De 5 procent religie. Tien geboden om een kansresultaat toch publiceerbaar te maken. En een voorstel om dat te voorkomen.

Laatst gebeurde het weer eens. Een student had haar resultaten aan de supervisor getoond en de supervisor had gezegd, kijk daar zit een 'significant' effect, daar moet je de aandacht op richten. Ik vroeg haar: maar in je werkstuk concept had je toch hypotheses? Die moet je toch gewoon toetsen? Nee daar kwam niet zoveel uit dus dat was niet zo nodig meer. Ze was verbaasd dat ik die benadering niet OK vond.
Onderzoekers aan deze afdeling lijden aan de 5 procent fixatie en ze zijn helaas niet de enigen in psychologenland. Ik zie diezelfde obsessie ook bij nagenoeg alle studenten in OP-groepjes, dus het begint al vroeg.
De reden achter deze aanbidding van het 5% niveau is de misvatting dat resultaten die statistisch niet 'significant' lijken, wetenschappelijk niet relevant zouden zijn en dus door de meeste *editors* van de tijdschriften bij voorbaat al naar de prullenmand worden verwezen. Aangezien een wetenschapper wordt afgerekend op het aantal publicaties moet hij of zij dus proberen in ieder geval minstens één $p<0.05$ resultaat per jaar te 'produceren'.
Het gebruik van het woord 'significant' werkt de misvatting in de hand. Het betreft hier een subjectief label dat in de opleiding en het taalgebruik vermeden zou moeten worden. Het is heus niet zo moeilijk om het voluit te zeggen: "de kans dat dit resultaat op toeval berust is 1 op 20". Dan kan ieder voor zichzelf beslissen of het resultaat zijn of haar aandacht waard is.

Er zijn talloze manieren om een toevalseffect (zoals ik dan maar een $p>0.05$ resultaat zal noemen) toch nog 'significant', te krijgen. Sommige van deze manieren raken aan pure, zij het zeker niet altijd bewuste, fraude, anderen zijn wat subtieler. Sommigen schijnen zelfs aangeleerd te worden in het OP!

Er zijn geen harde kwantitatieve gegevens bekend over al de manieren waarop gesjoemeld wordt, maar uit de dagelijkse praktijk weet ik dat binnen het gebouw A op het Roeterseiland de volgende zaken niet zeldzaam zijn.

Data massage

1. *Verwijder studies post hoc.*
   Je doet drie studies en je publiceert er 1. Driemaal raden welke dat is. Juist, die studie die de 'mooiste' p-waarden heeft. Deze praktijk heet ook wel publikatie bias en het resulteert in een filedrawer effect, resultaten die nooit gepubliceerd worden en dus in de la blijven. Een speciaal geval is het promoveren van de 'status' van een studie. Je kunt een studie, bijvoorbeeld uitgevoerd door een OP-groepje, die je oorspronkelijk had beschouwd als een 'pilot', promoveren tot hoofdstudie als de uitkomst je bevalt.

2 Verwijder proefpersonen post hoc.
Dat kan op grond van hele aannemelijke criteria gebeuren. Bijvoorbeeld een exit-vragenlijstresultaat waaruit blijkt dat de proefpersoon de instructie niet goed begrepen heeft. Zelden zie je echter proefpersonen verwijderd worden als daardoor de p-waarde groter (minder significant) wordt. Dus hier is gewoon sprake van sjoemelen als niet voor de aanvang van het experiment precies is aangegeven op welke gronden proefpersonen verwijderd MOETEN worden.

3. Doe nog enkele proefpersonen als je dicht bij de 5% grens zit.
Gelukkig kan er niet alleen gesjoemeld worden door proefpersonen te verwijderen;-). Het kan ook door er nog een paar bij te doen. Dit valt onder de noemer 'optional stopping' waarmee bedoeld wordt net zo lang door te gaan met nieuwe proefpersonen totdat 5% is bereikt. In de praktijk is het vaak zo dat de onderzoeker dat als volgt goedpraat. Laten we zeggen dat er na 20 ppn een 7% nivo was bereikt. De onderzoeker zegt dan 'aha er is wel een effect maar de 'power' van het experiment was net niet groot genoeg'. Dan doen we er toch nog maar een paar proefpersonen bij. Soms kan dat gedrag nog verder 'gerechtvaardigd' worden doordat er wat ppn waren uitgevallen (zie 2). Dat lijkt legitiem maar het is sjoemelen want men vult die 'uitgevallen' proefpersonen niet aan als de p-waarde 0.045 is. Je kijkt wel uit, het zou zo maar eens 0.055 kunnen worden!

