De Klassieke Steekproeftheorie - Een Overzicht

Jelke Bethlehem [1]

Samenvatting

Steekproefonderzoek biedt een oplossing voor situaties waarin integraal onderzoek te tijdrovend, te kostbaar of om andere redenen onmogelijk is. Dit rapport biedt een inleiding in het fenomeen steekproefonderzoek. Uitgelegd wordt hoe je op basis van steekproeven toch redelijk nauwkeurige schattingen kunt maken. Een aantal aspecten van de steekproeftheorie wordt behandeld aan de hand van de enkelvoudige aselecte steekproef. Verder wordt uitgelegd hoe met gebruikmaken van hulpinformatie in de trekkingsprocedure of schattingsprocedure nauwkeuriger schatters kunnen worden verkregen. Een aantal veelgebruikte procedures worden meer in detail besproken.

1. Over steekproefonderzoek

1.1. Waarom van steekproefonderzoek?

In onze complexe maatschappij bestaat een groeiende behoefte aan informatie op allerlei gebied. Die informatie stelt politici, wetenschappers en anderen in staat om goed onderbouwde beslissingen te nemen voor een betere toekomst. Soms kunnen die gegevens worden gehaald uit bestaande administratieve bronnen, maar veel vaker komt het voor dat de gewenste gegevens niet beschikbaar zijn. In dat geval is het *survey-onderzoek* het aangewezen middel om nieuwe gegevens te verzamelen. Een survey-onderzoek richt zich op het verzamelen van gegevens over een specifieke populatie. Een dergelijke populatie kan bestaan uit personen, maar hoeft daar zeker niet tot beperkt te blijven. Andere populaties kunnen bijvoorbeeld bestaan uit huishoudens, bedrijven, scholen of boerderijen. Het is kenmerkend voor een survey-onderzoek dat de gegevens worden verzameld met een vragenlijst. Dat is een gestandaardiseerde wijze van afnemen van een reeks vragen. En die vragen moeten worden beantwoord door vertegenwoordigers van de elementen in de doelpopulatie.

---

[1] Centraal Bureau voor de Statistiek, Sector Methoden en Ontwikkeling, Postbus 4000, 2270 JM Voorburg, E-mail: jbtm@cbs.nl

De meest voor de hand liggende manier om gegevens te verzamelen, is alle elementen in de populatie te vragen die gegevens te verstrekken. Dat wordt een *integraal onderzoek* genoemd. Deze benaderingswijze heeft een aantal nadelen. Zeker voor grote populaties is een integraal onderzoek erg kostbaar. Er zijn veel mensen en computers nodig om al die gegevens te verzamelen en te verwerken. Bovendien is een dergelijk onderzoek erg tijdrovend, en dat is niet erg bevorderlijk voor de actualiteit van de uitkomsten. Een ander nadeel van een integraal onderzoek is dat er erg veel mensen voor moeten worden lastig gevallen. Dit leidt tot een grote enquêtedruk. En een grote enquêtedruk is niet bevorderlijk voor de bereidheid van mensen tot medewerking aan dergelijk survey-onderzoek. Het gevolg is steeds meer non-respons en dat tast de betrouwbaarheid van de uitkomsten aan.

Het *steekproefonderzoek* biedt een oplossing voor veel problemen van het integrale onderzoek. In een steekproefonderzoek worden gegevens verzameld bij slechts een klein deel van de elementen in populatie. In principe komen alleen gegevens beschikbaar over die steekproef van elementen. Toch is het mogelijk om op basis van deze beperkte hoeveelheid gegevens uitspraken te doen over de gehele populatie. Als die steekproef op correcte wijze is getrokken, kunnen betrouwbare schattingen worden gemaakt van allerlei karakteristieken van de populatie als geheel.

1.2.Steekproefonderzoek en statistiek

Steekproefonderzoek is een toepassingsgebied van de statistiek. De manier waarop gebruikt wordt gemaakt van de statistische theorie en de kansrekening is echter wel anders dan wat men doorgaans aantreft in de standaardboeken over statistiek. Zo is er in de standaardboeken meestal sprake van onafhankelijke, gelijk verdeelde trekkingen uit een zekere verdeling. Bij steekproefonderzoek wordt een steekproef getrokken uit een eindige populatie van elementen. Doordat overwegend steekproeven zonder teruglegging worden getrokken, is er geen sprake van onafhankelijke trekkingen.

Verder worden in de standaard statistiekboeken de gegevens meestal gemodelleerd als kansvariabelen. Uitgaande van modelveronderstellingen kunnen vervolgens uitspraken worden gedaan over de exacte vorm van de verdeling van deze variabelen. In steekproefonderzoek zijn de gegevens echter vaste waarden die exact kunnen worden gemeten

bij de elementen in de populatie. De stochastiek wordt veroorzaakt door de onderzoeker zelf, doordat hij door middel van een lotingsprocedure een selectie maakt uit de populatie.

Het modelleren van de gegevens volgens een kansmodel draagt een zeker risico in zich. Als niet is voldaan aan de veronderstellingen die aan het model ten grondslag liggen, dan kan dit leiden tot fundamenteel onjuiste conclusies. De aanpak binnen de steekproeftheorie is veel robuuster De geldigheid van de conclusies is niet afhankelijk van modelveronderstellingen. Het is zeker zo dat ook in de steekproeftheorie gebruik wordt gemaakt van modellen. Het al of niet passen van deze modellen leidt echter alleen tot grotere of kleinere varianties van schattingen. De uitspraken die worden gedaan, blijven valide.

In dit rapport wordt wat dieper ingegaan op een aantal aspecten van de theorie van het steekproefonderzoek. We leggen uit hoe een steekproef moeten worden getrokken, hoe schattingen kunnen worden gemaakt op basis van steekproefgegevens, en ook wat kan worden gezegd over de nauwkeurigheid van die schattingen.

2. Het theoretisch raamwerk

2.1. De doelpopulatie

Een steekproef wordt getrokken uit een *populatie* U. We veronderstellen dat die populatie eindig is. De omvang van de populatie wordt aangegeven met N. Als we elk element een volgnummer geven, dan kunnen we de populatie noteren als een verzameling

$$U = \{1,2,,...,N\}. \tag{2.1.1}$$

De *doelvariabele* stelt het verschijnsel voor dat we willen onderzoeken. De doelvariabele noemen we Y. Hij neemt voor elk element in de populatie een zekere waarde aan. Die waarden geven we aan met

$$Y_1, Y_2, ..., Y_N \tag{2.1.2}$$

Is de doelvariabele bijvoorbeeld het inkomen van de te onderzoeken personen, dan is Y_1 het inkomen van persoon 1, Y_2 het inkomen van persoon 2, enz.

Doel van het onderzoek is het doen van uitspraken over bepaalde karakteristieken van de doelpopulatie. Zulke karakteristieken worden meestal aangeduid als *populatiegrootheden* of

populatieparameters. Een belangrijke populatiegrootheid is het populatiegemiddelde. Het *populatiegemiddelde* van de doelvariabele Y is gedefinieerd als

$$\overline{Y} = \frac{1}{N}\sum_{k=1}^{N} Y_k. \tag{2.1.3}$$

Een ander voorbeeld van een populatiegrootheid is het populatietotaal, gedefinieerd door

$$Y_T = \sum_{k=1}^{N} Y_k = N\overline{Y}. \tag{2.1.4}$$

Een andere populatiegrootheid die nog moet worden genoemd, is de *populatievariantie*. Deze grootheid zegt iets over de mate van variatie van de waarden van de doelvariabele. De populatievariantie is gedefinieerd als

$$S^2 = \frac{1}{N-1}\sum_{k=1}^{N}\left(Y_k - \overline{Y}\right)^2 \tag{2.1.5}$$

De populatievariantie speelt een belangrijke rol bij het bepalen van de nauwkeurigheid van schattingen. Deze grootheid kan ook worden opgevat als een maat voor de homogeniteit van de populatie. Heeft bijvoorbeeld iedereen in de populatie hetzelfde inkomen, dan is elk inkomen ook gelijk aan het gemiddelde inkomen. In deze situatie is de populatievariantie dus gelijk aan 0. Naarmate de inkomensverschillen groter zijn, zal ook de populatievariantie toenemen.

Soms wordt in een steekproefonderzoek ook gebruik gemaakt van *hulpinformatie*. Hierbij gaat het om *hulpvariabelen* die ook in het onderzoek worden gemeten, en waarover extra informatie beschikbaar is op het niveau van de populatie. Hulpvariabelen meten vaak achtergrondkenmerken van de onderzochte personen. Voorbeelden van zulke variabelen zijn geslacht, leeftijd, burgerlijke staat en provincie waarin men woont. Voor deze variabelen is vaak de verdeling, of het gemiddelde, in de populatie bekend. In de analyse van de onderzoeksgegevens bieden hulpvariabelen de mogelijk de uitkomsten nader te differentiëren naar verschillende groepen. Daarnaast kunnen hulpvariabelen worden gebruikt voor het verkrijgen van nauwkeuriger schattingen van populatiegrootheden. Een hulpvariabele zal worden aangeduid met de letter X. De waarden van de hulpvariabele in de populatie worden genoteerd met $X_1, X_2, \ldots, X_N$. Net zoals dat het geval is voor een doelvariabele, kunnen ook voor een hulpvariabele populatiegrootheden zoals het gemiddelde worden uitgerekend.

2.2. De steekproef

Als we een populatiegrootheid willen schatten, dan moeten we een steekproef trekken. Hiervoor moet een lotingsprocedure worden gebruikt. Een dergelijke procedure garandeert dat in de selectieprocedure niemand (bewust of onbewust) wordt bevoordeeld of benadeeld. Er wordt bereikt dat, gemiddeld gesproken, de steekproef representatief is voor de populatie waaruit hij wordt getrokken.

