SVR - Methoden en technieken 2010 / 3

Aanmaak en gebruik van gewichten voor surveydata

Met toepassing in SPSS

Jan Pickery

Studiedienst van de Vlaamse Regering

Vlaamse overheid

Aanmaak en gebruik van gewichten voor surveydata

Met toepassing in SPSS

Jan Pickery

Vlaamse overheid

Samenstelling
Diensten voor het Algemeen Regeringsbeleid
Studiedienst van de Vlaamse Regering (SVR)

Jan Pickery

Leescomité
Marc Callens, Ann Carton, SVR
Jelke Bethlehem, CBS Nederland
Denis Luminet, ADSEI
Geert Molenberghs, UHasselt

Verantwoordelijke uitgever
Josée Lemaître
Administrateur-generaal
Boudewijnlaan 30 bus 23
1000 Brussel

Lay-out cover
Diensten voor het Algemeen Regeringsbeleid
Communicatie
Patricia Van Dichel

Druk
Agentschap voor Facilitair Management

Depotnummer
D/2010/3241/356
http://www.vlaanderen.be/svr

Inhoudstafel

1. Inleiding

De Studiedienst van de Vlaamse Regering (SVR) wil bijdragen aan een kwaliteitsverhoging van de statistiekproductie binnen de Vlaamse overheid. Onze brochure over de principes van kwaliteitszorg in het statistische productieproces met heel wat aanbevelingen in verband met het verzamelen, verwerken en documenteren van statistische gegevens kadert daarin (APS, 2003). Verder willen wij ook concretere handleidingen aanbieden over het juiste gebruik van statistische technieken. Zo verschenen er *Technische rapporten* over de interpretatie van interactie-effecten in regressiemodellen (Pickery, 2008) en over het gebruik van contextuele regressiemodellen bij het vergelijken van landen (Callens, 2010).

In deze nota gaan we dieper in op de aanmaak van gewichten voor surveydata en het gebruik ervan bij de analyse. In deze uiteenzetting beperken we ons tot surveys van personen waarbij slechts een deel van de populatie bevraagd werd en waarbij die steekproef aselect getrokken is, een toevalssteekproef of een kanssteekproef dus. De principes zijn over het algemeen wel vlot overdraagbaar naar surveys van organisaties en sommige zijn ook toepasbaar als in eerste instantie de volledige populatie geselecteerd werd voor de survey. De beperking tot surveys gehouden bij een toevalssteekproef van personen, maakt het echter veel eenvoudiger om in deze tekst een eenvormig taalgebruik aan te houden, wat dan weer hopelijk de leesbaarheid ten goede zal komen.

De theoretische uiteenzetting wordt eerder beperkt. We verwijzen op verschillende plaatsen naar bestaande literatuur, waar de geïnteresseerde lezer meer uitleg kan vinden. Wel zullen we uitvoerig aandacht besteden aan twee toepassingen. De eerste betreft de survey naar "Sociaal-culturele verschuivingen in Vlaanderen" (SCV-survey), de tweede de survey van de stadsmonitor 2008. Deze praktische voorbeelden verduidelijken hopelijk de gebruikte werkwijze bij de berekening van de gewichten en de te volgen analysestrategie.

De rest van de tekst is als volgt opgebouwd: in sectie 2 verduidelijken we waarom gewichten nuttig en nodig kunnen zijn. In sectie 3 beschrijven we de assumpties die gelden bij gewogen analyses. In sectie 4 en 5 beschrijven we de verschillende stappen die gevolgd kunnen worden bij de berekening van de gewichten. In sectie 6 tonen we aan de hand van een fictief voorbeeld aan hoe onoordeelkundig gebruik van gewichten bij de analyse van surveydata tot verkeerde conclusies kan leiden. Sectie 7 illustreert zowel de berekening van gewichten als het gebruik ervan bij de analyse voor de SCV-survey 2008. Sectie 8 doet hetzelfde voor de survey van de stadsmonitor. Sectie 9 ten slotte besluit deze nota.

2. Waarom wegen?

Het gebruik van gewichten bij de analyse van surveydata heeft tot doel een aantal systematische fouten die gemaakt worden zo klein mogelijk te houden. Uitspraken gebaseerd op een survey die slechts een deel van de populatie bevraagd heeft, bevatten immers altijd fouten. Die afwijkingen of fouten worden doorgaans opgedeeld in twee componenten: steekproefvariabiliteit en vertekening. De eerste component geeft gewoon aan dat een andere, op dezelfde wijze getrokken, steekproef tot een ander resultaat had kunnen leiden. Dat is nu eenmaal het gevolg van een niet-uitputtende bevraging van de volledige populatie en is dus eerder een onzekerheid dan een fout. Deze onzekerheid neemt af naarmate de steekproefomvang groter wordt. De tweede component, vertekening of "bias", duidt erop dat de schatting die in de uitspraak vervat is, in een bepaalde richting verkeerd kan zijn, dit als gevolg van bepaalde tekorten bij de meetmethode of het onderzoeksopzet. Het is een fout die bij een herhaling van de onderzoeksopzet met nieuwe steekproeven telkens weer in dezelfde richting gaat. Biemer en Lyberg (2003, 38-43) delen fouten die tot vertekening kunnen leiden op in 5 componenten: specificatiefouten, fouten met betrekking tot het steekproefkader, non-responsfouten, meetfouten en procesfouten. *Specificatiefouten* treden op als het concept dat gemeten wordt door de vraag van de survey niet (volledig) overeenstemt met het concept dat gemeten zou moeten worden. Een inadequaat steekproefkader, de (administratieve) omschrijving van de te onderzoeken populatie, kan ertoe leiden dat niet de volledige populatie gedekt wordt in het onderzoek (*dekkingsfout*). Er is *non-respons* als niet bij alle eenheden (volledige) informatie kan bekomen worden. Respondenten, interviewers en vragenlijsten kunnen *meetfouten* veroorzaken. Respondenten kunnen foute informatie verstrekken (al dan niet gewild), interviewers kunnen het antwoordgedrag beïnvloeden en ambigue vragen of

verwarrende instructies kunnen verkeerde antwoorden uitlokken. Tot slot kan er bij het *verwerkingsproces* van de data ook nog een en ander mislopen met de invoer, codering, bewaring…

Gewichten hebben tot doel de dekkingsfouten en (vooral) de non-responsfouten te verkleinen. Een voorbeeld van vertekening als gevolg van non-respons treedt op als we met een survey het gemiddelde inkomen van een populatie willen schatten, maar armere mensen systematisch minder geneigd waren om deel te nemen aan zo'n survey. De schatting van het gemiddelde inkomen is vertekend als gevolg van de non-respons. De grootte van de vertekening is een functie van de hoogte van de non-respons en van de mate waarin respondenten en non-respondenten verschillen. In een overzichtsartikel dat onderzoek naar non-responsvertekening bundelt, komt Groves (2006) tot de bevinding dat de vertekening slechts beperkt gecorreleerd is met de hoogte van de non-respons. Responsverhogende maatregelen hebben ook niet noodzakelijk een positieve impact op de non-responsvertekening. Ook Schouten e.a. (2009) geven een voorbeeld waarbij een survey met een hogere respons juist meer vertekende schattingen oplevert van het aantal uitkeringstrekkers en het aantal allochtonen dan een survey met een lagere respons.

Een maat voor de totale fout, die dus zowel de steekproefvariabiliteit als de vertekening omvat, is de "Mean Squared Error" (MSE). De MSE kwantificeert de mate waarin een op basis van een toevalssteekproef geschatte parameter verschilt van de werkelijke populatiewaarde. In hetzelfde voorbeeld geeft de MSE dus een inschatting van de mate waarin het gemiddelde inkomen, geschat met behulp van een survey, verschilt van het werkelijke populatiegemiddelde. Omdat de werkelijke populatiewaarde (meestal) niet gekend is, kan de exacte waarde van de MSE niet berekend worden, maar schattingen zijn wel mogelijk.

Het gebruik van gewichten heeft tot doel deze MSE te minimaliseren. Gewichten moeten er vooral voor zorgen dat surveyschattingen minder vertekend zijn en spelen dus voornamelijk in op de tweede component van de MSE. Concreet proberen gewichten de over- of ondervertegenwoordiging van bepaalde groepen respondenten te corrigeren en dus vertekening die het gevolg is van non-respons, van een verkeerde dekking van de populatie of van ongelijke selectiekansen te remediëren. De niet-evenredige vertegenwoordiging van sommige groepen respondenten in de (gerealiseerde) survey die hiervan het resultaat is, maakt dat uitspraken gebaseerd op die survey vertekend zullen zijn. Het gebruik van gewichten kan die vertekening (of bias) verminderen. De mogelijke winst die het wegen oplevert, gaat wel uit van een belangrijke assumptie, namelijk dat de over- of ondervertegenwoordiging niet rechtstreeks gerelateerd is aan de variabelen waarin de survey geïnteresseerd is. In de volgende sectie komen we terug op deze assumptie. Als aan deze assumptie niet voldaan is, zijn er eventueel nog modelmatige oplossingen, maar zulke uitbreidingen vallen buiten het bestek van deze tekst. Voorbeelden van zulke modelmatige oplossingen kunnen gevonden worden in Little & Rubin (2002).

De winst die geboekt kan worden door het gebruik van gewichten, gaat meestal ten koste van de precisie. De variantie van de bekomen parameters (de steekproefvariabiliteit) is groter als er (sterk) verschillende gewichten worden toegepast (Kish, 1992). Sommige auteurs stellen dat dit niet altijd het geval hoeft te zijn bij wegen voor non-respons (Little and Vartivarian, 2005), maar door hun effect op de eerste component van de MSE kunnen gewichten toch ook een negatieve impact hebben op de totale fout, ook al reduceren ze misschien de vertekening.

Beide effecten van de gewichten moeten tegen elkaar afgewogen worden. In principe kan zulke afweging ("***bias reduction versus variance inflation***") voor elke afzonderlijke analyse gebeuren, maar dat is uiteraard niet zo realistisch. Omdat de steekproefvariabiliteit afneemt naarmate de steekproef groter wordt en de vertekening niet, kan er wel van uitgegaan worden dat bij grotere steekproeven de vertekening een belangrijkere component is van de MSE dan de steekproefvariabiliteit. Daarom vinden sommige auteurs de (mogelijke) impact van de gewichten op de precisie van minder belang (zie bijvoorbeeld Schouten, 2004).

3. Assumpties bij gewogen analyses

Wegen voor non-respons gaat uit van een belangrijke assumptie met betrekking tot het non-responsmechanisme. Little & Rubin (2002, 11-19) onderscheiden drie mechanismen die tot ontbrekende data kunnen leiden, op basis van de vraag of het ontbreken van waarden voor een variabele gerelateerd is aan die variabele zelf of aan andere variabelen in het databestand. Die mechanismen zijn belangrijk omdat ze bepalen of een methode om met de ontbrekende data om te gaan adequaat is.

Een eerste mechanisme wordt Missing Completely at Random (MCAR) genoemd. Dit is van toepassing als het ontbreken van waarden niet samenhangt met de variabelen die voor alle eenheden geobserveerd zijn en evenmin met de waarden van de variabele met ontbrekende waarnemingen. Dit betekent niet noodzakelijk dat het patroon volledig toevallig is, wel dat het niet samenhangt met de waarden die de data in het bestand kunnen aannemen.

Het tweede mechanisme wordt Missing at Random (MAR) genoemd. Dat is van toepassing als het ontbreken van de waarden wel samenhangt met de waarden van de variabelen die voor alle eenheden geobserveerd zijn, maar, onder controle daarvan, niet met de waarden van de variabelen waarvoor er data ontbreken.

Het mechanisme wordt Missing Not at Random (MNAR) genoemd als het ontbreken van waarden afhankelijk is van de variabelen met ontbrekende waarden zelf.

Little & Rubin geven het voorbeeld van een databestand met twee variabelen: leeftijd en inkomen. Stel dat leeftijd gekend is voor iedereen; bij inkomen zijn er ontbrekende waarden. Het mechanisme is MCAR als het ontbreken van een waarde voor inkomen niet gerelateerd is aan iemands leeftijd, noch aan zijn/haar inkomen. Als de kans op het ontbreken van een waarde voor inkomen samenhangt met de leeftijd van de respondent, maar niet varieert tussen respondenten met dezelfde leeftijd zijn de data MAR. De data zijn MNAR als de waarschijnlijkheid dat het inkomen niet gekend is, varieert volgens het inkomen van mensen met dezelfde leeftijd. Bemerk dat MAR dus niet uitsluit dat er een samenhang is tussen de hoogte van het inkomen en het ontbreken van een waarde ervoor. Gecontroleerd voor leeftijd moet die samenhang wel verdwijnen om aan de assumptie te voldoen.

Een probleem bij de vertaling van deze mechanismen naar een surveycontext is dat de variabelen van de survey verschillende rollen kunnen spelen. Eigenlijk moeten de mechanismen daarom analyse per analyse bekeken worden. Voor een bepaalde analyse met een aantal X en Y-variabelen kan het mechanisme MCAR zijn, terwijl dat voor een andere analyse MAR of zelfs MNAR is. Een afzonderlijke inschatting van het non-responsmechanisme voor elke nieuwe analyse is echter niet altijd praktisch haalbaar. Bovendien beschouwen sommige auteurs het gebruik van één enkel gewicht voor een hele dataset juist als één van de grote voordelen van het werken met gewichten (zie bijvoorbeeld Lohr, 2007), ook al is herhaaldelijk aangetoond dat dezelfde gewichten niet noodzakelijk even efficiënt zijn voor alle doelvariabelen (zie bijvoorbeeld Schouten, 2004). Het gebruik van een gewicht levert dan onvertekende schatters op als aan een bepaalde vorm van de MAR-assumptie voldaan is, namelijk binnen een bepaalde weegcategorie of -klasse zijn de respondenten een toevalssteekproef van de geselecteerde personen. Binnen die bepaalde categorie geldt dus MCAR. De weegklassen kunnen gebaseerd zijn op variabelen van het surveydesign, of van de steekproefeenheden (zowel respondenten als non-respondenten) en kunnen het resultaat zijn van een eenvoudige categorisering of van een meer modelmatige aanpak.

Het probleem is natuurlijk dat de MAR-assumptie niet kan getest worden. Ze kan wel aannemelijker gemaakt worden door voor alle eenheden van het bestand meer informatie te verzamelen die voorspellende waarde heeft, zowel voor de variabelen waarin de survey geïnteresseerd is als voor het al dan niet ontbreken van een waarde ervoor. Als die informatie opgenomen kan worden in de analyse wordt de samenhang tussen het ontbreken en de waarde op zich onder controle gehouden. Een mogelijke schending van deze assumptie gebruiken als een argument om niet te wegen kan ook gevaarlijk zijn. Het is inderdaad zo dat de non-respons waarschijnlijk nooit volledig verklaard kan worden door de gekende variabelen. Evenmin zal de relatie tussen de gekende populatievariabelen en de variabelen waarin de survey geïnteresseerd is, altijd correct gespecificeerd zijn. Maar veralgemeningen van surveyresultaten naar de populatie

zonder weging of zonder toepassing van een andere techniek voor ontbrekende waarden gaan eigenlijk uit van de nog strengere MCAR-assumptie. Gewichten worden inderdaad best niet aanzien als de oplossing voor alle non-responsproblemen, maar het niet gebruik ervan als gevolg van een gebrek aan vertrouwen in een bepaald non-responsmodel kan een nog strenger model impliceren (Lohr, 1999, 272).

Er zijn een aantal complexe regels over welke technieken wanneer toegepast kunnen worden bij de verschillende non-responsmechanismen. Die regels vallen buiten het bestek van deze tekst. Meer informatie kan gevonden worden in de tekst van Little & Rubin (2002) en ook in Molenberghs & Kenward (2007).

4. Verschillende stappen bij het berekenen van gewichten

Gewichten zijn doorgaans het resultaat van een stapsgewijze aanpak en zo luiden ook de gangbare aanbevelingen. Zulke aanbevelingen zijn bijvoorbeeld te vinden in de richtlijnen voor de European Social Survey (ESS) of voor de Survey on Income and Living Conditions (SILC) en ook in enkele wetenschappelijke artikels (bijvoorbeeld Kalton & Flores-Cervantes, 2003; Biemer & Christ, 2008).

Drie stappen komen altijd terug (in min of meer dezelfde vorm):

1) basisgewichten (design weights)
2) compensatie voor unit-non-respons
3) een vorm van poststratificatie

1) Basisgewichten (design weights)
Het basisgewicht is gelijk aan of proportioneel aan de inverse van de selectiekans, de kans dat een bepaalde persoon opgenomen wordt in de (oorspronkelijke) steekproef. Bij een enkelvoudige aselecte steekproef is dit basisgewicht voor iedereen gelijk. Dat geldt ook voor zogenaamde epsem-designs (*equal probability of selection of elementary units*). In die designs heeft iedereen immers dezelfde kans om getrokken te worden. Het basisgewicht is dan gelijk aan de populatieomvang gedeeld door de steekproefomvang:

$$\frac{N}{n} \qquad (1)$$

Ook bij tweetrapssteekproeven waarbij de kans om getrokken te worden in de eerste stap proportioneel is aan de bevolkingsomvang zijn *epsem*. Bij complexere steekproefdesigns met bijvoorbeeld disproportionele stratificatie, meerdere "trappen" of bepaalde vormen van clustering kan dit basisgewicht wel variëren. Lohr (1999, 225-227) geeft enkele formules voor de berekening van het basisgewicht bij complexere steekproefdesigns.

2) Compensatie voor unit-non-respons
De tweede stap bij het wegen, heeft tot doel de vertekening die het gevolg is van selectieve respons te verkleinen. Op basis van informatie die beschikbaar is voor zowel respondenten als niet-respondenten wordt de responskans[1] geschat. De gewichten in deze stap 2 zijn de inversen van die responskansen, of proportioneel daaraan. De informatie die zowel voor respondenten als voor niet-respondenten beschikbaar is, is vaak gelimiteerd. Belangrijk is dat deze stap zich volledig beperkt tot data over de eigen steekproef.

3) Een vorm van poststratificatie
In de derde stap wordt er ook gekeken naar wat er gekend is voor de volledige populatie. In de meest bekende vorm van deze derde stap wordt de verdeling van de steekproef voor bepaalde variabelen gelijkgesteld met (gekende) populatieverdelingen. In de literatuur wordt deze vorm van poststratificatie celweging genoemd ("cel weighting"). De informatie die je hiervoor kan gebruiken is echter relatief beperkt (bijvoorbeeld een kruistabel met 2 à 3 variabelen, afhankelijk van het aantal categorieën/cellen en het aantal eenheden). Andere methoden laten toe om meer informatie op te nemen bij het berekenen van de gewichten. Voorbeelden van zulke andere methoden zijn:

1 In het Engels wordt taalkundig vaak een onderscheid gemaakt tussen selectiekans ("selection probability") en responskans ("response propensity"). Deze laatste term zou ook vertaald kunnen worden als geneigdheid. De taalkundige finesse duidt aan dat er bij responskansen een grotere mate van onzekerheid is. Responskansen worden immers geschat op basis van beschikbare data, terwijl selectiekansen gekend zijn.

- "raking", ook wel iteratief proportioneel fitten of multiplicatief wegen genoemd:
  Hierbij wordt alleen gekeken naar de marginale verdelingen van de variabelen en niet naar celfrequenties of –percentages. De werkwijze werd 70 jaar geleden al geïllustreerd door Deming en Stephan (1940).
- logistische regressie:
  De kans op opname in de survey wordt afhankelijk gesteld van een reeks variabelen en de gewichten zijn gelijk aan de inverse van die kans. Hierbij kunnen de respondenten ook gegroepeerd worden op basis van hun opnamekansen, waarna per groep één gewicht wordt berekend (Little & Rubin, 2002, 48-49).
- GREG (generalised regression estimation) ook wel lineair wegen genoemd:
  Hierbij wordt de informatie van de weeg- of hulpvariabelen gebruikt in een regressieschatter voor de aanpassing van gewichten (zie Bethlehem en Keller, 1987; Bethlehem, 2002).

Naast de hierboven vermelde referenties kan de geïnteresseerde lezer voor een overzicht van deze methoden ook terecht bij Kalton & Flores-Cervantes (2003) en Bethlehem (2008).

Stappen 2 en 3 beogen vergelijkbare effecten. Beide hebben tot doel om selectiviteit als gevolg van non-respons of een verkeerde dekking van de populatie te remediëren voor zover dat mogelijk is. Poststratificatie kan daarnaast ook toevallige onevenwichten in de steekproef remediëren. Het belangrijke verschil is dat voor stap 2 alle informatie afkomstig is van de steekproef. Voor deze stap moeten dezelfde variabelen beschikbaar zijn zowel voor respondenten als voor non-respondenten, of op z'n minst de marginale totalen voor deze variabelen bij de non-respondenten. Voor stap 3 wordt er gebruik gemaakt van populatie-informatie. Voor die stap is informatie van de non-respondenten dus niet nodig als voor de volledige populatie de verdeling van enkele variabelen van de respondenten gekend is. Een bijkomend verschil is dat wegen voor non-respons doorgaans de steekproefvariabiliteit verhoogt, terwijl poststratificatie bij intern homogene strata die variabiliteit juist kan verlagen.

Deze verschillende stappen komen eigenlijk overeen met drie verschillende redenen om te wegen: 1) wegen moet corrigeren voor ongelijke selectiekansen; 2) wegen kan corrigeren voor selectieve non-respons en 3) wegen kan door middel van poststratificatie de precisie verhogen. Het zijn de twee laatste redenen die samen maken dat wegen de MSE kan verlagen. De effectiviteit van die twee stappen staat of valt wel met de beschikbaarheid van goede weegvariabelen. Die moeten enerzijds gecorreleerd zijn met de variabelen waarnaar de interesses van het onderzoek uitgaan en ook respons kunnen verklaren. Deze verschillende stappen worden bijvoorbeeld ook onderscheiden in het Nederlandse Centraal Bureau voor de Statistiek, maar daar worden stap 2 en 3 meestal gecombineerd in één weging. Dat kan omdat er veel informatie beschikbaar is op populatieniveau.

5. Bijkomende aanpassing van de gewichten

In sectie 2 werd reeds gesteld dat het gebruik van (extreme) gewichten de steekproefvariabiliteit kan verhogen. Poststratificatiegewichten (stap 3) daarentegen kunnen de steekproefvariantie soms ook verlagen (Biemer & Christ, 2008, 331-332), vergelijkbaar met stratificatie bij het trekken van de steekproef, zij het doorgaans minder doeltreffend. Ook bij poststratificatie kunnen intern homogene en extern heterogene strata met betrekking tot de variabelen waarin de survey geïnteresseerd is, de steekproefvariantie verkleinen. Maar de zorg om het verkleinen van de vertekening kan soms een andere keuze van strata opdringen. Twee potentieel tegenstrijdige strategieën dragen bij aan de doelstelling van de weging om de MSE zo klein mogelijk te maken. De vertekening kan verkleind worden door weegcategorieën te kiezen die de non-respons voorspellen. De variantie kan onder controle gehouden worden door weegcategorieën te kiezen die de variabelen waarin de survey geïnteresseerd is voorspellen (Little & Rubin, 2002, 48). De afweging "bias reduction versus variance inflation" blijft ook hier van tel.

Vanuit de overweging over de impact van gewichten op de variantie wordt het **matigen van extreme gewichten** ("***weight trimming***") soms als een bijkomende afzonderlijke stap vermeld naast de drie in sectie 3 vernoemde stappen. Extreme gewichten hebben een zodanige negatieve invloed op de precisie van de schatters dat de eventuele winst van de verminderde vertekening volledig teniet gedaan kan worden. Het resultaat is dat een parameter misschien wel wat minder

vertekend wordt geschat, maar dat het betrouwbaarheidsinterval errond zeer fel vergroot. De winst is dan nihil (of zelfs negatief). Potter (1990) beschrijft enkele methoden om extreme gewichten te milderen, maar vaak worden ook heel eenvoudige vuistregels gebruikt om extreme gewichten af te toppen. Voorbeelden van zulke vuistregels zijn: het gewicht mag niet groter zijn dan het mediane gewicht plus 5 of 6 keer de interkwartielafstand; of het gewicht mag niet groter zijn dan 5 keer het gemiddelde gewicht (Battaglia e.a. 2004); of het gewicht mag niet groter zijn dan het gemiddelde gewicht plus 3 keer de standaardafwijking (Biemer en Christ, 2008, 338). Bij celweging zijn extreme gewichten vaak het resultaat van cellen met (zeer) weinig eenheden. Het samenvoegen van cellen kan dan ook een manier zijn om extremere gewichten te vermijden en stabielere schattingen van de gewichten te bekomen.

