



## **Centraal Bureau voor de Statistiek**

Divisie Methodologie en Kwaliteit  
Sector Methodologie

*Postbus 4000  
2270 JM Voorburg*

---

# **Representativiteit van web-surveys - Een illusie?**

**Jelke Bethlehem**

Kennisgeving:

De in dit rapport weergegeven opvattingen zijn die van de auteurs en komen niet noodzakelijk overeen met het beleid van het Centraal Bureau voor de Statistiek.

---

*Projectnummer:*

*205340/09*

*Datum:*

*11 oktober 2006*

## 1. Inleiding

Het uitvoeren van een goed survey-onderzoek is ingewikkeld, kostbaar en tijdrovend. Traditioneel werd daarbij gebruik gemaakt van papieren vragenlijsten. Die konden op drie manieren worden gebruikt bij het verzamelen van gegevens: mondeling enquêteren bij de mensen thuis, telefonisch enquêteren en schriftelijk enquêteren.

Een van de nadelen van het gebruik van papieren vragenlijsten was dat de zo verkregen gegevens veel fouten bevatten. Om informatie van voldoende kwaliteit te krijgen, was het daarom noodzakelijk de gegevens systematisch te controleren en, daar waar nodig, te corrigeren. Deze controle- en correctieactiviteiten consumeerden vaak een substantieel deel van het beschikbare survey-budget, zie bijvoorbeeld Bethlehem (1997).

De ontwikkelingen in de informatietechnologie in de jaren tachtig van de vorige eeuw maakten het mogelijk om de computer te gaan gebruiken voor het verzamelen van gegevens. Zo ontstond *computergestuurd enquêteren*, of in het Engels: *Computer Assisted Interviewing* (CAI). Essentieel hierbij is dat de vragenlijst niet meer op papier staat, maar in een computerprogramma zit. Dit programma bepaalt welke vragen moeten worden gesteld en controleert ook de antwoorden.

Het gebruik van computers tijdens het verzamelen van de gegevens heeft drie belangrijke voordelen. In de eerste plaats wordt tijdens het interview de enquêteur ontlast. Deze hoeft niet steeds de volgende vraag op te zoeken. In de tweede plaats kan de computer tijdens het gesprek een aantal controles uitvoeren. Geconstateerde fouten kunnen onmiddellijk worden hersteld. Dat leidt tot gegevens van betere kwaliteit. In de derde plaats zitten de gegevens na afloop van het vraaggesprek meteen al in de computer. Dat hoeft dus naderhand niet meer te gebeuren, zodat het verwerkingsproces aanzienlijk wordt versneld.

De computer werd het eerst ingezet bij telefonisch enquêteren. Deze vorm van enquêteren wordt algemeen aangeduid met de term CATI (*Computer Assisted Telephone Interviewing*). Nadat de eerste shootcomputers op de markt verschenen, werd het mogelijk de computer in te zetten bij mondeling enquêteren. Zo ontstond *Computer Assisted Personal Interviewing* (CAPI). Nadat steeds meer huishoudens en bedrijven zelf de beschikking kregen over een computer, kon ook het schriftelijk enquêteren worden vervangen door een elektronische variant. Deze wordt vaak aangeduid met *Computer Assisted Self Interviewing* (CASI). Meer informatie over de diverse aspecten van computer gestuurd interviewen kan bijvoorbeeld worden gevonden in Couper et al. (1998).

Met de razendsnelle opkomst van het internet heeft een nieuw type gegevensverzameling zijn intrede gedaan: *Computer Assisted Web Interviewing* (CAWI). Hierbij wordt de vragenlijst aangeboden aan de respondenten via het internet. Dit

soort surveys wordt kortweg aangeduid als *websurveys*. In feite is een websurvey een speciale vorm van CASI.

Op het eerste gezicht heeft een websurvey een aantal aantrekkelijke eigenschappen:

- Internet biedt de mogelijkheid om op betrekkelijk eenvoudige wijze toegang te krijgen tot een zeer grote groep potentiële respondenten;
- De vragenlijsten kunnen tegen zeer lage kosten worden aangeboden. Verder hoeven er geen interviewers te worden ingeschakeld en er hoeven geen portokosten te worden betaald;
- Er gaat weinig tijd verloren tussen maken en aanbieden van de vragenlijst;
- Een websurvey biedt mogelijkheden om zaken als beeld (foto, video) en geluid op eenvoudige wijze in de vragenlijst op te nemen.

In dit rapport wordt ingegaan op een aantal methodologische aspecten van websurveys. In het bijzonder wordt onderzocht wat de effecten zijn van onderdekking en zelfselectie op de representativiteit van de uitkomsten. Tevens wordt nagegaan in hoeverre correctietechnieken in staat zijn om vertekening in de uitkomsten te reduceren.

Paragraaf 2 geeft een systematisch overzicht van mogelijke problemen die zich kunnen voordoen bij het uitvoeren van survey-onderzoek. Daarbij wordt aangegeven welke van die problemen vooral betrekking hebben op websurveys. Paragrafen 3 en 4 gaan in op twee specifieke problemen die vooral de kwaliteit van de uitkomsten van websurveys aantasten: het gebruik van het internet als steekproefkader en zelfselectie van respondenten. In paragraaf 5 wordt een theoretisch kader opgezet. Dat wordt in paragrafen 6 en 7 toegepast om een uitdrukking te vinden voor de vertekening van schattingen op basis van een websurvey. Daarmee wordt duidelijk hoe groot een vertekening kan zijn en waar die van afhangt. In paragrafen 8, 9 en 10 wordt een aantal mogelijkheden onderzocht om voor die vertekening te corrigeren: een weegprocedure, een controle-survey en een steekproef uit de mensen zonder internet. De theorie wordt in paragraaf 11 geïllustreerd aan de hand van een simulatie-experiment. Daaruit wordt duidelijk dat het soms wel en soms niet mogelijk is om een vertekening weg te werken. Het rapport besluit met discussie en conclusies.

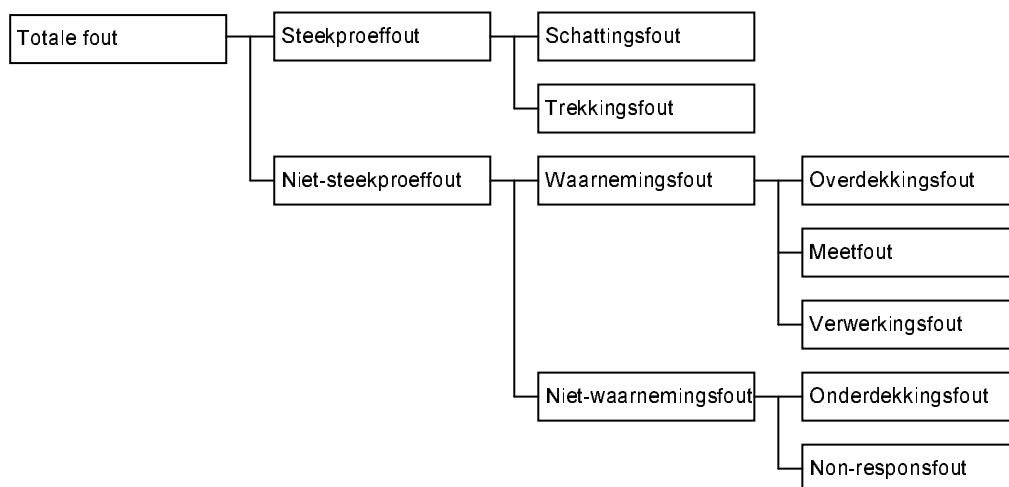
## **2. Fouten in survey-onderzoek**

Bij het uitvoeren van een survey-onderzoek krijgt de onderzoeker met allerlei verschijnselen te maken die de kwaliteit van de uitkomsten kan beïnvloeden. Het optreden van een aantal verstoringen is niet te voorkomen, zodat inspanningen er dan vooral op gericht zullen zijn de invloed ervan beperkt te houden.

Verstoringen in de uitvoering van een survey kunnen leiden tot onjuiste schattingen van allerlei populatiekenmerken. Verschijnselen die tot dit soort verstoringen leiden, worden *foutenbronnen* genoemd en de afwijking die ze met zijn allen

veroorzaken heet de *totale fout*. In Bethlehem (1999) wordt een taxonomie van mogelijke foutenbronnen gegeven. Deze taxonomie is gereproduceerd in figuur 2.1.

*Figuur 2.1. Taxonomie van foutenbronnen*



De totale fout wordt eerst gesplitst in de twee hoofdcategorieën, die worden aangeduid met steekproeffout en niet-steekproeffout. De *steekproeffout* is de fout die ontstaat doordat niet de gehele populatie wordt onderzocht maar slechts een beperkt deel (een steekproef) daarvan. De steekproeffout treedt dus niet op bij een integraal onderzoek.

De *niet-steekproeffout* is de afwijking die ontstaat door verschijnselen die ook zouden zijn opgetreden indien een integraal onderzoek was gehouden (met dezelfde onderzoekstechnieken). De niet-steekproeffout kan vanzelfsprekend ook optreden bij onderzoeken waarbij met steekproeven wordt gewerkt.

De steekproeffout wordt onderverdeeld in twee componenten: de trekkingsfout en de schattingsfout. De *trekkingsfout* is de fout die wordt geïntroduceerd doordat de werkelijke trekkingskansen van de elementen afwijken van de trekkingskansen die worden gebruikt bij de berekening van een schatting. Het is mogelijk dat door het gebruiken van een verkeerde trekkingsprocedure of door de opbouw van het steekproefkader sommige elementen een andere trekkingskans krijgen dan was bedoeld. Een veel voorkomende oorzaak is het dubbel of nog vaker voorkomen van elementen in het steekproefkader.

Wanneer onder gelijk blijvende omstandigheden het (op correcte wijze) trekken van een steekproef een aantal malen zou worden herhaald, dan zullen de schattingen op grond van deze steekproeven niet steeds aan elkaar gelijk zijn. Dit wordt veroorzaakt door het feit dat de steekproeven verschillende elementen kunnen bevatten en de daarbij behorende waarden van de doelvariabelen zullen ook niet steeds dezelfde zijn. De fout die ontstaat doordat toevallig andere elementen in de steekproef worden getrokken, heet de *schattingsfout*. Deze fout wordt gekwantificeerd in de variantie of standaardfout van de schatter.

Een foutenbron die verband houdt met de trekkingsfout en het trekken van steekproeven uit steekproefkaders is het feit dat elementen soms niet met de juiste gegevens in het steekproefkader zijn opgenomen. Men spreekt in dit verband wel van *kaderfouten*. Bij kaderfouten kan worden gedacht aan het feit dat bevolkingsadministraties soms achterlopen en mutaties zoals verhuizingen die nog niet verwerkt zijn. Wordt uit een dergelijke bevolkingsadministratie een personensteekproef getrokken dan kunnen personen in de steekproef terecht komen die niet meer op het door de gemeente opgegeven adres wonen. Kaderfouten kunnen niet worden beschouwd als alleen een onderdeel van de steekproeffout; zij kunnen ook optreden bij integraal onderzoek.