Om het lotingsmechanisme op eerlijke en objectieve wijze te kunnen laten werken, is een soort apparaat nodig. Een dergelijke apparaat heet een *aselecte getallen generator*, of kortweg *aselector*. De aselector moet voldoen aan de volgende eigenschappen:

- Het apparaat kan herhaaldelijk worden gebruikt;
- Iedere keer dat het apparaat in werking wordt gesteld geeft het één van de getallen 1 t/m N als uitkomst, waarbij N bekend wordt verondersteld;
- Elke keer opnieuw hebben alle mogelijk uitkomsten dezelfde kansen. Kennis over eerdere uitkomsten helpt niet bij het voorspellen van een volgende uitkomst. Kortom, elk voorspellingssysteem faalt.

De aselector is een theoretisch concept. In de praktijk bestaat de ideale aselector niet. Er zijn wel 'apparaten' die dicht in de buurt komen. Voorbeelden hiervan zijn een dobbelsteen (voor het loten van een getal uit de reeks 1 t/m 6), een roulette (voor het loten van een getal uit de reeks 0 t/m 36) en een lottomachine (voor het loten van een getal uit de reeks 1 t/m 41).

Deze aselectoren zijn natuurlijk leuk voor het gebruik in spelletjes waarin kansen een rol spelen, maar in de praktijk van het trekken van steekproeven zijn ze niet erg bruikbaar. Om een grote steekproef te selecteren uit een echte populatie is een andere aselector nodig. In de vroege historie van het trekken van steekproeven werd gebruik gemaakt van tabellen met aselecte getallen. Tegenwoordig worden meestal computers en rekenmachines gebruikt voor het genereren van aselecte getallen.

We laten nu eerst zien hoe de tabel met aselecte getallen kan worden gebruikt voor het trekken van een steekproef. Daartoe reproduceren we in figuur 2.2.1 een deel van de tabel met aselecte getallen.

Stel dat we een steekproef van 10 leden moet trekken uit het ledenbestand van een vereniging. Er is bekend dat de vereniging 682 leden heeft. We moeten dan dus 10 aselecte getallen hebben uit de reeks 1 t/m 682.

Figuur 2.2.1. Aselecte getallen

00822	63134	04080	29373	68731	34282	41827	94880	11505	07677	52659	69262
79771	19758	62062	81259	11215	42167	70001	78364	74388	10001	62523	83366
58614	41056	09869	27746	12931	93018	56160	39534	93340	87194	18428	51695
71287	49101	03330	45468	52358	62658	33674	26879	17227	49102	83879	23802
12073	76580	28601	14410	57528	04036	28540	91001	89127	94058	95697	78977

In de tabel zoeken we nu een willekeurig beginpunt en nemen een willekeurige route door de aselecte getallen. We nemen steeds drie opeenvolgende cijfers en zien dat als een getal van drie cijfers. Is dat getal groter dan 682, dan negeren we het en pakken het volgende getal. Is het getal uit de reeks van 1 t/m 682, dan is dat het volgnummer van een lid dat in de steekproef komt. Stel dat we linksboven beginnen, van links naar rechts gaan, en steeds van elke groep van vijf cijfers de eerste drie nemen. We krijgen dan als eerste 10 aselecte getallen:

008, 631, 040, 293, 687, 342, 418, 948, 115, 076 ...

De getallen 687 en 948 zijn groter dan 682, en doen daarom niet mee. De eerste 8 geselecteerde leden zijn dus de leden met volgnummers

8, 631, 40, 293, 342, 418,115 en 76.

Veel programmeertalen en rekenmachines hebben tegenwoordig de mogelijkheid om aselecte getallen te genereren. Heel vaak is er een routine aanwezig die een aselecte waarde genereert uit het interval [0, 1). Deze routine kan worden gebruikt voor het trekken van een willekeurig volgnummer uit de reeks 1 t/m N. Dat gaat als is weergegeven in figuur 2.2.2:

Figuur 2.2.2. Maken van aselecte getallen met de computer

Stap 1:	Trek een aselecte waarde uit [0, 1).
Stap 2:	Vermenigvuldig die waarde met N.
Stap 3:	Rond de uitkomst naar beneden af op een geheel getal.
Stap 4:	Tel bij de uitkomst 1 op.

Zou de routine achtereenvolgens de waarden

0,12073 0,76580 0,28601 0,14410

produceren, dan leidt toepassing van het bovenstaande algoritme voor N = 682 tot de nummers

83, 523, 196, 99

We kunnen de tabel met aselecte getallen ook gebruiken in combinatie met het bovenstaande algoritme. We nemen dan steeds een groepje opeenvolgende cijfers, en zien dat als het deel achter de komma van een getal tussen 0 en 1. Nemen we de eerste 4 groepen van vijf cijfers in de eerste rij, dan krijgen we

0,00822 0,63134 0,04080 0,29373,

waarna toepassing van het algoritme leidt tot de nummers

6, 431, 28, 201.

Daar waar de computer intensief wordt gebruikt voor het genereren van aselecte getallen, is wel enige voorzichtigheid geboden, want ze produceren meestal geen echte aselecte getallen, maar *pseudo-aselecte* getallen. Dat betekent dat vanaf eenzelfde startwaarde steeds dezelfde reeks getallen wordt gegenereerd. Bovendien hebben deze routines een cyclus. Na een vast aantal getallen komt de routine weer terug bij zijn beginwaarde. De lengte van deze cyclus moet voldoende groot te zijn willen de geproduceerde aselecte getallen goed bruikbaar zijn. De boodschap is dat bij professionele toepassingen het aanbeveling verdient dit soort aselectoren goed te testen voor ze worden gebruikt.

Steekproeven kunnen worden getrokken *met teruglegging* of *zonder teruglegging*. Bij steekproeven met teruglegging wordt een geselecteerd element weer teruggelegd in de populatie alvorens een volgende element wordt geselecteerd. Het is dan mogelijk dat een element meer dan één keer wordt getrokken in de steekproef.

Aangezien dit minder informatie oplevert dan allemaal verschillende elementen, wordt meestal de voorkeur gegeven aan een steekproef zonder teruglegging. Nog even terugkerend naar de simpele aselectoren, levert gebruik van een roulette een steekproef met teruglegging, en een lottomachine een steekproef zonder teruglegging.

De omvang van de te trekken steekproef geven we aan met n. Een steekproef van omvang n die is getrokken uit een populatie van omvang N kunnen we aangeven door middel van een reeks indicatoren

$$t_1, t_2, ..., t_N. \tag{2.2.1}$$

De indicator t_k geeft aan hoe vaak element k uit de populatie is getrokken. Voor steekproeven zonder teruglegging kan t_k alleen de waarden 0 en 1 aannemen (niet of wel getrokken). Voor steekproeven met teruglegging kan t_k ook waarden groter dan 1 aan. We beperkingen ons hier overwegend tot steekproeven zonder teruglegging. Aangezien de waarden van deze indicatoren het resultaat zijn van de werking van een kansmechanisme, noemen we ze *stochastische variabelen* of *toevalsvariabelen*. De steekproefomvang n kan worden teruggevonden door optellen van de indicatoren:

$$n = \sum_{k=1}^{N} t_k. \tag{2.2.2}$$

In het steekproefontwerp wordt vastgelegd welke kansen de elementen in de populatie krijgen om in de steekproef te komen. De kans om in de steekproef te komen wordt de insluitkans genoemd. Er wordt onderscheid gemaakt tussen eerste orde insluitkansen en tweede orde insluitkansen. De *eerste order insluitkans* π_k van element k is gedefinieerd als

$$\pi_k = P(t_k = 1) = E(t_k), \tag{2.2.3}$$

voor i = 1, 2, ..., N. Hierin is $E(t_k)$ de verwachting van de stochastische variabele t_k. Dat de verwachting van t_k gelijk is aan de kans dat deze variabele de waarde 1 aanneemt, is het directe gevolg van het feit dat hij voor steekproeven zonder teruglegging slechts de waarden 0 en 1 kan krijgen. Combinatie van formules (2.2.2) en (2.2.3) leidt tot de nuttige relatie

$$\sum_{k=1}^{N} \pi_k = n. \tag{2.2.4}$$

Dus de som van de insluitkansen is altijd gelijk aan de steekproefomvang.

Voor het bepalen van de nauwkeurigheid van een schatting zijn ook nog de tweede orde insluitkansen nodig. De tweede orde insluitkans π_{kl} van twee elementen k en l is gedefinieerd als de kans dat de twee elementen samen in de steekproef komen. In formule:

$$\pi_{kl} = P(t_k t_l = 1) = E(t_k t_l), \tag{2.2.5}$$

met $k \neq l$. Voor $k = l$ definiëren we $\pi_{kk} = \pi_k$. Voor de tweede orde insluitkansen geldt

$$\sum_{k=1}^{N} \sum_{l=k+1}^{N} \pi_{kl} = \frac{n(n-1)}{2}. \qquad (2.2.6)$$

Voor steekproeven met teruglegging kunnen analoge formules worden opgeschreven, zij het dat we dan niet meer praten over insluitkansen maar over insluitverwachtingen. De *eerste orde insluitverwachting* is bijvoorbeeld de verwachting van het aantal keer dat een element in de steekproef wordt getrokken.

2.3. Schatters

Voor de geselecteerde elementen (dus de elementen met de geselecteerde volgnummers) meten we de waarde van de doelvariabele. Dit zijn de metingen die in het steekproefonderzoek beschikbaar komen. De beschikbare metingen geven we aan met

$$y_1, y_2, \ldots, y_n. \qquad (2.3.1)$$

Merk op dat we zoveel mogelijk kleine letters gebruiken voor alles wat met de steekproef te maken heeft, en hoofdletters voor alles wat betrekking heeft op de populatie. Dus y_1 is de waarde van het eerste element in de steekproef, dus van de eerste indicator met een waarde groter dan 0.

