Blindstaren op significantie

Vraag een zaaltje huisartsen wie kan uitleggen wat 'significantie' precies inhoudt, en na een ongemakkelijke stilte zullen er een paar beginnen te stamelen dat het te maken heeft met 'de kans dat iets toeval is'. Vraag een zaal vol huisartsen waarom die nadruk in de wetenschap op significantie zo verderfelijk is, en het blijft stil.

Om bij het begin te beginnen: significantie gaat, op z'n simpelst, over de waardering van een wedstrijdresultaat. Heeft Ajax met 1-0 gewonnen van PSV, dan zullen Eindhovenaren zelfs in de hoofdstad nog wel kunnen volhouden dat het een gelukkige overwinning was, is het 6-0 geworden, dan worden ze weggehoond.

Zo is in de wetenschap elk experiment uiteindelijk een wedstrijdje, bijvoorbeeld tussen twee geneesmiddelen. Zonder verdere informatie moeten we ervan uitgaan dat ze even goed zijn, vervolgens organiseren we een, zeg, tienkamp – tien patiënten krijgen medicijn A, tien medicijn B. Uitslag: A wint 4 keer, B 6 keer. Is dit voldoende om te zeggen dat B beter is dan A? Dat zullen weinig artsen willen aannemen. Een kansberekening (onze tienkamp is een soort steekproef uit alle mogelijke tienkampen) laat zien dat het in 25% van alle tienkampen precies 5-5 wordt als de teams/medicijnen even sterk zijn. Dus in 75 van de 100 tienkampen zal een van de twee met ten minste 6-4 winnen. Een krachtsverschil van 7-3 of groter is al beter: dat gebeurt slechts in 34% van de gevallen door toevalsfluctuaties.

Losjes gezegd, als we nu – terwijl er geen verschil is – besluiten dat de ene beter is dan de andere, zitten we in 34% van de gevallen fout. Vinden we dat te veel risico, dan moeten we een extremer resultaat eisen. Als we niet vaker dan in 5% van de tienkampen de uitslag fout willen interpreteren, zullen we op een uitslag van 9-1 of 10-0 moeten wachten voor we overtuigd zijn. Dat zijn 'significante' uitslagen: puur door toeval wint een team slechts 1 op de 500 keer alle tien de wedstrijden van een precies even sterk team, en slechts 1 op 50 keer met een punt verschil.

Die grens van 5% ($p < 0{,}05$) is historisch zo gegroeid, maar arbitrair. Het ligt er in feite aan hoe erg het is om fout te zitten. Veel mensen zullen een afslankmiddel dat 94% kans van slagen belooft, toch wel willen proberen – en wie is de wetenschapper om te zeggen dat het resultaat 'niet significant' is? Sterker, als tijdschriften geen resultaten willen publiceren die niet significant op 5%-niveau zijn, komt niemand het bestaan van zo'n middel ooit te weten.

Dat is een eerste bezwaar dat tegen het gebruik van significantie: wat significant is voor de wetenschap, is misschien niet significant voor het publiek.

Ingewikkelder is het probleem, dat significantie niet alleen afhangt van de grootte van het effect, maar ook van het aantal wedstrijden – de grootte van de steekproef. Als we een significantieniveau eisen van 5%, heeft het weinig zin om vier wedstrijden te laten spelen: de kans op 4-0 of 0-4 is 12,5%, dus dat halen we nooit. Maar als we 10.000 wedstrijden laten spelen, is zelfs het kleinste procentuele verschil al significant.

Hoe realistisch is het idee dat de teams even sterk zijn? Niemand denkt echt dat een tot in fase II getest geneesmiddel even goed is als een placebo, maar die schijn wordt zo wel opgehouden – significantietests bevoordelen dus middelen waarvan we al vermoeden dat ze beter zijn. Andersom, een trial waaruit blijkt dat een homeopathisch middel significant beter scoort dan placebo, zal door weinigen serieus worden genomen: de 'voorafkans' dat ze even weinig doen, is te groot.

Zo zitten er nog veel meer, en veel gemenere, haken en ogen aan significantietesten – er zijn artikelen[1,2] en zelfs boeken[3] over volgeschreven. Hele vakgebieden zijn bijna te gronde gericht door de nadruk op p-waarden en significantie. Maar de medische wetenschap kan er moeilijk afstand van doen: het is de bel voor de hond van Pavlov geworden.

Wie alleen op significantie let, laat zich om de tuin leiden of is te lui om naar de echt belangrijke cijfers, zoals de grootte van het effect en de nauwkeurigheid van de meting, te kijken. Daarover wellicht een andere keer. ■

Hans van Maanen

1 Cohen J. The earth is round ($p < 0.05$). Am Psychol 1994;49:997-1003.
2 Hubbard R, Lindsay RM. Why P values are not a useful measure of evidence in statistical significance testing. Theory Psychol 2008;18:69-88.
3 Ziliak ST, McCloskey DN. The cult of statistical significance. Michigan: University of Michigan Press, 2008.

Hans van Maanen is wetenschapsjournalist.