De ESS-richtlijnen ten slotte vermelden nog het **herschalen van de gewichten** als een laatste stap. De bekomen gewichten kunnen herberekend worden zodat ze een gemiddelde hebben van 1. Dit kan gebeuren bij elke stap of eventueel alleen op het einde. Als de gewichten een gemiddelde hebben van 1 dan is het gewogen aantal eenheden van de survey gelijk aan het oorspronkelijke aantal. Dit geeft de indruk dat de gewichten geen impact zullen hebben op de resultaten van de statistische tests uitgevoerd op de surveydata. Zoals het voorbeeld in de volgende sectie zal aantonen, is die indruk echter niet altijd correct. Je ziet ook steeds vaker gewichten die voor stap 1 de inverse van de selectiekans nemen én die niet herschaald zijn (Vlaamse voorbeelden kunnen gevonden worden bij SILC[2], Gezondheidsenquête[3],…). Bij een eenvoudig gebruik van die gewichten (bijvoorbeeld de *defaultwijze* in SPSS) krijg je dan resultaten voor zo'n 6 miljoen Vlamingen, wat voor enkele onderzoekers en gebruikers van surveydata waarschijnlijk eigenaardig zal overkomen. Maar het zijn niet die gewichten die het probleem vormen, wel hoe SPSS en vele andere statistische software er standaard mee omspringen. Het is wel zo dat de mogelijke fouten normaal gezien kleiner zijn als de gewogen steekproefomvang gelijk is aan de ongewogen steekproefomvang. Herschaling is daarom niet slecht. Maar het herschalen alleen zorgt niet voor een correcte berekening van standaardfouten. Daarvoor zijn andere rekentechnieken en specifieke softwareoplossingen vereist. Die softwareoplossingen zijn bovendien meestal ongevoelig voor de absolute grootte van de gewichten, zodat het herschalen in dat geval niet noodzakelijk is.

6. Gewogen analyses

6.1. Eenvoudig voorbeeld

In deze sectie presenteren we een fictief voorbeeld. Volgens dit eenvoudige voorbeeld hebben we bij een steekproef van 1500 Vlamingen een houding gemeten (voor of tegen een beleidsmaatregel). Maar de geslachtsverdeling in onze steekproef is (helemaal) verkeerd. In de populatie zijn er exact evenveel mannen als vrouwen (telkens 2 miljoen). In de steekproef hebben we echter meer dan 80% mannen en minder dan 20% vrouwen. Er zijn bovendien aanwijzingen dat mannen en vrouwen een andere mening hebben over de beleidsmaatregel. We gaan tenslotte uit van de MAR-assumptie. De respons wordt bepaald door geslacht, maar binnen de geslachtscategorieën niet door de houding. Gegeven die assumptie weten we dat onze inschatting van de houding in de populatie vertekend zal zijn als we niet wegen voor geslacht.

We bespreken in dit voorbeeld de berekening van de gewichten, maar voornamelijk de impact die die gewichten hebben op statistische toetsen. Hoe die gewichten tot stand zijn gekomen, doet voor dit voorbeeld eigenlijk niet ter zake. Zij kunnen zowel een resultante zijn van de drie verschillende stappen die in sectie 4 vermeld werden (een combinatie van een compensatie voor een ongelijke selectiekans, selectieve non-respons volgens geslacht en bijkomende aanpassing aan de gekende populatieverdeling) als enkel en alleen het gevolg van poststratificatie (zoals de berekening hieronder doet vermoeden). Conceptueel zijn dat natuurlijk verschillende gewichten. Bovendien hebben ze niet helemaal dezelfde impact op de geschatte varianties. Maar dat onderscheid is hier niet essentieel om mogelijke ongewenste gevolgen van het gebruik van gewichten bij statistische toetsen te illustreren.

---

[2] Survey voor Statistics on Income and Living Conditions, zie http://statbel.fgov.be/nl/statistieken/gegevensinzameling/enquetes/silc/index.jsp

[3] Gezondheidsenquête door middel van Interview voor België in opdracht van het Wetenschappelijk Instituut Volksgezondheid, zie http://www.iph.fgov.be/EPIDEMIO/EPINL/index4.htm

Tabel 1 geeft de ongewogen verdeling naar geslacht van onze hypothetische steekproef.

Tabel 1 Verdeling volgens geslacht in de ***ongewogen*** steekproef

	Frequentie	Percentage
Man	1.207	80,5
Vrouw	293	19,5
Totaal	1.500	100,0

Als we bij die steekproef kijken naar de houding t.o.v. de beleidsmaatregel, ziet de ongewogen verdeling eruit zoals in tabel 2.

Tabel 2 Houding tegenover beleidsmaatregel (***ongewogen***)

	Frequentie	Percentage
Voor	734	48,9
Tegen	766	51,1
Totaal	1.500	100,0

We gaan ervan uit dat we de geslachtsverdeling van de populatie kennen op basis van registers (tabel 3).

Tabel 3 Verdeling volgens geslacht in de populatie

	Frequentie	Percentage
Man	2.000.000	50,0
Vrouw	2.000.000	50,0
Totaal	4.000.000	100,0

We kunnen op basis van deze verdelingen gewichten berekenen. Op basis van de frequenties zeggen we dat de 1.207 mannen in onze steekproef 2.000.000 mannen in de populatie weergeven. De 293 vrouwen representeren eveneens 2.000.000 vrouwen in de populatie. De niet-herschaalde gewichten zijn dan:

$$\text{voor de mannen: } \frac{2.000.000}{1.207} = 1.657{,}00$$
$$\text{voor de vrouwen: } \frac{2.000.000}{293} = 6.825{,}94 \qquad (2)$$

Als we deze gewichten op de "default"wijze gebruiken in SPSS krijgen we in onze tabellen aantallen van 4 miljoen mensen (en zal alles wat we testen natuurlijk ook significant zijn). Daarom is het gebruikelijker om te werken met herschaalde gewichten. Dat kan eenvoudig door de gewichten te vermenigvuldigen met de factor "totale steekproefomvang gedeeld door totale bevolkingsomvang", oftewel: $\frac{1.500}{4.000.000}$. Dan krijgen we volgende gewichten:

$$\text{voor de mannen: } \frac{2.000.000}{1.207} \times \frac{1.500}{4.000.000} = 0{,}62$$
$$\text{voor de vrouwen: } \frac{2.000.000}{293} \times \frac{1.500}{4.000.000} = 2{,}56 \qquad (3)$$

Bemerk dat we die herschaalde gewichten ook direct kunnen bekomen, als we voor de berekening van de gewichten werken met percentages in plaats van met aantallen.

$$\text{voor de mannen: } \frac{50}{80,5} = 0,62$$
$$\text{voor de vrouwen: } \frac{50}{19,5} = 2,56 \qquad (4)$$

Ook de interpretatie van die herschaalde gewichten is evident. Mannen krijgen een gewicht kleiner dan 1 omdat ze in de steekproef oververtegenwoordigd zijn. Vrouwen krijgen een gewicht groter dan 1 omdat ze ondervertegenwoordigd zijn.

Als we deze herschaalde gewichten zullen gebruiken, krijgen we in onze tabellen terug totale aantallen die gelijk zijn aan de ongewogen aantallen (1.500 eenheden). Dat voelt waarschijnlijk comfortabeler aan, maar wil daarom nog niet zeggen dat de analyse ook correct wordt uitgevoerd.

Het gebruik van die gewichten bij de analyses geeft natuurlijk andere cijfers. De verdeling van geslacht is nu exact gelijk aan de populatieverdeling.

Tabel 4 Verdeling volgens geslacht in de ***gewogen*** steekproef

	Niet-herschaald gewicht		**Herschaald gewicht**	
	Frequentie	Percentage	Frequentie	Percentage
Man	2.000.000	50,0	750	50,0
Vrouw	2.000.000	50,0	750	50,0
Totaal	4.000.000	100,0	1.500	100,0

Tabel 4 toont dat de weging de respons representatief heeft gemaakt met betrekking tot de variabele geslacht. Vanuit onze MAR-assumptie weten we nu dat de respons ook representatief zal zijn voor de houding tegenover de beleidsmaatregel. Om de MAR-assumptie aannemelijk te maken bij reële data, proberen we de respons op dezelfde of op vergelijkbare wijze representatief te maken voor een reeks variabelen om zo ook voor de doelvariabelen van de survey zo representatief mogelijke resultaten te bekomen. In ons voorbeeld laat de houdingvariabele alvast een andere verdeling zien. Het aantal tegenstanders is (een beetje) groter dan op basis van de ongewogen frequentietabel leek.

Tabel 5 Houding tegenover beleidsmaatregel (***gewogen***)

	Niet-herschaald gewicht		**Herschaald gewicht**	
	Frequentie	Percentage	Frequentie	Percentage
Voor	1.893.369	47,3	710	47,3
Tegen	2.106.631	52,7	790	52,7
Totaal	4.000.000	100,0	1.500	100,0

De aantallen in de tweede kolom van tabellen 4 en 5 komen misschien misleidend over. Het is inderdaad ook zo dat er geen 4.000.000 personen bevraagd werden. Maar de aantallen in de vierde kolom zijn eigenlijk even virtueel. In ons bestand zijn er immers geen 750 vrouwen en geen 790 personen tegen de beleidsmaatregel. Als we met deze tabel de houding in de populatie willen schatten, kunnen we ons natuurlijk wel altijd baseren op het percentage en dat is identiek bij gebruik van beide gewichten.

De percentages van tabel 2 en tabel 5 verschillen omdat geslacht samenhangt met de houding. Uitgaande van de MAR-assumptie is de gewogen frequentieverdeling beter, want niet-vertekend. Zij geeft een correctere inschatting van de houding in de populatie. Maar deze weging zal ook – ongewild – een belangrijke rol spelen als we gaan kijken (toetsen) of mannen en vrouwen significant verschillen van houding. De meest eenvoudige toets om te kijken of er een verschil is,

is een chi-kwadraattoets. Moeten we dan opteren voor een gewogen of een ongewogen toets? Kruistabellen 6 tot 8 laten de resultaten van drie verschillende toetsen zien, gewogen met de 2 verschillende gewichten en ongewogen.

Tabel 6 Kruistabel van geslacht en houding tegenover beleidsmaatregel (***gewogen met niet-herschaalde gewichten***)

		Houding		
Geslacht		Voor	Tegen	Totaal
Man	Aantal	999.171	1.000.829	2.000.000
	Rijpercentage	49,96%	50,04%	
Vrouw	Aantal	894.198	1.105.802	2.000.000
	Rijpercentage	44,71%	55,29%	
Totaal	Aantal	1.893.369	2.106.631	4.000.000
	Rijpercentage	47,33%	52,67%	

Tabel 6 maakt duidelijk dat er inderdaad een verschil is tussen mannen en vrouwen. Bij de mannen is 50% tegenstander, bij de vrouwen meer dan 55%. De p-waarde van de chi-kwadraat is zeer klein (**p<0,001**). Ook 20 cijfers achter de komma staat er nog een 0 bij deze p-waarde. Het is natuurlijk niet toevallig dat die p-waarde zo klein is. SPSS denkt dat er werkelijk zulke grote aantallen zitten in de onderscheiden cellen. Deze significantietoets is dan ook waardeloos.

Als we werken met de herschaalde gewichten krijgen we dezelfde percentages, maar een ander resultaat voor de chi-kwadraattoets.

Tabel 7 Kruistabel van geslacht en houding tegenover beleidsmaatregel (***gewogen met herschaalde gewichten***)

		Houding		
Geslacht		Voor	Tegen	Totaal
Man	Aantal	375	375	750
	Rijpercentage	50,0%	50,0%	
Vrouw	Aantal	335	415	750
	Rijpercentage	44,7%	55,3%	
Totaal	Aantal	710	790	1.500
	Rijpercentage	47,3%	52,7%	

Ook op basis van tabel 7 zouden we besluiten tot een significant verschil (op niveau $\alpha = 0,05$) tussen mannen en vrouwen. De p-waarde van de chi-kwadraattoetsstatistiek is gelijk aan **0,039**. Maar als we dezelfde significantietoets opvragen voor de ongewogen data, krijgen we echter andere resultaten.

Tabel 8 Kruistabel van geslacht en houding tegenover beleidsmaatregel (***ongewogen***)

		Houding		
Geslacht		Voor	Tegen	Totaal
Man	Aantal	603	604	1.207
	Rijpercentage	49,96%	50,04%	
Vrouw	Aantal	131	162	293
	Rijpercentage	44,71%	55,29%	
Totaal	Aantal	734	766	1.500
	Rijpercentage	48,93%	51,07%	

Hoewel het totaal aantal tegenstanders in kruistabel 8 verschillend is van dat in tabellen 6 en 7, is het aandeel geschatte tegenstanders bij de mannen en de vrouwen afzonderlijk exact hetzelfde als in de gewogen analyse. Dat is natuurlijk logisch omdat de weging wel een impact heeft op de

verhouding van het aantal mannen tot het aantal vrouwen, maar niet op het aandeel tegenstanders voor elk geslacht. Belangrijker is dat de chi-kwadraattoets een ander resultaat oplevert. Op basis van de ongewogen analyse zouden we besluiten dat het verschil tussen mannen en vrouwen niet significant is. De p-waarde bedraagt nu **0,107** en bevindt zich dus aan de andere kant van de – arbitraire maar universeel gebruikte – grens van $\alpha = 0,05$.

Het verschil tussen mannen en vrouwen in aandeel tegenstanders is bij de ongewogen analyse niet significant terwijl dat bij de gewogen analyse wel zo was. Nochtans is het verschil altijd exact even groot. De vraag is welke toets de juiste is. Niemand zal zich baseren op de toets bij tabel 6, maar een keuze tussen de toets bij tabel 7 en deze bij tabel 8 is minder evident. De reden voor het andere resultaat is nochtans eenvoudig. Een toets die twee groepen vergelijkt, heeft een groter statistisch onderscheidingsvermogen als beide groepen even groot zijn. Onze weging heeft dus niet alleen geleid tot een betere inschatting van het aandeel tegenstanders onder alle Vlamingen, maar heeft ook een impact op het onderscheidingsvermogen van de statistische toets die mannen en vrouwen vergelijkt.

Een vergelijkbare impact treedt trouwens eveneens op voor analyses met verscheidene onafhankelijke variabelen. Dat blijkt bijvoorbeeld uit een logistische regressie waarbij we die houding tegen de beleidsmaatregel opnemen als afhankelijke variabele (0= voor; 1= tegen) en geslacht en leeftijd als onafhankelijke variabelen. Vrouw is een dummy die waarde 0 heeft voor mannen en waarde 1 voor vrouwen. Leeftijd hebben we gecentreerd rond het gemiddelde van 46 jaar ("devleeftijd").

We voeren twee logistische regressies uit, één keer gewogen en één keer ongewogen. Bij de gewogen analyse tonen we alleen deze met het herschaalde gewicht. De analyse met het oorspronkelijke gewicht levert identieke parameterschattingen op, maar alle p-waarden zijn in dat geval zeer klein ($p<0,001$). Ook in de ongewogen analyse zijn de geschatte parameters vrijwel dezelfde, maar het verschil tussen mannen en vrouwen is er niet significant, terwijl dat wel zo is bij de gewogen analyse (ook deze met herschaalde gewichten).

Tabel 9 Resultaten van de logistische regressie met houding als afhankelijke variabele en leeftijd en geslacht als onafhankelijke variabelen (***gewogen met herschaalde gewichten***)

	b	Standaardfout	p-waarde	e^b
Intercept	0,001	0,073	0,989	1,001
Devleeftijd	0,002	0,003	0,490	1,002
Vrouw	0,210	0,104	***0,042***	1,234

Tabel 10 Resultaten van de logistische regressie met houding als afhankelijke variabele en leeftijd en geslacht als onafhankelijke variabelen (***ongewogen***)

	b	Standaardfout	p-waarde	e^b
Intercept	0,002	0,058	0,974	1,002
Devleeftijd	0,000	0,003	0,809	0,999
Vrouw	0,211	0,131	***0,107***	1,235

De vraag is nu natuurlijk welke toets we moeten geloven. In de meeste gevallen is het correcte antwoord op deze vraag "geen van beide". Langs de ene kant zorgde de weging voor een toename van het onderscheidingsvermogen door de twee groepen even groot te maken. De eigenlijk verzamelde data zijn op dat vlak een correctere inschatting van het werkelijke onderscheidingsvermogen. Dit pleit dus voor de toets op ongewogen data. Langs de andere kant gaat het weegmodel uit van de veronderstelling dat de gewogen percentages niet vertekend en dus correcter zijn dan de ongewogen percentages. Dat impliceert dat we voor statistische toetsen gewogen percentages met elkaar vergelijken.

Je kan deze tegenstelling enigszins begrijpen vanuit het principe dat een betrouwbaarheidsinterval rond een percentage niet alleen bepaald wordt door het aantal eenheden, maar ook door de hoogte van het geschatte percentage. In de regel zijn betrouwbaarheidsintervallen voor

percentages rond de 50% bijvoorbeeld groter dan intervallen voor percentages rond de 20% – als de andere omstandigheden van de steekproef gelijk blijven.

Om een correcte statistische toets te hebben, moeten we dus een niet (of minder) vertekende inschatting van de percentages hebben én qua aantallen een man-vrouwverhouding die overeenstemt met de ongewogen data. Dat is mogelijk, ook met SPSS, maar via specifieke procedures. De defaultwijze waarop SPSS met gewichten omgaat, voldoet niet. SPSS gaat er standaard vanuit dat die gewichten frequentiegewichten zijn. SPSS denkt dat een eenheid met een gewicht van 3 betekent dat er eigenlijk 3 eenheden zijn met dezelfde waarden, wat natuurlijk niet correct is. Dit geldt trouwens ook voor de manier waarop vele andere statistische software standaard met gewichten omgaat.

Bij dit concrete geval kunnen we nog opmerken dat er eigenlijk wel een "beste keuze" bestaat tussen de twee vermelde opties, namelijk de toets op ongewogen data. Dat komt omdat de gewichten berekend zijn op slechts 1 variabele (geslacht), die ook opgenomen is in de analyse als onafhankelijke variabele. Als de gewichten gebaseerd waren op meerdere variabelen, die niet allemaal in het model opgenomen zijn, is het wel zo dat geen van beide opties goed is.

6.2. Correcte gewogen analyses en analyses van data afkomstig van steekproeven die afwijken van enkelvoudig aselecte toevalssteekproeven

Het probleem bij het standaardgebruik van gewichten is dus dat die gewichten geïnterpreteerd worden als "herhalingen" van de betreffende persoon in de steekproef. Het resultaat is dat, indien individu A een gewicht heeft dat dubbel zo groot is als het gewicht van individu B, de standaardberekening de meting bij individu A ook als dubbel zo nauwkeurig interpreteert. Dit is natuurlijk niet correct.

Er bestaan verschillende methoden om op een correcte manier significantietesten uit te voeren en/of schattingen te verkrijgen van standaardfouten bij gewogen analyses. Die methoden zijn overigens ruimer toepasbaar en algemeen gericht op "complexe steekproeven". Zij kunnen met gewichten corrigeren voor ongelijke selectiekansen, compenseren voor non-response, maar ook de stratificatie of clustering bij de steekproeftrekking in rekening brengen bij de berekening van de standaardfouten. Kortom vrijwel alles wat maakt dat we niet van een enkelvoudige toevalssteekproef kunnen spreken, kan bij analyse in rekening gebracht worden. De courant gebruikte technieken in deze context zijn eigenlijk benaderend. Exacte rekenformules worden immers snel te rekenintensief (Höfler e.a., 2005, 294).

Rust (1985) geeft een overzicht en een beschrijving van zulke methoden, die reeds langer gekend zijn, maar eerder recentelijk ruimer ingang gevonden hebben. Ruwweg kunnen de methoden ingedeeld worden in linearisatiemethoden ("Taylor series linearisation") en replicatieve methoden (onder andere "Balanced Repeated Replication" en de "jacknife method"). Deze verschillende methoden zouden vergelijkbare resultaten opleveren (Rodgers-Farmer & Davis, 2001), ook al zijn de replicatieve methoden vanuit theoretisch oogpunt beter omdat zij de gewichten als toevalsvariabelen beschouwen in tegenstelling tot de linearisatietechnieken.

Ook SPSS biedt tegenwoordig een aangepaste berekening van standaardfouten en significantietesten aan voor data afkomstig van steekproeven die niet voldoen aan de eis van een enkelvoudige toevalssteekproef. Die berekening zit vervat in de afzonderlijke module "COMPLEX SAMPLES" en gebruikt de "Taylor series linearisation"-methode. We hebben de analyse van sectie 6.1 uitgevoerd met die module Complex Samples en daarbij gebruik gemaakt van twee verschillende gewichten, het oorspronkelijke gewicht (zie formule (2)) en een gewicht dat herschaald is zodanig dat het gemiddelde van alle gewichten gelijk is aan 1 (zie (3)).

De geschatte aantallen en percentages in tabellen 11 en 12 zijn dezelfde als deze in tabellen 6 en 7, al werden de aantallen daar wel automatisch afgerond tot gehele getallen. Zowel voor het geschatte aantal als voor het geschatte percentage kunnen we met COMPLEX SAMPLES ook eenvoudig een betrouwbaarheidsinterval opvragen. Zo blijkt het betrouwbaarheidsinterval voor vrouwen heel wat groter (we kunnen met 95% betrouwbaarheid stellen dat het aandeel tegenstanders zich in het interval [49,5%-60,9%] bevindt) dan voor mannen (95%-

betrouwbaarheidsinterval van [47,2%-52,9%] voor aandeel geschatte tegenstanders). Dat is een evident gevolg van het feit dat er in onze steekproef meer mannen dan vrouwen zitten.

Tabel 11 Kruistabel van geslacht en houding tegenover beleidsmaatregel, analyse met COMPLEX SAMPLES (***niet-herschaald gewicht***)

			Houding		
Geslacht			Voor	Tegen	Totaal
Man		Geschatte aantal	999.171,5	1.000.828,5	2.000.000,0
	95% betrouw-	*ondergrens*	937.430,2	939.070,5	
	baarheidsinterv.	*bovengrens*	1.060.912,8	1.062.586,5	
		Geschatte percentage	50,0%	50,0%	
	95% betrouw-	*ondergrens*	47,1%	47,2%	
	baarheidsinterv.	*bovengrens*	52,8%	52,9%	
		Ongewogen aantal	603	604	1.207
Vrouw		Geschatte aantal	894.198,0	1.105.802,1	2.000.000,0
	95% betrouw-	*ondergrens*	747.745,1	944.794,6	
	baarheidsinterv.	*bovengrens*	1.040.650,8	1.266.809,5	
		Geschatte percentage	44,7%	55,3%	
	95% betrouw-	*ondergrens*	39,1%	49,5%	
	baarheidsinterv.	*bovengrens*	50,5%	60,9%	
		Ongewogen aantal	131	162	293
Totaal		Geschatte aantal	1.893.369,5	2.106.630,6	4.000.000,0
	95% betrouw-	*ondergrens*	1.749.585,8	1.951.530,7	
	baarheidsinterv.	*bovengrens*	2.037.153,1	2.261.730,4	
		Geschatte percentage	47,3%	52,7%	
	95% betrouw-	*ondergrens*	44,2%	49,5%	
	baarheidsinterv.	*bovengrens*	50,5%	55,8%	
		Ongewogen aantal	734	766	1.500

Op de geschatte aantallen en de betrouwbaarheidsintervallen daarvan na, zijn tabellen 11 en 12 identiek. Deze vaststelling geldt eveneens voor de resultaten van de significantietoets. Bij beide tabellen of analyses is de p-waarde horend bij de toets van onafhankelijkheid[4] gelijk aan **0,107**. Bij deze correctere manier van schatten doet het er dus helemaal niet toe of we onze gewichten nu herschaald hebben of niet.