Voor alle elementen in de steekproef kan de waarde van de doelvariabele worden waargenomen en vastgelegd. Het zijn deze waarden die moeten worden gebruikt voor het schatten van de populatiekarakteristieken. Het recept voor de berekening van een schatting wordt een *schatter* genoemd. Het is een functie van de in de steekproef beschikbaar gekomen waarden van de doelvariabele. Bruikbare schatters moeten enkele speciale eigenschappen hebben:

Laat z een schatter zijn voor een populatiegrootheid Z. Dan moet z een *zuivere schatter* zijn. Dit wordt aangegeven met de formule

$$E(z) = Z \qquad (2.3.2)$$

De verwachte waarde van de schatter moet dus gelijk zijn aan de te schatten grootheid. De verwachte waarde kan worden gezien als de gemiddelde waarde van de schatter over alle mogelijke realisaties van de steekproef. Het is bij benadering het gemiddelde van alle

schattingsuitkomsten als de steekproeftrekking een zeer groot aantal malen zou worden herhaald. De eis van zuiverheid garandeert dat de schatter nooit de waarde van de populatiegrootheid systematisch over- of onderschat.

De schatter moet ook *nauwkeurig* zijn. Dit houdt in dat de variantie V(z) van de schatter z klein moet zijn:

$$V(z) = E(z - Z)^2 \text{ is klein} \qquad (2.3.3)$$

De variantie kan worden gezien als het gemiddelde van de kwadraten van de verschillen tussen de mogelijke waarden van de schatter en de waarde van de populatiegrootheid. In het ideale geval levert de schatter altijd de juiste waarde op, zodat de variantie dan 0 is. Naarmate de variantie kleiner is, is de schatter nauwkeuriger.

We zullen het hier vooral hebben over schatters voor het populatiegemiddelde. Om redenen van eenvoud wordt bij deze schatters vaak de voorkeur gegeven aan lineaire schatters, of een combinatie van lineaire schatters. Een lineaire schatter is een schatter waarvan de waarde wordt verkregen als een gewogen gemiddelde van de in de steekproef verkregen waarden van de doelvariabele. Uiteraard zijn die makkelijker te berekenen, maar dat is met de huidige generatie computers niet meer zo'n punt. Wat ook meespeelt is dat voor eenvoudige schatters de eigenschappen van de verdeling makkelijker kunnen worden bestudeerd. Als we de eis van zuiverheid opleggen aan een lineaire schatter op basis van een willekeurig steekproefontwerp, dan leidt dit een schatter die door Horvitz-Thompson (1952) is voorgesteld. De Horvitz-Thompson-schatter is als volgt gedefinieerd:

$$\bar{y}_{HT} = \frac{1}{N}\sum_{k=1}^{N} t_k \frac{Y_k}{\pi_k}. \qquad (2.3.4)$$

Door gebruik te maken van de relatie $E(t_k)=\pi_k$ kan eenvoudig worden aangetoond dat (2.3.4) inderdaad een zuivere schatter is. In de formule zorgen de steekproefindicatoren $t_1, t_2, \ldots, t_N$ ervoor dat alleen de steekproefgegevens worden meegenomen in de berekening. Verder wordt elke waarneming Y_k gewogen met de insluitkans π_k. Intuïtief valt wel duidelijk te maken waarom dit nodig is. Immers, elementen met een grotere insluitkans hebben een grotere kans om in de steekproef te komen, en zullen daarom oververtegenwoordigd zijn. Om deze onevenwichtigheid recht te trekken, moeten deze waarnemingen minder zwaar meetellen. Dat wordt gerealiseerd door ze te delen door hun (grotere) insluitkansen.

De variantie van de Horvitz-Thompson-schatter is gelijk aan

$$V(\bar{y}_{HT}) = \frac{1}{N^2}\sum_{k=1}^{N}\sum_{l=1}^{N}(\pi_{kl} - \pi_k\pi_l)\frac{Y_k Y_l}{\pi_k\pi_l}. \qquad (2.3.5)$$

Voor situaties waarin elke mogelijke steekproef dezelfde steekproefomvang heeft (en dat is meestal het geval), kan de variantie (2.3.5) ook worden geschreven als

$$V(\bar{y}_{HT}) = \frac{1}{2N^2}\sum_{k=1}^{N}\sum_{l=1}^{N}(\pi_k\pi_l - \pi_{kl})\left(\frac{Y_k}{\pi_k} - \frac{Y_l}{\pi_l}\right)^2. \qquad (2.3.6)$$

Aan deze alternatieve schrijfwijze is te zien dat de variantie van de Horvitz-Thompson-schatter kleiner is naarmate de insluitkansen meer proportioneel zijn met de waarden van de doelvariabele. Immers, in die situatie zijn de quotiënten Y_k/π_k bijna constant, en dus hun verschillen klein. In het geval dat de quotiënten exact constant zijn, wordt de variantie zelfs 0. Deze ideale situatie zal zich in de praktijk niet voordoen, omdat dit zou betekenen dat de waarden van de doelvariabele (op een constante na) bekend zijn. En het onderzoek werd juist uitgevoerd om die waarden te weten te komen.

De boodschap die we kunnen halen uit de eigenschappen van de Horvitz-Thompson-schatter is dat het niet vanzelfsprekend is om steekproeven met gelijke kansen te trekken. Weliswaar wordt dit toch uit het oogpunt van eenvoud het meest gedaan, maar een steekproefontwerp gebaseerd op ongelijke insluitkansen kan in veel situaties tot nauwkeuriger schattingen leiden.

Een bepaalde keuze voor de insluitkansen leidt tot een specifiek steekproefontwerp en de uitwerking daarvan in de theorie van Horvitz en Thompson tot een specifieke schatter. Enkele daarvan zullen in de volgende paragrafen aan de orde komen.

3. De enkelvoudige aselecte steekproef

3.1. De steekproeftrekking

De aselecte steekproef is de meest eenvoudige steekproef die kan worden getrokken, en misschien is het ook wel de meest bekende manier om dat te doen. Deze wijze van trekken staat het dichtst bij wat intuïtief onder het trekken van een steekproef wordt verstaan, namelijk dat elk element in de populatie dezelfde kans moet hebben om in de steekproef te komen. Een aselecte steekproef kan met en zonder teruglegging worden getrokken, maar we beperken ons hier tot steekproeven zonder teruglegging.

Voor het trekken van een enkelvoudige aselecte steekproef kan een van de simpele aselectoren worden gebruikt zoals beschreven in hoofdstuk 2. Voor grootschalige toepassing wordt uiteraard gebruik gemaakt van computerprogramma's.

3.2. De schattingsprocedure

De enkelvoudige aselecte steekproef gebruikt gelijke insluitkansen voor alle elementen. Toepassing van formule (2.2.4) geeft dat de eerste orde insluitkans π_k van element i gelijk moet zijn aan $\pi_k = n / N$. Verder volgt uit formule (2.2.6) dat de tweede orde insluitkansen dat π_{kl} gelijk moet zijn aan $\pi_{kl} = n(n-1) / N(N-1)$. Invullen van deze insluitkansen in formule (2.3.4) voor de Horvitz-Thompson-schatter levert dan de schatter gelijk is aan

$$\bar{y} = \frac{1}{n}\sum_{k=1}^{N} t_k Y_k = \frac{1}{n}\sum_{i=1}^{n} y_i \qquad (3.2.1)$$

De schatter is dus gelijk aan het gemiddelde van de waarnemingen in de steekproef, of wel het steekproefgemiddelde.

We zien hier dat het populatiegemiddelde wordt geschat met de analoge grootheid, maar dan berekent op basis van de steekproef. Dit analogieprincipe blijkt vaak (maar niet altijd) op te gaan in de steekproeftheorie.

Uitwerken van de formule voor de variantie van de Horvitz-Thompson-schatter voor dit steekproefontwerp leidt tot

$$V(\bar{y}) = \frac{1-f}{n} S^2 = \frac{1-f}{n} \frac{1}{N-1} \sum_{k=1}^{N} \left(Y_k - \bar{Y}\right)^2 . \qquad (3.2.2)$$

Hierin is f = n/N de *steekproeffractie*, en S^2 is de in (2.1.5) geïntroduceerde populatie-variantie. Aan deze formule zijn twee interessante dingen te zien.

- Aangezien (1-f) / n geschreven kan worden als (1/n - 1/N), wordt de variantie kleiner als de steekproefomvang groter wordt. Dat betekent dat nauwkeuriger schattingen worden verkregen naarmate grotere steekproeven wordt getrokken.
- In situaties waarin de omvang van de populatie veel groter is dan die van de steekproef, is (1-f) / n ongeveer gelijk aan 1/n. De variantie is dan onafhankelijk van de omvang van de populatie. Als de waarde van de populatievariantie maar gelijk blijft, dan doet het er voor

de nauwkeurigheid van de schatting niet toe of een steekproef wordt getrokken uit een populatie van 1000 elementen of uit een populatie van een 1000000 elementen.

Intuïtief komt de tweede constatering in eerste instantie misschien wat onwaarschijnlijk over. Vergelijk dit echter maar eens met het proeven van de soep. Dat ene lepeltje soep is voldoende om een oordeel te vellen over zowel een klein pannetje soep als een grote teil soep, mits er uiteraard eerst maar goed geroerd is.

Om iets te kunnen zeggen over de nauwkeurigheid van de berekende schatting, is de waarde van de variantie nodig. Helaas is die in het algemeen niet bekend, omdat de variatieformule de populatievariantie S^2 bevat. Deze kan alleen worden berekend als alle waarden van de doelvariabele in de populatie bekend zijn. De oplossing van dit probleem is om de populatievariantie te schatten op basis van de steekproefgegevens. Ook hierbij gaat het analogieprincipe weer op. Er kan worden aangetoond dat de steekproefvariantie s^2, gedefinieerd door

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 \qquad (3.2.3)$$

een zuivere schatter is voor de populatievariantie S^2. Dus is

$$v(\bar{y}) = \frac{1-f}{n}s^2 \qquad (3.2.4)$$

een zuivere schatter voor de variantie van het steekproefgemiddelde. Met behulp van de (geschatte) variantie kan een betrouwbaarheidsinterval worden afgeleid. Zie hiervoor paragraaf 3.3.