De niet-steekproeffout kan worden gesplitst in een waarnemingsfout en een niet-waarnemingsfout. De *waarnemingsfout* is dat gedeelte van de niet-steekproeffout dat ontstaat door het op incorrecte wijze verkrijgen, vastleggen en verwerken van de gegevens. De waarnemingsfout kan ontstaan *doordat* waarnemingen worden verricht. Een *niet-waarnemingsfout* ontstaat doordat het onmogelijk is om waarnemingen te verrichten.

De waarnemingsfout kan worden gesplitst in drie componenten: overdekking, meetfout en verwerkingsfout. *Overdekking* kan zich voordoen als het steekproefkader elementen bevat die niet tot de doelpopulatie behoren.

De *meetfout* is dat gedeelte van de waarnemingsfout dat ontstaat doordat de gegevens die van de steekproefelementen na waarneming en verwerking beschikbaar komen, niet met de werkelijkheid overeenkomen. Een meetfout kan door een groot aantal verstoringen worden veroorzaakt. Deze verstoringen kunnen verschillende invloeden op de uitkomsten hebben. Een onduidelijke vraagstelling kan leiden tot misverstanden, en daardoor tot verkeerde antwoorden. Het stellen van irrelevante vragen (door het volgen van de verkeerde route) kan irritatie opwekken, met alle gevolgen van dien voor de beantwoording van de vragen. Er kunnen verstoringen optreden als gevolg van interacties tussen respondent, enquêteur en andere aanwezigen bij het vraaggesprek, en door de situatie (omgeving) waarin het vraaggesprek plaatsvindt. Factoren die hierbij een rol spelen zijn huidkleur, geslacht, leeftijd, opleiding en sociale klasse van enquêteur en/of respondent. Waar de onderwerpen gevoelig liggen, bestaat de mogelijkheid dat de respondent alleen sociaal-wenselijke antwoorden geeft.

Problemen kunnen ook ontstaan bij het stellen van vragen die betrekking hebben op het verleden van de respondent. Gebeurtenissen kunnen worden vergeten of op een verkeerd moment in de tijd worden geplaatst. Dit soort verschijnselen, die ook wel *geheugeneffecten* worden genoemd, doen zich voor bijvoorbeeld voor bij vragen over aankopen van goederen, bioscoopbezoek en raadplegen van een arts.

De derde component van de waarnemingsfout is de verwerkingsfout. De *verwerkingsfout* is dat gedeelte van de waarnemingsfout dat ontstaat door verstoringen bij de verwerking van de gegevens. Hierbij kan worden gedacht aan fouten bij het interpreteren, controleren en coderen van de antwoorden, het intypen

van de codes, het programmeren, het uitvoeren van wegingprocedures en analysetechnieken.

Net als de waarnemingsfout, is de niet-waarnemingsfout opgebouwd uit een aantal componenten, namelijk onderdekking en de non-respons. *Onderdekking* doet zich voor als niet alle elementen uit de doelpopulatie terug te vinden zijn in het steekproefkader.

Onder *non-respons* verstaat men het verschijnsel dat van elementen die tot de doelpopulatie behoren en die in de steekproef zijn getrokken niet alle gegevens kunnen worden verkregen.

Veel van bovenstaande problemen kunnen zich ook voordoen bij websurveys. Vooral onderdekking kan een ernstig probleem zijn als het internet wordt gebruikt als steekproefkader. Dan kunnen nooit personen in het onderzoek worden geselecteerd die geen internet hebben.

Trekkingsfouten kunnen zich voordoen wanneer bij de werving voor een websurvey gebruik wordt gemaakt van *zelf-selectie*. Hierbij trekt de onderzoeker de steekproef niet zelf, maar laat het aan de mensen over om te reageren op een oproep om mee te doen met het onderzoek. De trekkingskansen zijn dan onbekend. Als ervan wordt uitgegaan dat er sprake is van een aselechte steekproef, dan zal er een discrepantie zijn tussen de werkelijke en veronderstelde trekkingskansen.

Ook non-respons kan zich voordoen bij websurveys, bijvoorbeeld als eerst een online-panel wordt opgebouwd, en vervolgens daaruit steekproeven worden getrokken. Technische problemen kunnen zelfs een extra oorzaak van non-respons zijn. Trage modems, onbetrouwbare verbindingen, hoge verbindingskosten, verouderde browsers, het niet houden aan HTML-standaards bij het ontwikkelen van vragenlijsten, en onduidelijke navigatie-instructies kunnen allemaal leiden tot het afhaken van respondenten. Zie bijvoorbeeld Couper (2000), Dillman & Bowker (2001), Fricker & Schonlau (2002), en Heerwegh & Loosveldt (2002).

In de volgende paragrafen worden de hiervoor genoemde dekkings- en selectieproblemen verder uitgediept.

### **3. Dekkingsproblemen**

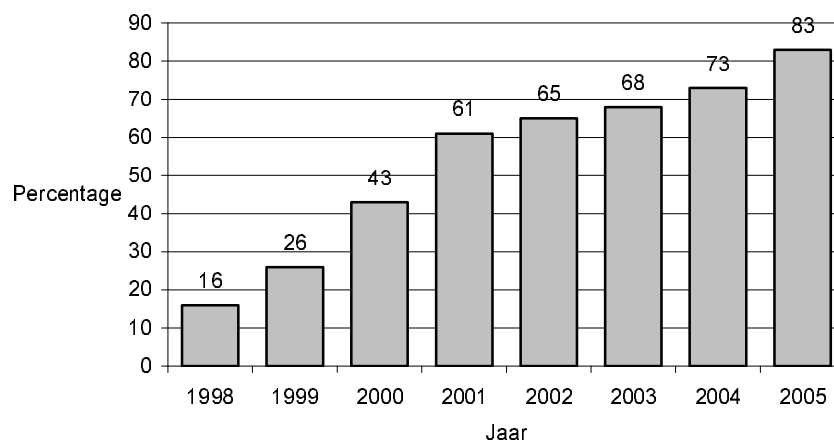
Steekproeven worden getrokken uit een steekproefkader. De verzameling van alle mensen die kunnen worden bereikt via het steekproefkader, wordt de *kaderpopulatie* genoemd. Aangezien de steekproef afkomstig is uit die kaderpopulatie, kunnen de conclusies van het onderzoek alleen maar betrekking hebben op die kaderpopulatie, en niet noodzakelijk op de eigenlijke doelpopulatie. Dekkingsproblemen zullen zich voordoen wanneer de kaderpopulatie afwijkt van de doelpopulatie.

*Onderdekking* doet zich voor wanneer mensen uit de doelpopulatie niet voorkomen in de kaderpopulatie. Deze mensen zullen nooit geselecteerd worden in de steekproef. Hun selectiekans is 0. Bij veel websurveys is het internet in feite het

steekproefkader. Onderdekking zal dan een groot probleem vormen als de doelpopulatie groter is dan de internetpopulatie.

*Figuur 3.1. Percentage personen in Nederland met internetaansluiting*

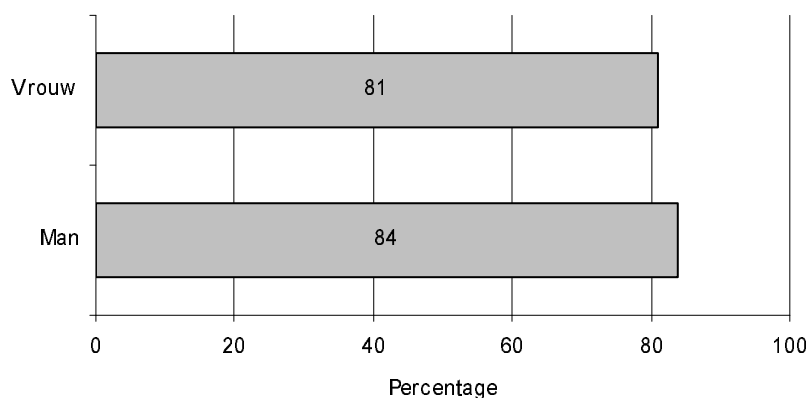
*(Bron: CBS / StatLine)*



Als de doelpopulatie de gehele Nederlandse bevolking is, dan zal er sprake zijn van onderdekking. Immers, niet iedereen heeft een internetaansluiting. In figuur 3.1 is te zien dat het percentage mensen met internet wel stijgt in de loop der jaren. In zeven jaar tijd is het percentage omhoog geschoten van 16 procent naar 83 procent. Toch heeft nog niet iedereen internet. De verwachting is wel dat het probleem in de toekomst steeds kleiner zal worden.

*Figuur 3.2. Personen in Nederland met internet in 2005 naar geslacht*

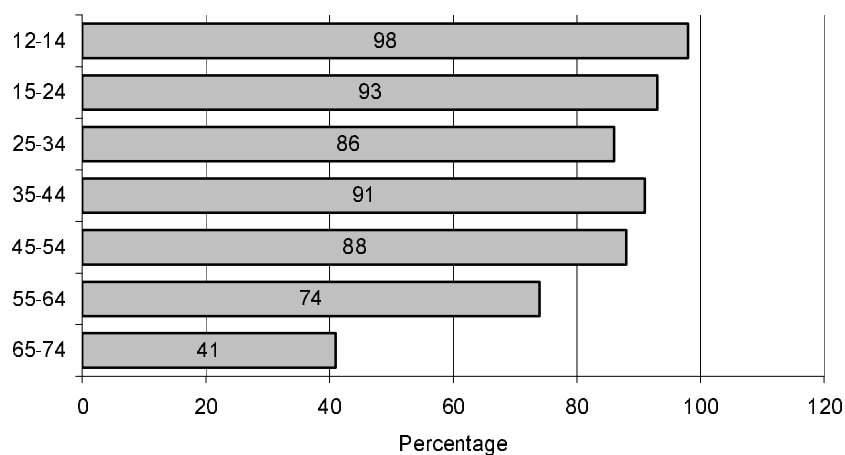
*(Bron: CBS / StatLine)*



Onderzoek naar gegevens over de samenstelling van de internetpopulatie in 2005 toont aan dat die populatie niet kan worden gezien als een afspiegeling van de Nederlandse bevolking.

