Geurproef niet meer in gebruik bij strafzaken

Geurt Jongbloed & Frank van der Meulen

In *de Volkskrant* van vrijdag 22 april 2011 was op de voorpagina te lezen:

> De omstreden 'geuridentificatieproef', waarin speurhonden van de politie verdachten van een misdrijf aanwijzen, is van tafel. Hondengeleiders die met de proef sjoemelden om verdachten veroordeeld te krijgen, worden echter niet vervolgd.

Het Volkskrantartikel stelt dat 'uit onderzoek van statistici van de TU Delft blijkt dat hondengeleiders de hand moeten hebben gelicht met de regels'. In deze bijdrage willen we de context en inhoud van ons onderzoek wat breder uiteenzetten.

In wat volgt leggen we eerst uit hoe de geurproef in zijn werk gaat. Vervolgens schetsen we in het kort welke punten van kritiek er in de voorbije jaren zoal naar voren zijn gebracht als het gaat om de bewijskracht van de uitkomst van een geurproef. Deze kritiek, met name geuit door prof. J.E.R. Frijters, deed het College van Procureurs Generaal (PG) tot een onafhankelijk onderzoek naar de geurproef besluiten. Het deelonderzoek dat bij de TU Delft werd uitgezet gaat over de vraag of bepaalde rangschikkingen wel zijn bepaald zoals in het protocol is voorgeschreven. Daarom zullen we de procedure die volgens het protocol gevolgd moet worden nader uitleggen en tot slot onze aanpak en conclusies van de analyse beschrijven.

Procedure van de geurproef

Het uitgangspunt is dat op de plaats delict een voorwerp is gevonden en dat er een verdachte is. Met behulp van een daartoe getrainde hond wil men vaststellen of de verdachte een 'geurovereenkomst' met het voorwerp heeft. Zonder details volledig te willen weergeven, komt de procedure van de geurproef op het volgende neer. Voor de proef leveren zeven personen een geurdrager aan. Dit zijn de verdachte, X, en zes figuranten, A, B, t/m F. De zeven personen houden nagenoeg gelijktijdig ieder twee geurdragers vast; ze houden er één in iedere hand stevig vast gedurende 1 à 2 minuten, en wisselen daarna de buizen om naar de andere hand. Vervolgens worden twee rijen gemaakt. In elke rij worden de zeven geurdragers in willekeurige volgorde gelegd. De hond ruikt vervolgens aan een voorwerp dat door persoon A van geur is voorzien en daarna wordt de hond door de hondengeleider langs de zeven geurdragers in rij 1 gevoerd. Als de hond geur A identificeert, wordt dezelfde proef herhaald bij rij 2. Correcte identificatie bij beide ronden wordt als kwalificatie voor de tweede stap gezien. De eerste stap wordt dus gebruikt om te zien of de hond 'in vorm' is. In de tweede stap wordt de geur van de controlepersoon (A) verwijderd uit beide rijen. Nu geeft de geleider zijn hond lucht van een voorwerp dat waarschijnlijk door de dader op de plek van het misdrijf is achtergelaten, alvorens de gang langs de eerste rij met geurdragers te maken. Als de hond precies bij object X een geurovereenkomst aangeeft, wordt dit als identificatie gerapporteerd. Vervolgens wordt hetzelfde gedaan bij rij 2. Als bij beide rijen correcte identificatie optreedt spreekt men van 'herkenning van de verdachte'.