[4] In COMPLEX SAMPLES is die kans niet gebaseerd op een gewone chi-kwadraat, maar op een aangepast F-toetsstatistiek.

Tabel 12 Kruistabel van geslacht en houding tegenover beleidsmaatregel, analyse met COMPLEX SAMPLES (***herschaald gewicht***)

Geslacht			**Houding** Voor	Tegen	Totaal
Man		Geschatte aantal	374,7	375,3	750,0
	95% betrouw-	*ondergrens*	351,5	352,2	
	baarheidsinterv.	*bovengrens*	397,8	398,5	
		Geschatte percentage	50,0%	50,0%	
	95% betrouw-	*ondergrens*	47,1%	47,2%	
	baarheidsinterv.	*bovengrens*	52,8%	52,9%	
		Ongewogen aantal	603	604	1.207
Vrouw		Geschatte aantal	335,3	414,7	750,0
	95% betrouw-	*ondergrens*	280,4	354,3	
	baarheidsinterv.	*bovengrens*	390,2	475,0	
		Geschatte percentage	44,7%	55,3%	
	95% betrouw-	*ondergrens*	39,1%	49,5%	
	baarheidsinterv.	*bovengrens*	50,5%	60,9%	
		Ongewogen aantal	131	162	293
Totaal		Geschatte aantal	710,0	790,0	1.500,0
	95% betrouw-	*ondergrens*	656,1	731,8	
	baarheidsinterv.	*bovengrens*	763,9	848,1	
		Geschatte percentage	47,3%	52,7%	
	95% betrouw-	*ondergrens*	44,2%	49,5%	
	baarheidsinterv.	*bovengrens*	50,5%	55,8%	
		Ongewogen aantal	734	766	1.500

Met COMPLEX SAMPLES kan eveneens een logistische regressie uitgevoerd worden. Ook hier sluit het resultaat van de significantietoets aan bij het resultaat voor de ongewogen toets. In een logistische regressie met leeftijd en geslacht als onafhankelijke variabelen is de p-waarde voor het effect van geslacht gelijk aan 0,108.

Dit fictieve voorbeeld maakt duidelijk dat het gebruik van geavanceerde software eigenlijk noodzakelijk is voor gewogen analyses van surveydata. Specifieke rekentechnieken zijn nodig om goede inschattingen te krijgen van varianties en p-waarden van zodra afgeweken wordt van een enkelvoudige aselecte toevalssteekproef en dus ook bij het gebruik van gewichten, ongeacht de herkomst van die gewichten. De specifieke rekenmethoden zijn nodig zowel voor designgewichten als voor non-respons- of poststratificatiegewichten. Idealiter maakt de software overigens een onderscheid tussen verschillende soorten gewichten omdat de impact van bijvoorbeeld non-responsgewichten op de standaardfouten anders is dan deze van poststratificatiegewichten (Dillman e.a., 2002, 19). SPSS Complex Samples gaat echter nog niet zover. Höfler e.a. (2005) geven een overzicht van de mogelijkheden van Stata, SAS, SUDAAN, S-Plus en SPSS voor de analyse van surveys waarbij de data niet afkomstig zijn van enkelvoudig aselecte steekproeven. Een inzicht in de concrete mogelijkheden van de verschillende softwareprogramma's is belangrijk. Binnen hetzelfde programma bieden verschillende procedures soms andere mogelijkheden. Molenberghs (2009, 588-739) vergelijkt bijvoorbeeld uitvoerig de mogelijkheden van verschillende SAS-procedures. Een vergelijkbare analyse van de mogelijkheden van SPSS is ons onbekend.

In de volgende 2 secties bespreken we reële voorbeelden. In sectie 7 kijken we naar de SCV-survey en tonen we hoe voor die survey gewichten berekend werden en hoe die gewichten gebruikt moeten worden bij de analyse. In sectie 8 doen we de weegstrategie voor de survey van de stadsmonitor uit de doeken.

7. Illustratie 1 – de SCV-survey

7.1. Beschrijving van de survey

Sinds 1996 wordt er in opdracht van de Studiedienst van de Vlaamse Regering jaarlijks een face-to-face survey uitgevoerd naar "Sociaal-culturele verschuivingen in Vlaanderen". Deze survey beoogt een steekproefomvang van ongeveer 1.500 respondenten te realiseren met de Belgische nationaliteit, die 18 tot 85 jaar oud zijn en (ook) Nederlands spreken. Het steekproefdesign verloopt in twee fasen. In de eerste fase worden postsectoren geselecteerd op basis van hun bevolkingsomvang (meervoudige trekking is mogelijk). In de volgende stap worden groepen van respondenten uit die sectoren getrokken op basis van het Rijksregister.

Tot 2003 werd er voor deze survey non-responsvervanging ("non-response substitution") toegepast. Personen die niet geïnterviewd konden worden, werden vervangen door een andere inwoner van de gemeente[5] met een vergelijkbare leeftijd. Vanaf 2004 heeft de Studiedienst van de Vlaamse Regering die subsitutie verlaten en sindsdien wordt er één bestand getrokken, dat uitputtend benaderd wordt. Bij iedere getrokken persoon wordt gepoogd om een interview af te nemen. Om vertekening als gevolg van lagere respons in enkele regio's (bijvoorbeeld grote steden) in de mate van het mogelijke tegen te gaan, wordt het aantal geselecteerde personen per sector variabel bepaald op basis van de responscijfers van voorgaande jaren (zie Pickery & Carton, 2008). In de meeste postsectoren worden 15 personen geselecteerd (met het oog op het bereiken van 10 respondenten), in sommige sectoren kunnen dat er meer zijn, in andere minder. Dit kan "differentiële oversampling" genoemd worden. De keuze voor deze procedure wordt mede ondersteund door de vaststelling dat sommige bevraagde variabelen samenhangen met verstedelijking (bijvoorbeeld mobiliteitsgedrag en houding tegenover vreemdelingen). Bovendien is de differentiële oversampling beperkt. Er wordt weinig afgeweken van het algemeen vooropgezette aantal zodat de impact op de varianties ook klein is. Een gevolg van deze procedure is natuurlijk wel dat er bij het steekproefdesign geen gelijke selectiekans van personen is. Dat zal onze eerste bezorgdheid zijn bij het bepalen van de gewichten.

De respons op de SCV-survey is behoorlijk (63,6% afgewerkte interviews op het totaal aantal geselecteerde personen in 2008), maar non-responsvertekening is niettemin waarschijnlijk net zoals in vrijwel elke survey (zie Groves, 2006). In de volgende stappen van de wegingsprocedure zullen we die non-responsvertekening in de mate van het mogelijke proberen te verminderen.

Bijkomend aan de SCV-survey is er ook nog een internationaal vergelijkbare module in het kader van het International Social Survey Program (ISSP). Deze ISSP-module wordt afgenomen met een zogenaamde drop-off. De interviewer laat een vragenlijst achter bij de respondent die deze moet terugsturen. De deelname aan deze extra module is meestal goed, maar natuurlijk niet 100% volledig, zodat er hiervoor een aparte weging aangewezen is.

7.2. Berekening van de gewichten voor de SCV-survey

Voor de berekening van de gewichten gaan we stapsgewijs te werk. Stappen 1), 2) en 3) zoals besproken in sectie 3 komen alle aan bod.

Stap 1
In stap 1 worden de designgewichten berekend. Deze gewichten zijn afhankelijk van selectiekansen, en die verschillen – voor Vlaanderen – van postsector tot postsector.

In het **Vlaamse Gewest** is het designgewicht de inverse van de selectiekans. We hebben 2.233 personen getrokken uit een populatie van 4.547.344 inwoners van het Vlaamse Gewest (aantal Belgische inwoners van 18-85 jaar op 01/01/2007). Het designgewicht weegt de 2.233 personen eigenlijk terug op tot 4.547.344, maar het is niet voor iedereen gelijk aan 2.036,4 (dit is de breuk tussen beide getallen: 4.547.344/2.233) omdat sommige postsectoren differentieel oversampled zijn. Dit gewicht moet die oversampling rechttrekken. Het designgewicht kan op volgende wijze berekend worden:

[5] Tot 2003 was de gemeente nog primaire steekproefeenheid en niet de postsector.

$$\text{oversamplingfactor} = \frac{\#\ geselecteerde\ personen\ voor\ deze\ cluster}{gemiddelde\ \#\ geselecteerde\ personen\ per\ cluster}$$

$$\text{selectiekans} = \frac{2.233}{4.547.344} \times \text{oversamplingfactor} \qquad (5)$$

$$\text{designgewicht} = \frac{1}{selectiekans}$$

Om het effect van dit gewicht te illustreren, kijken we naar de ongewogen en de gewogen verdeling van de 2.233 geselecteerde personen over de geselecteerde postsectoren. Omdat de tabel anders te omvangrijk zou worden, geven we in tabel 13 slechts enkele postsectoren weer.

Tabel 13 Verdeling van de geselecteerde personen over de geselecteerde postsectoren (***ongewogen***)

Postsector	Frequentie	Percentage
1502	15	0,7
1570	15	0,7
1600	15	0,7
1640	20	0,9
…	…	…
1800	18	0,8
…	…	…
2470	13	0,6
…	…	…
2800	35	1,6
…	…	…
Totaal	2.233	100,0

Uit deze tabel blijkt dat het aantal geselecteerde personen niet gelijk is voor alle postsectoren. In de meeste sectoren worden er 15 personen geselecteerd. Maar sector 1640 wordt op basis van voorgaande surveys als moeilijker bestempeld en hier worden 20 personen geselecteerd (eveneens met het oog op 10 interviews). Ook in sector 1800 zouden meer respondenten nodig zijn om 10 interviews te bekomen (18 in plaats van 15). Sector 2470 geldt dan weer als gemakkelijker, hier zouden 13 geselecteerde personen volstaan. Sector 2800 is qua bevolkingsomvang een grotere sector en eveneens “moeilijker te interviewen” dan gemiddeld. Deze sector werd tweemaal geselecteerd en per selectie zouden 17,53 personen getrokken moeten worden. Zo komen we dus aan 35 te selecteren respondenten voor deze sector.

Tabel 14 geeft dezelfde verdeling als tabel 13, maar ditmaal gewogen met het designgewicht zoals dat berekend is met de formules (5) op de voorgaande pagina.

Tabel 14 Verdeling van de geselecteerde personen over de geselecteerde postsectoren (***gewogen met het designgewicht***)

Postsector	Frequentie	Percentage
1502	31.146	0,7
1570	31.146	0,7
1600	31.146	0,7
1640	31.146	0,7
…	…	…
1800	31.146	0,7
…	…	…
2470	31.146	0,7
…	…	…
2800	62.292	1,4
…	…	…
Totaal	4.547.34	100,0

Tabel 14 maakt duidelijk dat het designgewicht de differentiële oversampling rechttrekt. Voor alle postsectoren waar één cluster geselecteerd werd, is het gewogen aantal gelijk aan 31.146. Ook blijkt uit de tabel dat de gewogen totale omvang gelijk is aan de populatieomvang waarvan vertrokken werd.

Het berekenen van het designgewicht voor **Brussel** is minder evident en eigenlijk wat "politiek". Hoeveel Nederlandstalige Brusselaars van 18 tot 85 jaar zijn er? Omdat talentellingen verboden zijn in België, moet die vraag onbeantwoord blijven. Wel weten we dat er op 1 januari 2007 voor die leeftijdscategorieën 552.308 Belgische inwoners waren in het Brusselse Hoofdstedelijke Gewest.

Stel dat we uitgaan van 15% Nederlandstalige Brusselaars onder de inwoners met Belgische nationaliteit[6], dan komen we uit op een populatie van 82.846. Omdat we in Brussel niet differentieel oversampled hebben, is de selectiekans voor alle Brusselaars gelijk. We hebben 88 Brusselaars geselecteerd voor een interview, wat maakt dat de selectiekans gelijk is aan 88/82.846 en het designgewicht aan de inverse daarvan.

$$\begin{aligned} \text{selectiekans} &= \frac{88}{82.846} \\ \text{designgewicht} &= \frac{1}{selectiekans} = 941{,}43 \end{aligned} \qquad (6)$$

De som van alle gewichten voor de 2.321 geselecteerde personen (2.233 in het Vlaamse Gewest en 88 in Brussel) is nu gelijk aan 4.630.190. Dat is exact gelijk aan de omvang van de theoretische populatie waaruit de steekproef getrokken werd (4.547.344 Belgische inwoners in het Vlaamse Gewest + 82.846 Nederlandstalige Brusselaars). Deze populatieomvang is natuurlijk theoretisch omdat het aantal Nederlandstalige Brusselaars niet gekend is.

Stap 2

In de tweede stap proberen we te compenseren voor selectieve non-respons. De informatie die we ter beschikking hebben om non-responsgewichten te berekenen komt gedeeltelijk van het Rijksregister. Onze steekproef bevat naast de naam en het adres van de geselecteerde personen ook nog het geslacht en de leeftijd. Voorts kunnen we ook gebruik maken van het rapport dat elke interviewer moet bijhouden over (alle contactpogingen met) de geselecteerde personen. Met deze informatie kunnen we responskansen berekenen en de inverse van die kansen gebruiken als non-responsgewicht. De beschikbare informatie is eigenlijk beperkt. Zoals gesteld in sectie 3 vertrekt

[6] Vaak wordt een percentage van 10% Nederlandstalige Brusselaars vooropgesteld, zie bijvoorbeeld ook de Belgische Gezondheidsenquête. Maar omdat we ons hier beperken tot de inwoners met de Belgische nationaliteit is die 10% misschien een onderschatting. Zo'n kleine 30% van de inwoners van het Brusselse Gewest heeft niet de Belgische nationaliteit.

deze weging van de MAR-assumptie en kan die assumptie aannemelijker gemaakt worden door meer variabelen op te nemen in het non-responsmodel. Het is duidelijk dat meer en betere hulpvariabelen eigenlijk aangewezen zijn, maar die zijn hier niet voorradig.

Voor het berekenen van de responskansen maken we gebruik van logistische regressie. Bij het steekproefdesign vertrokken we van de assumptie van een variabele non-respons per postsector. Voor het berekenen van de responskansen kunnen we die assumptie aanhouden door gebruik te maken van een multilevel logistische regressie met postsector als tweede niveau (zie bijvoorbeeld Agresti e.a., 2000; Rice, 2001). In zo'n model wordt voor elke postsector een residu geschat. Bij significant verschillende postsectoren zou het residu dan in rekening gebracht kunnen worden. Het steekproefdesign gaat inderdaad uit van een verschillende respons per postcode, zo niet zou de oversampling contraproductief zijn. We voorspellen dus dat het non-responsgewicht hoger is voor de postsectoren met lager designgewicht en vice versa. Een lager designgewicht werd toegekend aan de sectoren die oversampled waren, het is bij die sectoren dat we een lagere respons en een hoger non-responsgewicht verwachten. Uiteindelijk opteren we toch niet voor een multilevel logistische regressie, maar modelleren we de regionale variabiliteit met een variabele die gebaseerd is op een ruimtelijke structuurindeling van Vlaamse gemeenten (op basis van het Ruimtelijk Structuurplan Vlaanderen – RSV). Als we deze variabele opnemen in een multilevel-analyse verdwijnt vrijwel alle variantie op het niveau van de postsectoren. Een multilevelanalyse is bijgevolg niet nodig en we beperken ons tot een gewone logistische regressie. Toch hebben we door de opname van de ruimtelijke structuurindeling ook verschillende gewichten voor onderscheiden postsectoren. Maar sectoren behorende tot dezelfde RSV-categorie hebben zo wel dezelfde gewichten.

We volgen zo wel een andere strategie bij de analyse voor de steekproeftrekking (differentiële oversampling op basis van residu's uit een multilevelanalyse) dan bij de analyse voor het berekenen van de non-responsgewichten (geen multilevelanalyse, opname van een gemeente-indeling als onafhankelijke variabele). Dit verschil kan echter beargumenteerd worden vanuit het feit dat we bij deze non-responsanalyse bijkomende informatie beschikbaar hebben op het individuele niveau (bijvoorbeeld type woning) en vanuit de resultaten van de analyse (geen overblijvende variantie op niveau 2). Het resultaat is ook dat de non-responsgewichten minder sterk zullen variëren omdat er geen postcoderesiduelen zullen gebruikt worden. Vanuit het oogpunt van de precisie van de schatters, is minder variatie bij de gewichten een voordeel (zie sectie 2).

Zo lossen we trouwens ineens een andere belangrijke vraag bij deze stap op: wat te doen met Brussel? Brussel zat door de kleinere aantallen niet in de multilevel non-responsanalyse. Gegeven de andere procedure voor de designgewichten zou het bij een multilevelanalyse ook logischer zijn om Brussel uit de berekening te houden. Maar voor een aparte analyse voor Brussel blijven er dan wel heel weinig cases over (n = 88), met potentieel heel variabele (en vanuit precisieoogpunt onwenselijke) gewichten. Daarom nemen we de Brusselaars op in de algemene non-responsanalyse en beschouwen we Brussel als een aparte RSV-categorie.

Het non-responsgewicht is de inverse van de responskans. Maar, afhankelijk van het aantal gebruikte onafhankelijke variabelen bij de logistische regressie kunnen de geschatte responskansen en dus ook de gewichten onstabiel zijn. Een rechtstreeks gebruik van de responskansen bij de berekening van de gewichten gaat ook uit van een groot vertrouwen in de correcte specificatie van het non-responsmodel (Little & Rubin, 2002, 49). Daarom wordt er vaak voor geopteerd om de geselecteerde personen te groeperen volgens responskans (bijvoorbeeld in 5 groepen) en dan per groep een gemiddelde kans en gewicht te berekenen. Wij zullen beide procedures illustreren.

We vertrekken van een logistische regressie met respons (0/1) als afhankelijke variabele en als onafhankelijke variabelen leeftijd en geslacht uit het Rijksregister. Van het interviewerrapport behouden we type woning. Woonomgeving werd ook opgenomen in de logistische regressie. Maar beide variabelen zijn sterk gecorreleerd en omdat type woning een grotere impact blijkt te hebben op de responskans en meer dan waarschijnlijk ook betrouwbaarder geregistreerd is, behouden we alleen die variabele. Tot slot nemen we ook de RSV-indeling mee als onafhankelijke variabele.

De dichotome afhankelijke variabele van onze analyse is respons.

Tabel 15 Respons bij de SCV-survey

	Frequentie	Percentage
Non-respons	846	36,4
Respons	1.475	63,6
Totaal	2.321	100,0

Bron: SCV-survey 2008

Tabellen 16 tot en met 19 geven de frequentieverdelingen van de onafhankelijke variabelen voor zowel de volledige steekproef als voor alleen de respondenten.

Tabel 16 Geslachtsverdeling van de steekproef en van de respondenten

	Volledige steekproef		**Respondenten**	
	Frequentie	Percentage	Frequentie	Percentage
Man	1.141	49,2	736	49.9
Vrouw	1.180	50,8	739	50.1
Totaal	2.321	100,0	1.475	100.0

Bron: SCV-survey 2008

Tabel 17 Leeftijdsverdeling van de steekproef en van de respondenten

	Volledige steekproef		**Respondenten**	
	Frequentie	Percentage	Frequentie	Percentage
18-24j	233	10,0	163	11.1
25-34j	353	15,2	211	14.3
35-44j	412	17,8	273	18.5
45-54j	431	18,6	290	19.7
55-64j	369	15,9	233	15.8
65-74j	283	12,2	183	12.4
75-85j	240	10,3	122	8.3
Totaal	2.321	100,0	1.475	100.0

Bron: SCV-survey 2008

Tabel 18 Verdeling van de steekproef en van de respondenten volgens type woning

	Volledige steekproef		**Respondenten**	
	Frequentie	Percentage	Frequentie	Percentage
Eengezinswoning: open bebouwing of vrijstaande woning	899	38,7	646	43,8
Eengezinswoning: halfopen bebouwing	426	18,4	285	19,3
Eengezinswoning: gesloten bebouwing of rijwoning	532	22,9	333	22,6
Gebouw met appartementen of studio's + overige	464	20,0	211	14,3
Totaal	2.321	100,0	1.475	100,0

Bron: SCV-survey 2008

Tabel 19 Verdeling van de steekproef en van de respondenten volgens ruimtelijke structuurcategorieën

	Volledige steekproef		Respondenten	
	Frequentie	Percentage	Frequentie	Percentage
Centrumgemeente grootstedelijk gebied	248	10,7	134	9,1
Centrumgemeente regionaalstedelijk gebied	296	12,8	173	11,7
Grootstedelijk gebied	135	5,8	89	6,0
Regionaalstedelijk gebied	120	5,2	85	5,8
Structuurondersteunend kleinstedelijk gebied	240	10,3	165	11,2
Kleinstedelijk gebied op provinciaal niveau	240	10,3	168	11,4
Buitengebied	856	36,9	579	39,3
Vlaams stedelijk gebied rond Brussel	98	4,2	41	2,8
Brussels Hoofdstedelijk Gewest	88	3,8	41	2,8
Totaal	2.321	100,0	1475	100,0

Bron: SCV-survey 2008

Deze tabellen tonen al verschillen in respons. Zo wordt er een hogere respons opgetekend bij mannen, bij de jongste leeftijdscategorie en bij de groep van 35 tot 54 jaar, bij de bewoners van eengezinswoningen in open bebouwing en bij Vlamingen in het buitengebied. Deze variabelen worden opgenomen in de logistische regressie, die ook nog de interactie leeftijd-geslacht bevat. De resultaten van die logistische regressie zijn beschreven in tabel 20. Alle onafhankelijke variabelen in het model zijn categorisch en werden telkens dummygecodeerd met de laatste categorie als referentiecategorie. Ook al zijn de verschillen met de referentiecategorie zeker niet altijd significant, toch hebben met uitzondering van het hoofdeffect voor geslacht alle onafhankelijke variabelen een significant effect op de respons. Dat hoofdeffect van geslacht laten we ook in het model, omdat de interactie met leeftijd wel significant blijkt.

Tabel 20 Resultaten van de logistische regressie met respons als afhankelijke variabele*

	b	Stand. fout	p-waarde	e^b
Intercept	-0,917	0,272	0,001	0,400
Leeftijdsklasse			0,001	
18-24j	0,620	0,268	0,021	1,859
25-34j	0,706	0,238	0,003	2,026
35-44j	0,904	0,234	0,000	2,470
45-54j	0,834	0,229	0,000	2,304
55-64j	0,309	0,233	0,184	1,362
65-74j	0,514	0,244	0,035	1,672
Referentie: 75-85j	-	-	-	-
Geslacht			0,115	
Man	0,431	0,274	0,115	1,538
Referentie: vrouw	-	-	-	-
RSV-indeling			0,000	
Centrumgemeente grootstedelijk gebied	0,104	0,257	0,684	1,110
Centrumgemeente regionaalstedelijk gebied	0,072	0,255	0,778	1,075
Grootstedelijk gebied	0,352	0,293	0,228	1,423
Regionaalstedelijk gebied	0,469	0,309	0,129	1,598
Structuurondersteunend kleinstedelijk gebied	0,489	0,267	0,067	1,631
Kleinstedelijk gebied op provinciaal niveau	0,474	0,270	0,079	1,606
buitengebied	0,254	0,244	0,299	1,289
Vlaams stedelijk gebied rond Brussel	-0,698	0,311	0,025	0,498
Referentie: Brussels Hoofdstedelijk Gewest	-	-	-	-
Type woning			0,000	
Eengezinswoning: open bebouwing	0,984	0,134	0,000	2,674
Eengezinswoning: halfopen bebouwing	0,746	0,150	0,000	2,109
Eengezinswoning: gesloten bebouwing	0,607	0,134	0,000	1,836
Referentie: gebouw met appartementen	-	-	-	-
Man * leeftijdsklasse			0,004	
Man * 18-24j	0,052	0,401	0,898	1,053
Man * 25-34j	-0,791	0,355	0,026	0,454
Man * 35-44j	-0,833	0,348	0,017	0,435
Man * 45-54j	-0,641	0,345	0,063	0,527
Man * 55-64j	0,094	0,353	0,791	1,098
Man * 65-74j	-0,113	0,376	0,763	0,893
Referentie: vrouw, 75-85j	-	-	-	-

Bron: SCV-survey 2008

* De p-waarde naast vetgedrukte variabele refereert naar een toets voor de volledige variabele, een toets dat alle parameters geassocieerd met die variabele gelijk zijn aan 0.

