Rechtstreeks verwachtingen evalueren of de nulhypothese toetsen?

Nulhypothesetoetsing versus Bayesiaanse modelselectie

In veel psychologische artikelen wordt klassieke nulhypothesetoetsing (NHT) gebruikt om onderzoeksvragen te beantwoorden. De resultaten kunnen echter onbevredigend zijn. Rechtstreeks de verwachtingen evalueren zou beter zijn, maar is niet mogelijk met NHT. We laten zien wat de nadelen zijn van NHT en hoe het beter kan, namelijk met Bayesiaanse modelselectie die we introduceren voor niet-statistici.

Rens van de Schoot, Herbert Hoijtink en Sibe Doosje

Foto: Herman Wouters

Een praktiserend NIP-psycholoog houdt zijn of haar wetenschappelijke literatuur bij en blijft op de hoogte van recente ontwikkelingen op zijn of haar vakgebied. In dit artikel willen we praktiserende psychologen op de hoogte houden van een minder voor de hand liggende ontwikkeling, namelijk op het gebied van statistiek. In de wetenschappelijke literatuur die praktiserende psychologen vaak lezen, wordt door vrijwel alle onderzoekers klassieke nulhypothesetoetsing (NHT) gebruikt om antwoord te geven op de onderzoeksvraag. In dit artikel zullen we laten zien waarom NHT niet per se de beste keuze hoeft te zijn om de onderzoeksvraag te beantwoorden (zie ook het stuk van Eric-Jan Wagenmakers in *De Psycholoog* van juli/augustus 2008).

We laten zien wat de consequenties zijn als het mis gaat met NHT en introduceren vervolgens een recent ontwikkelde methode die een veelbelovend alternatief biedt voor NHT, namelijk Bayesiaanse modelselectie (BMS) (Hoijtink, Klugkist & Boelen, 2008; Klugkist, Laudy & Hoijtink, 2005; Van de Schoot et al., 2008). Aan de hand van een voorbeeld leggen we kort uit hoe BMS werkt en welke voordelen deze methode biedt ten opzichte van NHT. Om te laten zien hoe NHT 'faalt' en BMS beter werkt, presenteren we een relevant voorbeeld uit de arbeids- en gezondheidspsychologie. Op deze manier wordt duidelijk hoe BMS in de (wetenschappelijke) praktijk gebruikt kan worden.

Informatieve hypothesen

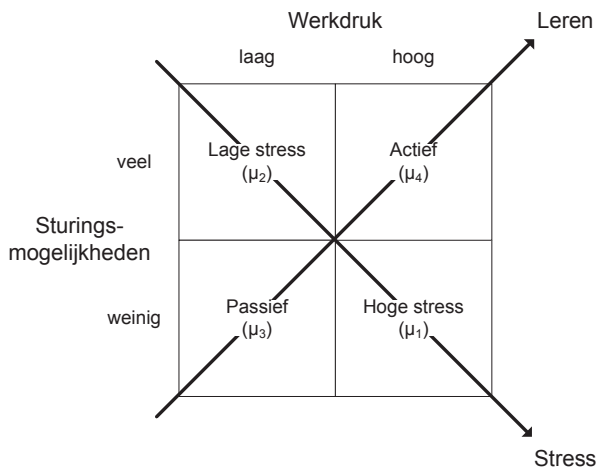
Onderzoekers hebben bepaalde verwachtingen over hoe de werkelijkheid er uit ziet. Verwachtingen en hypothesen kunnen gebaseerd zijn op eerder (literatuur)onderzoek, wetenschappelijk debat of zelfs (subjectieve) meningsverschillen. Het laatste kan bijvoorbeeld als de ene onderzoeker overtuigd is van het effect van een nieuwe interventie of nieuwe behandelmethode die nog niet eerder is onderzocht en een andere onderzoeker niet. Het is natuurlijk wel van belang dat alleen nuttige verwachtingen met elkaar worden vergeleken en niet alle mogelijke verwachtingen. Wij houden juist een pleidooi voor het rechtstreeks evalueren van de voorkennis die een onderzoeker heeft.

Het doel van veel onderzoekers is het evalueren van een aantal 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 wij *informatieve hypothesen* zullen noemen. Dit omdat er *a priori*, dat is voordat er data zijn verzameld, *informatie* bestaat. Bijvoorbeeld over de ordening tussen twee (of meer) groepsgemiddelden: $\mu_1 < \mu_2$, waarbij het teken '<' aangeeft dat het eerste gemiddelde (μ_1) lager is dan het tweede gemiddelde (μ_2). Wij zullen laten zien dat onderzoekers deze verwachtingen wel *willen* evalueren, maar dit niet zo maar *kunnen* doen. Het is namelijk vrijwel onmogelijk om met NHT complexe informatieve hypothesen te evalueren (zie ook: Van de Schoot et al., 2008). Als onderzoekers dit toch proberen omdat er geen alternatieven voorhanden zijn, dan ontstaan er enkele problemen die we nader zullen toelichten aan de hand van een voorbeeld.