Punten van kritiek

Vanuit verschillende invalshoeken zijn in het verleden kritische kanttekeningen geplaatst bij de bewijskracht van de uitkomst van een geurproef. Deze kanttekeningen hebben te maken met de filosofie achter de geurproef, de techniek van de geurproef en de vraag of de procedure wel goed gevolgd is. Een kwestie van het eerste type is bijvoorbeeld dat de conclusie die in het proces verbaal wordt opgenomen te sterk is voor wat er feitelijk gebeurt. Ook is er geen experimenteel onderzoek geweest naar de relatie tussen de perceptie van menselijke lichaamsgeur en het keuzegedrag van honden en is het een probleem dat de proef om technische redenen niet herhaald kan worden. Procedurele kwesties zijn bijvoorbeeld de vraag of de hondengeleider daadwerkelijk niet weet waar voorwerp X ligt in de rij (blinde proef). Bij sommige van deze vragen zijn goed opgezette proeven te bedenken die in statistische zin tot een antwoord kunnen leiden. De door het College van PG geformuleerde probleemstelling beperkte zich echter tot het procedurele punt van de randomisering. Worden de posities van de geurdragers wel volgens de daarvoor opgestelde procedure bepaald? Deze vraag kan op basis van de beschikbare gegevens worden beantwoord. Voordat we naar de data-analyse gaan, is het van belang om de procedure op het punt van de randomisering nog wat nader te beschouwen.

De randomisering

Zoals hierboven al aangegeven, vindt de geurproef in twee etappes plaats. Hiertoe moeten de zeven geurdragers aan een positie worden toegewezen in beide rijen. Om de keuze te maken voor deze twee rangschikkingen, zijn 36 zogenaamde uitlegschema's vastgesteld. Voorafgaand aan de proef dient het uitlegschema door loting te worden bepaald. Volgens de procedure moet dit worden gedaan door een (zuivere) dobbelsteen twee keer te werpen. Iedere mogelijke uitkomst van dit experiment is eenduidig aan een van de 36 uitlegschema's gekoppeld. Zo correspondeert een uitkomst (5,1) met een schema XFDCBEA bij de eerste rij en CAXFDBE in de tweede rij.

