Steekproeftheorie

Steekproefontwerpen en Ophoogmethoden

10

Reinder Banning, Astrea Camstra en Paul Knottnerus

Statistische Methoden (10009)

Centraal Bureau voor de Statistiek

Den Haag/Heerlen, 2010

Verklaring van tekens

.	=	gegevens ontbreken
*	=	voorlopig cijfer
**	=	nader voorlopig cijfer
x	=	geheim
–	=	nihil
–	=	(indien voorkomend tussen twee getallen) tot en met
0 (0,0)	=	het getal is kleiner dan de helft van de gekozen eenheid
niets (blank)	=	een cijfer kan op logische gronden niet voorkomen
2008–2009	=	2008 tot en met 2009
2008/2009	=	het gemiddelde over de jaren 2008 tot en met 2009
2008/'09	=	oogstjaar, boekjaar, schooljaar enz., beginnend in 2008 en eindigend in 2009
2006/'07–2008/'09	=	oogstjaar, boekjaar enz., 2006/'07 tot en met 2008/'09

In geval van afronding kan het voorkomen dat het weergegeven totaal niet overeenstemt met de som van de getallen.

Colofon

Uitgever
Centraal Bureau voor de Statistiek
Henri Faasdreef 312
2492 JP Den Haag

Prepress
Centraal Bureau voor de Statistiek - Grafimedia

Omslag
TelDesign, Rotterdam

Inlichtingen
Tel. (088) 570 70 70
Fax (070) 337 59 94
Via contactformulier: www.cbs.nl/infoservice

Bestellingen
E-mail: verkoop@cbs.nl
Fax (045) 570 62 68

Internet
www.cbs.nl

ISSN: 1876-0333

6016510009 X-37

Inhoudsopgave

1. Inleiding op de deelthema's

1.1 Algemene beschrijving

In dit document gaan we nader in op de verschillende steekproefontwerpen en ophoogmethoden die bij het CBS in gebruik zijn. Het gaat hierbij vooral om zogenaamde kanssteekproeven. Dat wil zeggen, steekproeven waarbij de selectie van de eenheden uit de populatie plaatsvindt in overeenstemming met een welomschreven selectiemechanisme.

De eenvoudigste en meest bekende kanssteekproef is een enkelvoudige, aselecte steekproef zonder teruglegging. Hierbij heeft elke mogelijke steekproef (van dezelfde omvang) uit de populatie exact dezelfde trekkingskans. Enkelvoudige, aselecte trekking is de basisselectiemethode en alle andere kanssteekproeven kunnen worden gezien als een uitbreiding of aanpassing hiervan. Deze complexere ontwerpen hebben bijna altijd als doel de efficiëntie te verhogen ofwel de schattingen nauwkeuriger te maken, maar ook praktische en/of economische redenen spelen een rol bij de keus voor een ontwerp. Voor al deze steekproeven geldt echter, dat elk element van de (doel)populatie een positieve en bekende kans moet hebben om in de steekproef terecht te komen. Die laatste kans wordt ook wel de insluitkans genoemd.

Nadat er een steekproefontwerp is gekozen, moet er vervolgens worden vastgesteld hoe het desbetreffende populatietotaal moet worden geschat of, anders geformuleerd, hoe de steekproefwaarnemingen moeten worden opgehoogd naar het populatietotaal. Om die reden spreekt men ook wel van een ophoogmethode. Maar vaak zal men evenwel ook geïnteresseerd zijn in het schatten van een populatiegemiddelde. De term ophogen is dan wat minder geschikt.

De keuze voor een bepaalde ophoogmethode of schattingsmethode wordt in belangrijke mate bepaald door de aanwezigheid van eventuele hulpinformatie en het karakter van de samenhang tussen de doelvariabele enerzijds en de hulpinformatie anderzijds. Afhankelijk daarvan kan men dan bijvoorbeeld kiezen voor de quotiëntschatter, de regressieschatter of de poststratificatieschatter.

Onder het deelthema Steekproefontwerpen vallen de hoofdstukken 2, 3, 4, en 5. Het deelthema Ophoogmethoden omvat de hoofdstukken 6, 7, en 8. Ten slotte merken we op, dat een vergelijking of formule genummerd is als er in de tekst naar wordt verwezen. Als dit nummer van een asterisk (*) is voorzien, dan is het bijbehorende bewijs in het appendix van het desbetreffende hoofdstuk opgenomen.

1.2 Afbakening en relatie met andere thema's

Het spreekt voor zich dat er relaties bestaan tussen aan de éne kant ophogen en aan de andere kant *Wegen als correctie voor non-respons*. De daar besproken methoden en technieken bouwen voort op de elementaire ophoogmethoden die in dit document worden besproken maar zijn meer toegespitst op het probleem van de non-respons.

Daarnaast is er een samenhang met het thema *Panels* waarbij populatie-elementen gedurende diverse aaneengesloten waarnemingsperioden in de steekproef zijn opgenomen. Dit gebeurt veelal wanneer men is geïnteresseerd in ontwikkelingen over de tijd. Verder bestaan er ook duidelijke relaties met het thema *Experimenten* en het thema *Modelmatig schatten* ten behoeve van bijvoorbeeld kleine subpopulaties.

1.3 Plaats in het statistische proces

Uit het voorafgaande moge duidelijk zijn dat het steekproefontwerp zich afspeelt aan het begin van het statistische proces, terwijl het ophogen zich meer aan het einde van statistische proces bevindt vlak voor het publiceren van de cijfers.

2. Enkelvoudige, aselecte steekproeven zonder teruglegging

2.1 Korte beschrijving

De enkelvoudige, aselecte steekproef zonder teruglegging (EAZT) is het meest bekende steekproefontwerp. De steekproef wordt enkelvoudig genoemd omdat zij uit de gehele populatie wordt getrokken. Naast de EAZT-steekproeven worden ook nog wel de enkelvoudige, aselecte steekproeven *met* teruglegging (EAMT) onderscheiden. Deze worden echter niet gebruikt bij het CBS. Omdat de relatief eenvoudige EAMT-variantieformules onder bepaalde voorwaarden toch goed bruikbaar zijn in geval van EAZT-steekproeven, zullen we in paragraaf 2.3.2 enige aandacht schenken aan EAMT-steekproeven en de bijbehorende variantieformules.

2.1.1 Definitie EAZT-steekproef

De wat formele omschrijving van het EAZT-steekproefontwerp is als volgt. Beschouw een populatie U van N elementen ofwel $U = \{1,2,\cdots,N\}$. We spreken van een EAZT-steekproef van omvang n uit populatie U wanneer alle mogelijke deelverzamelingen van U met omvang n even veel kans hebben om als steekproef te worden getrokken. Merk op dat er $\binom{N}{n}$ mogelijke deelverzamelingen van U bestaan met omvang n.

2.1.2 Trekkingsschema voor EAZT-steekproef

In de praktijk kan een EAZT-steekproef worden verkregen door achter elkaar en op aselecte wijze getallen tussen 1 en N te kiezen, en steeds het bijbehorende element op te nemen in de steekproef totdat men *n* elementen heeft geselecteerd. Indien een getal al eerder blijkt te zijn gekozen, dan kiest men opnieuw op aselecte wijze een getal.

Een goed alternatief voor de EAZT-steekproef is de, zogenaamde, *systematische* steekproef. Een beschrijving van de systematische steekproef is voorhanden in paragraaf 4.3.6.

2.2 Toepasbaarheid

Bij het CBS vindt het EAZT-steekproefontwerp vooral toepassing bij de bedrijfsstatistieken. Bij deze statistieken wordt de populatie van bedrijven veelal in strata ingedeeld. Vervolgens wordt er per stratum een EAZT-steekproef getrokken; zie het volgende hoofdstuk over gestratificeerde steekproeven. Verder wordt het EAZT-ontwerp ook dikwijls genomen als een soort ijkpunt waarmee andere steekproefontwerpen worden vergeleken voor wat betreft de varianties van de diverse schatters.

2.3 Uitgebreide beschrijving

2.3.1 Veronderstellingen en notatie

Eerst zullen we in het kort ingaan op een aantal belangrijke populatieparameters. Met betrekking tot de *doelvariabele* kunnen de volgende parameters worden onderscheiden: het *populatietotaal* Y, het *populatiegemiddelde* $\bar{Y}$, de *populatievariantie* σ_y^2, de *aangepaste populatievariantie* S_y^2, en de *populatievariatiecoëfficiënt* CV_y. De aangepaste populatievariantie wordt vaak gebruikt ter vereenvoudiging van de formules bij de EAZT-steekproef. Verder staat Y_k voor de waarde van de doelvariabele van element k, $k = 1,\cdots,N$. De genoemde parameters zijn als volgt gedefinieerd:

$$\begin{aligned} Y &= Y_1 + \cdots + Y_N = \sum_{k=1}^{N} Y_k \\ \bar{Y} &= \frac{1}{N} Y = \frac{1}{N} \sum_{k=1}^{N} Y_k \\ \sigma_y^2 &= \frac{1}{N} \sum_{k=1}^{N} (Y_k - \bar{Y})^2 \\ S_y^2 &= \frac{N}{N-1} \sigma_y^2 = \frac{1}{N-1} \sum_{k=1}^{N} (Y_k - \bar{Y})^2 \\ CV_y &= \frac{S_y}{\bar{Y}} \quad . \end{aligned} \tag{2.1}$$

De n (i.e. de steekproefomvang) waarnemingen van de doelvariabele in de steekproef duiden we aan met behulp van kleine letters $y_1,\cdots,y_n$. Het *steekproefgemiddelde* $\bar{y}_s$, de *steekproefvariantie* s_y^2 en de *steekproefvariatiecoëfficiënt* cv_y zijn de drie belangrijkste *steekproefparameters*. Deze drie steekproefparameters zijn als volgt gedefinieerd:

$$\begin{aligned} \bar{y}_s &= \frac{1}{n} \sum_{k=1}^{n} y_k \\ s_y^2 &= \frac{1}{n-1} \sum_{k=1}^{n} \left(y_k - \bar{y}_s\right)^2 \\ cv_y &= \frac{s_y}{\bar{y}_s} \quad . \end{aligned} \tag{2.2}$$

2.3.2 Schatters voor het populatiegemiddelde, het populatietotaal en het aangepaste populatievariantie

Voor drie van de vijf populatieparameters gedefinieerd in (2.1) gaan we een expliciete schatter formuleren. Het gaat daarbij om het populatietotaal, het populatiegemiddelde en de aangepaste populatievariantie. Intuïtief zal het duidelijk zijn dat het steekproefgemiddelde $\bar{y}_s$ in principe een redelijke schatter is voor het populatiegemiddelde. Dit noteren we als volgt:

$$\hat{\bar{Y}} = \bar{y}_s = \frac{1}{n}\sum_{k=1}^{n} y_k \quad . \tag{2.3}$$

Het dakje ^ in het linkerlid wordt gebruikt om aan te geven dat het om een schatter gaat van de desbetreffende parameter.

Op basis van de schatter voor het populatiegemiddelde komen we tot het volgende voorstel voor de schatter van het populatietotaal:

$$\hat{Y} = N\,\hat{\bar{Y}} = N\,\bar{y}_s = \sum_{k=1}^{n}\left(\frac{N}{n}\right) y_k \quad . \tag{2.4}$$

De parameter N/n die in deze uitdrukking voorkomt wordt het *ophooggewicht* van de steekproef genoemd omdat hij de verhouding tussen populatieomvang en steekproefomvang weergeeft.

We presenteren, ten slotte, de steekproefvariantie als de schatter voor de aangepaste populatievariantie, i.e.:

$$\hat{S}_y^2 = s_y^2 = \frac{1}{n-1}\sum_{k=1}^{n}\left(y_k - \bar{y}_s\right)^2 \quad . \tag{2.5}$$

Zuiverheidsanalyses

Een schatter is zuiver als deze in verwachtingswaarde overeenkomt met de te schatten grootheid. Om de verwachtingen van de *(directe) schatters* voor het populatiegemiddelde en het populatietotaal bij de EAZT-steekproef te kunnen bepalen, is het handig de schatters in een alternatieve vorm te herschrijven. Zo herschrijven we de schatter voor het populatiegemiddelde als:

$$\hat{\bar{Y}} = \bar{y}_s = \frac{1}{n}\sum_{k=1}^{N} a_k Y_k \tag{2.6}$$

waarbij de dichotome random variabele a_k is gedefinieerd als:

$$a_k = \begin{cases} 1 & \text{als element } k \textit{ wel} \text{ is getrokken in de steekproef} \\ 0 & \text{als element } k \textit{ niet} \text{ is getrokken in de steekproef.} \end{cases} \tag{2.7}$$

De dichotome random variabele a_k wordt ook wel de *insluitindicator* genoemd. De insluitindicator heeft een tweetal aantrekkelijke eigenschappen, die hieronder zijn weergegeven:

$$E(a_k) = P(a_k = 1) = \frac{n}{N} \qquad k = 1, \cdots, N \tag{2.8}$$

en:

$$E(a_k a_l) = \begin{cases} \dfrac{n}{N} & k = l \quad \wedge \quad k = 1, \cdots, N \\ \dfrac{n(n-1)}{N(N-1)} & 1 \le k \ne l \le N \quad . \end{cases} \tag{2.9}$$

Neem de alternatieve schrijfwijze (2.6) voor de schatter van het populatiegemiddelde als uitgangspunt. Dan kan op basis van eigenschappen (2.8) en (2.9) van insluitindicator a_k worden aangetoond, dat deze schatter een *zuivere* schatter is voor $\overline{Y}$:

$$\begin{aligned} E\left(\hat{\overline{Y}}\right) &= E\left(\frac{1}{n}\sum_{k=1}^{N} a_k Y_k\right) = \frac{1}{n}\sum_{k=1}^{N} E(a_k) Y_k \\ &= \frac{1}{n}\sum_{k=1}^{N} \frac{n}{N} Y_k = \frac{1}{N}\sum_{k=1}^{N} Y_k = \overline{Y} \quad . \end{aligned}$$

Door gebruik te maken van dit resultaat kunnen we ook de zuiverheid van de schatter van het populatietotaal eenvoudig verifiëren, nl.:

$$E\left(\hat{Y}\right) = E\left(N\,\hat{\overline{Y}}\right) = N\,E\left(\hat{\overline{Y}}\right) = N\,\overline{Y} = Y \quad .$$

Het gebruik van de insluitindicatoren is eveneens nuttig bij het bewijzen, dat s_y^2 in (2.5) een zuivere schatter is voor de aangepaste populatievariantie. Dat wil zeggen:

$$E\left(s_y^2\right) = E\left(\frac{1}{n-1}\sum_{k=1}^{n}\left(y_k - \overline{y}_s\right)^2\right) = S_y^2 \quad . \tag{2.10*}$$

De asterisk (*) bij het nummer van de laatste vergelijking geeft aan, dat het bewijs van de vergelijking in het appendix van dit hoofdstuk wordt gegeven.

Variantieberekeningen

Er kan worden bewezen, dat de variantie van de schatter voor het populatiegemiddelde een functie van de aangepaste populatievariantie is:

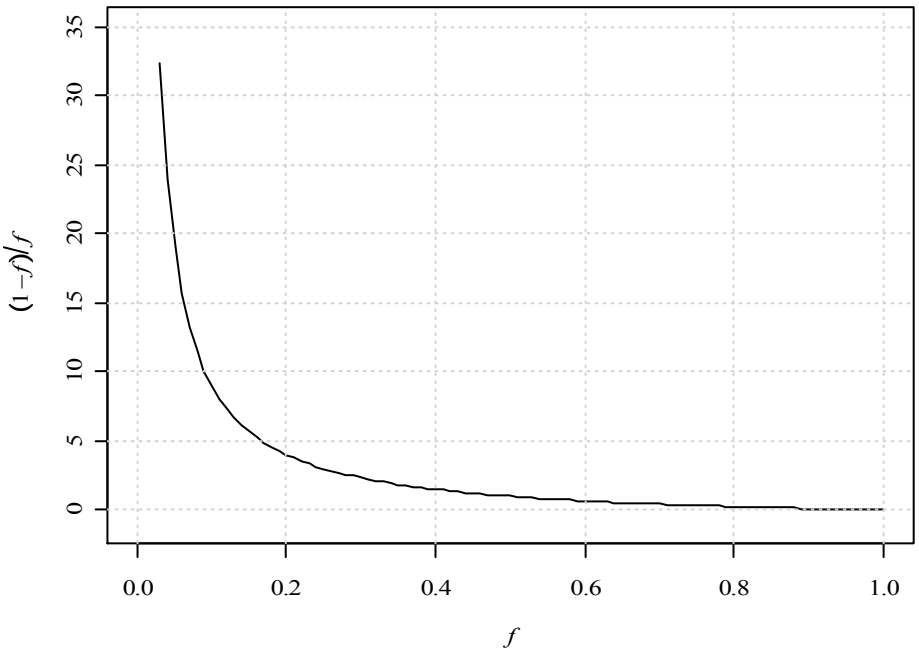
$$\operatorname{var}\left(\hat{\overline{Y}}\right) = \frac{1}{n}(1-f)S_y^2 \quad = \frac{1}{N}\left(\frac{1-f}{f}\right)S_y^2 \qquad \left(f = \frac{n}{N}\right) \quad . \tag{2.11*}$$

De factor $(1-f)$ wordt de *eindigheidscorrectie* genoemd terwijl de factor f *de steekproeffractie* wordt genoemd. Op basis van dit resultaat vinden we voor de variantie van de schatter van het populatietotaal:

$$\operatorname{var}\left(\hat{Y}\right) = N^2 \operatorname{var}\left(\hat{\overline{Y}}\right) = N\left(\frac{1-f}{f}\right)S_y^2 \quad . \tag{2.12}$$

De waarde van parameter f is uitgezet tegen de waarde van parameter $(1-f)/f$ in figuur 2.1. Als voor (vaste) populatieomvang N, de steekproefomvang n in waarde toeneemt, gaat steekproeffractie f naar 1. Uit de grafiek lezen we af dat de parameter $(1-f)/f$ dan naar 0 gaat. Met andere woorden, de variantie van de schatter van het populatietotaal gaat naar 0 bij een groter wordende steekproef, zie uitdrukking (2.12).

Figuur 2.1. Parameter f uitgezet tegen parameter $(1-f)/f$



De variantie van de schatter voor het populatiegemiddelde zoals uitgerekend in (2.11), kan worden geschat met behulp van de volgende schatter:

$$\hat{\text{var}}\left(\hat{\bar{Y}}\right) = \frac{1}{N}\left(\frac{1-f}{f}\right)\hat{S}_y^2 = \frac{1}{N}\left(\frac{1-f}{f}\right)s_y^2 \tag{2.13}$$

waarin s_y^2 de steekproefvariantie representeert. De zuiverheid van deze zogenaamde *variantieschatter* volgt uit de zuiverheid van steekproefvariantie s_y^2 als schatter van de aangepaste populatievariantie S_y^2.

De variantie van schatter $\hat{Y}$ zoals die in (2.12) is uitgerekend, kunnen we op geheel analoge wijze zuiver schatten met behulp van de schatter

$$\hat{\text{var}}\left(\hat{Y}\right) = N\left(\frac{1-f}{f}\right)s_y^2 \tag{2.14}$$

met s_y^2 de steekproefvariantie. De zuiverheid van deze variantieschatter is een direct gevolg van de (bewezen) zuiverheid van s_y^2 als schatter van S_y^2.

Variatiecoëfficiënten

Naast het uitrekenen (en schatten) van de variantie van een schatter, is het ook mogelijk de bijbehorende variatiecoëfficiënt te bepalen. De variatiecoëfficiënten van de schatter voor het populatiegemiddelde, notatie $CV\left(\hat{\bar{Y}}\right)$, en de schatter voor het populatietotaal, notatie $CV\left(\hat{Y}\right)$, zijn als volgt gedefinieerd:

$$CV\left(\hat{\bar{Y}}\right) = \frac{\sqrt{\text{var}\left(\hat{\bar{Y}}\right)}}{\bar{Y}}$$
$$CV(\hat{Y}) = \frac{\sqrt{\text{var}\left(\hat{Y}\right)}}{Y} \quad . \tag{2.15}$$

In formule (2.11) is een uitdrukking gegeven voor de variantie van de schatter van het populatiegemiddelde onder een EAZT-steekproef. Door gebruik te maken van deze formule, kan een relatie worden afgeleid, zie hieronder, op basis waarvan de variatiecoëfficiënt van deze schatter kan worden uitgerekend.

$$CV^2\left(\hat{\bar{Y}}\right) = \left(\frac{1}{N}\right)\left(\frac{1-f}{f}\right)CV_y^2 \tag{2.16}$$

Met andere woorden, er bestaat een direct verband tussen de variatiecoëfficiënt van de schatter voor het populatiegemiddelde en de populatievariatiecoëfficiënt.

Gegeven het verband tussen het populatietotaal en het populatiegemiddelde, en de relatie tussen de schatter voor het populatiegemiddelde en de schatter voor het populatietotaal, kan het bestaan van de volgende relatie tussen de variatiecoëfficiënten van de beide schatters worden aangetoond.

$$CV\left(\hat{Y}\right) = CV\left(\hat{\bar{Y}}\right) \tag{2.17}$$

Deze relatie brengt tot uitdrukking, dat de directe schatters voor het populatietotaal en het populatiegemiddelde even nauwkeurig zijn.

Helaas is het niet mogelijk om de beide variatiecoëfficiënten in (2.15) daadwerkelijk uit te rekenen. Dit betekent, dat ze zullen moeten worden geschat. De voor de hand liggende schatter voor de variatiecoëfficiënt van de schatter voor het populatiegemiddelde is:

$$\hat{CV}\left(\hat{Y}\right) = \left(\frac{s_y}{\bar{y}_s}\right)\sqrt{\frac{1}{N}\left(\frac{1-f}{f}\right)} \tag{2.18}$$

omdat:

$$\hat{CV}_y = cv_y = \frac{s_y}{\bar{y}_s} \quad .$$

Het is van belang op te merken dat de steekproeffractie in de praktijk vaak heel klein is, d.w.z. $f \approx 0$. In feite komt dit er op neer, dat men doet alsof de steekproef met teruglegging is getrokken. Intuïtief is het ook niet verwonderlijk dat wel/niet terugleggen geen effect meer heeft op de uitkomst van de schatter wanneer de populatieomvang veel groter is dan de omvang van de steekproef. In de volgende deelparagraaf gaan we nader in op de enkelvoudige, aselecte steekproef met teruglegging.

2.3.3 Enkelvoudige, aselecte steekproeven met teruglegging

Veronderstel dat we de steekproef enkelvoudig, aselect, met gelijke trekkingskansen $(1/N)$ en *met* teruglegging (EAMT) trekken. Dan kunnen $y_1, \cdots, y_n$ als *n onafhankelijke* trekkingen uit $Y_1, \cdots, Y_N$ worden gezien met de volgende verwachting en variantie:

$$E(y_k) = \frac{1}{N}\sum_{l=1}^{N} Y_l = \bar{Y} \qquad\qquad k = 1, \cdots, n$$

en:

$$\text{var}(y_k) = \frac{1}{N}\sum_{l=1}^{N} (Y_l - \bar{Y})^2 = \sigma_y^2 \qquad\qquad k = 1, \cdots, n \quad . \tag{2.19}$$

Merk op dat deze twee formules ook van toepassing zijn op een EAZT-steekproef. In geval van een EAMT-steekproef volgen uit deze twee formules direct respectievelijk de verwachting en de variantie van de schatter $\hat{\bar{Y}} = \bar{y}_s$:

$$E(\bar{y}_s) = \frac{1}{n}\sum_{k=1}^{n} E(y_k) = \bar{Y}$$

en:

$$\text{var}(\bar{y}_s) = \frac{1}{n}\sigma_y^2 \quad . \tag{2.20*}$$

Vergelijking van de variantieformules in deze deelparagraaf met die in de vorige deelparagraaf bevestigt de eerder genoemde eigenschap dat de variantieformules ongeveer hetzelfde zijn wanneer f klein is.

Ten slotte zij nog vermeld dat bij een EAMT-steekproef geldt dat $E(s_y^2) = \sigma_y^2$. Het bewijs is min of meer hetzelfde als het bewijs bij de EAZT-steekproef en daarom laten we het achterwege.

2.3.4 Bepaling steekproefomvang

Een veel voorkomende vraag in de praktijk is hoe groot een steekproef moet zijn om met een bepaalde nauwkeurigheid een uitspraak te kunnen doen over Y, het populatietotaal. Als maatstaf voor de nauwkeurigheid wordt vaak het 95%-betrouwbaarheidsinterval $I_{95}(Y)$ gehanteerd. De definitie van $I_{95}(Y)$ houdt in dat $I_{95}(Y)$ met een kans van 95% het onbekende populatietotaal omvat. Onder de aanname dat de verdeling van de statistische grootheid $\hat{Y}$ goed benaderd kan worden met de normale verdeling, kan het 95%-betrouwbaarheidinterval voor Y worden benaderd voor voldoende grote n met:

$$I_{95}(Y) = \left(\hat{Y} - 1{,}96\sqrt{\hat{\text{var}}\left(\hat{Y}\right)}, \hat{Y} + 1{,}96\sqrt{\hat{\text{var}}\left(\hat{Y}\right)} \right) \quad .$$

Dit wordt met behulp van procentuele onzekerheidsmarges ook wel geformuleerd als:

$$\begin{aligned} I_{95}(Y) &= \hat{Y} \pm 1{,}96 \frac{\sqrt{\hat{\text{var}}\left(\hat{Y}\right)}}{\hat{Y}} 100\% \\ &= \hat{Y} \pm \hat{CV}\left(\hat{Y}\right) \times 196\% \quad . \end{aligned} \tag{2.21}$$

Aan de hand van een klein voorbeeld zullen we dit toelichten. Stel $\hat{Y} = 120$ en $\hat{\text{var}}\left(\hat{Y}\right) = 25$. We krijgen dan $I_{95}(Y) = (120 - 9{,}8; 120 + 9{,}8) = (110{,}2; 129{,}8)$. Overeenkomstig (2.21) kunnen we dit interval ook weergeven als $120 \pm 8{,}2\%$.

Met behulp van bovenstaande relaties kan ook worden nagegaan hoe groot de steekproef moet zijn om binnen een gegeven onzekerheidsmarge te blijven, zeg $r\%$. Uit

$$196 \times CV_y \sqrt{\frac{1}{N}\left(\frac{1-f}{f}\right)} < r$$

volgt:

$$\frac{196^2 \times CV_y^2}{N \times r^2 + 196^2 \times CV_y^2} < f \quad .$$

Indien een onzekerheidsmarge van 5% toelaatbaar wordt gevonden, moet de steekproeffractie f voor een populatie met $N = 1.000$ en een $CV_y = 0{,}7$ aan de volgende ongelijkheid voldoen:

$$\frac{196^2 \times 0{,}49}{1.000 \times 25 + 196^2 \times 0{,}49} \approx 0{,}43 = f_{\min} < f \quad .$$

Bij de gegeven populatieomvang $N = 1.000$, komt deze minimale steekproeffractie overeen met een minimale steekproefomvang $n_{\min} = 430$.

In de praktijk moet CV_y uiteraard eerst op de een of andere manier worden geschat. Dit kan vaak aan de hand van recente data uit het verleden worden gedaan. Soms kan men gebruik maken van vergelijkbare data die eenzelfde mate van volatiliteit hebben.

2.4 Voorbeeld: het schatten van fracties

Indien we het percentage personen met een bepaalde eigenschap willen schatten, bijvoorbeeld het aantal personen met een hoog salaris, neemt Y_k de volgende waarden aan:

$$Y_k = \begin{cases} 1 & \text{als persoon } k \text{ een hoog salaris heeft} \\ 0 & \text{als persoon } k \text{ geen hoog salaris heeft} \end{cases} \qquad k = 1, \cdots, N \quad .$$

Als $P = \overline{Y}$ de fractie hooggesalarieerden in de populatie is, geldt er (merk op, dat in deze situatie geldt: $Y_k^2 = Y_k$):

$$\begin{aligned} \overline{Y} &= \frac{1}{N}\sum_{k=1}^{N} Y_k = P \\ \sigma_y^2 &= \frac{1}{N}\sum_{k=1}^{N}(Y_k - P)^2 = \frac{1}{N}\sum_{k=1}^{N} Y_k^2 - P^2 = P - P^2 = PQ \qquad (Q = 1 - P) \\ S_y^2 &= \frac{N}{N-1} PQ \quad . \end{aligned}$$

Voor de steekproeffractie van hooggesalarieerden $p = \hat{P}$ geldt, in overeenstemming met formules (2.3) en (2.11):

$$\begin{aligned} p &= \overline{y}_s \\ \text{var}(p) &= \frac{1}{N}\left(\frac{1-f}{f}\right) S_y^2 = \left(\frac{1}{N-1}\right)\left(\frac{1-f}{f}\right) PQ \quad . \end{aligned}$$

Het laatste resultaat, de uitgerekende variantie van de schatter, kan worden geschat met (2.13):

$$\begin{aligned} \hat{\text{var}}(p) &= \frac{1}{N}\left(\frac{1-f}{f}\right) s_y^2 = \frac{1}{N}\left(\frac{1-f}{f}\right)\left(\frac{1}{n-1}\right)\sum_{k=1}^{n}(y_k - p)^2 \\ &= (1-f)\left(\frac{1}{n-1}\right) p(1-p) \quad . \end{aligned}$$

De variatiecoëfficiënt van schatter p kunnen we dus op de volgende wijze schatten, zie (2.15) en ook (2.18):

$$\hat{CV}(p) = \sqrt{(1-f)\left(\frac{1}{n-1}\right)\left(\frac{1-p}{p}\right)} \quad .$$

Uit formule (2.21) voor het 95%-betrouwbaarheidsinterval voor p volgt dan, dat voor voldoende grote n, $I_{95}(p)$ geschat kan worden als:

$$I_{95}(P) = p \pm 196\% \times \sqrt{(1-f)\left(\frac{1}{n-1}\right)\left(\frac{1-p}{p}\right)} \quad .$$

Deze laatste formule geeft aan dat de relatieve onzekerheidsmarge sterk toeneemt wanneer p kleiner wordt. Bijvoorbeeld, wanneer op basis van een EAZT-steekproef, parameter P wordt geschat op 0,001 met $n = 50.000$, dan is onder de aanname dat $f \approx 0$ de relatieve onzekerheidsmarge gelijk aan 28% Dit illustreert hoe moeilijk het is om met voldoende nauwkeurigheid zgn. kleine domeinen te schatten.

2.5 Kwaliteitsindicatoren

Kwaliteitsindicatoren voor wat betreft een EAZT-steekproef zijn:

- de onzekerheidsmarges van de desbetreffende schatters;
- de omvang van de non-respons.

De non-respons kan de kwaliteit van de resultaten vooral aantasten wanneer (i) de omvang van de non-respons groot is en (ii) de non-respons selectief is. Soms kan de vertekening ten gevolge van selectieve non-respons voor een deel worden gecorrigeerd door gebruik te maken van hulpvariabelen die samenhangen met zowel de responskans als met de doelvariabele.