Voorbeeld 1. Werkdruk, sturingsmogelijkheden en verkoudheid

In deze paragraaf presenteren we een voorbeeld dat we eerst met behulp van NHT evalueren en daarna met BMS. Karasek (1979) stelde dat de gezondheid van werknemers wordt bepaald door combinaties van de mate van werkdruk ('job demands') en de beschikbare sturingsmogelijkheden ('job control'). Daarnaast zijn er twee onderliggende mechanismen die van invloed zijn op de werkdruk en sturingsmogelijkheden, namelijk leren en stress, zie Figuur 1.

Karasek voorspelde dat met name een combinatie van hoge werkdruk en weinig sturingsmogelijkheden (een 'hoge stress'-werksituatie) het risico op gezondheidsklachten zou vergroten ten opzichte van een 'lage stress'-werksituatie' (lage werkdruk, veel sturingsmogelijkheden) en ten opzichte van een 'actieve' (hoge werkdruk, veel sturingsmogelijkheden) en een 'passieve' (lage werkdruk, weinig sturingsmogelijkheden) werksituatie. In Figuur 1 zijn de vier werksituaties weergegeven: *hoge stress*, *lage stress*, *passief*, en *actief*.



Figuur 1. Het interactie model van Karasek

Het interactiemodel van Karasek veronderstelt dat de hoge werkdruk een toestand van fysiologische opwinding teweegbrengt, bijvoorbeeld door een verhoogde hartslag en adrenalineproductie, die door de gebrekkige sturingsmogelijkheden niet kan worden omgezet in een effectieve copingrespons (Buunk, De Jonge, Ybema & Wolff, 1998).

Omdat er aanwijzingen zijn dat het interactiemodel een goede verklaring biedt van cardiovasculaire klachten (Schnall, Landsbergis & Baker, 1994), zouden we kunnen veronderstellen dat dit ook geldt voor andere ziektebeelden, zoals het risico om verkouden te worden. Karaseks interactiemodel is hiervoor echter niet geheel empirisch ondersteund omdat alleen een hoofdeffect van werkdruk (Hao, Duan & Zhang, 2002; Mohren, Swaen, Borm, Bast & Galama, 2001) of van sturingsmogelijkheden (Doosje, Goede, Doornen, Goldstein & Van de Schoot, 2008) werd gevonden. De empirische steun voor de interacties die Karasek veronderstelt, is daarom beperkt.

Onderzoekers hebben meer verwachtingen dan je met een nulhypothese kunt toetsen.

In het voorbeeld voor dit artikel gebruiken we de dataset beschreven in het artikel van Doosje en collega's (Doosje et al., 2008). De onderzoeksvraag is hoe de vier typen werksituaties die Karasek beschrijft (zie Figuur 1), verschillen met betrekking tot het aantal keren dat iemand verkouden is geweest in het afgelopen halfjaar. Daarover hebben we drie verwachtingen opgesteld naar aanleiding van eerder onderzoek.

Verwachting A. Vanuit de oorspronkelijke theorie van Karasek (1979) zouden we verwachten dat de groep *hoge stress* (μ_1) het ongezondst is ten opzichte van de groepen *lage stress* (μ_2), *passief* (μ_3) en *actief* (μ_4). De groep *hoge stress* heeft dan een hoger gemiddelde op het aantal keren verkouden zijn in het afgelopen halfjaar dan de overige drie groepen. De informatieve hypothese ziet er dan zo uit, waarbij '>' verwijst naar een hoger gemiddelde en dus vaker verkouden zijn en '=' naar een gelijk gemiddelde:
HA : $\mu_1 > \{\mu_2 = \mu_3 = \mu_4\}$

Verwachting B. Als werkdruk de meest relevante variabele is, zoals Mohren et al. (2001) en Hao et al. (2002) hebben gevonden, dan is een hoge werkdruk gerelateerd aan het ongezondst zijn en dus vaker verkouden zijn. De groepen *actief* (μ_4) en *hoge stress* (μ_1) zouden dan een hoger gemiddelde hebben dan de andere twee groepen:
HB: $\{\mu_1 = \mu_4\} > \{\mu_2 = \mu_3\}$

Verwachting C. Zoals is gesuggereerd door Doosje et al. (2008) spelen zowel werkdruk als sturingsmogelijkheden een rol bij verkouden worden. In dat geval zou de groep *hoge stress* (μ_1) een hoger gemiddelde hebben op de variabele verkouden zijn gevolgd door de groep *passief* (μ_3), gevolgd door de groep *actief* (μ_4). De groep *lage stress* (μ_2) zou dan het minst vaak verkouden zijn omdat zij weinig stress en veel sturingsmogelijkheden hebben. De bijbehorende hypothese ziet er dan zo uit:
HC: $\mu_1 > \mu_3 > \mu_4 > \mu_2$