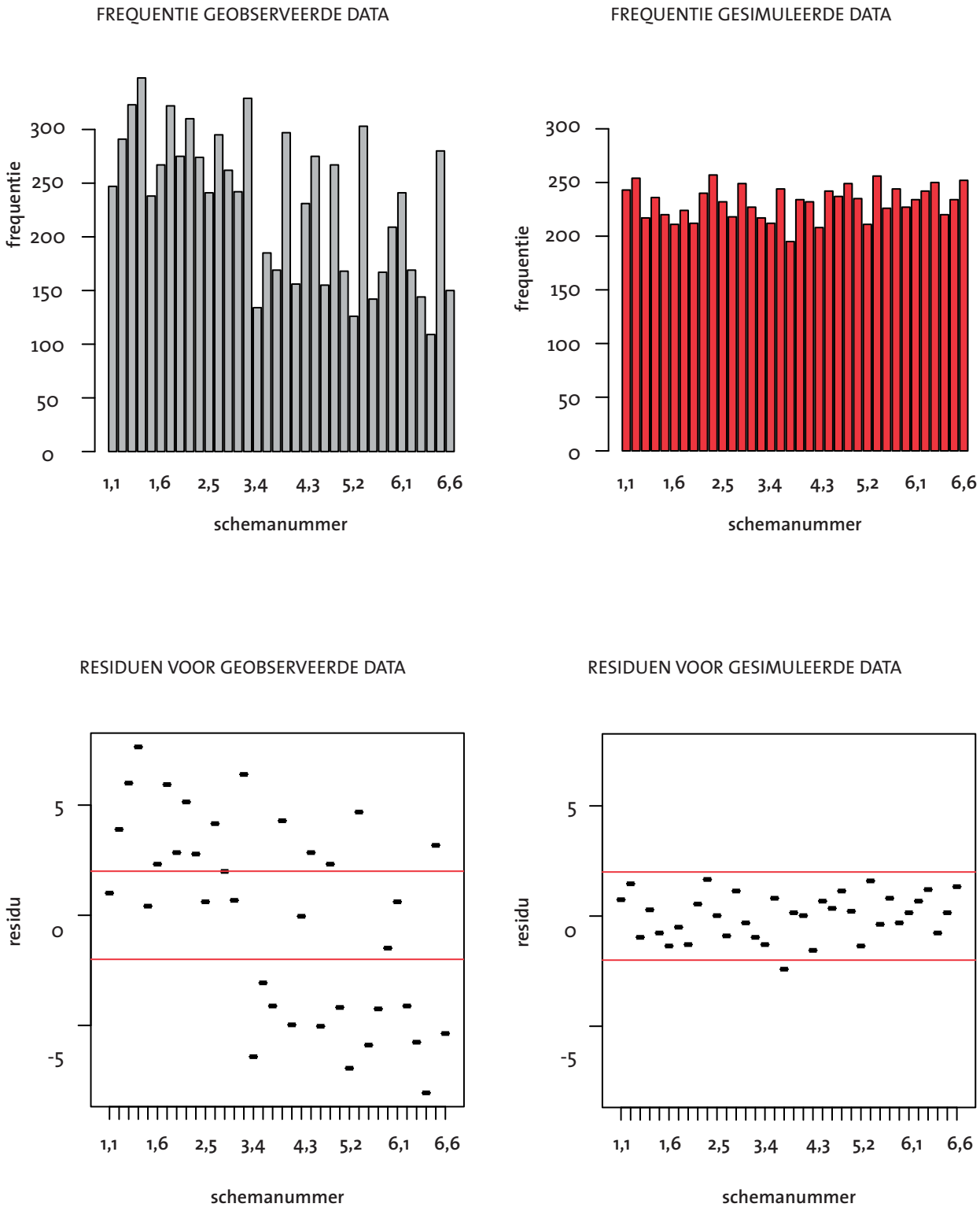
Data en toetsen

Voor het onderzoek kregen we de beschikking over de gegevens van alle 8341 geurproeven die tussen 1999 en 2006 zijn uitgevoerd door de oefengroepen Limburg, Nunspeet, Oost en Rotterdam. Van iedere proef zijn, naast het gekozen uitlegschema, ook de betrokken hondengeleider en de betrokken hond bekend. De uitlegschema's zijn gelabeld door de bijbehorende uitkomsten van de twee worpen: (1,1) tot en met (6,6). Als de procedure van dobbelen met een zuivere dobbelsteen wordt gevolgd, zal de geobserveerde vector van frequenties multinomiaal verdeeld zijn met parameters 8341 en 1/36 (voor iedere cel). Voor iedere hondengeleider geldt dat de bijbehorende vector van frequenties ook multinomiaal verdeeld is, met parameter n (aantal proeven door hem/haar gedaan) en 1/36 voor de celkansen. De toets waarvoor we hebben gekozen is de klassieke Chi-kwadraat toets voor multinomiale kansen. Deze hebben we uitgevoerd voor de gehele dataset, alsmede apart voor iedere helper, althans als de betreffende helper minimaal 180 proeven heeft gedaan. We gaan dus in feite uit van een model waarin de helpers via loting hun uitlegschema kiezen en toetsen de nulhypothese dat de kansen uniform zijn over de 36 uitlegschema's. Op andere afwijkingen van de procedure, zoals het deterministisch kiezen van uitlegschema's, wordt dus niet getoetst. De Chi-kwadraat toetsingsgrootheid gebaseerd op de gehele dataset is gelijk aan 703. Vergelijk dit met het 99,9% kwantiel van de Chi-kwadraat (35)-verdeling, welke gelijk is aan 66,6. Om voor de juristen de uitzonderlijkheid van de gerealiseerde keuzes van schema's nog wat inzichtelijker te maken, zijn frequentiediagrammen gemaakt van zowel de gerealiseerde worpen met de dobbelsteen alsook door ons gegenereerde schema's, volgens het protocol. Afwijkingen kunnen visueel worden geconstateerd. In figuren op deze pagina is dit voor de gehele dataset gedaan.

