidelijker is geworden wat het potentiële gevaar is van het gebruik van nulhypothesetoetsing, en meer in het algemeen van het gedachteloos toepassen van standaard statistische procedures[23]. Daarom vind ik dat studenten al vanaf het begin vooral moeten leren statistisch te *denken*, en moet het aanleren van statistische standaard procedures op het tweede plan komen. Met statistisch denken bedoel ik dan het nadenken over wat voor data-vereenvoudiging er gepleegd zou kunnen worden, en over onzekerheid in onderzoeksuitkomsten.
Dat vereist echter wel een actieve inzet van de student, waar het momenteel nogal eens aan lijkt te ontbreken. Niet dat ik denk dat studenten lui zijn, integendeel, ze zijn juist enorm actief op velerlei terrein, hebben meer dan ooit bijbaantjes, en onderhouden een communicatief netwerk via internet en gsm wat niet onderdoet voor het netwerk van menig klein bedrijf enkele decennia geleden. Ik denk echter dat het nodig en mogelijk is om de studenten, in deze tijd vol distractors, actief aan te zetten tot kritisch nadenken over de analyse van onderzoeksuitkomsten, door ze daartoe zowel via de inhoud als de opzet van het onderwijs beter toe te motiveren. Daartoe kunnen en moeten we aansluiten bij hun intrinsieke nieuwsgierigheid naar het “hoe” en “waarom” van gedragingen van mensen.
Dat dergelijke inhoudelijke motivaties inderdaad aan kunnen zetten tot nadenken over methodologische kanten van onderzoek bleek laatst nog maar weer eens in reactie op de recente uitspraken van burgemeester Wallage over de criminaliteit van asielzoekers. Na inzage in het onderzoeksrapport waarop Wallage zijn uitspraken baseerde, werden door menigeen methodologische vraagtekens geplaatst bij de juistheid van de conclusies in het rapport. Zoals we in het begin van dit cursusjaar al hebben gemerkt, kunnen we die kritische houding bij *studenten* aanspreken via de beantwoording van *eigen* onderzoeksvraagstellingen en onderzoeksgegevens. Juist wanneer de student direct geconfronteerd wordt met onderzoeksgegevens, kan deze zich serieus de vraag stellen wat hier nu wel en niet uit te concluderen valt, en zal de student ook duidelijk worden dat statistische technieken meer zijn dan een noodzakelijkerwijs uit te voeren ritueel.
In de toekomst zullen we de studenten veel intensiever moeten aanzetten tot “statistisch denken” dan we nu doen, door in de studie meer kritische discussies te arrangeren onder leiding van docenten, aio’s, en ouderejaarsstudenten, zodat studenten leren omgaan met zowel de statistische keuzevrijheid als de daaraan gekoppelde verantwoording. Ter ondersteuning daarvan dienen individuele en groepsgewijze opdrachten buiten de bijeenkomsten om te worden uitgevoerd, zodat de student actief bezig gaat met onderzoeksvraagstellingen en het statistisch rapporteren van uitkomsten. Ik ben er van overtuigd dat, door het statistiekonderwijs te intensiveren en het te koppelen aan motiverende onderzoeksvraag-stellingen, studenten daadwerkelijk kunnen leren statistisch te denken en dit ook niet zo gemakkelijk zullen vergeten. Ik heb niet de illusie dat dit voor alle studenten zal gelden. Studenten zonder intrinsieke belangstelling voor wat ze studeren, zullen op geen enkele manier kunnen worden aangezet tot actief studeren, maar voor hen zal dan ook het verblijf aan de opleiding van korte duur zijn, wanneer de studie dan strengere eisen zal stellen in de vorm van vervulling van opdrachten zowel binnen als buiten de lessen.

[23] Dit gedachteloos toepassen wordt ook duidelijk uit de column van psychologie-studente Eefje Boerhave de Groot, Universiteitskrant, 12 oktober 2000: “… die bergen statistiek die wij – arme psychologen –moeten verzetten, maar veel indruk maak ik er niet mee. Terecht, want die vakken kan je halen door volgordes van handelen uit je hoofd te leren, in plaats van het te snappen.”