Het *schatten van een percentage* is iets wat misschien nog wel vaker wordt gedaan dan het schatten van het gemiddelde. Voorbeelden zijn het schatten van het percentage voorstanders van een TGV door het Groene Hart van Holland, of het percentage werklozen, of het percentage gezinnen met een Internetaansluiting, enz. Het aardige van het schatten van percentages is dat de theorie hiervoor heel simpel kan worden afgeleid uit die voor het schatten van gemiddelden. Bovendien worden de formules een stuk eenvoudiger dan die voor het gemiddelde.

Bij het schatten van een percentage gaat het om het wel of niet hebben van een zeker kenmerk. Heeft een element het kenmerk wel, dan krijgt de doelvariabele de waarde 1, en heeft het element het kenmerk niet, dan wordt de waarde van de doelvariabele 0. Het

populatiegemiddelde van deze doelvariabele is dan gelijk aan de fractie enen, en dus gelijk aan de fractie elementen met dat kenmerk. Vermenigvuldigen van dat gemiddelde met 100 levert het percentage elementen met dat kenmerk. Wordt het populatiepercentage aangegeven met de letter P, dan geldt:

$$P = 100\overline{Y} \tag{3.2.5}$$

Voor het schatten van dit populatiepercentage moet eerst het populatiegemiddelde worden geschat. Daarvoor wordt het steekproefgemiddelde gebruikt. In dit geval is dat gelijk aan de fractie elementen in de steekproef met het betreffende kenmerk. Vermenigvuldigen van dit steekproefgemiddelde met 100 geeft het steekproefpercentage. Dit wordt aangegeven met

$$p = 100\overline{y} \tag{3.2.6}$$

Aangezien het steekproefgemiddelde een zuivere schatter is voor het populatiegemiddelde, is het steekproefpercentage een zuivere schatter voor het populatiepercentage.

De variantie van het steekproefgemiddelde wordt gevonden door in de variantieformule (3.2.2) de term S^2 uit te werken voor een populatie waarin het percentage elementen met de waarde 1 gelijk is aan P en het percentage elementen met waarde 0 gelijk is aan 100 - P. De variantie-formule wordt dan heel eenvoudig. Het resultaat is:

$$V(p) = \frac{1-f}{n}\frac{N}{N-1}P(100-P) \tag{3.2.7}$$

Deze variantie kan worden geschat op basis van de steekproefgegevens met behulp van de formule

$$v(p) = \frac{1-f}{n-1}p(100-p) \tag{3.2.8}$$

En deze geschatte variantie kan op zijn beurt weer worden gebruikt voor het maken van een betrouwbaarheidsinterval.

Als een schatting is verkregen voor de variantie van de schatter, dan moet die uitkomst worden geïnterpreteerd in termen van nauwkeurigheid. Dat is niet echt eenvoudig. Om te beginnen heeft de variantie een andere meeteenheid dan de schatter zelf. Is de doelvariabele bijvoorbeeld het inkomen in guldens, dan wordt de variantie uitgedrukt in guldens in het kwadraat. De meeteenheid is dus anders. Die meeteenheid kan eenvoudig gelijk worden

gemaakt door de wortel te nemen uit de variantie. De grootheid die zo wordt verkregen, heet de *standaardfout van de schatter*. Deze wordt genoteerd met

$$S(\bar{y}) = \sqrt{V(\bar{y})} \tag{3.2.9}$$

Deze standaardfout kan worden geschat door in deze formule de populatievariantie te vervangen door de schatter voor de populatievariantie. Het resultaat is

$$s(\bar{y}) = \sqrt{v(\bar{y})} \tag{3.2.10}$$

De standaardfout heeft wel dezelfde meeteenheid als de schatter, maar hij blijft lastig te interpreteren. Voor welke waarde van de standaardfout kunnen is een schatter nu nauwkeurig? Als de standaardfout bekend is, kan dan iets worden gezegd over de maximale afwijking tussen schatting en de te schatten populatiewaarde?

De statistische theorie biedt een antwoord op dit soort vragen in de vorm van het *betrouwbaarheidsinterval*. Dit wordt gekenmerkt door een onder- en een bovengrens die zijn bepaald op grond van de beschikbare gegevens, en wel zo dat de kans dat dit interval de (onbekende) populatiewaarde bevat, minstens gelijk is aan een van te voren vastgestelde (grote) kans 1 - α. De grootheid 1 - α wordt de *betrouwbaarheid* genoemd. Vaak wordt voor α de waarde 0,05 gekozen. Daaruit volgt dat de betrouwbaarheid dan gelijk is aan 0,95. De betekenis daarvan is de volgende: als de steekproeftrekking en de berekening van de schatting een groot aantal malen zou worden herhaald, dan zou in gemiddeld 95 van de 100 gevallen het betrouwbaarheidsinterval de te schatten populatiewaarde bevatten. Als dus de uitspraak wordt gedaan dat het betrouwbaarheidsinterval de onbekende populatiewaarde bevat, dan is die inspraak in gemiddeld 5% van de gevallen een onjuiste uitspraak. Anders geformuleerd: de onderzoeker loopt het risico in gemiddeld 1 op de 20 gevallen een verkeerde uitspraak te doen.

De keuze van de betrouwbaarheid is in principe vrij. Is een uitspraak met een hoge betrouwbaarheid vereist, dan moet de waarde van α kleiner worden genomen. Een waarde α=0,01 zou bijvoorbeeld kunnen worden overwogen. Daarvoor moet wel een prijs worden betaald. Die prijs is dat het resulterende betrouwbaarheidsinterval groter zal zijn. Er is in feite sprake van een uitruil tussen betrouwbaarheid en nauwkeurigheid: òf er wordt een minder nauwkeurige uitspraak met een grote betrouwbaarheid gedaan, òf een nauwkeurige uitspraak met een minder grote betrouwbaarheid.

De grenzen van het betrouwbaarheidsinterval zijn betrekkelijk eenvoudig te bepalen. Met gebruikmaking van de Centrale Limietstelling kan worden aangetoond dat het steekproefgemiddelde in een enkelvoudige aselecte steekproef zonder teruglegging voor niet te kleine steekproeven bij benadering een normale verdeling heeft. Het midden van het betrouwbaarheidsinterval interval is dan het steekproefgemiddelde. Daarbij wordt een bepaalde marge M opgeteld voor de bovengrens, en aftrekken van de marge geeft de ondergrens. Die marge is gelijk aan de standaardfout van de schatter, vermenigvuldigd met een constante. Voor een betrouwbaarheid van 0,95 is deze constante gelijk aan 1,96, zodat het betrouwbaarheidsinterval dan gelijk is aan:

$$\left(\bar{y} - 1{,}96S(\bar{y}), \bar{y} + 1{,}96S(\bar{y})\right) \qquad (3.2.11)$$

In de praktijk is de standaardfout niet bekend. Daarom wordt deze grootheid in formule (3.2.11) vervangen door de schatter van de standaardfout.

3.3. Het bepalen van de steekproefomvang

Tijdens het opzetten van een enquête komt op een gegeven moment onherroepelijk de vraag naar voren hoe groot de steekproef moet zijn. Dat is een belangrijke beslissing. Immers, als de steekproef groter wordt genomen dan echt noodzakelijk, dan wordt veel tijd en geld verkwist. Wordt de steekproef te klein genomen, dan worden de schatters minder nauwkeurig. Het vaststellen van de steekproefomvang is geen eenvoudige zaak. Er is geen regel die zegt hoe groot een steekproef moet zijn.

Al eerder is gebleken dat er een verband bestaat tussen de omvang van de steekproef en de nauwkeurigheid van de schatting. Hoe groter de steekproef is, des te nauwkeuriger de schatting zal zijn. Daarom kan de vraag naar de steekproefomvang eigenlijk pas worden beantwoord als duidelijk is welke nauwkeurigheid is vereist. Dat is dan ook de procedure die meestal wordt gevolgd. Eerst wordt vastgesteld welke nauwkeurigheid is gewenst. Vervolgens kan dan worden berekend welke steekproefomvang nodig is voor die nauwkeurigheid.

Uitgaande van aselecte steekproeven met gelijke kansen en zonder teruglegging zullen hieronder enkele formules worden afgeleid voor het vaststellen van de steekproefomvang. Eerst wordt dat gedaan voor het schatten van percentages, en daarna voor het schatten van het gemiddelde van een kwantitatieve variabele.

Uitgangspunt bij deze berekeningen is dat de onderzoeker aangeeft hoe groot de marge M in de schatting maximaal mag zijn. Deze marge is gelijk aan het verschil tussen de schatting en de bovengrens of ondergrens van het betrouwbaarheidsinterval. De formules geven dan aan welke steekproefomvang minimaal nodig is om deze nauwkeurigheid te bereiken. Voor een 95%-betrouwbaarheidsinterval is de marge gelijk aan

$$M = 1{,}96\, S(\bar{y}) \tag{3.3.1}$$

en voor een 99%-betrouwbaarheidsinterval moet de waarde 1,96 worden vervangen door 2,58.

