Casper Albers
Psychometrie & Statistiek
Rijksuniversiteit Groningen
c.j.albers@rug.nl

Column **Casper grijpt een kans**

Slim tellen

Caspers Albers schrijft op regelmatige basis in dit blad een column over alledaagse statistische onderwerpen.

Rekenen is de basis van de wiskunde. En tellen is de basis van rekenen. Daarnaast is tellen ook gewoon leuk (zoals Sir Francis Galton zei: "Whenever you can, count" [4]). Je zou dus verwachten dat we ondertussen dat ook gewoon kunnen. Maar tellen is soms nog best moeilijk. Bijvoorbeeld wanneer je niet (goed) kan zien wat je moet tellen.

Een klassiek voorbeeld van veertig jaar geleden is het *proofreading problem*. Aan twee reviewers wordt gevraagd om, onafhankelijk van elkaar, een tekst te lezen en de (spel)fouten te corrigeren. Geen van beide is perfect en ze zien mogelijk fouten over het hoofd. Uiteindelijk worden er a fouten gevonden door de ene reviewer en b door de andere. Er worden c fouten gezamenlijk gevonden. Uiteindelijk worden dus $a+b-c$ fouten gevonden. De vraag is of nu alle fouten gevonden zijn. Via een elegante berekening komt Pólya [5] op

$$\frac{(a-c)(b-c)}{c}$$

als puntschatter voor het aantal ongevonden fouten. De puntschatter is aan te vullen met standaardfouten via de delta-methode, of zelfs om te zetten in een verdelingsschatter [1]. Het aantal fouten dat niet gezien is, valt dus te tellen.

Een nog ouder voorbeeld is de volgende ouderwetse kermisattractie. Een grote pot is gevuld met knikkers en tegen betaling van een stuiver mag je raden hoeveel knikkers er in de pot zitten. Aan het eind van de dag wint degene die het dichtste bij het juiste aantal zit een rollade (of wat vroeger ook maar de prijzen op de kermis waren). Wat bij dit soort problemen vaak blijkt, is dat de mediaan van alle pogingen verrassend dicht bij de echte waarde zit. Galton beschreef dit probleem al ruim een eeuw geleden in zijn stuk 'Vox Populi' [3]. Hier ging het om het raden van het gewicht van een os op de markt van Plymouth. De mediaan van de schattingen van $N=787$ deelnemers bleek op 1 procent nauwkeurig te zijn. (Galton pleitte voor de mediaan in plaats van het gemiddelde, omdat de schattingen niet symmetrisch verdeeld waren.)

In zijn resultaat zag Galton bewijs van de kracht van democratische besluiten door een groep mensen: afwijkende meningen strepen tegen elkaar weg en gemiddeld gesproken weet men wel hoe het moet. Dit concept is later *the wisdom of the crowd* gaan heten. In essentie zegt dit concept eigenlijk weinig meer dan dat het steekproefgemiddelde doorgaans een betere schatting voor de (onbekende) populatiewaarde is dan individuele schattingen. Het werkt echter niet altijd en men is er de afgelopen jaren achter gekomen dat het vaker dan verwacht niet werkt. De belangrijkste eis is dat men — gemiddeld genomen — wel de juiste waarde schat. De schattingen moeten *zuiver* zijn (of, op zijn minst statistisch consistent): als dat zo is dan geldt, onder milde voorwaarden, dat de *mean squared error* van het steekproefgemiddelde een factor N kleiner is dan die van de individuele meting. Bij een pot met knikkers of het gewicht van een os gaat die zuiverheid nog wel op, maar bij complexere situaties niet.

In het dagelijks leven maken veel mensen gebruik van dezelfde informatiebronnen — kranten, tv, et cetera — en lopen daarmee het risico dezelfde vertekende informatie te ontvangen, die een verstoring in het wereldbeeld kan opleveren. Onlangs is een internationale peiling uitgevoerd [2] met onder meer vragen als "Vindt u abortus moreel verwerpelijk?" en "Van welk percentage landgenoten verwacht u dat ze abortus moreel verwerpelijk vinden?" In Nederland gaf 8 procent van de deelnemers aan abortus verwerpelijk te vinden, maar het gemiddelde van de schattingen voor de tweede vraag was 37 procent, ruim vier keer zo veel. Ook schattingen van het percentage inwoners dat moslim is (schatting: 18 procent, werkelijkheid: 6 procent) en of homoseksualiteit verwerpelijk is (5 procent gaf aan dat te vinden, de inschatting was dat 34 procent van de landgenoten het vinden) lagen *ver* van de juiste waarde: de wisdom van de crowd is hier ver te zoeken.

