Edith de Leeuw,
Universiteit Utrecht en
MethodikA, Amsterdam,
edithl@xs4all.nl

Peter van der Heijden,
Universiteit Utrecht,
p.vanderheijden@fss.uu.nl

Gevoelige vragen[1]

Iedere onderzoeker weet dat vragen naar gevoelige onderwerpen problemen met zich mee kunnen brengen. We zeggen wel dat alles bespreekbaar is, maar in de praktijk valt dit toch tegen. De respondenten willen eigenlijk het ware antwoord toch niet geven en dit resulteert dan in sociaal wenselijke antwoorden of verkapte weigeringen als 'weet-niet'. Survey-methodologen houden zich al lang bezig met de vraag hoe het beste naar gevoelige informatie te vragen, en een van de meer succesrijke methoden is de randomized-response-procedure.[2]

Hoe werkt nu randomized response'? Stelt u zich voor dat een interviewer u vraagt of u wel eens plagiaat hebt gepleegd. In plaats van direct met 'ja' of 'nee' te antwoorden, wordt u gevraagd met twee dobbelstenen te gooien zodat de interviewer niet ziet wat u gooit. U krijgt daarbij de volgende instructie: 'als u 2, 3, of 4 gooit zeg altijd ja, als u 11, of 12 gooit zegt u altijd nee, en bij 5, 6, 7, 8, 9 of 10 geeft u het echte antwoord op de vraag'. De interviewer noteert het antwoord 'ja' of 'nee', maar weet niet wat het antwoord betekent: hij heeft immers niet gezien wat u precies gooide! Niemand weet of u 'ja' zei omdat u toevallig een 5 gooide of omdat u een 4 gooide; uw privacy is beschermd door de dobbelstenen. Hierdoor wordt de extrinsieke dreiging van de vraag geminimaliseerd. De respondent is veilig.

Maar hoe moet de onderzoeker nu verder? Zij wil natuurlijk graag weten welk percentage van de wetenschappers plagieert, en of plagiaat vaker voorkomt in de sociale wetenschappen of in de natuurwetenschappen, of vaker onder hoogleraren dan onder Aio's. Het antwoord op de eerste onderzoeksvraag is met behulp van de kansrekening vrij eenvoudig te geven. Het antwoord op de andere vragen is wat lastiger, het gaat hier immers om het verband tussen twee variabelen waarbij een van deze twee variabelen door middel van randomized response tot stand is gekomen. Maar ook hier kan de statistiek uitkomst bieden, al is het niet eenvoudig.

Eerst maar de wat simpeler berekeningen voor de vraag naar het percentage plagiërende wetenschappers. Stel dat 1/3 van de wetenschappers 'ja' heeft gezegd, hoeveel hiervan heeft nu 'ja' gezegd omdat ze plagieerden en hoeveel omdat het nu eenmaal van de dobbelsteen moest? Als men 2, 3, of 4 gooide moest men 'ja' zeggen; nu zijn er zes manieren om 2, 3 of 4 met twee dobbelstenen te gooien, terwijl er in totaal 36 mogelijke worpen zijn. Dus, de kans op een door de dobbelsteen afgedwongen 'ja' is 6/36, oftewel 1/6. De kans op een gedwongen 'nee', bij een worp van 11 of 12, is 3/36, oftewel 1/9. De kans op 5, 6, 7, 8, 9 of 10 gelijk is aan 1-(6/36 + 3/36) = 3/4. Dus, de kans op het moeten beantwoorden van de echte vraag is 3/4.

Hoe schatten we nu de kans op plagiaat? Stel P(forced ja) is de kans op geforceerd 'ja' moeten zeggen.P(echte vraag) is de kans dat de echte vraag beantwoord moet worden.P(plagiaat) is de kans dat een respondent plagiaat heeft gepleegd, en deze kans willen we schatten. P(observed ja) is de kans dat een respondent 'ja' zegt. We kunnen nu schrijven: P(observed ja) = P(forced ja) + P(echte vraag) * P(plagiaat). Dit leidt tot de gezochte formule: P (plagiaat) = {P(observed ja)-P(forced ja)} / P(echte vraag). Invullen van onze voorbeeld gegevens geeft dan (1/3-1/6)/(3/4)= 4/18 = .222. Oftewel, volgens onze schatting, antwoordt tweeëntwintig procent van de respondenten dat ze het 'ongewenste', gevoelige gedrag wel eens hebben vertoond.[4]