Verder is een belangrijke aanname bij een kanssteekproef dat het *kader* waaruit de steekproef wordt getrokken goed overeenkomt met de *doelpopulatie* waarover men een uitspraak wil doen.

2.6 Appendix

Bewijs van (2.10)

Alvorens een formule voor de variantie van schatter (2.3) af te leiden merken we op, dat de variantie van de schatter voor het populatiegemiddelde niet verandert als we een constante optellen bij alle Y_k uit de populatie. Met andere woorden, er mag worden aangenomen zonder het algemene karakter van de resultaten te schaden, dat het populatiegemiddelde gelijk is aan nul, ofwel $\overline{Y} = 0$. Uitgaande van deze aanname herschrijven we de uitdrukking voor de steekproefvariantie met behulp van de insluitindicator als:

$$s_y^2 = \frac{1}{n-1}\sum_{k=1}^{n} y_k^2 - \frac{n}{n-1}\overline{y}_s^2 = \frac{1}{n-1}\sum_{k=1}^{N} a_k Y_k^2 - \frac{n}{n-1}\overline{y}_s^2 \quad .$$

De verwachting van s_y^2 kunnen we met behulp van definities (2.2) en (2.6) uitrekenen als (merk op, dat $E(s_y^2) = \text{var}(\overline{y}_s)$ aangezien $\overline{Y} = 0$):

$$
\begin{aligned}
E\left(s_y^2\right) &= \frac{1}{n-1}\sum_{k=1}^{N} E\left(a_k\right)Y_k^2 - \frac{n}{n-1}E\left(\bar{y}_s^2\right) \\
&= \frac{1}{n-1}\times\frac{n}{N}\sum_{k=1}^{N} Y_k^2 - \frac{n}{n-1}\operatorname{var}\left(\bar{y}_s\right) \\
&= \left(\frac{n}{n-1}\right)\sigma_y^2 - \frac{n}{n-1}\operatorname{var}\left(\hat{\bar{Y}}\right) \quad .
\end{aligned}
$$

Met andere woorden, de verwachting van de steekproefvariantie is een functie van de variantie van de schatter voor het populatiegemiddelde. Om de verwachting van de steekproefvariantie uit te kunnen rekenen, is dus kennis van de variantie van de schatter voor het populatiegemiddelde vereist. In (2.11) is een formule voor de variantie van de schatter voor het populatiegemiddelde afgeleid.

Gebruik makend van bovenstaand resultaat en uitdrukking (2.11) vinden we:

$$
\begin{aligned}
E\left(s_y^2\right) &= \frac{n}{n-1}\sigma_y^2 - \frac{n}{n-1}\times\frac{1}{n}\left(1-f\right)S_y^2 \\
&= \frac{n\left(N-1\right)}{N\left(n-1\right)}S_y^2 - \frac{N-n}{N\left(n-1\right)}S_y^2 \\
&= S_y^2 \quad .
\end{aligned}
$$

Het bewijs van de zuiverheid van s_y^2 als schatter van S_y^2 is hiermee afgerond. ■

Bewijs van (2.11)

Voor het bewijs van formule (2.11) voor de variantie van het steekproefgemiddelde, nemen we opnieuw ter vereenvoudiging aan, dat $\bar{Y} = Y = 0$. Uit (2.7)-(2.9) volgt:

$$
\begin{aligned}
\operatorname{var}\left(\hat{\bar{Y}}\right) = E\left(\hat{\bar{Y}}\right)^2 &= E\left(\frac{1}{n}\sum_{k=1}^{N} a_k Y_k\right)^2 \\
&= \frac{1}{n^2}\left\{\sum_{k=1}^{N} E\left(a_k^2\right)Y_k^2 + \sum_{k=1}^{N}\sum_{\substack{l=1 \\ l\neq k}}^{N} E\left(a_k a_l\right)Y_k Y_l\right\} \\
&= \frac{1}{n^2}\left\{\frac{n}{N}\sum_{k=1}^{N} Y_k^2 + \frac{n(n-1)}{N(N-1)}\sum_{k=1}^{N} Y_k\left(0-Y_k\right)\right\} \\
&= \frac{n}{n^2}\left\{\sigma_y^2 - \frac{n-1}{N-1}\sigma_y^2\right\} \\
&= \frac{1}{n}\left(\frac{N-n}{N-1}\right)\sigma_y^2 \\
&= \left(\frac{1}{N}\right)\left(\frac{N}{n}\right)\left(\frac{N-n}{N-1}\right)\left(\frac{N-1}{N}\right)S_y^2 \\
&= \frac{1}{N}\left(\frac{1-f}{f}\right)S_y^2 \quad .
\end{aligned}
$$

Hiermee is de juistheid van uitdrukking (2.11) voor de variantie van het steekproef-gemiddelde bewezen. ■

Bewijs van (2.20)

Bij een EAMT-steekproef zijn trekkingen y_k en $y_l\,(k \neq l)$ onafhankelijk, en dus ongecorreleerd (dit is het grote verschil met een EAZT-steekproef waarbij y_k en y_l negatief zijn gecorreleerd met een correlatiecoëfficiënt van $-1/(N-1)$). Voor de variantie van het steekproefgemiddelde kunnen we dus schrijven:

$$\mathrm{var}(\bar{y}_s) = \mathrm{var}\left(\frac{1}{n}\sum_{k=1}^{n} y_k\right) = \frac{1}{n^2}\sum_{k=1}^{n}\mathrm{var}(y_k) \quad .$$

Uit formule (2.19) voor de variantie van trekking k volgt dan:

$$\mathrm{var}(\bar{y}_s) = \frac{1}{n^2}\sum_{k=1}^{n}\sigma_y^2 = \frac{1}{n^2}n\sigma_y^2 = \frac{1}{n}\sigma_y^2 \quad .$$

Het bewijs is nu compleet. ■

3. Gestratificeerde steekproeven

3.1 Korte beschrijving

Het uitgangspunt van de definitie van de gestratificeerde, aselecte steekproef is een opsplitsing van de doelpopulatie in, zogenaamde, *strata* (het enkelvoud is *stratum*). Deze opsplitsing is dekkend en dusdanig uitgevoerd, dat individuele strata elkaar niet overlappen. In een gestratificeerde, aselecte steekproef wordt vervolgens uit ieder stratum een EAZT-steekproef getrokken. De verschillende EAZT-steekproeven zijn onderling onafhankelijk. Met andere woorden, in plaats van het trekken van één grote EAZT-steekproef, worden nu een aantal kleinere EAZT-steekproeven getrokken.

3.2 Toepasbaarheid

Bij het CBS worden vaak regionale, demografische of sociaal-economische factoren gebruikt voor stratificatie. Om te kunnen *stratificeren*, i.e. het opdelen van de doelpopulatie in strata, moet de benodigde hulpinformatie wel beschikbaar zijn in het steekproefkader. Wanneer we bijvoorbeeld willen stratificeren naar bedrijfstak moet voor elk bedrijf uit het steekproefkader bekend zijn tot welke bedrijfstak het behoort. In het ABR staat onder meer de SBI (Standaard Bedrijfs Indeling) en grootteklasse van een onderneming, hulpvariabelen die op het CBS vaak gebruikt worden om te stratificeren. Het GBA bevat voor ieder persoon een groot aantal persoonsgegevens zoals geslacht, geboortedatum, burgerlijke staat, adres en woonplaats die kunnen worden gebruikt om de Nederlandse bevolking onder te verdelen.

Er kunnen verschillende redenen zijn om een gestratificeerde steekproef te trekken. Ten eerste, stratificatie wordt veelal toegepast om de precisie van de schatters te vergroten (dat wil zeggen de variantie te verkleinen), vooral als het gaat om karakteristieken van de gehele populatie te schatten. Sommige variabelen kunnen een dermate grote populatievariantie hebben, bijvoorbeeld de omzet van bedrijven, dat een zeer grote enkelvoudige, aselecte steekproef moet worden getrokken om uitspraken met een redelijke betrouwbaarheid te kunnen doen. Als het mogelijk is groepen te vormen waarbinnen de doelvariabele weinig varieert zullen gestratificeerde steekproeven tot preciezere uitkomsten leiden dan enkelvoudige (bij gelijke steekproefgrootte). De precisie neemt toe omdat de variantie binnen de strata kleiner is dan de variantie voor de populatie als geheel.

Ten tweede, in onderzoek is men vaak niet alleen geïnteresseerd in de populatie als geheel, maar wil men ook uitspraken doen over deelpopulaties of deelpopulaties met elkaar vergelijken. Bij enkelvoudige steekproeven bepaalt de toevallige getrokken steekproef hoeveel elementen in de strata terechtkomen. Vooral kleine deelpopulaties kunnen dan slecht in de steekproef vertegenwoordigd zijn. Met behulp van stratificatie kan men er voor zorgen dat alle deelpopulaties waarin men geïnteresseerd

is, voldoende in de steekproef gerepresenteerd zijn om er betrouwbare uitspraken over te kunnen doen.

Ten derde, bij stratificatie is het mogelijk dat verschillende benaderingstechnieken worden gebruikt in verschillende strata. Zo kan het in een onderzoek onder bedrijven wenselijk zijn om kleine bedrijven met behulp van een (korte) schriftelijke vragenlijst te benaderen en grote bedrijven een uitgebreid telefonisch of persoonlijk interview af te nemen. Ook kunnen de trekking- en schattingsmethoden per stratum verschillen.

Ten vierde, om administratieve redenen bestaan steekproefkaders vaak al uit 'natuurlijke' onderverdelingen die zich soms zelfs op verschillende plaatsen kunnen bevinden. Het kan dan voordeliger zijn om aparte steekproeven te trekken.

3.3 Uitgebreide beschrijving

3.3.1 Veronderstellingen en notatie

De populatie wordt verdeeld in H *strata*. Een afzonderlijk *stratum* wordt aangeduid met de index $h, h = 1,\cdots,H$, en bevat N_h elementen. De strata mogen elkaar niet overlappen. Met andere woorden, ieder element mag slechts tot één enkel stratum behoren. Samen vormen de strata de populatie zodat

$$\sum_{h=1}^{H} N_h = N$$

waarin N weer de totale populatieomvang is. Verder nemen we aan dat voor ieder stratum de omvang N_h bekend is. De waarde van de doelvariabele voor element k in stratum h wordt genoteerd met Y_{hk}, voor $h = 1,\cdots,H$, en $k = 1,\cdots,N_h$. De omvang van de steekproef uit stratum h geven we weer met n_h; per definitie geldt $\sum_h n_h = n$. De notatie voor de steekproefwaarnemingen in stratum h is y_{hk}, met $h = 1,\cdots,H$, en $k = 1,\cdots,n_h$.

Met betrekking tot de doelvariabelen onderscheiden we de volgende, zogenaamde, *stratumparameters*: het *stratumtotaal* Y_h, het *stratumgemiddelde* $\overline{Y}_h$, de *stratumvariantie* σ^2_{yh}, *de aangepaste stratumvariantie* S^2_{yh}, en de *stratumvariatiecoëfficiënt* CV_{yh}. Deze parameters zijn als volgt gedefinieerd:

$$
\begin{aligned}
Y_h &= \sum_{k=1}^{N_h} Y_{hk} \\
\overline{Y}_h &= \frac{1}{N_h} Y_h \\
\sigma_{yh}^2 &= \frac{1}{N_h} \sum_{k=1}^{N_h} \left(Y_{hk} - \overline{Y}_h \right)^2 \\
S_{yh}^2 &= \frac{1}{N_h - 1} \sum_{k=1}^{N_h} \left(Y_{hk} - \overline{Y}_h \right)^2 \\
CV_{yh} &= \frac{S_{yh}}{\overline{Y}_h} \quad .
\end{aligned}
\tag{3.1}
$$

Merk op, dat het bij bovenstaande parameters in principe gaat om populatieparameters in stratum h.

Omdat er in het gestratificeerde steekproefontwerp sprake is van het trekken van een steekproef in ieder stratum, is het zinvol om de zogenaamde steekproefparameters per stratum te introduceren. De steekproefparameters binnen een stratum zijn: het *steekproefgemiddelde* $\overline{y}_h$ in stratum h, de *steekproefvariantie* s_{yh}^2 in stratum h, en de *steekproefvariatiecoëfficiënt* cv_{yh} in stratum h. Zij zijn als volgt gedefinieerd:

$$
\begin{aligned}
\overline{y}_h &= \frac{1}{n_h} \sum_{k=1}^{n_h} y_{hk} \\
s_{yh}^2 &= \frac{1}{n_h - 1} \sum_{k=1}^{n_h} \left(y_{hk} - \overline{y}_h \right)^2 \\
cv_{yh} &= \frac{s_{yh}}{\overline{y}_h} \quad .
\end{aligned}
\tag{3.2}
$$

3.3.2 Relaties tussen populatieparameters en stratumparameters

Ieder element in de populatie behoort tot één enkel stratum. Deze bijzondere eigenschap van de strata maakt het mogelijk, dat relaties kunnen worden afgeleid tussen populatieparameters (2.1) enerzijds en stratumparameters (3.1) anderzijds. Zo vinden we voor het populatietotaal en het populatiegemiddelde:

$$
\begin{aligned}
Y &= \sum_{h=1}^{H} \sum_{k=1}^{N_h} Y_{hk} = \sum_{h=1}^{H} Y_h \\
\overline{Y} &= \frac{1}{N} Y = \frac{1}{N} \sum_{h=1}^{H} Y_h = \frac{1}{N} \sum_{h=1}^{H} N_h \overline{Y}_h = \sum_{h=1}^{H} \left(\frac{N_h}{N} \right) \overline{Y}_h \quad .
\end{aligned}
\tag{3.3}
$$

Met andere woorden, het populatietotaal kunnen we schrijven als de som van de stratumtotalen, en het populatiegemiddelde blijkt niets anders te zijn dan een gewogen gemiddelde van de stratumgemiddelden. De weegfactor van het stratumgemiddelde komt overeen met de relatieve omvang van het desbetreffende stratum.

Met betrekking tot de populatievariantie vinden we het volgende verband tussen σ_y^2, zie (2.1), en de verschillende stratumvarianties σ_{yh}^2:

$$\sigma_y^2 = \sum_{h=1}^{H}\left(\frac{N_h}{N}\right)\sigma_{yh}^2 + \sum_{h=1}^{H}\left(\frac{N_h}{N}\right)\left(\overline{Y}_h - \overline{Y}\right)^2 \quad . \tag{3.4*}$$

De uitdrukking voor de aangepaste populatievariantie S_y^2 kan worden herschreven in termen van de aangepaste stratumvarianties. Het resultaat is hieronder weergegeven:

$$S_y^2 = \sum_{h=1}^{H}\left(\frac{N_h - 1}{N - 1}\right)S_{yh}^2 + \sum_{h=1}^{H}\left(\frac{N_h}{N-1}\right)\left(\overline{Y}_h - \overline{Y}\right)^2 \quad . \tag{3.5*}$$

De eerste term aan de rechterkant van het '='-teken, wordt de *binnenstratavariantie* genoemd omdat zij is samengesteld uit de afzonderlijke stratumvarianties. De tweede term aan de rechterkant van het '='-teken staat bekend als de *tussenstratavariantie*. Deze tussenstratavariantie geeft aan hoe de stratumgemiddelden rond het populatiegemiddelde gespreid liggen.

Met behulp van relatie (3.5) leiden we een uitdrukking af voor de populatievariatiecoëfficiënt in termen van de verschillende stratumvariatiecoëfficiënten:

$$CV_y^2 = \sum_{h=1}^{H}\left(\frac{N_h - 1}{N-1}\right)\left(\frac{\overline{Y}_h}{\overline{Y}}\right)^2 CV_{yh}^2 + \sum_{h=1}^{H}\left(\frac{N_h}{N-1}\right)\left(\left(\frac{\overline{Y}_h}{\overline{Y}}\right) - 1\right)^2 \quad . \tag{3.6*}$$

3.3.3 Schatters voor het populatiegemiddelde en het populatietotaal

In de gestratificeerde, aselecte steekproef wordt per stratum een EAZT-steekproef getrokken. Onze analyse richt zich daarom op het analyseren van de steekproef in stratum h. Uit de populatie van N_h elementen in stratum h, trekken we een EAZT-steekproef van omvang n_h. Schatters voor achtereenvolgens het stratumgemiddelde en het stratumtotaal zijn direct voor handen (zie bijvoorbeeld hoofdstuk 2):

$$\begin{aligned} \hat{\overline{Y}}_h &= \overline{y}_h \\ \hat{Y}_h &= N_h \overline{y}_h \quad . \end{aligned} \tag{3.7}$$

Deze twee schatters zullen we benutten bij het schatten van het populatietotaal en het populatiegemiddelde.

Het populatietotaal Y komt overeen met de som van de stratumtotalen Y_h, zie (3.3). Daarom ligt het voor de hand het populatietotaal te schatten met behulp van schatters voor de stratumtotalen. Een dergelijke schatter die expliciet gebruik maakt van het stratificatieontwerp, noemen we een *stratificatieschatter*. De *stratificatieschatter voor het populatietotaal*, notatie $\hat{Y}_{ST}$, is daarom per definitie gelijk aan:

$$\hat{Y}_{ST} = \sum_{h=1}^{H} \hat{Y}_h = \sum_{h=1}^{H} N_h \bar{y}_h \quad . \tag{3.8}$$

Op analoge wijze definiëren we de *stratificatieschatter voor het populatiegemiddelde*, notatie $\hat{\bar{Y}}_{ST}$, als:

$$\hat{\bar{Y}}_{ST} = \frac{1}{N} \hat{Y}_{ST} = \sum_{h=1}^{H} \left(\frac{N_h}{N} \right) \bar{y}_h \quad . \tag{3.9}$$

Kort samengevat, zowel het populatietotaal als het populatiegemiddelde kan worden geschat met een gewogen som van de stratumgemiddelden. De stratum-afhankelijke gewichtsfactoren zijn respectievelijk de absolute omvang van het stratum en de relatieve omvang van het stratum.

Zuiverheidsanalyses

Per stratum trekken we een EAZT-steekproef van omvang n_h. Een aantal eigenschappen van dit type steekproef is bekend, zie hoofdstuk 2. Zo kunnen we nu zonder meer stellen, dat de schatters voor het stratumgemiddelde en het stratumtotaal, zie (3.7), zuivere schatters zijn. Uitgaande van de zuiverheid van de schatter voor het stratumtotaal, is het mogelijk de verwachtingswaarde van de stratificatieschatter voor het populatietotaal (3.8) te berekenen:

$$E\left(\hat{Y}_{ST}\right) = \sum_{h=1}^{H} E\left(\hat{Y}_h\right) = \sum_{h=1}^{H} Y_h = Y \quad .$$

Stratificatieschatter (3.8) is dus een zuivere schatter voor het populatietotaal. Van dit resultaat maken we gebruik bij het berekenen van de verwachtingswaarde van de stratificatieschatter voor het populatiegemiddelde:

$$E\left(\hat{\bar{Y}}_{ST}\right) = \frac{1}{N} E\left(\hat{Y}_{ST}\right) = \frac{1}{N} Y = \bar{Y} \quad .$$

Hiermee is de zuiverheid van $\hat{\bar{Y}}_{ST}$ als schatter voor het populatiegemiddelde aangetoond.

De bewijzen voor de zuiverheid van schatters (3.8) en (3.9) (i.e. de formele afleidingen van de berekende verwachtingswaarden) die hierboven gegeven zijn, bevestigen wat intuïtief duidelijk is. Immers, als de schatters voor de stratumtotalen zuivere schatters zijn, dan is de gewogen som van de schatters voor de stratumtotalen ook een zuivere schatter. Een zelfde redenering geldt voor de schatters voor de stratumgemiddelden en de gewogen som van de schatters voor de stratumgemiddelden.

Variantieberekeningen

In ieder stratum wordt een EAZT-steekproef van omvang n_h getrokken. De variantie van zeg, de schatter voor het stratumtotaal, is daarom direct te bepalen, zie uitdrukking (2.12):

$$\text{var}\left(\hat{Y}_{yh}\right) = N_h\left(\frac{1-f_h}{f_h}\right)S_{yh}^2 \quad . \tag{3.10}$$

In deze formule is f_h de *steekproeffractie* in stratum h, i.e. $f_h = n_h / N_h$. De EAZT-steekproeven in de gestratificeerde steekproef worden onafhankelijk van elkaar getrokken. Als gevolg daarvan is de variantie van de stratificatieschatter voor het populatietotaal, gelijk aan de som van de varianties van de individuele schatters voor de stratumtotalen. We kunnen dus schrijven, dat:

$$\text{var}\left(\hat{Y}_{ST}\right) = \sum_{h=1}^{H}\text{var}\left(\hat{Y}_h\right) = N\sum_{h=1}^{H}\left(\frac{1-f_h}{f_h}\right)\left(\frac{N_h}{N}\right)S_{yh}^2 \quad . \tag{3.11}$$

Op basis van (3.11) en de natuurlijke relatie tussen de stratificatieschatters voor het populatietotaal en het populatiegemiddelde, kan de variantie van de laatstgenoemde stratificatieschatter als volgt worden uitgerekend

$$\text{var}\left(\hat{\bar{Y}}_{ST}\right) = \frac{1}{N^2}\text{var}\left(\hat{Y}_{ST}\right) = \frac{1}{N}\sum_{h=1}^{H}\left(\frac{1-f_h}{f_h}\right)\left(\frac{N_h}{N}\right)S_{yh}^2 \quad . \tag{3.12}$$

Bij nadere bestudering van formules (3.11) en (3.12) valt op, dat beide varianties klein zijn d.e.s.d.a de individuele aangepaste stratumvarianties S_{yh}^2 klein zijn. Het klein zijn van de aangepaste stratumvariantie wil zeggen, dat de doelvariabele weinig varieert binnen het stratum oftewel, dat het stratum intern *homogeen* is. Daarnaast blijkt de variantie van de schatter afhankelijk te zijn van het gebruikte allocatieschema, aangezien de individuele aangepaste stratumvarianties afhangen van de omvang van de steekproeven.

De variantie van de stratificatieschatter voor het populatietotaal, zie (3.10), is een gewogen lineaire combinatie van de aangepaste stratumvarianties. Een schatter voor deze variantie kan daarom worden verkregen door iedere individuele aangepaste stratumvariantie te schatten. Voor de EAZT-steekproef is het bekend, dat de steekproefvariantie een zuivere schatter is voor de aangepaste populatievariantie. Gebruik makend van dit resultaat krijgen we:

$$\hat{\text{var}}\left(\hat{Y}_{ST}\right) = N\sum_{h=1}^{H}\left(\frac{1-f_h}{f_h}\right)\left(\frac{N_h}{N}\right)\hat{S}_{yh}^2 = N\sum_{h=1}^{H}\left(\frac{1-f_h}{f_h}\right)\left(\frac{N_h}{N}\right)s_{yh}^2 \quad . \tag{3.13}$$

Geheel analoog aan dit resultaat kunnen we de variantie van de stratificatieschatter voor het populatiegemiddelde als volgt schatten:

$$\hat{\text{var}}\left(\hat{\bar{Y}}_{ST}\right) = \frac{1}{N^2}\hat{\text{var}}\left(\hat{Y}_{ST}\right) = \frac{1}{N}\sum_{h=1}^{H}\left(\frac{1-f_h}{f_h}\right)\left(\frac{N_h}{N}\right)s_{yh}^2 \quad . \tag{3.14}$$

Omdat de steekproefvariantie een zuivere schatter is voor de aangepaste populatievariantie bij een EAZT-steekproef, zie formule (2.10) weten we, dat de schatters (3.13) en (3.14) zuiver zijn.

Variatiecoëfficiënten

De variatiecoëfficiënten van de stratificatieschatters voor het populatietotaal en het populatiegemiddelde zijn, analoog aan definitie (2.15), gedefinieerd als:

$$\begin{aligned} CV\left(\hat{Y}_{ST}\right) &= \frac{\sqrt{\text{var}\left(\hat{Y}_{ST}\right)}}{Y} \\ CV\left(\hat{\bar{Y}}_{ST}\right) &= \frac{\sqrt{\text{var}\left(\hat{\bar{Y}}_{ST}\right)}}{\bar{Y}} \quad . \end{aligned} \tag{3.15}$$

Op basis van formules (3.11) en (3.12) kunnen we vaststellen dat bovenstaande variatiecoëfficiënten aan elkaar gelijk zijn, immers:

$$CV\left(\hat{\bar{Y}}_{ST}\right) = \frac{\sqrt{\text{var}\left(\hat{\bar{Y}}_{ST}\right)}}{\bar{Y}} = \frac{\left(\frac{1}{N}\right)\sqrt{\text{var}\left(\hat{Y}_{ST}\right)}}{\left(\frac{1}{N}\right)Y} = CV\left(\hat{Y}_{ST}\right) \quad .$$

De stratificatieschatters voor het populatietotaal en het populatiegemiddelde zijn dus even nauwkeurig.

Voor het bepalen van variatiecoëfficiënt $CV(\hat{Y}_{ST})$, maken we gebruik van uitdrukking (3.11). Het resultaat kunnen we herschrijven in termen van de stratumvariatiecoëfficiënten, zie definitie (3.1). Na kwadratering van de verkregen gelijkheid krijgen we de volgende relatie tussen de variatiecoëfficiënt van de stratificatieschatter van het populatiegemiddelde, en de individuele stratumvariatiecoëfficiënten:

$$CV^2\left(\hat{Y}_{ST}\right) = N\sum_{h=1}^{N}\left(\frac{1-f_h}{f_h}\right)\left(\frac{N_h}{N}\right)\left(\frac{\bar{Y}_h}{\bar{Y}}\right)^2 CV_{yh}^2 \quad .$$

3.3.4 Evaluatie van het steekproefontwerp

Het leidende idee achter het gestratificeerde steekproefontwerp is, dat met behulp van stratificatie een preciezere schatter kan worden verkregen dan bij een enkelvoudig, aselect steekproefontwerp mogelijk is. Wanneer de waarnemingen binnen de meeste strata homogener zijn dan de populatie als geheel zal, op basis van intuïtie, de reductie van de variantie in ieder van deze strata leiden tot een kleinere variantie van de schatter.

Het is mogelijk een steekproefontwerp te evalueren, door de variantie van een schatter op basis van deze steekproef te vergelijken met de variantie van de desbetreffende schatter op basis van een EAZT-steekproef van gelijke omvang (ook wel EAZT-schatter genoemd). De ratio van de twee varianties wordt ook wel het *ontwerpeffect* genoemd, in het Engels 'design effect', en wordt vaak afgekort tot *DEFF*. Voor de stratificatieschatters van het populatietotaal respectievelijk het populatiegemiddelde is het ontwerpeffect gedefinieerd als (onder de noodzakelijke voorwaarde, dat $\text{var}(\hat{Y}_{EAZT}) \neq 0$):

$$\begin{aligned} DEFF\left[\hat{Y}_{ST}\right] &= \frac{\text{var}\left(\hat{Y}_{ST}\right)}{\text{var}\left(\hat{Y}_{EAZT}\right)} \\ DEFF\left[\hat{\bar{Y}}_{ST}\right] &= \frac{\text{var}\left(\hat{\bar{Y}}_{ST}\right)}{\text{var}\left(\hat{\bar{Y}}_{EAZT}\right)} \quad . \end{aligned} \tag{3.16}$$

Bij een DEFF van 1, zijn de varianties van de twee schatters even groot; de precisie van de stratificatieschatter is dan net zo groot als de precisie van de EAZT-schatter. Mutatis mutandis, een gestratificeerde steekproef is efficiënter dan een EAZT-steekproef van dezelfde omvang bij een *DEFF* kleiner dan 1, en minder efficiënt bij een *DEFF* groter dan 1. Om een ontwerpeffect daadwerkelijk te berekenen, zijn van beide varianties expliciete berekeningen nodig.

Het ontwerpeffect voor de stratificatieschatter voor het populatiegemiddelde kunnen we als volgt uitrekenen:

$$DEFF\left[\hat{\bar{Y}}_{ST}\right] = \frac{\text{var}\left(\hat{\bar{Y}}_{ST}\right)}{\text{var}\left(\hat{\bar{Y}}_{EAZT}\right)} = \frac{\left(\frac{1}{N^2}\right)\text{var}\left(\hat{Y}_{ST}\right)}{\left(\frac{1}{N^2}\right)\text{var}\left(\hat{Y}_{EAZT}\right)} = DEFF\left[\hat{Y}_{ST}\right] \quad .$$

Met andere woorden, de zojuist gedefinieerde ontwerpeffecten voor de stratificatieschatters van het populatietotaal en het populatiegemiddelde, zijn aan elkaar gelijk. We kunnen dus volstaan met het uitrekenen van het ontwerpeffect voor de stratificatieschatter van het populatietotaal. Achteraf gezien is dit geen verrassing omdat het ontwerpeffect in feite niets anders is dan het quotiënt van de kwadraten van de variatiecoëfficiënten van de stratificatieschatter respectievelijk, de EAZT-schatter.

Een compacte uitdrukking voor het ontwerpeffect van de stratificatieschatter van het populatietotaal kan worden afgeleid als tussen $\text{var}(\hat{Y}_{ST})$ en $\text{var}(\hat{Y}_{EAZT})$ een eenvoudig verband zou bestaan. Een dergelijk verband bestaat voor de algemene situatie helaas niet. Het bestaat aantoonbaar wel onder twee, niet al te strikte, voorwaarden.

Neem voor het vervolg van de discussie in deze deelparagraaf als eerste aan, dat steekproeffractie f_h constant is. Onder deze aanname vereenvoudigt relatie (3.11) op de volgende wijze:

$$\text{var}\left(\hat{Y}_{ST}\right) = N\sum_{h=1}^{H}\left(\frac{1-f_h}{f_h}\right)\left(\frac{N_h}{N}\right)S_{yh}^2 = N\left(\frac{1-f}{f}\right)\sum_{h=1}^{H}\left(\frac{N_h}{N}\right)S_{yh}^2 \quad . \tag{3.17}$$

De variantie van de EAZT-schatter op basis van een EAZT-steekproef van dezelfde omvang voor het populatietotaal, is afhankelijk van de aangepaste populatievariantie S_y^2, zie (2.12). Neem daarom in tweede instantie aan, dat aan de volgende voorwaarden is voldaan:

$$N_h \approx (N_h - 1) \qquad \wedge \qquad N \approx (N-1) \quad . \tag{3.18}$$

Op basis van deze aanname kan de uitdrukking voor $\text{var}(\hat{Y}_{EAZT})$ als volgt worden herschreven:

$$\text{var}\left(\hat{Y}_{EAZT}\right) \approx \text{var}\left(\hat{Y}_{ST}\right) + \left(\frac{1-f}{f}\right)\sum_{h=1}^{H} N_h\left(\overline{Y}_h - \overline{Y}\right)^2 \quad . \tag{3.19*}$$

Kort samengevat, we kunnen de variantie van de EAZT-schatter voor het populatietotaal herschrijven als de som van de variantie van de stratificatieschatter voor het populatietotaal en een term die rechtevenredig is met de tussenstratavariantie. Dankzij dit resultaat is het mogelijk het ontwerpeffect van de stratificatieschatter voor het populatietotaal bij benadering uit te rekenen, i.e.:

$$DEFF\left[\hat{Y}_{ST}\right] \approx \frac{\text{var}\left(\hat{Y}_{ST}\right)}{\text{var}\left(\hat{Y}_{ST}\right) + \left(\frac{1-f}{f}\right)\sum_{h=1}^{H} N_h\left(\overline{Y}_h - \overline{Y}\right)^2} \quad . \tag{3.20}$$

Uit benadering (3.20) leiden we af, dat het ontwerpeffect van de stratificatieschatter altijd kleiner dan of gelijk is aan 1. Ook blijkt, dat een grotere tussenstratavariantie aanleiding geeft tot stratificatieschatters met een grotere precisie. Het is dus voordelig om bij het ontwerp van de stratificatie te streven naar een zo groot mogelijke spreiding van de stratumgemiddelden $\overline{Y}_h$.