Om erachter te komen welke van deze drie verwachtingen het meest waarschijnlijk is, is in het artikel van Doosje et al. (2008) een variantieanalyse (ANOVA) uitgevoerd. In Tabel 1 zijn de groepsgemiddelden weergegeven. Er bleken significante verschillen te bestaan tussen de vier groepen ($F(3) = 9.51$; $p < .001$) en uit een posthocanalyse met Bonferronicorrectie blijkt dat sommige maar niet alle groepen onderling van elkaar verschillen (zie Tabel 1). Als twee gemiddeldes dezelfde letter hebben, dan is het verschil significant. Zo hebben μ_1 en μ_2 beide de letter a en verschillen zij significant van elkaar ($p < .05$).

	Hoge stress (μ_1)	Lage stress (μ_2)	Passief (μ_3)	Actief (μ_4)
Gemiddelde	2.37^{ab}	2.17^{ac}	2.42^{cd}	2.18^{bd}
SD	1.03	0.93	0.98	0.97
n	289	594	517	638

Noot: Gemiddelden met dezelfde letter verschillen significant van elkaar ($p < .05$)

Tabel 1. Groepsgemiddelden en standaarddeviaties (SD) voor de vier groepen van het Karasekmodel

Voor de ANOVA is de volgende nulhypothese getoetst: Ho: $\mu_1 = \mu_2 = \mu_3 = \mu_4$. Er werden significante groepsverschillen gevonden en de nulhypothese is verworpen. Merk op dat dit tot nu toe nog steeds geen informatie geeft over welke van de informatieve hypothesen (HA; HB; HC) de beste is. Om hierover toch een uitspraak te doen, kan gekeken worden naar de ordening van de gemiddelden uit Tabel 1:
$\mu_3 > \mu_1 > \mu_4 > \mu_2$

Er zijn ook post-hoctoetsen uitgevoerd en als we bij niet significant resultaat de ordening van de gemiddeldes aanpassen, dan wordt de ordening:
$\{\mu_3 = \mu_1\} > \{\mu_4 = \mu_2\}$

Het is nu op basis van dit resultaat erg lastig om te kiezen tussen de drie informatieve hypothesen HA, HB, en HC. Geen van de hypothesen wordt namelijk volledig ondersteund door de gevonden ordening van de data. De resultaten geven slechts in meer of mindere mate steun voor elk van de informatieve hypothesen. Voor HA geldt dat μ_1 groter is dan μ_2 en μ_4, maar niet groter is dan μ_3. Alleen de gemiddeldes μ_1 en μ_4 voor HB zijn groter dan μ_2, maar dit geldt niet voor μ_3. Beide hypothesen worden dus niet echt ondersteund door de data. Hypothese C komt echter dicht in de buurt, zeker als gekeken wordt naar de ordening op basis van de groepsgemiddelden. Als echter naar de significante resultaten

Bayes Factors

De berekening van de PMK's (posterior modelkansen) geschiedt aan de hand van Bayes Factors, uitgevonden door Thomas Bayes in 1764 (Bayes, 1764) en is verder ontwikkeld in 1774 door Laplace (zie Laplace's 1774 *Memoir on inverse probability*, in Stigler, 1986). Pas in de twintigste eeuw werd de Bayesiaanse benadering opnieuw ontdekt, door onder anderen Ramsey, De Finetti, Jeffreys en Jaynes (voor een overzicht zie Corfield & Williamson, 2001). Pas in de jaren negentig werden computers snel genoeg om de berekeningen ook daadwerkelijk uit te voeren (zie bijvoorbeeld Bayarri & Berger, 2000; Kass & Raftery, 1995; Raftery, 1995). Het omzetten van de drie ingrediënten in Bayes Factors en daarna in PMK's, wordt uitgebreid beschreven in het boek van Hoijtink et al. (2008).

Bij de nulhypothese toets je: 'niks aan de hand' versus 'er gebeurt iets, maar we weten niet wat'.

wordt gekeken, dan klopt ook deze hypothese niet meer.

Hypothese C lijkt dus op het eerste gezicht de meeste steun te krijgen, maar de vraag is of dit ook werkelijk zo is. Het is nog lastiger, of zelfs onmogelijk om te zeggen *hoeveel waarschijnlijker* de ene hypothese is ten opzichte van een andere. Met andere woorden, nulhypothesetoetsing geeft in dit voorbeeld geen bevredigend antwoord op de onderzoeksvraag, in de volgende paragraaf gaan we nader in op wat er precies misgaat.

Wat gaat er mis?