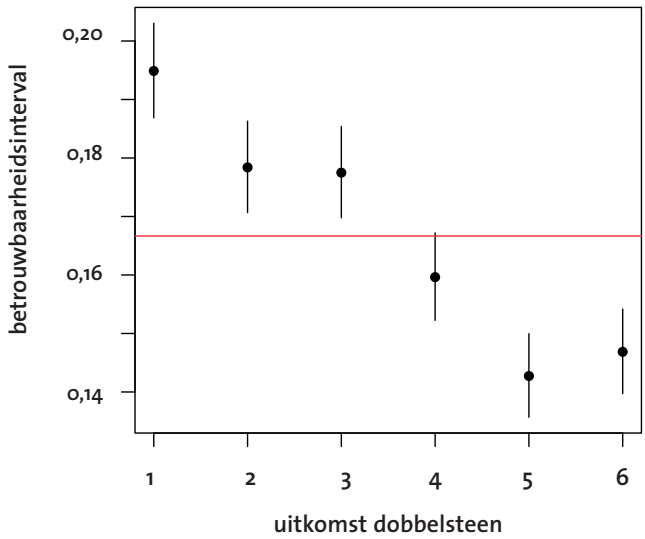
De beschreven procedure kan voor iedere helper afzonderlijk worden uitgevoerd. We hebben ons beperkt tot de 12 helpers die meer dan 180 proeven hebben uitgevoerd. Bij significantieniveau 0,01 en Bonferroni correctie voor meervoudig toetsen, blijkt dat de nulhypothese voor slechts één helper niet verworpen wordt.

Laagst- en hoogstvoorkomende frequentie per helper

Een andere manier die we gebruikt hebben om de uitkomsten van de toetsen te verduidelijken is door de eenvoudige toetsingsgrootheden laagst- en hoogstvoorkomende frequentie te beschouwen. Dit kunnen we voor iedere helper doen. We doen dit aan de hand van helper H. Er zijn gege-



Figuur 1. Linksboven: staafdiagram van de frequenties waarmee de verschillende dobbelsteenuitkomsten in de dataset voorkomen; linksonder: bijbehorend plaatje met residuen; de 2 figuren rechts zijn op dezelfde wijze verkregen, zij het dat de data nu onder de nulhypothese gesimuleerd zijn



Figuur 2. Betrouwbaarheidsintervallen voor kansen op de zes verschillende dobbelsteenuitkomsten

vens van 472 proeven voor deze helper. Bij zuiver dobbelen zouden we verwachten dat ieder van de schema's met frequentie $472/36 \approx 13{,}1$ voorkomt. Door toevalsvariatie zal dit nooit precies gebeuren. We concentreren ons nu op de hoogste en laagste frequentie. Deze geven immers de grootste afwijkingen van 13,1 naar boven en beneden respectievelijk. Voor helper A zijn dit 26 (bij schemanummer 2,1) en 1 (bij schemanummer 5,2). We vragen ons af hoe waarschijnlijk deze extreme frequenties zijn, indien we zouden werpen met twee zuivere dobbelstenen. Om dit te onderzoeken bootsen we het experiment met 2 zuivere dobbelstenen op de computer na. In ieder experiment simuleren we 472 worpen met 2 zuivere dobbelstenen. Vervolgens noteren we de laagste en hoogste frequentie. Dit gehele experiment herhalen we een groot aantal keren (we hebben gekozen voor 10.000 keer).

In de simulaties wordt slechts 7 keer als laagste frequentie 1 of lager verkregen.

Voor de hoogste frequentie, vinden we in 331 van de 10.000 gevallen een frequentie groter dan of gelijk aan 26.

Bij de uitgevoerde Chi-kwadraat-toets vinden we voor deze helper een veel kleinere p-waarde, namelijk 0,0000333. Dit wordt veroorzaakt doordat deze toets alle afwijkingen van 13,1 meeneemt, en we hier slechts kijken naar de laagste en hoogste frequentie.