4. Verwijder trials post hoc
Vooral bij psychofysiologisch werk zijn hier veel mogelijkheden, bijvoorbeeld bij de zogenaamde artefact verwijdering. Als het artefact niet van tevoren **precies** gespecificeerd is kun je hier prachtig mee spelen totdat je p-waarde er weer wat mooier uitziet. Hetzelfde zie je bij reactietijden onderzoek. Uitbijters in reactietijden moeten verwijderd worden, maar dat kan op heel veel manieren en er wordt bijna nooit vantevoren precies aangegeven hoe dat zal gaan gebeuren.

Over-analyse en ge-biasde analyse procedures.

5. Double check marginale resultaten
Resultaten die bijna significant zijn worden altijd ge'doublechecked'. Je zal toch maar een fout gemaakt hebben! Resultaten die net wel 'significiant' zijn worden echter zelden ge'doublechecked'. Kennelijk mag je wel fouten maken als het eindresultaat maar publiceerbaar is. Dit gebeurt overigens absoluut niet bewust.

6. Neem veel factoren op in je design dan is er altijd wel wat 'interessants'.
Als de vantevoren gespecificeerde analyse duidelijk corrigeert voor het feit dat er nogal wat met elkaar vergeleken wordt dan is dat op zich toelaatbaar, alhoewel je eigenlijk de voorspellingen op een model zou moeten baseren en dus een aantal van de vergelijkingen theoretisch zinloos zijn. Toch worden niet zelden onvoorspelde interacties gepresenteerd als zijnde voorspeld. Gedrag van mensen is zo complex dat er uiteindelijk bij elk resultaat wel een verhaal te bedenken is. Soms gaat dit zo ver

dat de inleiding wordt ‘herschreven’ en de incidentele interactie wordt gepromoveerd tot een ‘verwacht’ effect.

7. Probeer alle analyses die je kunt bedenken en kies de beste.
Je voorspelt dat leeftijd en V samenhangen. Je doet eerst een correlationele analyse en als daar ‘niets uitkomt’ splits je de proefpersoongroep in tweeën, 1 groep met leeftijden groter dan de gemiddelde leeftijd, 1 groep met leeftijden kleiner dan het gemiddelde en dan vergelijk je de gemiddelde V –scores van die twee groepen. En als daar ‘niets uitkomt’ splits je de groep op grond van de mediaan en als daar ‘niets uitkomt’ kies je drie groepen waarvan je de middelste verwijdert enzovoorts en zoverder. Je kunt ook nog voor alle analyses kijken naar de non-parametrische en de parametrische versie. Op zich is dit alles niet verboden indien het explorations betreft, maar dan moet worden aangegeven hoeveel analyses hebben plaats gevonden.

8. *Rommel met de afhankelijk variabele: Transformeer de data*
Vaak worden data die theoretisch niet normaal zijn verdeeld, zoals reactietijddata, getransformeerd om ze normaal verdeeld te krijgen. Prima, mits vantevoren aangekondigd. Maar ook achteraf kun je data ofwel transformeren ofwel combineren zodat er nieuwe variabelen worden gecreëerd die wellicht iets betere p-waarden leveren.

9. *Kies de optimale specificatie van de afhankelijke variabele*
In psychofysiologisch onderzoek bij het vergelijken van bijvoorbeeld Evoked Potentials bij twee condities zijn er erg veel mogelijkheden voor deze vergelijking. De Evoked Potential bestaat uit wel 128 meetpunten. En zelfs als er alleen naar piek amplitudes gekeken wordt kun je vaak nog kiezen uit 3 of 4 pieken. In fMRI zijn deze mogelijkheden zo groot dat in de meeste analyse-pakketten daarvoor corrigeren zodat over-analyse voorkomen wordt.