Voorts wordt het ook hoog tijd dat het onderwijs de moderne ontwikkelingen in de statistiek volgt, zeker als het gaat om zulke krachtige procedures als de bootstrap-procedure. Tom Snijders merkte onlangs bij één van zijn "Zin in statistiek bijeenkomsten" op dat alle statistiek die wij (niet alleen hier, maar bijna wereldwijd) in de basis onderwijzen is ontwikkeld vóór de oorlog. Schmidt (1996, p.128) verwoordt het nog wat krasser. Naar aanleiding van de juist ook in het gangbare onderwijs centraal staande nulhypothesetoetsing zegt hij: "We are teaching them the same methods that for over 40 years made it impossible to discern the real meaning of data and research literatures and have therefore retarded the development of cumulative knowledge[24] in psychology and the social sciences". U zult inmiddels hebben begrepen dat ik daarin, als verantwoordelijke voor het eerstejaars statistiekonderwijs, graag verandering wil brengen.

Tot slot

Bij het doen van wetenschappelijke uitspraken past een flinke dosis bescheidenheid. Wij allemaal, wetenschappers en niet-wetenschappers geven maar al te gauw toe aan de verleiding om uitspraken stelliger te doen dan we waar kunnen maken, en onze onzekerheden niet te erkennen, soms zo erg dat we dat niet eens meer doorhebben. "Ik hoop echter dat de psychologie in de éénentwintigste eeuw, met gepaste trots op haar verworvenheden, ook volmondig uitkomt voor de onzekerheidsmarges in alle uitspraken die zij doet, en niet langer genoegen neemt met de schijnnauwkeurigheid van ongenuanceerde getalsweergaven of hypothesetoetsingen" (Kiers, 2000).

Aan het slot van mijn rede gekomen wil ik nog graag enige dankwoorden uitspreken. Allereerst ben ik het college van bestuur van deze universiteit, het bestuur van de Stichting Groninger Universiteitsfonds, en het bestuur van de faculteit PPSW zeer erkentelijk voor het instellen van deze bijzondere leerstoel, en het in mij gestelde vertrouwen. Het is wereldwijd de eerste leerstoel die is ingesteld op het snel in belangstelling groeiende gebied van de analyse van meerweggegevens.
Vervolgens wil ik mijn leermeesters bedanken. Dat zijn er natuurlijk te veel om op te noemen, maar enkele wil ik hier toch in het bijzonder noemen. Waarde ten Berge, beste Jos. Ik weet dat je het volstrekt belachelijk vindt dat ik je zo formeel toespreek, omdat je niets moet hebben van dikdoenerij, maar je zult er maar even aan moeten geloven. Ik heb heel veel te danken aan jouw begeleiding van mij als student en later promovendus, waarin je kritische toetsing van mijn ideeën een nauwkeurige en heldere, van alle overdaad ontdane formulering van mijn ideeën afdwong. Met mijn ongeduldige aard zou ik daar misschien niet mee om hebben kunnen gaan als daar niet ook een enorme dosis vertrouwenwekkende ondersteuning en hulpbereidheid van jou tegenover hadden gestaan. Je houding ten aanzien van ondersteuning en begeleiding strekt mij als voorbeeld, ook al heb ik niet de illusie even hulpvaardig en enthousiasmerend als jij te kunnen zijn. Nu ik recentelijk door onderwijstaken en door mij aangegane externe samenwerkingsverbanden, niet of nauwelijks meer met je samenwerk, kan ik al gauw overkomen als een ondankbaar kind. Kind ben ik volgens de

[24] De redenatie hierachter loopt ongeveer als volgt. Bij een eerste onderzoek is het wel prettig om alleen maar te horen dat men wel of geen significant effect heeft gevonden, maar wanneer onderzoek wordt herhaald zal men vaak conflicterende uitkomsten aantreffen, want soms wordt er wel en soms wordt er geen significant effect gevonden. Dit werkt demotiverend, want het lijkt te impliceren dat, hoe meer we onderzoeken hoe minder we weten. Gelukkig is dat onzin: door resultaten te combineren, via meta-analyse, kom je wel degelijk verder wanneer er meer onderzoek naar een bepaald fenomeen is gedaan. Daarvoor is echter wel nodig dat we beschikken over betrouwbaarheidsintervallen of soortgelijke informatie, want het combineren van alleen maar significantie-resultaten is niet genoeg, en kan zelfs tot totaal verkeerde conclusies luiden.