Betrouwbaarheidsintervallen

Als we veronderstellen dat bij iedere proef een gelijkwaardige dobbelsteen is gebruikt (dat wil zeggen, met identieke kansen op een één, twee, drie, vier, vijf of zes), dan kunnen we op grond van de beschikbare data betrouwbaarheidsintervallen construeren voor de kansen op ieder van de zes dobbelsteenuitkomsten. Hiervoor zijn meerdere methoden voorgesteld in de literatuur. We maken hier gebruik van de methode zoals voorgesteld in Bailey, welke gebaseerd is op een betrouwbaarheidsinterval voor de succesparameter van een multinomiale verdeling, gebruik-

makend van een variantie-stabiliserende transformatie. Een Bonferroni correctie waarborgt het gewenste betrouwbaarheidsniveau, waarvoor we 95% gekozen hebben. In figuur 2 is voor iedere mogelijke dobbelsteenuitkomst (horizontaal), verticaal het verkregen betrouwbaarheidsinterval weergegeven. De rondjes zijn de geobserveerde fracties; de horizontale lijn ter hoogte 1/6 dient als referentie voor een zuivere dobbelsteen. We beschouwen hier de gehele dataset met alle proeven, waarop ook figuur 1 gebaseerd is.

We zien dat de data een duidelijk onzuivere dobbelsteen suggereren. Als er echt gedobbeld is door de helpers, dan zien we wederom dat het onwaarschijnlijk is dat dit gebeurd is met een zuivere dobbelsteen.

Conclusie

De geobserveerde frequenties van de uitlegschema's corresponderen niet met wat je zou mogen verwachten bij werpen met een zuivere dobbelsteen, hetgeen het protocol voorschrijft.

Literatuur

Bailey, B. J. R. (1980). Large sample simultaneous confidence intervals for the multinomial probabilities based on transformations of the cell frequencies. *Technometrics* 22(4), 583–589.

Frijters, J. E. R. (2006). De geuridentificatieproef in het licht van het falsificatiebeginsel. *Nederlands Juristenblad* 17, 945--948.

Frijters, J. E. R. (2008). Dobbelen en positievoorkeuren bij canine geuridentificatieproeven. *Expertise en Recht* 2008-1, 27--34.

Keuringsreglement politiespeurhond menselijke geur. Bijlage 1 bij de regeling politiehonden.

Geurt Jongbloed *is hoogleraar Mathematische Statistiek aan de Technische Universiteit Delft.*
E-mail: <G.Jongbloed@tudelft.nl>.

Frank van der Meulen *is universitair docent Statistiek aan de de Technische Universiteit Delft.*
E-mail: <f.h.vandermeulen@tudelft.nl>.

AGENDA

14-16 november 2011

Voor de **Stochastics Meeting Lunteren 2011** zijn zes sprekers uitgenodigd, die elk twee lezingen zullen geven. De sprekers zijn: L Peter Bickel (Asymptotic statistical inference for unlabelled graphs), Alexander Holroyd (Invariant Matching. Part), Niels Keiding (The current duration approach to estimating the distribution of time to pregnancy; Standardization vs. modelling for confounder control in observational studies: A historical perspective), Wilfrid Kendall (Random lines and effective transportation networks; Shy couplings, CAT(0) spaces, and the lion and man), Sylvie Meleard (Random modeling for adaptive dynamics), Markus Reiss (Estimation of the Levy triplet: Statistical inverse problem and Donsker Theorem).
Voor meer informatie: Marie-Colette van Lieshout, telefoon (020) 5924008; <http://homepages.cwi.nl/~colette/lunteren2011.html>.

12 december 2011

Lezingmiddag over **Nulhypothesetoetsing in de Sociale Wetenschappen** op 12 december van 13.30-17.30 uur, Universiteit van Amsterdam. Sprekers zijn Cees Glas, Rink Hoekstra, Rens van de Sloot, Eric-Jan Wagenmakers en Jelte Wicherts.

9-14 juli 2012

Het achtste **World Congress in Probability and Statistics** vindt volgend jaar in Istanbul plaats. Het programma biedt een breed scala aan onderwerpen. Zo worden onder meer de recente ontwikkelingen in de statistiek en kansrekening belicht en krijgen actuele onderzoeksthema's veel aandacht.
Voor meer informatie: <www.worldcong2012.org>.