Te mooi om waar te zijn (p > 0,95)

Regelmatig worden onderzoekers geconfronteerd met data die niet geheel aan hun verwachtingen of hypotheses voldoen, of nog erger, met een tekort aan geschikte data. Zeker in vakgebieden waar de publicatiedruk hoog is en de concurrentie moordend, wordt dan, zo blijkt uit de recente geschiedenis, geregeld gebruik gemaakt van dataverdichting, ook wel 'fraude' genoemd: het gegevensbestand wordt net zo lang opgewerkt en verrijkt tot er alsnog een publicabel resultaat uitrolt. De recente geschiedenis leert echter ook, dat deze fraude niet altijd even oordeelkundig geschiedt, waardoor de pleger na verloop van tijd toch tegen de lamp loopt en zijn handelen alom en luidkeels wordt veroordeeld. De perfecte misdaad blijkt nog niet zo eenvoudig te plegen (of wel natuurlijk, dat kunnen we niet weten).

Om in deze kennelijke lacune te voorzien geven wij een korte handleiding voor aspirant-fraudeurs. Waarop moeten zij vooral letten? Wat onderscheidt echte data van verdichte data? Wat zijn de beste technieken, en hoe werden beroemde voorgangers ontmaskerd?

Het belangrijkste principe van fraude is: doe het niet. Althans, niet zelf. Gebruik altijd een computer om data te genereren of te adjusteren. Mensen zijn fameus slecht in het verzinnen van willekeurige getallen, zeker als die ook nog aan bepaalde statistische regels moeten voldoen. Het is voor mensen al onmogelijk een echte reeks kop-of-muntworpen te verzinnen, laat staan een lijst van twintig normaal verdeelde scores met een gemiddelde van 86,7 en een standaardafwijking van 23,1. Leerzaam in dit verband zijn de lotgevallen van de Indiase onderzoeker Ram Bahadur Singh, die in de jaren negentig zeer veel schreef over voeding en cholesterol. Zoveel, dat de redacties van tijdschriften argwaan kregen, zijn ruwe data opvroegen, en er statistici op zetten.[1] Het gemiddelde cholesterol was in de dieetgroep 5,46 mmol/l, in de controlegroep 5,43 mmol/l. Dat kan, maar de standaardafwijkingen bleken 0,352 en 0,296 mmol/l, en dat kan niet: $p = 1 \times 10^{-7}$. Ook andere data konden onmogelijk afkomstig zijn uit echte steekproeven. De onderzoeker viel hard door de mand. Tegenwoordig zijn op internet overigens uitstekende programma's te vinden die tot op elke gewenste decimaal nauwkeurig dit soort data wel goed kunnen aanmaken – maak daarvan gebruik.

Dat voorkomt ook twee andere veel gemaakte fouten. De laatste cijfers van Singhs getallen bleken in het geheel niet uniform verdeeld, wat wel zou moeten bij echt willekeurige getallen. Wederom voor cholesterol bijvoorbeeld was het resultaat bepaald onaannemelijk: $p = 6 \times 10^{-221}$. Andersom mogen begincijfers juist niet altijd even vaak voorkomen: die dienen volgens de wet van Benford logaritmisch verdeeld te zijn.[2]

Het meest moeten aspirant-fraudeurs echter letten op resultaten die 'te mooi zijn om waar te zijn' – nauwkeuriger dan de meettheorie toestaat. Beroemd is de affaire rond de Franse immunoloog Jacques Benveniste, die in *Nature* publiceerde over het homeopathisch geheugen van water.[3] Hij stelde basofielen bloot aan steeds verder verdunde anti-IgE, en liet zijn medewerkster tot driemaal toe tellen hoeveel er waren gedegranuleerd. Dan vond zij bijvoorbeeld een gemiddelde van 49, met een standaardfout van 1,7. Dat kan weer niet: de standaardfout moet rond de √49 liggen. Ook hier kan men uitrekenen hoe groot de kans is op zo'n kleine standaardfout als er echt blind geteld is: $p = 2,6 \times 10^{-20}$. (Had *Nature* dat meteen gezien, dan was de hoofdredacteur vast niet zo potsierlijk het lab van Benveniste gaan inspecteren.)

Wie beweert dat zijn munt zo eerlijk is dat hij na 360 keer gooien precies 180 keer kop kreeg, wordt vreemd aangekeken, en wie beweert dat zijn gekruiste erwten precies in de mendeliaanse verhouding 313 : 106 : 103 : 34 uitkomen, is al even ongeloofwaardig: p = 0,996. Vandaar het bekende advies altijd tweezijdig te toetsen: niet alleen zien of $p < 0,05$, maar ook of $p < 0,95$.

Het is duidelijk: vrijwel alle recente gevallen van verdichting die de datapolitie de laatste jaren aan het licht bracht, zijn te wijten aan het feit dat de delinquent de meettheorie onvoldoende beheerste. Het zou al enorm schelen als die een vast onderdeel van het curriculum wordt. ■

Hans van Maanen

LITERATUUR

1 Al-Marzouki S, Evans S, Marshall T, Roberts I. Are these data real? Statistical methods for the detection of data fabrication in clinical trials. BMJ 2005;331:267-70.
2 Hill TP. The first digit phenomenon. Amer Sci 1998;86:358-63.
3 Davenas E, Beauvais F, Amara J et al. Human basophil degranulation triggered by very dilute antiserum against IgE. Nature 1988;333: 816-8.

Hans van Maanen is wetenschapsjournalist.