Prof.dr. Jelke Bethlehem

De kwaliteit van internetpeilingen

Universiteit Leiden

Universiteit Leiden. Universiteit om te ontdekken.

De kwaliteit van internetpeilingen

Oratie (in verkorte vorm) uitgesproken door

Prof.dr. Jelke Bethlehem

bij de aanvaarding van het ambt van bijzonder hoogleraar
op het gebied van Survey-Methodologie,
in het bijzonder met Behulp van het Internet,
vanwege het Centraal Bureau voor de Statistiek
aan de Universiteit Leiden
op maandag 28 januari 2013

Universiteit Leiden

Mijnheer de Rector Magnificus, leden van de directie van het Centraal Bureau voor de Statistiek, leden van het Curatorium van deze bijzondere leerstoel, dames en heren,

Meten peilingen wat ze moeten meten?

Op 12 september 2012 waren er verkiezingen voor de Tweede Kamer. Een korte, maar hevige, verkiezingscampagne ging er aan vooraf. Daarin volgden de peilingen elkaar in hoog tempo op. Er waren minsten vier peilers actief. Dat waren Peil.nl (Maurice de Hond), Ipsos Synovate (Politieke Barometer), TNS NIPO en GfK Intomart (De Stemming). Er waren zelfs peilers die elke dag met een nieuwe peiling kwamen.

Hoe goed zijn die peilingen? Kunnen ze echt de verkiezingsuitslag voorspellen? Dit soort vragen komen op als je de verkiezingsuitslag vergelijkt met de laatste peilingen. Er zijn behoorlijke verschillen te zien, vooral bij de politieke partijen waarvan de aanhang flink in beweging is.

Tabel.1. Prognoses voor de zetelverdeling bij de verkiezingen voor de Tweede Kamer van 12 september 2012.

	Verkiezings-uitslag	Peil.nl	Politieke Barometer	TNS NIPO	De Stemming
VVD	41	36	37	35	35
PvdA	38	36	36	34	34
PVV	15	18	17	17	17
CDA	13	12	13	12	12
SP	15	20	21	21	22
D66	12	11	10	13	11
GroenLinks	4	4	4	4	4
ChristenUnie	5	5	5	6	7
SGP	3	3	2	2	3
PvdD	2	3	3	2	2
50PLUS	2	2	2	4	3
Zetels verschil		18	18	24	24
Gemiddeld verschil		1,6	1,6	2,2	2,2

Tabel 1 toont de prognoses van de peilers vlak voor de verkiezingen. Die prognoses zitten er soms flink naast. Het verschil is zelfs zeven zetels bij de voorspelling van De Stemming voor de SP. Vier keer is het verschil zes zetels en twee keer is het verschil vijf zetels. Die afwijkingen zijn te groot. De peilingen zijn weliswaar gebaseerd op steekproeven, waardoor er onzekerheid in voorspellingen zit, maar dat verklaart nog geen verschil van vijf of meer zetels. Er moet dus iets anders aan de hand zijn.

De peilers zelf vonden dat er niets mis was. Eén peiler vond de peilingen juist heel goed. Volgens hem hadden de peilers niet de ambitie de verkiezingsuitslag te voorspellen. Het ging niet om een prognose, maar om een slotpeiling. Volgens een andere peiler hadden de kiezers op het laatste moment, dus na de laatste peiling, ineens besloten om strategisch te gaan stemmen.

Er is nog een andere mogelijke verklaring voor de afwijkingen in de voorspellingen en dat is dat die peilingen niet goed meten wat ze zouden moeten meten. Het zijn geen valide meetinstrumenten. Er zit een systematische vertekening in de uitkomsten. Aan het opzetten en uitvoeren van een peiling zitten nogal wat methodologische haken en ogen. Een peiling die niet voldoet aan de methodologische regels kan al snel afwijkende uitkomsten opleveren.

Het probleem van de niet-valide peilingen is veel groter geworden door de opkomst van het internet. Daardoor is het wel erg eenvoudig geworden om een peiling op te zetten en uit te voeren. Een internetpeiling levert weliswaar snel en goedkoop veel gegevens op, maar het is de vraag of de uitkomsten valide zijn. Als een onderzoeker niet de methodologische principes voor peilingen toepast, dan kan hij niet op wetenschappelijk verantwoorde wijze conclusies trekken uit het onderzoek. Helaas heeft de praktijk van de laatste jaren geleerd dat er bij peilingen veel kaf zit tussen het koren.

Ik wil het vandaag met u hebben over de kwaliteit van internetpeilingen. Ik wil proberen een antwoord te vinden op de vraag wat er voor nodig is om van een internetpeiling een valide meetinstrument te maken. Daarbij gaat het niet alleen om peilingen die de publieke opinie meten. Ook overheidsinstituten als het CBS en het SCP voeren peilingen uit. Alleen worden ze dan surveys of enquêtes genoemd. Die onderzoeken zijn vaak groter en ingewikkelder, maar het principe blijft hetzelfde. Een steekproef van personen vult een vragenlijst in. Vervolgens worden de zo verzamelde gegevens gebruikt om uitspraken te doen over de hele populatie.

Peilingen door de eeuwen heen

De mensheid is al van oudsher bezig met het verzamelen van statistische gegevens, overigens zonder het zo te noemen. Eigenlijk gebeurt het al sinds het begin van de menselijke beschaving. Het waren vooral koningen, keizers en andere heersers die gegevens nodig hadden voor het besturen van hun landen of rijken.

Al 1000 jaar voor Christus gaven de heersers van China en Egypte opdracht om statistieken te maken. Ze gebruikten deze gegevens voornamelijk voor het heffen van belastingen en militaire zaken. Ook de Romeinse keizers organiseerden regelmatige tellingen van mensen en hun bezittingen. Ze gebruikten de verzamelde gegevens om de politieke status van de inwoners te bepalen en om hun militaire en financiële verplichtingen vast te stellen. Bekend is het verhaal van keizer Augustus die omstreeks het jaar 0 een volkstelling uitschreef waarvoor Jozef en Maria naar Bethlehem moesten reizen.

Bij al dit soort onderzoek ging het om integraal onderzoek. Iedereen in de populatie moest meedoen. Er was geen sprake van steekproeven. Het idee was nog niet opgekomen dat je ook wel nauwkeurige statistieken zou kunnen maken op basis van slechts een deel van de gegevens.

Een mooi voorbeeld van een integrale peiling, maar wel van al wat latere datum, is het *Domesday Book*. Willem de Veroveraar gaf opdracht tot het samenstellen van dit boek nadat hij in 1086 Engeland had veroverd vanuit het Franse Normandië. Hij wilde weten wat hij had veroverd. Het Domesday Book was het resultaat van een integraal onderzoek van de bezittingen van de koning en zijn vazallen in Engeland. Iedereen die informatie kon verstrekken, moest voor een commissie verschijnen. Die werkte met een vaste vragenlijst. Daarin werd, bijvoorbeeld, gevraagd wie de eigenaar van een stuk land was, hoeveel vrije mensen en slaven er waren, wat er aan oppervlakte bos en grasland was, hoeveel molens en visvijvers er waren, wat de totale waarde van het gebied was, en wat de winstvooruitzichten waren. Zo kreeg Willem de Veroveraar een statistisch overzicht met gegevens over meer dan 13.000 dorpen en riddergoederen.

Een ander mooi voorbeeld van integraal onderzoek vinden we terug in het rijk der Inca's, dat zo tussen de 10e en 14e eeuw zijn hoogtepunt bereikte. Elk Incadistrict had een *quipucamayoc*. Dit was in feite een soort statisticus. Hij verzamelde gegevens over zaken als het aantal onderdanen, het aantal huizen dat zij bewoonden, hoeveel lama's er op de weiden graasden en het aantal jonge mannen dat geschikt was voor het leger. De quipucamayocs legden dat allemaal vast op quipu's. Een *quipu* was een systeem van geknoopte koorden van diverse kleuren. Elke kleur gaf een bepaald onderwerp aan, en de knopen de aantallen, volgens het decimale systeem. De quipu was dus een vroege voorganger van een enquêteformulier.

