



Hoofdafdeling Statistische Methoden

Postbus 959, 2270 AZ VOORBURG

NON-RESPONS EN POST-STRATIFICATIE

Jelke Bethlehem

Bert Kersten

SAMENVATTING

Indien het optreden van non-respons in enquête-onderzoek samenhangt met het te onderzoeken verschijnsel, kunnen bij het schatten van populatieparameters vertekeningen optreden. Door een dergelijke vertekening wordt de nauwkeurigheid van de schattingen in negatieve zin beïnvloed. In dit rapport wordt het non-respons probleem binnen een wiskundig model nader uitgewerkt. Duidelijk wordt in hoeverre de hierboven genoemde samenhang de vertekening beïnvloedt. Nagegaan wordt of de vertekening gereduceerd kan worden door middel van de methode van post-stratificatie. Indien de voor post-stratificatie benodigde gewichten niet bekend zijn, kunnen ze soms op grond van de steekproef worden geschat. Het blijkt dat ook met gebruikmaking van geschatte gewichten winst kan worden geboekt bij de nauwkeurigheid van de schatters.

De in dit rapport weergegeven opvattingen zijn die van de auteurs en komen niet noodzakelijkerwijs overeen met het beleid van het Centraal Bureau voor de Statistiek. De auteurs danken Albert Verbeek voor een aantal waardevolle suggesties ter verbetering.

1. Inleiding

De laatste jaren valt een duidelijke stijging van het percentage non-respons in enquête-onderzoek waar te nemen. Non-respons leidt niet alleen tot een reductie van het beschikbare aantal waarnemingen, het kan ook leiden tot ernstige vertekeningen in de schattingen voor de waarden van de populatieparameters. Dergelijke vertekeningen treden op indien er samenhang bestaat tussen de te onderzoeken variabele(n) en het wel of niet responderen. Alleen indien er gegronde redenen bestaan om te kunnen veronderstellen dat een dergelijke samenhang niet bestaat, is het verantwoord om gebruik te maken van alleen de responsgegevens bij het maken van schattingen voor de populatie. In de praktijk is de veronderstelling van een onafhankelijk optredende non-respons meer dan eens een hachelijke zaak. Vaak valt het bestaan van een dergelijk verband aannemelijk te maken. Moeten toch uitspraken over de populatie gedaan worden, dan is dit alleen mogelijk indien eerst een correctieprocedure wordt uitgevoerd.

In dit rapport wordt onderzocht in hoeverre de methode van post-stratificatie daarvoor een geschikte methode is. Grondslag voor de in dit rapport ontwikkelde theorie wordt gevormd door het Stochastische Respons Model, een wiskundig model dat het optreden van non-respons beschrijft. Eerst wordt binnen dit model in paragraaf 2 het non-respons probleem nader uitgewerkt. Daarna wordt dit in paragraaf 3 toegepast voor een enkelvoudige aselechte steekproef. In paragraaf 4 wordt dan bekeken hoe post-stratificatie de situatie in paragraaf 3 kan verbeteren. In paragraaf 5 tenslotte wordt nagegaan wat er nog te doen valt indien de voor post-stratificatie benodigde gewichten niet bekend zijn.

In iedere paragraaf wordt de daar ontwikkelde theorie geïllustreerd aan de hand van een speciaal geval. Door er vanuit te gaan dat de te onderzoeken variabele dichotoom is, komen vanwege de eenvoudiger formules allerlei aspecten van de theorie duidelijker naar voren.

Teneinde de leesbaarheid van dit rapport te bevorderen, zijn de bewijzen van de beweringen achterwege gelaten. Voorts is van de veronderstelling uitgegaan dat $1 \ll n \ll N$. Hierdoor vervalt de eindigheidscorrectie en het verschil tussen trekken met en zonder teruglegging.

Voor de normale steekproeftheorie (zonder optreden van non-respons) is

gebruik gemaakt van Cochran (1977). Meer over post-stratificatie is te vinden in Holt & Smith (1979).

2. Het non-respons probleem

2.1. Notatie

Laat de eindige populatie die onderzocht moet worden uit N elementen bestaan. Deze populatie kan worden gesplitst in H strata $1, 2, \dots, H$. Stratum h bevat N_h elementen, zodat

$$\sum_{h=1}^H N_h = N. \quad (1)$$

Het k^{de} element in stratum h wordt aangegeven met U_{hk} en de daarbij behorende waarde van de te onderzoeken variabele met Y_{hk} ($h = 1, 2, \dots, H$; $k = 1, 2, \dots, N_h$).