Er is door de tijd heen veel literatuur verschenen met kritiek op het gebruik van nulhypothesetoetsing (NHT) en het gebruik van p-waarden (Cohen, 1990, 1992, 1994; Balluerka, Gómez & Hidalgo, 2005; Krantz, 1999; Rozenboom, 1960; Sterne & Smith, 2001; Lee & Wagenmakers, 2005). Wij zullen ons voornamelijk richten op waar het mis gaat bij het evalueren van informatieve hypothesen met behulp van NHT.

Bij NHT is de hypothese die daadwerkelijk getoetst wordt de bekende nulhypothese *er is niks aan de hand* versus het alternatief *er gebeurt iets, maar we weten niet wat*. In het eerste voorbeeld van de vorige paragraaf was de onderzoeksvraag welke informatieve hypothese het meest waarschijnlijk was, HA, HB of HC:

HA: $\mu_1 > \{\mu_2 = \mu_3 = \mu_4\}$,
HB: $\{\mu_1 = \mu_4\} > \{\mu_2 = \mu_3\}$,
HC: $\mu_1 > \mu_3 > \mu_4 > \mu_2$.

De hypothesen die daadwerkelijk getoetst worden met NHT zijn echter:

H0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$,
H1: niet H0 .

Merk op dat deze nulhypothese (H0) en alternatieve hypothese (H1) niet hetzelfde zijn als de informatieve hypothesen HA, HB en HC die de onderzoekers eigenlijk wilden evalueren. Als de nulhypothese en de alternatieve hypothese geen onderdeel zijn van de onderzoeksvraag, dan is er geen directe relatie tussen de hypothesen waar een onderzoeker in geïnteresseerd is en de hypothesen die daadwerkelijk getoetst worden met NHT. De resultaten van NHT geven in dat geval geen antwoord op de onderzoeksvraag.

Daar komt bij dat onderzoekers vaak helemaal niet geïnteresseerd zijn in de nulhypothese. Onderzoekers hebben namelijk vrijwel altijd verwachtingen over hoe de relatie tussen variabelen eruit zou moeten zien. Het is dan raar dat een nulhypothese 'er is niks aan de hand' wordt getoetst aangezien de onderzoeker van tevoren al weet dat er wel 'iets' aan de hand is. Het is dan veel logischer om de informatieve hypothesen rechtstreeks te evalueren in plaats van de nulhypothese te toetsen.

Als dan toch een nulhypothese wordt getoetst, dan wordt de traditionele p-waarde gebruikt om deze nulhypothese te verwerpen of niet te verwerpen. Het omslagpunt van deze dichotome beslissing ligt bij de welbekende waarde van $p < .05$. Deze drempelwaarde van .05 is niet alleen willekeurig gekozen (zie bijvoorbeeld Cohen, 1994; Rozenboom, 1960), maar laat alleen ruimte voor de conclusie dat een nulhypothese wel of niet wordt verworpen, met niks daartussenin. Dit kan leiden tot vreemde beslissingen, bijvoorbeeld in het geval dat de waarde $p = .051$ of $p = .049$ is. In het eerste geval wordt de nulhypothese niet verworpen en in het tweede geval wel. Het mag duidelijk zijn dat beide situaties niet veel van elkaar verschillen. Het is dan vreemd dat de conclusie voor beide situaties totaal anders is.

Wanneer de nulhypothese wordt verworpen, dan weten we eigenlijk nog steeds niks over de informatieve hypothesen, aangezien de alternatieve hypothese geen informatie bevat over de ordening tussen de gemiddelden. Ook een visuele inspectie van bijvoorbeeld de groepsgemiddelden is niet altijd voldoende en is in ieder geval subjectief. Hoe kan een onderzoeker dan toch uitspraken doen over de informatieve hypothesen? Het resultaat zou geen dichotome 'ja/'nee'-beslissing moeten zijn, maar een *kans per hypothese* dat deze de beste is. In de volgende paragraaf presenteren we een methode, Bayesiaanse modelselectie, die hiertoe wel in staat is.

Bayesiaanse modelselectie

Omdat veel artikelen over Bayesiaanse modelselectie (BMS) lastig te lezen zijn voor een niet-statisticus, geven wij een zeer korte en vereenvoudigde introductie. Hiervoor gebruiken we een tweede voorbeeld dat is gebaseerd op Doosje et al. (2008) met slechts twee groepen en één variabele, zodat we ook grafisch kunnen weergeven wat er gebeurt. Voor een uitgebreidere introductie en voor een overzicht van publicaties zie Hoijtink et al. (2008) en Van de Schoot et al. (2008) en voor een meer technische introductie Klugkist et al. (2005).