Systematische verstoringen komen niet alleen voor bij het volk (de *vox populi*), maar ook bij de elite (de professionals). Zo hadden bijna alle peilingbureau's de uitslag van de Amerikaanse presidentsverkiezingen fout ondanks dat de peilmethoden opgezet waren door slimme en ervaren politicologen en methodologen. Een van de grootste oorzaken hiervoor was dat men in alle staten dezelfde systematische fout maakte (en daarmee de grootte van het Trump-kamp onderschatte). De grootte van die onderschatting zat nog wel in de foutenmarge per staat, maar bij het samenstellen van de resultaten voor het gehele land niet meer. Hiervoor werd het (gewogen) gemiddelde genomen van de individuele staten en werd ervan uitgegaan dat de meetfouten ongeveer een factor $\sqrt{50}$ kleiner zouden worden. Die vlieger gaat echter alleen op wanneer de meetfouten onafhankelijk zijn (in de ene staat overschat je de Trump-aanhang, in de andere onderschat je deze, en uiteindelijk streept alles grotendeels tegen elkaar weg). Die onafhankelijkheid ging hier niet op: in elke staat lagen dezelfde redenen achter de onderschatting.

Een recenter voorbeeld uit Amerika over telproblemen is de discussie rond de opkomst bij de inauguratie van Donald Trump op 20 januari. Op social media verschenen foto's die de opkomst bij Obama en Trump vergeleken: bij Obama leek het overvol te staan, bij Trump waren grote stukken leeg. Bij die foto's zijn wat vraagtekens te zetten (zijn ze bijvoorbeeld wel op hetzelfde tijdstip genomen; waren de weersomstandigheden bij Obama beter, et cetera), maar ook de experts waren duidelijk: de opkomst acht jaar geleden was (flink) groter. Watson en Yip [6] leggen verschillende methoden uit om goede schattingen te krijgen van de opkomst en hoe je bij deze methoden de standaardfout kan schatten. Met informatie over de oppervlakte ($\hat{O}$; aantal m^2) alsmede de gemiddelde dichtheid ($\hat{D}$; aantal personen per m^2) in de vorm van schattingen aangevuld met standaardfouten (SE), kan je de totale opkomst $\hat{N} = \hat{O} \times \hat{D}$ alsmede de standaardfout van de opkomst

$$\frac{SE(\hat{N})}{\hat{N}} = \sqrt{\frac{SE^2(\hat{O})}{\hat{O}^2} + \frac{SE^2(\hat{D})}{\hat{D}^2}}$$

schatten. Deze ligt doorgaans op circa 10 procent van $\hat{N}$. Ook via andere methoden van schatten van de opkomst kom je op soortgelijke standaardfouten.

De technologie en wiskunde is er dus om redelijk nauwkeurige schattingen te geven van de grootte van een menigte. Helaas zitten koppige politici de werkelijkheid in de weg: de voorlichter van het Witte Huis stond er op dat de opkomst bij Trump toch echt groter was dan die bij Obama — iets dat door Trumps hoogste assistent een 'alternatief feit' genoemd werd. Trump heeft geen alleenrecht op het glashard ontkennen van de waarheid. In 2011 omschreven Watson en Yip al "However, it seems there is too much politics in the mix for crowd estimation to be made precise in the near future. The public has a view of the truth that is coloured by their beliefs. This applies particularly to crowd estimation."

Referenties

1 C.J. Albers, G.Th. de Roos en W. Schaafsma, Estimating a frequency unseen: an example from ornithology, *Statistica Neerlandica* 59(3) (2005), 397–413.

2 P. Duncan, Europeans greatly overestimate Muslim population, poll shows, *The Guardian*, 13 december 2016.

3 F. Galton, Vox Populi, *Nature* 1949(75) (1907), 450–451.

4 F. Galton, in J.R. Newman, ed., *The World of Mathematics – Volume 2*, Simon and Schuster, 1956, p. 1169.

5 G. Pólya. Probabilities in proofreading, *The American Mathematical Monthly* 83(1) (1976), 42.

6 R. Watson en P. Yip, How many were there when it mattered? *Significance*, 2011.