De populatieparameter die geschat moet worden is het gemiddelde van de te onderzoeken variabele:

$$\bar{Y} = \frac{1}{N} \sum_{h=1}^H \sum_{k=1}^{N_h} Y_{hk} = \sum_{h=1}^H W_h \bar{Y}_h, \quad (2)$$

waarin het gewicht W_h gelijk is aan

$$W_h = \frac{N_h}{N} \quad (3)$$

en $\bar{Y}_h$ het gemiddelde is in stratum h :

$$\bar{Y}_h = \frac{1}{N_h} \sum_{k=1}^{N_h} Y_{hk}. \quad (4)$$

De theorie wordt ontwikkeld uitgaande van de veronderstelling dat de non-respons gegenereerd wordt volgens het Stochastische Respons Model. Dat houdt in dat iedere element U_{hk} in de populatie een zekere (onbekende) kans P_{hk} heeft om te responderen. Alle P_{hk} 's hoeven niet noodzakelijk dezelfde waarde te hebben. Het Stochastische Respons Model wordt wat minder vaak toegepast dan het Vaste Respons Model. Daarbij wordt er vanuit gegaan

dat de populatie gesplitst kan worden in twee strata: een responsstratum en een non-responsstratum. Trekt men een steekproef uit deze populatie dan zullen sommige elementen uit het responsstratum afkomstig zijn, en andere uit het non-responsstratum.

Stel $\underline{u}$ is een aselekt getrokken element uit de populatie (stochastische variabelen worden onderstreept). De daarbij behorende waarde van de te onderzoeken variabele wordt aangegeven met $\underline{y}$. De responsvariabele $\underline{r}$ krijgt de waarde 0 in geval van non-respons, en de waarde 1 in geval van respons. Uit de definitie van de responskans volgt:

$$P(\underline{r} = 1 \mid \underline{u} = U_{hk}) = P_{hk}. \quad (5)$$

Hieruit kan worden afgeleid dat

$$P(\underline{u} = U_{hk} \text{ en } \underline{r} = 1) = P_{hk}/N \quad (6)$$

en

$$P(\underline{u} = U_{hk} \mid \underline{r} = 1) = P_{hk}/N\bar{P}. \quad (7)$$

De gemiddelde responskans in de populatie is gelijk aan

$$\bar{P} = \frac{1}{N} \sum_{h=1}^H \sum_{k=1}^{N_h} P_{hk} = \sum_{h=1}^H W_h \bar{P}_h = P(\underline{r} = 1) \quad (8)$$

waarin $\bar{P}_h$ de gemiddelde responskans in stratum h is:

$$\bar{P}_h = \frac{1}{N_h} \sum_{k=1}^{N_h} P_{hk}. \quad (9)$$

De verwachting van $\underline{y}$ is gelijk aan

$$E\underline{y} = \bar{y} \quad (10)$$

en de variantie van $\underline{y}$ is gelijk aan

$$V(\underline{y}) = \frac{1}{N} \sum_{h=1}^H \sum_{k=1}^{N_h} (y_{hk} - \bar{y})^2. \quad (11)$$

De verwachting van $\underline{r}$ is gelijk aan

$$E \underline{r} = \bar{P} \quad (12)$$

en de variantie van $\underline{r}$ is gelijk aan

$$V(\underline{r}) = \bar{P}(1-\bar{P}) \quad (13)$$

Een andere grootheid, die gebruikt zal worden, is de covariantie tussen $\underline{y}$ en $\underline{r}$. Voor willekeurige stochastische variabelen $\underline{a}$ en $\underline{b}$ geldt:

$$\text{Cov}(\underline{a}, \underline{b}) = E(\underline{a}(\underline{b} - E\underline{b})). \quad (14)$$

Uitwerken van (14) voor $\underline{a} = \underline{y}$ en $\underline{b} = \underline{r}$ leidt tot

$$\text{Cov}(\underline{y}, \underline{r}) = \frac{1}{N} \sum_{h=1}^H \sum_{k=1}^{N_h} y_{hk} (P_{hk} - \bar{P}). \quad (15)$$

2.2. Non-respons

Als in een enquête-onderzoek non-respons optreedt, zijn alleen metingen beschikbaar voor de responderende elementen. Daarom is het zinvol om de verdeling van $\underline{y}$ te bekijken, conditioneel op het responderen van het betreffende element. Definieer daarom $\underline{z}$ als $\underline{z} = (\underline{y} | \underline{r} = 1)$. De verwachting van $\underline{z}$ is gelijk aan

$$E \underline{z} = \frac{1}{NP} \sum_{h=1}^H \sum_{k=1}^{N_h} y_{hk} P_{hk}, \quad (16)$$

en de variantie van $\underline{z}$ is gelijk aan

$$V(\underline{z}) = \frac{1}{NP} \sum_{h=1}^H \sum_{k=1}^{N_h} y_{hk}^2 P_{hk} - \left(\frac{1}{NP} \sum_{h=1}^H \sum_{k=1}^{N_h} y_{hk} P_{hk} \right)^2. \quad (17)$$

De vertekening B van een schatter wordt gedefinieerd als het verschil tussen de verwachting van die schatter en de waarde van de populatieparameter die hij beoogt te schatten. De vertekening van $\underline{z}$ is gelijk aan

$$B(\underline{z}) = E \underline{z} - \bar{Y} \quad (18)$$

Voor een willekeurige stochastische variabele $\underline{a}$ en een Bernoulli-variabele $\underline{b}$ geldt dat

$$\text{Cov}(\underline{a}, \underline{b}) = (E(\underline{a} \mid \underline{b} = 1) - E\underline{a})E\underline{b}. \quad (19)$$

Door invullen van $\underline{y}$ voor $\underline{a}$ en $\underline{r}$ voor $\underline{b}$ volgt hieruit dat

$$B(\underline{z}) = \frac{\text{Cov}(\underline{y}, \underline{r})}{\bar{p}} \quad (20)$$

Het is duidelijk dat de vertekening van $\underline{z}$ bepaald wordt door twee factoren:

- (1) De samenhang tussen $\underline{y}$ en $\underline{r}$: Hoe sterker de samenhang, hoe groter de vertekening.
- (2) De gemiddelde responskans: Gegeven een bepaalde samenhang tussen $\underline{y}$ en $\underline{r}$, leidt een lagere gemiddelde responskans tot een grotere vertekening.