Concluderend, in een gestratificeerde steekproef waarbij de steekproeffractie per stratum constant is en zowel de populatie als de individuele strata voldoende groot zijn, zijn de stratificatieschatters voor het populatietotaal en het populatiegemiddelde minstens zo precies als de overeenkomstige EAZT-schatters.

3.3.5 Allocatie

In voorgaande paragrafen werd telkens een gestratificeerde steekproef getrokken, waarbij zowel de omvang N_h van de strata als de steekproefgrootte n_h in de strata gegeven waren. In de volgende paragraaf zullen we bekijken hoe de strata het best gekozen kunnen worden, maar eerst zullen we nagaan hoeveel waarnemingen we in elk stratum moeten nemen. Het kan zijn dat we stratumgemiddelden of andere stratumparameters met een bepaalde, van te voren vastgelegde nauwkeurigheid willen schatten. In dat geval kunnen we de benodigde steekproefaantallen voor de individuele strata bepalen met behulp van de formules uit paragraaf 2.3.4. Deze steekproeven vormen dan samen de totale steekproef. Vaak ligt echter de totale steekproefomvang al vast en is de vraag hoe deze *n* elementen over de strata te verdelen.

Dit is het zogenaamde allocatieprobleem. In deze paragraaf komen verschillende allocatiemethoden aan bod.

Proportionele allocatie

Proportionele allocatie is gebaseerd op het oorspronkelijke idee van een representatieve steekproef. De gestratificeerde steekproef wordt zodanig getrokken dat de steekproef de populatie reflecteert met betrekking tot de stratificatievariabele(n). In elk stratum wordt een steekproef getrokken waarvan de omvang evenredig is met de grootte van het stratum:

$$n_h = \frac{N_h}{N} \times n \quad . \tag{3.21}$$

Als de op deze wijze berekende stratumomvang n_h geen geheel getal is, wordt zij op een gehele waarde afgerond. Gegeven de niet afgeronde steekproefomvang per stratum, kunnen we de steekproeffractie per stratum uitrekenen als:

$$f_h = \frac{n_h}{N_h} = \frac{n}{N} \quad .$$

Met andere woorden, de steekproeffractie per stratum is constant. Hiermee is automatisch voldaan aan de voorwaarde waaronder (3.17) geldig is. Bovendien geldt, dat resultaten (3.19) – (3.20) waar zijn onder, extra, voorwaarde (3.18).

Ten slotte, de insluitkans van element k in stratum h kunnen we berekenen als:

$$\pi_{hk} = \frac{n_h}{N_h} = \frac{N_h}{N}\frac{n}{N_h} = \frac{n}{N} \quad .$$

Alle elementen in de populatie, ongeacht in welk stratum ze thuishoren, hebben dus dezelfde kans in de steekproef terecht te komen en wegen dus even zwaar mee bij de ophoging. We spreken in dit geval van een *zelfwegende* steekproef.

Bij een EAZT-steekproef heeft ieder element uit de populatie ook een kans van n/N om in de steekproef terecht te komen. Bij een gestratificeerde steekproef zijn echter een aantal extreme steekproeven uitgesloten, die bij een EAZT-steekproef wel mogelijk zijn (bijvoorbeeld een steekproef van alleen hoge of lage inkomens).

Optimale allocatie

Wanneer de verschillende aangepaste stratumvarianties aan elkaar gelijk zijn, is proportionele allocatie de beste allocatiemethode om de precisie te verhogen. In een steekproef waarbij de orde van grootte van de eenheden verschilt (bijv. ondernemingen, scholen, gemeenten) zullen de grotere eenheden over het algemeen meer variëren dan de kleinere eenheden met betrekking tot de doelvariabelen. Denk bijvoorbeeld aan de internationale handel van bedrijven; grotere ondernemingen zullen

grotere export- c.q. importverschillen te zien geven dan kleinere bedrijven. We zijn in dit geval gebaat bij een hoger percentage grotere bedrijven in de steekproef.

Wanneer de aangepaste stratumvarianties sterk variëren leidt de zogenaamde *optimale* allocatie van de steekproef tot minimale varianties bij de schatters voor het populatietotaal en gemiddelde. Bij deze allocatiemethode wordt de steekproefomvang in stratum *h* gelijk genomen aan:

$$n_h = \frac{N_h S_{yh}}{\sum_{h=1}^{H} N_h S_{yh}} n \quad . \tag{3.22}$$

waarbij n_h weer op een gehele waarde wordt afgerond. Bij optimale allocatie zijn de insluitkansen niet gelijk voor alle elementen van de populatie; de kans om in de steekproef getrokken te worden is evenredig met S_{yh} en varieert dus over de strata.

Het aantal elementen dat in een stratum wordt getrokken is groter naarmate het stratum een groter deel uitmaakt van de populatie *en* wanneer het stratum minder homogeen is, i.e. bij een grotere S_{yh}. Bij toepassing van formule (3.22) kan het voorkomen dat de berekende steekproefomvang groter is dan de eigenlijke omvang van het stratum. In dat geval wordt het stratum integraal waargenomen.

Om de optimale allocatie te kunnen bepalen moeten (de verhoudingen tussen) de aangepaste stratumvarianties bekend zijn. In de praktijk zal deze informatie zelden bekend zijn maar op grond van eerder onderzoek kunnen soms benaderingen van de varianties worden gegeven. Als verwacht mag worden dat de stratumvarianties niet al te ver uiteenlopen, zodat men kan veronderstellen $S_{yh} = S_y$, reduceert formule (3.22) tot de proportionele allocatie.

Kosten

Naast de precisie van schatters spelen de kosten natuurlijk ook een belangrijke rol bij steekproefonderzoek. Als de kosten per waarneming verschillen tussen de strata, bijvoorbeeld door gebruik van verschillende waarnemingstechnieken in de strata, zal men hiermee ook rekening willen houden bij de allocatie door minder waarnemingen te verrichten in relatief dure strata. Stel dat een bepaald budget beschikbaar is voor het veldwerk, dat wil zeggen dat de totale kosten ten hoogste *C* mogen bedragen. Een simpele kostenfunctie zou dan kunnen zijn

$$C = c_0 + \sum_{h=1}^{H} n_h c_h \ , \tag{3.23}$$

waarbij c_0 de vaste algemene onkosten zijn en c_h de (variabele) kosten van een waarneming in stratum h. We willen nu de steekproef zodanig over de strata verdelen, dat de variantie van de schatter minimaal is gegeven de totale kosten *C*. Voorwaarde (3.23) vervangt de voorwaarde, dat de totale steekproefomvang gelijk moet

zijn aan n. We kunnen bewijzen, gegeven deze voorwaarde, dat door de steekproefgrootte in stratum h gelijk te kiezen aan

$$n_h = \frac{\frac{N_h S_{yh}}{\sqrt{c_h}}}{\sum_{h=1}^{H} \frac{N_h S_{yh}}{\sqrt{c_h}}} n \quad , \tag{3.24}$$

de variantie van de stratificatieschatter voor het populatietotaal minimaal is. Het bewijs is te vinden in Cochran (1977, paragraaf 5). We trekken dus meer elementen uit een stratum wanneer het stratum een groot deel van de populatie uitmaakt, de variantie binnen het stratum groot is en de waarneming in het stratum goedkoop is. De optimale allocatie (3.22) is in feite een speciaal geval van (3.24) waarbij de kosten per stratum gelijk zijn. Zij wordt dan ook wel de Neyman allocatie genoemd.

3.3.6 Fracties

Als de doelvariabele Y een indicatorvariabele is die slechts de waarden 0 en 1 kan aannemen, dan is het gemiddelde gelijk aan een fractie (zie ook paragraaf 2.4). De fractie elementen met een bepaalde eigenschap (de fractie 'successen') in de steekproef uit stratum h is dus gelijk aan $p_h = \bar{y}_h$. De fractie successen in de populatie kunnen we schatten met behulp van stratificatieschatter (3.9), i.e.:

$$\hat{P}_{ST} = \sum_{h=1}^{H} \left(\frac{N_h}{N} \right) p_h \quad .$$

De variantie van deze stratificatieschatter is uitgerekend in formule (3.12). In praktijksituaties kunnen we deze variantie schatten met behulp van formule (3.14), i.e.:

$$\hat{\text{var}}\left(\hat{P}_{ST}\right) = \sum_{h=1}^{H} \left(1 - f_h\right) \left(\frac{N_h}{N} \right)^2 \left(\frac{p_h (1 - p_h)}{(n_h - 1)} \right)$$

omdat de stratumvariantie kan worden uitgerekend als:

$$s_h^2 = \left(\frac{n_h}{n_h - 1} \right) p_h \left(1 - p_h\right) \quad .$$

3.4 Kwaliteitsindicatoren

Bij een gestratificeerde steekproef is een belangrijk kwaliteitscriterium de variantiereductie ten opzichte van de enkelvoudige aselecte steekproef zonder teruglegging. Dit komt tot uitdrukking in het design effect *DEFF,* besproken in paragraaf 3.3.4. Hoe groter de variantiereductie hoe kleiner de *DEFF*. De variantiereductie is vooral groot wanneer de strata zeer uiteenlopende gemiddelden voor de doelvariabele Y hebben.

3.5 Appendix

Bewijs van (3.4)

Uit definitie (2.1) voor de populatievariantie leiden we af dat:

$$\begin{aligned}\sigma_y^2 &= \frac{1}{N}\sum_{h=1}^{H}\sum_{k=1}^{N_h}\left(Y_{hk}-\overline{Y}\right)^2 \\ &= \frac{1}{N}\sum_{h=1}^{H}\sum_{k=1}^{N_h}\left(Y_{hk}-\overline{Y}_h+\overline{Y}_h-\overline{Y}\right)^2 \\ &= \frac{1}{N}\left[\sum_{h=1}^{H}\sum_{k=1}^{N_h}\left(Y_{hk}-\overline{Y}_h\right)^2+\sum_{h=1}^{H}\sum_{k=1}^{N_h}\left(\overline{Y}_h-\overline{Y}\right)^2\right]+ \\ &\quad +\frac{1}{N}\sum_{h=1}^{H}\sum_{k=1}^{N_h}2\left(Y_{hk}-\overline{Y}_h\right)\left(\overline{Y}_h-\overline{Y}\right) \quad .\end{aligned}$$

Het dubbele product in bovenstaande term rekenen we als volgt uit:

$$\begin{aligned}\sum_{h=1}^{H}\sum_{k=1}^{N_h}\left(Y_{hk}-\overline{Y}_h\right)\left(\overline{Y}_h-\overline{Y}\right) &= \sum_{h=1}^{H}\left(\overline{Y}_h-\overline{Y}\right)\sum_{k=1}^{N_h}\left(Y_{hk}-\overline{Y}_h\right) \\ &= \sum_{h=1}^{H}\left(\overline{Y}_h-\overline{Y}\right)\left[\sum_{k=1}^{N_h}Y_{hk}-N_h\overline{Y}_h\right]=0 \quad .\end{aligned}$$

Voor populatievariantie σ_y^2 vinden we daarom:

$$\begin{aligned}\sigma_y^2 &= \frac{1}{N}\left[\sum_{h=1}^{H}\sum_{k=1}^{N_h}\left(Y_{hk}-\overline{Y}_h\right)^2+\sum_{h=1}^{H}\sum_{k=1}^{N_h}\left(\overline{Y}_h-\overline{Y}\right)^2\right] \\ &= \sum_{h=1}^{H}\left(\frac{N_h}{N}\right)\sigma_{yh}^2+\sum_{h=1}^{H}\left(\frac{N_h}{N}\right)\left(\overline{Y}_h-\overline{Y}\right)^2\end{aligned}$$

volgens definitie (3.1). Hiermee is (3.4) bewezen. ■

Bewijs van (3.5)

Uit de definitie voor de aangepaste populatievariantie in (2.1) leiden we af:

$$\begin{aligned}S_y^2 &= \frac{1}{N-1}\sum_{h=1}^{H}\sum_{k=1}^{N_h}\left(Y_{hk}-\overline{Y}\right)^2 \\ &= \frac{1}{N-1}\left[\sum_{h=1}^{H}\sum_{k=1}^{N_h}\left(Y_{hk}-\overline{Y}_h\right)^2+\sum_{h=1}^{H}\sum_{k=1}^{N_h}\left(\overline{Y}_h-\overline{Y}\right)^2\right]+ \\ &\quad +\frac{1}{N-1}\sum_{h=1}^{H}\sum_{k=1}^{N_h}2\left(Y_{hk}-\overline{Y}_h\right)\left(\overline{Y}_h-\overline{Y}\right) \quad .\end{aligned}$$

Van het dubbelproduct dat in deze laatste expressie voorkomt, weten we dat het gelijk is aan 0. Bovenstaande uitdrukking kan dus als volgt worden vereenvoudigd:

$$S_y^2 = \frac{1}{N-1}\left[\sum_{h=1}^{H}\sum_{k=1}^{N_h}\left(Y_{hk}-\overline{Y}_h\right)^2 + \sum_{h=1}^{H}\sum_{k=1}^{N_h}\left(\overline{Y}_h-\overline{Y}\right)^2\right]$$
$$= \sum_{h=1}^{H}\left(\frac{N_h-1}{N-1}\right)S_{yh}^2 + \sum_{h=1}^{H}\left(\frac{N_h}{N-1}\right)\left(\overline{Y}_h-\overline{Y}\right)^2$$

volgens definitie (3.1) voor S_{yh}^2. Relatie (3.5) is hiermee bewezen. ■

Bewijs van (3.6)

Volgens definitie (2.1) van de populatievariatiecoëfficiënt geldt dat:

$$CV_y^2 = \frac{S_y^2}{\overline{Y}^2} \quad .$$

Gebruik makend van relatie (3.5) voor de aangepaste populatievariantie kunnen we schrijven:

$$CV_y^2 = \sum_{h=1}^{H}\left(\frac{N_h-1}{N-1}\right)\frac{S_{yh}^2}{\overline{Y}^2} + \sum_{h=1}^{H}\left(\frac{N_h}{N-1}\right)\frac{\left(\overline{Y}_h-\overline{Y}\right)^2}{\overline{Y}^2}$$
$$= \sum_{h=1}^{H}\left(\frac{N_h-1}{N-1}\right)\left(\frac{\overline{Y}_h}{\overline{Y}}\right)^2 CV_{yh}^2 + \sum_{h=1}^{H}\left(\frac{N_h}{N-1}\right)\left(\left(\frac{\overline{Y}_h}{\overline{Y}}\right)-1\right)^2 \quad .$$

Hiermee is het bewijs van (3.6) compleet. ■

Bewijs van (3.19)

Relatie (3.5) tussen de aangepaste populatievariantie en de stratumvarianties, kunnen we op basis van benadering (3.18) als volgt herschrijven:

$$S_y^2 = \sum_{h=1}^{H}\left(\frac{N_h-1}{N-1}\right)S_{yh}^2 + \sum_{h=1}^{H}\left(\frac{N_h}{N-1}\right)\left(\overline{Y}_h-\overline{Y}\right)^2$$
$$\approx \sum_{h=1}^{H}\left(\frac{N_h}{N}\right)S_{yh}^2 + \sum_{h=1}^{H}\left(\frac{N_h}{N}\right)\left(\overline{Y}_h-\overline{Y}\right)^2 \quad .$$

De variantie van de EAZT-schatter op basis van een EAZT-steekproef van dezelfde omvang is uit te rekenen volgens formule (2.12). Op basis van bovenstaand resultaat kunnen we de uitkomst op de volgende wijze benaderen:

$$\mathrm{var}\left(\hat{\overline{Y}}_{EAZT}\right) \approx N\left(\frac{1-f}{f}\right)\sum_{h=1}^{H}\left(\frac{N_h}{N}\right)S_{yh}^2 + \left(\frac{1-f}{f}\right)\sum_{h=1}^{H}N_h\left(\overline{Y}_h-\overline{Y}\right)^2$$
$$= \mathrm{var}\left(\hat{\overline{Y}}_{ST}\right) + \left(\frac{1-f}{f}\right)\sum_{h=1}^{H}N_h\left(\overline{Y}_h-\overline{Y}\right)^2 \quad .$$

Hiermee is het bewijs van relatie (3.19) voltooid. ■

4. Clustersteekproeven en Tweetrapssteekproeven

4.1 Korte beschrijving

Het uitgangspunt bij de twee types steekproeven die in dit hoofdstuk worden besproken is een opsplitsing van de doelpopulatie in zogenaamde *clusters*. Deze opsplitsing is dekkend en zodanig uitgevoerd, dat de clusters elkaar niet overlappen.

In een clustersteekproef wordt een aantal clusters geselecteerd op basis van een kanssteekproef. Vervolgens wordt ieder van de getrokken clusters in zijn geheel waargenomen. De clustersteekproef kan daarom worden geïnterpreteerd als een kanssteekproef van groepen. Ook nu hebben we de keus om de clusters met of zonder teruglegging te trekken.

In een tweetrapssteekproef worden in de eerste trap, net als in de clustersteekproef, clusters geselecteerd op basis van een kanssteekproef. Vervolgens worden in de tweede trap van de tweetrapssteekproef in iedere geselecteerde cluster alleen de via een nieuwe kanssteekproef geselecteerde populatie-eenheden waargenomen.

4.2 Toepasbaarheid

In het vorige hoofdstuk is stratificatie geïntroduceerd als een steekproefmethode waarbij we eerst de populatie opsplitsen in deelpopulaties (de strata), vervolgens uit ieder stratum apart een steekproef trekken. Een dergelijke gestratificeerde steekproef vergroot over het algemeen de precisie van schatters. Hoewel de clustersteekproef op het eerste gezicht erg veel lijkt op de gestratificeerde steekproef, heeft zij heel andere eigenschappen. Zo leidt de keuze voor de clustersteekproef meestal tot een afname van de precisie vergeleken met die van een enkelvoudige, aselecte steekproef. Clustersteekproeven worden dan ook alleen toegepast wanneer de praktijk daartoe noodzaakt of wanneer het verlies in precisie wordt gecompenseerd door een aanzienlijke reductie in de (waarnemings)kosten. Dit verlies aan precisie komt doordat in tegenstelling met de gestratificeerde steekproef bij een clustersteekproef niet alle deelpopulaties worden geselecteerd.

Soms kunnen de schatters evenwel weer aan precisie winnen door in een tweede trap steekproeven te trekken uit de clusters die in de eerste trap van de steekproef zijn getrokken. Omdat de clusters niet meer integraal worden waargenomen, zou men kunnen overwegen in de eerste trap wat meer clusters te trekken zodat een beter beeld van de overall populatie ontstaat. Een tweetrapssteekproef is vooral aan te bevelen wanneer de clusters homogeen zijn voor wat betreft de doelvariabele.

Voorbeeld van een tweetrapssteekproef is een clustering van alle huishoudens in Nederland naar regio's of wijken. Uit iedere wijk, die in de eerste trap is getrokken, wordt vervolgens in de 2^{e} trap een steekproef van huishoudens getrokken. Het voordeel van een dergelijke, regionale clustersteekproef is dan dat de totale reiskosten van de enquêteurs / enquêtrices omlaag kunnen gaan in vergelijking met een aselec-

te steekproef uit alle Nederlandse huishoudens. De reiskosten kunnen omlaag gaan omdat een enquêteur / enquêtrice nu relatief veel personen kan interviewen die in clusters dichtbij elkaar wonen.

Bij de tot nu toe besproken steekproefontwerpen gingen we er steeds van uit dat er een geschikt steekproefkader voor de doelpopulatie aanwezig is. Dit zal natuurlijk niet altijd het geval zijn. Het kan echter voorkomen dat er wel kaders beschikbaar zijn voor deelpopulaties waarin de doelpopulatie kan worden opgesplitst. Stel, we willen een onderzoek doen onder middelbare scholieren. Er is geen lijst met alle middelbare scholieren voorhanden maar we kunnen misschien wel op een relatief eenvoudige manier achter een lijst met middelbare scholen komen. Per school zijn natuurlijk wel de leerlingen bekend. Andere voorbeelden van 'natuurlijke' groepen in de populatie zijn gemeenten, huishoudens, fabrieken, verzorgingstehuizen etc. Deze groepen worden dan clusters genoemd.

4.3 Uitgebreide beschrijving

4.3.1 Veronderstellingen en notatie

Met betrekking tot doelvariabele Y onderscheiden we voor de populatie de vijf parameters gedefinieerd in (2.1). Ook voor de getrokken steekproefelementen beschouwen we, analoog aan hoofdstuk 2, het drietal parameters gedefinieerd in (2.2).

Net als in het gestratificeerde steekproefontwerp wordt bij een clustersteekproef de populatie onderverdeeld in deelpopulaties. Een deelpopulatie wordt een *cluster* of *primaire eenheid* genoemd (*primary sampling units* (psu's)). Het impliciete uitgangspunt is dat de onderverdeling van de populatie dekkend is, en dat de individuele clusters geen elementen gemeenschappelijk hebben.

In het steekproefontwerp worden de clusters in eerste instantie geïnterpreteerd als steekproefeenheden (daarom de eis, dat de clusters in de populatie elkaar niet overlappen). Een specifieke cluster wordt aangegeven met de index d, $d = 1, \cdots, N$; het aantal elementen in cluster d geven we weer met M_d. De elementen uit een primaire eenheid worden de s*ecundaire eenheden* genoemd (*secondary sampling units* (ssu's)). De aanname is dat clusteromvang M_d in principe bekend is.

Het totale aantal elementen in de populatie geven we weer met *M.* Er geldt dus

$$\sum_{d=1}^{N} M_d = M \quad .$$

Deze totale populatieomvang kunnen we middelen over het aantal clusters. Het resultaat noemen we de *gemiddelde clusteromvang*, notatie $\overline{M}$:

$$\overline{M} = \frac{M}{N} \quad .$$

De waarde van de doelvariabele voor element k in cluster d wordt genoteerd als Y_{dk} voor $d = 1,\cdots,N$ en $k = 1,\cdots,M_d$. Voor de doelvariabele onderscheiden we per cluster de volgende parameters: het *clustertotaal* Y_d, het clustergemiddelde $\overline{Y}_d$, de *clustervariantie* σ^2_{yd}, en de *aangepaste clustervariantie* S^2_{yd}. Deze parameters zijn gedefinieerd als:

$$\begin{aligned} Y_d &= \sum_{k=1}^{M_d} Y_{dk} \\ \overline{Y}_d &= \frac{1}{M_d} Y_d \\ \sigma^2_{yd} &= \frac{1}{M_d} \sum_{k=1}^{M_d} \left(Y_{dk} - \overline{Y}_d\right)^2 \\ S^2_{yd} &= \frac{1}{M_d - 1} \sum_{k=1}^{M_d} \left(Y_{dk} - \overline{Y}_d\right)^2 \quad . \end{aligned} \tag{4.1}$$

Aan deze vier clusterparameters voegen we nog het *gemiddelde clustertotaal* $\overline{Y}_{CT}$, de *variantie van clustertotalen* σ^2_{CT}, en de *aangepaste variantie van clustertotalen* S^2_{CT} als parameters toe. Zij zijn als volgt gedefinieerd:

$$\begin{aligned} \overline{Y}_{CT} &= \frac{1}{N} Y = \frac{1}{N} \sum_{d=1}^{N} Y_d \\ \sigma^2_{CT} &= \frac{1}{N} \sum_{d=1}^{N} \left(Y_d - \overline{Y}_{CT}\right)^2 \\ S^2_{CT} &= \frac{1}{N - 1} \sum_{d=1}^{N} \left(Y_d - \overline{Y}_{CT}\right)^2 \quad . \end{aligned} \tag{4.2}$$

Uit de populatie, die is opgedeeld in N clusters, wordt een enkelvoudige, aselecte steekproef van omvang n getrokken. Elk van de geselecteerde clusters wordt volledig waargenomen; met andere woorden, per cluster zijn er m_d waarnemingen met $m_d = M_d$. We kunnen de aselecte clustersteekproef dus interpreteren als een enkelvoudige, aselecte steekproef van n uit N primaire eenheden met de clustertotalen als waarnemingen.

De gemeten waarde van de doelvariabele voor element k in getrokken cluster d wordt weergegeven met y_{dk}. Voor ieder getrokken cluster d zijn het *steekproeftotaal* y_d, het *steekproefgemiddelde* $\overline{y}_d$, en de *steekproefvariantie* s^2_{yd} formeel gedefinieerd in uitdrukking (4.3). Merk op, dat deze drie steekproefparameters overeenkomen met de corresponderende clusterparameters aangezien ieder geselecteerd cluster volledig wordt waargenomen.

$$
\begin{aligned}
y_d &= \sum_{k=1}^{m_d} y_{dk} = Y_d \\
\bar{y}_d &= \frac{1}{m_d} \sum_{k=1}^{m_d} y_{dk} = \bar{Y}_d \\
s_{yd}^2 &= \frac{1}{m_d - 1} \sum_{k=1}^{m_d} \left(y_{dk} - \bar{y}_d \right)^2 = S_{yd}^2 \quad .
\end{aligned}
\tag{4.3}
$$

Aan deze lijst van steekproefparameters in cluster d, voegen we nog het *steekproefgemiddelde van clustertotalen* $\bar{y}_{CT}$, de *steekproefvariantie van clustertotalen* s_{CT}^2 toe:

$$
\begin{aligned}
\bar{y}_{CT} &= \frac{1}{n} \sum_{d=1}^{n} y_d \\
s_{CT}^2 &= \frac{1}{n-1} \sum_{l=1}^{n} \left(y_d - \bar{y}_{CT} \right)^2 \quad .
\end{aligned}
$$

4.3.2 Relaties tussen populatieparameters en clusterparameters

Tussen het populatietotaal en het populatiegemiddelde enerzijds en respectievelijk het clustertotaal en het clustergemiddelde anderzijds, bestaan directe verbanden. Deze verbanden zijn hieronder weergegeven.

$$
\begin{aligned}
Y &= \sum_{d=1}^{N} \sum_{k=1}^{M_d} Y_{dk} = \sum_{d=1}^{N} Y_d \\
\bar{Y} &= \frac{Y}{M} = \frac{1}{M} \sum_{d=1}^{N} Y_d = \frac{1}{M} \sum_{d=1}^{N} M_d \bar{Y}_d = \sum_{d=1}^{N} \left(\frac{M_d}{M} \right) \bar{Y}_d \quad .
\end{aligned}
$$

Volgens bovenstaande relaties komt het populatietotaal overeen met de som van alle clustertotalen en is het populatiegemiddelde te interpreteren als een gewogen som van de clustergemiddelden. Beide verbanden zijn intuïtief duidelijk.

Voor een clustersteekproef kan de populatievariantie worden bepaald op basis van de varianties in de deelpopulaties, net als bij de gestratificeerde steekproef, zie hoofdstuk 3. De volgende relatie kan worden aangetoond:

$$
\sigma_y^2 = \sum_{d=1}^{N} \left(\frac{M_d}{M} \right) \sigma_{yd}^2 + \sum_{d=1}^{N} \left(\frac{M_d}{M} \right) \left(\bar{Y}_d - \bar{Y} \right)^2 \tag{4.4*}
$$

Uitgaande van (4.4), kunnen we een relatie afleiden tussen de aangepaste populatievariantie S_y^2 en de individuele aangepaste clustervarianties S_{yd}^2:

$$
S_y^2 = N \left(\frac{\overline{M} - 1}{M - 1} \right) S_{binnen}^2 + \overline{M} \left(\frac{N - 1}{M - 1} \right) S_{tussen}^2 \tag{4.5*}
$$

met:

$$S^2_{binnen} = \frac{1}{N}\sum_{d=1}^{N}\left(\frac{M_d - 1}{\overline{M} - 1}\right)S^2_{yd}$$

$$S^2_{tussen} = \frac{1}{N-1}\sum_{d=1}^{N}\left(\frac{M_d}{\overline{M}}\right)\left(\overline{Y}_d - \overline{Y}\right)^2 \quad .$$

De eerste term (aan de rechterkant van het '='-teken) in (4.5) is een veelvoud van S^2_{binnen}, dat de *binnenclustervariantie* wordt genoemd. De binnenclustervariantie is een samenstelling van de N aangepaste clustervarianties. De tweede term (rechts van het '='-teken) in (4.5) is rechtevenredig met S^2_{tussen}, welke de *tussenclustervariantie* wordt genoemd. Deze tussenclustervariantie hangt af van het verschil van het clustergemiddelde en het populatiegemiddelde voor iedere cluster.

4.3.3 Schatters voor het populatiegemiddelde en het populatietotaal

De clustersteekproef is, in essentie, een enkelvoudige, aselecte steekproef met de clusters als eenheden en de clustertotalen als waarnemingen. Als we bovendien aannemen dat de steekproef waarmee de waar te nemen clusters worden geselecteerd zonder teruglegging wordt getrokken, dan is de clustersteekproef niets anders dan een EAZT-steekproef van (cluster)totalen. Dit type steekproef hebben we uitvoerig geanalyseerd in hoofdstuk 2. Ten slotte, een schatter voor het populatietotaal of het populatiegemiddelde, die is gebaseerd op de specifieke interpretatie van de clustersteekproef als een steekproef van clustertotalen, noemen we een *clusteringsschatter*.

Het uitgangspunt voor de analyse in deze deelparagraaf is een clustersteekproef die overeenkomt met een EAZT-steekproef van cluster(totalen). Gegeven dit uitgangspunt formuleren we de volgende schatter voor het gemiddelde clustertotaal:

$$\hat{\overline{Y}}_{CT} = \overline{y}_{CT} = \frac{1}{n}\sum_{d=1}^{n} y_d \quad .$$

Voor de aangepaste variantie van clustertotalen kan, dankzij ditzelfde uitgangspunt, zonder omwegen de hieronder weergegeven schatter worden geponeerd.

$$\hat{S}^2_{CT} = s^2_{CT} = \frac{1}{n-1}\sum_{d=1}^{n}\left(y_d - \overline{y}_{CT}\right)^2 \; .$$

Tussen het populatietotaal en het gemiddelde clustertotaal $\overline{Y}_{CT}$ bestaat een eenvoudige relatie, die kan worden afgeleid uit (2.1) en (4.2). Een soortgelijke relatie bestaat daarom ook tussen het populatiegemiddelde en het gemiddelde clustertotaal. Beide relaties zijn hieronder weergegeven.

$$Y = N\,\overline{Y}_{CT} \qquad \wedge \qquad \overline{Y} = \frac{1}{\overline{M}}\,\overline{Y}_{CT} \quad . \tag{4.6}$$

Op basis van deze relaties en op basis van de schatter voor het gemiddelde clustertotaal komen we tot een clusteringsschatter voor het populatietotaal Y, notatie $\hat{Y}_{CS}$, en een clusteringsschatter voor het populatiegemiddelde, notatie $\hat{\bar{Y}}_{CS}$.

$$\begin{aligned}\hat{Y}_{CS} &= N\,\hat{\bar{Y}}_{CT} = N\,\bar{y}_{CT} = \left(\frac{1}{f}\right)\sum_{d=1}^{n} y_d \\ \hat{\bar{Y}}_{CS} &= \frac{1}{M}\hat{\bar{Y}}_{CT} = \frac{1}{M}\bar{y}_{CT} = \left(\frac{1}{f}\right)\left(\frac{1}{M}\right)\sum_{d=1}^{n} y_d \quad .\end{aligned} \tag{4.7}$$

In deze twee formules geldt, dat $f = n/N$ de fractie clusters in de steekproef vertegenwoordigt. Merk op, dat het gemiddelde steekproefclustertotaal $\bar{y}_{CT}$ op natuurlijke wijze in deze schatters te voorschijn komt.