De interpretatie van de verschillende parameters is niet de hoofdbekommernis van deze tekst, maar uit de tabel blijkt alvast dat bij de vrouwen de laagste respons wordt opgetekend bij de oudste leeftijdscategorie. Bij de mannen is de respons ook minder goed in de categorieën 25 tot 44 jaar. De respons is ook vrijwel overal beter dan in Brussel, alleen in de Vlaamse rand rond Brussel is hij nog slechter. Bij eengezinswoningen ten slotte worden betere responscijfers opgetekend dan bij appartementen.

Voor het berekenen van de gewichten zijn de responskansen belangrijker dan de interpretatie van de verschillende parameters. We kunnen met dit model eenvoudigweg de voorspelde kans voor elke geselecteerde respondent wegschrijven en de inverse daarvan gebruiken als non-responsgewicht.

$$\text{non-responsgewicht} = \frac{1}{responskans} \qquad (7)$$

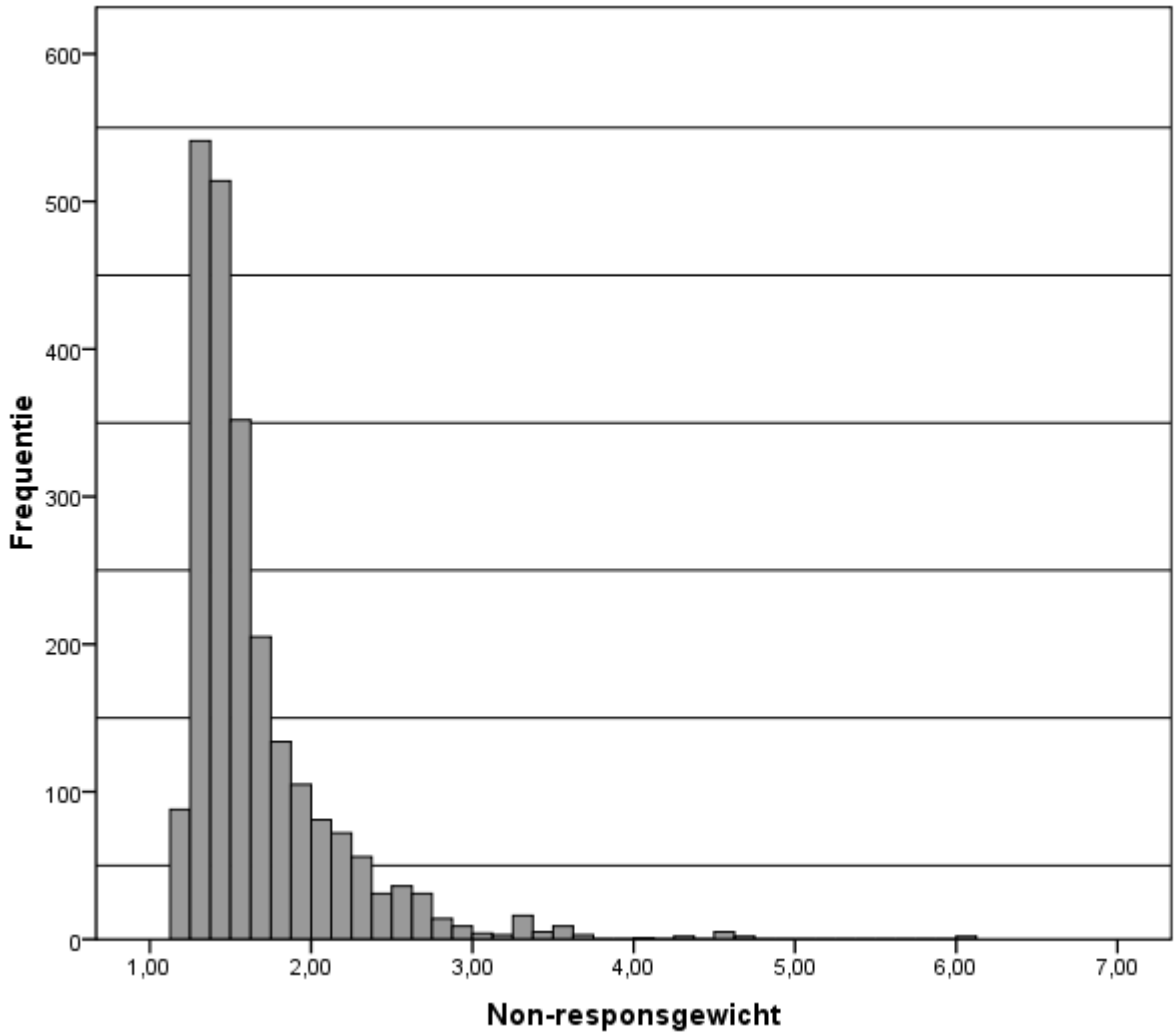
Het gemiddelde responsgewicht dat zo bekomen wordt, is gelijk aan 1,662 met een standaardafwijking van 0,476. De verdeling van de gewichten wordt getoond in tabel 21 en in grafiek 1. Om de tabel niet te overladen worden niet alle mogelijke waarden weergegeven.

Tabel 21 Verdeling van de non-responsgewichten voor alle geselecteerde personen

Gewicht	Frequentie	Percentage	Cumulatief percentage
1,19045	2	0,1	0,1
1.19343	3	0.1	0.2
1.19442	2	0.1	0.3
1.21838	2	0.1	0.4
1.23213	6	0.3	0.6
1.23577	12	0.5	1.2
1.23698	5	0.2	1.4
1.24102	27	1.2	2.5
…	…	…	
1.29378	50	2.2	9.7
1.29442	7	0.3	10.0
1.29904	4	0.2	10.1
1.30057	5	0.2	10.3
…	…	…	
3.18206	1	0.0	98.0
…	…	…	
4.69042	2	0.1	99.9
6.02648	2	0.1	100.0
Totaal	2.321	100,0	

Tabel 21 maakt duidelijk dat er zeer veel verschillende gewichten zijn. Ook al is het aantal onafhankelijke variabelen van de logistische regressie nog niet zo heel groot, toch kan er vanuit gegaan worden dat de bekomen gewichten niet zeer stabiel zijn. Dat is inherent aan deze aanpak. De meeste van die gewichten liggen wel dicht bij elkaar zoals grafiek 1 toont, maar toch zijn er ook enkele eerder extreme gewichten. Twee personen krijgen een gewicht dat gelijk is aan bijna vier keer het gemiddelde gewicht. Bemerk wel dat de gewichten uiteindelijk alleen toegekend zullen worden aan respondenten. Bij de twee geselecteerde personen met een gewicht groter dan 6, is geen interview afgenomen. Die gewichten zullen finaal dus verdwijnen. Maar dan nog zijn er een aantal gewichten die uit de band springen. Onbetwiste regels om extreme gewichten te duiden zijn er niet, maar als we de regel van Biemer en Christ (2008, 338) overnemen, dan is het criterium “gemiddeld gewicht +/- 3 keer de standaarddeviatie”. In dat geval zou 2% van de gewichten als eerder extreem kunnen bestempeld worden (alle gewichten groter dan 3,1).

Grafiek 1 Histogram van alle non-responsgewichten



Het gebrek aan stabiliteit en het voorkomen van extremere gewichten kunnen een aanleiding zijn om de responskansen op een andere wijze om te zetten in gewichten. De geselecteerde personen worden daarbij op basis van hun responskans in groepen gedeeld en voor elke groep wordt de gemiddelde responskans berekend en omgezet in een gewicht. Die kans kan berekend worden op basis van de resultaten van de logistische regressie of ook gewoon door het percentage respons te berekenen per groep. Bij deze laatste werkwijze kan je rechtstreeks verder werken met de aantallen die reeds gewogen zijn met het designgewicht (stap 1). Little & Rubin (2002, 48) suggereren hiervoor 5 of 6 groepen te nemen. Met 5 groepen levert die gegroepeerde procedure ons dus 5 non-responsgewichten op. Hoe die berekend worden kan afgeleid worden uit tabel 22.

De groepering in quintielen in de eerste kolommen is gebaseerd op de logistische regressie. De quintielen zijn niet exact even groot omdat sommige gewichten meer dan één keer voorkomen. De voorlaatste kolom geeft aan welk aandeel van de geselecteerde personen in het betreffende quintiel effectief geïnterviewd kon worden. Het gewicht in de laatste kolom is eenvoudigweg de inverse van die proportie respons.

Tabel 22 Berekening van de non-responsgewichten (na groepering van de kansen)

Groepering volgens respons-kans	Frequentie	Percentage	Gewogen aantal (design-gewicht)	Gewogen percentage (design-gewicht)	Proportie respons op de survey	Non-respons-gewicht
1ste quintiel	448	19,3	929.191	20,1	0,802	1,247
2de quintiel	454	19,6	940.837	20,3	0,723	1,383
3de quintiel	443	19,1	906.668	19,6	0,649	1,542
4de quintiel	470	20,2	926.542	20,0	0,580	1,724
5de quintiel	506	21,8	926.952	20,0	0,457	2,190
Totaal	2.321	100,0	4.630.190	100,0	0,640	1,617

Het gemiddelde non-responsgewicht van deze groepsgewijze berekening bedraagt 1,634 met een standaardafwijking gelijk aan 0,336. Deze is vanzelfsprekend kleiner dan bij de gewichten die rechtstreeks voortvloeien uit de logistische regressie. Bij deze nieuwe berekening situeert er zich geen enkel gewicht buiten het door Biemer en Christ (2008) voorgestelde interval – ook al is dat interval kleiner door de kleinere standaardfout.

Het is evident dat de correlatie tussen beide berekende gewichten hoog is (0,82). In die zin zal de keuze voor één van beide voor de meeste analyses en berekende parameters weinig uitmaken. Maar ter wille van de stabielere schattingen, gaat de voorkeur van Little & Rubin (2002) toch uit naar de gegroepeerde berekening.

Deze gewichten worden sowieso alleen toegepast op de uiteindelijke respondenten. Tabel 23 geeft de verdeling van de non-responsgewichten voor de 1.475 respondenten.

Tabel 23 Verdeling van de non-responsgewichten voor de respondenten

Gewicht	Frequentie	Percentage
1,247	359	24,3
1.383	328	22.2
1.542	287	19.5
1.724	272	18.4
2.190	229	15.5
Totaal	1.475	100,0

Het ***gewicht na stap 2*** is gelijk aan het product van het gewicht van stap 1 (designgewicht) en het gewicht van stap 2 (non-responsgewicht). Voor de berekening van het non-responsgewicht houden we rekening met het designgewicht. We gebruiken immers gewogen aantallen. Maar het non-responsgewicht omvat daardoor nog niet het designgewicht. Het gemiddelde gewicht na stap 2 is zo gelijk aan 3.139,1 en de som van alle gewichten voor de 1.475 respondenten is opnieuw gelijk aan 4.630.190, de theoretische omvang van onze populatie! Dit is logisch want door toepassing van het non-responsgewicht wordt de groep respondenten “opgewogen” tot de volledige steekproef en door het designeffect wordt de steekproef “opgewogen” tot de totale populatie. De mate waarin dit correct en succesvol is, is natuurlijk afhankelijk van de mate waarin aan de assumpties van het non-responsmodel voldaan is (zie sectie 3).

Een terzijde

Voor de berekening van het non-responsgewicht maken we geen gebruik van multilevelanalyse en de bijhorende postcoderesiduelen. De differentiële oversampling van het steekproefdesign was daar wel op gebaseerd. De vraag dringt zich dan op of onze steekproefaanpak wel nuttig was? Op dit moment hebben we een element om dat te evalueren. We kunnen de impact van de gewichten op de precisie van de schatters bij de analyse berekenen en doen dit voor het non-responsgewicht en het gewicht dat stap 1 en stap 2 combineert (het product van het designgewicht en het non-responsgewicht). Als het gecombineerde gewicht een kleinere impact heeft op de toename van de geschatte varianties, dan was ons oversampling design zinvol.

Een mogelijkheid om die impact te meten is het UWE-criterium (*Unequal Weighting Effect*)[7]. Dat gaat na wat de impact is van de spreiding van de gewichten op de schatting van de standaardfouten. Het UWE is altijd groter dan of gelijk aan 1 en geeft aan hoeveel groter de standaardfouten zijn door het gebruik van de gewichten in vergelijking met een ongewogen enkelvoudige aselecte steekproef. Een benaderende[8] formule voor het UWE is:

$$UWE \approx 1 + (cv)^2 \qquad \text{en} \qquad cv = \frac{std.dev(gewicht)}{gemid(gewicht)}$$

Voor de 1.475 respondenten is het gemiddelde non-responsgewicht gelijk aan 1,5689 met een standaarddeviatie van 0,3120. Het gemiddelde van het gecombineerde gewicht is gelijk aan 3.139,1119 met een standaarddeviatie van 613,4912. Dat geeft een UWE voor het non-responsgewicht die gelijk is aan 1,0396. Voor het gecombineerde gewicht wordt dat 1,0382. Het verschil is zeer klein, maar wel in het voordeel van het gecombineerde gewicht. We kunnen dus stellen dat ons steekproefdesign en het bijhorende designgewicht het non-responsgewicht en dus ook de impact ervan op de toename van de geschatte varianties "temperen".

Een correctere vergelijking van de impact van beide gewichten zou de Brusselse respondenten buiten beschouwing laten. De Brusselaars werden immers ook niet opgenomen in de multilevelanalyse van de respons. Bij die vergelijking worden de verschillen een beetje groter in het voordeel van het gecombineerde gewicht en dus van ons steekproefdesign (UWE voor gecombineerde gewicht zonder Brusselaars = 1,0338 en UWE voor non-responsgewicht = 1,0383).

Het verschil is zeer klein, maar ook zonder het gebruik van multilevelresiduelen bij de berekening van het non-responsgewicht, zorgt ons steekproefdesign dus voor minder variabele en vanuit precisieoogpunt betere gewichten. Dit besluit geldt natuurlijk gegeven het gebruikte non-responsmodel.

Stap 3

In stap 3 proberen we onze steekproef te conformeren aan de verdeling van enkele variabelen in de populatie, voor zover die gekend zijn. Veel onbetwistbare of onbetwiste externe bronnen ("golden standards") met bruikbare informatie zijn er echter niet. De variabelen die we uiteindelijk behouden hebben, zijn woonplaats (Brussels Gewest/Vlaams Gewest), huishoudgrootte en opleidingsniveau. Voor de eerste twee variabelen zijn de populatieverdelingen eenvoudig af te leiden uit de beschikbare Rijksregisterinformatie, ook al zal huishoudomvang in het Rijksregister waarschijnlijk niet 100% correct geregistreerd zijn. Voor de laatste variabele is er geen populatieverdeling beschikbaar. Wel is er de Belgische variant van de *Labour Force Survey* (Enquête naar de Arbeidskrachten, EAK). Deze grootschalige en verplichte enquête levert de best beschikbare informatie, maar het blijft natuurlijk een survey met waarschijnlijke fouten tot gevolg. Cijfers gebaseerd op de laatste volkstelling (de Socio-Economische Enquête van 2001) zijn onvoldoende actueel om bruikbaar te zijn.

[7] Vaak wordt er ook verwezen naar de effectieve steekproefomvang in plaats van naar het UWE-criterium. De effectieve steekproefomvang is de grootte van een enkelvoudige aselecte steekproef die dezelfde precisie zou hebben als de actuele steekproef. Om die effectieve steekproefomvang te berekenen kan de actuele steekproefomvang gedeeld worden door het UWE.

[8] Deze formule is alleen exact als aan een aantal assumpties voldaan is.

Woonplaats werd ook al gebruikt in stap 2 en met de ruimtelijke structuurcategorieën zelfs meer gedetailleerd dan de opdeling Vlaams Gewest/Brussel. Dat neemt echter niet weg dat die variabele ook in deze stap nog gebruikt kan worden. Omdat het logistische regressiemodel dat gehanteerd werd bij stap 2 niet 100% verzadigd was (niet alle mogelijke interactie-effecten werden opgenomen) en door de groepering van de responskansen in 5 groepen, kan de na stap 2 gewogen steekproef nog verschillen van de populatieverdeling. Dat kunnen we in deze stap rechttrekken.

Na stap 2 ziet de verdeling Brussel – Vlaams Gewest er zo uit:

Tabel 24 Verdeling van de respondenten volgens woonplaats (***gewogen met de gewichten na stap 2***)

Gewicht	Frequentie	Percentage
Vlaams Gewest	4.552.424	98,3
Brussels Gewest	77.766	1,7
Totaal	4.630.190	100,0

Met een eenvoudige breuk passen we de gewichten aan zodat we de aantallen krijgen waarvan we vertrokken waren (zie ook stap 1 – designgewichten). Voor de inwoners van het Vlaamse Gewest wordt het gewicht na stap 2 vermenigvuldigd met de breuk [4.547.344/4.552.424], voor de Brusselaars is dat [82.846/77.766]. Tabel 25 toont de nieuwe verdeling volgens woonplaats, gebruik makend van de aangepaste gewichten.

Tabel 25 Verdeling van de respondenten volgens woonplaats (***gewogen met de voor woonplaats aangepaste gewichten***)

Gewicht	Frequentie	Percentage
Vlaams Gewest	4.547.344	98,2
Brussels Gewest	82.846	1,8
Totaal	4.630.190	100,0

In theorie zouden we deze poststratificatie volgens woonplaats nog gedetailleerder kunnen uitvoeren. Binnen het Vlaamse Gewest kennen we bijvoorbeeld de populatieverdeling over de provincies of over de ruimtelijke structuurcategorieën. Maar bij ons steekproefdesign gaan we er niet alleen van uit dat er een ongelijke verdeling is van de respons, maar ook van het aandeel mensen dat in aanmerking komt voor een interview. Zieken, Nederlandsonkundigen,… kunnen niet geïnterviewd worden en de spreiding daarvan over de verschillende gemeenten en provincies is niet gelijkmatig. Poststratificeren met de officiële bevolkingsaantallen zou zo'n ongelijke verdeling negeren. Daarom beperken we ons bij de poststratificatie volgens woonplaats tot de opdeling Brussels Gewest/Vlaams Gewest.

De volgende variabelen die we in stap 3 gebruiken zijn huishoudomvang en opleidingsniveau. Een probleem hierbij is eens te meer het onderscheid Brussel–Vlaams Gewest. Zowel voor de informatie van het Rijksregister als voor de EAK zijn de verdelingen gekend per gewest, maar niet per taalgroep. Voor het Vlaamse Gewest is er dus geen probleem, maar het is onwaarschijnlijk dat het opleidingsniveau en de huishoudomvang van de Nederlandstaligen in Brussel overeenstemmen met deze voor het volledige Brusselse Gewest. Evenmin kunnen de Nederlandstalige Brusselaars gelijk gesteld worden met de inwoners van het Vlaamse Gewest. De populatietabellen tonen alvast grote verschillen tussen beide gewesten en ook in onze steekproef zijn de verdelingen van huishoudomvang en opleidingsniveau duidelijk verschillend voor de respondenten uit Brussel en de respondenten uit de rest van Vlaanderen. In Brussel hebben we bijvoorbeeld veel meer alleenstaanden en meer hooggeschoolden dan in het Vlaamse Gewest. Om deze redenen besluiten wij om de poststratificatie voor de variabelen huishoudgrootte en opleidingsniveau alleen toe te passen voor de inwoners van het Vlaamse Gewest. Dat betekent dat het verhaal voor de Brusselaars hier ophoudt. Het definitieve gewicht voor hen is datgene dat ook al gebruikt werd in tabel 25.

Voor de inwoners van het Vlaamse Gewest kijken we nog naar de populatieverdelingen voor huishoudgrootte en opleidingsniveau. Een vergelijking van tabellen 26 en 27 leert dat een weging volgens huishoudomvang slechts een beperkte impact zal hebben op de geschatte parameters en hun varianties. De verdeling van de voorlopig gewogen gerealiseerde steekproef is zeer vergelijkbaar met de populatieverdeling.

Tabel 26 Verdeling van de respondenten volgens huishoudomvang (***gewogen met de voor woonplaats aangepaste gewichten***)

Huishoudgrootte	Frequentie	Percentage
1	608.759	13,4
2	1.698.887	37,4
3	902.476	19,8
4	863.072	19,0
5	333.816	7,3
6 of meer	140.335	3,1
Totaal	4.547.344	100,0

Tabel 27 Populatieverdeling volgens huishoudomvang (***op basis van het Rijksregister***)

Huishoudgrootte	Frequentie	Percentage
1	673.148	14,8
2	1.593.984	35,1
3	927.689	20,4
4	851.043	18,7
5	314.324	6,9
6 of meer	187.156	4,1
Totaal	4.547.344	100,0

Eenzelfde vergelijking van de verdeling volgens opleidingsniveau laat veel grotere verschillen zien. Net zoals de meeste andere surveys (zie Billiet, 2007) kent de SCV-survey een duidelijke ondervertegenwoordiging van de laagstgeschoolden. Alle andere opleidingsniveaus zijn (licht) oververtegenwoordigd.

Tabel 28 Verdeling van de respondenten volgens opleidingsniveau (***gewogen met de voor woonplaats aangepaste gewichten***)

Huishoudgrootte	Frequentie	Percentage
Geen/lager onderwijs	578.069	12,7
Lager secundair onderwijs	940.142	20,7
Hoger secundair onderwijs	1.658.569	36,5
Niet-universitair hoger onderwijs	984.876	21,7
Universitair hoger onderwijs	385.689	8,5
Totaal	4.547.344	100,0

Bron: SCV-survey 2008

Tabel 29 Geschatte populatieverdeling volgens opleidingsniveau (***op basis van de EAK***)

Huishoudgrootte	Frequentie	Percentage
Geen/lager onderwijs	861.160	19,0
Lager secundair onderwijs	885.560	19,5
Hoger secundair onderwijs	1.616.478	35,6
Niet-universitair hoger onderwijs	847.892	18,7
Universitair hoger onderwijs	327.156	7,2
Totaal	4.538.246	100,0

Bron: EAK2007

Bemerk dat het totale aantal voor tabel 29 een beetje verschilt van dat van tabel 28. De EAK werkt met jaargemiddelden, terwijl de Rijksregisterinformatie geldt voor 1 januari van het betreffende jaar. Belangrijker dan dat kleine verschil is echter de procentuele verdeling, die gebruikt kan worden bij de poststratificatie. De laatste fase in de berekening van de gewichten voor de SCV-survey conformeert de voorlopig gewogen verdelingen volgens huishoudomvang en opleidingsniveau aan de populatieverdelingen. We hebben echter niet één gecombineerde verdeling, maar twee afzonderlijke verdelingen. Een mogelijke techniek is bijgevolg “raking”, ook wel iteratief proportioneel fitten of multiplicatief wegen genoemd. In essentie wordt daarbij telkens afwisselend herwogen op de twee variabelen tot beide verdelingen conform zijn. In SPSS kan raking toegepast worden met een Python-script dat kan afgehaald worden van de SPSS-website. Het resultaat is een gewicht waarvan de som voor de 1.434 respondenten uit Vlaanderen terug gelijk is aan 4.547.344. Bovendien worden de verdelingen van tabel 27 en tabel 29 perfect gereproduceerd.