ouderen onder u nog steeds, maar ondankbaar zeker niet. Jos, ik hoop dat we nog veel met elkaar over het vak en anderszins kunnen praten, en elkaar kunnen bekritiseren en enthousiasmeren.
Twee andere leermeesters vallen sinds kort onder de emiriti. Waarde Hofstee, beste Wim, je grootste invloed heb je op me gehad in de periode dat je me nog niet kende, toen ik als student je eerstejaarscolleges persoonlijkheidspsychologie volgde. Je betogen waren zeer uitdagend, in eerste instantie omdat ze ons ertoe dwongen op verstandelijk gebied alle zeilen bij te zetten, hetgeen veelal nog niet genoeg was, en in tweede instantie omdat ze dwarse en intrigerende ideeën uit de doeken deden. Met de rationale achter je weddenschapsmodel heb je mijn ideeën over statistiek direct of indirect sterk gevormd. Waarde Molenaar, beste Ivo. Jij hebt me geleerd dat statistiek vooral ook een toegepast vak is, en dat toepassing vooral ook gebaat is bij een flinke dosis gezond verstand. Je gevleugelde woorden “significant is niet relevant” en “wat gebeurt er als ik het overnieuw doe” zal ik nooit vergeten en de portee ervan draag ik nog keer op keer uit, zoals vandaag hier in de Aula, maar ook in eerstejaars en postacademisch onderwijs. Wat dat laatste betreft ben ik blij nog steeds met je samen te werken in de door jou opgezette postacademische cursus “Statistiek in Vogelvlucht”.
Ook dank verschuldigd ben ik aan de studenten en aio’s die ik onderwijs heb gegeven. Bij het klassikale onderwijs kan ik vooral genieten van jullie inzet, aandacht en leergierigheid. Ook de door mij individueel begeleide studenten en aio’s ben ik dankbaar voor hun inzet en enthousiasme, en voor het soms in een heel andere richting uitwerken van mijn oorspronkelijke ideeën. Marieke, ik ben heel blij met je fraaie proefschrift, wat nu, ruim binnen de ervoor gestelde tijd, naar de leescommissie gaat. Aangezien je in zekere zin mijn eerste aio was, ben je onbedoeld in een experiment beland. Ik heb veel van je onderzoek en onze discussies daarover geleerd.
Beste leden van de vakgroep Psychologie Algemeen. Al zijn we een zeer diverse groep, de collegialiteit en loyaliteit die we naar elkaar hebben is voor mij zeer waardevol. Vooral ook het teamwork aan het eerstejaars onderwijs met Frits, Ingrid, Marijtje, Marieke, Christian en Marieke waardeer ik zeer, en de vrolijke bijeenkomsten hierover verluchtigen het niet altijd even bemoedigende werk.
Dankzij de nodige afleiding, en de mogelijkheid om te klagen over oneerlijke procedures of andere problemen, vind ik ook buiten m’n werk veel steun voor op m’n werk. Ik ben dan ook m’n familie en vrienden dankbaar voor wat ze me op dit terrein bieden, ook al moeten sommigen onder hen me zo nodig elke zondag weer met tennis verslaan (dit was geschreven voor afgelopen zondag). Het meest dankbaar ben ik natuurlijk Jeanine. Ik zal niet zeggen dat ik hier zonder jou niet zou staan, en ik vind het ook wat goedkoop om nu te gaan zeggen dat het me spijt dat ik er de laatste tijd zo weinig voor je was. Ik wil wel zeggen dat ik je heel dankbaar ben dat je me op een manier hebt leren leven die buiten maar ook binnen m’n werk heel belangrijk voor me is, ook al bleek toen je de eennalaatste versie van deze oratie bekritiseerde dat ik nogal hardleers ben. Ik hoop dat we nog heel wat samen kunnen genieten, te beginnen met de receptie die zo meteen beneden plaats vindt.
Ik dank u allen heel hartelijk voor uw belangstelling en hoop uw geduld niet te lang op de proef te hebben gesteld. Ik raad u dan ook aan niet allemaal in de rij te gaan staan, maar vooral ook heerlijk onbescheiden meteen op de hapjes en drankjes af te gaan.

Ik heb gezegd

Literatuur

Berkson, J. (1938) Some difficulties of interpretation encountered in the application of the chi-square test. *Journal of the American Statistical Association, 33,* 526-542.

Berkson, J. (1942) Tests of significance considered as evidence. *Journal of the American Statistical Association, 37,* 325-335.