Zuiverheidsanalyses

Omdat de clustersteekproef een EAZT-steekproef van clustertotalen is, weten we dat de schatter voor het gemiddelde clustertotaal een zuivere schatter is. Op grond van hetzelfde argument mogen we ook concluderen, dat de steekproefvariantie van clustertotalen een zuivere schatter is voor de aangepaste variantie van clustertotalen.

Per definitie zijn de clusteringsschatter voor het populatietotaal en de clusteringsschatter voor het populatiegemiddelde geschaalde versies van de schatter voor het gemiddelde clustertotaal. De zuiverheid van de schatter voor het gemiddelde clustertotaal samen met dit gegeven en de afhankelijkheidsrelaties (4.6) maakt, dat de clusteringsschatters voor het populatietotaal, respectievelijk, het populatiegemiddelde in (4.7), beide zuivere schatters zijn.

Variantieberekeningen

We hebben gezien dat de zuiverheid van de schatter voor het gemiddelde clustertotaal ten grondslag ligt aan de zuiverheid van de twee clusteringsschatters. Zo ook ligt de variantieberekening voor de schatter van het gemiddelde clustertotaal ten grondslag aan de variantieberekeningen voor de beide clusteringsschatters.

De variantie van de schatter voor het gemiddelde clustertotaal kunnen we als volgt berekenen, zie ook (2.11):

$$\operatorname{var}\left(\hat{\bar{Y}}_{CT}\right) = \frac{1}{N}\left(\frac{1-f}{f}\right)S_{CT}^2 \quad . \tag{4.8}$$

Uit dit resultaat leiden we af, dat de variantie van de clusteringsschatter voor het populatietotaal, gelijk is aan:

$$\operatorname{var}\left(\hat{Y}_{CS}\right) = N^2 \operatorname{var}\left(\hat{\bar{Y}}_{CT}\right) = N\left(\frac{1-f}{f}\right)S_{CT}^2 \tag{4.9}$$

en, dat voor de variantie van clusteringsschatter voor het populatiegemiddelde geldt:

$$\mathrm{var}\left(\hat{\bar{Y}}_{CS}\right) = \mathrm{var}\left(\frac{1}{M}\hat{Y}_{CS}\right) = \frac{1}{M^2}\mathrm{var}\left(\hat{Y}_{CS}\right) = \left(\frac{1}{M}\right)\left(\frac{1}{\overline{M}}\right)\left(\frac{1-f}{f}\right)S_{CT}^2 \quad . \tag{4.10}$$

Schatters voor resultaten (4.8) - (4.10) kunnen worden gemaakt door in de desbetreffende formules S_{CT}^2 te vervangen door s_{CT}^2. Het is namelijk bekend dat de steekproefvariantie van clustertotalen een zuivere schatter is voor de aangepaste variantie van clustertotalen. De schatters voor de varianties (van de clusteringsschatters) die op deze wijze ontstaan, zijn dus zuivere schatters.

4.3.4 Evaluatie van het steekproefontwerp

Ter evaluatie van de clustersteekproef willen we de ontwerpeffecten van de clusteringsschatters voor het populatietotaal en het populatiegemiddelde berekenen. Een dergelijk ontwerpeffect vergelijkt de variantie van een clusteringsschatter met de variantie van de EAZT-schatter op basis van een EAZT-steekproef van dezelfde omvang. Helaas is de omvang van een clustersteekproef niet van te voren bekend. Gemiddeld genomen bestaat een clustersteekproef echter uit $(n \times \overline{M})$ individuele waarnemingen. Bij een vergelijking met de EAZT-schatter op basis van een EAZT-steekproef van dezelfde omvang, gaan we dan ook uit van een EAZT-steekproef van $(n \times \overline{M})$ uit M elementen.

De ontwerpeffecten van de clusteringsschatters voor het populatietotaal en het populatiegemiddelde zijn, analoog aan (3.16), gedefinieerd als (neem impliciet aan, dat geldt: $\mathrm{var}(\hat{Y}_{EAZT}) \neq 0$):

$$\begin{aligned} DEFF\left(\hat{Y}_{CS}\right) &= \frac{\mathrm{var}\left(\hat{Y}_{CS}\right)}{\mathrm{var}\left(\hat{Y}_{EAZT}\right)} \\ DEFF\left(\hat{\bar{Y}}_{CS}\right) &= \frac{\mathrm{var}\left(\hat{\bar{Y}}_{CS}\right)}{\mathrm{var}\left(\hat{\bar{Y}}_{EAZT}\right)} \quad . \end{aligned} \tag{4.11}$$

In definitie (4.11) zijn het ontwerpeffect van de clusteringsschatter voor het populatietotaal en het ontwerpeffect van de clusteringsschatter voor het populatiegemiddelde op formele wijze vastgelegd. Het hieronder weergegeven resultaat toont echter aan, dat het volstaat om in het vervolg van deze deelparagraaf nog slechts over het ontwerpeffect van de clusteringsschatter voor het populatietotaal te spreken.

$$DEFF\left[\hat{\bar{Y}}_{CS}\right] = \frac{\mathrm{var}\left(\hat{\bar{Y}}_{CS}\right)}{\mathrm{var}\left(\hat{\bar{Y}}_{EAZT}\right)} = \frac{\left(\frac{1}{M^2}\right)\mathrm{var}\left(\hat{Y}_{CS}\right)}{\left(\frac{1}{M^2}\right)\mathrm{var}\left(\hat{Y}_{EAZT}\right)} = DEFF\left[\hat{Y}_{CS}\right] \quad .$$

Voor een EAZT-schatter op basis van een EAZT-steekproef van dezelfde omvang, is de variantie van de schatter voor het populatietotaal gelijk aan:

$$\mathrm{var}\left(\hat{Y}_{EAZT}\right) = M\left(\frac{1-f}{f}\right)S_y^2$$

met $f = (n \times \overline{M}) / M = n / N$. Het ontwerpeffect van de clusteringsschatter voor het populatietotaal kunnen we op basis van dit resultaat en formule (4.9) nu als volgt uitrekenen:

$$DEFF\left[\hat{Y}_{CS}\right] = \frac{\mathrm{var}\left(\hat{Y}_{CS}\right)}{\mathrm{var}\left(\hat{Y}_{EAZT}\right)} = \frac{N\left(\frac{1-f}{f}\right)S_{CT}^2}{M\left(\frac{1-f}{f}\right)S_y^2} = \frac{1}{\overline{M}}\frac{S_{CT}^2}{S_y^2} \quad . \tag{4.12}$$

Het ontwerpeffect wordt dus bepaald door de verhouding van de aangepaste variantie van clustertotalen en de aangepaste populatievariantie. De aangepaste populatievariantie ligt vast en wordt niet beïnvloed door het steekproefontwerp. Dit is anders voor de aangepaste variantie van clustertotalen. De laatstgenoemde variantie is volledig bepaald door het ontwerp van de clustersteekproef.

Een gedetailleerde analyse van het ontwerpeffect is mogelijk onder een aanvullende aanname. Deze aanname garandeert een eenvoudig verband tussen S_{CT}^2 en S_y^2. Een gedetailleerde analyse van het ontwerpeffect is nu mogelijk geworden. In het vervolg van deze deelparagraaf maken we daarom deze aanname.

Clusters van gelijke omvang

We zetten de analyse van het ontwerpeffect voort onder de aanname, dat alle clusters evenveel elementen bevatten. Als gevolg van deze aanname weten we, dat:

$$M_d = \overline{M} = \frac{M}{N} \quad .$$

De resultaten voor de aselecte EAZT-steekproef van clustertotalen vereenvoudigen onder deze aanname aanzienlijk. Zo kan uitdrukking (4.5) voor de totale aangepaste variantie als volgt worden opgeschreven:

$$S_y^2 = \left(\frac{M-N}{M-1}\right)S_{binnen}^2 + \left(\frac{1}{\overline{M}}\right)\left(\frac{N-1}{M-1}\right)S_{CT}^2 \quad . \tag{4.13*}$$

Deze relatie levert het gezochte verband tussen de aangepaste variantie van clustertotalen en de aangepaste populatievariantie. Hiermee kunnen we formule (4.12) verder uitwerken. Kijkend naar formules (4.5) en (4.13) concluderen we, dat onder de huidige aanname, de tussenclustervariantie recht evenredig is met de aangepaste variantie van clustertotalen S_{CT}^2.

Verband (4.13) stelt ons in staat het ontwerpeffect van de clusteringsschatter voor het populatietotaal expliciet uit te rekenen. Het resultaat van de berekening is hieronder weergegeven:

$$DEFF\left[\hat{Y}_{CS}\right] = -\left(\frac{M-N}{N-1}\right)\left(\frac{S^2_{binnen}}{S^2_y}\right)+\left(\frac{M-1}{N-1}\right) \quad . \tag{4.14*}$$

De eerste term aan de rechterkant van het '='-teken is rechtevenredig met het quotient van de binnenclustervariantie en de populatievariantie. Met andere woorden, het ontwerpeffect (van de clusteringsschatter voor het populatietotaal) is een dalende lineaire functie in binnenclustervariantie S^2_{binnen}.

Volgens formule (4.14) bestaat er een indirecte relatie tussen de wijze van clusteren en het ontwerpeffect van de schatter, omdat de binnenclustervariantie afhankelijk is van clusterontwerp. Dit betekent, dat het ontwerpeffect maximaal is bij een minimale binnenclustervariantie en, dat het minimaal is bij een maximale binnenclustervariantie. Het bereik van het ontwerpeffect is daarom te bepalen als:

$$0 \le DEFF\left[\hat{Y}_{CS}\right] \le \left(\frac{M-1}{N-1}\right) \quad . \tag{4.15*}$$

De ondergrens van het bereik van het ontwerpeffect (van de clusteringsschatter voor het populatietotaal) wordt aangenomen d.e.s.d.a. de tussenclustervariantie gelijk is aan nul, zie ook relatie (4.5). In de praktijk blijkt een clustersteekproef vaak minder efficiënt te zijn dan een EAZT-steekproef. Dit verlies aan efficiëntie moet dan ook gecompenseerd worden door een kostenbesparing, willen we een clustersteekproef daadwerkelijk gaan gebruiken.

In algemene zin kunnen we stellen, dat een grotere binnenclustervariantie aanleiding geeft tot een clusteringsschatter (voor het populatietotaal) met een grotere precisie. Een grotere binnenclustervariantie komt, op zijn beurt, overeen met een kleinere interne homogeniteit van de clusters. Het is dus aan te bevelen om de clusters zo divers mogelijk samen te stellen.

In een alternatieve analyse van het ontwerpeffect van de clusteringsschatter voor het populatietotaal maken we gebruik van de, zogenaamde, *intraklassecorrelatiecoëfficiënt*. In het vervolg van deze deelparagraaf zullen we deze alternatieve analyse nader uitwerken.

De *intraklassecorrelatiecoëfficiënt* ρ_c is gedefinieerd als de Pearson correlatiecoëfficiënt voor de $N \times \overline{M}(\overline{M}-1)$ paren waarnemingen (Y_{dk}, Y_{dl}) binnen de clusters waarbij $k \neq l$ en $d = 1, \cdots, N$. De formele definitie luidt:

$$\rho_c = \frac{\sum_{d=1}^{N}\sum_{k=1}^{\overline{M}}\sum_{l\neq k}^{\overline{M}}\left(Y_{dk}-\overline{Y}\right)\left(Y_{dl}-\overline{Y}\right)}{\left(M-1\right)\left(\overline{M}-1\right)S^2_y} \tag{4.16}$$

De intraklassecorrelatiecoëfficiënt of *intraclustercorrelatiecoëfficiënt* is een maat voor de homogeniteit binnen de clusters. Hij geeft aan hoe gelijk of hoe verschillend

de elementen in een cluster zijn. Maximale homogeniteit van de clusters correspondeert met $\rho_c = 1$. Mutatis mutandis, minimale homogeniteit van de cluster komt overeen met een maximale tussenclustervariantie.

Nader onderzoek toont aan, dat het mogelijk is de intraclustercorrelatiecoëfficiënt uit te drukken in termen van het quotiënt van de aangepaste variantie van clustertotalen en de aangepaste populatievariantie, i.e.:

$$\rho_c = \left(\frac{N-1}{(M-1)(\overline{M}-1)} \right) \left(\frac{S_{CT}^2}{S_y^2} \right) - \frac{1}{\overline{M}-1} \quad . \tag{4.17*}$$

Uit dit resultaat blijkt, dat de intraclustercorrelatiecoëfficiënt een lineaire, stijgende functie is in de aangepaste variantie van clustertotalen S_{CT}^2. Met andere woorden, ρ_c is een stijgende functie in de tussenclustervariantie S_{tussen}^2, omdat laatstgenoemde een (positief) veelvoud is van S_{CT}^2. Concluderend, coëfficiënt ρ_c is minimaal bij de kleinst mogelijke tussenclustervariantie en hij is maximaal bij de grootst mogelijke tussenclustervariantie. Het bereik van de intraclustercorrelatiecoëfficiënt is zodoende direct te bepalen (zie onderstaande uitdrukking voor het resultaat).

$$-\left(\frac{1}{\overline{M}-1} \right) \leq \rho_c \leq 1 \quad . \tag{4.18*}$$

Ten slotte, door gebruik te maken van relatie (4.16) kunnen we het ontwerpeffect van de clusteringsschatter voor het populatietotaal schrijven als een functie van ρ_c. Analyse wijst uit, dat geldt:

$$DEFF\left[\hat{Y}_{CS}\right] = \left(\frac{(\overline{M}-1)(M-1)}{M-\overline{M}} \right) \rho_c + \frac{M-1}{M-\overline{M}} \quad . \tag{4.19*}$$

Het ontwerpeffect is dus een lineaire functie van ρ_c. Omdat het hier een lineair stijgend verband betreft, is het ontwerpeffect minimaal voor de kleinst mogelijke intraclustercorrelatiecoëfficiënt. Mutatis mutandis, het ontwerpeffect is maximaal voor de grootst mogelijke intraclustercorrelatiecoëfficiënt.

In de praktijk is de intraklassecorrelatiecoëfficiënt meestal positief. Vooral wanneer de clusters 'natuurlijke' groepen in de populatie vormen zullen de elementen binnen een cluster meer op elkaar lijken dan elementen die willekeurig uit de populatie zijn getrokken. Dit kan bijvoorbeeld het gevolg zijn van een gezamenlijke achtergrond of omgeving. Wanneer de elementen binnen een cluster op elkaar lijken zal binnenclustervariantie S_{binnen}^2 relatief klein zijn ten opzichte van aangepaste populatievariantie S_y^2, met als gevolg dat ρ_c positief zal zijn. Alleen als de elementen in een cluster een grotere spreiding vertonen dan een willekeurig gekozen groep elementen, is ρ_c negatief. Dit zal zelden het geval zijn maar kan voorkomen bij kunstmatig gevormde clusters waar elementen aselect aan clusters zijn toegewezen.

4.3.5 De tweetrapssteekproef

In de clustersteekproef zoals we die tot nu toe hebben behandeld onderzoeken we steeds alle elementen van ieder getrokken cluster. We hebben gezien, dat een clustersteekproef niet erg efficiënt is als de clusters vrij homogeen zijn. Wanneer de elementen binnen een cluster erg op elkaar lijken verspillen we in feite tijd en geld door alle elementen te onderzoeken; we hadden er ook een paar kunnen onderzoeken om dezelfde informatie te verkrijgen. Het rendement van een clustersteekproef is in dat geval dus vrij laag. In dit soort situaties kan men er toe overgaan om binnen de geselecteerde clusters een steekproef te nemen. We spreken dan van een *tweetrapssteekproef*.

In de eerste trap van een tweetrapssteekproef selecteren we clusters, ook wel *primaire eenheden* genoemd, op basis van een kanssteekproef. In de tweede trap van een tweetrapssteekproef trekken we uit iedere geselecteerde primaire eenheid, een kanssteekproef van elementen die we nu als *secundaire eenheden* aanduiden.

Onder de aanname, dat de clusters een homogeen karakter hebben, leidt een tweetrapssteekproef er in het algemeen toe, dat meer clusters in de steekproef kunnen worden getrokken in de eerste trap. Dit kan een verhogend effect op de precisie hebben. Aldus heeft men een betere controle over de steekproefomvang wanneer de primaire eenheden sterk in omvang verschillen. Merk op, dat ook nu weer alleen voor de getrokken primaire eenheden een steekproefkader vereist is.

Tweetrapssteekproeven liggen qua kosten en precisie een beetje tussen clustersteekproeven en gestratificeerde steekproeven in. Een tweetrapssteekproef is doorgaans duurder dan een even grote clustersteekproef maar de kosten zijn minder dan voor een gestratificeerde steekproef. Anderzijds is een tweetrapssteekproef in de regel preciezer dan een clustersteekproef en minder precies dan een gestratificeerde steekproef.

Een tweetrapssteekproef kan gemakkelijk worden uitgebreid naar meertrapssteekproeven (drie-, viertraps enz). Om een steekproef van middelbare scholieren te trekken kunnen we bijvoorbeeld eerst een steekproef van scholen trekken, vervolgens binnen een geselecteerde school een aantal klassen trekken en ten slotte binnen een geselecteerde klas een aantal leerlingen trekken. Een ander voorbeeld is een steekproef van achtereenvolgens regio's, gemeenten, postcodegebieden en adressen. De eenheden in de laatste trap zijn de elementen. We zullen ons hier beperken tot tweetrapssteekproeven, vanwege de eenvoud maar ook omdat dit ontwerp in de praktijk vaker voorkomt dan meertrapssteekproeven.

In het ontwerp van de tweetrapssteekproef moet men de selectiemethoden voor zowel de primaire als de secundaire eenheden vastleggen. Deze methoden hoeven niet aan elkaar gelijk te zijn. Per trap kan men de eenheden op verschillende manieren trekken: of enkelvoudig en aselect, of systematisch, of evenredig met de omvang (zie paragrafen 4.3.6 en 5.1.2 voor details). Ook kan men de eenheden eerst stratificeren.

Behalve de trekkingsmethoden in de beide trappen moet men ook een beslissing nemen over de verdeling van de steekproef over de eerste en de tweede trap. Trekt men relatief veel primaire eenheden met weinig secundaire eenheden of kiest men voor relatief weinig primaire eenheden met een groot aantal secundaire eenheden? In het eerste geval zal men doorgaans preciezere schatters verkrijgen maar daar staat tegenover dat de kosten ook vaak hoger zijn.

Veronderstellingen en notaties

In de huidige analyse gaan we er van uit, dat in beide trappen een EAZT-steekproef wordt getrokken. De gebruikte notatie is vrijwel analoog aan die voor het cluster-steekproefontwerp. Wederom is de populatie opgesplitst in N elkaar niet overlappende clusters of primaire eenheden. In de eerste trap wordt een steekproef van n primaire eenheden getrokken. Uit iedere primaire eenheid in de steekproef wordt vervolgens in de tweede trap een steekproef van m_d *secundaire eenheden* getrokken, waarbij in het algemeen geldt, dat $m_d \neq M_d$. Steekproefomvang m_d, $d = 1, \cdots, n$ is verondersteld van te voren te zijn vastgelegd. De definities van de vijf steekproefparameters zijn hieronder weergegeven:

$$\begin{aligned} y_d &= \sum_{k=1}^{m_d} y_{dk} \\ \bar{y}_d &= \frac{1}{m_d} \sum_{k=1}^{m_d} y_{dk} \\ s_{yd}^2 &= \frac{1}{m_d - 1} \sum_{k=1}^{m_d} \left(y_{dk} - \bar{y}_d \right)^2 \\ \bar{y}_{CT} &= \frac{1}{n} \sum_{d=1}^{n} y_d \\ s_{CT}^2 &= \frac{1}{n-1} \sum_{d=1}^{n} \left(y_d - \bar{y}_{CT} \right)^2 \quad . \end{aligned} \tag{4.20}$$

Schatters in de tweetrapssteekproef

Bij een tweetrapssteekproef nemen we niet alle secundaire eenheden in de getrokken primaire eenheden waar. Als gevolg daarvan moeten de clusterparameters van de (getrokken) primaire eenheden worden geschat. Het clustergemiddelde en het clustertotaal in primaire eenheid d kunnen we als volgt schatten:

$$\begin{aligned} \hat{\bar{Y}}_d &= \bar{y}_d = \left(\frac{1}{m_d} \right) \sum_{k=1}^{m_d} y_{dk} \\ \hat{Y}_d &= y_d = M_d \bar{y}_d = \left(\frac{M_d}{m_d} \right) \sum_{k=1}^{m_d} y_{dk} \quad . \end{aligned} \tag{4.21}$$

Met andere woorden, de schatting voor het clustertotaal komt overeen met een gewogen som van de waargenomen secundaire eenheden (in de eerder genoemde pri-

maire eenheid). De weegfactor is de reciproque van de steekproeffractie voor primaire eenheid d .

Het gemiddelde clustertotaal $\hat{\bar{Y}}_{CT}$ en de aangepaste variantie van clustertotalen kunnen we achtereenvolgens schatten als:

$$\hat{\bar{Y}}_{CT} = \bar{y}_{CT} = \frac{1}{n}\sum_{d=1}^{n} y_d = \left(\frac{1}{n}\right)\sum_{d=1}^{n}\sum_{k=1}^{m_d} y_{dk}$$

$$\hat{S}^2_{CT} = s^2_{CT} = \frac{1}{n-1}\sum_{d=1}^{n}\left(y_d - \bar{y}_{CT}\right)^2 \quad .$$

Ook bij de tweetrapssteekproef kunnen we spreken van clusteringsschatters. De definities van de clusteringsschatters voor het populatietotaal en het populatiegemiddelde geven we hieronder weer.

$$\hat{Y}_{CS} = N\hat{\bar{Y}}_{CT} = \left(\frac{1}{f}\right)\sum_{d=1}^{n}\sum_{k=1}^{m_d} y_{dk}$$

$$\hat{\bar{Y}}_{CS} = \frac{1}{\bar{M}}\hat{\bar{Y}}_{CT} = \left(\frac{1}{f}\right)\left(\frac{1}{M}\right)\sum_{d=1}^{n}\sum_{k=1}^{m_d} y_{dk} \quad .$$

Zuiverheidsanalyses

De omstandigheid, dat een geselecteerde cluster (primaire eenheid) niet meer volledig, maar door tussenkomst van een EAZT-steekproef van m_d uit M_d wordt waargenomen, is niet van invloed op de zuiverheidsanalyse m.b.t. tot de schatters. We kunnen zonder meer stellen, dat de schatters tot nu toe gedefinieerd in deze deelparagraaf, allemaal zuiver zijn.

Variantieberekeningen

Bij conventionele clustersteekproeven worden de varianties van de schatters $\hat{Y}_{CS}$ en $\hat{\bar{Y}}_{CS}$ alleen bepaald door de verschillen tussen de clusters. In een tweetrapssteekproef zijn de clustertotalen $\hat{Y}_d$ echter stochastische variabelen. Als gevolg daarvan draagt ook de variatie binnen de clusters bij aan de variantie van de schatters. De variantie van de clusteringsschatter voor het populatietotaal $\hat{Y}_{CS}$ bevat daarom een extra term (zie Muilwijk e.a, 1992, paragraaf 7.5.2):

$$\text{var}\left(\hat{Y}_{CS}\right) = N\left(\frac{1-f_1}{f_1}\right)S^2_{CT} + \sum_{d=1}^{N} M_d\left(\frac{1-f_{2,d}}{f_1 f_{2,d}}\right)S^2_{yd} \quad . \tag{4.22}$$

met:

$$f_1 = \frac{n}{N} \qquad \wedge \qquad f_{2,d} = \frac{m_d}{M_d} \quad .$$

De eerstgenoemde fractie wordt de *steekproeffractie in de eerste trap* genoemd; de tweedegenoemde fractie staat bekend als de *steekproeffractie in de tweede trap*.

De variantie van de clusteringsschatter voor het populatiegemiddelde wordt verkregen door de variantie van de clusteringsschatter voor het populatietotaal door M^2 te delen:

$$\begin{aligned} \mathrm{var}\left(\hat{\bar{Y}}_{CS}\right) &= \frac{1}{M^2}\,\mathrm{var}\left(\hat{Y}_{CS}\right) \\ &= \left(\frac{1}{M}\right)\left(\frac{1}{\overline{M}}\right)\left(\frac{1-f_1}{f_1}\right)S_{CT}^2 + \left(\frac{1}{M}\right)\sum_{d=1}^{N}\left(\frac{M_d}{M}\right)\left(\frac{1-f_{2,d}}{f_1 f_{2,d}}\right)S_{yd}^2 \quad . \end{aligned} \tag{4.23}$$

Het effect van het trekken van steekproeven in de secundaire eenheden op de variantie van bovenstaande clusteringsschatter wordt duidelijk na vergelijking van (4.22) met (4.9). De varianties van de schatters zijn opgebouwd uit twee termen. De eerste term (aan de rechterkant van het '='-teken) wordt veroorzaakt door de verschillen tussen de totalen van de primaire eenheden. De tweede term (aan de rechterkant van het '='-teken) wordt veroorzaakt door de spreiding binnen de primaire eenheden. In de vorige paragrafen speelde deze geen rol omdat de clusters integraal werden waargenomen ($f_{2,d} \equiv 1$).

In veel situaties zal voornoemde tweede term verwaarloosbaar zijn ten opzichte van de eerste term. Bij een sterk variërende M_d heeft de EAZT-schatter in de eerste trap het bezwaar, dat de verschillen tussen de clustertotalen meestal erg groot zijn. We kunnen dit effect tegengaan door de primaire eenheden evenredig met hun omvang te trekken of door de quotiëntschatter toe te passen.

De berekende varianties (4.22) en (4.23) kunnen we zuiver schatten door in de formules S_{CT}^2 en S_{yd}^2 te vervangen door de corresponderende steekproefparameters s_{CT}^2 en s_{yd}^2. In dit geval is s_{CT}^2 de variantie van de geschatte totalen van de primaire eenheden in de steekproef

$$s_{CT}^2 = \frac{1}{n-1}\sum_{d=1}^{n}\left(\hat{Y}_d - \hat{\bar{Y}}_{CT}\right)^2 \quad .$$

Evaluatie van het steekproefontwerp

In de tweetrapssteekproef worden gemiddeld $(n \times \overline{m})$ secundaire eenheden geïnterviewd; de gemiddelde steekproefomvang voor de N secundaire eenheden is gelijk aan $\overline{m} = \sum_{d=1}^{N}(m_d / N)$. Ter evaluatie van deze clustersteekproef vergelijken we de variantie van de clusteringsschatters met de variantie van de desbetreffende schatters

gebaseerd op een EAZT-steekproef van $(n \times \overline{m})$ uit M elementen. Daartoe berekenen we het ontwerpeffect van de clusteringsschatter voor het populatietotaal, zie definitie (4.11).

De variantie van de EAZT-schatter voor het populatietotaal op basis van een EAZT-steekproef van $(n \times \overline{m})$ elementen, kan als volgt worden uitgerekend:

$$\text{var}\left(\hat{Y}_{EAZT}\right) = M\left(\frac{1-f_1 f_2}{f_1 f_2}\right)S_y^2 \quad . \tag{4.24}$$

met:

$$f_1 = \frac{n}{N} \qquad \wedge \qquad f_2 = \frac{\overline{m}}{\overline{M}} \quad .$$

Op basis van uitdrukkingen (4.22) en (4.24) kunnen we het ontwerpeffect van de clusteringsschatter voor het populatietotaal als volgt uitrekenen:

$$\begin{aligned} DEFF\left[\hat{Y}_{CS}\right] = & \left(\frac{1}{\overline{M}}\right)\left(\frac{f_2 - f_1 f_2}{1-f_1 f_2}\right)\left(\frac{S_{CT}^2}{S_y^2}\right) + \\ & + \left(\frac{f_2}{1-f_1 f_2}\right)\sum_{d=1}^{N}\left(\frac{M_d}{M}\right)\left(\frac{1-f_{2,d}}{f_{2,d}}\right)\left(\frac{S_{yd}^2}{S_y^2}\right) \quad . \end{aligned} \tag{4.25*}$$

Dit resultaat kunnen we interpreteren als een breder inzetbare versie van uitdrukking (4.12). Dit blijkt onder andere uit het feit, dat onder de conditie van volledige waarneming van de geselecteerde primaire eenheden (i.e. $f_{2,d} = f_2 = 1$), beide formules met elkaar overeenkomen.

Met het doel formule (4.25) verder uit te werken nemen we aan, dat de primaire eenheden even groot zijn: i.e. $M_d = \overline{M}$ voor alle d. Het ligt dan voor de hand uit elke primaire eenheid hetzelfde aantal secundaire eenheden te trekken, i.e. $m_d = \overline{m}$ voor $d = 1, \cdots, N$. Onder deze omstandigheden leiden we eenvoudig af, dat geldt:

$$M_d = \overline{M} = \frac{M}{N} \qquad \wedge \qquad f_{2,d} = \frac{m_d}{M_d} = \frac{\overline{m}}{\overline{M}} = f_2 \quad . \tag{4.26}$$

De eerder afgeleide formule voor het ontwerpeffect van de clusteringsschatter voor het populatietotaal komt onder de huidige omstandigheden overeen met:

$$DEFF\left[\hat{Y}_{CS}\right] = \left(\frac{1}{\overline{M}}\right)\left(\frac{f_2\left(1-f_1\right)}{1-f_1 f_2}\right)\left(\frac{S_{CT}^2}{S_y^2}\right) + \left(\frac{1-f_2}{1-f_1 f_2}\right)\left(\frac{S_{binnen}^2}{S_y^2}\right) \quad . \tag{4.27*}$$

Formule (4.13) voor de aangepaste populatievariantie is nog steeds geldig, aangezien de eigenschappen van de populatie niet variëren met het steekproefontwerp. We kunnen dus zonder meer schrijven (zie ook (4.13)):

$$\left(\frac{S_{CT}^2}{S_y^2}\right) = -M\left(\frac{\overline{M}-1}{N-1}\right)\left(\frac{S_{binnen}^2}{S_y^2}\right) + \frac{\overline{M}(M-1)}{N-1} \quad . \tag{4.28}$$

Deze relatie levert ons het gewenste verband tussen de aangepaste variantie van clustertotalen en de aangepaste populatievariantie. Voor de tweetrapssteekproef kunnen we het ontwerpeffect van de clusteringsschatter voor het populatietotaal na substitutie van (4.28) in (4.27) als volgt bepalen:

$$\begin{aligned} DEFF\left[\hat{Y}_{CS}\right] = &\left(\frac{1}{1-f_1 f_2}\right)\left\{1-\left(\frac{M-1}{N-1}\right)f_2 + \left(\frac{M-N}{N-1}\right)f_1 f_2\right\}\left(\frac{S_{binnen}^2}{S_y^2}\right) + \\ &+\left(\frac{f_2 - f_1 f_2}{1-f_1 f_2}\right)\left(\frac{M-1}{N-1}\right) \quad . \end{aligned} \tag{4.29}$$

Zoals verwacht komt deze uitdrukking onder de voorwaarde van volledige waarneming in de geselecteerde primaire eenheden, i.e. $f_2 = 1$, overeen met formule (4.14).