De eerste moderne volkstelling in Noord-Amerika vond plaats in 1666 in Canada. Jean Talon was Intendant (gouverneur) van Nieuw-Frankrijk (Nouvelle-France). Hij wilde weten hoe het gebied zich had ontwikkeld sinds de stichting van Québec in 1608. Hij registreerde van alle inwoners geslacht, leeftijd, burgerlijke staat en beroep. Er bleken op dat moment 3.215 mensen te wonen in Nieuw-Frankrijk.

De Scandinavische landen liepen voorop bij de volkstellingen in Europa. De eerste volkstelling in Zweden vond plaats in 1748. De staat en de kerk hadden er belang bij. De staat wilde weten hoeveel mannen er konden worden opgeroepen voor militaire dienst en de kerk wilde in de gaten houden hoeveel mensen het protestantse geloof aanhingen. De eerste volkstelling werd in Denemarken georganiseerd in 1769.

Den Dulk en Van Maarsseveen (1990) beschrijven de ontwikkelingen in Nederland. De eerste volkstelling vond plaats in 1795. Nederland stond toen onder Franse invloed. Het was de periode van Bataafse Republiek. Er was een nieuw gecentraliseerde bestuurd en dat wilde nieuwe kiesdistricten maken. Daarvoor moest men weten hoe de verdeling van de bevolking over het land was.

Het ontstaan van steekproefonderzoek

In de jaren voor 1895 waren het voornamelijk de nationale statistische bureaus die zich bezig hielden met het verzamelen van statistische gegevens. Die bureaus deden altijd integraal onderzoek. Het trekken van steekproeven was taboe. Waarom zou je een steekproef trekken als het ook met een integrale telling kon? En bovendien was het ongepast om echte waarnemingen te vervangen door rekenkundige manipulaties. Het was een vorm van discriminatie om een groot deel van de mensen uit te sluiten van deelname aan een onderzoek. Ook leek het de statistici in die tijd een onmogelijke zaak om een uitspraak over een hele bevolking te doen als je maar gegevens over een klein deel daarvan had.

Een belangrijke doorbraak vond plaats in het jaar 1895. In dat jaar kwam het ISI (International Statistical Institute) bijeen in het Zwitserse Bern. Het was Anders Kiær (1895, 1997), de directeur van het Noorse Statistische Bureau, die daar een pleidooi hield voor het gebruik van steekproeven. Hij betoogde dat je met zijn "Representatieve Methode" goede resultaten kon behalen zolang de steekproef maar een kopie op kleine schaal was van de populatie. Op grond van kenmerken die voor elk object in de populatie bekend waren, selecteerde hij objecten zodanig dat de verdeling van de kenmerken in de steekproef overeen kwam met die in de populatie. Kiær zorgde er bijvoorbeeld voor dat de verhouding man-vrouw in de steekproef overeen kwam met de verhouding man-vrouw in de populatie. En ook de verdeling over grote steden en platteland maakte hij kloppend.

Bij de selectie van de steekproef maakte Kiær geen gebruik van loting. Hij zocht doelgericht net zolang mensen bij elkaar tot hij een steekproef had met de gewenste samenstelling. We zouden dat nu een *quota-steekproef* noemen.

Een probleem van de Representatieve Methode van Kiær was dat hij geen idee had hoe goed of slecht zijn schattingen waren. Andere statistici vonden dit een ernstig nadeel. Daarom was er jarenlang veel discussie over de toepassing van de Representatieve Methode in de praktijk.

Het was Arthur Bowley (1906, 1926) die voor het eerst een theorie ontwikkelde waarmee we de onzekerheid in de uitkomsten van een peiling konden kwantificeren. Hij stelde voor om steekproeven te loten. Voordeel daarvan is dat je dan allerlei resultaten uit de theorie van de kansrekening kan toepassen. Bowley toonde bijvoorbeeld aan dat schattingen bij benadering een *normale verdeling* hebben. Vervolgens kon hij uitrekeningen hoe ver schatting en werkelijke (te schatten) waarde maximaal van elkaar af kunnen liggen.

De Poolse wetenschapper Jerzy Neyman leverde in 1934 nog een aantal fundamentele bijdragen die de verdere ontwikkeling van de steekproeftheorie hebben bepaald. Zo introduceerde hij het *betrouwbaarheidsinterval* als instrument om de precisie van een schatting te bepalen. Dat betrouwbaarheidsinterval gebruiken we nog steeds om de onzekerheidsmarges van schattingen aan te geven.

Neyman liet ook zien wat de risico's zijn als je niet loot. Uit experimenten met gegevens uit een Italiaanse volkstelling bleek dat zulke steekproeven tot wezenlijk verkeerde schattingen kunnen leiden, ook al is de steekproef representatief naar een aantal achtergrondkenmerken. Neyman toonde in feite aan dat de kanssteekproef de enige wetenschappelijk verantwoorde manier is om op basis van een steekproef een conclusie te trekken over een hele populatie.

Dit wordt nog eens bevestigd door een belangrijk artikel van Horvitz en Thompson (1952). Zij lieten zien dat je altijd goede schattingen van kenmerken in de populatie kunt uitrekenen als aan twee voorwaarden is voldaan:

- de steekproef moet zijn geloot uit de hele populatie;
- iedereen in de populatie heeft een positieve (van nul verschillende) kans gehad om in de steekproef te komen.

Een goede schatting betekent dat er geen sprake is van systematische afwijkingen. Als je de trekking van de steekproef een groot aantal malen zou herhalen, dan levert dit, gemiddeld genomen, de te schatten waarde op. We noemen dat een *zuivere schatter*.

Een ander belangrijk voordeel van kanssteekproeven is dat je altijd de nauwkeurigheid van de schattingen kunt uitrekenen. Daardoor wordt het mogelijk de ruis van de steekproef te scheiden van echte effecten.

De eerste opiniepeilingen

De geschiedenis van de opiniepeilingen in de VS gaat terug naar 1824. In dat jaar probeerden twee Amerikaanse kranten, de *Pennsylvanian* in Harrisburg en de *Star* in Raleigh, de politieke voorkeur van de kiezers te meten in de periode voor de presidentsverkiezingen van dat jaar. De kranten gebruikten wel steekproeven maar ze hadden weinig aandacht voor de manier waarop die steekproef tot stand kwam. Daarom viel er weinig zinnigs te zeggen over de kwaliteit van de uitkomsten. Zulke peilingen werden *straw polls* genoemd. Die uitdrukking komt van het boerenland. Boeren gooiden een handvol strootjes in de lucht om te zien van welke kant de wind kwam. De kranten deden straw polls in de straten van de stad om te zien hoe de politieke wind waaide.

De Amerikaanse presidentsverkiezingen van 1936 waren een beslissend moment in de ontwikkeling van opiniepeilingen. In die verkiezingen namen de Democraat Franklin Roosevelt en de Republikein Alf Landon het tegen elkaar op. De leidende politieke peiler was in die tijd het tijdschrift *Literary Digest*. Het tijdschrift had 10 miljoen Amerikanen aangeschreven. De adressen waren afkomstig uit lijsten van eigenaren van auto's en uit telefoonboeken. Uiteindelijk vulden 2,4 miljoen Amerikanen de vragenlijst in.

In 1935 was er een nieuwe peiler in Amerika gekomen. Dat was *George Gallup*. Hij besefte dat je alleen goede voorspellingen kunt doen als de steekproef goed in elkaar zit. De steekproef moet representatief zijn. Hij maakte hiervoor gebruik van quota-steekproeven. Hij gaf instructies aan de enquêteurs over de aantallen mensen die ze in de verschillende groepen moesten enquêteren: zoveel vrouwen uit de middenklasse in de stad, zoveel mannen uit de lagere klasse op het platteland, enz. De omvang van de steekproef bedroeg bij Gallup 50.000. De steekproef van Gallup was dus aanzienlijk kleiner dan die van Literary Digest.

Tabel 2 bevat de voorspellingen van beide peilers en de echte uitslag van de verkiezingen. Literary Digest zat er helemaal naast. Die peiler voorspelde dat Landon de verkiezingen zou winnen met 57%. Maar het werd Roosevelt met 61%. Gallup voorspelde de winnaar wel goed, maar deze peiler zat er toch ook nog 5% naast.