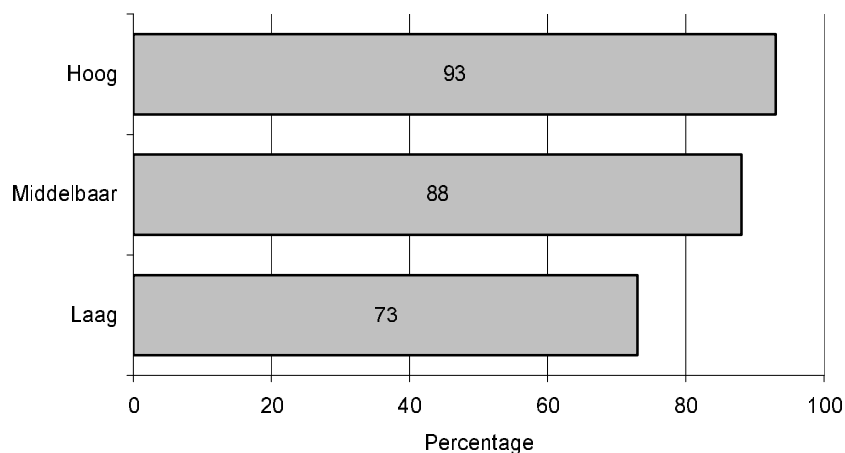
Figuur 3.3 geeft het internetbezit per leeftijdscategorie weer. Duidelijk is te zien dat internetbezit daalt met de leeftijd. Vooral mensen boven de 65 jaar zijn substantieel ondervertegenwoordigd.

*Figuur 3.3. Personen in Nederland met internet in 2005 naar leeftijdscategorie (Bron: CBS / StatLine)*



In figuur 3.4 staat het internetbezit uitgezet tegen het opleidingsniveau. Ook hier valt een duidelijke trend te onderscheiden: het internet bezit neemt toe met het opleidingsniveau.

*Figuur 3.4. Personen in Nederland met internet in 2005 naar opleidingsniveau (Bron: CBS / StatLine)*



Volgens De Haan & Van 't Hof (2006) is internetbezit onder allochtone jongeren veel lager dan onder autochtone jongeren: 91 procent van de jonge autochtonen heeft internettoegang thuis. Voor Surinaamse en Antilliaanse jongeren is het 80 procent, voor jongeren van Turkse afkomst is het 68 procent en voor jongeren van Marokkaanse afkomst slechts 64 procent.



Bovenstaande gegevens komen overeen met de bevindingen van onderzoekers in andere landen. Zie bijvoorbeeld Couper (2000) en Dillman & Bowker (2001).

Het zal duidelijk zijn dat het internet nog geen geschikt steekproefkader is voor het trekken van een steekproef uit de gehele bevolking. Specifieke groepen (ouderen, laag opgeleiden en allochtonen) zijn zwaar ondervertegenwoordigd. Daarom zal zelfs een kanssteekproef uit de internetpopulatie geen goede afspiegeling van de bevolking opleveren.

#### **4. Selectieproblemen**

Horvitz & Thompson (1952) hebben in hun fundamentele artikel laten zien dat zuivere schattingen (m.a.w. schattingen zonder systematische vertekening) alleen kunnen worden berekend als

- 1) de steekproef met een lotingsmechanisme wordt getrokken uit de doelpopulatie van het onderzoek (een kanssteekproef);
- 2) de trekkingskans van elke persoon in de doelpopulatie bekend is;
- 3) al die trekkingskansen positief zijn (dus niet gelijk aan 0 mogen zijn).

Ook is het zo dat alleen onder deze voorwaarden de nauwkeurigheid van de schatting kan worden bepaald.

Veel websurveys zijn niet gebaseerd op kanssteekproeven. Er wordt zelfs helemaal geen steekproef getrokken. De websurvey wordt op het internet gezet en de onderzoeker wacht af wie reageert. Dat zijn dan personen die de beschikking hebben over internet, toevallig de desbetreffende website bezoeken, en besluiten om mee te doen aan het onderzoek. Dit wordt *zelf-selectie* genoemd. Het is duidelijk dat de onderzoeker geen enkele controle heeft over het selectieproces.

Wat is de doelpopulatie van een dergelijk onderzoek? Het zijn in ieder geval mensen met een internetaansluiting, een positieve kans om bij de website terecht te komen en die besluiten mee te doen. Het is echter de vraag of dit ook de doelpopulatie is waarop de onderzoeker zich wil richten bij zijn onderzoek. Een consequentie van deze aanpak is dat de selectiekansen volledig onbekend zijn. Het is dus onmogelijk om zuivere schattingen te berekeningen voor wat dan ook maar de doelpopulatie mag zijn.

De effecten van zelf-selectie kunnen heel duidelijk worden geïllustreerd aan de hand van een voorbeeld dat betrekking heeft op de Tweede Kamerverkiezingen van 22 januari 2003. Diverse organisaties probeerden de zetelverdeling te voorspellen op basis van opiniepeilingen. De uitkomsten van die peilingen zijn samengevat in tabel 4.1.

Bij de peiling op de website *Kennisnet* was duidelijk sprake van zelf-selectie. Elke bezoeker van deze website voor het onderwijs kon zijn stem uitbrengen. Het zullen echter vooral mensen uit het onderwijs zijn geweest die op deze website terecht

kwamen. De verschillen tussen de peiling (op de dag voor de verkiezingen) en het uiteindelijke resultaat (in de kolom ‘Uitslag’) bleken enorm. De Mean Absolute Difference (MAD) geeft aan hoeveel zetels de voorspelling er gemiddeld naast zit. In dit geval is de MAD gelijk aan 12,5. Dus het verschil per partij is gemiddeld groter dan 12 zetels. Merk op dat veel mensen (17 duizend) hebben meegedaan aan deze peiling. Een grote steekproef is dus absoluut geen garantie voor een correcte uitkomst.

*Tabel 4.1. De uitkomsten van enkele opinieonderzoeken*

	<b>Uitslag</b>	<b>Kennisnet</b>	<b>RTL4</b>	<b>SBS6</b>	<b>Nederland 1</b>
Omvang steekproef		17 000	10 000	3 000	1 200
Zetels in de Tweede Kamer					
CDA	44	29	24	42	42
LPF	8	18	12	6	7
VVD	28	24	38	28	28
PvdA	42	13	41	45	43
SP	9	22	10	11	9
GroenLinks	8	26	9	6	8
D66	6	4	7	5	6
Andere partijen	5	14	9	7	7
Mean Absolute Difference		12,5	5,3	1,8	0,8

Een ander voorbeeld van een websurvey met zelf-selectie was de peiling van de TV-zender RTL4. Ook hier konden toevallige bezoeker van de website hun stem uitbrengen. De uitkomsten van RTL4 zijn wel beter dan die van Kennisnet, maar nog steeds niet om over naar huis te schrijven. De waarde van de MAD ligt nog steeds boven de 5. Dat de uitkomsten wat beter zijn, zou te maken kunnen hebben met de samenstelling van de groep bezoekers van website van RTL4. Dat zou wel eens een minder specifieke groep kunnen zijn.

Bij de TV-zender SBS6 werd gebruik gemaakt van een *access panel*. Zulke panels worden opgebouwd door het breed werven van mensen op goed bezochte websites en internetportals. Bij de werving is dus ook sprake van zelf-selectie. Voor allerlei surveys worden dan vervolgens steekproeven getrokken uit dit panel.

Het voordeel van een access panel is dat bij de werving al allerlei vragen kunnen worden gesteld over sociaal-demografische kenmerken van de respondenten. In het panel van SBS6 werd ook gevraagd naar het stemgedrag bij de voorgaande Tweede Kamerverkiezingen. Die informatie kan bij latere surveys worden gebruikt om de representativiteit te controleren (met betrekking tot deze kenmerken) en eventueel een correctie uit te voeren voor een gebrek aan representativiteit.

De waarde van de MAD voor de peiling van SBS6 was gelijk aan 1,8. En dat is al weer een stuk kleiner dan de waarden voor de peilingen van Kennisnet en RTL4, ondanks de veel kleinere steekproefomvang van 3 duizend personen. De mogelijkheid tot correctie op basis van de demografische kenmerken en het kiesgedrag zal daar zeker toe hebben bijgedragen. Toch is er nog steeds een gemiddeld verschil per partij van bijna twee zetels.

Het beste resultaat in tabel 4.1 wordt bereikt voor de peiling van de TV-zender Nederland 1. Hierbij is gebruik gemaakt van een echte kanssteekproef. Telefoonnummers werden op een volstrekt willekeurige manier gegenereerd door een computeralgoritme ('random digit dialling'). Met de bijbehorende personen werd contact gezocht. Met een steekproef van slechts 1200 mensen werd zo een voorspelling verkregen met een gemiddelde afwijking van minder dan één zetel.

De uitkomsten in tabel 4.1 laten nog eens zien dat het voor het verkrijgen van betrouwbare uitkomsten niet gaat om de omvang van de steekproef maar om de representativiteit ervan. Ook is duidelijk dat het trekken van een kanssteekproef een vitaal onderdeel vormt van een goede onderzoeksopzet.

Zelf-selectie draagt ook nog een extra gevaar in zich. Specifieke groepen in de doelpopulatie zouden deze mogelijkheid kunnen aangrijpen om ervoor te zorgen dat ze zijn oververtegenwoordigd in het onderzoek. Daarmee ontstaat de mogelijkheid om de uitkomsten van het onderzoek te manipuleren. Een concreet voorbeeld daarvan is de NS-Publieksprijs 2005. Via de website kon men zijn stem uitbrengen op één van de zes genomineerde boeken. Winnaar werd uiteindelijk een niet-genomineerd boek: de Nieuwe Bijbelvertaling. Bijna 72 procent van de ruim 92 duizend stemmers kozen voor dit boek. Er was 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. Zie ook het weekblad Elsevier van 3 november 2005.

## 5. Het theoretische kader

Laat de doelpopulatie  $U$  van het onderzoek bestaan uit  $N$  elementen (mensen, huishoudens), die worden genummerd van 1 t/m  $N$ . Bij elk element  $k$  hoort een waarde  $Y_k$  van de doelvariabele  $Y$ . Verondersteld wordt dat het doel van het onderzoek is het schatten van het populatiegemiddelde

$$\bar{Y} = \frac{1}{N} \sum_{k=1}^N Y_k \quad (5.1)$$

van de doelvariabele  $Y$ .

De doelpopulatie  $U$  is opgedeeld in twee deelpopulaties: de deelpopulatie  $U_I$  van elementen met toegang tot het internet en de deelpopulatie  $U_{NI}$  van elementen zonder toegang tot het internet. Aan elke element  $k$  in de doelpopulatie is een indicator  $I_k$  gekoppeld, waarbij  $I_k = 1$  als element  $k$  internettoegang heeft (en dus tot deelpopulatie  $U_I$  hoort) en waarbij  $I_k = 0$  als dat niet het geval is. De deelpopulatie  $U_I$  wordt hier de *internetpopulatie* genoemd.