Stel dat we de groep *hoge stress* (μ_1) willen vergelijken met de groep *lage stress* (μ_2) op het aantal keren verkouden zijn geweest in het afgelopen halfjaar. En stel dat we de volgende drie hypothesen hebben: (H1) er is geen verwachting over de twee groepen; (H2) beide groepen hebben dezelfde score; en (H3) de groep *hoge stress* is vaker verkouden dan groep de groep *lage stress*. De hypothesen zien er dan zo uit:

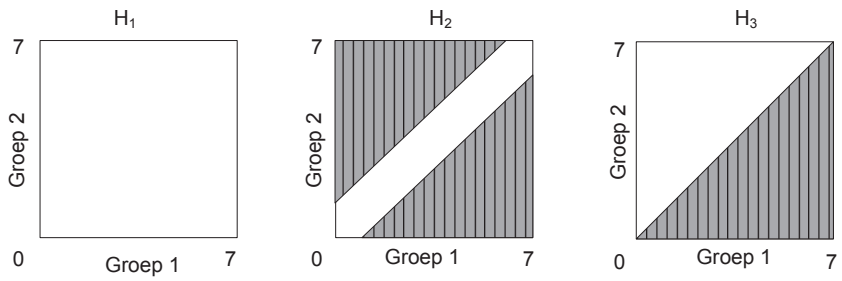
H1: μ_1 ; μ_2,
H2: $\mu_1 = \mu_2$,
H3: $\mu_1 > \mu_2$.

Om erachter te komen welke van de drie hierboven beschreven hypothesen het meest waarschijnlijk is, gaan we deze evalueren met zogenaamde posterior modelkansen (PMK). Om deze PMK's uit te rekenen zijn drie ingrediënten nodig, namelijk (1) de voorkennis die een onderzoeker heeft, (2) de *likelihood* (waarschijnlijkheid) van de data en (3) de steun in de data voor elk van de hypothesen.

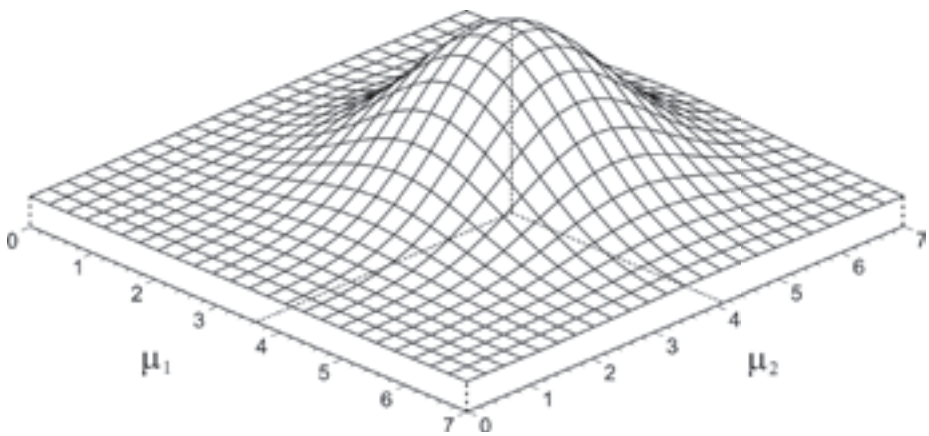
Het eerste ingrediënt is de kennis die er is over de ordening van de gemiddelde scores van de groepen *hoge stress* (μ_1) en *lage stress* (μ_2), voordat de data zijn gezien. Dit zijn de opgestelde hypothesen die in Figuur 2 grafisch zijn weergegeven. Het vierkant vertegenwoordigt alle mogelijke combinaties van gemiddelden die beide groepen kunnen hebben op de variabele 'verkouden zijn'.

Voor elke hypothese bepalen we nu wat de toegestane ruimte is binnen dit vierkant. Met andere woorden, we bepalen welke combinaties van gemiddelden toegestaan zijn voor elk van de opgestelde hypothesen. Voor Hypothese 1 is de gehele ruimte mogelijk, alle combinaties van gemiddelde scores op *verkouden* zijn toegestaan. Voor Hypothese 2 is een groot deel van het vierkant niet toegestaan, er zijn namelijk alleen combinaties van gemiddelden mogelijk waar beide groepen aan elkaar

Onderzoekers zijn vaak helemaal niet geïnteresseerd in de nulhypothese.



Figuur 2. Voorkennis vertaald in Informatieve Hypothesen



Figuur 3. Likelihood van de data

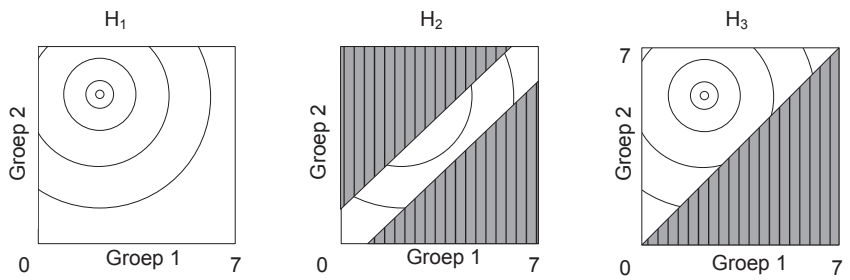
gelijk zijn, dit is de diagonaal van Figuur 2. Voor Hypothese 3 is het gedeelte van het vierkant toegestaan waar de groep *hoge stress* hoger scoort op het risico om verkouden te zijn dan de groep *lage stress*. Dit is de witte driehoek in Figuur 2. Binnen deze ruimte is elke combinatie van gemiddelden even waarschijnlijk. Dit wordt ook wel een uniforme priorverdeling genoemd die over de toegestane ruimte is gelegd (zie voor meer informatie hierover Klugkist et al., 2005).