Alleen indien er voldoende redenen zijn om aan te nemen dat er geen verband bestaat tussen de te onderzoeken variabele en de responsvariabele, kunnen de gegevens van de responderende elementen gebruikt worden om populatiewaarden te schatten.

De invloed van de vertekening op de variantie is minder duidelijk. Toepassen van relatie (19) met $\underline{a} = (\underline{y} - E\underline{y})^2$ en $\underline{b} = \underline{r}$ geeft

$$\text{Cov}((\underline{y} - E\underline{y})^2, \underline{r}) = (V(\underline{z}) - V(\underline{y}))E\underline{r} \quad (21)$$

zodat

$$V(\underline{z}) = V(\underline{y}) + \text{Cov}((\underline{y} - E\underline{y})^2, \underline{r})/E\underline{r}. \quad (22)$$

Het zal van de waarden der P_{hk} 's afhangen, en van de samenhang tussen P_{hk} en $(Y_{hk} - \bar{y})^2$, of $V(\underline{z})$ groter of kleiner is dan $V(\underline{y})$. Indien bijvoorbeeld kleine verschillen tussen Y_{hk} en $\bar{y}$ overeenkomen met grote responskansen, en grote verschillen met kleine responskansen, dan zal $V(\underline{z})$ kleiner zijn dan $V(\underline{y})$. Het is dus in principe mogelijk dat vertekening tot kleinere varianties leidt. In deze situatie is het gebruik van de variantie voor het aangeven van de nauwkeurigheid van de schatter nogal misleidend. Een betere indicatie voor de nauwkeurigheid van een mogelijk vertekende schatter wordt verkregen met behulp van de Gemiddelde Kwadratische Fout (GKF). De GKF is gedefinieerd als het verwachte kwadratische verschil tussen de

schatter en de waarde van de populatieparameter die hij beoogt te schatten. In de onderhavige situatie is de GKF van $\underline{z}$ dus gelijk aan

$$GKF(\underline{z}) = E(\underline{z} - \bar{Y})^2 \quad (23)$$

Uitschrijven van (23) leidt tot de bekende uitdrukking

$$GKF(\underline{z}) = V(\underline{z}) + B^2(\underline{z}). \quad (24)$$

Uitschrijven van (24) leidt tot

$$GKF(\underline{z}) = \frac{1}{N} \sum_{h=1}^H \sum_{k=1}^{N_h} (y_{hk} - \bar{Y})^2 \frac{P_{hk}}{\bar{P}} \quad (25)$$

Indien de factor $P_{hk}/\bar{P}$ in (25) gelijk genomen wordt aan 1, dan reduceert (25) tot $V(\underline{y})$. Dit is natuurlijk een gevolg van het feit dat de vertekening verdwenen is. In het algemeen zal het van de samenhang tussen de responskansen P_{hk} en de gekwadrateerde afwijking $(y_{hk} - \bar{Y})^2$ afhangen of de GKF van $\underline{z}$ groter of kleiner is dan de variantie van $\underline{y}$.

2.3. Speciaal geval

In het kader van een speciaal geval wordt de situatie wat nader bekeken. Veronderstel dat de te onderzoeken variabele een dichotome variabele is. Stel dat de twee waarden 0 en 1 even vaak voorkomen in de populatie. Voor een aselekt getrokken element $\underline{y}$ geldt dus

$$P(\underline{y} = 0) = P(\underline{y} = 1) = 0,5. \quad (26)$$

De te schatten parameter, het gemiddelde van de te onderzoeken variabele, is dus gelijk aan $\bar{Y} = 0,5$. Stel nu dat de samenhang tussen $\underline{y}$ en de responsvariabele $\underline{r}$ wordt vastgelegd door een parameter P op zodanige wijze dat

$$P(\underline{r} = 1 \mid \underline{y} = 0) = P \quad (27)$$

en

$$P(\underline{r} = 1 \mid \underline{y} = 1) = 1 - P \quad (28)$$

waarbij $0 \leq P \leq 0,5$. De waarde van de te onderzoeken variabele legt dus de waarde van de responskansen vast. Toepassen van (15) geeft

$$\text{Cov}(\underline{y}, \underline{r}) = 0,25 - 0,5P. \quad (29)$$

Voor $P=0$ is de samenhang het sterkst. Naarmate de waarde van P toeneemt, vermindert de covariantie. Voor $P=0,5$ is er geen enkele samenhang.