Laat M_{max} de waarde voorstellen van de maximale marge die de onderzoeker wil toelaten. Dan laat deze voorwaarde zich vertalen in

$$S(\bar{y}) \leq \frac{M_{max}}{1{,}96}. \tag{3.3.2}$$

Voor het schatten van een percentage wordt de variantie van de schatter gegeven door formule (3.2.7). Invullen in (3.6.2) leidt tot de voorwaarde

$$\sqrt{\frac{1-f}{n}\frac{N}{N-1}P(100-P)} \leq \frac{M_{max}}{1{,}96}. \tag{3.3.3}$$

Door n op te lossen uit deze ongelijkheid wordt een ondergrens gevonden voor de omvang van de steekproef. Dat kan echter niet zonder meer, want de voorwaarde (3.3.3) bevat een onbekende grootheid en dat is het te schatten populatiepercentage P. Om dit probleem op te lossen kunnen een tweetal wegen worden ingeslagen:

- Er is een globale indicatie van P bekend. Die indicatie kan afkomstig zijn uit eerder onderzoek, maar het kan ook zijn dat een inhoudelijk deskundige een dergelijke indicatie kan verschaffen. Die waarde kan dan worden ingevuld voor P, en vervolgens kan n worden opgelost uit (3.3.3).

- Er is helemaal niets bekend over de waarde van P. Dan kan worden geconstateerd dat het gedeelte P(100-P) een bergparabool beschrijft. In het interval tussen 0 en 100 neemt deze parabool minimum waarden aan voor P = 0 en P = 100. Precies daar tussenin, voor P = 50, neemt de parabool zijn maximum waarde aan. Dat betekent dat een bovengrens kan worden bepaald voor de variantie door in formule voor P de waarde 50 in te vullen. Wordt nu n zo bepaald dat deze maximale variantie niet wordt overschreden, dan is de werkelijk

variantie zeker kleiner. Overigens is het ook zo dat voor waarden van P in het interval van 30% tot 70% de werkelijke variantie niet veel zal afwijken van de maximale variantie.

Oplossen van n uit de ongelijkheid (3.3.3) leidt tot de volgende ondergrens voor n:

$$n \geq \frac{1}{\frac{N-1}{N}\left(\frac{M_{max}}{1{,}96}\right)^2 \frac{1}{P(100-P)} + \frac{1}{N}}. \qquad (3.3.4)$$

Een wat eenvoudiger benadering wordt verkregen als de omvang N van de populatie groot is. Dan kan de term (N-1) / N vervangen worden door 1, en de term 1 / N kan worden weggelaten. Uitdrukking (3.3.4) reduceert dan tot

$$n \geq \left(\frac{1{,}96}{M_{max}}\right)^2 P(100-P). \qquad (3.3.5)$$

We illustreren deze formule aan de hand van een voorbeeld. Stel dat uit eerder onderzoek is gebleken dat 38% van de kiezers gaat stemmen op een bepaalde partij. In een nieuw onderzoek wil men nagaan hoe de partij er nu voorstaat. De verwachting is dat de verschillen niet erg groot zullen zijn. Het is daarom niet onredelijk om voor P de waarde 38 in te vullen in formule (3.3.5). Verder wil men de steekproef zo groot hebben dat de maximale afwijking in de steekproef niet groter is dan 3 procentpunten. Dus moet voor M_{max} de waarde 3 worden ingevuld in formule (3.3.5). Dit leidt tot de uitkomst

$$n \geq \left(\frac{1{,}96}{3}\right)^2 38 \times 62 = 1005{,}6.$$

De steekproefomvang moet dus minstens gelijk zijn aan 1006. Zou men een uitspraak met een betrouwbaarheid van 99% in plaats van 95% willen hebben, dan moet 1,96 worden vervangen door 2,58, en wordt de steekproefomvang minimaal 1689.

Gaat het in het onderzoek niet om het schatten van de aan- of afwezigheid van een kenmerk, maar om het gemiddelde van de doelvariabele, dan is het uitgangspunt ook weer formule (3.3.2). Alleen is er nu geen simpele uitdrukking beschikbaar voor de standaardfout. Formule (3.3.2) wordt nu vertaald in

$$\sqrt{\left(\frac{1}{n} - \frac{1}{N}\right)S^2} \leq \frac{M_{max}}{1{,}96}, \qquad (3.3.6)$$

waarin S^2 de populatievariantie is. Het probleem is nu dat die populatievariantie in veel gevallen niet bekend is. Soms kan een schatting worden gemaakt op grond van voorgaand onderzoek, of misschien is er een indicatie van de waarde uit een proefonderzoek.. Dan kan die waarde worden ingevuld in (3.6.6) en leidt omwerken tot

$$n \geq \frac{1}{\left(\frac{M_{max}}{1{,}96S}\right)^2 + \frac{1}{N}} \qquad (3.3.7)$$

Voor grote waarden van N kan de formule worden vereenvoudigd tot

$$n \geq \left(\frac{1{,}96S}{M_{max}}\right)^2 . \qquad (3.3.8)$$

Als er totaal geen indicatie is voor de waarde van S, dan wordt het wat lastiger. De volgende vuistregels zouden wat uitkomst kunnen bieden:

- De waarden zijn min of meer normaal verdeeld over een interval ter lengte L Dan zal L ongeveer gelijk zijn aan 6S, en kan voor S dus de waarde L/6 worden ingevuld.
- De waarden zijn gelijkmatig (uniform) verdeeld over een interval ter lengte L. Dan zal S ongeveer gelijk zijn aan $L/\sqrt{12} = 0{,}29$ L.
- De waarden volgen een exponentiële verdeling, met heel veel kleine waarden en heel weinig grote waarden, en het overgrote deel van de waarden (zeg zo'n 99%) ligt in een interval ter lengte L. Dan zal S ongeveer gelijk zijn aan 0,22 L.
- De 'worst case', de verdeling met de grootste variantie, is die waarbij de waarden in een interval ter lengte L liggen, en wel zo dat de ene helft van de waarden ligt bij het linker uiteinde, en de andere helft bij het rechter uiteinde van het interval. In dit geval zal S ongeveer gelijk zijn aan 0,50 L.

4. Betere trekkingsprocedures

4.1. Gebruik van hulpinformatie

In het voorgaande hoofdstuk is beschreven hoe op basis van een aselecte steekproef een schatting worden gemaakt van het populatiegemiddelde. De schatter hiervoor was het steekproefgemiddelde. Dit is een zuivere schatter. Verder was aan de formule van de variantie

van deze schatter te zien, dat nauwkeuriger schattingen konden worden gemaakt door een grotere steekproef te trekken.

Er zijn ook nog andere methoden om de nauwkeurigheid van de schattingen te verbeteren. Enkele van deze methoden worden in dit en het volgende hoofdstuk besproken. Essentieel bij deze methoden is dat gebruik wordt gemaakt van extra informatie. Die extra informatie moet aanwezig zijn in de vorm van hulpvariabelen. Dat zijn variabelen die in de steekproef zijn gemeten, maar waarover ook informatie op het niveau van de populatie bekend is. Sommige methoden vereisen slechts dat het populatiegemiddelde van de hulpvariabele bekend is, maar andere methoden kunnen alleen worden toegepast als de waarden van de hulpvariabele voor elke element in de doelpopulatie bekend is.

De toepassingsmogelijkheden van een hulpvariabele worden voor een deel bepaald door het meetniveau ervan. We onderscheiden hier twee meetniveaus.

- Een *kwalitatieve hulpvariabele* neemt slechts een eindig aantal mogelijke waarden aan. Elke mogelijk waarde is niets anders dan een etiket voor de groep waarin de waarneming zit. Er mag dan ook niet met deze waarden worden gerekend. Voorbeelden van kwalitatieve variabelen zijn geslacht, burgerlijke staat, provincie en kerkgenootschap.
- Een *kwantitatieve hulpvariabele* kan binnen een bepaald gebied elke mogelijke waarde aannemen. Die waarden duiden meestal een hoeveelheid, omvang, waarde of grootte aan. Met deze waarden mag dan ook worden gerekend. Grootheden als gemiddelde en variantie zijn hier zinvol. Voorbeelden zijn leeftijd, lengte, gewicht en inkomen.

Hulpvariabelen kunnen op twee manieren worden toegepast voor het maken van nauwkeuriger uitspraken over de populatie. In de eerste plaats kunnen hulpvariabelen worden gebruikt voor een andere wijze van trekken van de elementen uit de steekproef. Het hangt van het meetniveau van de hulpvariabele af hoe dit in zijn werk gaat. Voor een kwalitatieve variabele kan een *gestratificeerde steekproef* worden getroffen en voor een kwantitatieve hulpvariabele kan een steekproef met *ongelijke kansen* worden getrokken. Deze twee steekproefontwerpen worden in het vervolg van dit hoofdstuk behandeld.

In de tweede plaats kunnen de hulpvariabelen worden gebruikt om bij een simpele aselecte steekproef een verbeterde schattingsprocedure te gebruiken. Hieraan is hoofdstuk 5 gewijd. Voor kwantitatieve hulpvariabelen zullen twee zulke schatters worden behandeld. Dat zijn de

quotiëntschatter en de *regressieschatter*. Voor kwalitatieve hulpvariabelen zal de *post-stratificatie-schatter* worden behandeld.

Welke methode ook wordt gehanteerd om de nauwkeurigheid van de uitkomsten van het onderzoek te verbeteren, het is niet zo dat elke hulpvariabele zonder meer kan worden gebruikt. De nauwkeurigheid zal met name beter worden naarmate de samenhang tussen doelvariabele en hulpvariabele groter is. Bij het vinden van geschikte hulpvariabele moet dus enerzijds worden gekeken of er populatie-informatie voor deze hulpvariabele beschikbaar is, en ook of er verband bestaat tussen doelvariabele en hulpvariabele.

Voor een kwantitatieve hulpvariabele betekent samenhang tussen doelvariabele en hulpvariabele dat de correlatie tussen beide variabelen hoog is. Voor een kwalitatieve hulpvariabele betekent samenhang dat de hulpvariabele de populatie indeelt in groepen waarbinnen de doelvariabele weinig varieert. Die deelpopulaties moeten dus homogeen zijn met betrekking tot de doelvariabele.