Het tweede ingrediënt is de informatie die er aanwezig is in de data over de waarschijnlijkheid van mogelijke combinaties in de populatie. Dit wordt ook wel *likelihood van de data* genoemd en kan gezien worden als een landschap met daarin een piek op de plek waar de meest waarschijnlijke combinatie van de gemiddelden zich bevindt (zie Figuur 3). Als we een uitspraak willen doen over het gemiddelde van de groepen *hoge stress* en *lage stress* in de populatie, dan zijn de gemiddelden die in de data zijn geobserveerd, het meest waarschijnlijk. Stel dat bijvoorbeeld $\mu_1 = 4.1$ en $\mu_2 = 3.6$ de gemiddeldes zijn die in de data set zijn geobserveerd, dan is de kans dat in de populatie groep 1 een gemiddelde heeft van 3.6 en groep 2 een gemiddelde van 4.1, maximaal. Merk op dat dit fictieve waardes zijn, gebaseerd op een hypothetische dataset. De maximale waarschijnlijkheid, *maximum likelihood* genoemd, is de piek van de curve in Figuur 3. Combinaties van gemiddelden die verder afliggen van deze piek, zullen steeds minder waarschijnlijk zijn in de populatie en leiden tot een steeds lagere curve in Figuur 3. De kans dat bijvoorbeeld in de populatie de groepsgemiddelden een waarde hebben van $\mu_1 = 1.2$ en $\mu_2 = 6.8$, is veel kleiner, wat te zien is aan de lagere curve op dit punt in de grafiek.

De methode maakt gebruik van de voorkennis die een onderzoeker heeft.

Het derde ingrediënt is de berekening van de hoeveelheid steun in de data voor elk van de hypothesen. Dit wordt gedaan door de gemiddelde hoogte van de likelihood uit te rekenen binnen de toegestane parameterruimte. Om dit grafisch weer te geven, leggen we Figuur 2 op Figuur 3, wat resulteert in Figuur 4. In deze laatste figuur is te zien hoeveel er van de likelihood in de toegestane ruimte van het vierkant ligt. Voor elk van de drie hypothesen kan vervolgens worden uitgerekend hoe groot de gemiddelde hoogte van de likelihood is in deze toegestane ruimte. Een groot deel van het lagere gebied van de likelihood valt bijvoorbeeld buiten de toegestane ruimte die bij Hypothese 3 hoort. Dit lagere gebied wordt echter wel meegenomen in de berekening van Hypothese 1 omdat hier het gehele oppervlak van het vierkant is toegestaan. Hierdoor zal de gemiddelde hoogte van de likelihood voor Hypothese 3 een stuk hoger zijn dan voor Hypothese 1. De gemiddelde hoogte voor Hypothese 2 zal juist heel erg klein zijn ten opzichte van Hypothese 1 en 3, omdat een groot gedeelte van de likelihood inclusief de piek van de curve niet in het toegestane gebied van Hypothese 2 ligt.

De drie ingrediënten die we hiervoor besproken hebben, worden omgezet in posterior modelkansen (PMK's). Wanneer de gemiddelde hoogte van de likelihood groter is, dan is er meer steun van de data voor de hypothese, wat vervolgens resulteert in een hogere PMK. Een PMK houdt niet alleen rekening met hoe goed de hypothese bij de data past, maar ook hoe complex de hypothese is. Dit resulteert in één enkel getal per hypothese op een kansschaal en kan geïnterpreteerd worden als de waarschijnlijkheid per hypothese dat deze de beste is van alle hypothesen die onderzocht worden. Een gebruiker van BMS 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, de bijbehorende software levert de uitkomsten van de analyses (zie http://www.fss.uu.nl/ms/informativehypothesis).



Figuur 4. De voorkennis en data met elkaar gecombineerd.

De resultaten van voorbeeld 2 (zie Tabel 2) laten zien dat Hypothese 3 de meeste ondersteuning krijgt door de informatie die in de dataset zit en dus de beste hypothese is vergeleken met de andere twee hypothesen. De conclusie is dan dat de verwachting dat de *hoge-stress*-groep hoger scoort dan de de *lage-stress*groep op 'verkouden zijn', het beste is met een waarschijnlijkheid van .61. De kans dat deze conclusie niet correct is, is 1- .61 =.39. Het is nu aan de onderzoeker om te beslissen of een waarschijnlijkheid van .61 en een foutmarge van .39 een interessante conclusie oplevert.