Definieer $\underline{z} = (\underline{y} \mid \underline{x}=1)$. Dan is $\underline{z}$ een Bernoulli variabele met $P(\underline{z}=0) = P$ en $P(\underline{z}=1) = 1 - P$. Hieruit volgt dat

$$E\underline{z} = 1 - P \quad (30)$$

en

$$V(\underline{z}) = P(1-P) \quad (31)$$

De vertekening van $\underline{z}$, berekend met behulp van (18) of (20), is gelijk aan

$$B(\underline{z}) = 0,5 - P \quad (32)$$

Het is duidelijk dat de vertekening van $\underline{z}$ toeneemt naarmate de relatie tussen $\underline{y}$ en $\underline{x}$ sterker wordt. De GKF van $\underline{z}$ is gelijk aan

$$\text{GKF}(\underline{z}) = 0,25. \quad (33)$$

In deze situatie is de GKF onafhankelijk van de sterkte van de samenhang tussen $\underline{y}$ en $\underline{x}$. Kennelijk wordt een grotere vertekening gecompenseerd door een kleinere variantie.

3. Enkelvoudige aselechte steekproef

Uit de populatie, zoals die beschreven is in de vorige paragraaf, wordt een enkelvoudige aselechte steekproef van omvang n getrokken. Als gevolg van de veronderstelling $1 \ll n \ll N$ kan het verschil tussen trekken met en zonder teruglegging verwaarloosd worden. De n getrokken elementen worden aangegeven met $\underline{u}_1, \underline{u}_2, \dots, \underline{u}_n$. De n bijbehorende waarden van de te onderzoeken variabele zijn gelijk aan $\underline{y}_1, \underline{y}_2, \dots, \underline{y}_n$ en de waarden van de responsvariabele worden genoteerd met $\underline{x}_1, \underline{x}_2, \dots, \underline{x}_n$.

3.1. Geen non-respons

Indien geen non-respons optreedt, kunnen de gebruikelijke schattings-technieken toegepast worden. De voor de hand liggende schatter voor $\bar{y}$ is

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n \underline{y}_i. \quad (34)$$

Dit is een zuivere schatter voor $\bar{Y}$ met variantie

$$V(\bar{y}) = \frac{V(y)}{n} . \quad (35)$$

3.2. Non-respons

Indien er non-respons optreedt, zijn alleen die waarden y_i bekend waarvoor $r_i = 1$. Het aantal responderende elementen is gelijk aan

$$m = \sum_{i=1}^n r_i \quad (36)$$

Onder de in paragraaf 1 gestelde voorwaarde ($1 \ll n \ll N$) kunnen de r_i 's als stochastisch onafhankelijk opgevat worden, zodat m een binomiale verdeling heeft met parameters n en $\bar{P}$. De waarde van het j^{de} responderende element wordt aangegeven met z_j ($j = 1, 2, \dots, m$). Aangetoond kan worden dat iedere z_j dezelfde verdeling heeft als z , zodat met behulp van (18) volgt dat

$$Ez_j = \bar{Y} + B(z) \quad (37)$$

en

$$V(z_j) = V(z) . \quad (38)$$

De in aanmerking komende schatter in deze situatie is

$$\bar{z} = \frac{1}{m} \sum_{j=1}^m z_j . \quad (39)$$

De schatter (39) is gedefiniëerd als het gemiddelde van een stochastisch aantal waarnemingen. In principe is het dus mogelijk dat de schatter niet bestaat, omdat er geen enkele beschikbare waarneming is. Dit gebeurt wanneer geen enkel element respondeert. Om deze situatie te voorkomen wordt de theorie ontwikkeld conditioneel op de situatie dat $m > 0$. Dit is nauwelijks een beperking, aangezien in de praktijk de kans $P(m=0)$ voor niet al te kleine n erg klein is. Ter illustratie: voor $n = 100$ en $\bar{P} = 0,75$ is de kans ongeveer 10^{-60} .

Als gevolg van het stochastische karakter van het aantal waarnemingen verloopt de berekening van verwachting en variantie van $\bar{z}$ niet op de gebrui-

kelijke wijze. Definieer $\phi(\underline{m})$ door

$$\phi(\underline{m}) = E(\bar{z} \mid \underline{m} = \underline{m}). \quad (40)$$

Uitschrijven van (40) leidt tot

$$\phi(\underline{m}) = \frac{1}{m} \sum_{j=1}^m E z_j = \bar{Y} + B(\underline{z}). \quad (41)$$

De verwachting van $\bar{z}$ is nu gelijk aan

$$E\bar{z} = E\phi(\underline{m}) = E(\bar{Y} + B(\underline{z})) = \bar{Y} + B(\underline{z}). \quad (42)$$

Definieer $\Psi(\underline{m})$ door

$$\Psi(\underline{m}) = V(\bar{z} \mid \underline{m} = \underline{m}). \quad (43)$$

Uitschrijven van (43) levert

$$\Psi(\underline{m}) = \frac{V(\underline{z})}{m}. \quad (44)$$

De variantie van $\bar{z}$ is nu gelijk aan

$$V(\bar{z}) = V(\phi(\underline{m})) + E\Psi(\underline{m}), \quad (45)$$

waarin $\phi(\underline{m})$ als gedefinieerd in (40). Aangezien $\phi(\underline{m})$ niet van m afhangt, geldt $V(\phi(\underline{m})) = 0$. Dus $V(\bar{z})$ kan bepaald worden uit

$$V(\bar{z}) = E \frac{V(\underline{z})}{m} = V(\underline{z}) E \frac{1}{m} \quad (46)$$