4.2. De gestratificeerde steekproef

Bij een gestratificeerde steekproef wordt een kwalitatieve hulpvariabele gebruikt om een *stratificatie* aan te brengen in de populatie. Strata zijn deelpopulaties die corresponderen met de verschillende categorieën van de hulpvariabele. Vervolgens wordt uit elk stratum apart een steekproef getrokken. De manier waarop de steekproef in elk stratum wordt getrokken is geheel vrij, zolang die steekproef maar een schatting oplevert voor de waarde van de populatiegrootheid in het betreffende stratum. Tenslotte worden dan de schattingen voor de strata gecombineerd tot een schatting voor de gehele populatie.

De flexibiliteit van de gestratificeerde steekproef biedt in allerlei opzichten interessante mogelijkheden. Een voorbeeld is de situatie waarin men niet alleen uitspraken wil doen over de gehele populatie, maar ook over bepaalde deelpopulaties. Denk hierbij aan een regio of een bedrijfstak. Bij het opstellen van het steekproefontwerp kan rekening worden gehouden met de nauwkeurigheid die is vereist voor de schattingen in die strata. Dat kan door de steekproefomvang in de betreffende strata aan te passen.

Stratificatie kan alleen worden uitgevoerd als van te voren voor elk element in de populatie bekend is in welk stratum het thuishoort. Aangezien uit elk stratum een aparte steekproef wordt getrokken, moet voor elk stratum apart een steekproefkader beschikbaar zijn. Als,

bijvoorbeeld, bij het trekken van een steekproef uit de bevolking van een stad wordt gestratificeerd naar wijk, dan moet voor elke wijk apart een steekproef worden getrokken uit de bevolking van die wijk. Het ontbreken van aparte steekproefkaders per stratum kan een struikelblok vormen voor het trekken van een gestratificeerde steekproef.

Deze paragraaf beperkt zich tot gestratificeerde steekproeven waarbij uit elk stratum een aselecte steekproef zonder teruglegging wordt getrokken. Eigenlijk biedt dan stratificatie niet zoveel nieuws. In plaats van één aselecte steekproef, worden een aantal aselecte steekproeven getrokken. Het nieuwe is alleen de wijze waarop stratumschattingen worden gecombineerd tot een populatieschatting.

Om de formules voor de schatter en de variantie van de schatter te kunnen opschrijven, wordt de notatie enigszins aangepast. Allerlei grootheden krijgen een extra index die het stratum aanduidt waarop die grootheid betrekking heeft.

Stel dat de populatie U wordt verdeeld in L strata. Die strata worden aangegeven met

$$U_1, U_2, ..., U_L \tag{4.2.1}$$

De strata hebben geen overlap met elkaar en vormen samen met elkaar precies de populatie U. De omvang van stratum h wordt aangegeven met N_h (voor h = 1, 2, ..., L). Dan geldt

$$\sum_{h=1}^{L} N_h = N_1 + N_2 + ... + N_L = N \tag{4.2.2}$$

De N_h waarden voor de doelvariabele in stratum h worden genoteerd met

$$Y_1^{(h)}, Y_2^{(h)}, ..., Y_{N_h}^{(h)} \tag{4.2.3}$$

Als het gemiddelde van de doelvariabele in stratum h wordt aangeduid met

$$\overline{Y}^{(h)} = \frac{1}{N_h} \sum_{k=1}^{N_h} Y_k^{(h)}, \tag{4.2.4}$$

dan kan het populatiegemiddelde worden geschreven als

$$\overline{Y} = \frac{1}{N} \sum_{h=1}^{L} N_h \overline{Y}^{(h)} \tag{4.2.5}$$

Het populatiegemiddelde is dus gelijk aan de gewogen som van de stratumgemiddelden. De (aangepaste) variantie in stratum h kan worden geschreven als

$$S_h^2 = \frac{1}{N_h - 1} \sum_{k=1}^{N_h} \left(Y_k^{(h)} - \overline{Y}^{(h)} \right)^2 \qquad (4.2.6)$$

Uit de aldus gestratificeerde populatie wordt een steekproef van omvang n getrokken. Die steekproef wordt gerealiseerd door het trekken van L substeekproeven, met respectievelijke omvangen n_1, n_2, ..., n_L, en wel zodanig dat $n_1 + n_2 + ... + n_L = n$. Elke substeekproef is een aselecte steekproef met gelijke kansen. De n_h waarnemingen die zo in stratum h beschikbaar komen, worden aangegeven met

$$y_1^{(h)}, y_2^{(h)}, ..., y_{n_h}^{(h)} \qquad (4.2.7)$$

Het steekproefgemiddelde in stratum h is gelijk aan

$$\overline{y}^{(h)} = \frac{1}{n_h} \sum_{i=1}^{n_h} y_i^{(h)} \qquad (4.2.8)$$

Dit steekproefgemiddelde in stratum h is een zuivere schatter voor het populatiegemiddelde in stratum h. De variantie van de schatter is gelijk aan

$$V(\overline{y}^{(h)}) = \frac{1 - f_h}{n_h} S_h^2 \qquad (4.2.9)$$

waarin $f_h = n_h / N_h$. Deze variantie kan zuiver worden geschat met

$$v(\overline{y}^{(h)}) = \frac{1 - f_h}{n_h} s_h^2, \qquad (4.2.10)$$

waarin

$$s_h^2 = \frac{1}{n_h - 1} \sum_{i=1}^{n_h} \left(y_i^{(h)} - \overline{y}^{(h)} \right)^2 \qquad (4.2.11)$$

Met behulp van de schatters voor de stratumgemiddelden kan nu een schatter voor het populatiegemiddelde worden gemaakt. De formule hiervoor luidt

$$\overline{y}_S = \frac{1}{N} \sum_{h=1}^{L} N_h \overline{y}^{(h)} \qquad (4.2.12)$$

Dit is een zuivere schatter voor het populatiegemiddelde. De schatter is dus gelijk aan het gewogen gemiddelde van de schatters voor de strata. De variantie van schatter (4.2.12) is gelijk aan

$$V(\bar{y}_S) = \frac{1}{N^2}\sum_{h=1}^{L} N_h^2 \frac{1-f_h}{n_h} S_h^2 \qquad (4.2.13)$$

Deze variantie kan zuiver worden geschat door

$$v(\bar{y}_S) = \frac{1}{N^2}\sum_{h=1}^{L} N_h^2 \frac{1-f_h}{n_h} s_h^2 \qquad (4.2.14)$$

Nadere beschouwing van formule (4.2.13) leert dat de variantie van de stratificatie-schatter klein is als de stratumvarianties klein zijn. Dit is het geval als de strata homogeen zijn. Hiermee wordt bedoeld dat de doelvariabele binnen de strata erg weinig varieert. De variantie van de doelvariabele zit dan met name tussen de strata. De variantie uit zich in niveauverschillen tussen de strata. Kortom, als de stratificatie zo wordt gekozen dat de strata homogeen zijn, dan wordt een nauwkeurige schatter verkregen.

Een belangrijk aspect bij het trekken van een gestratificeerde steekproef is het vaststellen van de *allocatie*. De verdeling van de totale steekproefomvang n over de L strata. Uitgangspunt zou kunnen zijn dat de stratumgemiddelden met een zekere nauwkeurigheid moeten worden geschat. Dat legt de steekproefomvang in elk stratum vast, en dus ook de totale steekproefomvang. Vaker echter ligt de totale steekproefomvang n al van tevoren vast, en gaat het om de verdeling van deze n over de strata.

Als het er om gaat de schatting van het populatiegemiddelde zo nauwkeurig mogelijk te krijgen, dan kan worden aangetoond dat de beste schatter wordt verkregen met *de optimale allocatie*. Deze allocatie wordt ook wel de *Neyman-allocatie* genoemd. Deze allocatie schrijft voor dat de steekproefomvang n_h in stratum h gelijk moet worden genomen aan

$$n_h = n\frac{N_h S_h}{\sum_{j=1}^{L} N_j S_j}. \qquad (4.2.15)$$

Daarbij moet n_h zonodig worden afgerond op een gehele waarde. De steekproefomvang in een stratum zal groter zijn naarmate de stratumvariantie groter is, dus naarmate het stratum minder homogeen is. In de praktijk kan het voorkomen dat de met formule (4.2.15) bepaalde steekproefomvang groter is dan de omvang van het stratum. In dit geval moet het stratum gewoon integraal worden waargenomen, waarna voor de resterende strata opnieuw de optimale allocatie wordt uitgerekend.

Bij stratificatie is de kans op selectie van een element in stratum h gelijk aan n_h / N_h. Voor optimale allocatie betekent dit dat de selectiekans in stratum h evenredig is aan S_h. Dat betekent dat niet elke element in de populatie eenzelfde trekkingskans heeft. Dit is geen probleem, want in de formule voor de schatter wordt voor deze ongelijke trekkingskansen gecorrigeerd.

De optimale allocatie kan alleen worden berekend als de stratumvarianties bekend zijn. Wordt stratificatie niet toegepast voor de verhoging van de nauwkeurigheid, maar meer om administratieve redenen (er is geen steekproefkader voor de gehele populatie, maar wel voor elk stratum apart, en het is technisch niet mogelijk deze kaders samen te voegen) of ter verkrijging van schattingen voor deelpopulaties, dan is het soms niet onaannemelijk te veronderstellen dat de stratumvarianties bij benadering gelijk zijn. Formule (4.2.15) reduceert dan tot

$$n_h = n\frac{N_h}{N}. \tag{4.2.16}$$

Een allocatie volgens formule (4.2.16) wordt een *evenredige allocatie* genoemd. Dit is eigenlijk wat de steekproefdeskundigen aan het begin van deze eeuw voor ogen stond toen ze het hadden over representatieve steekproeven. Immers, voor deze allocatie heeft elke element in de populatie dezelfde trekkingskans, en die is gelijk aan n / N. Daarom wordt een gestratificeerde steekproef met evenredige allocatie ook wel een *zelfwegende* steekproef genoemd.