Het lijkt misschien dat er een grote foutmarge is,

merk echter wel op dat Hypothese 3 ongeveer acht keer zo waarschijnlijk is als Hypothese 2. Dit kan op zichzelf al een bevredigend resultaat zijn. Dat Hypothese 3 'slechts' twee keer zo waarschijnlijk is als een model zonder enige beperkingen (Hypothese 1), is niet eens zo slecht. Dit omdat het enige verschil tussen beide hypothesen slechts één restrictie is. De conclusie dat Hypothese 3 de beste hypothese is in deze modelselectiecompetitie is dus geoorloofd. Het zou overigens best kunnen zijn dat bepaalde resultaten van BMS een ongeveer even grote PMK opleveren, bijvoorbeeld bij een PMK's van .49 en .50. In dit geval moet de onderzoeker terug naar de tekentafel en moet er gezocht worden naar een betere verwachting die wellicht meer ondersteuning krijgt van de data. Deze nieuwe hypothese kan dan worden toegevoegd aan de reeds bestaande set van hypothesen. Dit geldt natuurlijk ook als een andere onderzoeker een andere hypothese er op na houdt en zijn of haar eigen verwachting wil toevoegen.

Wat hebben we nu gedaan? We hebben de voorkennis voordat we data hebben verzameld, vertaald in een set van hypothesen. We hebben daarna uitgerekend hoe waarschijnlijk deze hypothesen zijn nadat we de data hebben gezien. Hierdoor is duidelijk gemaakt welke verwachting het beste wordt ondersteund door de data en ook hoeveel onzekerheid hierover bestaat. In de volgende paragraaf passen we deze methode toe op het uitgebreidere Karasekvoorbeeld.

Voorbeeld 1 opnieuw geanalyseerd

De verwachtingen van voorbeeld 1 zijn met behulp van BMS geanalyseerd. Ter herinnering: verwachting A stelt dat de groep *hoge stress* het vaakst verkouden is ten opzichte van de groepen *lage stress*, *passief* en *actief*; verwachting B stelt dat de groepen actief en hoge stress vaker verkouden zijn dan de andere twee groepen; verwachting C stelt dat de groep *hoge stress* het vaakst verkouden is gevolgd door respectievelijk de groep *passief*, *actief* en *lage stress*.

In Tabel 3 is voor elk van de drie verwachtingen aangegeven hoe waarschijnlijk deze is. Hypothese C heeft de hoogste waarschijnlijkheid en heeft een kans van .88 dat dit de beste hypothese is en een kans van .12 dat dit niet zo is.

Geconcludeerd kan worden dat de groep *hoge stress* (μ_1) het vaakst verkouden is. Mensen die gekarakteriseerd worden door veel stress op het werk en maar weinig sturingsmogelijkheden hebben, lopen dus het grootste risico op verkoudheid. Deze groep wordt gevolgd door mensen die een lage werkdruk ervaren maar tevens ook weinig sturingsmogelijkheden hebben. Heb je echter veel sturingsmogelijkheden dan heb je minder kans om verkouden te zijn, en heb je ook nog eens weinig stress op het werk, dan ben je relatief het gezondst en ben je het minst vaak verkouden.

Verschillende hypothesen kunnen direct met elkaar worden vergeleken.

Hypothese	PMK
H_1	.31
H_2	.08
H_3	.61

PMK = Posterior Model Kans

Tabel 2. Resultaten BMS voor Voorbeeld 2

Hypothese	PMK
HA	.11
Hb	.01
Hc	.88

PMK = Posterior Model Kans

Tabel 3. Resultaten van BMS voor Voorbeeld 1

Conclusie

Met klassieke nulhypothesetoetsing (NHT) moet een hele stapel output geëvalueerd worden om een onderzoeksvraag te beantwoorden: F-toets, post-hoctoetsen, groepsgemiddelden, et cetera. Deze stapel output kan makkelijk leiden tot verwarrende resultaten. Daarnaast geven de resultaten van NHT geen direct antwoord op de onderzoeksvraag en kunnen informatieve hypothesen niet direct met elkaar worden vergeleken, iets dat met BMS wel kan. Ook wanneer het niet voor de hand ligt welke hypothese de meeste steun krijgt van de data, bijvoorbeeld wanneer de ordening van de groepsgemiddelden niet geheel overeenkomt met elk van de hypothesen, geeft BMS nog steeds 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, enzovoort, geeft BMS nog steeds een enkel getal per hypothese.

Bayesiaanse modelselectie (BMS) resulteert in makkelijk te interpreteren resultaten en geeft een precies antwoord op de onderzoeksvraag. Dat is namelijk per verwachting/hypothese de kans dat deze hypothese de beste hypothese is en dus de meeste steun krijgt van de data. BMS is daardoor een veelbelovend alternatief voor NHT en zal steeds vaker opduiken in de wetenschappelijk literatuur.