Tabel 2. De Amerikaanse presidentsverkiezingen van 1936.

Kandidaat	Voorspelling Literary Digest	Voorspelling Gallup	Verkiezings-uitslag
Roosevelt (D)	43%	56%	61%
Landon (R)	57%	44%	37%

Waarom was de voorspelling van Literary Digest zo slecht? Dat kwam omdat de samenstelling van de steekproef niet goed was. De adressen waren die van eigenaren van auto's en van telefoonbezitters. Dat waren in die tijd mensen met wat hogere inkomens. Die mensen stemden vooral Republikeins. Dus de Republikeinen waren oververtegenwoordigd in de steekproef, met als gevolg dat er teveel Landon-stemmers in de peiling zaten.

De quota-steekproeven van Gallup werkten in de praktijk ook niet altijd goed. Dat bleek bij de presidentsverkiezingen van 1948. Toen nam de Democraat Harry Truman het op tegen de Republikein Thomas Dewey. In tabel 3 staat de voorspelling van Gallup en de werkelijke verkiezingsuitslag.

Tabel 3. De Amerikaanse presidentsverkiezingen van 1948.

Kandidaat	Voorspelling Gallup	Verkiezings-uitslag
Truman (D)	44%	50%
Dewey (R)	50%	45%

De steekproef van Gallup had een omvang van 3.250 personen. Op grond van de peiling voorspelde Gallup dat Dewey de verkiezingen zou winnen. Sommige kranten waren zo overtuigd van de voorspelling van Gallup dat ze in hun vroege edities Dewey al tot winnaar verklaarden.

Gallup voorspelde dat Dewey 50% van de stemmen zou krijgen, en dat was 5% meer dan Dewey in werkelijkheid kreeg. Net als bij de verkiezingen van 1936, zaten er teveel Republikeinen in de steekproef van Gallup. Alleen leidde dat in 1936 niet tot een verkeerde voorspelling, omdat daarvoor het verschil tussen Roosevelt en Landon te groot was. In 1948 waren de verschillen tussen de kandidaten kleiner. De afwijking in de steekproef van Gallup zorgde er toen wel voor dat Gallup met de verkeerde voorspelling kwam.

Oorzaak van de problemen met de voorspellingen van Gallup was dat hij met quota-steekproeven werkte. Dit soort steekproeven is niet gebaseerd op loting. Hij liet de enquêteurs porties mensen (quota) met bepaalde eigenschappen selecteren. Hij maakte zijn steekproeven representatief met betrekking tot variabelen als geslacht, leeftijd, opleidingsniveau en huidskleur. Maar dat betekent niet automatisch dat de steekproef ook representatief is met betrekking tot andere variabelen, zoals stemgedrag. Onderzoekers hebben inderdaad vastgesteld dat over een lange reeks van jaren de Republikeinen oververtegenwoordigd waren in dit soort quota-steekproeven.

Als gevolg van het fiasco van Gallup in 1948, besloot deze organisatie om af te stappen van het gebruik van quota-steekproeven. Vanaf dat moment werd alleen nog maar gebruik gemaakt van echte kanssteekproeven.

In Nederland zien we dat Unilever in 1934 het eerste marktonderzoekbureau opricht. Het heet Lintas (Lever's International Advertising Services). Na de Tweede Wereldoorlog onderzocht dit marktonderzoekbureau regelmatig het consumentengedrag met een panel van 600 huisvrouwen. Het is niet duidelijk hoe dit panel was opgezet en of het representatief was. In 1987 doopt Unilever Lintas om in Research International Nederland.

In 1940 ontstaat nog een ander onderzoeksbureau: de Nederlandse Stichting voor Statistiek (NSS). Dit bureau

kon worden gezien als de commerciële zuster van het CBS. De directeur van het CBS zat in de directie van het NSS. De belangrijkste activiteiten van het NSS waren marktonderzoek en opiniepeilingen.

In 1945 komt er nog een marktonderzoekbureau bij: het Nederlands Instituut voor de Publieke Opinie (NIPO). Het NIPO bracht in 1946 ook een tijdschrift uit: 'De Publieke Opinie'. Het eerste nummer legt uit dat je voor een goede peiling echt geen steekproef van 100.000 personen nodig hebt. Een omvang van 2.000 tot 10.000 is voldoende als de steekproef maar representatief is met betrekking tot kenmerken als inkomen, beroep, leeftijd en soms geloof. Het bureau was tegen schriftelijke peilingen. Dan zouden er namelijk teveel mensen in de steekproef zitten met een hoge intelligentie en een hogere sociaaleconomische positie. Zie ook NIPO (1946a).

Verkiezingen zijn altijd een mooie gelegenheid om te controleren of peilingen inderdaad doen wat ze moeten doen. De verkiezingen van 17 mei 1946 zijn een aardige illustratie daarvan. Tabel 4 vergelijkt de voorspelling van het NIPO met de werkelijke verkiezingsuitslag. De cijfers van het NIPO zijn gebaseerd op een peiling die twee weken voor de verkiezingen plaatsvond.

Tabel 4. De Nederlandse verkiezingen van 1946.

Partij	Verkiezings-uitslag	Voor-spelling NIPO	Verschil
Partij van de Arbeid	28,3 %	33,9 %	5,6 %
Katholieke Volkspartij	30,8 %	29,5 %	1,3 %
Anti-Revolutionaire Partij	12,9 %	10,3 %	2,6 %
Christelijk Historische Unie	7,8 %	6,6 %	1,2 %
Partij van de Vrijheid	6,4 %	9,5 %	3,1 %
Communistische Partij Nederland	10,6 %	7,9 %	2,7 %
Protestantsche Unie	0,7 %	0,5 %	0,2 %
Staatkundig Gereformeerde Partij	2,1 %	0,9 %	1,2 %
Bellamy-Partij	0,2 %	0,8 %	0,6 %
Groep Lopes	0,1 %	0,1 %	0,0 %
Gemiddelde verschil			1,8 %

NIPO was niet echt blij met de uitkomsten. Een gemiddeld verschil van 1,8% was toch wel erg groot. Vooral de voorspelling voor de Partij van de Arbeid zat er met een verschil van 5,6% behoorlijk naast. Als verklaring verwees NIPO naar de gebeurtenissen op de avond voor de verkiezingen. Toen was er een toespraak van premier Schermerhorn op de radio waarin hij aankondigde het leger te gaan inzetten om een staking te breken. Dat zou voor veel stemmers een reden zijn geweest om op het laatste moment van de Partij van de Arbeid over te stappen naar de Communistische Partij Nederland. Voor een meer gedetailleerde beschrijving van deze peiling wordt verwezen naar NIPO (1946b).

Meer over de introductie en het gebruik van kanssteekproeven en de discussie daarover is te vinden in Bethlehem (2009).

Peilingen via het internet

Met de razendsnelle opkomst van het internet deed een nieuwe methode van gegevensverzameling zijn intrede: peilingen via het internet. In de jaren 80 van de vorige eeuw werden al de eerste experimenten gedaan met e-mailpeilingen. Vragenlijsten in e-mails hebben echter nogal wat beperkingen. Het internet wordt pas echt interessant als in 1995 HTML 2.0 beschikbaar komt. HTML (HyperText Markup Language) is een opmaaktaal voor web-pagina's. De eerste versie van HTML was ontwikkeld door Tim Berners Lee in 1991. Sterk punt van HTML 2.0 was de mogelijkheid om formulieren op het scherm te maken. Ingevulde gegevens konden vervolgens worden verstuurd van de computer van de respondent naar de server van de onderzoeker.

Veel onderzoekers zien al snel de mogelijkheden van enquêteren via het World Wide Web. Maar de lage internetpenetratie maakt het lastig dit op grote schaal te gaan doen. Omdat in die begintijd bedrijven sneller overgingen tot het aanschaffen van internet dan huishoudens, werd eerst geëxperimenteerd met enquêtes onder bedrijven.