Bij het vaststellen van de allocatie zal het vaak niet alleen gaan om de nauwkeurigheid van de schatter, maar ook om de kosten die dit met zich meebrengt. Het is niet ondenkbaar dat enquêteren in het ene stratum kostbaarder is dan in het andere stratum. Dan zou de optimale allocatie wel eens niet de goedkoopste allocatie kunnen zijn.

Stel eens dat de totale kosten van het veldwerk het bedrag C niet mogen overschrijden. Stel ook dat het enquêteren van een element in stratum h kosten ter grootte van een bedrag c_h met zich meebrengt. Dan moet een allocatie aan de volgende voorwaarde voldoen:

$$C = \sum_{h=1}^{L} c_h n_h . \tag{4.2.17}$$

Het is dan deze voorwaarde die in de plaats komt van de voorwaarde dat de totale steekproefomvang gelijk moet zijn aan n. In feite is deze laatste voorwaarde een speciaal geval van formule (4.2.17), namelijk wanneer de kosten van enquêteren overal even groot zijn. Er kan worden aangetoond dat onder voorwaarde (4.2.17) de nauwkeurigste schatter wordt verkregen als de steekproefomvang in stratum h gelijk is aan

$$n_h = K\frac{N_h S_h}{\sqrt{c_h}}, \tag{4.2.18}$$

waarbij de constante K gelijk is aan:

$$K = \frac{C}{\sum_{h=1}^{L} N_h S_h \sqrt{c_h}}. \tag{4.2.19}$$

Uit formule (4.2.18) kan de voor de hand liggende conclusie worden getrokken dat er relatief minder elementen hoeven te worden geselecteerd in dure strata.

4.3. De steekproef met ongelijke kansen

In de enkelvoudige aselecte steekproef worden de elementen met gelijke kansen getrokken. En ook bij de gestratificeerde steekproef wordt binnen de strata (meestal) met gelijke kansen getrokken. De theorie van Horvitz en Thompson (1952) laat echter zien dat het voordelig kan zijn om met ongelijke kansen te trekken. De variantie van de schatter wordt kleiner naarmate de trekkingskansen meer evenredig worden genomen aan de waarden van de doelvariabele. Bij perfecte evenredigheid wordt de variantie zelfs 0. Deze situatie zal zich echter in de praktijk nooit voordoen. Immers, als de kansen evenredig aan de waarden van de doelvariabele konden worden genomen, dan zouden de waarden van de doelvariabele bekend zijn. Het onderzoek werd juist uitgevoerd om daarachter te komen.

In de praktijk wordt een steekproef met ongelijke kansen gerealiseerd door te zoeken naar een hulpvariabele die een sterke samenhang vertoont met de doelvariabele, en vervolgens de trekkingskansen te baseren op de waarden van die hulpvariabele. Een voorbeeld uit de praktijk is een onderzoek naar aantal en waarde van winkeldiefstallen, waarbij winkels worden getrokken met kansen die evenredig zijn met het vloeroppervlak van de winkel. De idee hierachter is dat in grote winkels meer diefstallen plaatsvinden dan in kleine winkels.

Het blijkt niet eenvoudig te zijn om praktische procedures op te stellen voor het zonder teruglegging trekken van steekproeven met ongelijke kansen. In principe zou kunnen worden getracht de procedures voor steekproeven zonder teruglegging met gelijke kansen te generaliseren, maar dat is niet altijd mogelijk. Daarom beperkt men zich meestal tot steekproeven met teruglegging met ongelijke kansen.

Stel dat een steekproef moet worden getrokken met kansen evenredig aan de waarden X_1, X_2, ..., X_N van een hulpvariabele X. Essentieel bij het trekken met ongelijke kansen is dat alle waarden positief zijn. Er worden nu twee procedures besproken om n elementen te trekken met teruglegging en met kansen evenredig aan deze waarden.

De eerste procedure wordt wel de *Cumulatieve Methode* genoemd. Hiervoor moeten eerst de N subtotalen T_1, T_2, …, T_N worden uitgerekend, waarbij T_k gelijk is aan

$$T_k = \sum_{i=1}^{k} X_i \qquad (4.3.1)$$

Verder wordt $T_0 = 0$ geïntroduceerd. Voor de selectie van een element worden de volgende stappen doorlopen:

- Trek een aselecte waarde t geloot uit het interval $(0, T_N]$;
- Bepaal het interval $(T_{i-1}, T_i]$ waarin deze waarde ligt;
- Selecteer element i in de steekproef.

Bij de tweede procedure, ontwikkeld door Lahiri in 1951, hoeven de subtotalen niet te worden berekend. Hier is het voldoende een bovengrens X_{max} voor de waarden van X_1, X_2, ..., X_N te kennen. Voor de selectie van een element worden de volgende stappen uitgevoerd:

- Trek een aselect nummer k uit de nummers 1 t/m N. Dit is het volgnummer van een kandidaat-element voor de steekproef.
- Trek een aselecte waarde x uit het interval $(0, X_{max}]$.
- Als $x \leq X_k$ (de waarde van de hulpvariabele X voor kandidaat-element k), selecteer dan element k in de steekproef.
- Als $x \geq X_k$, trek dan een nieuwe kandidaat en een nieuwe waarde van u.

Bij de methode van Lahiri moet veel vaker gebruik worden gemaakt van de aselecte getallen generator. Per kandidaat-element is dat in ieder geval twee keer. Regelmatig zullen

kandidaten worden verworpen, maar dit verwerpen gebeurt minder vaak naarmate X_{max} dichter in de buurt van het werkelijke maximum van $X_1, X_2, ..., X_N$ ligt.

Om het populatiegemiddelde te schatten kan geen gebruik worden gemaakt van het steekproefgemiddelde. In de schatter moet immers worden gecorrigeerd voor het feit dat sommige elementen grotere kansen hebben, en dus oververtegenwoordigd zijn in de steekproef. Die correctie vindt plaats door bij de berekening van de schatter de gemeten waarden van de doelvariabele te delen door de bijbehorende waarde van de trekkingskans. Het komt erop neer dat voor π_k de waarde X_k/NX_T wordt ingevuld (waarin X_T het populatietotaal is van de hulpvariabele) Om de formule voor de schatter wat eenvoudiger te kunnen opschrijven, wordt eerst de notatie

$$z_i = \frac{y_i}{x_i} \tag{4.3.2}$$

ingevoerd. De schatter voor het populatiegemiddelde van de doelvariabele wordt dan

$$\bar{y}_{OK} = \bar{X}\bar{z}. \tag{4.3.3}$$

Het steekproefgemiddelde van de waarden van de z_i wordt dus vermenigvuldigd met het populatiegemiddelde van de hulpvariabele.

De variantie van de schatter (4.3.3) kan worden geschreven als:

$$V(\bar{y}_{OK}) = \frac{\bar{X}}{Nn}\sum_{k=1}^{N} X_k \left(\frac{Y_k}{X_k} - \frac{\bar{Y}}{\bar{X}}\right)^2 \tag{4.3.4}$$

Aan deze formule is te zien dat als Y_k evenredig is aan X_k, de quotiënten Y_k / X_k constant zijn en dus ook gelijk aan het quotiënt van het gemiddelde van Y en het gemiddelde van X, zodat de variantie dan 0 is. Een dergelijke perfecte evenredigheid zal in de praktijk niet vaak voorkomen. Maar ook evenredigheid die slechts bij benadering geldt, helpt al om de variantie te verkleinen.

In de praktijk moet de variantie in (4.3.4) worden geschat op basis van de steekproefgegevens. Dat gaat met de volgende formule:

$$v(\bar{y}_{OK}) = \frac{\bar{X}^2}{n(n-1)}\sum_{i=1}^{n}(z_i - \bar{z})^2 \tag{4.3.5}$$

5. Andere schatters

5.1. De quotiëntschatter

Uitgangspunt van de *quotiëntschatter* is dat er een kwantitatieve hulpvariabele X is waarvan de waarden min of meer evenredig zijn met de waarden van de doelvariabele Y. Zouden we de waarden van beide variabelen tegen elkaar uitzetten in een grafiek, dan betekent dit dat de punten (X_1,Y_1), (X_2,Y_2),..., (X_N,Y_N) bij benadering op een rechte lijn liggen, en die rechte lijn moet ongeveer door de oorsprong gaan. Uit de populatie wordt een enkelvoudige aselecte steekproef getrokken.

Zowel de doelvariabele Y als de hulpvariabele X moeten kwantitatieve variabelen zijn. Beide variabelen moeten zijn gemeten in de steekproef en van de hulpvariabele X met het populatiegemiddelde bekend zijn.

In deze situatie kan de quotiëntschatter worden toegepast. De quotiëntschatter is als volgt gedefinieerd:

$$\bar{y}_Q = \bar{y}\frac{\overline{X}}{\bar{x}} \tag{5.1.1}$$

De quotiëntschatter is in feite gelijk aan het eenvoudige steekproefgemiddelde van de doelvariabele vermenigvuldigd met een correctiefactor. Die correctiefactor is het quotiënt van het populatiegemiddelde en het steekproefgemiddelde van de hulpvariabele. Zou, bijvoorbeeld, het steekproefgemiddelde van X lager zijn dan het populatiegemiddelde, dan vindt de quotiëntschatter, dat het steekproefgemiddelde van de doelvariabele te laag is uitgevallen, en maakt hem wat groter.