Aangezien $E(1/m)$ voor grote n benaderd kan worden door (zie Stephan (1945))

$$E \frac{1}{m} \approx \frac{1}{E m} = \frac{1}{n\bar{P}} \quad (47)$$

kan de variantie van $\bar{z}$ voor grote n benaderd worden door

$$V(\bar{z}) \approx \frac{V(\underline{z})}{n\bar{P}}. \quad (48)$$

De GKF van $\bar{z}$ is gelijk aan

$$\text{GKF}(\bar{z}) = \frac{V(z)}{nP} + B^2(z). \quad (49)$$

Nadere beschouwing van (49) maakt een belangrijk aspect van onzuivere schatters duidelijk. Naarmate de steekproefomvang groter wordt, neemt de GKF af. Maar aangezien de vertekeningsterm in (49) niet afhangt van n , zal de GKF nooit kleiner kunnen worden dan $B^2(z)$. In grote steekproeven hangt de nauwkeurigheid dus in hoge mate af van de grootte van de vertekening. Om nauwkeuriger schatters te verkrijgen, zal de vertekening gereduceerd moeten worden.

3.3. Speciaal geval

Om meer inzicht te verkrijgen in het gedrag van $\bar{z}$ wordt de analyse van het speciale geval in paragraaf 2.3 hier voortgezet. Uit de populatie zoals beschreven in (26) en (27) wordt een steekproef van omvang n getrokken (waarbij weer wordt aangenomen dat $1 \ll n \ll N$). Gebruikmakend van (30) en (31) wordt de verwachting van de schatter $\bar{z}$, zoals gedefinieerd in (39):

$$E\bar{z} = 1 - p \quad (50)$$

en de variantie van $\bar{z}$ is gelijk aan

$$V(\bar{z}) = \frac{V(z)}{nP} = \frac{2p(1-p)}{n}. \quad (51)$$

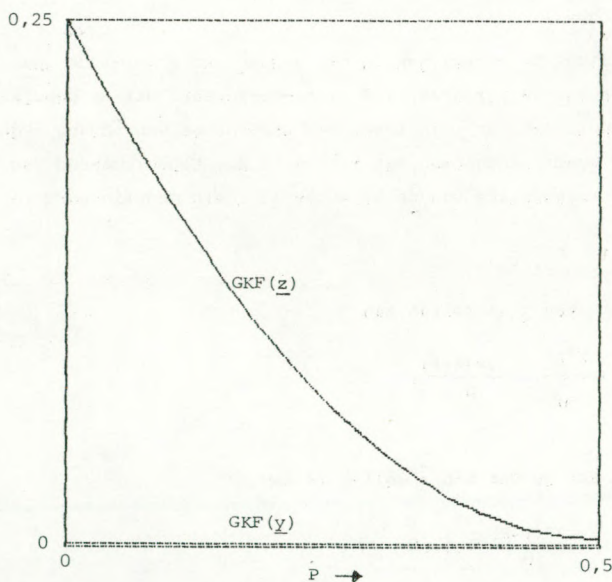
Hieruit volgt dat de GKF van $\bar{z}$ gelijk is aan

$$\text{GKF}(\bar{z}) = \frac{2p(1-p)}{n} + (0,5 - p)^2. \quad (52)$$

Voor vaste n is de GKF van $\bar{z}$ minimaal voor $p = 0,5$ (geen vertekening). De GKF is dan gelijk aan $1/4n$. De GKF bereikt zijn maximale waarde voor $p = 0$ (grootste vertekening). De waarde is dan $0,25$. De GKF van $\bar{y}$ is gelijk aan $V(\bar{y})$. De waarde is altijd gelijk aan $1/4n$. In de onderhavige situatie zal non-respons dus altijd leiden tot minder nauwkeurige schatters. Voor waarden van p in de buurt van $0,5$ is dit voornamelijk het gevolg van een geringer aantal beschikbare waarnemingen. Voor waarden van p in de buurt van

0 is de vertekening de voornaamste oorzaak. In het onderhavige geval is de verwachting van het beschikbare aantal waarnemingen gelijk aan $n/2$. Het effect van het geringere aantal waarnemingen zou dus ongedaan kunnen worden gemaakt door de steekproefomvang te verdubbelen. Toch zal, zelfs met een verdubbelde steekproefomvang, de GKF van $\bar{z}$ altijd minstens zo groot zijn als de GKF van $\bar{y}$ (zonder verdubbelde steekproefomvang). In figuur 1 is bij wijze van illustratie de grafiek gegeven van de GKF van $\bar{y}$ en $\bar{z}$ bij $n = 100$.

Figuur 1. De gemiddelde kwadratische fout van $\bar{y}$ en $\bar{z}$ in een populatie met $P(\bar{y} = 0) = P(\bar{y} = 1) = 0,5$. De responskans is P voor $\bar{y} = 0$ en $1 - P$ voor $\bar{y} = 1$. De steekproefomvang is 100.