Enquêteren via het internet heeft een aantal aantrekkelijke eigenschappen:

- Het is een betrekkelijk eenvoudige manier om toegang te krijgen tot een zeer grote groep potentiële respondenten, namelijk iedereen met internet.
- De vragenlijsten kunnen tegen zeer lage kosten worden aangeboden: er zijn geen enquêteurs nodig, er zijn geen drukkosten (van papieren vragenlijsten) en er zijn ook geen verzendkosten.
- Een peiling kan heel snel worden uitgevoerd. Er hoeft maar weinig tijd verloren te gaan tussen maken en aanbieden van de vragenlijst.
- Het internet biedt aantrekkelijke extra mogelijkheden om zaken als beeld (foto, video) en geluid in de vragenlijst op te nemen.

Een peiling via het internet lijkt dus een snelle, goedkope en aantrekkelijke manier om veel gegevens te verzamelen. Het is echter niet allemaal rozengeur en maneschijn. De relatieve eenvoud waarmee een internetpeiling kan worden opgezet, leidt tot een groot en nog steeds groeiend aanbod van dit soort peilingen. Er zijn websites (zoals *SurveyMonkey* and *LimeSurvey*) waarmee iedereen in korte tijd een enquête in de lucht kan brengen, ook al heb je geen enkel verstand van survey-methodologie. Veel van die enquêtes zijn niet op verantwoorde wijze opgezet. En door het grote aanbod is het moeilijk het kaf van het koren te scheiden.

Ik wil een aantal methodologische aspecten van internetpeilingen met u bespreken. Het gaat om aspecten die tot onjuiste uitkomsten kunnen leiden. Het gaat om

- *Onderdekking*: Bij een interpeiling kunnen mensen zonder internet niet deelnemen aan het onderzoek.
- *Zelfselectie*: Er wordt geen aselecte steekproef getrokken. In plaats daarvan zet de onderzoeker de vragenlijst op het internet. Vervolgens roept hij op allerlei manieren mensen op de vragenlijst in te vullen.
- *Non-respons*: Er zijn personen die zijn geselecteerd voor de peiling, maar die de vragenlijst toch niet invullen.
- *Meetfouten*: De respondenten vullen de vragenlijst wel in, maar geven niet de juiste antwoorden op vragen.

Daar waar de betrouwbaarheid is aangetast, komt de vraag op of voor deze problemen kan worden gecorrigeerd. Daarvoor komen in de eerste plaats wegingstechnieken in aanmerking. Maar die zijn geen garantie voor succes.

Onderdekking

Steekproeven worden gewoonlijk niet uit een populatie zelf getrokken maar uit een administratieve weergave daarvan. Dit wordt het *steekproefkader* genoemd. De gemeentelijke Basis

Administratie (GBA) is een voorbeeld van zo'n steekproefkader. Een ander voorbeeld is een telefoonboek (voor een telefonisch onderzoek).

Onderdekking doet zich voor wanneer het steekproefkader niet exact de populatie afdekt. Er zijn dan personen die wel deel uitmaken van de populatie, maar toch niet voorkomen in het steekproefkader. Bij het GBA is de onderdekking klein. Bij een telefoonboek is de onderdekking groter, omdat ongeveer 1 op de 3 Nederlanders niet in het telefoonboek staat.

Bij internetpeilingen op basis van zelfselectie wordt het internet zelf als steekproefkader gebruikt. Ook hier hebben we te maken met onderdekking. Immers, niet elke Nederlander heeft thuis de beschikking over internet. Dat betekent dat mensen zonder internet nooit in een internetpeiling terecht kunnen komen. Het is misschien wel fijn voor deze mensen dat ze nooit worden lastig gevallen, maar het is vervelend voor de onderzoeker, omdat hij een deel van zijn doelpopulatie mist.

Het is zeker zo dat steeds meer mensen internet hebben. Volgens cijfers van het CBS is in tien jaar tijd het percentage huishoudens met internet omhoog geschoten van 58% naar 94% (Bron: CBS, Statline). Toch heeft nog niet iedereen internet. Die groep wordt wel steeds kleiner, maar het is niet uitgesloten dat er een groep zonder internet overblijft die behoorlijk afwijkt van de mensen met internet. Zolang dit het geval is, zullen de uitkomsten van online-onderzoek een systematische afwijking kunnen vertonen.

Er zijn verschillen tussen mensen met en zonder internet. Je kunt de internetpopulatie niet zien als een goede afspiegeling van de Nederlandse bevolking. Groepen zoals bijvoorbeeld ouderen, laag opgeleiden en allochtonen zijn behoorlijk ondervertegenwoordigd.

In Nederland hebben heel veel mensen toegang tot internet, maar dat geldt niet voor andere landen. In tabel 5 staan de cijfers voor 29 landen in Europa. Nederland scoort het hoogst in Europa. Ook in de Scandinavische landen is de penetratie van internet hoog. Voor al in Oost-Europa en de Balkan is het internetbezit relatief laag. In Bulgarije en Roemenië heeft zelfs minder dan de helft van huishoudens toegang tot internet.

Tabel 5. Huishoudens met toegang tot internet. Bron: Eurostat (2011).

Land	Internet
Bulgarije	45%
Roemenië	47%
Griekenland	50%
Cyprus	57%
Portugal	58%
Italië	62%
Litouwen	62%
Spanje	64%
Letland	64%
Hongarije	65%
Tsjechië	67%
Polen	67%
Estland	71%
Slowakije	71%
Slovenië	73%
Malta	75%
Oostenrijk	75%
Frankrijk	76%
België	77%
Ierland	78%
Duitsland	83%
Ver. Koninkrijk	83%
Finland	84%
Denemarken	90%
Luxemburg	91%
Zweden	91%
Noorwegen	92%
IJsland	93%
Nederland	94%

In de Verenigde Staten heeft ongeveer 80% van de huishoudens toegang tot internet. Dat is een stuk minder dan in Noord-Europa. Volgens het U.S. Department of Commerce (2011) is er nog steeds sprake van een 'digital divide'. Mensen met minder of geen toegang tot het internet zijn vooral te vinden in bepaalde etnische groepen (Afro-Amerikanen en Latino-Amerikanen), op het platteland en in groepen met een lage opleiding en een laag inkomen. Dit maakt het internet minder geschikt voor representatieve peilingen.

Een interessant verschil tussen Nederland en Amerika is dat in Nederland de opiniepeilingen vrijwel uitsluitend via het internet worden uitgevoerd, terwijl veel Amerikaanse peilers nog gebruik maken van de telefoon. Dat is misschien niet zo verbazingwekkend als je bedenkt dat in Amerika nog maar ongeveer 80% van de mensen toegang heeft tot internet, terwijl dit in Nederland al op 94% ligt.

Peilen via de telefoon wordt in Nederland steeds problematischer. Ongeveer 30% van de huishoudens met een vaste telefoon staat niet in het telefoonboek. En ook veel jonge mensen met alleen maar een mobiele telefoon staan er niet in. Dus het is lastig om een goede, representatieve steekproef te trekken.

Ook in Amerika hebben ze problemen met het peilen via de telefoon. Zie bijvoorbeeld het artikel in de New York Times van Nate Silver op 19 september 2012. Er zijn twee soorten peilingen:

- Robopolls. Hierbij wordt het interview met de respondent helemaal door de computer afgehandeld. De computer belt mensen op en stelt de vragen. Er komen geen 'echte' interviewers aan te pas. Deze geautomatiseerde systemen mogen van de federale overheid geen mobiele nummers bellen.
- Traditionele CATI-interviews. Er zijn echte interviewers van vlees en bloed. In deze peilingen kunnen ook mobiele nummers worden gebeld.

Beide soorten peilingen werden gebruikt tijdens de campagne voor de presidentsverkiezing in de VS. De robopolls leverden substantieel andere uitkomsten op. In de robopolls werd het verschil tussen Obama en Romney geschat op 1,5% (op 3 november 2012). Dit verschil is veel groter (4,1%) als er ook mobieltjes in de peiling zitten.

De verklaring van deze verschillen is dat er in ongeveer een derde van de Amerikaanse huishouden geen vaste telefoon meer is. De leden gebruiken allemaal mobieltjes. Uit onderzoek is bleken dat hier vooral om mensen gaat met lage inkomens, financiële problemen, slechtere gezondheid, geen ziektekosten-verzekering, en meer alcoholproblemen. Dit zijn typisch bevolkingsgroepen waarin Obama veel aanhang heeft.