10. Kies eenzijdige toetsing zelfs als vantevoren geen richting is gespecificeerd
In exploratieve toetsingen die in het algemeen niet door een model worden gedreven zijn in principe altijd twee-zijdige toetsingen nodig. Standaard wordt in zeer veel onderzoek een sexe effect getoetst. Vaak wordt niet vantevoren aangegeven of de vrouwen hoger dan wel lager zullen scoren dan de mannen, er wordt slecht een verschil voorspeld.

Wat is nu het resultaat van al dit gesjoemel? Ten eerste: zelfbedrog. Het resultaat waarvan je stelt dat het kans heeft van 1 op de 20 om bij toeval op te treden is helemaal niet zo ‘significant’. In de medische literatuur wordt gesteld dat meer dan 50% van de gepubliceerde resultaten ‘incorrect’ zijn (Ionnidis, JPA (2005). Why most Published Research Findings are False, Plos 2-8, e124). Of anders gezegd: van elk van de 2 zogenaamd significante resultaten in de literatuur is er 1 een opgeklopt kansresultaat. Het zal in de psychologie niet veel anders zijn; kortom de helft van de artikelen die de studenten moeten lezen is volkomen kul! Dat is niet alleen een verspilling van papier

maar vooral een verspilling van tijd van de studenten en van andere lezers. Dus eigenlijk wordt de vooruitgang van de wetenschap gefrustreerd.

In nagenoeg alle vakgebieden zijn alle hierboven besproken praktijken "zonden" en weet men dat ook wel. Maar toch wordt er nauwelijks expliciet over gesproken en als dat gebeurt wordt nooit man en paard genoemd maar wordt er over experimenter 'bias' gesproken. In het onderwijs wordt er onvoldoende aandacht aan besteed. En het gaat gewoon maar door. Zoals hierboven betoogd moet de oorzaak gezocht worden in het publicatiebeleid waarbij alleen wat 'significant' is in aanmerking komt. Maar ook op congressen krijg je als onderzoeker alleen maar aandacht als je iets 'significants presenteert.
Kunnen we hier nu echt niets aan doen? Het is toch eigenlijk om je dood te schamen. Het ligt voor de hand omin de eerste plaats naar het publicatiebeleid te kijken. Immers daar worden deze praktijken door in de hand gewerkt. Er is bijvoorbeeld een vakgebied waar de onderzoeker de mogelijkheid heeft om een experiment VANTEVOREN bij een tijdschrift te deponeren voor acceptatie. **Onafhankelijk** van de resultaten. Een dergelijk beleid resulteert dan meteen in meer nadruk op de zogenaamde meta-analyse waarbij de effect grootte en niet de p-waarde de relevante variabele is. Dat is overigens iets wat ook door de Amerikaanse beroepsorganisatie van psychologen (de APA) wordt gestimuleerd. Helaas hebben we het publicatiebeleid niet in eigen hand en de vraag is dan wat we er **zelf** aan kunnen doen.
Zelf kunnen we als individuele onderzoeker streven naar het volgen van dubbel blinde onderzoeks-opzetten maar ook blinde analyses waarbij de condities gecodeerd zijn. En meer institutioneel stel ik het volgende voor:

Voorstel
Ik stel voor dat bij het indienen van onderzoek bij de ethische commissie ook een pagina wordt overhandigd met de specifieke voorspellingen en de geplande analyses alsmede criteria voor uitsluiting van proefpersonen en trials. Het is echt niet zoveel werk. Deze voorspellingen worden dan in standaardvorm in een database opgeslagen met de datum en eventueel met een geautomatiseerde power-analyse. Bij het indienen van onderzoek ter publicatie kan dan gerefereerd worden naar de vantevoren gespecificeerde voorspellingen. Ik hoop dat van deze praktijk dan ook een opvoedend werking uitgaat zodat onze studenten leren gesjoemel te herkennen en leren dat gesjoemel niet kan. Je hebt er uiteindelijk jezelf, je collega's en de wetenschap mee.

Dick Bierman

Voor enkele mooie voorbeelden zie ook: http://m0134.fmg.uva.nl/Utts_ethics.pdf
vanaf pag. S5.10