4. Post-stratificatie

Bij post-stratificatie wordt eerst uit de populatie een enkelvoudige aselechte steekproef getrokken. Daarna worden de getrokken elementen ingedeeld in strata. Indien de strata zodanig geconstrueerd worden dat de te onderzoeken variabele zo weinig mogelijk varieert binnen de strata, kan een aanzienlijke winst in nauwkeurigheid worden behaald. Om post-stratificatie te kunnen toepassen moeten wel de omvangen $N_1, N_2, \dots, N_H$ van de strata bekend zijn. In deze paragraaf wordt ingegaan op de vraag in hoeverre post-stratificatie de GKF van de schatter kan reduceren.

Laat $u_1, u_2, \dots, u_n$ de enkelvoudige aselechte steekproef zijn, $y_1, y_2, \dots, y_n$ de bijbehorende waarden van de te onderzoeken variabele, en $r_1, r_2, \dots, r_n$ de waarden van de responsvariabele.

Stel n_h is het aantal steekproefelementen dat in stratum h terecht komt. Het i^{de} steekproefelement in stratum h wordt aangegeven met u_{hi} , de bijbehorende waarde met y_{hi} en de waarde van de responsvariabele met r_{hi} ($h = 1, 2, \dots, H; i = 1, 2, \dots, n_h$).

4.1. Geen non-respons

Indien geen non-respons optreedt, is de bij post-stratificatie gebruikte schatter gelijk aan

$$\bar{y}_{PS} = \sum_{h=1}^H W_h \bar{y}_h. \quad (53)$$

Het gewicht W_h is gedefinieerd in (3) en $\bar{y}_h$ is het gemiddelde van de n_h waarden in stratum h :

$$\bar{y}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} y_{hi}. \quad (54)$$

De schatter (54) is een zuivere schatter voor $\bar{y}_h$. Conditioneel onder de voorwaarde $n_h > 0$ kan, voor ieder stratum h gebruikmakend van (35), de variantie van $\bar{y}_{PS}$ benaderd worden met

$$V(\bar{y}_{PS}) = \frac{1}{n} \sum_{h=1}^H W_h V_h(\bar{y}), \quad (55)$$

waarin $V_h(\underline{y})$ de variantie van $\underline{y}$ binnen stratum h .

4.2. Non-respons

Indien er non-respons optreedt, laat dan m_h het aantal responderende elementen in stratum h zijn. De waarde van de te onderzoeken variabele voor het j^{de} responderende element in stratum h wordt genoteerd met z_{hj} ($h = 1, 2, \dots, H$; $j = 1, 2, \dots, m_h$). Onder de voorwaarde $1 \ll n \ll N$ is m_h binomiaal verdeeld met parameters n en $W_h \bar{p}_h$. In geval van post-stratificatie is de voor de hand liggende schatter voor $\bar{Y}$:

$$\bar{z}_{PS} = \sum_{h=1}^H W_h \bar{z}_h, \quad (56)$$

waarin $\bar{z}_h$ het gemiddelde van de waarden van de responderende elementen in stratum h :

$$\bar{z}_h = \frac{1}{m_h} \sum_{j=1}^{m_h} z_{hj}. \quad (57)$$

De schatter $\bar{z}_h$ bestaat onder de conditie dat $m_h > 0$ (wat impliceert dat $n_h > 0$). De schatter $\bar{z}_{PS}$ is onzuiver met betrekking tot $\bar{Y}$. Toepassing van (18) en (20) voor elk stratum h geeft

$$E\bar{z}_h = \bar{Y}_h + B_h(\underline{z}) \quad (58)$$

met

$$B_h(\underline{z}) = \text{Cov}_h(\underline{y}, \underline{r}) / \bar{p}_h, \quad (59)$$

waarin $\text{Cov}_h(\underline{y}, \underline{r})$ de covariantie tussen $\underline{y}$ en $\underline{r}$ in stratum h . Invullen in (56) en toepassen van (2) geeft dan

$$E\bar{z}_{PS} = \bar{Y} + B(\bar{z}_{PS}), \quad (60)$$

waarin

$$B(\bar{z}_{PS}) = \sum_{h=1}^H W_h B_h(\underline{z}). \quad (61)$$

De variantie van $\bar{z}_{PS}$ kan onder de al eerder genoemde voorwaarden, door toepassing van (48) voor elk stratum, benaderd worden met

$$V(\bar{z}_{PS}) \approx \frac{1}{n} \sum_{h=1}^H w_h \frac{V_h(\bar{z})}{\bar{p}_h} \quad (62)$$

waarin $V_h(\bar{z})$ de variantie van $\bar{z}$ binnen stratum h .