Een terzijde

Door steekproeffluctuatie en door de gebruikte methode bij de non-responsweging (groepering volgens responskansen, logistisch regressiemodel niet 100% verzadigd) kunnen de volledige steekproef en de tot de steekproef “opgewogen” groep van respondenten ook nog (kleine) verschillen vertonen met de populatie voor de kenmerken leeftijd en geslacht. Die variabelen zouden dus eigenlijk ook opnieuw gebruikt kunnen worden in stap 3. Dat zou dan wel de vraag oproepen of die variabelen niet evengoed weggelaten kunnen worden uit stap 2.

Opname in stap 2 is zeker aangewezen. Het niet opnemen van relevante verklarende variabelen voor de respons in het logistische regressiemodel (zoals leeftijd en geslacht) zou immers leiden tot een verkeerde inschatting van het effect van andere variabelen (bijvoorbeeld type woning) op de respons en zo ook tot een niet correct gebruik ervan bij het bepalen van het non-responsgewicht. Het is dus best dat leeftijd en geslacht wel gebruikt worden in stap 2. Dat neemt echter niet weg dat ze ook in stap 3 nog zouden kunnen terugkomen. Maar de gecombineerde verdeling volgens leeftijd en geslacht na stap 2 is al vrijwel identiek aan de populatieverdeling. Een bijkomende weging volgens leeftijd en geslacht in stap 3 is bijgevolg overbodig.

Tabel 30 Verschillende stappen bij de berekening van de gewichten

Stap	N	Min	Max	Gem.	Som	UWE
Stap 1. Designgewicht	2.321	941,43	2.395,86	2.023,27	4.630.190,00	1,014
	1.475	941,43	2.395,86	2.016,40	2.974.187,79	1,011
Stap 2. Non-responsgewicht	1.475	1,25	2,19	1,57	2.314,18	1,040
Stap 1. + Stap 2. Designgewicht * non-responsgewicht	1.475	1.451,70	5.246,18	3.139,11	4.630.190,00	1,038
Stap 3a. Poststratificatie Brussel/Vlaams Gewest	1.475	1,00	1,07	1,00	1.476,08	1,000
Stap 1. + Stap 2. + Stap 3a. Designgewicht * non-responsgewicht * Poststratificatie Brussel/Vlaams Gewest	1.475	1.546,53	5.240,33	3.139,11	4.630.189,97	1,037
Stap 3b. Raking	1.434	2.427,43	6.156,49	3.171,09	4.547.344,39	1,058
Definitieve gewicht = Stap 3b. voor Vlaams Gewest = Stap 1. + Stap 2. + Stap 3a. voor Brussel	1.475	1.546,53	6.156,49	3.139,11	4.630.189,97	1,061
Definitieve gewicht herschaald	1.475	0,49	1,96	1,00	1475	1,061

Het definitieve gewicht voor de Vlaamse respondenten van de SCV-survey is dus datgene wat het resultaat is van de raking in stap 3. Aan de Brusselaars geven we het gewicht dat ook al gebruikt werd in tabel 25. De som van de gewichten voor alle 1.475 respondenten is dan opnieuw gelijk aan 4.630.190. In een allerlaatste stap kunnen we het gewicht nog herschalen zodat het gemiddelde ervan gelijk is aan 1 en de som gelijk is aan 1.475 (het ongewogen aantal respondenten). In tabel 30 kunnen alle verschillende stappen bij de berekening van de gewichten nog eens meegevolgd worden.

In stap 1 werd de steekproef herwogen naar de populatie. Maar omdat het gewicht uiteindelijk alleen toegekend wordt aan de respondenten is de som van de gewichten nog niet gelijk aan de omvang van de (doel)populatie. Vermenigvuldiging van het designgewicht met het non-responsgewicht maakt dat de rekensom wel klopt voor de groep van respondenten. De gewichten die berekend werden in stap 3 zijn neutraal wat de gemiddelde grootte betreft. De som na stap 3 blijft dus gelijk. Stap 3b. wordt alleen toegepast voor de respondenten die wonen in het Vlaamse Gewest. Dat zijn er 1.434 van de 1.475.

In de laatste stap herschalen we het gewicht, zodat de gewogen en de ongewogen steekproefomvang gelijk zijn en het gemiddelde gewicht gelijk is aan 1. Als je analyses uitvoert met de geëigende software (bijvoorbeeld COMPLEX SAMPLES binnen SPSS), maakt die herschaling niks uit. Maar voor de meeste gebruikers van surveybestanden zal dat herschaalde gewicht wat vertrouwder overkomen. Als de gewichten geen al te grote variatie kennen, zullen mogelijke fouten ook beperkt zijn bij default gebruik van het herschaalde gewicht. Maar eigenlijk geeft het toch een beetje een vals gevoel van veiligheid omdat bij gebruik van standaardsoftware ook het herschaalde gewicht een ongewenste impact kan hebben op significantietoetsen.

In de laatste kolom geven we met behulp van de UWE ook een inschatting van de impact van de gewichten op de schatting van de varianties. Die impact neemt globaal gezien toe naarmate we meer stappen zetten in de weegprocedure, maar blijft al bij al beperkt. Betrouwbaarheidsintervallen zouden zo'n 1,06 keer groter zijn door het gebruik van die gewichten in vergelijking met een enkelvoudig aselecte steekproef van dezelfde omvang. Het zijn stappen 2 en 3 (respectievelijk weging voor non-respons en postratificatie) die de voornaamste component uitmaken van dit UWE.

7.3. Berekening van de gewichten voor de ISSP-module

Naast de eigenlijke SCV-survey, is er ook nog de ISSP-module. Deze bijkomende vragenlijst wordt achtergelaten door de interviewer en de respondent wordt geacht om die vragenlijst in te vullen en terug te sturen. In 2008 stuurden 1.263 van de 1.475 respondenten (85,6%) deze vragenlijst terug. Voor dit ISSP-gedeelte berekenen we alleen een non-responsgewicht (stap 2 van de procedure) dat we zullen combineren met het definitieve gewicht voor de SCV-survey (de laatste 2 rijen van tabel 30). Een afzonderlijk designgewicht (stap 1) is overbodig want dat zit vervat in het SCV-gewicht. Bijkomende poststratificatie (stap 3) zou eventueel nog wel mogelijk zijn. Maar bij een voldoende gespecificeerd non-responsmodel, zou de poststratificatie van de weging van de eigenlijke SCV-survey moeten volstaan.

De berekening van het non-responsgewicht gebeurt net als in stap 2 van de berekening van het eigenlijke SCV-gewicht met een responsmodel. Daarvoor is er nu zeer veel informatie beschikbaar. Ook bij de 15% non-respondenten werd immers het face-to-face gedeelte afgenomen. In eerste instantie kijken we naar de klassieke achtergrondkenmerken: leeftijd, geslacht, huishoudtype, opleidingsniveau en het al dan niet hebben van betaald werk. Daar voegen we nog levensbeschouwing aan toe omdat religie het hoofdthema was van de ISSP. Bovendien zijn er nog een aantal surveygerelateerde variabelen waar we naar terug kunnen grijpen als verklarende variabelen. Zo is er de houding tegenover surveyonderzoek. In het mondelinge gedeelte zat immers een vraag met een reeks items over de geloofwaardigheid en bruikbaarheid van surveys. Die vraag bevatte 8 items. Vier items hiervan vormen min of meer een schaal (Cronbach's α = 0,70)[9]. Van deze 4 items werd het gemiddelde als onafhankelijke variabele opgenomen in de

[9] Het gaat om de items: de resultaten van dergelijke onderzoeken zijn bruikbaar om beleidsbeslissingen te nemen; de resultaten van dergelijke onderzoeken zijn steeds geloofwaardig; via dergelijk onderzoek kan men zijn mening kenbaar maken; met dergelijke onderzoeken krijgt de overheid een goed beeld van wat er leeft bij de bevolking.

logistische regressie die deelname aan de ISSP probeerde te verklaren. Maar ook de effecten van de andere items werden – afzonderlijk – getest. Het aantal ontbrekende waarden bij enkele batterijen met uitspraken werd ook opgenomen als een indicatie voor de mate van bereidwillige medewerking. Bij 12 batterijen werd telkens nagegaan of er voor minstens één item itemnon-respons was. Op basis hiervan werd een totale itemnon-respons indicator opgesteld die dus kon variëren van 0 tot 12. Tot slot is er het interviewerrapport. Bij elk afgenomen interview moest de interviewer enkele vragen beantwoorden over de motivatie van de respondent en zijn of haar weerstand bij het beantwoorden van enkele vragen. Natuurlijk speelt hier een subjectieve inschatting van de interviewer, maar die inschatting bleek ook vroeger al samen te hangen met de latere medewerking aan de bijkomende schriftelijke vragenlijst (Carton e.a., 2008, 87-93).

De resultaten van de logistische regressie met deelname aan de ISSP-module als afhankelijke variabele en bovengenoemde variabelen als onafhankelijke worden beschreven in tabel 31. De tabel behoudt alleen de variabelen met een significant effect.

Tabel 31 maakt duidelijk dat de jongste leeftijdscategorieën de laagste respons laten optekenen en ook mannen sturen de bijkomende enquête minder vaak terug dan vrouwen. Een interactie-effect leeftijd-geslacht werd er bij deze analyse niet gevonden. Verder wordt de hoogste respons opgetekend bij de houders van een diploma hoger onderwijs buiten de universiteit en de laagste respons bij de laagstgeschoolden. Mensen met betaald werk zijn ook minder geneigd om de vragenlijst in te dienen. De kleine schaal in verband met de attitude ten opzichte van surveys bleek geen effect te hebben op de ISSP-participatie evenmin als de meeste andere items van die batterij. Alleen blijken mensen die vinden dat ze betaald zouden moeten worden voor hun medewerking aan surveys en mensen die die vraag onbeantwoord laten, de vragenlijst minder vaak ingevuld terug te sturen. Door de variabele over de betaling als een categorische onafhankelijke te beschouwen, kan de groep met een ontbrekende waarde als een aparte groep in de analyse worden opgenomen. Ontbrekende antwoorden of “missing values” tijdens het mondelinge interview blijken trouwens mee het al dan niet deelnemen aan de ISSP te voorspellen. Het aantal ontbrekende waarden doorheen het face-to-face gedeelte heeft een effect op verdere participatie (hoe meer, hoe lager de kans). Ook met enkele variabelen uit het interviewerrapport is er een samenhang. De verschillende variabelen van dat rapport zijn wel sterk gecorreleerd en uiteindelijk werd de weerstand bij het beantwoorden van sommige vragen behouden. Hoe hoger die weerstand tijdens het mondelinge gedeelte, hoe lager de kans op terugsturen van de ISSP-vragenlijst. Door de opname van opleidingsniveau en twee surveygerelateerde variabelen is de MAR-assumptie in dit geval zeker aannemelijker dan bij de berekening van het SCV-gewicht. De situatie is vergelijkbaar met panelsurveys, waar er doorgaans ook meer informatie beschikbaar is om een gedetailleerd non-responsmodel te specificeren.

Tabel 31 Resultaten van de logistische regressie met ISSP-respons als afhankelijke variabele

	b	St. fout	p-waarde	e^b
Intercept	3,521	0,467	0,000	33,810
Leeftijdsklasse			0,000	
18-24j	-1,517	0,403	0,000	0,219
25-34j	-1,215	0,427	0,004	0,297
35-44j	-0,707	0,426	0,097	0,493
45-54j	-0,382	0,420	0,363	0,682
55-64j	-0,117	0,407	0,773	0,889
65-74j	0,382	0,438	0,382	1,466
Referentie: 75-85j	-	-	-	-
Geslacht			0,017	
Man	-0,382	0,159	0,017	0,683
Referentie: vrouw	-	-	-	-
Opleidingsniveau			0,025	
Geen/lager onderwijs	-0,714	0,375	0,057	0,490
Lager secundair onderwijs	-0,285	0,309	0,356	0,752
Hoger secundair onderwijs	0,089	0,282	0,752	1,093
Niet-universitair hoger onderwijs	0,286	0,303	0,344	1,331
Referentie: universitair onderwijs	-	-	-	-
Al dan niet betaald werk			0,027	
Ja	-0,502	0,227	0,027	0,605
Referentie: nee	-	-	-	-
Mensen zouden moeten betaald worden om mee te werken aan een dergelijk interview			0,005	
Referentie: (helemaal) oneens	-	-	-	-
Noch eens, noch oneens	0,255	0,246	0,300	1,291
(Helemaal) eens	-0,518	0,203	0,011	0,596
Geen antwoord	-2,016	0,954	0,035	0,133
Aantal ontbrekende waarden	-0,156	0,062	0,012	0,856
Interviewerrapport: weerstand bij het beantwoorden van sommige vragen (1: nooit; 5 : zeer veel)	-0,266	0,107	0,013	0,767

Bron: SCV-survey 2008

Voor het berekenen van de gewichten zijn de responskansen opnieuw belangrijker dan de interpretatie van de verschillende parameters. Ook in dit geval gebruiken we niet eenvoudigweg de inverse van de responskansen, maar groeperen we die kansen eerst in 5 gelijke groepen. Voor elk van die 5 groepen wordt nadien de proportie deelnemers aan de ISSP berekend en de inverse daarvan is het (bijkomende) ISSP-gewicht. Die berekening kan gevolgd worden in tabel 32.

Tabel 32 Berekening van de non-responsgewichten (na groepering van de kansen)

Groepering volgens responskans	Gewogen aantal (definitieve SCV-gewicht)	Gewogen percentage (definitieve SCV-gewicht)	Proportie respons op de ISSP	ISSP non-respons gewicht
1ste quintiel	925.743	20,0	0,689	1,453
2de quintiel	924.687	20,0	0,864	1,157
3de quintiel	919.320	19,9	0,859	1,165
4de quintiel	932.748	20,1	0,901	1,110
5de quintiel	927.693	20,0	0,956	1,046
Totaal	4.630.190	100,0	0,854	1,184

De groepering in quintielen in de eerste kolommen is gebaseerd op de logistische regressie. De voorlaatste kolom kijkt welk aandeel van de geselecteerde personen in het betreffende quintiel effectief de ISSP-vragenlijst heeft teruggestuurd. Het gewicht in de laatste kolom is eenvoudig de inverse van die proportie respons. Bemerk dat de proportie respons in het tweede quintiel een klein beetje hoger is dan deze in het derde quintiel, hoewel de voorspelde respons voor het tweede quintiel per definitie lager is. Beide liggen echter zeer dicht bij elkaar en de bijhorende gewichten zijn bijgevolg ook vrijwel gelijk.

Het gemiddelde ISSP-non-responsgewicht van deze groepsgewijze berekening bedraagt 1,184 met een standaardafwijking gelijk aan 0,140. Maar deze gewichten worden natuurlijk ook alleen toegepast op de uiteindelijke ISSP-respondenten. Voor die groep bedraagt het gemiddelde 1,170 met een standaarddeviatie van 0,131.

Het uiteindelijke ISSP-gewicht is dan de vermenigvuldiging van het ISSP-non-responsgewicht met het definitieve SCV-gewicht. Het gemiddelde ISSP-gewicht is zo gelijk aan 3.666,03 en de som van alle gewichten voor de 1.263 ISSP-respondenten is opnieuw gelijk aan de theoretische omvang van onze populatie: 4.630.190. Ook dit gewicht kunnen we herschalen zodat het gemiddelde gelijk is aan 1.

Tabel 33 toont de verschillende stappen bij het berekenen van dit ISSP-gewicht, net zoals tabel 30 deed voor het eigenlijke SCV-gewicht. De tabel toont dat ook hier de som van de ISSP-gewichten voor alle ISSP-respondenten gelijk is aan de theoretische omvang van de populatie.

Tabel 33 Verschillende stappen bij de berekening van de ISSP-gewichten

Stap	N	Min	Max	Gem.	Som
ISSP- non-responsgewicht	1.263	1,05	1,45	1,17	1.477,29
Definitieve ISSP-gewicht (= SCV-gewicht * ISSP-non-responsgewicht)	1.263	1.617,05	8.945,13	3.666,03	4.630.189,97
Definitieve ISSP-gewicht herschaald	1.263	0,44	2,44	1,00	1.263,00

Het UWE van het ISSP-non-responsgewicht is gelijk aan 1,012 wat resulteert in een totaal UWE voor het ISSP-gewicht gelijk aan 1,077. Dat is dus nog een beetje gestegen ten opzichte van het SCV-gewicht, maar blijft ook hier al bij al beperkt.

7.4. Een voorbeeldanalyse

Een eenvoudige analyse kan het effect van de gewichten illustreren. Bij dit voorbeeld kijken we naar het computergebruik. De vraag die aan de respondenten van de SCV-survey gesteld werd, luidde: "*Nu volgen enkele vragen over uw algemeen ICT-gebruik, waarmee we het gebruik bedoelen zowel thuis, op het werk, op school of elders. U kan de antwoordkaart gebruiken om te antwoorden. Wanneer gebruikte u voor het laatst een pc of een laptop?*". De respondenten moesten een antwoord kiezen uit vier alternatieven, gaande van *nooit* tot *gedurende de laatste drie maanden*, waarop er voor de laatste groep een aantal extra vragen volgden. Tabel 34 toont de ongewogen antwoordverdeling voor deze vraag.

Tabel 34 Computergebruik (***ongewogen***)

	Frequentie	Percentage
Nooit	371	25,2
Meer dan een jaar geleden	27	1,8
Tussen 3 maanden en een jaar geleden	16	1,1
Minder dan 3 maanden geleden	1.060	71,9
Weet niet	1	0,1
Totaal	1.475	100,0

Bron: SCV-survey 2008

Voor de eenvoud hebben we in dit voorbeeld het aantal categorieën teruggebracht tot twee. We zijn daarbij vooral geïnteresseerd in het aandeel Vlamingen dat nog nooit een computer heeft gebruikt. De verdeling van deze variabele met minder categorieën bevindt zich in tabel 35.

Tabel 35 Computergebruik (***ongewogen***)

	Frequentie	Percentage
Al wel computer gebruikt	1.103	74,8
Nog nooit computer gebruikt	371	25,2
Totaal	1.474	100,0

Bron: SCV-survey 2008

In tabel 36 bekijken we het percentage Vlamingen dat nog nooit een computer heeft gebruikt, maar ditmaal gewogen met de verschillende gewichten die in tabel 30 geïllustreerd werden. Uit de tabel blijkt duidelijk dat zowel het non-responsgewicht als de correctie voor de (geschatte) populatieverdelingen van huishoudomvang en (vooral) opleidingsniveau ervoor zorgen dat het geschatte aandeel Vlamingen zonder computerervaring stijgt. Dat is natuurlijk niet zo verwonderlijk aangezien ouderen (en daarvan voornamelijk de oudere vrouwen) en laaggeschoolden een hoger gewicht krijgen. Ter vergelijking hebben we in de tabel ook de resultaten meegegeven van de oude weegprocedure. Die beperkte zich tot een poststratificatie voor leeftijd, opleidingsniveau en geslacht. Het is niet verwonderlijk dat ook bij die weging het geschatte aandeel Vlamingen zonder computerervaring groter is dan de ongewogen frequentietabel doet vermoeden. Maar bij de uitgebreidere weegprocedure is het verschil tussen het ongewogen en het gewogen percentage nog groter. Ook bij de uitgebreide weegprocedure is het de poststratificatie die voor het belangrijkste verschil zorgt.

Tabel 36 Geschatte aandeel dat nog nooit een computer gebruikt heeft, **gewogen** met de verschillende gewichten (zie tabel 30)

Gewicht	Aandeel dat nog nooit een computer gebruikt heeft
Louter designgewicht	25,1
Gewicht na stap 2: designgewicht * non-responsgewicht	26,4
Gewicht na stap 3a: toevoeging poststratificatie Brussel	26,4
Gewicht na stap 3b: aanpassing gewicht aan marginale verdelingen van huishoudgrootte en opleidingsniveau (raking)	29,4
Oude weegprocedure ("eenvoudige" poststratificatie voor leeftijd, opleidingsniveau en geslacht)	28,1

We kijken ook naar eventuele verschillen die kunnen opduiken bij statistische toetsen. In dit geval gaan we na of het aandeel mensen zonder computerervaring even hoog is bij mannen als bij vrouwen. Tabel 37 toont dat meer vrouwen dan mannen nog nooit een computer gebruikt hebben.

Tabel 37 Aandeel Vlamingen dat nog nooit een computer gebruikt heeft, volgens geslacht (***ongewogen***)

Geslacht	Aandeel dat nog nooit een computer gebruikt heeft
Man	22,7
Vrouw	27,6
Totaal	25,2

Bron: SCV-survey 2008

Een eenvoudige chikwadraattoets op deze data zou ons leren dat het verschil significant is (p = 0,03076).

In tabel 38 maken we dezelfde vergelijking volgens geslacht, maar deze keer gewogen met het definitieve gewicht. Het verschil tussen mannen en vrouwen wordt groter en bijgevolg wordt ook de p-waarde voor de toets van gelijkheid volgens geslacht kleiner.

Tabel 38 Kruistabel van geslacht en mate van instemming met het optrekken van de pensioenleeftijd (**g*ewogen met definitieve SCV-gewichten***)

Geslacht	Aandeel dat nog nooit een computer gebruikt heeft
Man	26,3
Vrouw	32,4
Totaal	29,4
p-waarde, berekend met COMPLEX SAMPLES	0,00675

Bron: SCV-survey 2008

Bij deze toets zou de conclusie dezelfde zijn: meer vrouwen dan mannen hebben nog nooit een computer gebruikt. Maar de bekomen percentages zijn toch zeer verschillend en het verschil in p-waarde bij de toets is ook opvallend. Bij een toets waarvoor de bekomen p-waarde schommelt rond de universeel gebruikte 0,05, kan het al dan niet gebruik van gewichten bij die toets dus ook leiden tot verschillende conclusies. Het gebruik van geëigende software is bijgevolg noodzakelijk om correcte inferentiële besluiten te kunnen trekken.

8. Illustratie 2 – de survey van de stadsmonitor

8.1. Beschrijving van de survey

Het tweede voorbeeld in dit rapport betreft de survey van de Stadsmonitor 2008. In opdracht van het Vlaamse Stedenbeleid werd in 2004 een eerste editie van de "Stadsmonitor voor leefbare en duurzame Vlaamse steden" ontwikkeld. Deze monitor telt een 200-tal indicatoren voor de 13 Vlaamse centrumsteden, waarvan er een aanzienlijk aantal ingevuld wordt met behulp van een survey bij de inwoners van die steden. Alle inwoners ouder dan 16 jaar, ongeacht de nationaliteit, vormen het steekproefkader. Deze stadsmonitor werd een eerste maal geactualiseerd in 2006 en opnieuw in 2008. In 2004 en 2006 werd de uitvoering van de survey volledig uitbesteed. Voor beide edities werd er toen geopteerd voor een telefonische enquête. Ter wille van een aantal redenen (onder andere een dalende en selectievere respons, zie Schelfaut, 2009) werd er in 2008 gekozen voor een postenquête, waarvan de regie volledig opgenomen werd door de Studiedienst van de Vlaamse Regering. De steekproefomvang in de 13 steden werd in 2004 bepaald buiten de SVR om. In grotere steden werd er toen geopteerd voor een grotere steekproefomvang, maar de verschillen waren niet evenredig aan de werkelijke populatieverschillen. Bij de twee volgende edities werd er voort gebouwd op de initiële steekproefaantallen. Alleen werd in 2008 aan de steden de mogelijkheid geboden om de steekproefomvang in hun eigen stad op te trekken om zo intrastedelijke vergelijkingen mogelijk te maken. Antwerpen en Turnhout maakten gebruik van die mogelijkheid met de bedoeling om districten, respectievelijk stadsdelen met elkaar te kunnen vergelijken. Voor die steden werd de steekproefomvang dus verhoogd en werd er met het oog op de vergelijking een gelijke steekproefomvang per district respectievelijk stadsdeel vooropgesteld. Zo komen we tot de verdeling van de brutosteekproef over de steden (tabel 39) en voor Antwerpen over de districten (tabel 40) en voor Turnhout over de stadsdelen (tabel 41).