De quotiëntschatter is geen zuivere schatter. Hij heeft een kleine onzuiverheid, maar deze is voor een flinke steekproefomvang verwaarloosbaar. De variantie van de schatter kan niet exact worden uitgerekend. Meestal wordt gebruik gemaakt van een benadering die alleen geld voor een flinke steekproefomvang. Die benadering is:

$$V(\bar{y}_Q) = \frac{1-f}{n}\frac{1}{N-1}\sum_{i=1}^{N}(Y_i - \frac{\overline{Y}}{\overline{X}}X_i)^2 \tag{5.1.2}$$

De berekening van deze variantie vereist kennis van alle waarden van Y en X in de populatie, en die is niet aanwezig. Dus wordt deze variantie (net als de variantie van het steekproefgemiddelde) geschat op grond van de steekproefgegevens. De formule hiervoor luidt:

$$v(\bar{y}_Q) = \frac{1-f}{n}\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \frac{\bar{y}}{\bar{x}}x_i)^2 \qquad (5.1.3)$$

Naarmate de waarden van Y en X beter evenredig zijn, dus naarmate alle quotiënten Y_i/X_i meer op elkaar lijken, is de variantie in (5.1.2) kleiner, en dus de schatter nauwkeuriger.

5.2. De regressieschatter

De quotiëntschatter stelt ons in staat om aanzienlijk nauwkeuriger te schatten, maar het kan nog beter. Daarvoor moet gebruik worden gemaakt van de *regressieschatter*. De berekeningen voor deze schatter zijn wat omvangrijker, maar de beloning daarvoor is een nauwkeuriger schatting.

De regressieschatter kan in exact dezelfde situatie worden gebruikt als de quotiëntschatter. Zowel de doelvariabele Y als de hulpvariabele X moeten kwantitatieve variabelen zijn. Beide variabelen moeten zijn gemeten in de enkelvoudige aselecte steekproef en van de hulpvariabele X met het populatiegemiddelde bekend zijn.

De regressieschatter is gedefinieerd als

$$\bar{y}_R = \bar{y} - b(\bar{x} - \bar{X}) \qquad (5.2.1)$$

Hierin is b de richtingscoëfficiënt van de lijn die zo goed mogelijk door de puntenwolk gaat die we krijgen door de doelvariabele af te zetten op de y-as en de hulpvariabele op de x-as. De punten in de puntenwolk zijn de waarnemingsparen (y_i, x_i) die beschikbaar komen in de steekproef. De formule voor de berekening van b is gebaseerd op het Kleinste Kwadratencriterium:

$$b = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}. \qquad (5.2.2)$$

De regressieschatter is niet helemaal zuiver, maar ook hier geldt weer dat die onzuiverheid voor grotere steekproeven kan worden verwaarloosd.

Net zoals de quotiëntschatter kan de regressieschatter worden gezien als een correctie op het steekproefgemiddelde. Nu wordt echter gekeken naar het verschil van populatiegemiddelde en

steekproefgemiddelde van de hulpvariabele in plaats van naar het quotiënt. De coëfficiënt b zorgt voor de juiste schaling.

De variantie van de regressieschatter is gelijk aan

$$V(\bar{y}) = \frac{1-f}{n} S^2 (1 - R^2). \qquad (5.2.3)$$

Hierin duidt R de *productmoment correlatiecoëfficiënt* aan. Deze grootheid is gedefinieerd als

$$R = \frac{\sum_{i=1}^{N} (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{N} (X_i - \bar{X})^2 \sum_{i=1}^{N} (Y_i - \bar{Y})^2}} \qquad (5.2.4)$$

De correlatiecoëfficiënt is een maat voor de samenhang tussen twee variabelen. De waarde ligt altijd in het interval [-1, 1]. Een waarde van 1 betekent een perfecte samenhang die kan worden weergegeven door een stijgende lijn in de puntenwolk. Een waarde van -1 betekent ook een perfecte samenhang, maar dan in de vorm van een dalende lijn. Is de waarde 0, dan is er in het geheel geen samenhang.

Bekijken we nu nog eens de formule van de variantie van de regressieschatter, dan zien we dat die variantie gelijk is aan de variantie van het steekproefgemiddelde (zie formule (3.2.2)), maar dan vermenigvuldigd met de factor $(1 - R^2)$. Aangezien R ligt in het interval [-1, 1], ligt R^2 in het interval [0, 1], en dus is $(1 - R^2)$ altijd kleiner dan of gelijk aan 1. Dus kunnen we concluderen dat de regressieschatter een variantie heeft die nooit groter is dan die van het steekproefgemiddelde. Alleen in de situatie R = 0 (totaal geen samenhang) zijn de varianties gelijk. Verder is de regressieschatter altijd nauwkeuriger dan het steekproefgemiddelde. Zodra er sprake is van enige samenhang, verdient de regressieschatter dus de voorkeur uit het oogpunt van nauwkeurigheid.

In de voorafgaande paragrafen is steeds gebruik gemaakt van een kwantitatieve hulpvariabele voor het verbeteren van de schatter. Een kwantitatieve variabele is een variabele waarmee echt kan worden gerekend. Zo kan bijvoorbeeld het gemiddelde van zijn waarden worden uitgerekend Er zijn echter ook kwalitatieve variabelen. Dat zijn variabelen die elementen indelen in categorieën. Voorbeelden van een kwalitatieve variabelen zijn het geslacht van de respondent (2 categorieën), de provincie waarin de respondent woont (12 categorieën) of de burgerlijke staat van de respondent (4 categorieën: nog nooit gehuwd geweest, gehuwd,

gescheiden en weduwstaat). Omdat met dit type hulpvariabelen niet kan worden gerekend, kunnen ze ook niet worden gebruikt voor de quotiëntschatter of de regressieschatter.

5.3. Post-stratificatie

Er is een schattingstechniek die met name is bedoeld voor kwalitatieve hulpvariabelen, en dat is post-stratificatie. Merk op dat al eerder (in paragraaf 4.2) de gestratificeerde steekproef is behandeld. Dit is echter een heel andere techniek dan post-stratificatie. Bij de gestratificeerde steekproef wordt een kwalitatieve variabele gebruikt om de trekkingsprocedure te verbeteren. Post-stratificatie wordt in combinatie met een aselecte steekproef gebruikt om de schattingsprocedure te verbeteren.

De kwalitatieve variabele verdeelt de populatie in strata. Bij post-stratificatie wordt nu gekeken welke waarnemingen in welk stratum terecht zijn gekomen. De aantallen in de strata worden volledig door het toeval bepaald. Immers, van te voren is bij de steekproeftrekking geen rekening gehouden met een verdeling van de populatie in strata.

Stel de kwalitatieve variabele verdeelt de populatie in L strata. Laat n_h het aantal waarnemingen in stratum h voorstellen (voor h = 1, 2, ..., L). Aangezien de waarden van deze grootheden door het toeval worden bepaald, zijn het stochastische variabelen. Zij verder

$$\bar{y}^{(h)} \tag{5.3.1}$$

het gemiddelde van de waarnemingen aan de doelvariabele in stratum h. Dan is de post-stratificatie-schatter voor het populatiegemiddelde van de doelvariabele gedefinieerd als

$$\bar{y}_{PS} = \frac{1}{N}\sum_{h=1}^{L} N_h \bar{y}^{(h)} . \tag{5.3.2}$$

Om de schatter te kunnen uitrekenen, moet de kwalitatieve hulpvariabele in de steekproef zijn gemeten. Bovendien moet de populatieomvang per stratum bekend zijn. Dat is de hulpinformatie die bekend wordt verondersteld. De formule van de schatter in (5.3.2.) is hetzelfde als de formule van de schatter van de gestratificeerde steekproef in (4.2.12). De eigenschappen van de twee schatters zijn echter wel verschillend, en dat komt tot uiting in de formules voor de varianties.

De post-stratificatie-schatter kan dus worden gezien als een gewogen gemiddelde van de schatters voor de stratumgemiddelden. Onder de voorwaarde dat alle strata minstens één

waarneming bevatten is de post-stratificatie-schatter een zuivere schatter voor het populatiegemiddelde. De variantie van de schatter valt niet op eenvoudige wijze te berekenen. Voor grotere steekproeven kan hij worden benaderd met

$$V(\bar{y}_{PS}) = \frac{1-f}{n}\sum_{h=1}^{L} W_h S_h^2 + \frac{1}{n^2}\sum_{h=1}^{L}(1-W_h)S_h^2 \qquad (5.3.3)$$

Hierin is S_h^2 de populatievariantie in stratum h, en $W_h = N_h / N$. De variantie van de schatter zal met name klein zijn als de strata homogeen zijn met betrekking tot de doelvariabele. Hiermee wordt bedoeld dat de variatie in de waarden van de doelvariabele zich vooral manifesteert in niveauverschillen tussen de strata, terwijl de waarden binnen de strata relatief weinig variëren. In deze situatie zijn de stratumvarianties klein, en dus ook de variantie van de schatter.

Variantie (5.3.3) kan op basis van de steekproefgegevens worden geschat met de formule

$$v(\bar{y}_{PS}) = \frac{1-f}{n}\sum_{h=1}^{L} W_h s_h^2 + \frac{1}{n^2}\sum_{h=1}^{L}(1-W_h)s_h^2 \qquad (5.3.4)$$

6. Conclusie

In deze bijdrage is een overzicht gegeven van een aantal belangrijke aspecten van de steekproeftheorie. Het overzicht is echter bij lange na niet compleet. Voor meer methoden en meer details wordt verwezen naar de uitgebreide literatuur op dit gebied.

Het standaardwerk dat door veel steekproefdeskundigen in de praktijk wordt toegepast, is Cochran (1977). Voor meer theoretische diepgang kan worden verwezen naar Särndal et al. (1992).

Literatuur

Cochran, W.C., 1977, Sampling techniques, 3rd edition (Wiley, New York).

Horvitz, D.G., en D.J. Thompson, 1952, A generalization of sampling without replacement from a finite universe. Journal of the American Statistical Association 47, blz. 663-685.

Särndal, C-E., Swensson, B. en J. Wretman, 1992, Model Assisted Survey Sampling, Springer-Verlag, New York.