Bepaling steekproefgrootte

Net als bij een enkelvoudige, aselecte steekproef kan men de totale steekproefomvang bepalen aan de hand van de vereiste precisie voor een schatter. Heeft men de totale steekproefgrootte eenmaal bepaald dan kan men zich afvragen hoe deze over de primaire en secundaire eenheden te verdelen. Dat wil zeggen hoeveel secundaire eenheden trekken we in de geselecteerde primaire eenheden en hoeveel primaire eenheden trekken we in de steekproef. Deze vragen vereisen veelal een afweging van de kosten en de precisie. Trekt men veel primaire eenheden met weinig secundaire eenheden per primaire eenheid dan zal dat over het algemeen tot kleine varianties leiden. De kostenbesparing die ontstaat door een groot aantal elementen binnen dezelfde primaire eenheid waar te nemen gaat dan echter grotendeels verloren. Veel secundaire eenheden binnen een beperkt aantal primaire eenheden waarnemen is weliswaar goedkoper maar zal tot minder precieze schatters leiden. Om de vraag met betrekking tot de verdeling van de steekproef te kunnen beantwoorden moeten we weten wat de (waarnemings)kosten zijn in beide trappen en we moeten een idee hebben van de homogeniteit van de primaire eenheden.

Zijn de primaire eenheden perfect homogeen, dat wil zeggen $S_{binnen}^2 = 0$, dan kan je net zo goed één element per primaire eenheid trekken; meer elementen waarnemen leidt alleen tot hogere kosten zonder dat de precisie toeneemt. In alle andere gevallen hangt de verdeling van de steekproef af van de kosten voor het trekken van de eenheden in de twee trappen. Een simpele kostenfunctie is:

$$C = nc_1 + n\overline{m}c_2 \quad .$$

Bij de parameters c_1 en c_2 kan men denken aan de enquêteringskosten en / of reiskosten verbonden aan de eerste trap respectievelijk de tweede trap van de tweetrapssteekproef. Bij deze kostenfunctie en bij vaste totale kosten C, wordt de variantie van de clusteringsschatter voor het populatietotaal geminimaliseerd door in de steekproef:

$$n = \frac{C}{c_1 + c_2 m}$$

primaire eenheden te selecteren en:

$$m = \sqrt{\frac{\overline{M} S^2_{binnen}}{\left(\overline{M} S^2_{tussen} - S^2_{binnen}\right)} \frac{c_1}{c_2}} = \sqrt{\frac{c_1 / c_2}{S^2_{tussen} / S^2_{binnen} - \left(1/\overline{M}\right)}}$$

secundaire eenheden per primaire eenheid. Om deze aantallen expliciet te kunnen bepalen moet men wel een idee hebben van de verhouding $S^2_{tussen} / S^2_{binnen}$, dat wil zeggen van de homogeniteit van de primaire eenheden. In de praktijk zal men vaak andersom te werk gaan. Dit komt er op neer dat men verschillende waarden van n en m probeert, de bijbehorende variantie van een schatter berekent en kijkt of deze enigszins acceptabel is

4.3.6 De systematische steekproef

De systematische steekproef wordt vaak voorgesteld als een goed alternatief voor een EAZT-steekproef. Om een systematische steekproef van n elementen uit een populatie van N elementen te trekken bepalen we eerst een, zogenaamde, *stapleng-te* $L = N/n$. We gaan er voor het gemak van uit, dat de elementen genummerd zijn van 1 t/m N. Vervolgens kiezen we uit het interval (0, L) een aselect startgetal R. Het eerste element dat we nu in de steekproef trekken is het element met rangnummer k_1 dat zodanig is dat $k_1 - 1 < R \le k_1$. Het tweede element dat we zo trekken in de steekproef heeft rangnummer k_2 dat zodanig is dat $k_2 - 1 < R + L \le k_2$, enz. Het laatste n^e element in de steekproef heeft rangnummer k_n met de eigenschap $k_n - 1 < R + (n-1)L \le k_n$. Met andere woorden, de elementen met rangnummers $k_1, \cdots, k_n$ vormen de steekproef ofwel de rangnummers in de steekproef zijn gelijk aan de naar boven afgeronde getallen $R, R + L, R + 2L, \cdots, R + (n-1)L$. Het is duidelijk dat het willekeurig gekozen startgetal R de gehele steekproef vastlegt, en dat ieder element uit de populatie maar in één enkele steekproef voorkomt.

Theoretisch gesproken en aannemende dat L geheeltallig is, zijn er bij een vaste populatievolgorde L verschillende steekproeven mogelijk, zodat de trekkingskans van een systematische steekproef gelijk is aan:

$$P(s) = \frac{1}{L} = \frac{n}{N} \quad .$$

Aangezien elk element slechts in één steekproef voor kan komen, is de insluitkans van populatie-element k gelijk aan de trekkingskans van de steekproef, i.e.:

$$\pi_k = \frac{1}{L} = \frac{n}{N} \quad .$$

Bij een systematische steekproef van deze vorm is het aantal mogelijke steekproeven kleiner dan bij een EAZT-steekproef van dezelfde omvang. Niet elke mogelijke combinatie van n elementen is namelijk mogelijk, wat wel het geval is bij een EAZT-steekproef. De eerste-orde insluitkansen zijn, zie het resultaat hierboven, echter wel hetzelfde als bij de EAZT-steekproef. Ofschoon de hier beschreven systematische steekproef in feite een clustersteekproef is, wordt in de praktijk de variantie van deze schatter vaak benaderd door die van de EAZT-schatter. De veronderstelling hierbij is (dat de vaste) volgorde van de populatie-elementen niet leidt tot een systematisch verschil met de EAZT-schatter.

Het systematisch trekken van een steekproef komt dus neer op het trekken van één cluster uit *F* clusters van omvang *n*. De populatieparameters worden geschat aan de hand van de steekproefgrootheden van deze éne getrokken cluster. De clusteringsschatters voor het populatietotaal en populatiegemiddelde zijn respectievelijk:

$$\hat{Y}_{sys} = N\,\bar{y} \qquad \wedge \qquad \hat{\bar{Y}}_{sys} = \bar{y} \quad .$$

De zuiverheid van de schatters volgt uit de eigenschappen van een clustersteekproef. Volgens formule (4.10) is de variantie voor de clusteringsschatter van het populatiegemiddelde bij een trekking van 1 cluster gelijk aan:

$$\mathrm{var}\left(\hat{\bar{Y}}_{sys}\right) = \left(1 - \frac{1}{F}\right)\frac{S_{CT}^2}{n^2} \tag{4.30}$$

De variantie is geschreven in de notatie voor een systematische steekproef. Merk op dat N en n hier betrekking hebben op elementen en niet op clusters. De omvang van een cluster (systematische steekproef) is nu n; het totale aantal elementen bedraagt N; het aantal clusters is F. De variantie van de clusteringsschatter voor het populatietotaal verkrijgen we door de variantie van de clusteringsschatter voor het populatiegemiddelde te vermenigvuldigen met N^2.

Ten slotte, het ontwerpeffect van de clusteringsschatter voor het populatietotaal kunnen we op basis van formule (4.19) uitrekenen als:

$$DEFF\left[\hat{\bar{Y}}_{sys}\right] = \left(\frac{(N-1)(n-1)}{N-n}\right)\rho_c + \frac{N-1}{N-n} \quad .$$

Systematische steekproeven zijn dus beter dan enkelvoudige, aselecte steekproeven d.e.s.d.a. de intraklassecorrelatiecoëfficiënt negatief is, dat wil zeggen als voor de intraclustercorrelatiecoëfficiënt geldt:

$$-\frac{1}{n-1} \leq \rho_c \leq -\frac{1}{N-1} \quad .$$

Wanneer de variantie binnen de mogelijke systematische steekproeven groter is dan de populatievariantie zullen de clustergemiddelden elkaar niet al te veel ontlopen en is een systematische steekproef preciezer dan een enkelvoudig aselecte steekproef. Is de variantie binnen de systematische steekproeven echter klein ten opzichte van de populatievariantie $(0 < \rho_c)$ dan geven alle elementen in de steekproef dezelfde soort informatie en zal de variantie van de schatter voor een systematische steekproef groter zijn dan voor een enkelvoudige aselecte steekproef.

Een belangrijk nadeel van het systematische steekproefontwerp is echter het ontbreken van zuivere schatters voor de verschillende berekende varianties. Aangezien we maar één cluster trekken is de steekproefvariantie van de clustertotalen s_{CT}^2 niet gedefinieerd. We kunnen de variantie echter wel benaderen wanneer we iets weten over de structuur van de populatie. We onderscheiden de volgende 3 situaties:

1. willekeurige of aselecte volgorden van het steekproefkader;
2. positieve autocorrelatie;
3. periodieke fluctuaties.

Ad 1 (Aselecte volgorden van het steekproefkader)
In veel situaties is het aannemelijk om te veronderstellen dat de waarden van de doelvariabelen niet samenhangen met de volgorde in het steekproefkader. Denk bijvoorbeeld aan een kader waarin de personen op alfabetische volgorde gerangschikt staan. In dit geval zal een systematische steekproef grote overeenkomsten vertonen met een enkelvoudig aselecte steekproef ($\rho_c \approx 0$) en beide steekproeven zullen dezelfde soort resultaten leveren. We kunnen dan de variantieformule voor een enkelvoudig aselecte steekproef gebruiken om $\text{var}(\hat{\bar{Y}}_{sys})$ te schatten.

Ad 2 (Positieve autocorrelatie)
Soms komt het voor dat er een oplopend of aflopend volgorde-effect in het kader aanwezig is. Bedrijven kunnen bijvoorbeeld in een aflopende volgorde van grootte op een lijst voorkomen. We spreken in dit geval van een positieve autocorrelatie: opeenvolgende elementen lijken meer op elkaar dan elementen die verder van elkaar afliggen. Een systematische steekproef zorgt voor meer spreiding in de steekproefelementen en zal dan ook preciezer zijn dan een enkelvoudig aselecte steekproef ($\rho_c < 0$). De variantieformule voor een enkelvoudige, aselecte steekproef levert in dit geval een overschatting.

Ad 3 (Periodieke fluctuaties)

Problemen kunnen zich voordoen wanneer het steekproefkader een periodieke variatie vertoont ten aanzien van een doelvariabele. Systematische trekking zal dan aanzienlijk minder precies zijn dan een enkelvoudig aselecte trekking, met name wanneer de staplengte overeenkomt met (een veelvoud van) de periode. De variantie van de schatter zal dan erg groot zijn wat niet tot uitdrukking komt wanneer we de enkelvoudig aselecte variantieformule gebruiken om de variantie te schatten.

Stel de populatie-elementen zijn zo gerangschikt dat de doelvariabele het volgende patroon vertoont: 1,2,3,4,5,1,2,3,4,5,1,2,3,4,5,⋯⋯. Bij een staplengte van 5 krijgen we dan allemaal dezelfde elementen in de steekproef en de variantieformule voor een enkelvoudige steekproef schat de variantie als $\hat{\text{var}}(\hat{\bar{Y}}) = 0$. De werkelijke waarde van $\text{var}(\hat{\bar{Y}}_{sys})$ voor deze populatie is echter gelijk aan 2.

4.4 Kwaliteitsindicatoren

Een belangrijke kwaliteitsindicator voor (meertraps-) clustersteekproeven is de kostenreductie die het gevolg is van het feit dat de waar te nemen eenheden nu in clusters bij elkaar liggen. De prijs die hiervoor moet worden betaald in de vorm van een toename van de variantie ten opzichte van de enkelvoudige aselecte steekproef zonder teruglegging, is een ander belangrijk criterium.

4.5 Appendix

Bewijs van (4.4)

Uit de definitie van de populatievariantie leiden we af:

$$\begin{aligned}\sigma_y^2 &= \frac{1}{M}\sum_{d=1}^{N}\sum_{k=1}^{M_d}\left(Y_{dk}-\bar{\bar{Y}}\right)^2 \\ &= \frac{1}{M}\sum_{d=1}^{N}\sum_{k=1}^{M_d}\left(Y_{dk}-\bar{Y}_d\right)^2 + \frac{1}{M}\sum_{d=1}^{N}\sum_{k=1}^{M_d}\left(\bar{Y}_d-\bar{\bar{Y}}\right)^2\end{aligned}$$

omdat het dubbele product gelijk is aan nul, zie ook het bewijs van formule (3.4). Nadere uitwerking van deze uitdrukking geeft het volgende resultaat:

$$\begin{aligned}\sigma_y^2 &= \sum_{d=1}^{N}\left(\frac{M_d}{M}\right)\sigma_{yd}^2 + \frac{1}{M}\sum_{d=1}^{N}\sum_{k=1}^{M_d}\left(\bar{Y}_d-\bar{\bar{Y}}\right)^2 \\ &= \sum_{d=1}^{N}\left(\frac{M_d}{M}\right)\sigma_{yd}^2 + \sum_{d=1}^{N}\left(\frac{M_d}{M}\right)\left(\bar{Y}_d-\bar{\bar{Y}}\right)^2 \quad .\end{aligned}$$

Het bewijs van (4.4) is hiermee gegeven. ■

Bewijs van (4.5)

Uit definities (2.1) en (4.1) leiden we af, dat:

$$S_y^2 = \left(\frac{M}{M-1}\right)\sigma_y^2 \qquad \wedge \qquad S_{yd}^2 = \left(\frac{M_d}{M_d-1}\right)\sigma_{yd}^2 \quad .$$

Na vermenigvuldiging van gelijkheid (4.4) met de factor $(M/(M-1))$ wordt de volgende relatie verkregen:

$$\left(\frac{M}{M-1}\right)\sigma_y^2 = \sum_{d=1}^{N}\left(\frac{M_d}{M}\right)\left(\frac{M}{M-1}\right)\sigma_{yd}^2 + \sum_{d=1}^{N}\left(\frac{M}{M-1}\right)\left(\frac{M_d}{M}\right)\left(\overline{Y}_d - \overline{Y}\right)^2 \quad .$$

Deze gelijkheid kunnen we herschrijven als:

$$\begin{aligned} S_y^2 &= \sum_{d=1}^{N}\left(\frac{M_d}{M-1}\right)\sigma_{yd}^2 + \sum_{d=1}^{N}\left(\frac{M_d}{M-1}\right)\left(\overline{Y}_d - \overline{Y}\right)^2 \\ &= \sum_{d=1}^{N}\left(\frac{M_d}{M-1}\right)\left(\frac{M_d-1}{M_d-1}\right)\sigma_{yd}^2 + \sum_{d=1}^{N}\left(\frac{M_d}{M-1}\right)\left(\overline{Y}_d - \overline{Y}\right)^2 \\ &= \sum_{d=1}^{N}\left(\frac{M_d-1}{M-1}\right)S_{yd}^2 + \sum_{d=1}^{N}\left(\frac{M_d}{M-1}\right)\left(\overline{Y}_d - \overline{Y}\right)^2 \quad . \end{aligned}$$

Met dit resultaat is (4.5) bewezen. ■

Bewijs van (4.13)

Uitgangspunt voor het bewijs van (4.13) is formule (4.5) voor de aangepaste populatievariantie. Ook spelen de volgende relaties een rol van betekenis:

$$\overline{Y} = \frac{1}{\overline{M}}\overline{Y}_{CT} \qquad \wedge \qquad \overline{Y}_d = \frac{1}{\overline{M}}Y_d \quad .$$

De uitdrukking voor de aangepaste populatievariantie kunnen we nu herschrijven als:

$$\begin{aligned} S_y^2 &= \sum_{d=1}^{N}\left(\frac{M_d-1}{M-1}\right)S_{yd}^2 + \sum_{d=1}^{N}\left(\frac{M_d}{M-1}\right)\left(\overline{Y}_d - \overline{Y}\right)^2 \\ &= \left(\frac{\overline{M}-1}{M-1}\right)\sum_{d=1}^{N}S_{yd}^2 + \left(\frac{\overline{M}}{M-1}\right)\sum_{d=1}^{N}\left(\frac{1}{\overline{M}}\right)^2\left(Y_d - \overline{Y}_{CT}\right)^2 \\ &= \left(\frac{\overline{M}-1}{M-1}\right)\left(\frac{N}{N}\right)\sum_{d=1}^{N}S_{yd}^2 + \left(\frac{1}{\overline{M}(M-1)}\right)\left(\frac{N-1}{N-1}\right)\sum_{d=1}^{N}\left(Y_d - \overline{Y}_{CT}\right)^2 \\ &= \left(\frac{M-N}{M-1}\right)S_{binnen}^2 + \left(\frac{1}{\overline{M}}\right)\left(\frac{N-1}{M-1}\right)S_{CT}^2 \quad . \end{aligned}$$

Relatie (4.13) is dus bewezen. ■

Bewijs van (4.14)

Formule (4.13) voor de populatievariantie kan als volgt worden herschreven:

$$\left(\frac{1}{\overline{\overline{M}}}\right)\left(\frac{N-1}{M-1}\right)S_{CT}^2 = S_y^2 - \left(\frac{M-N}{M-1}\right)S_{binnen}^2$$

waaruit volgt:

$$\left(\frac{1}{\overline{\overline{M}}}\right)S_{CT}^2 = \left(\frac{M-1}{N-1}\right)S_y^2 - \left(\frac{M-N}{N-1}\right)S_{binnen}^2 \quad .$$

Substitutie van deze relatie in formule (4.12) geeft:

$$DEFF\left[\hat{Y}_{CS}\right] = \frac{\frac{1}{\overline{\overline{M}}}S_{CT}^2}{S_y^2}$$
$$= \left(\frac{M-1}{N-1}\right) - \left(\frac{M-N}{N-1}\right)\frac{S_{binnen}^2}{S_y^2} \quad .$$

Hiermee is het bewijs van (4.14) compleet. ■

Bewijs van (4.15)

Het ontwerpeffect van de clusteringsschatter voor het populatietotaal is minimaal als de binnenclustervariantie maximaal is. Volgens (4.5) is de binnenclustervariantie maximaal d.e.s.d.a. de tussenclustervariantie gelijk is aan nul. De waarde van S_{binnen}^2 is dan als volgt te bepalen:

$$S_{binnen}^2\Big|_{S_{tussen}^2=0} = \left(\frac{M-1}{M-N}\right)S_y^2 \quad .$$

Het ontwerpeffect is voor deze waarde van de binnenclustervariantie gelijk aan:

$$DEFF\left[\hat{Y}_{CS}\right]\Big|_{S_{tussen}^2=0} = -\left(\frac{M-N}{N-1}\right)\left(\frac{M-1}{M-N}\right) + \left(\frac{M-1}{N-1}\right) = 0 \quad .$$

Het ontwerpeffect van de clusteringsschatter voor het populatietotaal is maximaal als de binnenclustervariantie minimaal is. De minimale waarde die de binnenclustervariantie aan kan nemen is 0. De maximale waarde die het ontwerpeffect aan kan nemen is dan eenvoudig bepaald als:

$$DEFF\left[\hat{Y}_{CS}\right]\Big|_{S_{binnen}^2=0} = \frac{M-1}{N-1} \quad .$$

Het bewijs van (4.15) is hiermee compleet. ■

Bewijs van (4.17)

Algebraïsche manipulaties tonen aan, dat:

$$\sum_{k=1}^{\overline{M}}\sum_{l=1}^{\overline{M}}\left(Y_{dk}-\overline{Y}\right)\left(Y_{dl}-\overline{Y}\right)=\sum_{k=1}^{\overline{M}}\left(Y_{dk}-\overline{Y}\right)\sum_{l=1}^{\overline{M}}\left(Y_{dl}-\overline{Y}\right)$$
$$=\left(Y_{d}-\overline{Y}_{CT}\right)^{2} \quad .$$

Uit dit resultaat en definitie (4.2) leiden we vervolgens af, dat:

$$\sum_{d=1}^{N}\sum_{k=1}^{\overline{M}}\sum_{l=1}^{\overline{M}}\left(Y_{dk}-\overline{Y}\right)\left(Y_{dl}-\overline{Y}\right)=\sum_{d=1}^{N}\left(Y_{d}-\overline{Y}_{CT}\right)^{2}$$
$$=\left(N-1\right)S_{CT}^{2} \quad .$$

Voor het uitrekenen van de hierboven genoemde 3-voudige som, bestaat ook een alternatieve aanpak, namelijk:

$$\sum_{d=1}^{N}\sum_{k=1}^{\overline{M}}\sum_{l=1}^{\overline{M}}\left(Y_{dk}-\overline{Y}\right)\left(Y_{dl}-\overline{Y}\right)=\sum_{d=1}^{N}\sum_{k=1}^{\overline{M}}\sum_{l\neq k}^{\overline{M}}\left(Y_{dk}-\overline{Y}\right)\left(Y_{dl}-\overline{Y}\right)$$
$$+\sum_{d=1}^{N}\sum_{k=1}^{\overline{M}}\sum_{l=k}^{\overline{M}}\left(Y_{dk}-\overline{Y}\right)\left(Y_{dl}-\overline{Y}\right)$$
$$=\sum_{d=1}^{N}\sum_{k=1}^{\overline{M}}\sum_{l\neq k}^{\overline{M}}\left(Y_{dk}-\overline{Y}\right)\left(Y_{dl}-\overline{Y}\right)$$
$$+\sum_{d=1}^{N}\sum_{k=1}^{\overline{M}}\left(Y_{dk}-\overline{Y}\right)^{2}$$
$$=\sum_{d=1}^{N}\sum_{k=1}^{\overline{M}}\sum_{l\neq k}^{\overline{M}}\left(Y_{dk}-\overline{Y}\right)\left(Y_{dl}-\overline{Y}\right)+\left(M-1\right)S_{y}^{2} \quad .$$

Combinatie van deze twee resultaten geeft de volgende uitdrukking:

$$\sum_{d=1}^{N}\sum_{k=1}^{\overline{M}}\sum_{l\neq k}^{\overline{M}}\left(Y_{dk}-\overline{Y}\right)\left(Y_{dl}-\overline{Y}\right)=\left(N-1\right)S_{CT}^{2}-\left(M-1\right)S_{y}^{2} \quad .$$

Op basis van definitie (4.16) voor de intraclustercorrelatiecoëfficiënt en dit resultaat, kunnen we concluderen, dat:

$$\rho_{c}=\frac{\left(N-1\right)S_{CT}^{2}-\left(M-1\right)S_{y}^{2}}{\left(M-1\right)\left(\overline{M}-1\right)S_{y}^{2}}$$
$$=\left(\frac{N-1}{\left(M-1\right)\left(\overline{M}-1\right)}\right)\frac{S_{CT}^{2}}{S_{y}^{2}}-\left(\frac{1}{\overline{M}-1}\right) \quad .$$

Hiermee is het bewijs van formule (4.17) afgerond. ■

Bewijs van (4.18)

De tussenclustervariantie is maximaal d.e.s.d.a de binnenclustervariantie gelijk is aan 0. Uit relatie (4.13) volgt, dat onder deze omstandigheid de aangepaste variantie van clustertotalen gelijk is aan:

$$S_{CT}^{2}\Big|_{S_{binnen}^{2}=0}=\left(\frac{\overline{M}\left(M-1\right)}{N-1}\right)S_{y}^{2} \quad .$$

Invulling van dit resultaat in (4.17) resulteert in:

$$\rho_c\big|_{S^2_{binnen}=0} = \left(\frac{N-1}{(M-1)(\overline{M}-1)}\right)\left(\frac{\overline{M}(M-1)}{N-1}\right) - \frac{1}{\overline{M}-1}$$
$$= \frac{\overline{M}}{\overline{M}-1} - \frac{1}{\overline{M}-1} = 1 \quad .$$

De minimale waarde van de tussenclustervariantie is 0. Dit betekent, dat de minimale waarde van de aangepaste variantie van clustertotalen eveneens gelijk is 0. Hieruit volgt:

$$\rho_c\big|_{S^2_{tussen}=0} = -\frac{1}{\overline{M}-1} \quad .$$

Het bewijs van (4.18) is compleet. ■

Bewijs van (4.19)

Uit formule (4.17) leiden we af, dat:

$$\frac{S^2_{CT}}{S^2_y} = \left(\frac{(M-1)(\overline{M}-1)}{N-1}\right)\rho_c + \frac{M-1}{N-1} \quad .$$

Substitutie van dit resultaat in uitdrukking (4.12) voor het ontwerpeffect van de clusteringsschatter voor het populatietotaal geeft:

$$DEFF\left[\hat{Y}_{CS}\right] = \frac{1}{\overline{M}}\left(\frac{(M-1)(\overline{M}-1)}{N-1}\right)\rho_c + \frac{1}{\overline{M}} \times \frac{M-1}{N-1}$$
$$= \left(\frac{(M-1)(\overline{M}-1)}{M-\overline{M}}\right)\rho_c + \frac{M-1}{M-\overline{M}} \quad .$$

Hiermee is het bewijs van formule (4.19) voltooid. ■

Bewijs van (4.25)

Het ontwerpeffect van de clusteringsschatter voor het populatietotaal is gedefinieerd als het quotiënt van varianties (4.22) en (4.24). We leiden daarom eenvoudig af, dat:

$$DEFF\left[\hat{Y}_{CS}\right] = N\left(\frac{1-f_1}{f_1}\right)\left(\frac{1}{M}\right)\left(\frac{f_1 f_2}{1-f_1 f_2}\right)\left(\frac{S_{CT}^2}{S_y^2}\right)$$
$$+\left(\frac{1}{M}\right)\left(\frac{f_1 f_2}{1-f_1 f_2}\right)\sum_{d=1}^{N} M_d\left(\frac{1-f_{2,d}}{f_1 f_{2,d}}\right)\left(\frac{S_{yd}^2}{S_y^2}\right)$$
$$=\left(\frac{1}{\overline{M}}\right)\left(\frac{f_2(1-f_1)}{1-f_1 f_2}\right)\left(\frac{S_{CT}^2}{S_y^2}\right)+$$
$$+\left(\frac{f_2}{1-f_1 f_2}\right)\sum_{d=1}^{N}\left(\frac{M_d}{M}\right)\left(\frac{1-f_{2,d}}{f_{2,d}}\right)\left(\frac{S_{yd}^2}{S_y^2}\right) \quad .$$

Hiermee is het bewijs van (4.25) voltooid. ■

Bewijs van (4.27)

Het ontwerpeffect van de clusteringsschatter voor het populatietotaal is gedefinieerd als het quotiënt van varianties (4.22) en (4.24). Voordat we het quotiënt uitrekenen, gaan we eerst resultaat (4.22) herschrijven aangezien $f_{2,d} = f_2$:

$$\text{var}\left(\hat{Y}_{CS}\right) = N\left(\frac{1-f_1}{f_1}\right)S_{CT}^2 + \sum_{d=1}^{N}\overline{M}\left(\frac{1-f_2}{f_1 f_2}\right)S_{yd}^2$$
$$= N\left(\frac{1-f_1}{f_1}\right)S_{CT}^2 + M\left(\frac{1-f_2}{f_1 f_2}\right)\left(\frac{1}{N}\right)\sum_{d=1}^{N} S_{yd}^2$$
$$= N\left(\frac{1-f_1}{f_1}\right)S_{CT}^2 + M\left(\frac{1-f_2}{f_1 f_2}\right)S_{binnen}^2 \quad .$$

Het delen van dit resultaat door de variantie van de EAZT-schatter op basis van een EAZT-steekproef van gelijke omvang, zie (4.24), geeft de formule:

$$DEFF\left[\hat{Y}_{CS}\right] = N\left(\frac{1-f_1}{f_1}\right)\left(\frac{1}{M}\right)\left(\frac{f_1 f_2}{1-f_1 f_2}\right)\left(\frac{S_{CT}^2}{S_y^2}\right)$$
$$+\left(\frac{1}{M}\right)\left(\frac{f_1 f_2}{1-f_1 f_2}\right)M\left(\frac{1-f_2}{f_1 f_2}\right)\left(\frac{S_{binnen}^2}{S_y^2}\right)$$
$$=\left(\frac{1}{\overline{M}}\right)\left(\frac{f_2(1-f_1)}{1-f_1 f_2}\right)\left(\frac{S_{CT}^2}{S_y^2}\right)+\left(\frac{1-f_2}{1-f_1 f_2}\right)\left(\frac{S_{binnen}^2}{S_y^2}\right) \quad .$$

Hiermee is het bewijs van formule (4.27) voltooid. ■

5. Steekproeven met (on)gelijke insluitkansen

5.1 Korte beschrijving

5.1.1 Het voordeel

Soms is het raadzaam bij het trekken van een steekproef de elementen van een populatie niet allemaal dezelfde insluitkans te geven zoals we dat tot nu toe steeds hebben gedaan. Vooral wanneer de doelvariabele Y_k in de populatie min of meer evenredig is met een bepaalde hulpvariabele X_k $(k = 1, \cdots, N)$, verdient een steekproef met ongelijke insluitkansen de voorkeur. De laatste variabele moet dan wel bekend zijn voor alle elementen in de populatie. In de praktijk staat hulpvariabele X_k vaak voor de omvang of de relevantie van element k in de gehele populatie.

Er zijn vele manieren waarop steekproeven zonder teruglegging met ongelijke insluitkansen kunnen worden getrokken. In dit hoofdstuk zullen we vooral aandacht schenken aan de systematische PPS-steekproef waarbij de elementen in een random volgorde staan en de insluitkansen evenredig zijn met een gegeven hulpvariabele X_k, zie het voorbeeld in paragraaf 5.4. Dit laatste is de reden, dat dit steekproefontwerp ook wel wordt aangeduid met het acroniem PPS (Engels: *probability proportional to size*). Het grote voordeel van ongelijke insluitkansen is dat de varianties van de schatters aanzienlijk kunnen worden gereduceerd en derhalve ook de desbetreffende onzekerheidsmarges. Een ander voordeel is dat geen stratumindeling in grootteklassen meer nodig is waardoor bij een gegeven steekproefomvang het risico dat in sommige strata erg weinig waarnemingen vallen, wordt vermeden.