Tabel 39 Verdeling over de steden

	Steekproef		Populatie	
Gemeente	Aantal	Percentage	Aantal	Percentage
Aalst	1.500	4,7	65.747	5,2
Antwerpen	9.700	30,6	389.142	30,7
Brugge	1.750	5,5	98.420	7,8
Genk	1.350	4,3	52.324	4,1
Gent	2.300	7,3	196.847	15,5
Hasselt	1.450	4,6	61.294	4,8
Kortrijk	1.500	4,7	62.023	4,9
Leuven	1.649	5,2	78.178	6,2
Mechelen	1.500	4,7	64.487	5,1
Oostende	1.450	4,6	59.561	4,7
Roeselare	1.350	4,3	46.768	3,7
Sint-Niklaas	1.450	4,6	58.041	4,6
Turnhout	4.700	14,9	33.939	2,7
Totaal	31.649	100,0	1.266.771	100,0

Bron: Survey stadsmonitor 2008

Tabel 40 Verdeling over de Antwerpse districten

	Steekproef		Populatie	
District	Aantal	Percentage	Aantal	Percentage
Antwerpen	1.080	11,1	141.556	36,4
Berchem	1.078	11,1	34.435	8,8
BEZALI[10]	1.074	11,1	7.799	2,0
Borgerhout	1.078	11,1	33.484	8,6
Deurne	1.078	11,1	58.803	15,1
Ekeren	1.078	11,1	18.273	4,7
Hoboken	1.078	11,1	28.071	7,2
Merksem	1.078	11,1	34.425	8,8
Wilrijk	1.078	11,1	32.296	8,3
Totaal	9.700	100,0	389.142	100,0

Bron: Survey stadsmonitor 2008

Tabel 41 Verdeling over de Turnhoutse stadsdelen

	Steekproef		Populatie	
Stadsdeel	Aantal	Percentage	Aantal	Percentage
Blijkhoef	717	15,3	3.017	8,9
Centrum	717	15,3	11.461	33,8
Schorvoort	717	15,3	2.490	7,3
Stadsbos en Noorden	398	8,5	942	2,8
Stedelijk wonen Oost	717	15,3	6.916	20,4
Stedelijk wonen West	717	15,3	7.455	22,0
Zevendonk	717	15,3	1.658	4,9
Totaal	4.700	100,0	33.939	100,0

Bron: Survey stadsmonitor 2008

Zoals uit tabel 41 blijkt, werd er voor één stadsdeel van Turnhout afgeweken van de regel van gelijke steekproefomvang. In Stadsbos en Noorden wonen zo weinig mensen dat een kleinere steekproef dankzij de eindigheidscorrectie of finiteitscorrectie ook al vergelijkbare betrouwbaarheidsintervallen zal opleveren. Met de finiteitscorrectie kan de steekproeffractie in rekening gebracht worden bij de berekening bijvoorbeeld standaardfouten, naast de steekproefomvang. Bij grotere aandelen, zoals bijvoorbeeld in Stadsbos en Noorden maakt dat een verschil uit.

Voor elke stad afzonderlijk werd er een steekproefplan opgezet. Dat plan ging uit van twee principes: (1) expliciete stratificatie volgens deelgemeente/wijk/district en (2) impliciete stratificatie volgens nationaliteit (Belg/niet-Belg), leeftijd en geslacht. Deze dubbele stratificatie had niet tot doel om resultaten te geven per deelgemeente, nationaliteitsgroep,... Dat is vaak ook niet mogelijk omdat de groepen te klein zijn. Maar de dubbele stratificatie moest er wel voor zorgen dat de bevolking van de volledige stad zo goed mogelijk weerspiegeld werd in de steekproef.

Voor de ***expliciete*** stratificatie werden de steden opgedeeld in een aantal deelgebieden (op basis van deelgemeente, wijk, district of wat dan ook). Het aantal respondenten per deelgebied werd dan bepaald in functie van het aantal inwoners van het deelgebied in verhouding tot het totaal aantal inwoners van de centrumstad, een proportioneel gestratificeerde steekproef dus.

[10] Berendrecht, Zandvliet en Lillo

Ook bij de trekking van respondenten binnen een stad of deelgemeente werd het toeval een beetje geholpen. Hierbij werd gebruik gemaakt van de variabelen nationaliteit, geslacht en leeftijd. Om de populatie niet vooraf te moeten indelen volgens deze kenmerken (een indeling die heel verscheiden resultaten zou opleveren voor de verschillende centrumsteden), werd er hierbij gekozen voor ***impliciete*** stratificatie. Impliciete stratificatie kan toegepast worden door de populatielijst te ordenen volgens de relevante kenmerken en systematisch eenheden te trekken. Concreet werd een lijst gemaakt met eerst alle Belgen en dan alle niet-Belgen. Binnen de nationaliteitscategorieën werden dan alle mannen van jong naar oud gerangschikt en daarna alle vrouwen ook van jong naar oud. Afhankelijk van de steekproefomvang werd daarop elke 'zoveelste' persoon getrokken zodanig dat de hele lijst doorlopen werd. Zulke impliciete stratificatie geeft vergelijkbare voordelen als expliciete stratificatie (een betere weerspiegeling van de populatie van de stad) en het is eenvoudiger om de procedure constant te houden in alle steden. In de praktijk werd de impliciete stratificatie deelgebied per deelgebied uitgevoerd.

De dubbele stratificatiestrategie resulteerde steeds in een getrokken steekproef die de populatie van de centrumstad volledig weerspiegelde voor de kenmerken nationaliteit, leeftijd en geslacht. Voor Antwerpen en Turnhout geldt die weerspiegeling op districtsniveau, respectievelijk stadsdeelniveau.
De uiteindelijke steekproeftrekking gebeurde in de maand maart door SVR in de steden zelf. De steden stelden een extractie uit het bevolkingsregister ter beschikking die een steekproeftrekking door een SVR-medewerker toeliet. Deze extractie was steeds recent (maximaal een week oud) en de aanwezigheid van de SVR-medewerker garandeerde een uniforme steekproeftrekking voor alle steden.

Deze werkwijze liet ook toe om wat informatie voor het volledige bestand te verzamelen (leeftijdsverdeling, geslachtsverdeling en soms nog enkele bijkomende variabelen). Zo is er voor vrijwel alle steden nog meer informatie beschikbaar om de representativiteit van de steekproef te beoordelen dan louter de gekende statistieken.

8.2. Berekening van de gewichten

Ook voor de berekening van de gewichten van de survey van de stadsmonitor werden de drie in sectie 4 vermelde stappen overwogen. Uiteindelijk werden enkel de eerste twee uitgevoerd.

Stap 1 Basisgewichten (design weights)
Stap 1 vormt geen enkel probleem. Van elke stad is gekend hoe groot de populatie is waaruit de steekproef getrokken werd. Die informatie kan je ook halen uit de steekproefrapporten. Zo telde de doelpopulatie in Aalst 65.747 personen. Daarvan werden er 1.500 geselecteerd. De eerste component van het gewicht voor Aalst is de inverse van de selectiekans, dus:

$$\frac{65.747}{1.500}$$

Voor 10 andere steden gebeurt deze berekening op exact dezelfde manier. Voor Antwerpen en Turnhout gebeurt de berekening district per district, respectievelijk stadsdeel per stadsdeel.

Stap 2 Compensatie voor non-respons
Er was wel enige informatie beschikbaar over de steekproef. Voor alle steden hadden we van alle geselecteerde personen: geslacht, leeftijd, nationaliteit. In sommige steden was ook gezinsgrootte of burgerlijke staat beschikbaar, maar we opteerden ervoor om ook deze stap voor alle steden uniform uit te voeren. Informatie die niet voor iedereen beschikbaar was, lieten we dus vallen.

De beschikbare informatie kon het makkelijkst gebruikt worden in een logistisch regressiemodel. Dat model vertrekt van een bestand met alle geselecteerde personen en de variabele respons (nee/ja – 0/1). Dat is de afhankelijke variabele van de logistische regressie. Onafhankelijke variabelen zijn dan geslacht, leeftijd, nationaliteit (Belg/niet-Belg) en eventuele interacties tussen die variabelen. Met die logistische regressie kunnen we responskansen voor iedere eenheid in het bestand berekenen, de gewichten van deze stap 2 zijn de inverse van die kansen, waarbij we de kansen al dan niet groeperen.

De gemiddelde responskans in Aalst was gelijk aan 46,6% (699 antwoorden op 1.500 verstuurde enquêtes). De gewichten van stap 2 zullen dus gemiddeld gelijk zijn aan 2,15.

$$\frac{1}{\Pr(\text{respons} = 1)}$$

$$\frac{699}{1.500} = 0{,}466$$

$$\frac{1}{0{,}466} = 2{,}15$$

Afhankelijk van de leeftijd, het geslacht en/of de nationaliteit van de geselecteerde persoon zullen de responskansen groter of kleiner zijn.

Deze stap zorgde uiteindelijk voor enkele problemen.

Een eerste probleem deed zich voor met de variabele **nationaliteit** (Belg/niet-Belg). Die variabele heeft duidelijk een impact op de kans om al dan niet deel te nemen aan de survey.
Van de 31.649 geselecteerde personen hadden er 2.676 niet de Belgische nationaliteit. Daarvan hebben er 754 deelgenomen aan het interview. Dat is een respons van 28,2%. Bij Belgen stuurden 14.441 van de 28.973 geselecteerde mensen de vragenlijst terug in (49,8%). Een duidelijk verschil dat ook multivariaat (onder controle van leeftijd en geslacht) standhoudt. Maar het verschil was zo manifest dat de gewichten voor de niet-Belgen zeer veel groter waren dan voor de Belgen. Opname van nationaliteit in de logistische regressie leidde bijna altijd tot extreme gewichten. Daar komt bij dat het aantal niet-Belgen in sommige steden zeer klein is. Voor Aalst zitten er bijvoorbeeld slechts 10 niet-Belgen in de steekproef. Dat is dan ook nog eens een zeer heterogene groep (9 verschillende nationaliteiten). Die 10 personen dergelijk groot gewicht toekennen, leek niet zo'n goed idee.

Deze situatie (extreem hoge gewichten die aan een te kleine heterogene groep toegekend zouden moeten worden) deed zich in min of meerdere mate voor in elke stad. Uiteindelijk hebben we er daarom voor geopteerd om gewoon toe te geven dat de niet-Belgen niet goed vertegenwoordigd zijn in onze steekproef en dat dat niet op te lossen valt met een weging. Meer geavanceerde technieken waarbij restricties opgelegd kunnen worden aan de gewichten, werden hier buiten beschouwing gelaten.

Zo bleven dus alleen de variabelen leeftijd en geslacht over. Leeftijd hebben we in 7 categorieën opgedeeld. De logistische regressiemodellen bevatten verder ook telkens de interactie leeftijd-geslacht. Dat leidde dan eigenlijk tot een "fully saturated model", dat overeenkomt met celweging op de gecombineerde verdeling van leeftijd en geslacht. Die weging gebeurde dus ook stad per stad en voor Antwerpen en Turnhout district per district, respectievelijk stadsdeel per stadsdeel. Het aantal groepen (en dus ook berekende kansen) bleef zo al bij al beperkt (14) en het leek niet nodig om te groeperen. Er waren immers vrijwel geen extreme gewichten.

Twee stadsdelen in Turnhout vormden hierop een uitzondering. Met de gevolgde strategie waren er in die stadsdelen toch enkele extreme gewichten. Door voor één stadsdeel de drie jongste leeftijdscategorieën te herleiden tot twee en voor het andere stadsdeel de leeftijdsgrenzen een beetje op te schuiven konden de extreme gewichten eenvoudig gematigd worden. Het is natuurlijk niet toevallig dat deze problemen zich voordeden in Turnhout. In Turnhout is de steekproefomvang per stadsdeel relatief klein (de omvang voor de stad als geheel is wel groter). Dat zou een argument kunnen zijn om de weging niet per stadsdeel uit te voeren, maar in één keer voor de stad als geheel. Maar omdat de kleine wijzigingen van hierboven volstonden om de problemen op te lossen, hebben we ervoor geopteerd om de compensatie voor non-respons ook stadsdeel per stadsdeel te berekenen, het niveau waarop ook het steekproefplan opgesteld was.

Voor de volledigheid toont tabel 42 het non-responsmodel (de logistische regressie) van de stad Aalst.

Tabel 42 Resultaten van de logistische regressie met respons als afhankelijke variabele voor de stad Aalst

	b	Stand. fout	p-waarde	e^b
Intercept	-0,299	0,201	0,137	0,741
Leeftijd			0,000	
76 jaar en ouder	-0,508	0,301	0,091	0,602
66 tot 75 jaar	0,171	0,288	0,552	1,187
56 tot 65 jaar	0,795	0,286	0,005	2,213
46 tot 55 jaar	0,786	0,269	0,003	2,195
36 tot 45 jaar	0,109	0,264	0,681	1,115
26 tot 35 jaar	0,012	0,277	0,967	1,012
Referentie: 16 tot 25 jaar	-	-	-	-
Geslacht			0,340	
Man	-0,276	0,290	0,340	0,759
Referentie: vrouw	-	-	-	-
Geslacht * leeftijd			0,006	
Man * 76 jaar en ouder	1,393	0,461	0,003	4,028
Man * 66 tot 75 jaar	0,558	0,422	0,186	1,747
Man * 56 tot 65 jaar	-0,160	0,406	0,694	0,852
Man * 46 tot 55 jaar	-0,211	0,380	0,579	0,810
Man * 36 tot 45 jaar	0,085	0,379	0,822	1,089
Man * 26 tot 35 jaar	-0,078	0,398	0,845	0,925
Referentie: vrouw, 16 tot 25 jaar	-	-	-	-

Bron: survey stadsmonitor 2008

De analyse in tabel 42 maakt duidelijk dat er “maar” 14 verschillende gewichten zullen zijn, voor elke leeftijdscategorie één voor mannen en één voor vrouwen.

Stap 3 Een vorm van poststratificatie

De populatie van de survey voor de stadsmonitor is behoorlijk specifiek: 16+, wonende in één van de 13 centrumsteden. Voor deze populatie zijn er niet veel populatiegegevens direct beschikbaar. Eenvoudig beschikbaar zijn leeftijd, geslacht en nationaliteit. Maar voor leeftijd en geslacht hebben we al gewogen in stap 2 en omdat we voor die kenmerken (impliciet) gestratificeerd hadden bij de steekproeftrekking, is onze gewogen verdeling (na stap 2) al vrijwel 100% gelijk aan de gekende populatieverdeling. Voor nationaliteit struikelen we ongeveer meteen op het probleem dat ook in stap 2 werd aangehaald: extreme gewichten die aan kleine groepen zouden toegekend moeten worden. Er werd dus besloten om niet te poststratificeren voor deze 3 kenmerken.

Eén andere variabele zou eventueel nog in aanmerking komen: huishoudgrootte. In de survey werd ook gevraagd naar het aantal leden in het huishouden. Uit de bevolkingsregisters kan het aantal personen volgens het aantal gezinsleden worden afgeleid voor de specifieke populatie van de survey van de stadsmonitor, zelfs op het niveau van de statistische sector, en dus voor Antwerpen en Turnhout ook op districts/stadsdeelniveau. Maar de instructie bij de betreffende vraag in de survey was onvoldoende duidelijk, want de antwoorden op die vraag stroken zeer vaak niet met de antwoorden op de voorgaande vraag, die een omschrijving van de gezinssituatie betreft. Er waren zoveel onverenigbaarheden, dat een poststratificatie voor huishoudgrootte te gevaarlijk leek. Omdat er geen andere variabelen overbleven om poststratificatie uit te voeren, beperkt de weging zich dus tot de twee eerste stappen.

De gewichten van stap 1 en stap 2 werden eenvoudig gecombineerd door ze te vermenigvuldigen. Het definitieve gewicht maakt dat de gewogen steekproef dan representatief is voor de kenmerken leeftijd en geslacht in alle steden (en voor Antwerpen en Turnhout in de districten/stadsdelen) en ook voor alle steden samen. Ook de globale gecombineerde verdeling van beide kenmerken wordt gereproduceerd, net als de verdeling over de steden. De som van de gewichten voor alle 15.195 respondenten is gelijk aan 1.266.771, de populatieomvang. De gewogen omvang van het databestand is dus gelijk aan de doelpopulatie (alle inwoners van de 13 centrumsteden die aan de leeftijdsvereisten voldoen).

Om met deze gewichten zinvolle standaardfouten en statistische testen te krijgen is het natuurlijk ook noodzakelijk te werken met correctere schattingsmethoden, bijvoorbeeld de linearisatiemethode die vervat zit in de SPSS-module Complex Samples.

Zoals aangetoond met het fictieve voorbeeld in sectie 6 en met de voorbeeldanalyse van de SCV-survey in sectie 7, lost het herschalen van het gewicht de ongewenste effecten van de gewichten bij een verkeerd gebruik (bijvoorbeeld de defaultwijze in SPSS) niet altijd op. In dit geval zal een standaardgebruik van een herschaald gewicht zelfs contraproductief zijn. Het design was namelijk zo opgebouwd dat steden konden vergeleken worden of, voor Antwerpen en Turnhout, dat districten, respectievelijk stadsdelen vergeleken konden worden. Een weging in combinatie met een herschaling van dat gewicht en standaardgebruik daarvan doet eigenlijk het design teniet. In dat geval krijgen bijvoorbeeld in Turnhout kleinere stadsdelen een kleinere gewogen steekproefomvang, waardoor zij minder vaak (significant) zullen verschillen van andere stadsdelen. Zo'n herschaalde weging zal dus het statistische onderscheidingsvermogen bij die vergelijking verlagen en dat kan niet de bedoeling zijn. Een bijkomend probleem is dat verschillende herschalingen noodzakelijk kunnen zijn als slechts delen van het bestand geanalyseerd worden. Voor sommige analyses zullen bijvoorbeeld de twee grootste centrumsteden (Antwerpen en Gent) buiten beschouwing gelaten worden omdat die qua omvang toch te veel verschillen van de andere. Een analyse zonder Antwerpen en Gent impliceert een andere herschaling met mogelijk andere inferentiële conclusies als gevolg. Het voorbeeld dat we hiervan tonen in bijlage illustreert dat probleem.

9. Conclusie en discussie

In deze tekst werd de berekening en het gebruik van gewichten voor surveydata beschreven. Door gewichten te gebruiken proberen onderzoekers mogelijke vertekening van surveyresultaten tegen te gaan. Maar, hoewel het (her)wegen van databestanden ruim is ingeburgerd, is het vaak onduidelijk hoe die gewichten best berekend worden en hoe ze gebruikt moeten worden bij de analyse. Een Amerikaans auteur concludeerde naar aanleiding van vergelijkbare bevindingen: "Survey weighting is a mess!" (Gelman, 2007).

Deze tekst probeerde alvast duidelijk te maken dat het aantal mogelijke manieren om gewichten te berekenen veel uitgebreider is en veel verder gaat dan louter "poststratificeren via celweging". Courante aanbevelingen schrijven bovendien voor om verscheidene stappen te volgen bij het berekenen van de gewichten. Poststratificatie, al dan niet via celweging, is daarbij slechts de laatste stap.

Een belangrijke kanttekening en relativering bij de verschillende weegmethoden is dat de gebruikte methode vaak veel minder belangrijk is dan de gebruikte variabelen (Kalton & Flores-Cervantes, 2003). De keuze van de variabelen die gebruikt worden bij de weging hebben een grotere impact dan de keuze van de methode bij het gebruik ervan. Het al dan niet groeperen van de responskansen bij de logistische regressie met respons als afhankelijke variabele, zal bijvoorbeeld een kleinere impact hebben dan de keuze van de onafhankelijke variabelen voor die logistische regressie.

De keuze van de variabelen die gebruikt worden bij de weging is cruciaal. Vaak zijn alleen de traditionele demografische variabelen beschikbaar, maar het gebruik daarvan is waarschijnlijk onvoldoende om de vertekening te doen verdwijnen. Bovendien kunnen de ideale weegvariabelen verschillen naargelang de variabelen waarin de onderzoeker geïnteresseerd is, zoals ook het non-responsmechanisme anders kan zijn voor onderscheiden variabelen of analyses. Nochtans is het idee van één gewicht juist zeer aanlokkelijk en vanuit consistentieoverwegingen vaak ook wenselijk (Lohr, 2007, 176).

Meer en betere weegvariabelen moeten dus de doelstelling zijn. Bij de recentere edities van de SCV-survey hebben we alvast bij het Rijksregister meer informatie opgevraagd over de steekproefeenheden (bijvoorbeeld burgerlijke staat, al dan niet samenwonen met een partner, huishoudomvang,…). Een beetje creativiteit bij de zoektocht naar de geschikte variabelen is ook aangewezen. Zo gebruikt het CBS in Nederland steevast het al dan niet bezitten van een telefoon met gekend nummer. Die variabele zou goede resultaten laten optekenen bij het terugdringen van de vertekening (Bethlehem en Schouten, 2004). Zo bleek bijvoorbeeld de schatting van aandeel

uitkeringstrekkers minder vertekend als er gewogen werd voor onder andere het bezit van een telefoon met gekend nummer. Dan kon eenvoudig nagegaan worden omdat het al dan niet trekken van een uitkering ook gekend was voor de non-respondenten. Telefoonbezit kan ook voor Belgische huishoudens nagegaan worden, ook al bleek een koppeling van adresgegevens aan telefoongegevens voor een experimenteel onderzoeksopzet niet 100% waterdicht (De Waele e.a., 2008, 50-51).

In deze tekst illustreerden we het berekenen van gewichten met twee voorbeelden. In die voorbeelden hanteerden we een verschillende weegstrategie. Zo toonden we dat er niet zoiets bestaat als één uniforme berekening van gewichten die altijd gevolgd moet worden. Het steekproefplan, de beschikbaarheid van data, zowel van de steekproef als van de populatie,... er zijn heel wat variabelen die bepalen hoe de gewichten best berekend worden. Zoals steeds is een duidelijke en volledige documentatie het eerste kwaliteitskeurmerk.

Het succes van de weging hangt af van de plausibiliteit van het (respons)model. Maar het in vraag stellen van dat model, kan geen aanleiding zijn om het wegen zomaar volledig overboord te gooien. Een eenvoudige veralgemening naar een populatie op basis van een survey met non-respons zonder enige vorm van weging of zonder toepassing van een andere techniek voor ontbrekende waarden gaat eigenlijk uit van de MCAR-assumptie en een perfecte enkelvoudige aselecte toevalssteekproef. Dat is nog een veel strengere assumptie dan deze van het responsmodel. Wegen zal niet alle problemen van vertekening oplossen, maar niet wegen doet dat zeker niet.

De tekst toont ook dat voorzichtigheid geboden is bij gewogen analyses. De manier waarop vele softwarepakketten en met name SPSS, default omgaan met gewichten voldoet niet. Er is geen probleem met bijvoorbeeld de schattingen van gemiddelden, regressiecoëfficiënten,... maar wel met de bijbehorende standaardfouten. Bijgevolg zijn de significantietoetsen niet correct. Er zijn op dit moment echter voldoende alternatieven beschikbaar om met gewogen datasets wel correcte inferentiële resultaten te bekomen. Alle “grote” statistische softwareprogramma’s zoals bijvoorbeeld SAS, SPSS en STATA bieden tegenwoordig rekenmodules aan die kunnen corrigeren voor het feit dat de data niet afkomstig zijn van een enkelvoudige toevalssteekproef. Wanneer de geëigende procedures gebruikt worden, blijken de verschillen tussen die softwareprogramma’s klein tot onbestaande (Siller & Tompkins, 2006). Maar de verschillen tussen analyses die corrigeren voor het steekproefdesign en de gewichten en analyses die dat niet doen, zijn wel vaak substantieel (Tibaldi e.a. 2003). Gewichten zelf normeren om de specifieke rekenmethoden te vermijden, vormt geen afdoende oplossing. Het is daarom op zijn minst ongelukkig dat dit in het geval van SPSS de aankoop van een bijkomende (dure) module (Complex Samples) impliceert.