In Nederland is onderdekking niet het grootste probleem van internetpeilingen. Mocht het toch belangrijk zijn de groep zonder internet mee te nemen in een peiling, dan zijn er nog wel oplossingen te bedenken. Zo biedt het LISS-panel van de Universiteit Tilburg personen zonder internet een simpele PC met een gratis internetverbinding aan voor de duur van het onderzoek. Zie Scherpenzeel (2008) voor meer informatie. Ook kan worden overwogen om de mensen zonder internet op een andere manier te benaderen (per brief, per telefoon, of mondeling). Dit wordt een *mixed-mode peiling* genoemd.

Zelfselectie

De principes van de steekproeftheorie schrijven voor dat een steekproef door loting moet worden geselecteerd uit de doelpopulatie. Alleen dan krijg je valide uitkomsten en alleen dan kun je de onzekerheid in de uitkomsten kwantificeren. Om een steekproef te loten is een steekproefkader nodig.

Hoe loot je een steekproef bij een internetpeiling? Er is immers meestal geen lijst met alle e-mailadressen beschikbaar. Veel internetpeilers omzeilen dit probleem door de respondenten

zichzelf te laten selecteren voor de peiling. Dit wordt *zelfselectie* genoemd. Via banners en popup windows en andere advertenties worden internetgebruikers op de peiling attent gemaakt. De peiler hoopt dan dat deze gebruikers zich daardoor laten overhalen.

Typische voorbeelden van zelfselectie zijn te vinden bij de politieke peilingen in Nederland. De peiler heeft dan een panel opgezet en dat gevuld met mensen die zich daarvoor zelf hebben aangemeld. De mensen in dit panel worden regelmatig benaderd met peilingen. In feite zitten in zulke panels vooral mensen die het leuk vinden om regelmatig aan peilingen mee te doen en geïnteresseerd zijn in politieke zaken. De representativiteit van dit soort panels is op zijn minst twijfelachtig.

Een bijkomend probleem is ook dat mensen zich kunnen aanmelden voor het panel die niet tot de doelpopulatie behoren. Ook is het soms bij peilingen mogelijk om jezelf meer dan één keer aan te melden, hetzij onder dezelfde identiteit hetzij onder een andere identiteit. Uiteraard kan de onderzoeker via een of meer vragen proberen dit soort zaken te ontdekken en te elimineren, maar dat is niet zo eenvoudig. Een handige respondent is wel in staat deze controles te omzeilen.

Ook de American Association for Public Opinion Research (AAPOR) waarschuwt voor de gevaren van zelfselectie (Baker et al., 2010): "*Only when a web-based survey adheres to established principles of scientific data collection can it be characterized as representing the population from which the sample was drawn. But if it uses volunteer respondents, allows respondents to participate in the survey more than once, or excludes portions of the population from participating, it must be characterized as unscientific and is unrepresentative of any population*".

Een ander gevaar van zelfselectie is dat groepen mensen kunnen proberen de uitkomst van de peiling te manipuleren. Een voorbeeld daarvan is de NS publieksprijs. Dat is een literaire prijs die elk jaar wordt toegekend. De winnaar wordt via een internetpeiling bepaald. Voor de prijs van 2005 kon je stemmen op één van de zes genomineerde boeken, maar je kon ook zelf een boek opgeven. In totaal brachten 92.000 mensen hun stem uit. Tot verbazing van iedereen werd niet een van de genomineerde boeken tot winnaar gekozen. Van de stemmers koos 72% voor de Nieuwe Bijbelvertaling. Deze uitslag was het resultaat van een campagne gevoerd door onder anderen het dagblad Trouw, de Evangelische Omroep, het Nederlands Bijbelgenootschap, de Katholieke Bijbelstichting en de Protestantse Kerk om te stemmen op de nieuwe Bijbelvertaling.

Ook bij een peiling voor de Tweede Kamerverkiezingen van 12 september 2012 deed zich een poging tot manipulatie voor. Een groep van 2500 'infiltranten' probeerde zich aan te melden bij het panel van Maurice de Hond (Bronzwaer, 2012). Het idee van deze groep was om zich te presenteren als CDA-stemmer en dan langzamerhand over te gaan naar de partij 50PLUS. De actie werd ontdekt omdat er ineens wel heel veel aanmeldingen tegelijk binnen kwamen. Het laat echter wel zien dat met een wat subtielere aanpak peilingen op basis van zelfselectie te manipuleren zijn.

Vanuit een methodologisch perspectief is het probleem van zelfselectie dat je als onderzoeker de trekkingskansen niet meer in de hand hebt. Je kunt zelfs achteraf niet meer bepalen wat die trekkingskansen zijn geweest. Het is daarom onmogelijk om op basis van aldus verkregen gegevens valide schattingen te maken van kenmerken van de populatie

Voor een valide internetpeiling dient zelfselectie dus te worden vermeden. Hoe moet de steekproef dan worden getrokken? Het ligt voor de hand dit te doen op een andere manier dan via het internet. Een voorbeeld is de aanpak van het CBS. Hierbij wordt eerst een kanssteekproef getrokken uit het bevolkingsregister (GBA). De geselecteerde personen worden eerst benaderd per brief. Die brief bevat het adres van de

website en een unieke login-code. De respons is relatief laag. Daarom worden de non-respondenten opnieuw benaderd, maar dan per telefoon (CATI), als een telefoonnummer beschikbaar is, en anders komt een enquêteur bij hen langs (CAPI).

Bij het LISS-panel, een internetpanel van de Universiteit Tilburg, wordt de schriftelijke stap overgeslagen. Een steekproef uit het bevolkingsregister wordt benaderd via CATI (als een telefoonnummer beschikbaar is) en anders via CAPI. Zie Scherpenzeel (2009).

Voordeel van beide aanpakken is dat zo ook personen in de steekproef terecht kunnen komen die geen internet hebben. Bij het LISS-panel krijgen personen zonder internet een simpele PC aangeboden met een gratis internetverbinding voor de duur van het onderzoek.

De door zelfselectie veroorzaakte problemen kunnen dus worden opgelost, maar daarvoor moet wel een prijs worden betaald: het rekruteren van respondenten via een brief, telefoon of persoonlijk bezoek is kostbaarder en tijdrovender. Daarmee gaan enkele voordelen van een internetpeiling (het is goedkoop en snel) verloren.

Het blijft belangrijk om wetenschappelijk onderzoek te doen naar goedkope en snelle rekruteringstechnieken die representatieve steekproeven voor peilingen opleveren.

Non-respons

In elke peiling treedt non-respons op. Dat is het verschijnsel dat van de in de steekproef getrokken personen de gewenste informatie niet wordt verkregen. De vragenlijst blijft leeg. Een direct gevolg van non-respons is dat minder waarnemingen beschikbaar komen omdat de gerealiseerde steekproef kleiner is. Erger is echter dat non-respons vaak selectief is: sommige groepen zijn oververtegenwoordigd in het onderzoek (omdat ze goed meedoen) en andere groepen zijn ondervertegenwoordigd (omdat ze minder enthousiast zijn). Daardoor wordt de representativiteit van het onderzoek aangetast en kunnen onjuiste conclusies worden getrokken. Een bekend voorbeeld is een kiezersonderzoek. Mensen die gaan stemmen bij verkiezingen doen vaker mee peilingen dan mensen die niet gaan stemmen. Dat leidt ertoe dat stemmers oververtegenwoordigd zijn in peilingen. Een voorbeeld hiervan is te vinden in Bethlehem, Cobben en Schouten (2011, blz. 291).

Er zijn drie belangrijke oorzaken van non-respons:

- *Geen contact.* De uitnodiging om mee te doen bereikt de geselecteerde persoon niet. Dat kan diverse redenen hebben. Voorbeelden zijn een persoon die niet thuis is, een fout adres, of een persoon die is verhuisd.
- *Weigering.* Veel voorkomende redenen om te weigeren zijn dat de geselecteerde persoon geen zin of geen tijd heeft, dat het onderwerp van de peiling de persoon niet aan staat, of dat de persoon vindt dat het onderzoek teveel inbreuk maakt op zijn privacy.
- *Niet in staat.* Het gaat om personen die wel mee willen doen maar het niet kunnen. Dit kan bijvoorbeeld komen door ziekte, dronkenschap of een geestelijke handicap. Ook een taalbarrière kan een reden voor deze vorm van non-respons zijn.