5.1.2 Het trekkingsschema voor de systematische PPS-steekproef

Nadat de populatie-elementen in een random volgorde zijn gezet wordt aan ieder element k een interval met lengte X_k (op de reële rechte) toegekend $(k = 1, \cdots, N)$. Aan het eerste element wordt, per definitie, het interval $(0, X_1]$ toegewezen, aan het tweede element het interval $(X_1, X_1 + X_2]$, enz. Het interval $(0, X]$, waarbij X staat voor het populatietotaal van de hulpvariabele, wordt op deze wijze verdeeld in N aaneensluitende intervallen. Vervolgens wordt $(0, X]$ verdeeld in *n* gelijke stukken van lengte L $(L = X / n)$; L wordt ook wel de *staplengte* genoemd.

Het trekkingschema verloopt nu als volgt. Kies op aselecte wijze een getal tussen 0 en L. Dit getal ligt in precies één van de intervallen waaruit $(0, X]$ is samengesteld. Het populatie-element dat met dit interval correspondeert, komt als eerste in de steekproef terecht. Het populatie-element waarvan het toegekende interval het getal $(r + L)$ omvat, komt als 2e element in de steekproef, enz.. Het laatste element in de steekproef, is het element waarvan het toegekende interval het getal

$[r+(n-1)L]$ omvat. In formules kunnen we dit trekkingsschema als volgt weergeven:

$$S_k = X_1 + \cdots + X_k \quad (k = 1, \cdots, N)$$
$$S_0 = 0$$
$$J(x) = k \quad \text{voor} \quad S_{k-1} < x \le S_k \qquad (0 < x \le X) \qquad .$$

De systematische PPS-steekproef bestaat dan uit de *n* elementen met de rangnummers

$$J(r), J(r+L), \cdots\cdots, J(r+(n-1)L)$$

Om te voorkomen dat een element meerdere keren in de steekproef wordt getrokken, hebben we steeds stilzwijgend aangenomen, dat elementen waarvoor geldt, dat $L < X_k$, reeds uit de populatie zijn verwijderd en zijn ondergebracht in een apart stratum, dat integraal wordt waargenomen. Dit is het stratum van de, zogenaamde, *zelfselecterende* elementen. Nadat dergelijke elementen zijn verwijderd, moet X opnieuw worden berekend evenals de nieuwe *L* en de nieuwe steekproefomvang *n* zonder de zelfselecterende elementen. Indien nodig wordt dit net zolang herhaald totdat er geen nieuwe zelfselecterende elementen meer bijkomen.

In het hier beschreven trekkingsschema worden de elementen eerst in een random volgorde gezet. Soms kunnen er echter goede redenen zijn om een bepaalde vaste volgorde van de elementen te handhaven, bijvoorbeeld een volgorde waarbij de X_k stijgen. Indien Y_k / X_k en X_k nauw met elkaar samenhangen, kan een dergelijke volgorde leiden tot schatters met een bijzonder kleine variantie. Een probleem hierbij is evenwel het vinden van een goede variantieschatter. Standaard variantieschatters ontbreken, maar onder bepaalde modelveronderstellingen kunnen er redelijk betrouwbare variantieschatters worden afgeleid.

In aanvulling hierop zij nog opgemerkt dat een random volgorde van de elementen overbodig is wanneer er geen noemenswaardig verband bestaat tussen Y_k / X_k en X_k of, nauwkeuriger geformuleerd, wanneer er geen verband bestaat tussen Y_k / X_k en k, $k = 1, \cdots, N$. In een dergelijke situatie zal een random ordening van de elementen geen systematisch effect hebben op de uitkomsten.

Ten slotte schenken we nog aandacht aan het speciale geval, dat alle elementen dezelfde insluitkans hebben. Dat wil zeggen, aan alle elementen is een interval met dezelfde lengte toegekend, terwijl de X_k onderling verschillen. Wanneer de elementen random zijn geordend, wordt op deze manier een EAZT-steekproef verkregen. In dat geval zijn de formules van hoofdstuk 2 van toepassing. Staan de elementen daarentegen in een vaste volgorde, dan wordt het weer veel lastiger de variantie van de schatter voor het populatietotaal te schatten. Wel kan worden aangetoond dat de variantie omlaag gaat door de elementen zodanig te ordenen dat alle mogelijke steekproeven zoveel mogelijk op elkaar lijken (en dus ook op de populatie) voor wat

betreft de verdeling van de X_k. Dit kan in de praktijk min of meer worden gerealiseerd door de elementen te rangschikken naar grootte van de X_k. Voorwaarde voor de reductie van de variantie van de schatter voor het populatietotaal is wel, dat er voldoende samenhang bestaat tussen Y_k en X_k.

Een andere optie in dit verband is de elementen zoveel mogelijk naar regio te ordenen, zodat er in iedere steekproef sprake is van voldoende regionale spreiding (heterogeniteit). Met andere woorden, men dient met de (vaste) volgorde te voorkomen dat de aldus getrokken steekproeven homogeen gaan worden. Evenals bij ongelijke insluitkansen blijft het probleem evenwel het traceren van de variantie; zie Knottnerus (2003, pp. 190-1 en 237-43).

5.2 Toepasbaarheid

Bij het CBS vindt het PPS-steekproefontwerp vooral toepassing bij de sociale statistieken. Dit gebeurt onder meer als onderdeel van een tweetrapssteekproef van personen. In de eerste trap worden dan gemeenten getrokken met een insluitkans evenredig met het aantal inwoners van de gemeente. In de tweede trap worden vervolgens op basis van een EAZT-steekproef personen getrokken uit iedere gemeente die in de eerste trap is geselecteerd. De steekproeffractie van de laatste steekproef is op zijn beurt weer omgekeerd evenredig is met het aantal inwoners van de desbetreffende gemeente zodat per saldo alle personen dezelfde insluitkans hebben.

In navolging van andere landen gaat op korte termijn bij het CBS het PPS-steekproefontwerp ook gebruikt worden voor de raming van de maandelijkse PPI (producentenprijzenindex voor de industrie). In paragraaf 5.4 werken we aan de hand van een concreet voorbeeld met data van 2005 de theorie voor de PPI verder uit.

5.3 Uitgebreide beschrijving

5.3.1 Veronderstellingen en notatie

Het uitgangspunt is een populatie van N elementen waaruit een steekproef van omvang n wordt getrokken *zonder* teruglegging, tenzij anders wordt vermeld. In dit hoofdstuk gaan we er van uit, dat steekproefomvang n vast ligt. Evenals in vorige hoofdstukken, zijn we geïnteresseerd in het schatten van het populatiegemiddelde $\overline{Y}$ en het populatietotaal Y. De n steekproefwaarnemingen van de doelvariabele worden aangeven met $y_1, \cdots, y_n$.

De dichotome insluitindicator a_k, zie (2.7), is een random variabele die aangeeft of element k uit de populatie, uiteindelijk wel of niet in de steekproef is terecht gekomen.

De *kans* van populatie-element k om in de steekproef getrokken te worden, i.e. de *insluitkans*, geven we aan met het symbool π_k. De formele definitie van de insluitkans luidt:

$$\pi_k = P(a_k = 1) \qquad k = 1, \cdots, N \tag{5.1}$$

met:

$$a_k = \begin{cases} 1 & \text{als element } k \textit{ wel} \text{ is getrokken in de steekproef} \\ 0 & \text{als element } k \textit{ niet} \text{ is getrokken in de steekproef.} \end{cases}$$

De tweede-orde insluitkans voor de twee populatie-elementen k en l, notatie π_{kl}, is de kans, dat populatie-elementen k en l tezamen in de steekproef terecht komen. Wat formeler uitgedrukt, is dat:

$$\pi_{kl} = \begin{cases} P(a_k = 1 \wedge a_l = 1) & k \neq l = 1, \cdots, N \\ \pi_k & k = l = 1, \cdots, N \end{cases} . \tag{5.2}$$

5.3.2 Schatters voor het populatietotaal en het populatiegemiddelde

Gegeven de (ongelijke) insluitkansen π_k voor alle N elementen van de populatie, kan het populatietotaal Y worden geschat met de *Horvitz-Thompson schatter* (ook wel HT-schatter genoemd). De HT-schatter, notatie $\hat{Y}_{HT}$, is gedefinieerd als:

$$\hat{Y}_{HT} = \sum_{k=1}^{n} \frac{y_k}{\pi_k} \quad . \tag{5.3}$$

Een noodzakelijke voorwaarde voor het kunnen toepassen van deze schatter is, dat insluitkans π_k groter is dan 0 voor $k = 1, \cdots, N$. Uit definitie (5.3) kunnen we onmiddellijk de definitie van de Horvitz-Thompson schatter voor het populatiegemiddelde afleiden, notatie $\hat{\bar{Y}}_{HT}$, als:

$$\hat{\bar{Y}}_{HT} = \frac{1}{N} \hat{Y}_{HT} = \frac{1}{N} \sum_{k=1}^{n} \frac{y_k}{\pi_k} \quad ; \tag{5.4}$$

Zie Horvitz en Thompson (1952). De HT-schatter is zeer algemeen en kan worden toegepast op elke steekproef *zonder* teruglegging.

Zuiverheidsanalyses

Om de zuiverheid van HT-schatters (5.3) en (5.4) aan te tonen, gaan we over op een andere notatiewijze voor (5.3) en (5.4). Met behulp van de insluitindicator kunnen we de formules voor de beide HT-schatters herschrijven als:

$$\begin{aligned} \hat{Y}_{HT} &= \sum_{k=1}^{n} \frac{y_k}{\pi_k} = \sum_{k=1}^{N} \left(\frac{a_k}{\pi_k} \right) Y_k \\ \hat{\bar{Y}}_{HT} &= \frac{1}{N} \sum_{k=1}^{n} \frac{y_k}{\pi_k} = \frac{1}{N} \sum_{k=1}^{N} \left(\frac{a_k}{\pi_k} \right) Y_k \quad . \end{aligned} \tag{5.5}$$

Omdat $E(a_k) = \pi_k$ zijn de schatters in (5.5) zuiver ofwel:

$$E\left(\hat{Y}_{HT}\right) = Y \qquad \wedge \qquad E\left(\hat{\overline{Y}}_{HT}\right) = \overline{Y} \quad . \tag{5.6}$$

Variantieberekeningen

Bij het berekenen van de variantie van de HT-schatters (5.3) en (5.4), maken we gebruik van de volgende resultaten betreffende de (co)varianties van de insluitindicator a_k:

$$\begin{aligned} \operatorname{var}(a_k) &= \pi_k(1-\pi_k) \\ \operatorname{cov}(a_k, a_l) &= \pi_{kl} - \pi_k\pi_l \end{aligned} \qquad 1 \le k,l \le N \quad . \tag{5.7*}$$

Dit resultaat ligt in het verlengde van eigenschap (2.9) van de insluitindicator. Nu is het mogelijk geworden de volgende formule voor de variantie van de HT-schatter af te leiden:

$$\operatorname{var}\left(\hat{Y}_{HT}\right) = \sum_{k=1}^{N}\sum_{l=1}^{N}\left(\frac{\pi_{kl} - \pi_k\pi_l}{\pi_k\pi_l}\right) Y_k Y_l \quad . \tag{5.8*}$$

Met andere woorden, de variantie van de HT-schatter voor het populatietotaal is een lineaire combinatie van alle mogelijke producten $Y_k Y_l$ met $k,l = 1,\cdots,N$.

Om variantie (5.8) op basis van waarnemingen te kunnen bepalen, introduceren we de volgende schatter, die wel de *HT-schatter* van de variantie wordt genoemd:

$$\hat{\operatorname{var}}_{HT}\left(\hat{Y}_{HT}\right) = \sum_{k=1}^{n}\sum_{l=1}^{n}\frac{\pi_{kl} - \pi_k\pi_l}{\pi_{kl}\pi_k\pi_l} y_k y_l \quad . \tag{5.9}$$

Wanneer n vast is, bestaat er een alternatief voor deze schatter, die er als volgt uit ziet (Sen (1953) en Yates en Grundy (1953)):

$$\hat{\operatorname{var}}_{SYG}(\hat{Y}_{HT}) = -\frac{1}{2}\sum_{k=1}^{n}\sum_{l=1}^{n}\frac{\pi_{kl} - \pi_k\pi_l}{\pi_{kl}}\left(\frac{y_k}{\pi_k} - \frac{y_l}{\pi_l}\right)^2 \quad . \tag{5.10}$$

Het is mogelijk aan te tonen, dat schatter (5.9) in verwachtingswaarde overeenkomt met de variantie berekend in (5.8), zie Appendix (paragraaf 5.5). Eenzelfde resultaat geldt voor schatter (5.10). We hebben hiermee bewezen, dat HT-schatter (5.9) en SYG-schatter (5.10) zuivere schatters zijn voor de variantie van de HT-schatter van het populatietotaal.

Variantieschatter (5.10) heeft doorgaans een wat kleinere variantie dan (5.9), maar beide schatters kunnen in de praktijk een negatieve waarde aannemen. Beschouw bijvoorbeeld de extreme situatie, dat π_k precies evenredig is met Y_k, dan is de HT-schatter met kans 1 gelijk aan Y. Derhalve is de variantie van de HT-schatter in deze situatie gelijk aan 0. De HT-variantieschatter (5.9) is echter niet 0 met kans 1, maar is wel zuiver. Dit laatste betekent dat bij een deel van alle mogelijke steek-

proeven de HT-variantieschatter een negatieve waarde aanneemt. Daarentegen is variantieschatter (5.10) in deze situatie wel met kans 1 gelijk aan 0.

In de praktijk zijn vaak de exacte uitdrukkingen voor de π_{kl} moeilijk traceerbaar waardoor het wat lastig wordt de variantieformules (5.9) en (5.10) te gebruiken. Wel bestaan er in bepaalde situaties redelijke benaderingen voor de varianties. Een goede approximatie is bijvoorbeeld voorhanden wanneer n veel kleiner is dan N. In dat geval kan men doen alsof de steekproef met teruglegging is getrokken. In de volgende deelparagraaf komen we hier op terug.

Ten slotte zij nog opgemerkt dat voor een EAZT-steekproef geldt (zie hoofdstuk 2):

$$\pi_k = \frac{n}{N} \qquad \wedge \qquad \pi_{kl} = \frac{n(n-1)}{N(N-1)} \qquad\qquad k,l = 1,\cdots,N \quad .$$

Het wordt aan de lezer overgelaten na te gaan, dat substitutie van deze twee formules in (5.8)-(5.10) inderdaad leidt tot de variantieformules zoals die in hoofdstuk 2 zijn afgeleid voor de EAZT-steekproef. Overigens is het bij de afleidingen raadzaam eerst met de iets eenvoudigere SYG-formules te beginnen.

5.3.3 De PPS-steekproef

De PPS-steekproef zonder teruglegging

In een PPS-steekproef zonder teruglegging van omvang n, is de insluitkans π_k van populatie-element k per definitie evenredig met een gegeven hulpvariabele X_k. Aangezien de insluitkansen van de populatie-elementen per definitie optellen tot de steekproefomvang, is π_k als volgt gedefinieerd:

$$\pi_k = n\frac{X_k}{X} \qquad\qquad k = 1,\cdots,N \quad .$$

Met andere woorden, de PPS-steekproef zonder teruglegging van omvang n is niets anders dan een steekproef zonder teruglegging van omvang n met variabele insluitkansen. De resultaten uit de paragraaf 5.3.2 zijn dus zonder meer van kracht voor dit type steekproefontwerp.

De PPS-steekproef met teruglegging

Bij het trekken van een steekproef *met* teruglegging van omvang n, heeft bij alle n trekkingen het populatie-element k dezelfde kans om daadwerkelijk getrokken te worden. Deze kans noemen we de *trekkingskans* van populatie-element k en zij wordt aangeduid met p_k.

In het geval van een PPS-steekproef *met* teruglegging is de trekkingskans p_k van populatie-element k ($k = 1,\cdots,N$) gelijk aan:

$$p_k = \frac{X_k}{X} = \frac{\pi_k}{n} \qquad \wedge \qquad \sum_{k=1}^{N} p_k = 1 \quad .$$

Hierbij staat π_k voor de insluitkans van element k bij een PPS-steekproef *zonder* teruglegging. Met andere woorden, er geldt $\pi_k = np_k = nX_k / X$.

De standaard schatter voor het populatietotaal Y in een PPS-steekproef *met* teruglegging is de zgn. Hansen-Hurwitz schatter, kortweg HH-schatter genoemd; zie Hansen en Hurwitz (1943). De HH-schatter van het populatietotaal Y is als volgt gedefinieerd:

$$\hat{Y}_{HH} = \bar{z}_s = \frac{1}{n}\sum_{k=1}^{n} z_k \qquad \text{met} \qquad z_k \equiv \frac{y_k}{p_k} = \frac{y_k}{x_k} X \quad .$$

Merk op, dat z_k ($k = 1, \cdots, n$) kan worden gezien als een stochast die met kans p_l de waarde $Z_l = Y_l / p_l$ aanneemt ($l = 1, \cdots, N$); zie ook paragraaf 2.3.3. De zuiverheid van de HH-schatter volgt dan uit het feit, dat:

$$E(z_k) = \sum_{l=1}^{N} p_l Z_l = \sum_{l=1}^{N} p_l \frac{Y_l}{p_l} = Y \qquad k = 1, \cdots, n \quad .$$

Definieer verder:

$$\sigma_z^2 = \sum_{l=1}^{N} p_l \left(\frac{Y_l}{p_l} - Y \right)^2 \quad .$$

Omdat de variantie van z_k gelijk is aan:

$$\text{var}(z_k) = E(z_k - Y)^2 = \sum_{g=1}^{N} p_l \left(\frac{Y_l}{p_l} - Y \right)^2 = \sigma_z^2 \qquad k = 1, \cdots, n \quad .$$

en omdat de steekproef met teruglegging wordt getrokken, is $\text{var}(\hat{Y}_{HH})$ gelijk aan:

$$\text{var}\left(\hat{Y}_{HH}\right) = \frac{1}{n^2}\sum_{k=1}^{n} \text{var}(z_k) = \frac{\sigma_z^2}{n} \quad . \tag{5.11}$$

Deze variantie kan zuiver worden geschat met:

$$\begin{aligned} \hat{\text{var}}\left(\hat{Y}_{HH}\right) &= \frac{s_z^2}{n} \\ &= \frac{1}{n(n-1)}\sum_{k=1}^{n} (z_k - \bar{z}_s)^2 \\ &= \frac{1}{n(n-1)}\sum_{k=1}^{n} \left(\frac{y_k}{p_k} - \hat{Y}_{HH} \right)^2 \quad . \end{aligned} \tag{5.12}$$

De zuiverheid van deze variantieschatter volgt uit:

$$E\left(s_z^2\right) = \frac{n}{n-1} E\left\{\frac{1}{n}\sum_{k=1}^{n}(z_k - Y)^2 - (\bar{z}_s - Y)^2\right\}$$
$$= \frac{n}{n-1}\left\{\frac{1}{n}\sum_{k=1}^{n}\sigma_z^2 - \frac{\sigma_z^2}{n}\right\} = \sigma_z^2 \quad .$$

Verder geldt er, dat wanneer n veel kleiner is dan N, var($\hat{Y}_{HT}$) bij een PPS-steekproef *zonder* teruglegging redelijk goed kan worden geschat met:

$$\hat{\text{var}}_{HH}\left(\hat{Y}_{HT}\right) = \frac{s_z^2}{n} = \frac{1}{n(n-1)}\sum_{k=1}^{n}\left(\frac{y_k}{x_k / X} - \hat{Y}_{HT}\right)^2 \quad . \tag{5.13}$$

Bij deze benadering is de onderliggende aanname dat de variantie van de PPS-schatter van Y nauwelijks afhangt van het al dan niet terugleggen van de getrokken elementen.

5.3.4 Een simpele variantieschatter bij een systematische PPS-steekproef

Hoewel variantieschatter (5.13) is bedoeld voor PPS-steekproeven *met* teruglegging, kan hij met een geringe aanpassing ook worden gebruikt voor PPS-steekproeven zonder teruglegging wanneer *n* niet echt veel kleiner is dan *N*. Deze aangepaste variantieschatter wordt onder meer impliciet genoemd door Hájek (1964) voor een vergelijkbaar steekproefontwerp, *rejective sampling* genaamd, en is als volgt:

$$\hat{\text{var}}_{Haj}\left(\hat{Y}_{HT}\right) = \frac{1}{n(n-1)}\sum_{k=1}^{n}(1-\pi_k)\left(\frac{y_k}{x_k / X} - \hat{Y}_{HT}\right)^2 \quad . \tag{5.14}$$

De door Hájek (1964, p. 1520) voor *rejective sampling* afgeleide variantieformule is

$$\text{var}_{Haj}\left(\hat{Y}_{HT}\right) = \frac{1}{n^2}\sum_{k=1}^{N}\pi_k(1-\pi_k)\left(\frac{y_k}{x_k / X} - Y^*\right)^2$$
$$Y^* = \sum_{k=1}^{N}\alpha_k \frac{y_k}{x_k / X} \qquad \alpha_k = \frac{\pi_k(1-\pi_k)}{\sum_{k=1}^{N}\pi_k(1-\pi_k)} \quad .$$

Variantieschatter (5.14) kan worden gebruikt bij PPS zonder teruglegging onder de voorwaarde, dat de elementen van de populatie eerst in een random volgorde worden gezet voordat de systematische PPS-steekproef wordt getrokken. Hij kan ook worden gebruikt wanneer n en N van dezelfde ordegrootte zijn mits Y_k / X_k en X_k niet zijn gecorreleerd. Zijn Y_k / X_k en X_k wel gecorreleerd en zijn n en N bovendien van dezelfde ordegrootte dan is het beter de volgende variantieschatter te gebruiken:

$$\hat{\text{var}}_{Haj2}\left(\hat{Y}_{HT}\right)=\frac{1}{n(n-1)}\sum_{k=1}^{n}\left(1-\pi_k\right)\left(\frac{y_k}{x_k/X}-\hat{Y}^*\right)^2$$

$$\hat{Y}^*=\frac{\sum_{k=1}^{n}\left(1-\pi_k\right)\frac{y_k}{x_k/X}}{\sum_{k=1}^{n}\left(1-\pi_k\right)} \quad .$$

Voor een vergelijking van de varianties van de HT-schatter in een systematische PPS-steekproef en de quotiëntschatter in combinatie met een EAZT-steekproef, zie Knottnerus (2009). Ten slotte zij nog opgemerkt dat indien $\pi_k = n/N$ ofwel $x_k = X/N,$ (5.14) overgaat in de bekende schatter (2.12) voor de variantie van de directe schatter $\hat{Y}_{EAZT} = N\overline{y}_s$.

5.4 Voorbeeld

Om een indruk te krijgen van de variantiereductie bij een systematische PPS-steekproef gaan we in deze paragraaf nader in op het schatten van een (samengestelde) prijsindex van 70 bedrijven op basis van een steekproef met omvang $n=9$. Hierbij vergelijken we de variantie van een systematische PPS-steekproef met die van een EAZT-steekproef. De prijsmutaties zijn afkomstig van de prijswaarnemingen van Prodcom 27 van december 2004 tot en met december 2005. De data staan vermeld in tabel 5.1. Het voorbeeld is overgenomen uit Knottnerus (2009). Overigens zijn de grootste twee bedrijven weggelaten omdat hun omzetaandeel groter was dan 1/9. Zij worden als separaat stratum integraal waargenomen.

Beschouw nu de volgende prijsindex voor N bedrijven:

$$I=\sum_{k=1}^{N}W_kP_k \qquad (N=70)$$

P_k : de prijsmutatie bij bedrijf k in genoemde periode in procenten

W_k : het omzetaandeel van bedrijf k in het basisjaar $\quad(\sum_{k=1}^{N}W_k=1)$.

Merk op, dat bij een prijsindex de doelvariabele gelijk is aan $Y_k = W_kP_k$. Aan de hand van dit voorbeeld zullen we de twee steekproefontwerpen EAZT en PPS plus de bijbehorende schatters met elkaar vergelijken.

De standaard schatter voor een prijsindex op basis van een EAZT-steekproef is een pure quotiëntschatter en is gedefinieerd als:

$$\hat{I}_{EAZT}=\frac{\sum_{k=1}^{n}w_kp_k}{\sum_{k=1}^{n}w_k} \quad . \tag{5.15}$$

Ofschoon we later in hoofdstuk 6 het schatten van ratio's zullen behandelen, geven we hier alvast de vaak gebruikte benadering voor de variantie van een geschatte ratio:

$$\begin{aligned}\text{var}\left(\hat{I}_{EAZT}\right) &= \text{var}\left(\frac{\sum_{k=1}^{n} w_k p_k}{\sum_{k=1}^{n} w_k}\right) \\ &= \text{var}\left(\frac{1/n\sum_{k=1}^{n} w_k p_k}{1/n\sum_{k=1}^{n} w_k} - I\right) \\ &= \text{var}\left(\frac{1}{n}\sum_{k=1}^{n}\frac{w_k\left(p_k - I\right)}{\overline{w}_s}\right) \quad .\end{aligned}$$

Omdat de variantie van de noemer in de laatste uitdrukking relatief gering is in vergelijking met de volatiliteit van de teller, is het gebruikelijk de noemer te vervangen door het populatiegemiddelde van de W_k ofwel $1/N$. Dit geeft:

$$\begin{aligned}\text{var}\left(\hat{I}_{EAZT}\right) &= N^2\,\text{var}\left(\frac{1}{n}\sum_{k=1}^{n} w_k\left(p_k - I\right)\right) \\ &= \frac{N^2}{n}\left(1-\frac{n}{N}\right)\left(\frac{1}{N-1}\right)\sum_{k=1}^{N} W_k^2\left(P_k - I\right)^2 \quad .\end{aligned} \tag{5.16}$$

In de laatste regel hebben we gebruik gemaakt van formule (2.12) met:

$$Y_k = W_k\left(P_k - I\right) \qquad \wedge \qquad \overline{Y}_p = 0 \quad .$$

Passen we deze formule toe op de 70 prijsmutaties in tabel 5.1, dan krijgen we:

$$\text{var}\left(\hat{I}_{EAZT}\right) = 101 \quad . \tag{5.17}$$

Zou de steekproef met teruglegging zijn getrokken, dan zou de variantie zonder eindigheidscorrectie gelijk zijn geweest aan:

$$\text{var}\left(\hat{I}_{EAMT}\right) = 116 \quad .$$

Met andere woorden, de eindigheidscorrectie bedraagt $(1-f) = 0.87$.

De variantieformules in (5.16) en (5.17) zullen we nu gaan vergelijken met de variantie van een tweede schatter die is gebaseerd op een systematische PPS–steekproef, waarbij de insluitkansen evenredig zijn met de W_k. Zoals we aan het begin van paragraaf 5.3.3 gezien hebben, geldt er dan dat $\pi_k = W_k n$.

Tabel 5.1. De prijsmutaties (P_k) *en omzetaandelen* (W_k) *van 70 bedrijven*

k	prijsmutatie	omzetaandeel	k	prijsmutatie	omzetaandeel
1	-18,4%	0,0608	36	34,8%	0,0427
2	-16,0%	0,0784	37	13,1%	0,0121
3	3,3%	0,0762	38	31,7%	0,0351
4	12,5%	0,0100	39	-24,8%	0,0074
5	0,0%	0,0029	40	55,3%	0,0009
6	8,3%	0,0006	41	40,5%	0,0066
7	-39,0%	0,0182	42	34,6%	0,0022
8	-25,1%	0,0020	43	1,7%	0,0001
9	1,1%	0,0040	44	0,0%	0,0039
10	4,4%	0,0066	45	3,9%	0,0304
11	-4,9%	0,0039	46	25,4%	0,0209
12	-8,9%	0,0070	47	25,6%	0,0062
13	-7,0%	0,0148	48	0,0%	0,0033
14	-15,0%	0,0108	49	-0,3%	0,0019
15	-10,7%	0,0087	50	66,6%	0,0346
16	-9,0%	0,1079	51	0,0%	0,0039
17	-11,3%	0,0247	52	-2,9%	0,0007
18	10,6%	0,0024	53	15,8%	0,0011
19	-23,2%	0,0001	54	0,0%	0,0026
20	-25,4%	0,0001	55	0,0%	0,0018
21	-80,7%	0,0002	56	11,6%	0,0057
22	13,4%	0,0005	57	0,0%	0,0042
23	-42,5%	0,0010	58	0,0%	0,0236
24	-34,8%	0,0014	59	-1,5%	0,0015
25	-30,0%	0,0126	60	0,0%	0,0003
26	8,0%	0,0530	61	11,7%	0,0067
27	0,0%	0,0208	62	0,0%	0,0012
28	2,1%	0,0119	63	0,8%	0,0040
29	11,3%	0,0208	64	2,0%	0,0009
30	0,7%	0,0322	65	2,3%	0,0018
31	9,5%	0,0447	66	4,7%	0,0026
32	11,5%	0,0018	67	0,9%	0,0064
33	5,8%	0,0174	68	-1,0%	0,0309
34	-6,9%	0,0197	69	-0,5%	0,0005
35	0,0%	0,0124	70	0,0%	0,0006

De PPS-schatter van I wordt nu overeenkomstig (5.3):

$$\hat{I}_{PPS} = \sum_{k=1}^{n} \frac{y_k}{\pi_k} = \sum_{k=1}^{n} \frac{w_k p_k}{w_k n} = \frac{1}{n} \sum_{k=1}^{n} p_k = \bar{p}_s \quad .$$

Met andere woorden, bij een PPS-steekproef wordt het *gewogen* gemiddelde van de afzonderlijke prijsmutaties in de populatie geschat met het *ongewogen* gemiddelde van de prijsmutaties in de steekproef.

Eerst kijken we naar de variantie van een PPS-steekproef met teruglegging. Formule (5.11) wordt in dit voorbeeld:

$$\text{var}_{HH}\left(\hat{I}_{PPS}\right) = \frac{\sigma_z^2}{n} = \frac{1}{n} \sum_{k=1}^{N} W_k \left(P_k - I\right)^2 = 44 \quad .$$

Vervolgens kijken we naar wat de variantiebenadering oplevert die ten grondslag ligt aan (5.14):

$$\begin{aligned} \text{var}\left(\hat{I}_{PPS}\right) &= \frac{1}{n^2} \sum_{k=1}^{N} \pi_k \left(1 - \pi_k\right)\left(P_k - I\right)^2 \\ &= \frac{1}{n} \sum_{k=1}^{N} W_k \left(1 - n W_k\right)\left(P_k - I\right)^2 = 29 \quad . \end{aligned}$$

Met andere woorden, bij de PPS-steekproef is de eindigheidscorrectie van 0,66 ($= 29/44$) nog weer veel kleiner dan die van 0,87 bij de EAZT-steekproef. In tabel 5.2 staat een overzicht van de diverse varianties.