Literatuur

Agresti, A., Booth, J.G., Hobbert, J.P. & Caffo, B. (2000). Random-Effects Modeling of Categorical Response Data. In: *Sociological Methodology*, 30, 27-80.

APS (2003). *Kwaliteitszorg Statistisch Productieproces. Aanbevelingen.* Brussel: Ministerie van de Vlaamse Gemeenschap.

Battaglia, M.P., Izrael, D., Hoaglin, D.C., & Frankel, M.R. (2004). *Tips and Tricks for Raking Survey Data (a.k.a. Sample Balancing)*. Paper gepresenteerd op de jaarlijkse meeting van AAPOR, Phoenix, Arizona.

Bethlehem, J.G. (2002). Weighting Non-response Adjustments Based on Auxiliary Information. In: Groves, R.M., Dillman, D.A., Eltinge, J.L. & Little, R.J. (red.) *Survey Non-response*. New York: John Wiley, 275-288.

Bethlehem, J.G. (2008). *Wegen als correctie voor non-respons. Statistische Methoden (08005)*. Voorburg/Heerlen: Centraal Bureau voor de Statistiek.

Bethlehem, J.G. & Keller, W.J. (1987). Linear Weighting of Sample Survey Data. In: *Journal of Official Statistics*, 3 (2), 141-153.

Bethlehem, J.G. & Schouten, B. (2004). *Non-response Adjustment in Household Surveys*. Den Haag: Centraal Bureau voor de Statistiek.

Biemer, P.P. & Christ, S.L. (2008). Weighting Survey Data. In: Hox, J., de Leeuw, E. & Dillman, D.A. (red.) *The International Handbook of Survey Methodology*, New York: Lawrence Erlbaum Associates, 317-341.

Biemer, P.P. & Lyberg, L.E. (2003). *Introduction to Survey Quality*. New York: John Wiley.

Billiet, J. (2007). Het belang van regelmatig onderzoek naar opinies en houdingen in de bevolking. In: Pickery, J. (red.). *Vlaanderen gepeild! SVR-Studie 2007/2*. Brussel: Studiedienst van de Vlaamse Regering, 7-36.

Callens, M. (2010). *Contextuele regressiemethoden voor internationaal vergelijkend onderzoek SVR-Methoden en Technieken 2010/2*. Brussel: Studiedienst van de Vlaamse Regering.

Carton, A., Vander Molen, T. & Pickery, J. (2008). *Sociaal-culturele verschuivingen in Vlaanderen 2007. Basisdocumentatie. SVR-Technisch rapport 2008/3*. Brussel: Studiedienst van de Vlaamse Regering.

Deming, W.E. & Stephan, F.F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. In: *Annals of Mathematical Statistics*, 11 (4), 427-444.

De Waele, M., Heerwegh, D. & Loosveldt, G. (2008). *NOTESUMO: Non-response to a Telephone Survey such as the Security Monitor. Deel II : Evaluatie van een mixed mode survey design*. Leuven: Centrum voor Surveymethodologie / Katholieke Universiteit Leuven.

Dillman, D.A., Eltinge, J.L., Groves, R.M. & Little, R.J. (2002). Survey Response in Design, Data Collection, and Analysis. In: Groves R.M., Dillman D.A., Eltinge, J.L., Little R.J. (red.) *Survey Non-response*. New York: John Wiley, 3-26.

Gelman, A. (2007) Struggles with Survey Weighting and Regression Modeling. In: *Statistical Science*, 2007, 22 (2), 153–164.

Groves, R.M. (2006). Non-response Rates and Non-response Bias in Household Surveys. *Public Opinion Quarterly*, 70(5), 646-675.

Höfler, M., Pfister, H., Lieb, R. & Wittchen, H-U. (2005). The use of weights to account for non-response and drop-out. In: *Social Psychiatry and Psychiatric Epidemiology*, 40 (4), 291-299.

Hosmer, D. & S. Lemeshow (2000). *Applied Logistic Regression, 2nd Edition*. New York: John Wiley.

Kalton, G. & Flores-Cervantes, I. (2003). Weighting Methods. In: *Journal of Official Statistics*, 19 (2), 81-97.

Kish, L. (1992). Weighting for Unequal Pi. In: *Journal of Official Statistics*, 8 (2), 183-200.

Little, R.J. & Rubin, D.B. (2002). *Statistical Analysis with Missing Data, 2nd edition*. New York: John Wiley.

Little, R.J. & Vartivarian, S. (2005). Does Weighting for Non-response Increase the Variance of Survey Means? In: *Survey Methodology*, 31 (2), 161-168.

Lohr, S.L. (1999). *Sampling: Design and Analysis*. Pacific Grove: Duxburry Press.

Lohr, S.L. (2007). Comment: Struggles with Survey Weighting and Regression Modeling. In: *Statistical Science*, 2007, 22 (2), 175–178.

Molenberghs, G. (2009). *Survey Methods & Sampling Techniques*. Ongepubliceerde transparanten van de lessen van het gelijknamige college in het kader van het programma Master in Quantitative Methods aan de HUB.

Molenberghs, G. & Kenward, M.G. (2007). *Missing Data in Clinical Studies*. Chichester: John Wiley.

Pickery, J. (2008). *De interpretatie van interactie-effecten in regressiemodellen. SVR-Technisch rapport 2008/1*. Brussel: Studiedienst van de Vlaamse Regering.

Pickery, J. & Carton, A. (2008). Oversampling in Relation to Differential Regional Response Rates, In: *Survey Research Methods*, 2 (2), 83-92.

Potter, F. (1990). A study of Procedures to Identify and Trim Extreme Sampling Weights, *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 225-230.

Rice, N. (2001). Binomial Regression. In: Leyland, A.H. & Goldstein, H. (red.) *Multilevel Modelling of Health Statistics*. New York: John Wiley, 27-44.

Rodgers-Farmer, A.Y. & Davis, D. (2001). Analysing Complex Survey Data. In: *Social Work Research*, 25 (3), 185-192.

Rust, K. (1985). Variance Estimation for Complex Estimators in Sample Surveys. In: *Journal of Official Statistics*, 1 (4), 381-397.

Schelfaut, H. (2009). *Survey Stadsmonitor "Thuis in de stad 2008". Methodologisch Rapport. SVR-Technisch rapport 2009/1*. Brussel: Studiedienst van de Vlaamse Regering.

Schouten, B. (2004). *Adjustment for Bias in the Integrated Survey on Household Living Conditions (POLS) 1998. Discussion paper 04001*. Voorburg/Heerlen: Centraal Bureau voor de Statistiek.

Schouten, B., Cobben, F. & Bethlehem, J. (2009). Indicators for the representativeness of survey response. In: *Survey Methodology*, 35 (1), 101-113.

Siller, A.B. & Tompkins, L. (2006). *The Big Four: Analyzing Complex Sample Survey Data. Using SAS, SPSS, STATA and SUDAAN*. Poster gepresenteerd op het SAS Users Group International (SUGI), San Francisco.

Tibaldi, F., Bruckers, L., Van Oyen, H., Van der Heyden, J. & Molenberghs, G. (2003). Statistical software for calculating properly weighted estimates from Health Interview Survey Data. In: *Sozial- und Präventivmedizin*, 48 (4), 269-71.

Bijlage

Voorbeeldanalyses op de data van de survey van de stadsmonitor

Inleiding

Deze bijlage illustreert het gebruik van de gewichten bij de analyse van de data van de survey van de stadsmonitor. We presenteren zowel ongewogen analyses als gewogen analyses (met 2 verschillende gewichten) en ook enkele analyses met de module COMPLEX SAMPLES.

Deze bijlage bevat veel tabellen met verschillende resultaten en kansen. De tekst tussendoor zorgt al voor enige toelichting, maar het is vooral de samenvattende tabel op het einde die extra structuur in het geheel moet brengen.

Eigenlijk vermengt deze bijlage twee functies:
(1) hoe analyseer en toets je verschillen tussen gemeenten met de data van de stadsmonitor? en
(2) wat zijn mogelijke gevolgen van foutief gebruik van gewichten (in SPSS)?
De bedoeling is vooral om te focussen op (2), en (1) vormt daarbij een goede illustratie. In die illustratie wordt gebruik gemaakt van een kruistabelanalyse (met chi-kwadraattoets) en een logistische regressie.

We vertrekken van het ongewogen bestand dat 15.195 respondenten bevat. Die verdelen zich als volgt over de 13 centrumsteden.

Tabel B1 Ongewogen verdeling van de respondenten over de 13 centrumsteden

Centrumstad	Frequentie	Percentage
Aalst	699	4,6
Antwerpen	4.342	28,6
Brugge	913	6,0
Genk	598	3,9
Gent	1.081	7,1
Hasselt	758	5,0
Kortrijk	741	4,9
Leuven	801	5,3
Mechelen	675	4,4
Oostende	715	4,7
Roeselare	712	4,7
Sint-Niklaas	737	4,9
Turnhout	2.423	15,9
Totaal	15.195	100,0

Bron: Survey stadsmonitor 2008

Als we de gewogen tabel opvragen, ziet die er zo uit:

Tabel B2 Gewogen verdeling van de respondenten over de 13 centrumsteden (**niet-herschaalde gewicht**)

Centrumstad	Frequentie	Percentage
Aalst	65.747	5,2
Antwerpen	389.142	30,7
Brugge	98.420	7,8
Genk	52.324	4,1
Gent	196.847	15,5
Hasselt	61.294	4,8
Kortrijk	62.023	4,9
Leuven	78.178	6,2
Mechelen	64.487	5,1
Oostende	59.561	4,7
Roeselare	46.768	3,7
Sint-Niklaas	58.041	4,6
Turnhout	33.939	2,7
Totaal	1.266.771	100,0

Bron: Survey stadsmonitor 2008

Het is dus duidelijk dat het niet-herschaalde gewicht in tabel B2 de data "opblaast" tot de volledige doelpopulatie. Als we de aantallen van de verschillende steekproefkaders van de 13 steden bij elkaar optellen, krijgen we het totale steekproefkader dat 1.266.771 personen bevat. Daaruit hebben we in eerste instantie 31.649 personen geselecteerd en bij 15.195 personen een ingevulde vragenlijst bekomen. De gewichten die een combinatie zijn van de inverse van de selectiekans en de (geschatte) responskans geven ons in totaal dus terug die 1.266.771 personen. De absolute grootte van de gewichten is onbelangrijk als we ze goed gebruiken. De relatieve grootte speelt natuurlijk wel een rol.

De gewichten maken dat de data binnen elke stad (of stadsdeel) representatief zijn voor de gecombineerde verdeling leeftijd/geslacht, maar ook voor het geheel van alle steden samen. We moeten bij het berekenen van het gewicht geen keuze maken voor één van beide.

In de voorbeeldanalyse in deze bijlage kijken we naar de mate waarin de respondenten fier zijn op de eigen stad. Dat werd bevraagd in vraag 7 van de survey. De vraag luidde "*In welke mate ben je het met onderstaande uitspraken over jouw stad eens?*". De betreffende uitspraak was "*Ik ben echt fier op mijn stad*".

Ongewogen analyse

De ongewogen verdeling van de antwoorden van de respondenten ziet er als volgt uit:

Tabel B3 ***Ongewogen*** verdeling voor de vraag naar fierheid op de eigen stad

	Frequentie	Percentage
Helemaal oneens	446	2,9
Eerder oneens	1.010	6,6
Niet eens, niet	3.195	21,0
Eerder eens	5.555	36,6
Helemaal eens	4.186	27,5
Weet niet	225	1,5
Geen antwoord	578	3,8
Totaal	15.195	100,0

Bron: Survey stadsmonitor 2008

Tabel B3 maakt duidelijk dat er voor deze vraag ook nog wat item-non-respons is, in totaal bijna 4%. Voor de eenvoud hebben we deze variabele gedichotomiseerd (de 2 meest positieve categorieën samengenomen en de 3 minder positieve/negatieve ook) en ook "Weet niet" als ontbrekende waarde gedefinieerd. Tabel B4 toont de verdeling van die gedichotomiseerde variabele.

Tabel B4 Fierheid op de eigen stad (***ongewogen***)

	Frequentie	Percentage
Niet	4.651	32,3
Wel	9.741	67,7
Totaal	14.392	100,0

Bron: Survey stadsmonitor 2008

Bij deze nieuwe variabele staat "wel" dus voor "(heel) fier op mijn stad". 67,7% is dus (heel) fier op zijn stad. Bemerk dat dit een ongewogen percentage is.

De mate waarin mensen fier zijn op hun stad verschilt sterk van stad tot stad. In Turnhout is dat "slechts" 57%, terwijl van de Bruggelingen bijna 90% zegt (heel) fier te zijn op de stad. Een eenvoudige chi-kwadraattest leert ons dat die verschillen heel significant zijn, zeker gegeven de universeel gebruikte $\alpha = 0,05$. De eigenlijke p-waarde heeft alleen maar nullen tot 10 cijfers achter de komma. Deze conclusie is gebaseerd op een test op ongewogen data.

Tabel B5 Fierheid op de eigen stad naar centrumstad (***ongewogen***)

		Fier op stad		
Centrumstad		Niet fier	(Heel) fier	Totaal
Aalst	aantal	258	396	654
	rijpercentage	39,4%	60,6%	
Antwerpen	aantal	1.581	2.476	4.057
	rijpercentage	39,0%	61,0%	
Brugge	aantal	93	795	888
	rijpercentage	10,5%	89,5%	
Genk	aantal	165	399	564
	rijpercentage	29,3%	70,7%	
Gent	aantal	177	844	1.021
	rijpercentage	17,3%	82,7%	
Hasselt	aantal	105	637	742
	rijpercentage	14,2%	85,8%	
Kortrijk	aantal	273	428	701
	rijpercentage	38,9%	61,1%	
Leuven	aantal	140	630	770
	rijpercentage	18,2%	81,8%	
Mechelen	aantal	232	419	651
	rijpercentage	35,6%	64,4%	
Oostende	aantal	169	515	684
	rijpercentage	24,7%	75,3%	
Roeselare	aantal	201	475	676
	rijpercentage	29,7%	70,3%	
Sint-Niklaas	aantal	281	418	699
	rijpercentage	40,2%	59,8%	
Turnhout	aantal	976	1.309	2.285
	rijpercentage	42,7%	57,3%	
Totaal	aantal	4.651	9.741	14.392
	rijpercentage	32,3%	67,7%	

Bron: Survey stadsmonitor 2008

De chi-kwadraattoets levert ons een algemene indicatie van de afhankelijkheid tussen de variabelen centrumstad en fierheid en bevestigt dus dat de mate van fier zijn verschilt van stad tot stad. Maar je krijgt geen test om onderling steden te vergelijken of om één bepaalde stad met een gemiddelde te vergelijken. Zulke testen zijn op verschillende andere manieren wel mogelijk. Eén van die manieren is een logistische regressie waarbij de verschillende steden als dummy of effect in het model worden opgenomen.

Hoewel effectcodering vaak interessantere conclusies kan opleveren (er is immers geen arbitraire referentiecategorie nodig) zullen wij in deze illustratie toch dummycodering gebruiken. De resultaten van gewogen en ongewogen analyses met dummycodering laten immers enkele duidelijke en ongewenste verschillen zien. Bij dummycodering wordt er voor elke stad een nieuwe variabele aangemaakt die waarde 1 krijgt als de respondent inwoner is van die stad en waarde 0 in alle andere gevallen. Je kan zo dus 13 dummies aanmaken, maar er wel slechts 12 opnemen in de logistische regressie. De 13de is immers een perfecte lineaire combinatie van de 12 andere.

We gebruiken nu logistische regressie om de analyse van de kruistabel (tabel B5) te repliceren en bijkomende extra statistische testen te bekomen. We doen dit voorlopig nog altijd op de ongewogen data. We nemen als onafhankelijke variabelen 12 dummies op (allemaal behalve Aalst). De afhankelijke variabele van de logistische regressie is het al dan niet fier zijn op de eigen stad.

De logistische regressie geeft ons in eerste instantie een model test. Die is eigenlijk identiek aan één van de chi-kwadraattesten bij tabel B5 (niet de pearson chi-square, maar wel de likelihood ratio chi-square). Dit toont aan dat we uit onze logistische regressie dezelfde informatie halen als uit onze eenvoudige chi-kwadraattest. Ook hier is de algemene conclusie van deze ene test “Er is een verschil tussen de centrumsteden in de mate waarin hun inwoners fier zijn op hun stad”.

Belangrijker zijn de verschillende parameterschattingen en bijhorende significantietesten. De betekenis van de b en Exp(b) wordt in handboeken over logistische regressie voldoende uitgelegd (zie bijvoorbeeld Hosmer & Lemeshow, 2000). Hier kunnen we ons beperken tot welke verschillen significant zijn. We hebben in deze logistische regressie Aalst niet opgenomen, dat is dus de referentiecategorie waarmee we alle andere steden vergelijken. Op basis van de voorlaatste kolom zien we dat Antwerpen, Kortrijk, Mechelen, Sint-Niklaas en Turnhout niet significant verschillen van Aalst op niveau α = 0,05. Alle andere steden dus wel.

Tabel B6 Resultaten van de logistische regressie met fier op stad als afhankelijke variabele (***ongewogen***)

	b	Stand. fout	p-waarde	e^b
Intercept	0,428	0,080	0,000	1,535
Centrumstad				
Antwerpen	0,020	0,086	**0,815**	1,020
Brugge	1,717	,0136	0,000	5,569
Genk	0,455	0,122	0,000	1,575
Gent	1,134	0,115	0,000	3,107
Hasselt	1,374	0,132	0,000	3,953
Kortrijk	0,021	0,111	**0,849**	1,021
Leuven	1,076	0,123	0,000	2,932
Mechelen	0,163	0,114	**0,155**	1,177
Oostende	0,686	0,119	0,000	1,985
Roeselare	0,432	0,116	0,000	1,540
Sint-Niklaas	-0,031	0,111	**0,778**	0,969
Turnhout	-0,135	0,090	**0,136**	0,874
Referentie: Aalst	-	-	-	-

Bron: Survey stadsmonitor 2008

Wat inferentiële conclusies (veralgemeningen naar de populatie) betreft, geeft deze logistische regressie ons dus dezelfde informatie als de chi-kwadraattest ("er is een significant verschil tussen de steden"), maar ook nog wat meer ("Antwerpen verschilt niet significant van Aalst, maar Brugge bijvoorbeeld wel"). Een beetje ongelukkig bij deze logistische regressie met dummies is de noodzakelijke keuze voor een referentiecategorie (in dit geval Aalst). We krijgen geen test voor een verschil met het gemiddelde en kunnen uit de analyse hierboven ook niet afleiden of Hasselt en Leuven significant verschillen. Met effectcodering kunnen we wel een statistische test krijgen voor een verschil met een gemiddelde.

Gewogen analyse – standaardgebruik van niet-herschaalde gewichten in SPSS

In een volgende stap voeren we een gewogen analyse uit, waarbij we in SPSS gewoon op de defaultwijze wegen met het uiteindelijke gewicht. We krijgen dan aantallen die optellen tot 1.266.771. Door de ontbrekende waarden is het bekomen aantal voor de variabele fierheid op de eigen stad gelijk aan iets minder dan 1.200.000.

Tabel B7 Fierheid op de eigen stad (***standaardgebruik niet-herschaalde gewichten***)

	Frequentie	Percentage
Niet	347.733	29,0
Wel	851.802	71,0
Totaal	1.199.535	100,0

Bron: Survey stadsmonitor 2008

Bemerk dat het gewogen percentage inwoners dat fier is op de eigen stad toch redelijk sterk afwijkt van het ongewogen percentage (71,0% versus 67,7%). De reden hiervoor is relatief eenvoudig. De gewichten trekken de gecombineerde verdeling leeftijd-geslacht binnen elke stad recht, maar ook en vooral de verdeling volgens centrumstad. In het ongewogen bestand is Turnhout het sterkste oververtegenwoordigd (en in Turnhout zijn er minder mensen fier op de stad) en Gent het sterkste ondervertegenwoordigd (en in Gent zijn er relatief meer mensen fier op de stad). Het is dus logisch dat het gewogen percentage "fier op de stad" hoger ligt dan het ongewogen percentage.

Als we steden vergelijken met een eenvoudige kruistabel, zien we in de gewogen analyse opnieuw grote verschillen (bijvoorbeeld 56,3% in Turnhout versus 89,3% in Brugge). Bemerk dat deze percentages per stad sterk aanleunen bij de ongewogen percentages. Er zijn natuurlijk verschillen omdat binnen elke stad gewogen is volgens leeftijd en geslacht, maar dat deel van de weging is minder ingrijpend dan de weging volgens bevolkingsaantal van de centrumstad. Het effect van de weging laat zich dus veel meer voelen op het totale percentage dan op de percentages binnen elke stad.

Tabel B8 Fierheid op de eigen stad naar centrumstad (***standaardgebruik niet-herschaalde gewichten***)

		Fier op stad		
Centrumstad		Niet fier	(Heel) fier	Totaal
Aalst	Aantal	24.072	37.546	61.618
	Rijpercentage	39,1%	60,9%	
Antwerpen	Aantal	134.803	228.410	363.213
	Rijpercentage	37,1%	62,9%	
Brugge	Aantal	10.264	85.494	95.758
	Rijpercentage	10,7%	89,3%	
Genk	Aantal	14.799	34.617	49.416
	Rijpercentage	29,9%	70,1%	
Gent	Aantal	31.987	153.730	185.717
	Rijpercentage	17,2%	82,8%	
Hasselt	Aantal	8.805	51.136	59.941
	Rijpercentage	14,7%	85,3%	
Kortrijk	Aantal	22.877	35.595	58.472
	Rijpercentage	39,1%	60,9%	
Leuven	Aantal	14.173	61.112	75.285
	Rijpercentage	18,8%	81,2%	
Mechelen	Aantal	22.213	39.907	62.120
	Rijpercentage	35,8%	64,2%	
Oostende	Aantal	14.362	42.400	56.762
	Rijpercentage	25,3%	74,7%	
Roeselare	Aantal	13.171	31.154	44.325
	Rijpercentage	29,7%	70,3%	
Sint-Niklaas	Aantal	22.260	32.749	55.009
	Rijpercentage	40,5%	59,5%	
Turnhout	Aantal	13.947	17.952	31.899
	Rijpercentage	43,7%	56,3%	
Totaal	Aantal	347.733	851.802	1.199.535
	Rijpercentage	29,0%	71,0%	

Bron: Survey stadsmonitor 2008

Ook hier vinden we dat het verschil significant is. De p-waarde vertoont alleen maar nullen tot heel wat cijfers achter de komma (zover als SPSS wil gaan). Maar omdat SPSS bij default gebruik van deze gewichten denkt dat er werkelijk 1.200.000 eenheden in onze steekproef zitten, is deze significantietoets natuurlijk waardeloos.

We kunnen ook een gewogen logistische regressie uitvoeren om bijkomende significantietoetsen te verkrijgen. Het is logisch dat deze logistische regressie ook meer significante verschillen toont

dan de ongewogen analyse. Uit tabel B9 blijkt dat volgens deze toets alleen Kortrijk niet significant verschilt van Aalst. Maar het gewogen percentage inwoners dat fier is op de eigen stad, is in Kortrijk en in Aalst dan ook exact gelijk tot op één cijfer na de komma. Bemerk dat er in de ongewogen analyse nog 5 steden waren die geen significant verschil met Aalst vertoonden. Het hoeft niet meer gezegd dat ook deze significantietoetsen uit de logistische regressie waardeloos zijn.