De onderzoekster wil echter niet alleen de onderzochte groep respondenten beschrijven, ze wil ook een schatting maken van het percentage met ongewenst gedrag in de populatie: de inferentiële statistiek komt nu om de hoek kijken. De volgende stap is het berekenen van het betrouwbaarheidsinterval rondom de schatting. Stel dat de schatting gebaseerd is op een steekproef van 300 mensen. Bij elke aselecte steekproef heeft het toeval een rol gespeelt in die zin dat je net zo goed andere personen in je steekproef had kunnen aantreffen. Als er geen randomized response was gebruikt, maar recht toe rechtaan gevraagd was naar plagiaat dan kan de basisstatistiek zonder problemen worden toegepast. Het 95% betrouwbaarheids-interval rond de .222 is dan (.175;.269). Als de onderzoekster dit betrouwbaarheidsinterval te groot vindt, dan moet zij zorgen dat de steekproefomvang groter is dan 300. Want, hoe groter de steekproef, hoe kleiner de standaardfout en hoe smaller het betrouwbaarheidsinterval!

Bij gebruik van randomized response als vraagvorm in een steekproef speelt het toeval op twee manieren een rol. In de eerste plaats hadden andere mensen in de steekproef terecht kunnen komen dan de mensen die er nu inzitten. Dit is niet verschillend van steekproefonderzoek met directe vragen (zie boven). In de tweede plaats hadden de dobbelstenen anders kunnen rollen. Het zal duidelijk zijn dat de toevalsfluctuatie in een randomized-response-onderzoek daarom groter is. In ons plagiaat voorbeeld is het betrouwbaarheidsinterval rondom de schatting van .222 gelijk aan (.151; .293). Een aanzienlijk verschil! In beide gevallen was de steekproefgrootte 300, maar de manier van vraagstellen was anders. De conclusie is dat, om in een randomized response onderzoek een even groot betrouwbaarheidsinterval te krijgen als in een gewoon onderzoek, er in het randomized response onderzoek een veel grotere steekproef nodig is. Randomized response geeft meer valide antwoorden, maar brengt door de grotere steekproef meer kosten met zich mee: 'there is no such thing as a free lunch'. Toch kiezen veel onderzoekers voor de duurdere randomized-response-procedure. Ze kunnen het zich niet permitteren de randomized response niet te gebruiken, want dan zouden de respondenten waarschijnlijk niet eerlijk antwoorden op gevoelige vragen en zou ongewenst gedrag onderschat worden.[5]

Bij gebruik van randomized response kunnen we het percentage ongewenst gedrag goed schatten, maar hoe staat het nu met de overige onderzoeksvragen, de vragen naar verbanden tussen variabelen? Wij bespreken dit voor categorische en continue variabelen afzonderlijk.

Een voorbeeld voor een categorische variabele in het plagiaatonderzoek is de variabele 'type personeel' met categorieën 'aio', 'hoogleraar' en 'overig wetenschappelijk personeel'. Als onze voorbeeld onderzoekster wil onderzoeken of het percentage plagieerders onder hoogleraren verschilt van het percentage onder de aio's en de overigen, dan is dit eenvoudig te doen door de berekeningen die we zojuist hebben laten zien uit te voeren voor elk van deze drie groepen afzonderlijk. Dus, voor de gehele groep werd gevonden dat 1/3 'ja' had gezegd, maar het kan zijn dat dit voor hoogleraren .40 is, voor aio's .25, en voor overigen .35. Door deze getallen in de formule P (plagiaat) = {P(observed ja)-P(forced ja)} / P(echte vraag) te stoppen, vinden we als schatting van de kans op plagiaat bij de hoogleraren .31, bij aio's .11 en bij de overigen .24. Ook nu weer kunnen betrouwbaarheids intervallen voor deze schattingen opgesteld worden. Een belangrijke vraag is of de gevonden verschillen tussen .31, .11 en .24 op toeval berusten of niet, met andere woorden zijn deze verschillen statistisch significant? Dit kan onderzocht worden door een chi-kwadraattoets uit te voeren op de 3 x 2 kruistabel van 'type personeel' bij *geobserveerde* 'ja/nee'-antwoorden.[6] Is deze toets significant, dan is het verschil tussen .31, .11 en .24 significant.