Tabel 5.2. Varianties bij diverse steekproefontwerpen

	met teruglegging	zonder teruglegging
EA	116	101
PPS	44	29

Ten slotte zij vermeld, dat een simulatie waarbij 20.000 systematische PPS-steekproeven van omvang $n = 9$ werden getrokken uit een (steeds opnieuw) random geordende populatie van 70 bedrijven een variantie van 30 voor $\hat{I}_{PPS}$ opleverde. Met andere woorden, (5.14) is in deze situatie een nagenoeg zuivere variantieschatter. Ook illustreert dit voorbeeld nog eens dat de variantiereductie van een systematische PPS-steekproef in vergelijking met een EAZT-steekproef aanzienlijk kan zijn, ca. 70% . ■

5.5 Kwaliteitsindicatoren

Kwaliteitsindicatoren voor wat betreft een PPS-steekproef zijn:

- de onzekerheidsmarges van de desbetreffende schatters;
- de omvang van de non-respons;
- het random karakter van de volgorde van de elementen in de populatie en de afhankelijkheid tussen Y_k / X_k en X_k.

De non-respons kan de kwaliteit van de resultaten vooral aantasten wanneer (*i*) de omvang van de non-respons groot is en (*ii*) de non-respons selectief is. Soms kan de vertekening ten gevolge van selectieve non-respons voor een deel worden gecorrigeerd door gebruik te maken van hulpvariabelen die samenhangen met zowel de responskans als met de doelvariabele.

Verder is een belangrijke aanname bij een kanssteekproef dat het *kader* waaruit de steekproef wordt getrokken goed overeenkomt met de *doelpopulatie* waarover men een uitspraak wil doen.

In het bovenstaande hebben we vooral aandacht besteed aan systematische PPS-steekproeven waarbij de elementen in een random volgorde staan of kunnen worden geacht te staan. Wanneer er geen noemenswaardig verband bestaat tussen Y_k / X_k en k ($k = 1, \cdots, N$), zal een random ordening van de populatie geen systematisch effect hebben op de uitkomsten zoals we eerder in paragraaf 5.1.2 hebben opgemerkt. In een dergelijke situatie is een random ordening derhalve niet echt nodig. Wanneer de elementen in een vaste volgorde blijven staan, dient te worden voorkomen dat de steekproeven homogeen gaan worden voor wat betreft de Y_k^* ($\equiv Y_k / \pi_k$).

5.6 Appendix

Voor de N insluitindicatoren geldt per definitie:

$$\sum_{k=1}^{N} a_k = a_1 + \cdots + a_N = n \quad .$$

Het nemen van de verwachting aan beide zijden geeft:

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

Evenzo geldt er per definitie:

$$a_k \left(a_1 + \cdots + a_N \right) = n\, a_k \quad .$$

Door ook hier aan weerszijden de verwachtingen te nemen verkrijgt men meteen:

$$\sum_{l=1}^{N} \pi_{kl} = n \pi_k \quad .$$

Deze twee resultaten hebben we later in deze paragraaf nodig.

Bewijs van (5.7)

Voor insluitindicator a_k geldt:

$$\begin{aligned}\text{var}(a_k) &= E(a_k - E(a_k))^2 \\ &= E(a_k^2) - E^2(a_k) \\ &= E(a_k) - \pi_k^2 = \pi_k - \pi_k^2\end{aligned}$$

en:

$$\begin{aligned}\text{cov}(a_k, a_l) &= E((a_k - E(a_k))(a_l - E(a_l))) \\ &= E(a_k a_l) - E(a_k)E(a_k) \\ &= \pi_{kl} - \pi_k \pi_l \quad .\end{aligned}$$

Het bewijs van (5.7) is hiermee compleet. ■

Bewijs van (5.8)

Uit de definitie van de HT-schatter voor het populatietotaal, zie (5.5), en (5.7) volgt dat zijn variantie gelijk is aan:

$$\begin{aligned}\text{var}(\hat{Y}_{HT}) &= \sum_{k=1}^{N} \pi_k (1-\pi_k) \frac{Y_k^2}{\pi_k^2} + \sum_{k=1}^{N} \sum_{\substack{l=1 \\ l \neq k}}^{N} (\pi_{kl} - \pi_k \pi_l) \frac{Y_k Y_l}{\pi_k \pi_l} \\ &= \sum_{k=1}^{N} \sum_{l=1}^{N} \left(\frac{\pi_{kl} - \pi_k \pi_l}{\pi_k \pi_l} \right) Y_k Y_l \quad .\end{aligned}$$

Het bewijs van (5.8) is hiermee gegeven. ■

Voor schatters (5.9) en (5.10) kunnen we de volgende resultaten bewijzen:

$$E(\hat{\text{var}}_{HT}(\hat{Y}_{HT})) = \text{var}(\hat{Y}_{HT}) \quad \wedge \quad E(\hat{\text{var}}_{SYG}(\hat{Y}_{HT})) = \text{var}(\hat{Y}_{HT}) \quad .$$

De bewijzen gaan als volgt. De HT-schatter $\hat{\text{var}}(\hat{Y}_{HT})$ (zie (5.9)) herschrijven we in termen van de insluitindicatoren als:

$$\hat{\text{var}}_{HT}(\hat{Y}_{HT}) = \sum_{k=1}^{n} \sum_{l=1}^{n} \frac{\pi_{kl} - \pi_k \pi_l}{\pi_{kl} \pi_k \pi_l} y_k y_l = \sum_{k=1}^{N} \sum_{l=1}^{N} \left(a_k a_l \frac{\pi_{kl} - \pi_k \pi_l}{\pi_{kl} \pi_k \pi_l} \right) Y_k Y_l \quad .$$

De zuiverheid van de variantieschatter volgt nu uit de eigenschap: $E(a_k a_l) = \pi_{kl}$ (zie (2.9)). We kunnen immers berekenen, dat geldt:

$$E\left(\hat{\text{var}}_{HT}\left(\hat{Y}_{HT}\right)\right) = \sum_{k=1}^{N}\sum_{l=1}^{N}\left(E(a_k a_l)\frac{\pi_{kl}-\pi_k\pi_l}{\pi_{kl}\pi_k\pi_l}\right)Y_k Y_l$$

$$= \sum_{k=1}^{N}\sum_{l=1}^{N}\left(\pi_{kl}\times\frac{\pi_{kl}-\pi_k\pi_l}{\pi_{kl}\pi_k\pi_l}\right)Y_k Y_l$$

$$= \sum_{k=1}^{N}\sum_{l=1}^{N}\left(\frac{\pi_{kl}-\pi_k\pi_l}{\pi_k\pi_l}\right)Y_k Y_l = \text{var}_{HT}\left(\hat{Y}_{HT}\right) \quad .$$

Voorwaarde voor de zuiverheid is wel weer, dat $0 < \pi_{kl} \quad 1 \le k,l \le N$.

Omdat $E(a_k a_l) = \pi_{kl}$, is schatter (5.10) op zich een zuivere schater van:

$$\text{var}_{SYG}\left(\hat{Y}_{HT}\right) = -\frac{1}{2}\sum_{k=1}^{N}\sum_{l=1}^{N}\left(\pi_{kl}-\pi_k\pi_l\right)\left(\frac{Y_k}{\pi_k}-\frac{Y_l}{\pi_l}\right)^2$$

mits $0 < \pi_{kl}$. Met andere woorden, er geldt:

$$E\left(\hat{\text{var}}_{SYG}\left(\hat{Y}_{HT}\right)\right) = -\frac{1}{2}\sum_{k=1}^{N}\sum_{l=1}^{N}\left(\pi_{kl}-\pi_k\pi_l\right)\left(\frac{Y_k}{\pi_k}-\frac{Y_l}{\pi_l}\right)^2 \quad .$$

Door het kwadraat in het rechterlid van bovenstaande relatie uit te werken en gebruik te maken van de twee in het begin van deze paragraaf afgeleide eigenschappen van de insluitindicatoren $\sum_{k=1}^{N}\pi_k = n$ en $\sum_{l=1}^{N}\pi_{kl} = n\pi_k$, komt men weer uit op:

$$E\left(\hat{\text{var}}_{SYG}\left(\hat{Y}_{HT}\right)\right) = \sum_{k=1}^{N}\sum_{l=1}^{N}\left(\frac{\pi_{kl}-\pi_k\pi_l}{\pi_k\pi_l}\right)Y_k Y_l \quad .$$

Met andere woorden, schatter (5.10) is ook een zuivere schatter van (5.8).

6. Quotiëntschatter

6.1 Korte beschrijving

In hoofdstuk 2 hebben we gezien dat het steekproefgemiddelde een zuivere schatter is voor het populatiegemiddelde. Het wordt ook wel de *directe* schatter voor $\overline{Y}$ genoemd, omdat alleen gebruik wordt gemaakt van de waargenomen waarden van de variabele Y. In de hoofdstukken daarna hebben we gezien hoe de precisie van een schatter verhoogd kan worden door over de populatie beschikbare informatie te verwerken in het steekproefontwerp. Zo werd in hoofdstuk 3 hulpinformatie gebruikt om de populatie in homogene groepen of strata in te delen waaruit vervolgens afzonderlijke steekproeven worden getrokken. Hoofdstuk 5 toonde dat hulpinformatie met betrekking tot de omvang van de elementen het mogelijk maakt de elementen met ongelijke kansen te trekken. In dit hoofdstuk en in de volgende twee hoofdstukken, zullen we laten zien hoe hulpinformatie gebruikt kan worden in de schattingsprocedure. Afhankelijk van het soort hulpinformatie dat we gebruiken zal dit tot verschillende schatters leiden. Aan de hand van het eenvoudigste steekproefontwerp, het enkelvoudige, aselecte steekproefontwerp (zonder terugleggen), zullen we enkele veelgebruikte alternatieve schatters voor het populatiegemiddelde (of totaal) introduceren. We beperken ons daarbij tot schattingsmethoden waarbij gebruikt wordt gemaakt van één hulpvariabele. Net als bij gestratificeerde steekproeven en steekproeven met ongelijke kansen geldt hier dat het gebruik van hulpinformatie tot preciezere schatters zal leiden als er voldoende sterke samenhang is tussen de doelvariabele(n) en de hulpvariabele(n).

Tot nu toe hebben we de belangrijkste bouwstenen van steekproefonderzoek besproken: enkelvoudige, aselecte steekproeven, stratificatie, clustering, trekken met ongelijke kansen. Quotiënt- en regressieschatters komen hierna nog aan de orde. De meeste steekproefontwerpen die op het CBS gebruikt worden, zijn opgebouwd uit meerdere bouwstenen. De formules voor de schatters en met name die voor de standaardfouten kunnen dan al vrij snel zeer ingewikkeld worden. In de praktijk worden gewichten gebruikt om puntschattingen van parameters te verkrijgen. Dit maakt de zaak meestal een stuk eenvoudiger. De gewichten kunnen met behulp van het weegpakket Bascula automatisch worden bepaald. Dit pakket bevat ook verschillende methoden om varianties te berekenen. In de resterende hoofdstukken gaan we nog kort in op de relatie tussen schatten en wegen.

6.2 Toepasbaarheid

De quotiëntschatter is een relatief eenvoudige en daardoor ook veel gebruikte alternatieve schatter voor de directe schatter zoals die is besproken in hoofdstuk 2. De quotiëntschatter, die gebruik maakt van een kwantitatieve hulpvariabele, leunt op de gedachte dat het quotiënt Y_k / X_k van de waarden van de doelvariabele en de hulpvariabele in de populatie minder varieert dan de doelvariabele zelf.

Denk bijvoorbeeld aan de omzet per werknemer. Indien deze ratio of dit quotiënt weinig varieert per bedrijf kan men daar gebruik van maken om de totale omzet in Nederland te schatten door die ratio te schatten op basis van een steekproef en vervolgens die geschatte ratio te vermenigvuldigen met het aantal werknemers in Nederland.

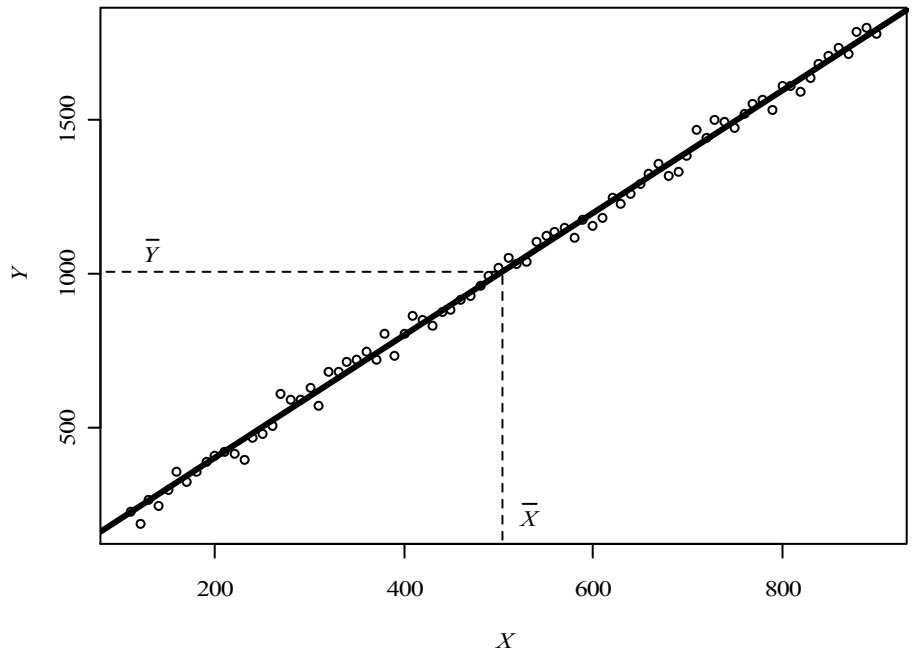
6.3 Uitgebreide beschrijving

Het weinig variëren van het quotiënt van de doelvariabele Y en de hulpvariabele X kan worden uitgedrukt door:

$$\frac{Y_k}{X_k} \approx B \quad \text{ofwel} \quad Y_k \approx BX_k$$

met B een constante. De punten (X_k, Y_k) liggen bij benadering op een rechte lijn door de oorsprong, ook wel de regressielijn genoemd (zie figuur 6.1).Wanneer de regressielijn de puntenwolk goed beschrijft kunnen we de individuele waarden van Y goed voorspellen aan de hand van X. En omdat we van X populatie-informatie hebben ligt het voor de hand dit te verwerken in de schatter.

Figuur 6.1. Lineair verband tussen Y en X door de oorsprong



Wanneer we nu sommeren over alle elementen van de populatie en delen door *N*, krijgen we:

$$\overline{Y} = B\overline{X}.$$

Wanneer we het populatiegemiddelde $\overline{X}$ kennen en de waarde van B is beschikbaar, dan kunnen we $\overline{Y}$ dus schatten met $B\overline{X}$. De ideale waarde voor B is natuurlijk $\overline{Y}/\overline{X}$, maar deze is niet te berekenen aangezien het kennis van $\overline{Y}$ veronderstelt. We kunnen deze waarde echter wel schatten op basis van de steekproefgegevens met:

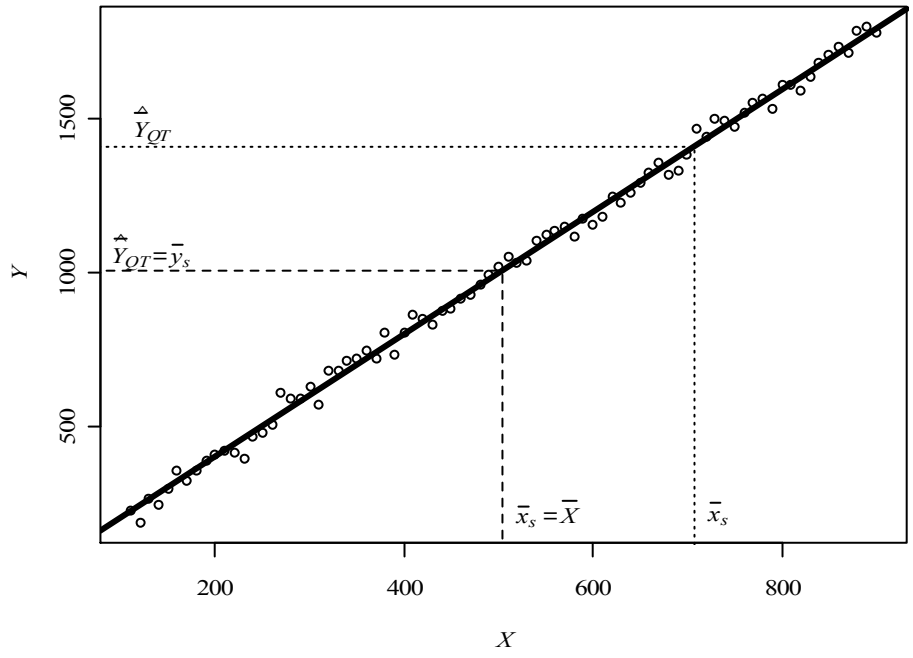
$$\hat{B} = b = \frac{\overline{y}_s}{\overline{x}_s} = \frac{\hat{\overline{Y}}}{\hat{\overline{X}}} \quad .$$

Dit geeft ons de quotiëntschatter $\hat{\overline{Y}}_{QT}$ voor het populatiegemiddelde $\overline{Y}$:

$$\hat{\overline{Y}}_{QT} = b\overline{X} = \frac{\overline{y}_s}{\overline{x}_s}\overline{X} \quad . \tag{6.1}$$

Bij de quotiëntschatter wordt het steekproefgemiddelde $\overline{y}_s$ vermenigvuldigd met een term $\overline{X}/\overline{x}_s$, die corrigeert voor het verschil tussen het steekproefgemiddelde en het populatiegemiddelde van de hulpvariabele. De aanname hierbij is dat deze correctie ook te projecteren is op de doelvariabele. Er is immers een verband tussen de doelvariabele en de hulpvariabele. In figuur 6.2 is te zien hoe de quotiëntschatter werkt.

Figuur 6.2. De quotiëntschatter



Uit figuur 6.2 blijkt, dat voor het verkrijgen van een quotiëntschatting, bij het steekproefgemiddelde $\bar{y}_s$ een term $b(\bar{X} - \bar{x}_s)$ opgeteld moet worden, waarbij $b = \bar{y}_s / \bar{x}_s$ de richting van de regressielijn aangeeft. De quotiëntschatter (6.1) kan dus ook geschreven worden als:

$$\hat{\bar{Y}}_{QT} = \bar{y}_s + b(\bar{X} - \bar{x}_s) = b\bar{X} \quad .$$

Merk op dat wanneer we $Y = X$ kiezen, de quotiëntschatter gelijk is aan $\bar{X}$. Voor wat betreft de hulpvariabele is de steekproefuitkomst dus consistent met de populatie. Dit is in de praktijk een zeer gewenste eigenschap. Het stelt ons in staat verschillende statistieken op elkaar af te stemmen door dezelfde hulpvariabele te gebruiken.

De quotiëntschatter in (6.1) kan ook in termen van gewichten worden geschreven. Dat wil zeggen:

$$\hat{\bar{Y}}_{QT} = b\bar{X} = \frac{\bar{y}_s}{\bar{x}_s}\bar{X} = \sum_{k=1}^{n} w_k y_k$$

$$w_k = \frac{\bar{X}}{n\bar{x}_s} \quad .$$

De gewichten zijn voor alle waarnemingen in de steekproef hetzelfde en voldoen aan de voorwaarde dat $\sum_{k=1}^{n} w_k x_k = \bar{X}$. Het gewicht representeert als het ware de desbetreffende fractie elementen in de populatie. Tevens geeft het gewicht een indruk van de mogelijke bijdrage (invloed) van de bewuste waarneming op de uitkomst.

We gaan er gemakshalve van uit dat de doel- en hulpvariabele positief zijn, zodat B en b ook positief zijn. Op zich mag de hulpvariabele ook de waarde nul aannemen zolang het steekproefgemiddelde $\bar{x}_s$ maar positief is. Is Y_k bijvoorbeeld de omzet van een bedrijf k in een bepaalde periode en X_k de omzet in de periode ervoor dan heeft X_k de waarde nul als het bedrijf destijds nog niet bestond. Omgekeerd geldt $Y_k = 0$ als een bedrijf inmiddels niet meer bestaat.

De quotiëntschatter is geen zuivere schatter voor het populatiegemiddelde $\bar{Y}$. Cochran (1977, blz. 160) bewijst dat de vertekening omgekeerd evenredig is met de steekproefomvang, zodat bij grote steekproeven de vertekening te verwaarlozen is. In de formule voor de schatter (6.1) is $\hat{B} = b$ een stochastische variabele en $\bar{X}$ een constante. De vertekening van $\hat{B}$ is bij benadering:

$$E[\hat{B}] - B \approx \frac{1-f}{n}\frac{1}{\bar{X}^2}\left[BS_x^2 - S_{xy}\right] = \frac{1}{N}\left(\frac{1-f}{f}\right)B\left[CV_x^2 - CV_{xy}\right] \quad .$$

Hierin is CV_x de variatiecoëfficiënt van X, zie ook (2.1). Verder geldt:

$$CV_{xy} = \frac{S_{xy}}{\bar{X}\bar{Y}} \quad .$$

De vertekening is dus bij benadering nul wanneer $CV_x^2 = CV_{xy}$. In dat geval gaat de regressielijn door de oorsprong en is Y_k ruwweg evenredig met X_k.

Zuiverheid is één van de criteria om de kwaliteit van schatters te beoordelen. Bij een zuivere schatter zijn de afwijkingen van de schattingen tot de populatiewaarde, berekend voor alle mogelijke steekproeven, gemiddeld nul. Een tweede en meestal belangrijker kwaliteitscriterium is, dat de schattingen niet teveel van steekproef tot steekproef mogen fluctueren en allen in de buurt van de werkelijke populatiewaarde liggen. Dit uit zich in een kleine standaardfout. Voor onzuivere schatters is de gemiddelde kwadratische fout vaak een belangrijker criterium dan de variantie. De (kleine) vertekening van de quotiëntschatter wordt gewoonlijk gecompenseerd door een kleine gemiddelde kwadratische fout. De gemiddelde kwadratische fout van $\hat{B}$ kunnen we uitrekenen als:

$$G\left(\hat{B}\right) = E\left(\left(\hat{B} - B\right)^2\right) = \text{var}\left(\hat{B}\right) + \left(E\left(\hat{B}\right) - B\right)^2 \quad .$$

Wanneer de vertekening van $\hat{B}$ klein is ten opzichte van de standaardfout van $\hat{B}$ is de variantie een goede benadering van de gemiddelde kwadratische fout. Er geldt (zie bijvoorbeeld, Muilwijk, 1992, paragraaf 5.8.3):

$$G(\hat{B}) \leq \left[1 + \frac{1}{N}\left(\frac{1-f}{f}\right)CV_x\right]\text{var}\left(\hat{B}\right)$$

zodat de benadering beter is, naarmate f groter en/of CV_x kleiner zijn.

De variantie van de quotiëntschatter voor het populatiegemiddelde kan goed benaderd worden door:

$$\text{var}\left(\hat{\bar{Y}}_{QT}\right) = \frac{1}{N(N-1)}\left(\frac{1-f}{f}\right)\sum_{k=1}^{N}\left(Y_k - BX_k\right)^2 \quad . \tag{6.2}$$

De variantie is kleiner naarmate het verband tussen Y en X groter is, dat wil zeggen naarmate de waarden Y_k minder afwijken van de regressielijn. De variantie (6.2) kan echter niet berekend worden omdat immers de populatiegegevens van Y ontbreken, maar we kunnen wel weer een schatting maken aan de hand van de steekproef. De variantieschatter van (6.2) is:

$$\hat{\text{var}}\left(\hat{\bar{Y}}_{QT}\right) = \frac{1}{N}\left(\frac{1-f}{f}\right)\left\{\frac{1}{n-1}\sum_{k=1}^{n}\left(y_k - bx_k\right)^2\right\} \quad . \tag{6.3}$$

Voor grote steekproeven is de variantie van de quotiëntschatter doorgaans kleiner dan die van de directe schatter. We kunnen zelfs afleiden wanneer de variantie van de quotiëntschatter kleiner is dan die van de directe schatter. Daartoe herschrijven we de variantie van de quotiëntschatter in:

$$\text{var}\left(\hat{\bar{Y}}_{QT}\right) = \frac{1}{N}\left(\frac{1-f}{f}\right)\left(S_y^2 + B^2 S_x^2 - 2BS_{xy}\right) \quad .$$

Wanneer we deze variantie vergelijken met de variantie voor het steekproefgemiddelde bij een enkelvoudig aselecte steekproef, zie (2.11), krijgen we:

$$\begin{aligned}\text{var}\left(\hat{\bar{Y}}_{QT}\right) - \text{var}\left(\hat{\bar{Y}}\right) &\approx \frac{1}{N}\left(\frac{1-f}{f}\right)\left(S_y^2 + B^2 S_x^2 - 2BS_{xy} - S_y^2\right) \\ &= \frac{1}{N}\left(\frac{1-f}{f}\right)BS_x\left(BS_x - 2\rho S_y\right) \quad .\end{aligned}$$

Hierin is ρ de populatiecorrelatiecoëfficiënt van X en Y. De quotiëntschatter is dus te prefereren boven de directe schatter als geldt:

$$\frac{1}{2}B\left(\frac{S_x}{S_y}\right) = \frac{1}{2}\left(\frac{CV_x}{CV_y}\right) < \rho \quad .$$

Indien de variatiecoëfficiënt van X dus meer dan twee keer zo groot is dan die van Y, $2CV_y < CV_x$, is de quotiëntschatter minder precies dan de directe schatter (zelfs bij een correlatie van $\rho = 1$). Is de variabele X gelijk aan de variabele Y maar op een eerder tijdstip gemeten, dan zijn de variatiecoëfficiënten ongeveer gelijk en is de quotiëntschatter beter dan de directe schatter als $0{,}5 < \rho$. Wanneer de waarden van X allemaal hetzelfde zijn $(S_x^2 = 0)$, komt de quotiëntschatter overeen met de directe schatter bij een enkelvoudige, aselecte steekproef.

De quotiëntschatter voor het populatiegemiddelde levert ons weer eenvoudig de quotiëntschatter voor het populatietotaal door met de populatieomvang te vermenigvuldigen:

$$\hat{Y}_{QT} = N\bar{Y}_{QT} = bN\bar{X} = bX \quad .$$

De variantie van deze schatter kan worden berekend als:

$$\text{var}\left(\hat{Y}_{QT}\right) = \left(\frac{N}{N-1}\right)\left(\frac{1-f}{f}\right)\sum_{k=1}^{N}\left(Y_k - BX_k\right)^2 \quad .$$

Ook de quotiëntschatter van het populatietotaal kan in termen van gewichten worden geschreven:

$$\begin{aligned}\hat{Y}_{QT} &= Nb\bar{X} = \frac{\bar{y}_s}{\bar{x}_s}X = \sum_{k=1}^{n} w_k^{\#} y_k \\ w_k^{\#} &= \frac{X}{n\bar{x}_s} \quad .\end{aligned}$$

De gewichten $w_k^{\#}$ voldoen aan $\sum_{k=1}^{n} w_k^{\#} x_k = X$. Analoog aan de quotiëntschatter voor het gemiddelde representeert het gewicht nu als het ware het desbetreffende aantal elementen in de populatie.

Meestal wordt de quotiëntschatter gebruikt om de precisie te verhogen. Daarnaast zorgt de quotiëntschatter ervoor dat de steekproefschattingen consistent zijn met populatieaantallen of totalen voor de hulpvariabelen (zie ook hoofdstuk 8) en corrigeert de schatter voor non-respons. Soms kan het voorkomen dat de parameter die we willen schatten een quotiënt is (het aantal dodelijke verkeersongevallen in verhouding tot het totale aantal ongevallen). Ook kunnen we de quotiëntschatter gebruiken wanneer we een totaal willen schatten maar de populatieomvang is onbekend. Dit komt met name voor wanneer we schattingen voor deelpopulaties willen maken.

Soms zijn er meer hulpvariabelen die correleren met de doelvariabele Y. Dan kan gekozen worden voor één hulpvariabele, bijvoorbeeld degene met de hoogste correlatie, maar het is ook mogelijk meerdere hulpvariabelen te gebruiken via de zogenaamde algemene regressieschatter. Tot slot merken we op, dat vaak niets bekend is over het verband tussen Y en X in de populatie. We kunnen dan alleen afgaan op de steekproefgegevens, mits de steekproef door toeval of non-respons geen scheve verdeling ten aanzien van Y vertoont. Ook informatie uit eerdere onderzoeken kan helpen inzicht te krijgen in het verband tussen Y en X. De quotiëntschatter is het meest geschikt wanneer een rechte lijn door de oorsprong de relatie tussen Y en X het best beschrijft en wanneer de variantie van de waarden Y_k om de regressielijn evenredig is met X_k.

6.4 Voorbeeld

Uit een populatie van 42 gemeenten wordt een enkelvoudig aselecte steekproef van 8 gemeenten getrokken. Tabel 6.1 geeft de inwoneraantallen en het aantal huisartsen per gemeente. We willen het totale aantal huisartsen in de populatie schatten. Het aantal huisartsen zal samenhangen met het aantal inwoners in een gemeente en de inwoneraantallen zullen dan ook als hulpinformatie bij de schatting worden gebruikt. Gemiddeld zijn er 23,5 huisartsen in de 8 getrokken gemeenten. Wanneer we geen gebruik maken van de hulpvariabele schatten we het totale aantal huisartsen op $\hat{Y} = N\bar{y} = 987$ met een standaardfout van 80,48. Het gemiddelde aantal inwoners in de populatie is 50 en in de steekproef 47,13. In verhouding tot de populatie is het gemiddelde aantal inwoners in de populatie dus iets te laag en waarschijnlijk is het gemiddelde aantal huisartsen in de steekproef dus ook iets te laag.

Schatten we het totale aantal huisartsen in de populatie met behulp van de quotiëntschatter dan komen we uit op:

$$\hat{Y}_{QT} = \frac{\bar{y}_s}{\bar{x}_s} X = 0{,}499 \times 2100 = 1047 \quad .$$

De standaardfout van de schatter is:

$$\hat{S}_e\left(\hat{Y}_{QT}\right) = \sqrt{N^2 \frac{1-f}{n} \frac{1}{n-1} \sum_{k=1}^{n} \left(y_k - bx_k\right)^2}$$
$$= \sqrt{42^2 \times 0{,}10 \times 114{,}68} = 54{,}08 \quad .$$

Tabel 6.1 Inwoneraantallen en aantal huisartsen in een steekproef van 8 gemeenten

Gemeente	Inwoners (x1000)	Huisartsen	$(y_k - bx_k)^2$
1	38	20	1,10
2	52	23	8,59
3	75	35	5,76
4	32	20	16,34
5	60	28	3,69
6	43	25	12,65
7	57	22	41,27
8	20	15	25,27
Totaal	377	188	114,68

■

6.5 Kwaliteitsindicatoren

Kwaliteitsindicatoren voor wat betreft de quotiëntschatter zijn:

- de onzekerheidsmarges van de desbetreffende schatters;
- de omvang van de non-respons
- de omvang van de steekproef.

De non-respons kan de kwaliteit van de resultaten vooral aantasten wanneer (i) de omvang van de non-respons groot is en (ii) de non-respons selectief is. Soms kan de vertekening ten gevolge van selectieve non-respons voor een deel worden gecorrigeerd door gebruik te maken van hulpvariabelen die samenhangen met zowel de responskans als met de doelvariabele zoals met de in dit hoofdstuk beschreven quotiëntschatter.