De omvang van de internetpopulatie wordt aangegeven met

$$N_I = \sum_{k=1}^N I_k \quad (5.2)$$

Evenzo geeft  $N_{NI}$  de omvang van deelpopulatie  $U_{NI}$ , aan waarbij  $N_I + N_{NI} = N$ . Het gemiddelde van de doelpopulatie in de internetpopulatie is gelijk aan

$$\bar{Y}_I = \frac{1}{N_I} \sum_{k=1}^N I_k Y_k \quad (5.3)$$

## 6. Een aselechte steekproef uit de internetpopulatie

Eerst wordt nu de situatie bekeken waarin het mogelijk is om een aselechte steekproef zonder teruglegging (en met gelijke kansen) te trekken uit de internetpopulatie. Daarvoor is een steekproefkader nodig waarin alle elementen met internet staan vermeld. Een dergelijke lijst bestaat (nog) niet. Er zijn echter manieren om deze situatie te benaderen. Een aanpak zou kunnen zijn om eerst een steekproef te trekken uit een groter steekproefkader (bijvoorbeeld een bevolkingsregister of een adressenlijst), de geregistreerde personen op een klassieke wijze te benaderen (per brief of telefoon) en zo de personen te selecteren die internettoegang hebben. De geselecteerde personen krijgen vervolgens het internetadres toegestuurd van de vragenlijst. Het is duidelijk dat er bij een dergelijk steekproefkader sprake is van overdekking. Er zitten in het steekproefkader personen die niet tot de doelpopulatie behoren. Het is echter wel zo dat iedereen in de internetpopulatie een bekende en positieve kans op selectie heeft, zodat het mogelijk is om zuivere schatters te berekenen en de onzekerheidsmarge in de uitkomsten te kwantificeren.

Een aselechte steekproef zonder teruglegging uit de internetpopulatie kan worden voorgesteld door een reeks

$$a_1, a_2, \dots, a_N \quad (6.1)$$

van  $N$  indicatoren, waarbij  $k$ -de indicator  $a_k$  de waarde 1 aanneemt als persoon  $k$  wordt geselecteerd, en de waarde 0 is als die persoon niet wordt geselecteerd (voor  $k = 1, 2, \dots, N$ ). Uiteraard geldt dat  $a_k = 0$  voor elke persoon  $k$  zonder internet. De omvang van deze steekproef wordt genoteerd met  $n_I$ .

De verwachte waarde  $\pi_k = E(a_k)$  wordt de *eerste orde insluitkans* van element  $k$  genoemd. Horvitz & Thompson (1952) hebben aangetoond dat het altijd mogelijk is om een zuivere schatter te construeren, mits alle eerste orde insluitkansen bekend en positief zijn. De *Horvitz-Thompson-schatter* voor het gemiddelde van de internetpopulatie is gedefinieerd als

$$\bar{y}_{HT} = \frac{1}{N_I} \sum_{k=1}^N a_k I_k \frac{Y_k}{\pi_k}, \quad (6.2)$$

waarbij per definitie  $Y_k / \pi_k = 0$  als  $I_k = 0$ . In het geval van een aselechte steekproef zonder teruglegging met gelijke kansen zijn alle eerste orde insluitkansen gelijk aan  $n / N$ . Formule (6.2) reduceert dan tot

$$\bar{y}_I = \frac{I}{n_I} \sum_{k=1}^N a_k I_k Y_k \quad (6.3)$$

Dit is een zuivere schatter voor het gemiddelde  $\bar{Y}_I$  van de internetpopulatie, maar niet noodzakelijk voor het gemiddelde  $\bar{Y}$  van de doelpopulatie. De vertekening is gelijk aan

$$B(\bar{y}_{HT}) = E(\bar{y}_{HT}) - \bar{Y} = \bar{Y}_I - \bar{Y} = \frac{N_{NI}}{N} (\bar{Y}_I - \bar{Y}_{NI}) \quad (6.4)$$

De omvang van de vertekening wordt bepaald door twee factoren. De eerste factor is de relatieve omvang van de deelpopulatie zonder internet. Naarmate deze omvang groter is, zal de vertekening ook groter zijn. De tweede factor is het *contrast*  $\bar{Y}_I - \bar{Y}_{NI}$  tussen de internetpopulatie en de niet-internetpopulatie. Naarmate de gemiddelden van de doelvariabele meer verschillen voor de twee deelpopulaties, zal de vertekening groter zijn.

Op het ogenblik kan de omvang van de deelpopulatie van mensen zonder internet nog niet worden genegeerd. Figuur 3.1 laat zien dat het percentage mensen zonder internet afneemt, maar nog steeds in de buurt van de 20 procent ligt.

Verder is in paragraaf 3 beschreven dat er nog steeds substantiële verschillen zijn tussen de twee deelpopulaties. Vooral ouderen, laag opgeleiden en allochtonen zijn slecht vertegenwoordigd in de internetpopulatie. Daarom zullen schatters voor kenmerken van de doelpopulatie op basis van een steekproef uit de internetpopulatie niet zuiver zijn.

## 7. Zelf-selectie uit de internetpopulatie

Bij veel websurveys is er geen sprake van het trekken van een steekproef uit de internetpopulatie. Bij deze surveys laat de onderzoekers het aan de personen zelf over om te participeren in het onderzoek. Dit wordt *zelf-selectie* genoemd. Deelname wordt in de eerste plaats bepaald door de bekendheid van personen met het onderzoek. Ze moeten toevallig terecht komen op de website van het onderzoek, of ze moeten ingaan op een uitnodiging daartoe (bijvoorbeeld via een banner op een andere website). In de tweede plaats moet een dergelijke persoon ook besluiten om inderdaad aan dat onderzoek deel te nemen.

Zelf-selectie betekent dat elk persoon  $k$  in de internetpopulatie een onbekende kans  $\rho_k$  heeft om te participeren in een onderzoek. De responderende personen kunnen worden aangegeven met een reeks

$$r_1, r_2, \dots, r_N \quad (7.1)$$

van  $N$  indicatoren. De  $k$ -de indicator  $r_k$  neemt de waarde 1 aan als persoon  $k$  aan het onderzoek meedoet, en anders krijgt de indicator de waarde 0 (voor  $k = 1, 2, \dots, N$ ). De verwachte waarde  $\rho_k = E(r_k)$  zal de *responskans* van persoon  $k$  worden genoemd. Ook voor personen zonder internet kan het begrip responskans worden geïntroduceerd. Uiteraard zijn de responskansen voor al deze personen per definitie gelijk aan 0.

De omvang van de steekproef die uiteindelijk via zelf-selectie tot stand komt, is gelijk aan

$$n_S = \sum_{k=1}^N r_k \quad (7.2)$$

Een naïeve onderzoeker zou nu kunnen veronderstellen dat elk element in de internetpopulatie dezelfde kans heeft om in de steekproef terecht te komen. In dit geval zou het steekproefgemiddelde

$$\bar{y}_S = \frac{1}{n_S} \sum_{k=1}^N r_k Y_k \quad (7.3)$$

een goede schatter moeten zijn voor het gemiddelde van de internetpopulatie. De verwachte waarde van deze schatter is (bij benadering) gelijk aan

$$E(\bar{y}_S) \approx \bar{Y}_I^* = \frac{1}{N_I \bar{\rho}} \sum_{k=1}^N \rho_k I_k Y_k. \quad (7.4)$$

Hierin is  $\bar{\rho}$  het gemiddelde van alle responskansen in the internetpopulatie, zie bijvoorbeeld Bethlehem (1988).

In het algemeen is de verwachte waarde van het steekproefgemiddelde niet gelijk aan het gemiddelde van de internetpopulatie. De enige situatie waarin de vertekening verdwijnt, is die waarin alle responskansen gelijk zijn. Dat is ook niet verwonderlijk omdat dan elke persoon in de populatie dezelfde selectiekans heeft.

Bethlehem (1988) laat zien dat de vertekening van het steekproefgemiddelde in (7.3) kan worden geschreven als

$$B(\bar{y}_S) = E(\bar{y}_S) - \bar{Y}_I = \bar{Y}_I^* - \bar{Y}_I = \frac{C(\rho, Y)}{\bar{\rho}}, \quad (7.5)$$

waarin

$$C(\rho, Y) = \frac{1}{N_I} \sum_{k=1}^N I_k (\rho_k - \bar{\rho})(Y_k - \bar{Y}) \quad (7.6)$$

de covariantie in de internetpopulatie is tussen de waarden van de doelvariabele en de responskansen. De omvang van deze vertekening wordt bepaald door twee factoren:

- De hoogte van de gemiddelde responskans. Naarmate personen meer bereid zijn om deel te nemen aan het onderzoek, worden hun responskansen groter, en zal de vertekening afnemen.

- De samenhang tussen de waarden van het kenmerk dat wordt onderzocht en de waarden van de responskansen. Naarmate de samenhang sterker is, zal de vertekening groter zijn.

Nadere beschouwing van formule (7.5) levert drie situaties op waarin de vertekening verdwijnt:

- 1) Alle responskansen zijn gelijk. Dan is waarde van de covariantie gelijk aan 0. Dit is weer de situatie waarin het selectieproces kan worden vergeleken met een aselechte steekproef;
- 2) Alle waarden van het te onderzoeken kenmerk zijn gelijk. Dit is een hoogst onwaarschijnlijke situatie. Als dit zich zou voordoen, is één waarneming genoeg. Een peiling zou dan verder overbodig zijn;
- 3) Er is geen samenhang tussen het te onderzoeken kenmerk en de responskansen. Deelname van personen aan het onderzoek hangt dan dus niet af van de waarde van het te onderzoeken kenmerk voor die personen

In de meeste gevallen zal het onderzoek zich niet concentreren op de internetpopulatie maar op de gehele populatie. Als het gaat om het schatten van het gemiddelde in de totale populatie, dan is de vertekening van het steekproef-gemiddelde gelijk aan

$$\begin{aligned} B(\bar{y}_S) &= E(\bar{y}_S) - \bar{Y} = E(\bar{y}_S) - \bar{Y}_I + \bar{Y}_I - \bar{Y} = \\ &= \frac{N_{NI}}{N} (\bar{Y}_I - \bar{Y}_{NI}) + \frac{C(\rho, Y)}{\bar{\rho}} \end{aligned} \quad (7.7)$$

De vertekening is nu opgebouwd uit twee termen: een bijdrage veroorzaakt door het feit dat niet de gehele populatie wordt onderzocht maar slechts de internetpopulatie (onderdekking) en een bijdrage veroorzaakt door zelf-selectie.

Theoretisch is het niet uitgesloten dat deze twee termen elkaar compenseren waardoor de totale vertekening klein is. Als personen zonder internet echter lijken op personen met internet die niet graag aan onderzoek meedoen, dan zullen de twee bijdragen elkaar juist versterken. Praktische ervaringen lijken te suggereren dat dit laatste eerder het geval zal zijn. Als  $Y$  bijvoorbeeld een variabele is die de intensiteit meet van zekere activiteit op het internet (surfen, het spelen van onlinespellen), dan is een positieve correlatie tussen deze variabele en de responskansen niet onwaarschijnlijk. Verder zal het gemiddelde van  $Y$  voor de internetpopulatie een positieve waarde hebben, terwijl die waarde voor personen zonder internet 0 zal zijn. In beide gevallen is de vertekeningsterm dan dus positief. De vertekeningen verdwijnen niet, maar versterken elkaar.