Non-respons komt bij elk type peiling voor, of het nu een schriftelijke, mondelinge, telefonische of een online peiling betreft. Wel is het zo dat responspercentages hoger zijn bij peilingen waarbij enquêteurs worden ingezet. De respons is vaak laag (niet boven de 40%) bij internetpeilingen. Dat verhoogt het risico van grote afwijkingen in de uitkomsten van de peiling.

Een onderzoeker moet altijd nadenken over maatregelen om de respons in een peiling te verhogen. Een mogelijkheid

is, bijvoorbeeld, het geven van een beloning (geldbedrag, VVV-bon, donatie aan een goed doel) aan de geselecteerde personen. De onderzoeker moet daarbij wel in het oog houden dat niet alleen een hoge respons belangrijk is, maar ook de representativiteit ervan. Er zijn voorbeelden waarbij beloningen de respons verhoogden, maar tegelijk tot een minder evenwichtige samenstelling van de respons leidden.

In een poging zo goed mogelijk te corrigeren voor akelige effecten van non-respons, kan een weegprocedure worden uitgevoerd. Daarop wordt dieper ingegaan in paragraaf 9.

Ondertussen blijft het probleem van de non-respons in Nederland onverminderd groot. En in het buitenland nemen de problemen alleen maar toe. Daarom blijft het belangrijk om onderzoek te doen naar de mechanismen die leiden tot non-respons. Meer inzicht daarin biedt ook meer mogelijkheden om daarvoor te corrigeren.

Correctie

Uit het voorgaande is gebleken dat de representativiteit van een internetpeiling op diverse manieren kan worden aangetast: door onderdekking, door zelfselectie en door non-respons. Door de peiling goed op te zetten, is het mogelijk om onderdekking en zelfselectie te vermijden. Maar dan nog blijft non-respons over als een verschijnsel dat heel moeilijk te bestrijden is.

Het is daarom belangrijk om de respons in een peiling te onderzoeken op mogelijke aantasting van de representativiteit. Daarvoor zijn hulpvariabelen nodig. Dat zijn kenmerken van personen die in de peiling zijn geregistreerd en waarvan de verdeling in de hele populatie bekend is. Een voorbeeld is het geslacht. Het is bekend dat de bevolking voor ongeveer 50% uit vrouwen bestaat. Als de respons op een peiling nu slechts 40% vrouwen bevat, dan kan worden geconcludeerd dat er te weinig vrouwen en teveel mannen in de peiling zitten. Als het onderwerp van de peiling dan ook nog samenhangt met geslacht, dan zullen de uitkomsten vertekend zijn.

Als de analyse van de non-respons voldoende aanwijzingen oplevert voor een mogelijke aantasting van de representativiteit, dan is het niet verantwoord is om zonder verdere correcties over te gaan tot publicatie van de uitkomsten. Een veel toegepaste methode om de uitkomsten te corrigeren is het uitvoeren van een *weegprocedure*. Daarbij wordt elke waargenomen persoon een *correctiegewicht* toegekend. In de schattingsprocedures worden vervolgens deze correctiegewichten meegenomen.

In het geval van de variabele geslacht zou dit betekenen dat alle vrouwen een gewicht van 50/40= 1,250 krijgen en alle mannen een gewicht van 50/60=0,833.

Het uitvoeren van een weegprocedure biedt geen garantie op succes. Het werkt alleen als alle relevante weegvariabelen worden gebruikt voor het berekenen van correctiegewichten. Dit betekent dat:

- Weegvariabelen moeten worden meegenomen in de weging die van invloed zijn op het responsgedrag van de geselecteerde personen;
- Weegvariabelen moeten worden meegenomen in de weging die van invloed zijn op de variabelen die worden gemeten in de peiling.

Het is nog maar de vraag of in de praktijk deze variabelen altijd beschikbaar zijn. Voor veel relevante variabelen zal de verdeling in de bevolking helaas niet bekend zijn.

Er zijn allerlei methoden ontwikkeld om een weegprocedure uit te voeren. Voorbeelden zijn post-stratificatie, lineair wegen, multiplicatief wegen en wegen met responskansen. Zie hiervoor Bethlehem en Biffignandi (2012). Meer onderzoek is nodig om te zoeken hoe weegprocedures effectief kunnen

worden ingezet bij internetpeilingen. Er is nog onvoldoende antwoord op vragen als welke weegvariabelen het beste werken en in welke weegtechnieken ze het beste kunnen worden gebruikt.

Meetfouten

Vanuit een cognitief perspectief gezien is het beantwoorden van vragen in een peiling geen simpele taak. Schwarz et al. (2008) beschrijven welke stappen de respondenten daarbij doorlopen: (1) het begrijpen van de vraag, (2) het ophalen uit het geheugen van de benodigde informatie, (3) het vertalen van de informatie in het correcte antwoord, en (4) besluiten of dit antwoord inderdaad wordt gegeven.

Er kan van alles misgaan in dit proces, vooral bij internetpeilingen. Bij mondeling en telefonisch enquêteren zijn interviewers aanwezig. Die kunnen de respondenten bij de les houden en ondersteunen bij het formuleren van het juiste antwoord. Dit levert gegevens van goede kwaliteit op. Dat is heel anders bij internetpeilingen. Dan moet de respondent het helemaal in zijn eentje doen. Krug (2006) beschrijft hoe mensen websites lezen. Dat is uiteraard ook van toepassing op internetpeilingen:

- Respondenten hebben weinig of geen belangstelling voor onderwerp van de peiling.
- Meedoen aan de peiling is niet echt belangrijk voor hen.
- Ze lezen de vragen niet nauwgezet, maar scannen ze slechts globaal.
- Ze weten dat er geen straf staat op het geven van onjuiste antwoorden.
- Ze zoeken niet uit hoe de vragenlijst precies werkt, maar modderen voort van vraag naar vraag en proberen zo het eind te halen.

Dit alles leidt ertoe dat respondenten niet het beste antwoord geven (het antwoord dat zo dicht mogelijk bij de werkelijkheid ligt), maar dat ze met een minimale inspanning het eerste redelijke antwoord kiezen dat ze acceptabel vinden. Dit wordt wel *satisficing* genoemd. De term is een samenvoeging van 'satisfy' en 'suffice'. Satisficing komt voor in allerlei vormen:

- *Primacy effect*: De respondent kiest in een lange lijst van mogelijke antwoorden een antwoord vooraan in de lijst.
- *Acquiescence*: De respondent is geneigd in te stemmen met een stelling in een opinievraag ongeacht de inhoud ervan.
- *Endorsing the status quo*: Als een respondent wordt gevraagd zijn mening te geven over een mogelijk verandering, zal hij geneigd zijn te kiezen voor de optie 'geen verandering'.
- *Preference for the middle option*: Als een respondent naar zijn mening wordt gevraagd, zal hij geneigd zijn te vluchten naar de middelste, neutrale optie die aangeeft dat hij geen duidelijke mening heeft.
- *Straightlining*: Als een reeks vragen met allemaal dezelfde antwoordmogelijkheden wordt aangeboden in de vorm van een matrix, dan is de respondent geneigd alle antwoorden in dezelfde kolom te kiezen. Zie figuur 1 voor een voorbeeld.

Figuur 1. Een matrixvraag met straightlining.

	Uitstekend	Heel goed	Goed	Redelijk	Slecht
1. Wat vindt u in het algemeen van de kwaliteit van de omroep?	○	○	◉	○	○
2. Wat vindt u van de kwaliteit van de nieuwsprogramma's?	○	○	◉	○	○
3. Wat vindt u van de kwaliteit van de sportprogramma's?	○	○	◉	○	○
4. Wat vindt u van de kwaliteit van de muziekprogramma's?	○	○	◉	○	○
5. Wat vindt u van de kwaliteit van de culturele programma's?	○	○	◉	○	○
6. Wat vindt u van de kwaliteit van de kinderprogramma's?	○	○	◉	○	○

- *Don't know*: De respondent is bij een opinievraag geneigd te kiezen voor de optie 'Weet niet' indien deze wordt aangeboden als een van de mogelijke antwoorden.
- *Arbitrary answers*: Vooral bij vragen waarbij alle in aanmerking komende antwoorden moeten worden aangekruist, beperkt de respondent zich tot maar een paar mogelijkheden.