4.3. Vergelijking van enkelvoudige steekproef en post-stratificatie

Wordt geen post-stratificatie toegepast, dan is de vertekening van de schatter $\bar{z}$ gelijk aan

$$B(\bar{z}) = \frac{\text{Cov}(\bar{y}, \bar{r})}{\bar{p}} \quad (63)$$

Wordt wel post-stratificatie toegepast, dan is de vertekening van de schatter $\bar{z}_{PS}$ gelijk aan

$$B(\bar{z}_{PS}) = \sum_{h=1}^H w_h \frac{\text{Cov}_h(\bar{y}, \bar{r})}{\bar{p}_h} \quad (64)$$

Vergelijking van (63) en (64) laat zien dat, indien de strata zodanig geconstrueerd worden dat de te onderzoeken variabele zo weinig mogelijk varieert binnen ieder stratum, de covariantie in ieder stratum klein wordt en daarmee de totale vertekening. Uitschrijven van (63) en (64) levert

$$B(\bar{z}) = \frac{1}{N} \sum_{h=1}^H \sum_{k=1}^{N_h} y_{hk} \left(\frac{p_{hk}}{\bar{p}} - 1 \right) \quad (65)$$

en

$$B(\bar{z}_{PS}) = \frac{1}{N} \sum_{h=1}^H \sum_{k=1}^{N_h} y_{hk} \left(\frac{p_{hk}}{\bar{p}_h} - 1 \right). \quad (66)$$

De relaties (65) en (66) verschillen alleen in de termen $p_{hk}/\bar{p}$ en $p_{hk}/\bar{p}_h$. Vergelijking laat zien dat de vertekening gereduceerd kan worden door de strata zodanig te construeren dat binnen elk stratum de responskansen zo weinig mogelijk variëren.

Er zijn dus twee manieren om de vertekening te reduceren. De eerste manier, het maken van strata die homogeen zijn met betrekking tot $\underline{y}$, is niet uit te voeren op grond van het materiaal van de respondenten, aangezien deze gegevens niet voldoende zijn om een mogelijke samenhang tussen $\underline{y}$ en $\underline{r}$ op te sporen. Dit is bij de tweede manier, het maken van strata die homogeen zijn met betrekking tot de responskansen, wel mogelijk. Daarom verdient de tweede manier de voorkeur.

Teneinde de varianties van $\bar{z}$ en $\bar{z}_{PS}$ te kunnen vergelijken, wordt die van $\bar{z}$ herschreven in termen van stratum varianties:

$$V(\bar{z}) = \frac{1}{n} \sum_{h=1}^H w_h V_h(z) \frac{\bar{p}_h}{\bar{p}^2} + \frac{1}{n} \sum_{h=1}^H (\bar{z}_h - \bar{z})^2 w_h \frac{\bar{p}_h}{\bar{p}^2}. \quad (67)$$

Hierin is $\bar{z}_h$ de verwachting van $\underline{z}$ binnen stratum h en $\bar{z}$ de verwachting van $\underline{z}$.

Vergelijking van (67) en (62) toont aan dat de variantie van $\bar{z}_{PS}$ kleiner is dan de variantie van $\bar{z}$ indien

$$\sum_{h=1}^H w_h V_h(z) \left(\frac{\bar{p}_h}{\bar{p}} - \frac{\bar{p}_h}{\bar{p}} \right) \leq \sum_{h=1}^H (\bar{z}_h - \bar{z})^2 w_h \frac{\bar{p}_h}{\bar{p}}. \quad (68)$$

Dit geschiedt met name in de volgende drie situaties:

- (1) De varianties $V_h(z)$ binnen de strata met lager dan gemiddelde responskansen zijn erg klein.
- (2) De gemiddelde responskans is in alle strata ongeveer hetzelfde.
- (3) De verwachtingen van $\underline{z}$ binnen de strata verschillen onderling sterk.

De GKF van $\bar{z}_{PS}$ is gelijk aan

$$GKF(\bar{z}_{PS}) = V(\bar{z}_{PS}) + B^2(\bar{z}_{PS}). \quad (69)$$

Aangezien de variantie-component in (69) afneemt met een toenemende steekproefomvang, zal voor een grote steekproef de vertekeningcomponent in hoge mate de nauwkeurigheid bepalen. Het is dus van groot belang om met name de vertekening te verminderen. Indien de strata zo geconstrueerd worden, dat ze homogeen zijn met betrekking tot de responskansen, wordt de vertekening gereduceerd. Een waarschuwing is hier echter op zijn plaats. Het is

in principe mogelijk dat aldus enkele strata geconstrueerd worden waarin alle elementen dezelfde lage responskans hebben. Daardoor komen in dit stratum maar erg weinig waarnemingen beschikbaar. Dit werkt weer variantieverhogend (zie relatie (62)).

4.4. Speciaal geval

De situatie wordt wat nader bekeken in het al eerder behandelde speciale geval. Bij het overwegen van een stratificatie voor een dichotome onderzoeksvariabele lijkt het zinvol om te zoeken naar een indeling in twee strata van gelijke omvang, een stratum dat zoveel mogelijk elementen met waarde 0 bevat, een stratum dat zoveel mogelijk elementen met waarde 1 bevat. Stel dat het mogelijk is de strata zodanig te construeren dat in het eerste stratum

$$P(\underline{y}=0) = Q \text{ en } P(\underline{y}=1) = 1-Q \quad (70)$$

en in het tweede stratum

$$P(\underline{y}=0) = 1-Q \text{ en } P(\underline{y}=1) = Q, \quad (71)$$

waarbij $0 \leq Q \leq 0,5$. De waarde van Q is in feite een indicatie voor de doeltreffendheid van de stratificatie. De beste stratificatie wordt verkregen voor $Q=0$. Dan bevat het eerste stratum alle elementen met waarde 1 en het tweede stratum alle elementen met waarde 0. Voor $Q=0,5$ is de stratificatie het slechtst. Dan zitten in beide strata evenveel elementen met de waarden 0 en 1.