## 8. Correctie met een weegprocedure

Weegtechnieken kunnen worden toegepast om schattingen voor allerlei populatiekenmerken te verbeteren. Omdat te kunnen doen, is *hulpinformatie* nodig.

Hierbij wordt hulpinformatie gedefinieerd als een reeks van variabelen die in het onderzoek zijn gemeten en waarvan de verdeling in de populatie bekend is. Door de verdeling van een hulpvariabele in de steekproef te vergelijken met de verdeling van diezelfde variabele in de populatie, kan worden vastgesteld of de steekproef *representatief* is met betrekking tot die variabele. Als aanzienlijke verschillen tussen beide verdelingen worden geconstateerd, dan zal de conclusie moeten luiden dat de steekproef *selectief* is.

Om voor die selectiviteit te corrigeren, worden *gewichten* uitgerekend. Ondervertegenwoordigde groepen krijgen een groot gewicht, en oververtegenwoordigde groepen een klein gewicht. Vervolgens kunnen schattingen voor populatiekenmerken worden verkregen door de gewogen waarden te gebruiken in plaats van de ongewogen waarden.

Wegen is een techniek die veel wordt toegepast bij het corrigeren van onderzoeken die te leiden hebben van non-respons. Zie Bethlehem (2002) voor een overzicht.

De eenvoudigste en waarschijnlijk meest toegepaste weegtechniek is *post-stratificatie*. Hier wordt deze correctietechniek uitgelegd voor het geval van één hulpvariabele. Uitbreiding naar meer hulpvariabelen is niet essentieel anders.

Stel er is een (kwalitatieve) hulpvariabele  $X$  beschikbaar met  $L$  categorieën. Deze variabele verdeelt de populatie dus in  $L$  *strata* (deelpopulaties). De strata worden aangegeven met  $U_1, U_2, \dots, U_L$ . Het aantal personen in stratum/deelpopulatie  $U_h$  wordt genoteerd met  $N_h$  (voor  $h = 1, 2, \dots, L$ ). De omvang  $N$  van de doelpopulatie is dus gelijk aan  $N = N_1 + N_2 + \dots + N_L$ . Er wordt bij post-stratificatie verondersteld dat deze informatie beschikbaar is. Is  $X$  bijvoorbeeld de variabele geslacht, dan is  $L$  gelijk aan 2 en  $N_1$  en  $N_2$  zijn het aantal mannen en vrouwen in de doelpopulatie.

Veronderstel nu dat een aselechte steekproef van omvang  $n$  is getrokken uit de internetpopulatie. Als  $n_h$  het aantal geselecteerde elementen is in stratum  $h$ , dan geldt  $n = n_1 + n_2 + \dots + n_L$ . Merk op dat aangezien de steekproef wordt getrokken uit de internetpopulatie, alleen gegevens worden verkregen van personen in de sub-strata  $U_l \cap U_h$  (voor  $h = 1, 2, \dots, L$ ).

Bij post-stratificatie krijgen alle participerende personen in hetzelfde stratum hetzelfde gewicht. Het gewicht  $w_k$  voor een persoon  $k$  in stratum  $h$  is gelijk aan

$$w_k = \frac{N_h / N}{n_h / n}. \quad (8.1)$$

Het eenvoudige (ongewogen) steekproefgemiddelde

$$\bar{y}_l = \frac{1}{n} \sum_{k=1}^N a_k I_k Y_k \quad (8.2)$$

wordt nu vervangen door het gewogen steekproefgemiddelde

$$\bar{y}_{PS} = \frac{1}{n} \sum_{k=1}^N a_k w_k I_k Y_k. \quad (8.3)$$



Als hierin formule (8.1) voor de gewichten wordt gesubstitueerd, dan leidt de schatter (8.3) tot de definitie van de *post-stratificatie-schatter*:

$$\bar{y}_{PS} = \frac{1}{N} \sum_{h=1}^L N_h \bar{y}_I^{(h)} = \sum_{h=1}^L W_h \bar{y}_I^{(h)}. \quad (8.4)$$

Hierin is  $\bar{y}_I^{(h)}$  het steekproefgemiddelde in stratum  $h$  en  $W_h = N_h / N$  is de relatieve omvang van stratum  $h$ . De verwachte waarde van deze post-stratificatie-schatter is gelijk aan

$$E(\bar{y}_{PS}) = \frac{1}{N} \sum_{h=1}^L N_h E(\bar{y}_I^{(h)}) = \sum_{h=1}^L W_h \bar{Y}_I^{(h)} = \tilde{Y}_I. \quad (8.5)$$

Hierin is  $\bar{Y}_I^{(h)}$  het gemiddelde van de doelvariabele in stratum  $h$  van de internetpopulatie. In het algemeen zal deze verwachting niet gelijk zijn aan het gemiddelde  $\bar{Y}^{(h)}$  van de doelvariabele in stratum  $h$  van de doelpopulatie. De vertekening van deze schatter is gelijk aan

$$\begin{aligned} B(\bar{y}_{PS}) &= E(\bar{y}_{PS}) - \bar{Y} = \tilde{Y}_I - \bar{Y} = \sum_{h=1}^L W_h (\bar{Y}_I^{(h)} - \bar{Y}^{(h)}) = \\ &= \sum_{h=1}^L W_h \frac{N_{NI,h}}{N_h} (\bar{Y}_I^{(h)} - \bar{Y}_{NI}^{(h)}), \end{aligned} \quad (8.6)$$

waarin  $N_{I,h}$  de omvang van stratum  $h$  in de internetpopulatie.

De vertekening zal klein zijn als binnen de strata de gemiddelden van de personen met en zonder internet elkaar niet veel ontlopen. Deze situatie doet zich voor als er een sterke samenhang tussen de doelvariabele  $Y$  en de hulpvariabele  $X$  die wordt gebruikt voor de stratificatie. De variatie in waarden van  $Y$  zal zich dan vooral manifesteren tussen de strata en niet binnen de strata.

Samenvattend kan worden gesteld dat post-stratificatie tot een reductie van de vertekening kan leiden. Daarvoor moeten hulpvariabelen worden gebruikt met de volgende eigenschappen:

- Ze moeten worden gemeten in het onderzoek;
- Hun populatie verdeling ( $N_1, N_2, \dots, N_L$ ) moet bekend zijn;
- Ze moeten sterk samenhangen met de doelvariabele van het onderzoek.

Helaas liggen geschikte hulpvariabelen niet altijd voor het oprapen. Als er al hulpvariabelen beschikbaar zijn, dan is hun samenhang met de doelvariabele meestal niet zo sterk.

De variantie van de post-stratificatie-schatter is gelijk aan

$$V(\bar{y}_{PS}) = \sum_{h=1}^L W_h^2 V(\bar{y}_I^{(h)}). \quad (8.7)$$

Cochran (1977) toont aan dat in het geval van een aselechte steekproef met gelijke kansen uit de hele doelpopulatie deze variantie kan worden uitgewerkt tot

$$V(\bar{y}_{PS}) = \frac{1-f}{n} \sum_{h=1}^L W_h S_h^2 + \frac{1}{n^2} \sum_{h=1}^L (1-W_h) S_h^2, \quad (8.8)$$

waarin  $f = n / N$  en  $S_h^2$  de variantie in stratum  $h$ . Als de strata homogeen zijn met betrekking tot  $Y$  (d.w.z. de doelvariabele varieert maar weinig binnen de strata), dan is de variantie klein.

In de situatie van een aselechte steekproef uit de internetpopulatie is de variantie van de schatter gelijk aan

$$V(\bar{y}_{PS}) = \sum_{h=1}^L W_h^2 \left( \frac{1}{nW_{1,h}} + \frac{1-W_{1,h}}{(nW_{1,h})^2} - \frac{1}{N_{1,h}} \right) S_{1,h}^2, \quad (8.9)$$

waarin  $N_{1,h}$  de omvang van stratum  $h$  in de internetpopulatie,  $W_{1,h} = N_{1,h} / N_1$  en  $S_{1,h}^2$  de variantie in stratum  $h$  van de internetpopulatie.

## 9. Correctie met een controle-survey

In paragraaf 8 is aangegeven dat wegen een effectieve correctietechniek kan zijn, maar alleen als er geschikte hulpvariabelen beschikbaar zijn. Als dat niet het geval is, dan kan worden overwogen om geschikte hulpvariabelen aan te maken met behulp van een controle-survey.

Een *controle-survey* is een onderzoek waarvoor een kleine steekproef via een nette aselechte steekproef is getrokken uit de totale doelpopulatie. Elke persoon in die doelpopulatie moet dezelfde kans op selectie hebben. Er mag geen non-respons optreden in de survey, of de non-respons moet zodanig zijn dat deze niet leidt tot vertekening in de uitkomsten (zogenaamde ‘ignorable nonresponse’).

Een controle-survey kan dus geen websurvey zijn. Dan zouden immers personen zonder internet nooit getrokken kunnen worden in de steekproef. Voor een controle-survey kan worden gedacht aan een CAPI- of CATI-survey.

Als er geen nonresponse optreedt in de controle-survey (of als de non-response ‘ignorable’ is), zal de deze survey zuivere schattingen opleveren van allerlei populatiekenmerken. Het is dus ook mogelijk om bijvoorbeeld het populatie-gemiddelde van de doelvariabele zuivere schatten. Daarbij moet echter wel worden aangetekend dat deze schatting vanwege de kleine omvang van de steekproef een grote variantie zal hebben. De vraag is nu of het mogelijk is om de schattingen te verbeteren door de grote steekproefomvang van de websurvey te combineren met de zuiverheid van de controle-survey.

Om dit nader te onderzoeken wordt verondersteld dat er één kwalitatieve hulpvariabele beschikbaar is, dat deze hulpvariabele in zowel de websurvey als de controle-survey is gemeten, en dat deze hulpvariabele sterk samenhangt met de doelvariabele. Dan kan een vorm van post-stratificatie worden toegepast waarbij stratumgemiddelden worden berekend met gegevens uit de websurvey en voor de

bepaling van de gewichten worden de gegevens uit de controle-survey gebruikt. Dit leidt dan tot de aangepaste post-stratificatie-schatter

$$\bar{y}_{CS} = \sum_{h=1}^L \frac{m_h}{m} \bar{y}^{(h)}. \quad (9.1)$$

Hierin is  $\bar{y}^{(h)}$  het steekproefgemiddelde in stratum  $h$  (voor de websurvey) en  $m_h / m$  is de relatieve steekproefomvang in stratum  $h$  van de controle-survey (voor  $h = 1, 2, \dots, L$ ). De controle-survey is een nette, aselechte steekproef. Daarom is de  $m_h / m$  een zuivere schatter voor  $W_h = N_h / N$ .