Al deze effecten kunnen leiden tot onjuiste en minder juiste antwoorden, wat een negatief effect heeft op de kwaliteit van de uitkomsten van de peiling.

Hierbij moet wel worden aangetekend dat er ook aspecten zijn waarop een internetpeiling het beter doet dan peilingen met interviewers. Een voorbeeld daarvan is sociaalwenselijke antwoorden. Die worden eerder gegeven op gevoelige vragen als een interviewer de vragen stelt. Bij een internetpeiling is de respondent geneigd om eerlijker te antwoorden.

Problemen met de vragenlijst van een internetpeiling kunnen nog worden vergroot doordat de weergave ervan van respondent tot respondent kan verschillen, als gevolg van gebruik van verschillende browsers, van verschillende instellingen van dezelfde browser, of van verschillende apparaten (computer, laptop, netbook, tablet, smartphone).

Het onderzoek op dit terrein is gefragmenteerd, en vaak gebaseerd op kleine experimenten bij universiteiten met groepjes studenten. Het is niet duidelijk of en hoe groot deze effecten zijn in brede en grootschalige surveys onder de hele bevolking. Hier is duidelijk meer onderzoek nodig.

Een pleidooi voor betere peilingen

We peilen heel wat af in Nederland. Dat is vooral te merken in de periode voor de Tweede Kamerverkiezingen. In die campagnetijd volgen de politieke peilingen elkaar in hoog tempo op. Maar ook buiten de verkiezingen wordt steeds vaker de mening van 'de Nederlander' gevraagd over allerlei onderwerpen.

Vooral internet is de oorzaak van een sterke toename van het aantal peilingen. Internet maakt het mogelijk eenvoudig, snel en goedkoop bij heel veel mensen gegevens te verzamelen. Er zijn websites waarop iedereen zonder enige kennis van onderzoeksmethoden snel een peiling kan opzetten. De vraag is echter of al die peilingen wel goed in elkaar zitten. En als dat niet zo is, dan dienen grote vraagtekens te worden gezet bij de validiteit van de uitkomsten.

Een checklist

Er zijn goede en slechte peilingen. Voor de gebruikers van de uitkomsten van peilingen (journalisten, bestuurders, beleidsmakers, enz.) is het lastig om op eenvoudige wijze het kaf van het koren te scheiden. Daarom is een checklist ontwikkeld. Zie Bethlehem (2012). De checklist bestaat uit negen vragen:

1. Is duidelijk wie de *opdrachtgever* en/of *financier* van het onderzoek is?
2. Is er een *onderzoeksverantwoording* waarin precies staat aangegeven hoe het onderzoek is opgezet en uitgevoerd?
3. Is duidelijk wat de *doelpopulatie* is?
4. Is de volledige *vragenlijst* is opgenomen in de onderzoeksverantwoording? Is deze vragenlijst voor de start van het onderzoek getest?
5. Hoe is de *steekproef* getrokken? Is hij geloot of is er sprake van zelfselectie?
6. Is de omvang van de gerealiseerde steekproef (het *aantal respondenten*) vermeld?
7. Is het *percentage respons* voldoende hoog, zeg hoger dan 50%?
8. Is een *correctie* (weging) uitgevoerd voor de opgetreden non-respons?
9. Worden *onzekerheidsmarges* in de uitkomsten vermeld?

Door de negen vragen in de checklist één voor één af te lopen, kan een eerste indruk worden verkregen de kwaliteit de peiling. Lijkt de kwaliteit goed te zijn, dan verdienen de uitkomsten misschien wel nadere aandacht. Roept het doorlopen van de checklist veel vragen op, dan is het misschien maar beter het onderzoek te negeren.

De checklist is een gezamenlijk initiatief van het Centraal Bureau voor de Statistiek (CBS), het Nederlandstalig Platform voor Survey-onderzoek (NPSO) en de Vereniging voor Onderzoeksjournalisten (VVOJ). Op elke nieuwsredactie zou die checklist goede diensten kunnen bewijzen.

Transparantie

Het is pas mogelijk iets meer te zeggen over de kwaliteit van peilingen als er informatie beschikbaar is over de wijze waarop de peiling is opgezet en uitgevoerd. Daarom is het belangrijk dat een peiling goed wordt gedocumenteerd. Er moet een rapport zijn waarin de hele onderzoeksmethodologie beschreven staat.

Al in 1948 stelde de Verenigde Naties een commissie in die richtlijnen moest maken voor het documenteren van de methodologische aspecten van dit soort onderzoek. In 1964 werden deze richtlijnen verder aangepast (United Nations, 1964).

De documentatie zou in principe voldoende moeten zijn om een peiling in de toekomst op exact dezelfde wijze opnieuw te kunnen uitvoeren. Er moet in ieder geval precies worden aangegeven op welke populatie de uitkomsten betrekking hebben, welk steekproefkader is gebruikt, hoe de steekproef is getrokken, of er non-respons is opgetreden en hoe daarvoor is gecorrigeerd. Verder moet de documentatie een kopie van de vragenlijst bevatten, zodat duidelijk is hoe en in welke volgorde de vragen zijn gesteld.

Het is in Nederland niet goed gesteld met het documenteren van peilingen. De beschikbare informatie is vaak minimaal en meestal onvoldoende voor een goed oordeel. Misschien is het goed eens een blik over de grens te werpen. Febelmar, de Belgische Federatie van Marktonderzoekbureaus, geeft op haar website (www.febelmar.be) richtlijnen en aanbevelingen voor het rapporteren over politieke peilingen. Van alle recente 'kiesintentiemetingen' kan deze documentatie ook op de website worden geraadpleegd. Helaas, ontbreekt een dergelijke website in Nederland.

Kennis van zaken

Nieuwsredacties worden geconfronteerd met een voortdurende stroom van persberichten over de uitkomsten van onderzoek. Zodra die berichten zinnen bevatten als "Uit onderzoek is gebleken dat …", dan moeten de alarmbellen van de redactie gaan rinkelen. Iemand van de redactie moet uitzoeken, bijvoorbeeld aan de hand van de checklist, of het om een goede of slechte peiling gaat. Nog beter zou het zijn als kennis van de meer methodologische aspecten van peilingen bij de redactie of de journalist zelf aanwezig is.

Ook al zit een peiling methodologisch goed in elkaar, is de steekproef netjes geloot uit de hele populatie, en veroorzaakt non-respons geen vertekening, dan nog is er sprake van een onzekerheidsmarge in de uitkomsten. De lezers, luisteraars of kijkers moeten weten dat die onzekerheid er is. Daarom moeten de onzekerheidsmarges worden vermeld. Al is het maar om te voorkomen dat zinloze discussies worden gevoerd over niet-significante verschillen veroorzaakt door de 'ruis' in de steekproef. In de Verenigde Staten is het vrij gebruikelijk om onzekerheidsmarges erbij te zetten. En in de Editorial Guidelines van de BBC staat: "*We should report the expected margin of error in voting intention polls if the gap between the contenders is within the margin. Television and online graphics should always show the margin of error*".

Tijdens de campagne voor de Tweede Kamerverkiezingen van 12 september 2012 hebben peilers hier en daar voor het eerst iets over onzekerheidsmarges in hun peilingen gezegd. Dat is een goede ontwikkeling. Het kan echter nog veel beter, al is het maar om te voorkomen dat in TV-programma's zinloze gesprekken worden gevoerd over de oorzaken van één zetel erbij of eraf.

Als er sprake is van onderdekking, zelfselectie of een substantiële hoeveelheid non-respons, dan kan er ook nog een vertekening optreden in de uitkomsten. Helaas kunnen we die vertekening niet berekenen. Dat betekent dat de onzekerheid in de uitkomsten nog veel groter is, en bovendien weten we niet hoe groot die is. Het enige dat we kunnen zeggen is dat werkelijke onzekerheid groter is dan de berekende onzekerheid.