Tabel B9 Resultaten van de logistische regressie met fier op stad als afhankelijke variabele (***standaardgebruik niet-herschaalde gewichten***)

	b	Stand. fout	p-waarde	e^b
Intercept	0,445	0,008	0,000	1,560
Centrumstad				
Antwerpen	0,083	0,009	0,000	1,086
Brugge	1,675	0,013	0,000	5,340
Genk	0,405	0,013	0,000	1,500
Gent	1,125	0,010	0,000	3,081
Hasselt	1,315	0,014	0,000	3,723
Kortrijk	-0,002	0,012	**0,836**	0,998
Leuven	1,017	0,012	0,000	2,765
Mechelen	0,141	0,012	0,000	1,152
Oostende	0,638	0,013	0,000	1,893
Roeselare	0,416	0,013	0,000	1,516
Sint-Niklaas	-0,058	0,012	0,000	0,943
Turnhout	-0,192	0,014	0,000	0,825
Referentie: Aalst	-	-	-	-

Bron: Survey stadsmonitor 2008

Gewogen analyse – standaardgebruik van herschaalde gewichten in SPSS

Eén van de opties die overwogen kan worden om onzinnige significantietoetsen te vermijden, is het herschalen van de gewichten, zodanig dat ze een gemiddelde hebben dat gelijk is aan 1 en het gewogen aantal respondenten gelijk is aan het ongewogen aantal. Deze herschaling deelt dus het gewicht door z'n gemiddelde. Tabel B10 toont die eenvoudige herschaling.

Tabel B10 Beschrijvende statistieken van het definitieve gewicht en het herschaalde gewicht

	N	Min	Max	Som	Gemid	St.Afw.
Niet-herschaalde gewicht	15.195	3,47	530,24	1.266.771	83,37	65,41
Herschaalde gewicht	15.195	0,04	6,36	15.195	1,00	0,78

In de volgende analyses zullen we nu dit herschaalde gewicht toepassen, nog altijd op de klassieke manier waarop SPSS omgaat met gewichten. Het totale aantal in tabel B11 is nu inderdaad gelijk aan dat van de ongewogen analyse (ook al kunnen er nog kleine afwijkingen zijn als gevolg van ontbrekende waarden), maar de percentages zijn gelijk aan deze van de gewogen analyse. Zo lijkt het herschaalde gewicht een oplossing te bieden voor de onzinnige significantietoetsen op de vorige pagina's.

Tabel B11 Fierheid op de eigen stad (***standaardgebruik herschaalde gewichten***)

	Frequentie	Percentage
Niet	4.171	29,0
Wel	10.217	71,0
Totaal	14.388	100,0

Bron: Survey stadsmonitor 2008

Als we de centrumsteden vergelijken, blijken eens te meer de verschillen. En ook hier zijn die verschillen significant. De p-waarde bevat alleen maar nullen tot 10 cijfers achter de komma.

Tabel B12 Fierheid op de eigen stad naar centrumstad (***standaardgebruik herschaalde gewichten***)

		Fier op stad		
Centrumstad		Niet fier	(Heel) fier	Totaal
Aalst	Aantal	289	450	739
	Rijpercentage	39,1%	60,9%	
Antwerpen	Aantal	1.617	2.740	4.357
	Rijpercentage	37,1%	62,9%	
Brugge	Aantal	123	1.026	1.149
	Rijpercentage	10,7%	89,3%	
Genk	Aantal	178	415	593
	Rijpercentage	30,0%	70,0%	
Gent	Aantal	384	1.844	2.228
	Rijpercentage	17,2%	82,8%	
Hasselt	Aantal	106	613	719
	Rijpercentage	14,7%	85,3%	
Kortrijk	Aantal	274	427	701
	Rijpercentage	39,1%	60,9%	
Leuven	Aantal	170	733	903
	Rijpercentage	18,8%	81,2%	
Mechelen	Aantal	266	479	745
	Rijpercentage	35,7%	64,3%	
Oostende	Aantal	172	509	681
	Rijpercentage	25,3%	74,7%	
Roeselare	Aantal	158	374	532
	Rijpercentage	29,7%	70,3%	
Sint-Niklaas	Aantal	267	393	660
	Rijpercentage	40,5%	59,5%	
Turnhout	Aantal	167	215	382
	Rijpercentage	43,7%	56,3%	
Totaal	Aantal	4.171	10.218	14.389
	Rijpercentage	29,0%	71,0%	

Bron: Survey stadsmonitor 2008

De specifiekere testen voor de vergelijking van steden, die we halen uit de logistische regressie, blijken in de lijn te liggen van de ongewogen analyse. Antwerpen, Kortrijk, Mechelen, Sint-Niklaas en Turnhout verschillen niet significant van Aalst, de andere steden wel.

Tabel B13 Resultaten van de logistische regressie met fier op stad als afhankelijke variabele (***gewogen met het herschaalde gewicht***)

	b	Stand. fout	p-waarde	e^b
Intercept	0,445	0,075	0,000	1,560
Centrumstad				
Antwerpen	0,083	0,082	**0,311**	1,086
Brugge	1,675	0,122	0,000	5,340
Genk	0,405	0,117	0,001	1,500
Gent	1,125	0,094	0,000	3,081
Hasselt	1,315	0,130	0,000	3,723
Kortrijk	-0,002	0,108	**0,982**	0,998
Leuven	1,017	0,114	0,000	2,765
Mechelen	0,141	0,107	**0,188**	1,152
Oostende	0,638	0,116	0,000	1,893
Roeselare	0,416	0,121	0,001	1,516
Sint-Niklaas	-0,058	0,109	**0,593**	0,943
Turnhout	-0,192	0,128	**0,133**	0,825
Referentie: Aalst	-	-	-	-

Bron: Survey stadsmonitor 2008

Gegeven het non-responsmodel gaan we ervan uit dat de gewogen percentages niet of minder vertekend zijn dan de ongewogen percentages en daarom is het beter die met elkaar te vergelijken dan de ongewogen percentages. Bovendien is de gewogen totale steekproefomvang gelijk aan de ongewogen totale steekproefomvang. Maar is dit ook werkelijk voldoende voor een correcte statistische toets? Het herschalen herstelt wel het totale aantal, maar niet de respectievelijke aantallen van de steden. Zo zijn er ongewogen 2.285 respondenten in Turnhout, gewogen met het herschaald gewicht nog slechts 382 (zie tabel B12). Dat heeft natuurlijk zijn implicaties bij de statistische toetsen.

COMPLEX SAMPLES analyse

De enige echt afdoende oplossing om in SPSS zinvolle statistische toetsen te verkrijgen bij gewogen analyses, is werken met de module COMPLEX SAMPLES. Hieronder tonen we enkele van die Complex Samples analyses.

Tabel B14 geeft de gewone frequentietabel van de variabele “fier op mijn stad”, opgevraagd via Complex Samples. Je kan in zo’n tabel verschillende schattingen opvragen. In dit geval kozen we voor het geschatte aantal en percentage in de populatie en bijkomend voor beide, de standaardfout, het betrouwbaarheidsinterval en het ongewogen aantal in het databestand. Zo zien we dat het geschatte percentage inwoners van de centrumsteden dat (heel) fier is op z’n stad 71,0% bedraagt en we kunnen met 95% betrouwbaarheid zeggen dat het werkelijke percentage zich tussen 70,1% en 71,9% bevindt. De 71,0% is gelijk aan het percentage dat we vonden in de gewogen analyses (al dan niet herschaald gewicht). Het betrouwbaarheidsinterval geeft wat extra informatie. Uit deze tabel blijkt overigens ook dat we wel degelijk een gewogen percentage krijgen. Een eigen berekening op de ongewogen aantallen levert andere percentages op (9.741 = 67,7% van 14.392).

Tabel B14 Fierheid op de eigen stad (***analyse met Complex Samples***)

		Fier op stad		
		Niet fier	(Heel) fier	Totaal
Geschatte aantal		347.732,8	851.801,8	1.199.534,6
95% betrouw-	*ondergrens*	336.080,9	837.588,9	
baarheidsinterv.	*bovengrens*	359.384,7	866.014,6	
Geschatte percentage		29,0%	71,0%	
95% betrouw-	*ondergrens*	28,1%	70,1%	
baarheidsinterv.	*bovengrens*	29,9%	71,9%	
Ongewogen aantal		4.651	9.741	14.392

Bron: Survey stadsmonitor 2008

Met een kruistabel (tabel B15) kunnen we terug centrumsteden vergelijken en een test uitvoeren. De geschatte percentages die we per stad bekomen, zijn opnieuw gelijk aan de gewogen percentages. Daarrond wordt ook een betrouwbaarheidsinterval geschat. Logischerwijze is dat interval groter als er minder respondenten zijn in de betreffende stad. In Antwerpen bedraagt het betrouwbaarheidsinterval bijvoorbeeld 4 procentpunten [60,8% - 65,0%] terwijl dat in Genk bijna 8 procentpunten is [66,0% - 73,8%].

Tabel B15 Fierheid op de eigen stad naar centrumstad (***analyse met Complex Samples***)

		Fier op stad		
Centrumstad		Niet fier	(Heel) fier	Totaal
Aalst	Geschatte percentage	39,1%	60,9%	
	95% betr. ondergr. interv.	35,3%	57,1%	
	bovengr.	42,9%	64,7%	
	Ongewogen aantal	258	396	654
Antwerpen	Geschatte percentage	37,1%	62,9%	
	95% betr. ondergr. interv.	35,0%	60,8%	
	bovengr.	39,2%	65,0%	
	Ongewogen aantal	1.581	2.476	4.057
Brugge	Geschatte percentage	10,7%	89,3%	
	95% betr. ondergr. interv.	8,8%	87,0%	
	bovengr.	13,0%	91,2%	
	Ongewogen aantal	93	795	888
Genk	Geschatte percentage	29,9%	70,1%	
	95% betr. ondergr. interv.	26,2%	66,0%	
	bovengr.	34,0%	73,8%	
	Ongewogen aantal	165	399	564
Gent	Geschatte percentage	17,2%	82,8%	
	95% betr. ondergr. interv.	15,0%	80,3%	
	bovengr.	19,7%	85,0%	
	Ongewogen aantal	177	844	1.021
Hasselt	Geschatte percentage	14,7%	85,3%	
	95% betr. ondergr. interv.	12,2%	82,5%	
	bovengr.	17,5%	87,8%	
	Ongewogen aantal	105	637	742
Kortrijk	Geschatte percentage	39,1%	60,9%	
	95% betr. ondergr. interv.	35,5%	57,1%	
	bovengr.	42,9%	64,5%	
	Ongewogen aantal	273	428	701
Leuven	Geschatte percentage	18,8%	81,2%	
	95% betr. ondergr. interv.	16,1%	78,2%	
	bovengr.	21,8%	83,9%	
	Ongewogen aantal	140	630	770
Mechelen	Geschatte percentage	35,8%	64,2%	
	95% betr. ondergr. interv.	32,1%	60,4%	
	bovengr.	39,6%	67,9%	
	Ongewogen aantal	232	419	651
Oostende	Geschatte percentage	25,3%	74,7%	
	95% betr. ondergr. interv.	22,1%	71,2%	
	bovengr.	28,8%	77,9%	
	Ongewogen aantal	169	515	684
Roeselare	Geschatte percentage	29,7%	70,3%	
	95% betr. ondergr. interv.	26,4%	66,7%	
	bovengr.	33,3%	73,6%	
	Ongewogen aantal	201	475	676
Sint-Niklaas	Geschatte percentage	40,5%	59,5%	
	95% betr. ondergr. interv.	36,8%	55,8%	
	bovengr.	44,2%	63,2%	
	Ongewogen aantal	281	418	699
Turnhout	Geschatte percentage	43,7%	56,3%	
	95% betr. ondergr. interv.	41,2%	53,7%	
	bovengr.	46,3%	58,8%	
	Ongewogen aantal	976	1.309	2.285
Totaal	Geschatte percentage	29,0%	71,0%	
	95% betr. ondergr. interv.	28,1%	70,1%	
	bovengr.	29,9%	71,9%	
	Ongewogen aantal	4.651	9.741	14.392

Bron: Survey stadsmonitor 2008

Ook de significantietoets van deze Complex Samples analyse geeft een zeer kleine p-waarde, met nullen tot 10 cijfers achter de komma. De conclusie blijft dus dezelfde. Er zijn significante verschillen tussen de centrumsteden.

Een Complex Samples logistische regressie geeft een aantal bijkomende significantietesten. We zijn daarbij het meest geïnteresseerd in de significantie van de parameters van de logistische regressie. De parameters zelf (b en Exp(b)) zijn exact dezelfde als bij de analyses waarbij de defaultweging van SPSS werd toegepast. Daaraan verandert er niks. Op basis van de geschatte probabiliteiten zien we dat Antwerpen, Kortrijk, Mechelen en Sint-Niklaas niet significant verschillen van Aalst. Alle andere centrumsteden wel, ook Turnhout! Dat laatste is dus een verschil ten opzichte van vroegere testen – hetzij ongewogen, hetzij gewogen met het herschaalde gewicht. We komen hier in het besluit op terug.

Tabel B16 Resultaten van de logistische regressie met fier op stad als afhankelijke variabele (***complex samples analyse***)

	b	Stand, fout	p-waarde	e^b
Intercept	0,445	0,082	0,000	1,560
Centrumstad				
Antwerpen	0,083	0,094	**0,377**	1,086
Brugge	1,675	0,138	0,000	5,340
Genk	0,405	0,126	0,001	1,500
Gent	1,125	0,117	0,000	3,081
Hasselt	1,315	0,135	0,000	3,723
Kortrijk	-0,002	0,114	**0,983**	0,998
Leuven	1,017	0,125	0,000	2,765
Mechelen	0,141	0,117	**0,229**	1,152
Oostende	0,638	0,122	0,000	1,893
Roeselare	0,416	0,118	0,000	1,516
Sint-Niklaas	-0,058	0,113	**0,606**	0,943
Turnhout	-0,192	0,097	**0,048**	0,825
Referentie: Aalst	-	-	-	-

Bron: Survey stadsmonitor 2008

Analyse van slechts een deel van de steekproef

Stel dat je de data van de stadsmonitor wil analyseren zonder Antwerpen en Gent.
Dat is niet zo'n eigenaardige veronderstelling. Veel beleidsmensen van de (andere) centrumsteden vinden dat die twee grootsteden echt wel anders zijn en niet zomaar vergeleken mogen worden met hun eigen stad (en vice versa). We gaan na wat het effect is van het analyseren van een deel van de dataset, zowel bij een ongewogen analyse, een default gewogen analyse (al dan niet met herschaald gewicht) en een complex samples analyse. We rapporteren niet meer alle tabellen, maar belichten alleen de belangrijkste verschillen.

Ongewogen zijn er helemaal geen verschillen. De parameterschattingen en significantietesten zijn identiek aan deze in tabel B6. Of we nu Antwerpen en Gent in de analyse opnemen of niet, de testen voor het verschil tussen Aalst en Mechelen en tussen Aalst en Turnhout geven exact dezelfde p-waarden als voorheen.

Dezelfde conclusie gaat op voor de analyse **gewogen met het niet-herschaalde gewicht**. De resultaten daarvan zijn identiek aan deze in tabel B9.
Voor een analyse met **COMPLEX SAMPLES** is het weglaten van de twee grootsteden eveneens vrijwel volledig "neutraal". Er zijn enkele zeer kleine verschillen ten opzichte van tabel B16, maar die bevinden zich alle minstens 5 cijfers na de komma. Deze kleine verschillen zijn eigenlijk verwaarloosbaar en kunnen ook verklaard worden doordat de totale steekproefomvang een rol speelt bij de schattingsmethode.

Als je een deel van de data wil analyseren met het standaardgebruik van het **herschaalde gewicht**, ziet het plaatje er wel heel anders uit. In dat geval moet je opnieuw herschalen, je ongewogen aantal is immers veranderd en er is geen enkele garantie dat het gewogen aantal volgens het

oorspronkelijke herschaalde gewicht daar nog gelijk aan is. In het ongewogen bestand zonder Antwerpen en Gent zitten er nog 9.772 respondenten. De logica van het herschalen gaat ervan uit dat je voor deze partiële analyse je gewichten dus opnieuw herschaalt, zodanig dat ze ook hier een gemiddelde gelijk aan 1 hebben en het gewogen totaal gelijk is aan het ongewogen totaal. Zoals je hieronder kan zien, voldoet het 'gewone' herschaalde gewicht inderdaad niet meer. Het gewogen totale aantal zou te klein zijn.

Tabel B17 Beschrijvende statistieken van het definitieve gewicht, het herschaalde gewicht en het opnieuw herschaalde gewicht voor de data zonder Antwerpen en Gent

	n	Min	Max	Som	Gemid	St.Afw.
Niet-herschaalde gewicht	9.772	3,47	161,93	**680.782**	69,67	37,21
Herschaalde gewicht (hele dataset)	9.772	0,04	1,94	**8.166**	0,84	0,45
Herschaalde gewicht (zonder Antwerpen en Gent)	9.772	0,05	2,32	**9.772**	1,00	0,53

Maar standaardgebruik van dit nieuwe gewicht heeft enkele ongewenste implicaties. De logistische regressie met dummies toont bijvoorbeeld enkele opvallende verschillen met voorgaande analyses.

Tabel B18 Resultaten van de logistische regressie met fier op stad als afhankelijke variabele (***zonder Antwerpen en Gent, gewogen met het opnieuw herschaalde gewicht***)

	b	Stand, fout	p-waarde	e^b
Intercept	0,445	0,069	0,000	1,560
Centrumstad				
Brugge	1,675	0,111	0,000	5,340
Genk	0,405	0,107	0,000	1,500
Hasselt	1,315	0,118	0,000	3,723
Kortrijk	-0,002	0,099	0,980	0,998
Leuven	1,017	0,104	0,000	2,765
Mechelen	0,141	0,098	**0,150**	1,152
Oostende	0,638	0,106	0,000	1,893
Roeselare	0,416	0,111	0,000	1,516
Sint-Niklaas	-0,058	0,100	0,559	0,943
Turnhout	-0,192	0,117	**0,100**	0,825
Referentie: Aalst	-	-	-	-

Bron: Survey stadsmonitor 2008

We bekijken voornamelijk de test van het verschil tussen Mechelen en Aalst en deze van het verschil tussen Turnhout en Aalst. Bij deze nieuwe analyse zijn de p-waarden respectievelijk gelijk aan 0,15 en 0,10, terwijl ze in de voorgaande analyse met een herschaald gewicht gelijk waren aan 0,19 en 0,13. Dat zijn toch al beduidende verschillen. Het is niet onlogisch dat we – als we slechts een deel van de data analyseren – de grootste (en de enige fundamentele) verschillen vinden bij de gewogen analyse met het herschaald gewicht. Het gewicht is daar een functie van de ongewogen grootte van de steekproef en die varieert tussen beide. Het is dus niet onlogisch, maar wel onwenselijk. Een test voor het verschil tussen Aalst en Turnhout zou niet in dergelijke mate bepaald mogen worden door het al dan niet opnemen van de data van Antwerpen en Gent in de analyse. Analyses met een herschaald gewicht kunnen dus leiden tot potentieel tegenstrijdige conclusies, afhankelijk van de populatie die je onderzoekt.

Samenvatting

In tabel B19 op de volgende pagina vatten we de verschillende analyses van de voorgaande pagina's samen. De makkelijkste kolom is eigenlijk die van de "Gewogen Analyse – definitief gewicht". Op één test na is alles hier significant. Maar al deze testen zijn het resultaat van de opgeblazen steekproefomvang en zijn dus waardeloos.

De "waarheid" bevindt zich in de voorlaatste kolom. Complex Samples is eigenlijk de enige manier om binnen SPSS voldoende correct om te gaan met gewichten.

De "Ongewogen analyse" en de "Gewogen analyse – definitief herschaald gewicht" benaderen die "waarheid" relatief vaak, maar niet altijd. De analyse met het herschaald gewicht sluit meestal het dichtst aan bij de resultaten van de Complex Samples analyse.

Een interessante case is de test "verschilt Turnhout van Aalst?". Als we uitgaan van de universeel gebruikte grens ($\alpha = 0,05$) is het antwoord volgens de ongewogen analyse en de analyse met het herschaald gewicht telkens nee. Complex Samples vertelt ons dat er wel een significant verschil is (zij het nipt). De verschillen tussen de drie testen zijn eigenlijk eenvoudig te verklaren. Ongewogen is het verschil tussen Aalst en Turnhout gelijk aan 3,3 procentpunten. In Aalst is 60,6% (heel) fier op de eigen stad, in Turnhout 57,3%. Gegeven het design met een grote steekproef in Turnhout (2.285 respondenten) en een kleinere steekproef in Aalst (654) blijkt dit verschil niet significant. Gewogen is het verschil tussen beide steden groter, namelijk 4,6 procentpunten (Aalst 60,9% fier en Turnhout 56,3%). Maar in de analyse met het herschaald gewicht is ook dit grotere verschil niet significant omdat er rekening wordt gehouden met een gewogen steekproefomvang in beide steden en die is voor Aalst een beetje groter (739), maar voor Turnhout veel kleiner (382). Beide heffen elkaar enigszins op; het grotere verschil dat blijkt uit de gewogen analyse weegt niet op tegen de (in z'n geheel) kleinere aantallen. Complex Samples combineert beide. Het kijkt naar het verschil tussen de steden volgens de gewogen percentages, maar houdt rekening met de effectieve steekproefomvang (ongewogen dus) in beide steden.

De conclusie is dat je voor je statistische toetsen op gewogen surveydata in SPSS eigenlijk steeds Complex Samples moet gebruiken. Als je alleen maar percentages, gemiddeldes of wat dan ook wil berekenen, is de defaultmanier van het gebruik van gewichten in SPSS eveneens in orde. En daarbij doet het er niet toe of die gewichten nu herschaald zijn of niet. Herschalen heeft wel een impact op de statistische toetsen, maar maakt ze daarom niet correct, ook al benaderen ze vaker de juiste resultaten. Soms geeft dat herschalen een vals gevoel van veiligheid. Bovendien moet je die herschaling (in theorie) telkens opnieuw uitvoeren voor afzonderlijke analyses. Dat is niet alleen ingewikkelder, maar kan ook tot tegenstrijdige resultaten leiden, zoals blijkt uit de laatste kolom van tabel B19. Herschalen is dus zeker geen afdoende oplossing.

Tabel B19 Vergelijkende tabel van verschillende significantietesten volgens verschillende methoden

Analyse / Test	Ongewogen analyse	Gewogen analyse – definitief gewicht	Gewogen analyse – definitief gewicht herschaald	Complex Samples	Analyse zonder Antwerpen en Gent – opnieuw herschaald gewicht	Complex Samples Analyse zonder Antwerpen en Gent
Algemene test voor verschil tussen centrumsteden						
p	< 0,000	< 0,000	< 0,000	< 0,000	< 0,000	< 0,000
Logistische Regressie met dummies (Aalst = referentie)						
p(Antwerpen verschilt niet van Aalst)	0,815	< 0,000	0,311	0,378	-	-
p(Brugge verschilt niet van Aalst)	< 0,000	< 0,000	< 0,000	< 0,000	< 0,000	< 0,000
p(Genk verschilt niet van Aalst)	0,002	< 0,000	< 0,001	0,001	< 0,000	0,001
p(Gent verschilt niet van Aalst)	< 0,000	< 0,000	< 0,000	< 0,000	-	-
p(Hasselt verschilt niet van Aalst)	< 0,000	< 0,000	< 0,000	< 0,000	< 0,000	< 0,000
p(Kortrijk verschilt niet van Aalst)	0,849	0,836	0,982	0,983	0,980	0,983
p(Leuven verschilt niet van Aalst)	< 0,000	< 0,000	< 0,000	< 0,000	< 0,000	< 0,000
p(Mechelen verschilt niet van Aalst)	0,155	< 0,000	0,188	0,229	0,150	0,229
p(Oostende verschilt niet van Aalst)	< 0,000	< 0,000	< 0,000	< 0,000	< 0,000	< 0,000
p(Roeselare verschilt niet van Aalst)	< 0,000	< 0,000	0,001	< 0,000	< 0,000	< 0,000
p(St-Niklaas verschilt niet van Aalst)	0,778	< 0,000	0,593	0,606	0,559	0,606
p(Turnhout verschilt niet van Aalst)	**0,136**	< 0,000	**0,133**	**0,048**	**0,100**	**0,048**

Foto: Image 100