Bethlehem (2006) laat zien dat de verwachting van schatter (9.1) gelijk is aan

$$E(\bar{y}_{CS}) = \tilde{Y}_I = \sum_{h=1}^L W_h \bar{Y}_I^{(h)} \quad (9.2)$$

Vergelijking van formule (9.2) met formule (8.5) wijst uit dat de verwachte waarde van schatter (9.1) gelijk is aan die van de gewone post-stratificatie-schatter. Ook de vertekening van schatter (9.1) is dus gelijk aan de vertekening van de gewone post-stratificatie-schatter:

$$\begin{aligned} B(\bar{y}_{CS}) &= E(\bar{y}_{CS}) - \bar{Y} = \tilde{Y}_I - \bar{Y} = \sum_{h=1}^L W_h (\bar{Y}_I^{(h)} - \bar{Y}^{(h)}) = \\ &= \sum_{h=1}^L W_h \frac{N_{NI,h}}{N_h} (\bar{Y}_I^{(h)} - \bar{Y}_{NI}^{(h)}). \end{aligned} \quad (9.3)$$

De vertekening zal klein zijn als binnen de strata de gemiddelden van de personen met en zonder internet elkaar niet veel ontlopen. Deze situatie doet zich voor als er een sterke samenhang tussen de doelvariabele  $Y$  en de hulpvariabele  $X$  die wordt gebruikt voor de stratificatie.

Zo op het eerste gezicht lijkt de controle-survey dus niets te verbeteren ten opzichte van gewone post-stratificatie. Er is echter wel degelijk een voordeel, en dat is dat de onderzoeker helemaal zelf zijn hulpvariabelen kan uitkiezen. Elke hulpvariabele die in beide surveys wordt gemeten, kan worden gebruikt voor wegen. Als hij een hulpvariabele kiest die leidt tot veel variatie tussen de strata en weinig variatie binnen de strata, dan zal een aanmerkelijke reductie in de vertekening het geval zijn.

Bethlehem (2006) toont aan dat de variantie van schatter (9.1) gelijk is aan

$$V(\bar{y}_{CS}) = \frac{1}{m} \sum_{h=1}^L W_h (\bar{Y}_I^{(h)} - \tilde{Y}_I)^2 + \frac{1}{m} \sum_{h=1}^L W_h (1 - W_h) V(\bar{y}^{(h)}) + \sum_{h=1}^L W_h^2 V(\bar{y}^{(h)}) \quad (9.4)$$

De eerste term in de variantie is van de orde  $1/m$  (waarin  $m$  de steekproefomvang van de controle-survey). De derde term in de variantie is van de orde  $1/n$  (waarin  $n$  de steekproefomvang van de websurvey). De middelste term is van de orde  $1/mn$ . Aangezien  $n$  in de praktijk meestal veel groter is dan  $m$ , zal de eerste term in (9.4) de variantie domineren. In feite wordt de variantie bepaald door de omvang van de controle-survey. Concluderend kan worden vastgesteld dat het met een controle-

survey mogelijk is de vertekening te reduceren, maar wel ten koste van een veel grotere variantie.

## 10. Onderzoek van de niet-internetpopulatie

Het wezenlijke probleem van websurveys is dat er geen waarnemingen beschikbaar komen van personen die geen internet hebben. Dit probleem zou kunnen worden aangepakt door het trekken van een gestratificeerde steekproef. De doelpopulatie bestaat uit twee strata: de internetpopulatie  $U_I$  van omvang  $N_I$  en de niet-internetpopulatie  $U_{NI}$  van omvang  $N_{NI}$ .

Uit dat internetpopulatie wordt met een websurvey een grote steekproef van omvang  $n$  getrokken. Als dit een aselechte steekproef met gelijke kansen is, dan is het steekproefgemiddelde

$$\bar{y}_I = \frac{1}{n} \sum_{k=1}^N a_k I_k Y_k \quad (10.1)$$

een zuivere schatter voor het gemiddelde in de internetpopulatie.

Veronderstel nu eens dat het mogelijk is om een aselechte steekproef met gelijke kansen van omvang  $m$  uit de niet-internetpopulatie te trekken. Helaas is er geen steekproefkader voor de personen zonder internet. Dit zou kunnen worden opgelost door het trekken van een steekproef uit de gehele doelpopulatie en alleen personen zonder internet te selecteren. Eventueel zouden geselecteerde personen met internet nog bij de websurvey kunnen worden gevoegd. Het onderzoek onder mensen zonder internet kan uiteraard niet met een websurvey worden uitgevoerd. Alternatieven zijn een CAPI- of CATI-survey.

Het gemiddelde van de steekproef uit de niet-internetpopulatie kan worden genoteerd met

$$\bar{y}_{NI} = \frac{1}{m} \sum_{k=1}^N b_k (1 - I_k) Y_k . \quad (10.2)$$

Hierin geeft de indicator  $b_k$  aan of persoon  $k$  wel of niet is getrokken in de steekproef. Verder geldt dat

$$m = \sum_{k=1}^N b_k (1 - I_k) . \quad (10.3)$$

De *stratificatie-schatter* is nu gedefinieerd als

$$\bar{y}_{ST} = \frac{N_I}{N} \bar{y}_I + \frac{N_{NI}}{N} \bar{y}_{NI} . \quad (10.4)$$

Dit is een zuivere schatter voor het gemiddelde in de gehele doelpopulatie. Hierbij wordt wel verondersteld dat de omvang  $N_I$  van de internetpopulatie en de omvang

$N_{NI}$  van de niet-internetpopulatie bekend zijn, of nauwkeurig kunnen worden geschat. De variantie van de schatter is gelijk aan

$$V(\bar{y}_{ST}) = \left(\frac{N_I}{N}\right)^2 V(\bar{y}_I) + \left(\frac{N_{NI}}{N}\right)^2 V(\bar{y}_{NI}). \quad (10.5)$$

De variantie bestaat uit twee bijdragen. De eerste bijdrage komt van de variantie van het steekproefgemiddelde in de internetpopulatie. Die variantie is van de orde  $1/n$ . Het gaat hier om de websurvey. De steekproefomvang daarvan is meestal groot. Daarom is de bijdrage aan de variantie klein.

De tweede bijdrage aan de variantie in formule (10.5) is afkomstig van de variantie van het steekproefgemiddelde in de niet-internetpopulatie. Deze variantie is van de orde  $1/m$ . Om de kosten van dit onderdeel van de gegevensverzameling binnen de perken te houden, zal de omvang van de steekproef niet te groot mogen zijn. In de praktijk zal het er dus op neerkomen dat  $n$  veel groter is dan  $m$ . Dat betekent dat de tweede bijdrage in de variantie (10.5) zal overheersen.

Net als bij een controlesurvey, zal stratificatie leiden tot het verdwijnen van de vertekening. Er dient echter wel een prijs voor te worden betaald. In de eerste plaats is extra onderzoek onder personen zonder internet kostbaar. In de tweede plaats gaat het voordeel van de grote websurvey verloren. De nauwkeurigheid wordt bepaald door de omvang van de kleine steekproef onder de niet-internetters.

## 11. Een simulatie-experiment

Om te onderzoeken wat de mogelijk effecten kunnen zijn van diverse correctietechnieken, is een klein simulatie-experiment uitgevoerd. Er is een fictieve doelpopulatie geconstrueerd. Voor die populatie zijn een beperkt aantal variabelen gegenereerd. Dat is zo gebeurd dat allerlei samenhangen bestaan zoals die zich ook in werkelijkheid kunnen voordoen. Door een groot aantal maal een steekproef te trekken uit deze populatie, kan de verdeling van allerlei schatters worden geanalyseerd.

Eerst wordt de verdeling van de schatter bepaald voor de ideale situatie van een aselechte steekproef uit de gehele doelpopulatie. Vervolgens wordt bekeken wat er gebeurt als de aselechte steekproef alleen uit de internetpopulatie wordt getrokken (een websurvey). Tenslotte wordt onderzocht wat het effect is op de schatter van diverse correctietechnieken (wegen, controle-survey).

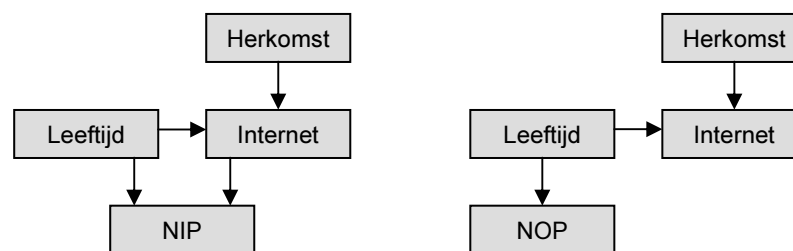
De omvang van de fictieve populatie was 30 duizend personen. Er zijn vijf variabelen gegenereerd:

- Leeftijd in drie categorieën: Jong (met kans 0,40), Middelbaar (met kans 0,35) en Oud (met kans 0,35);
- Herkomst in twee categorieën: Autochtoon (met kans 0,85) en Allochtoon (met kans 0,15).

- Internettoegang in twee categorieën Ja en Nee. De kans op internet hangt af van de twee variabelen Leeftijd en Herkomst. Voor autochtonen zijn deze kansen 0,90 (voor Jong), 0,70 (voor Middelbaar) en 0,50 (voor Oud). Internettoegang neemt dus af met de leeftijd. Voor allochtonen zijn deze kansen 0,20 (voor Jong), 0,10 (voor Middelbaar) en 0,00 (voor Oud). Kans op internet is voor allochtonen dus duidelijk lager dan voor autochtonen.
- Stemt bij de verkiezingen op de Nationale Ouderen Partij (NOP). De kans om op deze partij te stemmen hangt alleen af van de leeftijd. De kansen zijn 0,00 (voor Jong), 0,40 (voor Middelbaar) en 0,60 (voor Oud). De kans om op de NOP te stemmen neemt dus toe met de leeftijd.
- Stemt bij de verkiezingen op de Nieuwe Internet Partij (NIP). De kans om op deze partij te stemmen hangt af van zowel de leeftijd als van het wel of niet hebben van internet. Voor de internetbezitters zijn de kansen zijn 0,80 (voor Jong), 0,40 (voor Middelbaar) en 0,20 (voor Oud). Voor mensen zonder internet zijn alle kansen gelijk aan 0,10. De kans om op de NIP te stemmen neemt voor internetbezitters dus af met de leeftijd en is erg laag voor mensen zonder internet.

In figuur 11.1 zijn de samenhangen tussen de verschillende variabelen nog eens grafisch voorgesteld.