Dankwoord

Dames en heren,

Aan het slot gekomen van mijn rede, wil ik graag nog een kort dankwoord uitspreken.

Ik wil allen danken die aan de totstandkoming van mijn benoeming hebben bijgedragen. Dat zijn het College van Bestuur, het curatorium van deze bijzondere leerstoel, de directie van het Centraal Bureau voor de Statistiek en de besturen van het Instituut Politieke Wetenschap en het Instituut Bestuurskunde.

Ik wil ook graag mijn nieuwe collega Joop van Holsteijn bedanken. We zijn beiden methodologische adviseur van het EénVandaag Opiniepanel. En we delen een zekere zendingsdrang in onze streven de kwaliteit van de peilingen te verbeteren. Ook hij heeft bijdragen aan mijn benoeming bij de Faculteit Sociale Wetenschappen.

Ik hoop dat mijn benoeming leidt tot meer samenwerking tussen het CBS en de Universiteit Leiden. Enerzijds hoop ik dat die samenwerking bijdraagt aan het oplossen van de methodologische uitdagingen waarvoor het CBS staat. En anderzijds hoop ik ook dat de jarenlange ervaring van het CBS op het gebied van survey-onderzoek zijn weg vindt in de universiteit.

Ik heb gezegd.

Literatuur

Baker, R., Blumberg, S.J., Brick, J.M., Couper, M.P., Courtright, M., Dennis, J.M., Dillman, D., Frankel,, M.R., Garland, P., Groves, R.M., Kennedy, C., Krosnick, J., Lavrakas, P.J., Lee, S., Link, M., Piekarski, L., Rao, K., Thomas, R.K. & Zahs, D. (2010), Research Synthesis: AAPOR Report on Online Panels. *Public Opinion Quarterly* 74, blz. 711-781.

Bethlehem, J.G. (2009), *The Rise of Survey Sampling*. Discussion Paper 09015, Statistics Netherlands, The Hague/Heerlen.

Bethlehem, J.G. (2012), *Peilingen beoordelen - Een checklist*. Centraal Bureau voor de Statistiek, Den Haag/Heerlen.

Bethlehem, J.G. & Biffignandi, S. (2012), *Handbook of Web Surveys*. John Wiley & Sons, Hoboken, NJ.

Bethlehem, J.G., Cobben, F. & Schouten, B. (2011), *Handbook of Nonresponse in Household Surveys*. John Wiley & Sons, Hoboken, NJ, USA.

Bowley, A.L. (1906), Address to the Economic Science and Statistics Section of the British Association for the Advancement of Science. *Journal of the Royal Statistical Society* 69, blz. 548-557.

Bowley, A.L. (1926): Measurement of the Precision Attained in Sampling. *Bulletin of the International Statistical Institute*, XII, Book 1, blz. 6-62.

Bronzwaer, S. (2012), *Infiltranten probeerden de peilingen van Maurice de Hond te manipuleren*. NRC, 13 september 2012.

Den Dulk, C.J. & Van Maarseveen, J.G.S.J. (1990), Volkstellingen 2795-1971. De ontwikkeling van beleid en methode van onderzoek. In: Erwich, B. & Van Maarsseveen, J.G.S.J. (red.), *Een eeuw statistieken*, Centraal Bureau voor de Statistiek, Voorburg/Heerlen, blz. 329-366.

Eurostat (2011), *Internet Use in Households and by Individuals in 2011*. Statistics in Focus 66/2011, Eurostat, Luxembourg.

Horvitz, D.G. & D.J. Thompson (1952). A Generalization of Sampling without Replacement from a Finite Universe. *Journal of the American Statistical. Association* 47, blz. 663-685.

Kiær, A. N. (1895), Observations et Expériences Concernant des Dénombrements Représentatives. *Bulletin of the International Statistical Institute*, IX, Book 2, blz. 176-183.

Kiær, A. N. (1997, herdruk): Den Repräsentative Undersökelsesmetode. *Christiania Videnskabsselskabets Skrifter. II. Historiskfilosofiske klasse*, Nr 4 (1897). Statistisk Sentralbyrå, Oslo, Noorwegen.

Krug, S. (2006), *Don't Make Me Think! A Common Sense Approach to Web Usability, Second Edition*. New Riders, Berkeley, California, USA.

Neyman, J. (1934), On the Two Different Aspects of the Representative Method: the Method of Stratified Sampling and the Method of Purposive Selection. *Journal of the oyal Statistical Society* 97, blz. 558-606.

NIPO (1946a), Wat denkt het publiek ervan? *De Publieke Opinie*, 1e jaargang, No. 1, blz. 1-2.

NIPO (1946b), Rekening en Verantwoording? *De Publieke Opinie*, 1e jaargang, No.2, blz. 1.

Scherpenzeel (2008). An Online Platform for Multi-disciplinary Research. In: Stoop, I. & Wittenberg, M. (red.), *Access Panels and Online Research, Panacea or Pitfall?* Aksant, Amsterdam, blz. 101-106.

Scherpenzeel (2009). *Innovations and New Technologies in Panel Research.* Paper presented at the 5th International Conference on Panel Data Users, Lausanne, Switzerland.

Schwarz, N., Knäuper, B., Oyserman, D. & Stich, C. (2008), The Psychology of Asking Questions. In: De Leeuw, E.D., Hox, J.J. & Dillman, D.A. (red.), *International Handbook of Survey Methodology*. Lawrence Erlbaum Associates, New York / London, blz 18-34.

Silver, N. (2012), *Obama's Lead Looks Stronger in Polls that Include Cellphones.* The New York Times, 19 september 2012.

United Nations (1964), *Recommendations for the Preparation of Sample Survey Reports.* Provisional Report, Sales No. 64.XVII.4, United Nations, New York.

U.S. Department of Commerce (2011), *Exploring the Digital Nation, Computer and Internet Use at Home.* U.S. Department of Commerce, Economics and Statistics Administration and National Telecommunications and Information Administration, Washington DC.

Prof.dr. Jelke Bethlehem

1967-1974	Studie Mathematische Statistiek aan de Universiteit van Amsterdam
1974-1978	Wetenschappelijk medewerker bij de afdeling Mathematische Statistiek van het Mathematisch Centrum in Amsterdam
1978-1987	Onderzoeker bij de hoofdafdeling Statistische Methoden van het Centraal Bureau voor de Statistiek in Voorburg
1986	Promotie bij de Universiteit van Amsterdam (Werken met Non-respons)
1987-1996	Hoofd van de sector Statistische Informatica van het Centraal Bureau voor de Statistiek in Voorburg
1997-heden	Senior-methodoloog bij de sector Methodologie van het Centraal Bureau voor de Statistiek in Voorburg/Den Haag
1991-2011	Parttime hoogleraar in de Statistische Informatieverwerking bij de Faculteit Economie en Bedrijfswetenschappen van de Universiteit van Amsterdam
2012	Benoeming tot bijzonder hoogleraar in de survey-methodologie, in het bijzonder met behulp van het internet, bij de Faculteit Sociale Wetenschappen van de Universiteit Leiden.

Er wordt in Nederland veel gepeild. Dat is vooral te merken in de periode voor de Tweede Kamerverkiezingen. In die campagnetijd volgen de politieke peilingen elkaar in hoog tempo op. Maar ook buiten de verkiezingen om wordt de mening van 'de Nederlander' steeds vaker gevraagd over allerlei onderwerpen.

Vooral het internet is er de oorzaak van dat het aantal peilingen sterk is toegenomen. Internet maakt het mogelijk eenvoudig, snel en goedkoop bij heel veel mensen informatie te verzamelen. De vraag is echter of al die peilingen wel een goed beeld geven van de werkelijkheid.

De representativiteit van internetpeilingen kan op allerlei manieren worden aangetast. Doordat nog niet iedereen in Nederland internet heeft, vallen bepaalde groepen buiten de boot. Bij veel peilingen wordt de steekproef niet netjes geloot, maar is de werving gebaseerd op zelfselectie van respondenten. Verder zijn de percentages non-respons hoog. En de kwaliteit van de antwoorden laat vaak te wensen over, omdat er geen enquêteurs worden ingeschakeld. Daarom vereist het gebruik van een internetpeiling grote zorgvuldigheid.

Universiteit Leiden