Msc A.G.J. van de Schoot en prof.dr. H. Hoijtink zijn, respectievelijk als promovendus en als hoogleraar, verbonden aan de Afdeling Methoden en Technieken van de Universiteit Utrecht, Postbus 80.140, 3508 TC, Utrecht. Drs. S. Doosje is als promovendus verbonden aan de Afdeling Klinische en Gezondheidspsychologie van dezelfde universiteit. Correspondentie naar de eerste auteur, e-mailadres: <a.g.j.vandeschoot@uu.nl>.

Noot

Geschreven met behulp van een subsidie van het Nederlands Wetenschappelijk Onderzoeksinstituut: NOW-VICI-453-05-002.

Literatuur

Balluerka, N., Gómez, J. & Hidalgo, D. (2005). The controversy over null hypothesis significance testing revisited. *Methodology*, *1*, 55-70.

Bayarri, M.J. & Berger, J.O. (2000). P-values for composite null models. *Journal of the American Statistical Association*, *95*, 1127-1142.

Bayes, T. (1764). An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, *53*, 370-418.

Buunk, B.P., Jonge, J. de, Ybema, J.F. & Wolff, C.J. (1998). Psychosocial aspects of occupational stress. In H. Thierry, P.J. Drenth, P.J. Willems & C.J. de Wolff (Eds.), *Handbook of work and organizational psychology*. Hove, England: Psychology Press/ Erlbaum (UK) Taylor & Francis.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist*, *45*, 1304-1312.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*, 155-159.

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

Corfield, D. & Williamson, J. (Eds.) (2001). *Foundations of bayesianism* (Vol. 24). London: Kluwer Academic Publishers.

Doosje, S., Goede, M.D., Doornen, L.V., Goldstein, J. & Schoot, R. van de (2008). Humorous coping styles, job characteristics and job-related affect as predictors of the incidence of upper respiratory infection. Article submitted for publication.

Hao, Q., Duan, Z. & Zhang, A. (2002). Relations between the occupational stress of nurses and their salivary immunoglobulin a level. *Chinese Nursing Research*, *16*, 207-208.

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

Karasek, R.A. (1979). Job demands, job decision latitude and mental strain. Implications for job redesign. *Administrative Science Quarterly*, *24*, 285-308.

Kass, R.E. & Raftery, A.E. (1995). Bayes factors. *Journal of the American Statistical Association, 90,* 773-795.

Klugkist, I., Laudy, O. & Hoijtink, H. (2005). Inequality constrained analysis of variance: a bayesian approach. *Psychological Methods*, *10*, 477-493.

Krantz, D.H. (1999). The null hypothesis testing controversy in psychology. *Journal of the American Statistical Association*, *94*, 1372-1381.

Lee, M.D. & Wagenmakers, E.-J. (2005). Bayesian statistical inference in psychology. Comment on Trafimow (2003). *Psychological Review*, *112*, 662-668.

Mohren, D.C., Swaen, G.M., Borm, P.J., Bast, A. & Galama, J.M. (2001). Psychological job demands as a risk factor for common cold in a dutch working population. *Journal of Psychosomatic Research*, *50*, 21-27.

Raftery, A.E. (1995). Bayesian model selection in social research. *Sociological Methodology*, *25*, 111-163.

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

Schnall, P., Landsbergis, P. & Baker, D. (1994). Job strain and cardiovascular disease. *Annual Review of Public Health*, *15*, 381-411.

Sterne, J.A.C. & Smith, G.D. (2001). Sifting the evidence – what's wrong with significance tests? *Physical Therapy*, *8*, 1464-1471.

Schoot, R. van de, Hoijtink, H., Mulder, J., Aken, M. van, Orobio de Castro, B., Meeus, W. et al. (2008). Should null, alternative or informative hypotheses be used to evaluate expectations in psychological research? Manuscript submitted for publication.

Stigler, S.M. (1986). Laplace's 1774 memoir on inverse probability. *Statistical Science*, *1*, 359-363.

summary

Directly evaluating expectations or testing the null hypothesis. Null hypothesis testing versus Bayesian Model Selection

A.G.J. van de Schoot, H. Hoijtink, S. Doosje

Researchers in psychology have specific expectations about their theories. These are called informative hypothesis because they contain information about reality. Note that these hypotheses are not necessarily the same as the traditional null and alternative hypothesis. Many researchers use traditional null-hypothesis testing to evaluate informative hypotheses. However, this can be problematic, as will become clear in this article. We offer an innovative solution to evaluate informed hypotheses based on Bayesian Model Selection. The method is introduced in non-statistical terms and its utility is illustrated by applying it to examples from occupational health psychology.