*Figuur 11.1. Samenhang tussen de variabelen*



De situatie waarbij personen zonder internet ontbreken in een onderzoek vertoont enige gelijkenis met het verschijnsel non-respons. Ook daar ontbreken gegevens, maar dat is omdat geselecteerde personen weigeren mee te doen, ze niet thuis zijn, of niet in staat zijn om mee te doen. In de theorie over non-respons worden drie mechanismen onderscheiden die kunnen leiden tot het ontbreken van waarnemingen:

- *Missing Completely At Random* (MCAR). Hierbij is er geen samenhang tussen het wel of niet ontbreken en de onderzoeksvariabelen. Dit is de ideale situatie waarin het mechanisme alleen maar tot minder waarnemingen leidt en niet tot een vertekening.
- *Missing At Random* (MAR). Hierbij is er samenhang tussen het wel of niet ontbreken en de doelvariabelen, maar deze samenhang loopt via een hulpvariabele die kan worden gebruikt voor correctie. In termen van post-stratificatie komt MAR erop neer dat zonder stratificatie de schatter een vertekening heeft, maar dat die vertekening zich niet voordoet binnen elke

categorie van de hulpvariabele apart. De vertekening ontstaat doordat de verdeling over de strata in de steekproef niet overeenkomt met de verdeling in de doelpopulatie. Door weging kan dit probleem worden opgelost.

- *Not Missing At Random* (NMAR). Hierbij is er sprake van een directe samenhang tussen het wel of niet ontbreken en de doelvariabele van het onderzoek. Dit is de slechtst denkbare situatie. De schatter heeft een vertekening die niet kan worden weggewerkt door weging.

In het hierboven beschreven simulatie-experiment is er bij de variabele NOP (Stemt op de Nationale Ouderenpartij) sprake van MAR. Er is een directe samenhang tussen NOP en leeftijd en er is samenhang tussen leeftijd en het hebben van een internetaansluiting. Hier zou correctie via weging met leeftijdverbetering moeten brengen.

Bij de variabele NIP is er sprake van NMAR. Er is (onder anderen) een directe samenhang tussen NIP en het hebben van internet. Weging zou dan niet effectief moeten zijn.

In het simulatie-experiment is in diverse situaties de verdeling van de schatter bepaald voor het percentage stemmen op de NOP en de NIP. De omvang van de steekproef was steeds  $n = 2000$ . Om een beeld te krijgen van de verdeling van de schatter, werd in elke situatie het trekken van de steekproef steeds 800 keer herhaald.

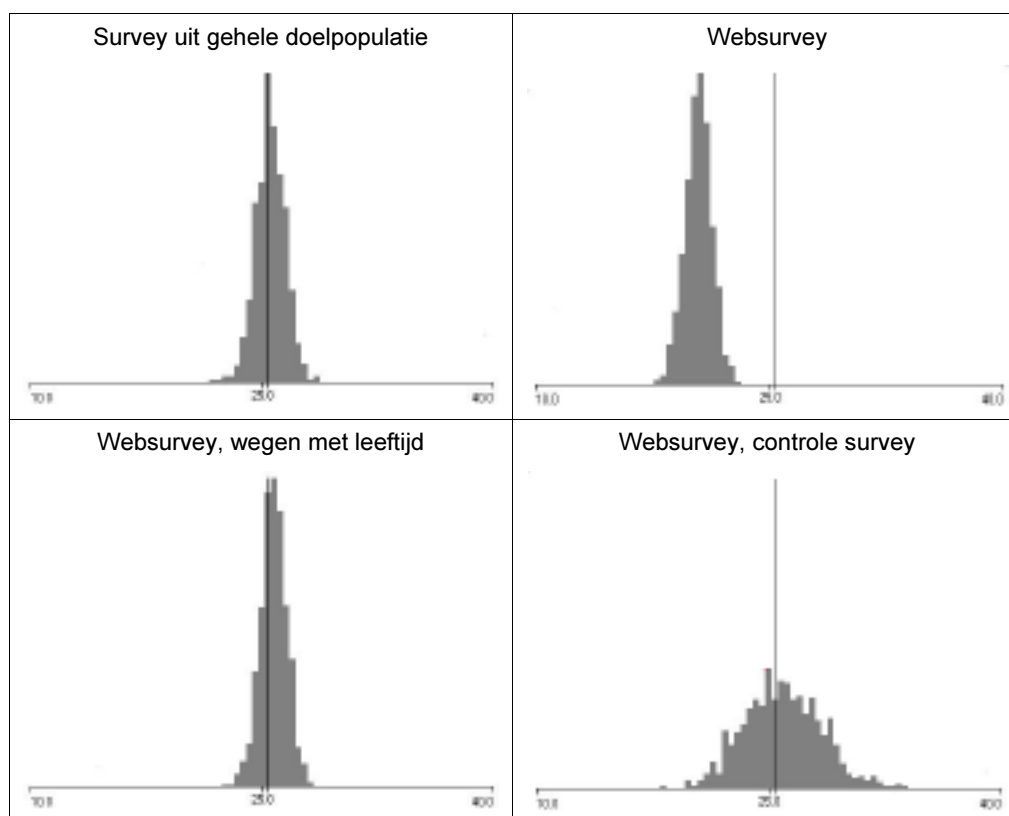
Figuur 11.2 bevat de uitkomsten voor de variabele NOP (stemt op Nationale Ouderen Partij). De grafiek linksboven toont de verdeling van de schatter bij aselechte trekking van een steekproef uit de gehele doelpopulatie. De verdeling ligt symmetrisch verdeeld om de populatiewaarde 25,4 procent (aangegeven door de verticale streep). De schatter is dus zuiver.

De grafiek rechtsboven laat zien wat er gebeurt als een aselechte steekproef alleen uit de internetpopulatie wordt getrokken. De verdeling is in zijn geheel naar links geschoven. De schattingen vallen systematisch te laag uit. Het percentage stemmers op de NOP in de internetpopulatie is slechts 20,3 procent. Er is dus sprake van een vertekening. Dat is verklaarbaar: relatief weinig ouderen hebben internet, waardoor ze dus ondervertegenwoordigd zijn in de steekproef. Dat zijn nu net de mensen die met grote kans op de NOP zullen stemmen.

De grafiek linksonder toont het effect van weging (post-stratificatie) met de variabele leeftijd. Omdat bij de websurvey sprake is van Missing At Random (MAR), is de correctie succesvol: door het weging verdwijnt de vertekening.

Post-stratificatie kan in deze situatie alleen worden toegepast als de verdeling van de variabele leeftijd in de populatie bekend is. Is dat niet het geval, dan kan worden overwogen een controle-survey uit te voeren. De grafiek rechtsonder laat zien wat voor effect dat op de verdeling van de schatter heeft. De vertekening wordt ook hier weggewerkt, maar dat gaat wel ten koste van een substantiële vergroting van de variantie van de schatter.

Figuur 11.2. Uitkomsten van de simulatie voor de variabele NOP



Figuur 11.3 bevat de uitkomsten voor de variabele NIP (Stemt op Nieuwe Internet Partij). De grafiek linksboven toont de verdeling van de schatter bij aselechte trekking van een steekproef uit de gehele doelpopulatie. De verdeling ligt symmetrisch verdeel om de populatiewaarde 39,5 procent. De schatter is dus zuiver.

De grafiek rechtsboven laat zien wat er gebeurt als een aselechte steekproef alleen uit de internetpopulatie wordt getrokken. De verdeling is in zijn geheel flink naar rechts geschoven. De schattingen vallen systematisch te hoog uit. Het percentage stemmers op de NIP in de internetpopulatie is 56,5 procent. Er is dus sprake van een vertekening. Dat is verklaarbaar. Immers, in de internetpopulatie zitten juist de mensen die met een grote kans op de NIP zullen stemmen.

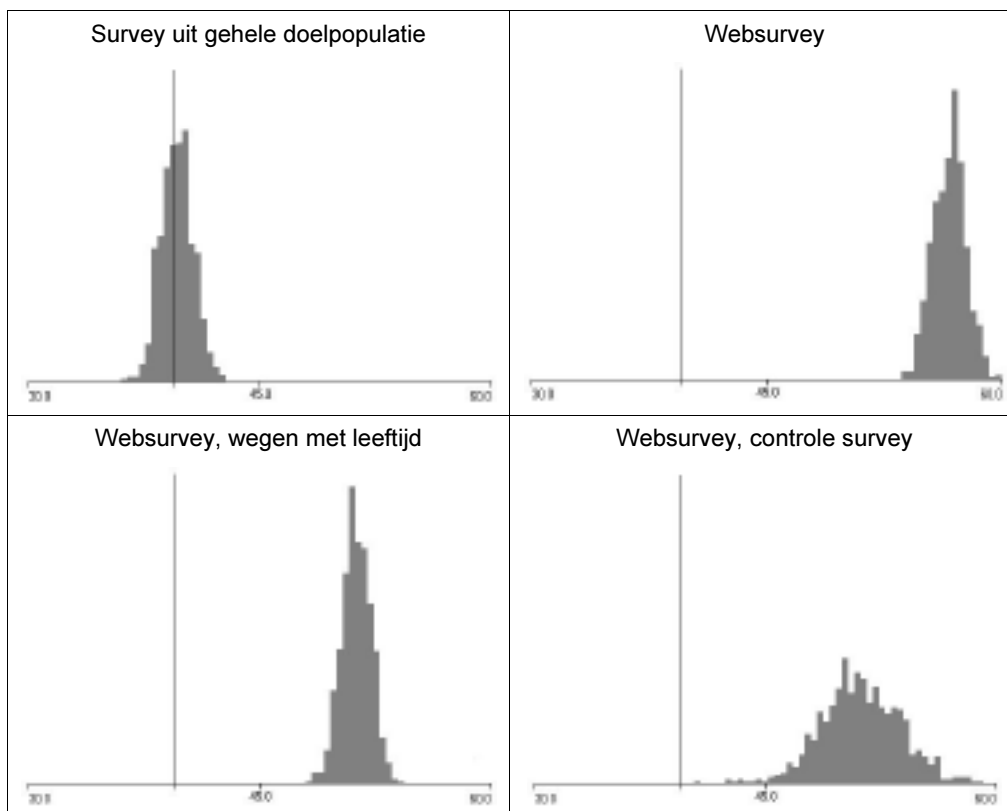
De grafiek linksonder toont het effect van wegen (post-stratificatie) met de variabele leeftijd. Omdat bij deze websurvey sprake is van Not Missing At Random (NMAR), is de correctie niet succesvol. Door het wegen verdwijnt maar een klein deel van de vertekening, namelijk dat deel dat veroorzaakt wordt door de variabele leeftijd. De vertekening veroorzaakt door de directe samenhang tussen stemgedrag en internettoegang verdwijnt niet.

Ook hier kan worden overwogen een controle-survey uit te voeren als de populatieverdeling van de hulpvariabele niet beschikbaar is. De grafiek rechtsonder in figuur 11.3 laat zien wat voor effect dat op de verdeling van de schatter heeft als leeftijd als hulpvariabele wordt gebruikt. De vertekening wordt ook hier maar voor



een klein deel weggewerkt. Bovendien neemt de variantie van de schatter aanzienlijk toe.