Een voorbeeld voor een continue variabele is in ons plagiaatonderzoek de variabele 'de kansinschatting dat het plagiaat ontdekt wordt', die zou kunnen lopen van 0% tot 100%. Normaliter is het zo dat, als logistische regressie gebruikt wordt om het verband te onderzoeken tussen een afhankelijke variabele met twee categorieën en een onafhankelijke variabele die continu is. Nu de afhankelijke variabele echter een door randomized response verkregen variabele is, gaat onze aandacht niet uit naar de relatie tussen de continue variabele en de *geobserveerde* 'ja/nee'-antwoorden, maar naar de relatie tussen de continue variabele en de *echte* antwoorden over plagiaat. Om dit te kunnen onderzoeken hebben we een logistische regressie-procedure moeten ontwikkelen die dit mogelijk maakt.[7] Uit die procedure kan bijvoorbeeld komen dat, bij een waarde van 0 % op de onafhankelijke variabele 'de kansinschatting dat het ontdekt wordt', de geschatte kans op plagiaat .60 is, bij een waarde van 50 % zou de geschatte kans op plagiaat .40 kunnen zijn, en bij

een waarde van 100 % zou de waarde .25 kunnen zijn. In dit geval lijkt de relatie tussen de continue onafhankelijke variabele en de randomized-response-variabele zeker aanwezig, maar een statistische toets zal moeten uitwijzen of de relatie significant is, d.w.z. niet op toeval berust.

Tot slot: Als algemene lijn bij de statistische analyses is het verhelderend het volgende beeld voor ogen houden. We proberen een gevoelig onderwerp te meten. Om de respondenten veiligheid te bieden, 'gooien' we een hoeveelheid ruis in de data door de dobbelsteen te laten bepalen of men antwoord op de gevoelige vraag of simpelweg 'ja' of 'nee' zegt. Hierdoor kunnen we *individuele geobserveerde* 'ja/nee'-antwoorden niet meer vertrouwen. Echter, wij begrijpen heel goed hoe de ruis die we zelf in de data hebben gegooid in elkaar zit, en daarom kunnen we die ruis eruit halen op het moment dat we over de *gehele* steekproef uitspraken willen doen. Voor dat doel zijn allerlei analyse technieken aangepast die men gewoonlijk ook toepast. Hierdoor zijn binnen analysetechnieken die men normaliter zou willen gebruiken, na aanpassing, toch de werkelijke antwoorden op de gevoelige onderwerpen te schatten. Zo verkrijgen we valide antwoorden op gevoelige vragen. Maar, we betalen hiervoor doordat we grotere steekproeven nodig hebben om eenzelfde betrouwbaarheid te bereiken als die we binnen een onderzoek zonder randomized-response-vragen mogen verwachten.

1. Dit is de laatste in een serie MEtT-kolommen over vragen naar gevoelige informatie in enquêtes.
Een algemene inleiding verscheen onder de titel *Vragen naar gevoelige informatie* van Gerty Lensvelt-Mulders en Edith de Leeuw in FACTA 10(2002), 3, p.34-35. Een reactie hierop verscheen in Facta 10 (2002), 4, p. 26. In Facta 10 (2002), 5, p.28-30 beschreven Gerty Lensvelt-Mulders & Edith de Leeuw in de kolom 'Beschermd door een dobbelsteen' hoe randomised-response-technieken werken om de privacy van de respondent te beschermen en opener antwoorden op gevoelige vragen te krijgen.
2. Zie ook: G.J.M. Lensvelt-Mulders, J.J. Hox, & P.G.M. van der Heijden (2003, under review) *Meta-analysis of randomised response research: 35 years of validation.*
3. In de Teleac-serie JOTA is een korte video beschikbaar over de randomised-response-techniek. U kunt deze gratis via internet bekijken op http://www.teleacnot.nl/sites/jota-tv/ Het betreft de uitzending van 03-02-2002 getiteld *Toevallig Eerlijk.*
4. Deze berekening ziet er misschien ingewikkeld uit, maar het is aardig te weten dat dit onderwerp enkele jaren geleden deel uit maakte van het VWO-eindexamen wiskunde.
5. G. Lensvelt-Mulders (2003) Randomized-response-technieken als instrument voor het onderzoek van sociaal gevoelige onderwerpen. In: A.E. Bronner et al (Red). Ontwikkelingen in het marktonderzoek, jaarboek van de Marktonderzoek Associatie 2003.Haarlem: De Vrieseborch.
6. Zie de appendix in: P.G.M. van der Heijden , G. van Gils, J. Bouts and J. Hox (2000). *A comparison of randomized response, CASAQ, and direct questioning; eliciting sensitive information in the context of welfare and unemployment benefit. Sociological Methods and Research, 28,* 505-537.
7. Zie A.van den Hout, en P.G.M. van der Heijden. *The analysis of multivariate misspecified data, with special attention to randomized response data.* Ter publicatie aangeboden.