Carver, R.P. (1978) The case against statistical significance testing. *Harvard Educational Review, 48,* 378-399.

Cooper, L.G., Klapper, D. & Inoue, A. (1996). Competitive-component analysis: A new approach to calibrating asymmetric market-share models. *Journal of Marketing Research, 33,* 224-238.

Cohen, J. (1994) The earth is round (p<.05). *American Psychologist, 49,* 997-1003.

Efron, B. (1979) Bootstrap methods: another look at the jackknife. *Annals of Statistics, 7,* 1-26.

Efron, B., & Tibshirani, R.J. (1993) *An introduction to the bootstrap.* London: Chapman Hall.

Goldberg, L.R. (1981) Language and individual differences: The search for universals in personality lexicons. In L. Wheeler (Ed.), *Review of personality and social psychology* (Vol. 2, pp.141-165). Beverly Hills, CA: Sage.

Guttman, L. (1985) The illogic of statistical inference for cumulative science. *Applied Stochastic Models And Data Analysis, 1,* 3-10.

Hand, D.J. (1999) Statistics and data mining: Intersecting disciplines, *4CM SIGKDD, 1,* 16-19.

Harlow, L.L., Mulaik, S.A., & Steiger, J.H. (Eds.) (1997) *What if there were no significance tests?* London: Lawrence Erlbaum.

Hofstee, W.K.B. (1996) *How to conceive of personality structure*. Internal note, RUG, Groningen.

Hofstee, W.K.B., ten Berge, J.M.F. & Hendriks, A.A.J. (1998) How to score questionnaires. *Personality and Individual Differences, 25,* 897-909.

Kiers, H.A.L. (2000) De beste wensen van ... . *De Psycholoog, 35, januari,* p. 23.

Kiers, H.A.L. (1998) Joint orthomax rotation of the core and component matrices resulting from three-mode principal components analysis. *Journal of Classification, 15,* 245-263.

Kiers, H.A.L., & van Mechelen, I. (2001) Three-way component analysis: Principles and illustrative application. *Psychological Methods, xx,* xxx-xxx.

Kiers, H.A.L., & Harshman, R.A. (1997) Relating two proposed methods for speedup of algorithms for fitting two- and three-way principal component and related multilinear models. *Chemometrics and Intelligent Laboratory Systems, 36,* 31-40.

Kiers, H.A.L., & Krijnen, W.P. (1991) An efficient algorithm for PARAFAC of three-way data with large numbers of observation units. *Psychometrika, 56,* 147-152.

Kiers, H.A.L., Kroonenberg, P.M., & Ten Berge, J.M.F. (1992) An efficient algorithm for TUCKALS3 with large numbers of observation units. *Psychometrika, 57,* 415-422.

Kroonenberg & de Leeuw (1980) Kroonenberg, P.M., & De Leeuw, J. (1980). Principal component analysis of three-mode data by means of alternating least squares algorithms. *Psychometrika, 45,* 69-97.

Meehl, P.E. (1967) Theory testing in psychology and physics: A methodological paradox. *Philosophy of Science, 34,* 103-115.

Porter, T. (1984). Making things quantitative, *Science in Context, 7,* 389-407.

Rozeboom, W.W. (1960) The fallacy of the null hypothesis significance test. *Psychological Bulletin, 57,* 416-428.

Sartre, J.-P. (1965) L'existentialisme est un humanisme. Paris: Les Éditions Nagel.

Schmidt, F.L. (1996) Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods, 2*, 115-129.

Ten Berge, J.M.F. (1986) Rotation to perfect congruence and the cross-validation of component weights across populations. *Multivariate Behavioral Research, 21*, 41-64; 262-266.

Tucker, L.R. (1966). Some mathematical notes on three-mode factor analysis. *Psychometrika, 31*, 279-311.

van Mechelen, I., & Kiers, H.A.L. (1999) Individual differences in anxiety responses to stressful situations: A three-mode component analysis model. *European Journal of Personality, 13*, 409-428.

van Peet, A.A.J., van den Wittenboer, G.L.H. & Hox, J.J. (1997) *Toegepaste Statistiek: Inductieve Technieken*. Groningen: Wolters-Noordhoff.

Wilkinson, L., & the Task Force on Statistical Inference (1999) Statistical methods in psychology journals. *American Psychologist, 54*, 594-604.