*Figuur 11.3. Uitkomsten van de simulatie voor de variabele NIP*



Uit bovenstaande simulatie-experimenten kunnen enkele belangrijke conclusies worden getrokken:

- In geval van Missing At Random (MAR) of Not Missing At Random (NMAR) hebben schatters voor een websurvey een vertekening;
- In het algemeen is er geen garantie dat wegen de vertekening doet verdwijnen. Alleen in geval van Missing At Random (MAR) kan de vertekening worden weggewerkt met een geschikt gekozen weegprocedure;
- Ook bij gebruik van een controle-survey kan de vertekening alleen worden weggewerkt indien er sprake is van Missing At Random (MAR);
- Gebruik van een controle-survey leidt altijd tot een aanzienlijke toename van de variantie van de schatter.

## 12. Discussie en conclusies

In dit rapport is ingegaan op enkele methodologische problemen bij het gebruik van websurveys. Lage kosten en snelheid lijken het belangrijkste argument te zijn om te kiezen voor een websurvey. De fundamentele vraag is echter of een websurvey kan

worden gebruikt om nauwkeurige, betrouwbare uitspraken te doen over een doel-populatie.

Voor het trekken van een steekproef is een steekproefkader nodig. Het internet is niet een ideaal steekproefkader, want er is sprake van onderdekking. Specifieke groepen (ouderen, laag opgeleiden, allochtonen) zijn slecht vertegenwoordigd op het internet. Daardoor hebben schattingen voor populatiekenmerken vaak een systematische vertekening. Er kan worden getracht te corrigeren voor een vertekening door toepassing van een weegtechniek. Deze correctie is alleen succesvol als het niet hebben van internettoegang kan worden gezien als een vorm van Missing At Random (MAR).

Niet onvermeld mag blijven dat ook traditionele methoden van gegevensverzameling dekkingsproblemen hebben. Zo vereist een CATI-survey een steekproefkader van telefoonnummers. Het CBS kan alleen vaste telefoonnummers gebruiken. Echter slechts 60 procent tot 70 procent van die nummers is beschikbaar. De overige nummers zijn geheim of staan niet in het telefoonboek, zie Cobben (2004). Dit probleem zal in de toekomst alleen maar groter worden, aangezien het aantal vaste telefoons afneemt, steeds meer mensen een mobiele telefoon hebben, en er geen steekproefkader van mobiele nummers is. Zie Kuusela (2003) voor een overzicht van de problematiek.

Het percentage personen met internettoegang ligt nu boven de 80 procent in Nederland. Het ligt in de lijn der verwachting dat dit percentage in de toekomst nog verder zal toenemen. Het valt daarom te verwachten dat de dekkingsproblemen voor websurveys in de toekomst minder belangrijk zullen worden. Daar staat tegenover dat de groep personen zonder internet wel eens steeds meer kan gaan afwijken van de groep personen met internet. Dit zou kunnen leiden tot een groter contrast, waardoor de vertekening van de schatters niet kleiner wordt.

Zuivere schatters voor populatiekenmerken zijn alleen maar te berekenen als iedere persoon in de doelpopulatie een bekende en positieve trekkingskans heeft. Dat is heel vaak niet het geval. Zo blijkt uit het Nederlandse Online Panelvergelijkingsonderzoek (NOPVO) 2006, dat voor de web-panels van commerciële marktonderzoekbureaus heel veel gebruik wordt gemaakt van zelf-selectie, zie van Ossenbruggen et al. (2006). De onderzoekers concluderen dat de representativiteit van de panels veel te wensen overlaat.

Kan een websurvey een CAPI- of CATI-survey volledig vervangen? De dekkingsproblemen zullen waarschijnlijk op termijn verminderen en zelf-selectie kan worden ondervangen door het trekken van aselechte steekproef. Toch is er dan toch nog een belangrijk verschil tussen CAPI en CATI enerzijds en CAWI anderzijds: bij CATI- en CATI-surveys worden enquêteurs ingezet en bij websurveys niet. De vraag is of dit invloed heeft op de kwaliteit van de antwoorden en de hoogte van de respons. De Leeuw & Collins (1997) geven aan dat responspercentages vaak hoger zijn bij de inzet van enquêteurs. Gevoelige vragen worden echter beter beantwoord zonder enquêteurs. Er is echter nog maar weinig bekend

over de kwaliteit van de uitkomsten van websurveys. Over dit aspect is nader onderzoek gewenst.

CAPI en CATI zijn beide een vorm van computergestuurd interviewen. Deze vorm van interviewen heeft het voordeel dat controles kunnen worden ingebouwd in het computerprogramma dat de vragen stelt. Dat betekent dat tijdens het interview fouten in de antwoorden kunnen worden ontdekt, en dus ook verbeterd. Daarmee kan de kwaliteit van de verzamelde gegevens worden verbeterd, zie bijvoorbeeld Couper (1998). De vraag komt nu op of deze vorm van controle/correctie ook kan worden ingebouwd in websurveys. Wat zal er gebeuren als respondenten thuis op hun computer wordt geconfronteerd met een foutmelding bij het invullen van de vragenlijst? Zullen ze hun antwoorden corrigeren, of raken ze zo geïrriteerd dat ze het invullen van de vragenlijst halverwege staken? Wellicht is hier sprake van een uitruil tussen de hoogte van de respons en de kwaliteit van de gegevens. Onderzoek zal moeten uitwijzen wat de beste aanpak is.

In dit rapport wordt een controle-survey beschreven als een mogelijk instrument voor het corrigeren van de vertekening in websurveys. Een voordeel van een controle-survey is dat de onderzoeker volledig vrij is in de keuze van de hulpvariabelen, zolang ze maar in beide surveys (websurvey en controle-survey) worden gemeten. Als hulpvariabelen worden gebruikt die sterk samenhangen met de doelvariabele, dan zal dit de vertekening aanzienlijk kunnen reduceren. Een nadeel van de controle-survey is wel dat de variantie van de schattingen aanzienlijk wordt vergroot. Er kan dus veel minder nauwkeurig worden geschat. Toepassing van een controle-survey betekent dat de vertekening wordt ingeruild tegen een grotere variantie. De vraag zou zelfs op kunnen komen of het nog de moeite loont om een websurvey uit te voeren als de nauwkeurigheid van de uitkomsten in feite wordt bepaald door de omvang van de relatief kleine en dure websurvey. De winst ten opzichte van alleen een controle-survey doen is niet zo groot.

Een controle-survey kan noodzakelijkerwijs niet worden uitgevoerd als een websurvey. Het ligt voor de hand er een CAPI- of CATI-survey van te maken. Dit leidt tot een aanzienlijke verhoging van de kosten. Bovendien kunnen er zogenaamde *mode-effecten* optreden. Dat betekent dat een zelfde vraag bij een andere wijze van gegevensverzameling anders wordt beantwoord.

Een noodzakelijke voorwaarde voor het gebruik van een controle-survey is dat deze kan worden uitgevoerd via een echte aselechte steekproef en zonder het optreden van non-respons, of de non-respons moet van het type MCAR zijn (Missing Completely At Random). Het is nog maar de vraag of dit in de praktijk haalbaar is.

Het zal duidelijk zijn dat een websurvey niet zonder meer een geschikt instrument is voor gegevensverzameling. Er zitten teveel haken en ogen aan de technieken die nodig zijn om voor gebrek aan representativiteit te corrigeren.

### 13. Referenties

- Bethlehem, J.G. (1988), Reduction of the nonresponse bias through regression estimation. *Journal of Official Statistics* 4, blz. 251-260.
- Bethlehem (1997), Integrated Control Systems for Survey Processing. In: Lyberg, L., Biemer, P., Collins, M., De Leeuw, E., Dippo, C., Schwarz, N. & Trewin, D. (eds.), *Survey Measurement and Process Control*. Wiley, New York, blz. 371-392.
- Bethlehem, J.G. (1999), Cross-sectional Research. In: Adèr, H.J. & Mellenbergh, G.J., *Research Methodology in the Social, Behavioural & Life Science*. Sage Publications, London, blz.110-142.
- Bethlehem, J.G. (2002), Weighting Nonresponse Adjustments Based on Auxiliary Information. In: Groves, R.M., Dillman, D.A., Eltinge, J.L. and Little, R.J.A. (Eds), *Survey Nonrespons*. Wiley, New York.
- Bethlehem, J.G. (2006), *Reducing the bias of web survey based estimates*. Discussion Paper, to be published. Statistics Netherlands, Voorburg/Heerlen, The Netherlands.
- Cobben, F.(2004), *Nonresponse correction techniques in household surveys at Statistics Netherlands: a CAPI-CATI comparison*. Technical report, Statistical Netherlands, Methods and Informatics Department, Voorburg, The Netherlands.
- Cochran, W.G. (1977), *Sampling Techniques*, Third Edition. John Wiley & Sons (1977).
- Couper, M.P. (2000), Web surveys: A review of issues and approaches. *Public Opinion Quarterly* 64, blz. 464-494.
- Couper, M.P., Baker, R.P., Bethlehem, J.G., Clark, C.Z.F., Martin, J., Nicholls II & W.L., O'Reilly, J.M. (eds.) (1998), *Computer Assisted Survey Information Collection*. Wiley, New York.
- Dillman, D A. & Bowker, D. (2001), The web questionnaire challenge to survey methodologists. In: Reips, U.D. and Bosnjak, M. (eds.), *Dimensions of Internet Science*, Pabst Science Publishers, Lengerich, Duitsland.
- Fricker, R. & Schonlau, M. (2002), Advantages and disadvantages of Internet research surveys: Evidence from the literature. *Field Methods* 15, blz. 347-367.
- Haan, J. de & Van 't Hof, C. (2006), *Jaarboek ICT en samenleving, de digitale generatie*. Sociaal en Cultureel PlanBureau, Den Haag.
- Heerwegh, D. & Loosveldt, G. (2002), An evaluation of the effect of response formats on data quality in web surveys. Paper gepresenteerd op de International Conference on Improving Surveys, Kopenhagen, 2002.

- Horvitz, D.G. & Thompson, D.J. (1952), A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* 47, blz. 663-685.
- Kuusela, V. (2003), Mobile phones and telephone survey methods. In: Banks, R., Currall, J., Francis, J., Gerrard, L., Kahn, R., Macer, T., Rigg, M., Ross, E., Taylor, S. & Westlake, A. (Eds.), *ASC 2003 - The impact of new technology on the survey process. Proceedings of the 4th ASC international conference*, blz. 317-327. Chesham Bucks, UK: Association for Survey Computing (ASC).
- Leeuw, E. de & Collins M. (1997), Data collection methods and survey quality. In: Lyberg, L., Biemer, P., Collins, M., De Leeuw, E., Dippo, C., Schwarz, N. & Trewin, D. (eds.), *Survey Measurement and Process Control*. Wiley, New York, blz. 199-220.
- Ossenbruggen, R. van, Vonk, T. & Willems, P. (2006), Online panel, goed bekeken. *Clou* 24, blz 28-34.