SISWO/Instituut voor Maatschappijwetenschappen
Plantage Muidergracht 4, 1018 TV Amsterdam
tel. 020 5270600, fax 020 6229430
e-mail: siswo@siswo.uva.nl
internet: http://www.siswo.nl

Bijeenkomsten

De ruimtelijke orde
SISWO, 16 oktober 2003
DE werkgroep Stedengeschiedenis organiseert in samenwerking met het SISWO een studiemiddag getiteld 'de ruimtelijke orde'. Dagvoorzitter: Dr. Michiel Wagenaar. Programma:
- Drs. Jan Hein Furnée - Standen, sekse en plaatsen van stedelijk vertier, Den Haag 1850-1890
- Dr. Koos Bosma - Een schuilplaats in de letteren, de eerste planoloog J.M. de Casseres, 1902-1990'
- Drs. Cordula Rooijendijk - Beeldvorming over de ruimtelijke ordening van Amsterdam en Rotterdam in publieke debatten 1945-2000
De studiedag is gratis toegankelijk. Informatie kan worden ingewonnen bij Jan van den Noort, tel. 010 4366014.

SWOME-bijeenkomst
SISWO, 28 oktober 2003
SWOME organiseert een studiemiddag over landenvergelijkend onderzoek naar duurzame energie met Valentina Dinica (proefschrift over liberaliseren en verduurzamen) en Bas van Vliet (proefschrift over privatisering van de watersector).
Informatie: Drs. Otto Nuys, tel. 020 5270621, e-mail: nuys@siswo.uva.nl

15e onderwijssociologische conferentie
De strijd om het curriculum
Lunteren, 12 en 13 november 2003
Inlichtingen: Henk Kleijer, tel. 020 5270647, e-mail: kleijer@siswo.uva.nl

NOSMO Methodologendag
'State of the art & science' van het sociaalwetenschappelijk onderzoek
Amsterdam (VU), 21 november 2003
Inlichtingen: Henk Kleijer, tel. 020 5270647, e-mail: kleijer@siswo.uva.nl

Elfde Sociaal-wetenschappelijke studiedagen
Amsterdam (Casa 400), 22 & 23 april 2004
Stuur uw abstract (maximaal 250 woorden) voor 15 januari 2004 op aan het congressecretariaat, e-mail: studiedagen@siswo.uva.nl
Voorzie uw abstract voor een paper van een titel en geef aan, in welke sessie u wilt participeren. Vermeld uw naam en uw correspondentieadres. Zie voor een uitvoeriger Call for papers, met thema's en coördinatoren de SISWO-site: www.siswo.nl
De Studiedagen hebben als plenair thema *De Ontregelde Samenleving*, toegelicht door prof. Mark Elchardus (Vrije Universiteit Brussel).
Organisatie: SISWO/Instituut voor Maatschappijwetenschappen, in samenwerking met de Vlaamse Vereniging voor Sociologie (VVS) en de Nederlandse Sociologische Vereniging (NSV). Inlichtingen: Marycke Borghardt, e-mail borghardt@siswo.uva.nl en Ente Breed, e-mail: studiedagen@siswo.uva.nl. Zie voor actuele informatie de SISWO-website: www.siswo.nl