Verder is een belangrijke aanname bij de quotiëntschatter dat de omvang van de steekproef voldoende groot is zodat de vertekening van de quotiëntschatter en die van de schatter van zijn variantie beperkt blijven.

7. Regressieschatter

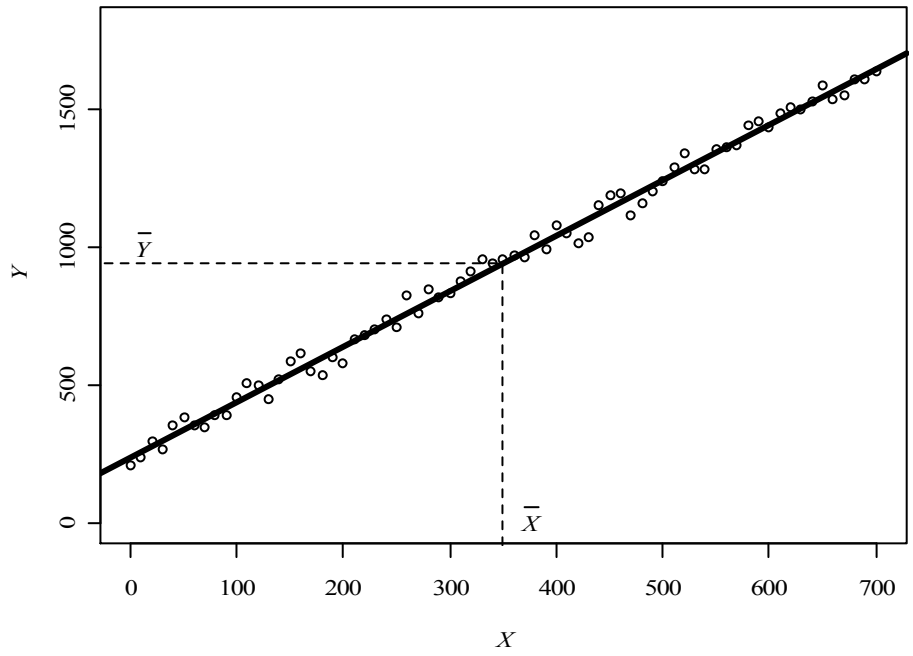
7.1 Korte beschrijving

Wanneer er sprake is van een lineair verband, echter niet door de oorsprong, is het beter in plaats van de quotiëntschatter de regressieschatter te gebruiken, zie figuur 7.1. Het verband tussen Y en X geven we nu weer door:

$$Y_k \approx A + BX_k \tag{7.1}$$

waarbij A en B constanten ongelijk aan nul zijn. Deze constanten worden de regressiecoëfficiënten genoemd.

Figuur 7.1. Lineair verband tussen Y en X



Dit lineaire verband vormt het uitgangspunt voor de regressieschatter. In paragraaf 7.3 volgt een meer gedetailleerde beschrijving van de regressieschatter.

7.2 Toepasbaarheid

De regressieschatter en allerlei varianten daarvan worden veelvuldig gebruikt bij de sociaal-economische statistieken. Voor de hulpvariabelen wordt veelal geput uit de GBA. Bij het deelthema *Herhaald wegen* (thema Steekproeftheorie van de Methodenreeks; Gouweleeuw en Knottnerus, 2008) wordt eveneens gebruik gemaakt van

de regressieschatter, zij het met een iets ander doel. Ook bij andere thema's wordt vaak de regressieschatter gehanteerd.

7.3 Uitgebreide beschrijving

Onder de aanname dat er bij benadering een lineair verband tussen Y en X geldt, geeft sommeren in (7.1) over de elementen van de populatie en delen door N de volgende benadering:

$$\overline{Y} \approx A + B\overline{X} \quad .$$

Wanneer de constanten A en B bekend zouden zijn en het populatiegemiddelde van de hulpvariabele X is eveneens bekend, dan kan het populatiegemiddelde van de doelvariabele worden geschat met $A + B\overline{X}$. In de regel beschikken we echter niet over de waarden A en B, en zullen we ze moeten schatten. Daarbij streven we naar waarden voor A en B die zodanig zijn dat voor elk element k in de populatie $A + BX_k$ zo dicht mogelijk bij Y_k ligt. Dit kunnen we bereiken door de methode van de kleinste kwadraten toe te passen, dat wil zeggen minimaliseren van de kwadratensom:

$$\sum_{k=1}^{N} \left(Y_k - A - BX_k\right)^2 \quad .$$

De kwadratensom is minimaal wanneer B gelijk is aan

$$B = \frac{\sum_{k=1}^{N} \left(X_k - \overline{X}\right)\left(Y_k - \overline{Y}\right)}{\sum_{k=1}^{N} \left(X_k - \overline{X}\right)^2} = \frac{S_{xy}}{S_x^2}$$

en wanneer A gelijk is aan:

$$A = \overline{Y} - B\overline{X} \quad .$$

De waarden van A en B kunnen op basis van de uitkomsten van de steekproef, geschat worden door:

$$\hat{B} = b = \frac{\sum_{k=1}^{n} \left(x_k - \overline{x}_s\right)\left(y_k - \overline{y}_s\right)}{\sum_{k=1}^{n} \left(x_k - \overline{x}_s\right)^2} = \frac{s_{xy}}{s_x^2} \tag{7.2}$$

en:

$$\hat{A} = a = \overline{y}_s - b\overline{x}_s \quad .$$

De regressieschatter voor het populatiegemiddelde, notatie $\hat{\overline{Y}}_{REG}$, is vervolgens gedefinieerd als:

$$\hat{\overline{Y}}_{REG} = a + b\overline{X} = \overline{y}_s + b\left(\overline{X} - \overline{x}_s\right) \quad . \tag{7.3}$$

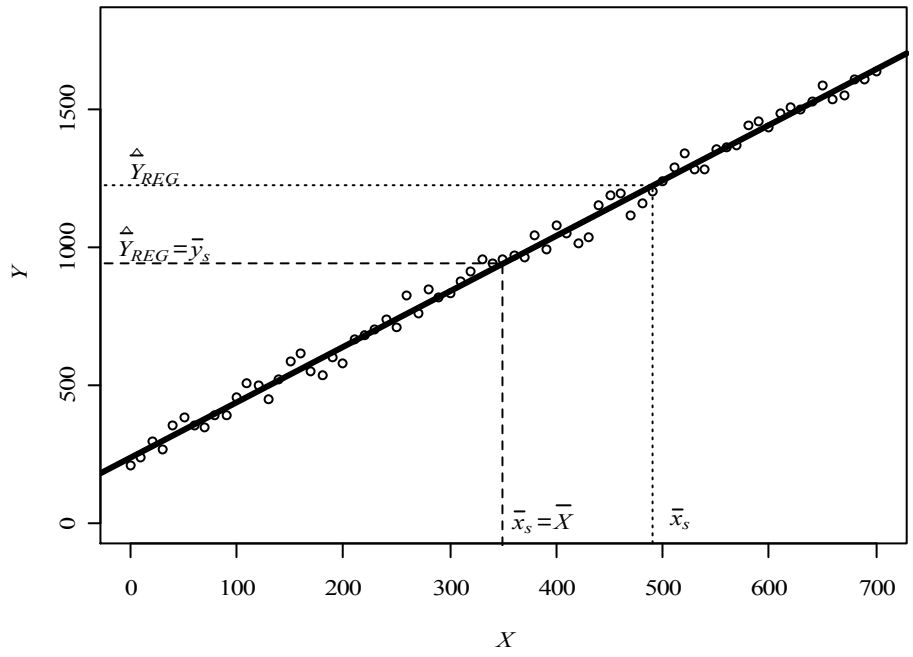
Merk op, dat wanneer we $a = 0$ kiezen, $\hat{\bar{Y}}_{REG} = b\bar{X}$ met b gelijk aan:

$$b = \frac{\sum_{k=1}^{n} x_k y_k}{\sum_{k=1}^{n} x_k^2} \quad .$$

Deze schatter is echter niet gelijk aan de quotiëntschatter. Dit heeft te maken met de verschillende uitgangspunten om schatters af te leiden (Särndal e.a. 1992, paragraaf 6.4).

Evenals de quotiëntschatter is de regressieschatter een aanpassing op het steekproefgemiddelde $\bar{y}_s$, zie figuur 7.2. Er wordt gecorrigeerd zodra er een verschil is tussen het populatiegemiddelde en het steekproefgemiddelde van de hulpvariabele X.

Figuur 7.2. De regressieschatter



Analoog aan de quotiëntschatter kan ook de regressieschatter voor het populatiegemiddelde weer in termen van gewichten worden geschreven. Substitutie van (7.2) in (7.3) geeft:

$$\hat{\bar{Y}}_{REG} = \sum_{k=1}^{n} w_k y_k$$
$$w_k = \frac{1}{n} + \frac{(x_k - \bar{x}_s)(\bar{X} - \bar{x}_s)}{\sum_{k=1}^{n} (x_k - \bar{x}_s)^2} \quad .$$

Net als bij de quotiëntschatter voldoen de gewichten w_k aan $\sum_{k=1}^{n} w_k x_k = \bar{X}$.

De regressieschatter is geen zuivere schatter voor het populatiegemiddelde. Dit komt omdat b geen zuivere schatter is voor B. De vertekening van $\hat{\bar{Y}}_{REG}$ bedraagt $-\text{cov}(\hat{B},\bar{x})$, waarbij $\text{cov}(\cdot)$ voor covariantie staat. Als de regressielijn door alle punten (X_k, Y_k) gaat, is $\hat{B} = B$ voor elke mogelijke steekproef en is $\text{cov}(\hat{B},\bar{x}) = 0$. In deze situatie is de vertekening gelijk aan nul. Evenals bij de quotiëntschatter is de vertekening echter omgekeerd evenredig met de steekproefomvang, zodat we bij grote steekproeven de vertekening kunnen verwaarlozen. De variantie van de regressieschatter kan goed benaderd worden door:

$$\text{var}\left(\hat{\bar{Y}}_{REG}\right) = \frac{1}{N(N-1)}\left(\frac{1-f}{f}\right)\sum_{k=1}^{N}\left(Y_k - \bar{Y} - B\left(X_k - \bar{X}\right)\right)^2 \quad .$$

De variantie wordt bepaald door de variatie in Y die niet door X verklaard kan worden, de zogenaamde residuele kwadratensom. Het residu voor element k, notatie e_k, is per definitie gelijk aan:

$$e_k = Y_k - \left(A + BX_k\right) = Y_k - \bar{Y} - B\left(X_k - \bar{X}\right) \quad .$$

Hoe meer van de variatie in Y verklaard wordt door X, dus hoe sterker het verband tussen Y en X, hoe kleiner de variantie. Een alternatieve manier om dit te zien is om de variantie te herschrijven in:

$$\text{var}\left(\hat{\bar{Y}}_{REG}\right) = \frac{1}{N}\left(1-\rho^2\right)\left(\frac{1-f}{f}\right)S_y^2 = \left(1-\rho^2\right)\text{var}\left(\bar{y}\right) \quad .$$

Naarmate de populatiecorrelatiecoëfficiënt dichter bij $+1$ of -1 ligt, zal de variantie kleiner worden. De variantie is altijd een factor $(1-\rho^2)$ kleiner dan de variantie van de directe schatter. Als er geen verband is tussen Y en X, dat wil zeggen $\rho = 0$, dan is de variantie van de regressieschatter gelijk aan die van de directe schatter. Het heeft in dat geval ook geen zin om X als hulpvariabele te gebruiken.

Een schatter voor de variantie van de regressieschatter op basis van de steekproef is:

$$\hat{\text{var}}\left(\hat{\bar{Y}}_{REG}\right) = \frac{1}{N}\left(\frac{1-f}{f}\right)\left\{\frac{1}{n-1}\sum_{k=1}^{n}\left(y_k - \bar{y}_s - b\left(x_k - \bar{x}_s\right)\right)^2\right\} \quad .$$

Op basis van regressieschatter (7.3) voor het populatiegemiddelde, kunnen we rechtstreeks een regressieschatter voor het populatietotaal afleiden:

$$\hat{Y}_{REG} = N\hat{\bar{Y}}_{REG} = N\bar{y}_s + Nb\left(\bar{X} - \bar{x}_s\right) \quad .$$

Ook deze regressieschatter voor het populatietotaal, kunnen we in termen van gewichten herschrijven:

$$\hat{Y}_{REG} = \sum_{k=1}^{n} w_k^{\#} y_k$$

$$w_k^{\#} = \frac{N}{n} + \frac{N(x_k - \bar{x}_s)\left(\bar{X} - \bar{x}_s\right)}{\sum_{k=1}^{n}\left(x_k - \bar{x}_s\right)^2} \quad .$$

De gewichten in deze regressieschatter voldoen aan de conditie $\sum_{k=1}^{n} w_k^{\#} x_k = X$. Zie Camstra en Nieuwenbroek (2002) en het deelthema *Herhaald wegen* voor gewichten indien er twee of meer hulpvariabelen zijn.

De variantie van de regressieschatter voor het populatietotaal kan als volgt worden uitgerekend:

$$\text{var}\left(\hat{Y}_{REG}\right) = N^2 \,\text{var}\left(\hat{\bar{Y}}_{REG}\right) \quad .$$

De regressieschatter is zowel bij positieve als bij negatieve correlatie bruikbaar; Y_k en X_k kunnen ook nul zijn of negatieve waarden aannemen.

7.4 Voorbeeld

Als we de steekproefgegevens uit voorbeeld 6.4 bekijken bestaat het vermoeden dat de samenhang tussen aantal inwoners en aantal huisartsen per gemeente weliswaar lineair is maar dat het verband niet kan worden samengevat door middel van een regressielijn door de oorsprong. Het ligt dan voor de hand de regressieschatter toe te passen. De steekproefparameters zijn:

$$\bar{x} = 47{,}13 \qquad s_x^2 = 304{,}13$$
$$\bar{y} = 23{,}50 \qquad s_y^2 = 36{,}24$$
$$s_{xy} = 95{,}97 \quad .$$

De schattingen van de regressiecoëfficiënten zijn:

$$b = \frac{s_{xy}}{s_x^2} = 0{,}31 \qquad \wedge \qquad a = \bar{y} - b\bar{x} = 8{,}66 \quad .$$

De regressieschatting voor het totale aantal huisartsen is daarom:

$$\hat{Y}_{REG} = N(a + b\bar{X}) = 42 \times (8{,}66 + 14{,}84) = 987$$

wat overeenkomt met de schatting die werd verkregen bij toepassing van de directe schatter. De standaardfout van de regressieschatter voor het totaal is:

$$\hat{S}_e\left(\hat{Y}_{REG}\right) = \sqrt{N\left(\frac{1-f}{f}\right)\left\{\frac{1}{n-1}\sum_{k=1}^{n}\left(y_k - \bar{y}_s - b\left(x_k - \bar{x}_s\right)\right)^2\right\}}$$
$$= \sqrt{42 \times \left(\frac{1-(8/42)}{(8/42)}\right) \times \left\{\frac{1}{7} \times 42{,}82\right\}} = 35{,}05 \quad .$$

De standaardfout is dus een stuk kleiner dan de standaardfout van de directe schatter en ook kleiner dan de standaardfout van de quotiëntschatter. Bij grote steekproeven is de regressieschatter nooit slechter dan de quotiëntschatter. ■

7.5 Kwaliteitsindicatoren

Kwaliteitsindicatoren voor wat betreft de regressieschatter lijken veel op die van de quotiëntschatter:

- de onzekerheidsmarges van de desbetreffende schatters;
- de omvang van de non-respons
- het aantal hulpvariabelen
- de omvang van de steekproef.

De non-respons kan de kwaliteit van de resultaten vooral aantasten wanneer (i) de omvang van de non-respons groot is en (ii) de non-respons selectief is. Soms kan de vertekening ten gevolge van selectieve non-respons voor een deel worden gecorrigeerd door gebruik te maken van hulpvariabelen, die samenhangen met zowel de responskans als met de doelvariabele zoals de in dit hoofdstuk beschreven regressieschatter.

Verder is een belangrijke aanname bij de regressieschatter, dat de omvang van de steekproef voldoende groot is zodat de vertekening van de regressieschatter en die van de schatter van zijn variantie beperkt blijven. Anders dan bij de quotiëntschatter is het bij de regressieschatter nog van belang dat het aantal hulpvariabelen in het model niet te groot mag zijn verhouding tot het aantal waarnemingen. In deze paragraaf hebben we ter wille van de eenvoud alleen het geval beschreven met één hulpvariabele, maar in werkelijkheid kunnen er afhankelijk van de omvang van de steekproef veel meer hulpvariabelen in het regressiemodel worden opgenomen.

8. Poststratificatieschatter

8.1 Korte beschrijving

Men spreekt van een Poststratitficatieschatter wanneer na het trekken van een EAZT-steekproef (zie hoofdstuk 2) de populatie alsnog in (post)strata wordt opsplitst. Ook de getrokken steekproef wordt overeenkomstig de poststrata opgedeeld. Met andere woorden, de steekproefomvang per poststratum heeft een stochastisch karakter. Dit is het essentiële verschil met de gewone stratificatie vooraf.

8.2 Toepasbaarheid

Poststratificatie wordt regelmatig gebruikt wanneer van te voren niet voor ieder element van de populatie bekend is tot welk stratum het behoort of wanneer bij het steekproefontwerp vooraf geen rekening is gehouden met de desbetreffende stratumindeling. Een dergelijke situatie kan zich voordoen wanneer men onderzoek wil doen naar de financiële situatie bij jonge adolescenten waarvan niet precies bekend is of ze studeren, werken of werkloos zijn. Hierdoor wordt het lastig vooraf een dergelijke stratificatie aan te brengen. Wanneer de steekproef eenmaal is getrokken, worden de getrokken elementen bij poststratificatie alsnog naar de verschillende strata ingedeeld. Wel dient er vooraf voor te worden gezorgd dat de kans op strata zonder steekproefwaarnemingen beperkt blijft door de strata niet te klein te kiezen. Mochten zich in de praktijk toch lege strata voordoen, dan worden alsnog verschillende strata bij elkaar gevoegd.

8.3 Uitgebreide beschrijving

8.3.1 De standaard poststratificatieschatter

In tegenstelling tot de quotiëntschatter en de regressieschatter wordt bij de poststratificatieschatter alleen gebruik gemaakt van kwalitatieve (categoriale hulpvariabelen), zoals geslacht, regio, burgerlijke staat enz. Stel dat de populatie aan de hand van een kwalitatieve hulpvariabele verdeeld is in L disjuncte groepen. Van ieder element in de populatie kunnen we achteraf (na waarneming van de waarde van de hulpvariabele) vaststellen tot welke groep (poststratum) het behoort. Het aantal elementen in de steekproef dat tot poststratum l behoort, geven we aan met n_l voor $l = 1,\cdots,L.$ Het aantal elementen in de populatie in poststratum l is gelijk aan N_l (dit aantal moet bekend zijn). Om de poststrata te onderscheiden van de standaard strata gebruiken we subscript l voor het l^{e} poststratum $(l = 1,\cdots,L)$ en subscript h voor het h^{e} standaard stratum $(h = 1,\cdots,H).$ De poststratificatieschatter voor het populatiegemiddelde $\overline{Y}$ is

$$\hat{\bar{Y}}_{post} = \sum_{l=1}^{L} \frac{N_l}{N} \bar{y}_l . \tag{8.1}$$

De poststratificatieschatter is zuiver onder de voorwaarde dat ieder poststratum ten minste één waarneming bevat. Op het eerste gezicht lijkt de schatter dezelfde als de directe schatter bij een gestratificeerde steekproef. Er is echter een conceptueel verschil. De steekproefomvang in een poststratum is stochastisch; vóór de steekproef getrokken is, is niet bekend hoeveel steekproefelementen in het poststratum zullen vallen. Bij een (vooraf) gestratificeerde steekproef wordt voordat de steekproef is getrokken precies vastgelegd hoeveel elementen per stratum getrokken moeten worden. Het stochastische karakter van de steekproefomvang per poststratum komt tot uitdrukking in de variantie van de poststratificatieschatter. Bij grote steekproeven kan de variantie goed benaderd worden door:

$$\text{var}\left(\hat{\bar{Y}}_{post}\right) = \frac{1-f}{n} \sum_{l=1}^{L} \frac{N_l}{N} S_l^2 + \frac{1-f}{n^2} \sum_{l=1}^{L} \left(1 - \frac{N_l}{N}\right) S_l^2 . \tag{8.2}$$

De eerste term in de variantieformule is gelijk aan de variantie bij stratificatie vooraf en evenredige allocatie. De tweede term reflecteert de onzekerheid met betrekking tot de steekproefaantallen in de poststrata. Deze term kan worden opgevat als de extra variantiebijdrage die ontstaat doordat de gerealiseerde allocatie afwijkt van de evenredige allocatie. Naarmate de strata *homogener* zijn ten aanzien van de doelvariabele, dus naarmate er binnen een stratum minder verschil is tussen de waarden van de doelvariabele, zal de variantie van de poststratificatieschatter kleiner zijn. Het is dus van belang de poststrata zodanig te kiezen dat de variatie van de doelvariabele zich voornamelijk manifesteert in de verschillen tussen de strata. De schatter voor de variantie wordt verkregen door in (8.2) S_l^2 te vervangen door de steekproefequivalenten s_l^2 .

8.3.2 De poststratificatieschatter bij complexe steekproefontwerpen

Bij poststratificatie wordt vaak uitgegaan van een (of meer) categoriale hulpvariabele X die de populatie in L categorieën (poststrata) verdeelt, elk ter grootte van N_l elementen. Het populatietotaal van Y in poststratum $l = 1, \cdots, L$ kan geschat worden met de Horvitz-Thompson schatter:

$$\hat{Y}_{HT,l} = \sum_{k=1}^{n_l} \frac{y_{lk}}{\pi_{lk}}$$

met n_l het aantal elementen in de steekproef dat tot poststratum l behoort en π_{lk} de insluitkans van element k in stratum *l*. Evenzo kunnen we de populatieomvang N_l van poststratum l schatten door:

$$\hat{N}_{HT,l} = \sum_{k=1}^{n_l} \frac{1}{\pi_{lk}}$$

zodat een schatter voor $\overline{Y}_l$ is:

$$\hat{\overline{Y}}_l = \frac{\hat{Y}_{HT,l}}{\hat{N}_{HT,l}} \quad .$$

De poststratificatieschatter voor het totaal is dan gelijk aan:

$$\hat{Y}_{post} = \sum_{l=1}^{L} \frac{N_l}{\hat{N}_{HT,l}} \sum_{k=1}^{n_l} \frac{y_{lk}}{\pi_{lk}} \quad .$$

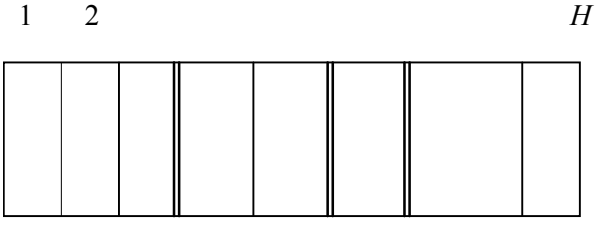
Merk op, dat we met *L* verschillende correctiegewichten te maken hebben en dat de correctiegewichten in een poststratum l gelijk zijn aan $N_l / \hat{N}_{HT,l}$. De poststratificatieschatter is bij benadering zuiver. Voor de variantie kan een goede benaderingsformule gegeven worden. Deze biedt voor specifieke steekproefontwerpen echter niet erg veel inzicht. We bespreken daarom de poststratificatieschatter alleen voor een gestratificeerde steekproef, waarbij we formules achterwege laten.

Poststratificatie bij een gestratificeerde steekproef

Stel de populatie is verdeeld in *H* strata en we hebben een gestratificeerde steekproef getrokken met in elk stratum n_h elementen. We onderscheiden vier gevallen die in de praktijk kunnen voorkomen.

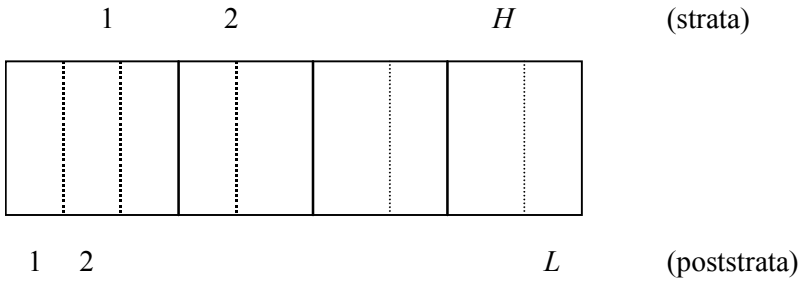
(a) De strata en de poststrata vallen samen (dus $H = L$).

(b) De poststrata zijn verenigingen van strata. Een voorbeeld is als de provincies de strata zijn en de landsdelen vormen de poststrata.



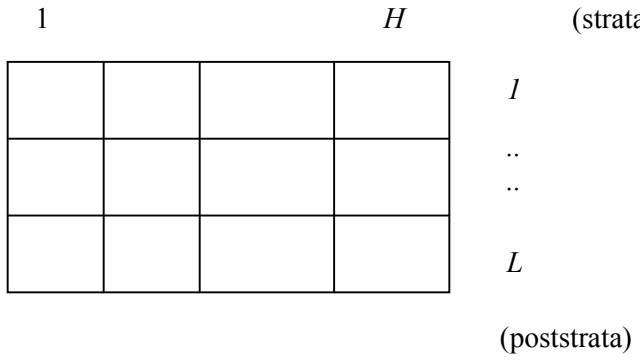
Poststrata zijn verenigingen van strata

(c) De poststrata zijn delen van strata. Een voorbeeld is als de provincies de strata zijn en de gemeenten vormen de poststrata.



Poststrata zijn delen van strata

(d) De poststrata doorkruisen de strata. Een voorbeeld is als de provincies de strata zijn en de poststrata worden bepaald door geslacht of door leeftijdsgroepen.



Poststrata doorkruisen strata

Ad (a)
We hebben hier te maken met de directe schatter (HT-schatter) bij een gestratificeerde steekproef (zie hoofdstuk 3). De bijbehorende variantie volgt ook uit Hoofdstuk 3.

Ad (b)
Poststratificatie voegt hier niets toe, omdat de stratificatie een fijnere indeling is dan de poststratificatie. Er geldt dat $\hat{N}_l = N_l$ voor alle l, dus we hebben weer de directe schatter bij een gestratificeerde steekproef.

Ad (c)
Op de enkelvoudig, aselect getrokken steekproef in een stratum h wordt poststratificatie toegepast met, zeg L_h strata. Binnen het stratum geldt dus de variantie van de poststratificatieschatter bij een enkelvoudig aselect getrokken steekproef. Omdat de steekproeven van de strata onafhankelijk van elkaar getrokken zijn, mogen deze

varianties opgeteld worden om de variantie van de poststratificatieschatter (met L poststrata) te krijgen.

Ad (d)
Binnen ieder poststratum worden aan de hand van de directe schatter bij een gestratificeerde steekproef het totaal van de doelvariabele in dat poststratum en het populatieaantal van het poststratum geschat. Daarmee worden de gemiddelden van de doelvariabele in de poststrata geschat en deze worden, gewogen met de werkelijke populatieomvang van de poststrata, opgeteld; zie Särndal et al. (1992, p. 268).

8.4 Voorbeeld

Een onderzoeker is geïnteresseerd in de financiële situatie van studenten. Hij heeft de beschikking over een register van een universiteit met 5.000 studenten en trekt hieruit een EAZT-steekproef van 500 studenten. Eén van de vragen heeft betrekking op het feit of iemand naast zijn studie een betaalde baan heeft van meer dan 10 uur. De steekproef bestaat uit 300 mannelijke en 200 vrouwelijke studenten waarvan 105 mannen en 50 vrouwen aangeven betaald werk te hebben voor meer dan 10 uur. Wanneer de onderzoeker alleen de informatie uit de steekproef gebruikt zou hij schatten, dat

$$\frac{5.000}{500} \times 155 = 1.550$$

studenten een betaalde baan van meer dan 10 uur hebben.

Het register van de universiteit bevat ook informatie met betrekking tot het geslacht van de studenten; er zijn in totaal 3.300 mannelijke en 1.700 vrouwelijke studenten aan de universiteit ingeschreven. Hieruit volgt, dat ten opzichte van de studentenpopulatie, de mannelijke studenten ondervertegenwoordigd zijn in de steekproef en de vrouwelijke studenten oververtegenwoordigd zijn in de steekproef. Omdat we vermoeden, dat het geslacht van de student samenhangt met het hebben van een betaalde baan van meer dan 10 uur, geven we de voorkeur aan een steekproef die representatief is in termen van het aantal mannelijke en vrouwelijke studenten. In de huidige situatie waar de steekproef al is getrokken, kan alleen achteraf (a posteriori) in de schatter voor de onder- en oververtegenwoordiging worden gecorrigeerd.

Een betere schatting van het aantal studenten met een betaalde baan is dan ook, dat:

$$\frac{105}{300} \times 3.300 + \frac{50}{200} \times 1.700 = 1.580 .$$ ■

8.5 Kwaliteitsindicatoren

Een belangrijke kwaliteitsindicator bij de Poststratificatieschatter is zijn onzekerheidsmarge. Zolang de strata steeds voldoende elementen bevatten, is er weinig aan de hand. Wanneer sommige strata evenwel bijna geen waarnemingen bevatten, kan de variantie van de Poststratificatieschatter oplopen. Om dit te voorkomen kan men strata bij elkaar voegen.

9. Literatuur

Camstra, A. en Nieuwenbroek, N.J. (2002), *Syllabus bij de cursus steekproeftheorie*. CBS-Academie, Voorburg.

Cochran, W.G. (1977), *Sampling Techniques*. John Wiley & Sons, New York.

Gouweleeuw, J.M. en Knottnerus, P. (2008), *Methodenreeks: thema Steekproeftheorie, deelthema Herhaald wegen*. CBS, Voorburg.

Hájek, J. (1964), Asymptotic theory of rejective sampling with varying probabilities from a finite population. *Annals of Mathematical Statistics* 35, 1491-1523.

Hansen, M.H. and Hurwitz, W.N. (1943), On the theory of sampling from finite populations. *Annals of Mathematical Statistics* 14, 333-362.

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

Knottnerus, P. (2003), *Sample Survey Theory: Some Pythagorean Perspectives*. Springer-Verlag, New York.

Knottnerus, P. (2009), *On the efficiency of randomized PPS sampling with an application to the Producer Price Index*. Discussion paper 09014, CBS, The Hague.

Muilwijk, J., Snijders, T.A.B. en Moors, J.J.A. (1992), *Kanssteekproeven*. Stenfert Kroese, Leiden.

Särndal, C.E, Swensson, B. and Wretman, J.H. (1992), *Model Assisted Survey Sampling*. Springer-Verlag, New York.

Sen, A.R. (1953), On the estimate of the variance in sampling with varying probabilities. *Journal of the Indian Society of Agricultural Statistics* 5, 119-127.

Yates, F. and Grundy, P.M. (1953), Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society* B, 15, 253-261.