In het eerste stratum zijn de verwachting en variantie van $\underline{z}$ gelijk aan

$$E_{\underline{z}} = \frac{(1-P)(1-Q)}{PQ + (1-P)(1-Q)} \quad (72)$$

en

$$V(\underline{z}) \approx \frac{PQ(1-P)(1-Q)}{(PQ + (1-P)(1-Q))^2} \cdot \quad (73)$$

In het tweede stratum zijn deze grootheden gelijk aan

$$E_{\underline{z}} = \frac{(1-P)Q}{(1-P)Q + P(1-Q)} \quad (74)$$

en

$$V(\underline{z}) \approx \frac{PQ(1-P)(1-Q)}{((1-P)Q + P(1-Q))^2} \cdot \quad (75)$$

Uit de hierboven beschreven populatie wordt een enkelvoudige aselechte steekproef van omvang n getrokken. Dan heeft de schatter $\bar{z}_{PS}$ hier de vorm

$$\bar{z}_{PS} = \frac{\bar{z}_1 + \bar{z}_2}{2}, \quad (76)$$

waarin $\bar{z}_1$ en $\bar{z}_2$ schatters voor de stratumgemiddelden. De verwachting van $\bar{z}_{PS}$ is gelijk aan

$$E\bar{z}_{PS} = \frac{(1-P)(1-Q)/2}{PQ + (1-P)(1-Q)} + \frac{(1-P)Q/2}{(1-P)Q + P(1-Q)}. \quad (77)$$

De variantie van $\bar{z}_{PS}$ kan worden benaderd door

$$V(\bar{z}_{PS}) \approx \frac{1}{2n} \left(\frac{PQ(1-P)(1-Q)}{(PQ + (1-P)(1-Q))^3} + \frac{PQ(1-P)(1-Q)}{((1-P)Q + P(1-Q))^3} \right). \quad (78)$$

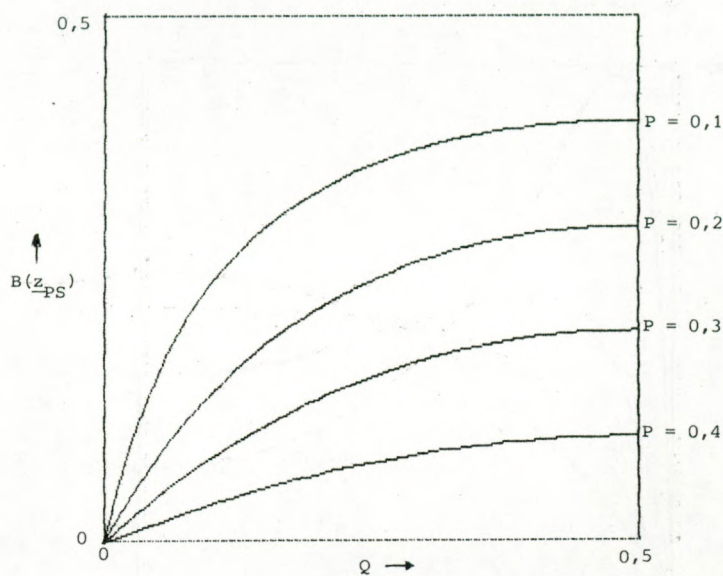
De vertekening van $\bar{z}_{PS}$ wordt uit (77) verkregen door aftrekken van 0,5:

$$B(\bar{z}_{PS}) = E\bar{z}_{PS} - 0,5. \quad (79)$$

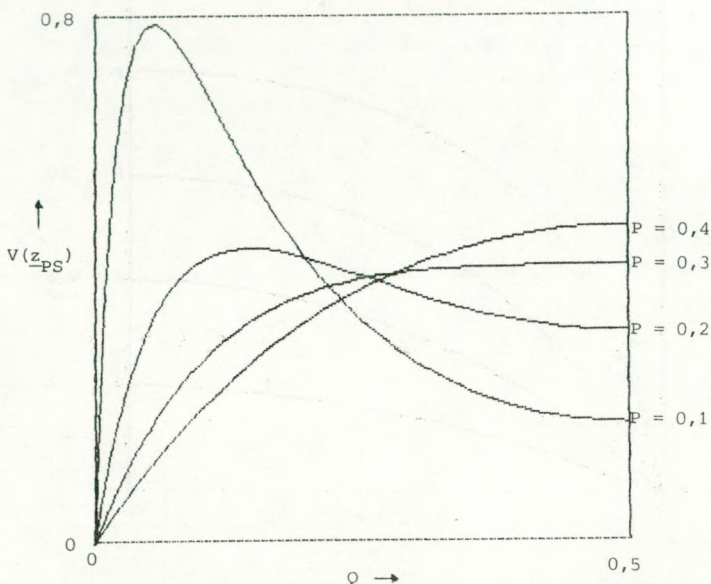
In figuur 2 is voor een aantal waarden van P de vertekening $B(\bar{z}_{PS})$ getekend als functie van Q . Voor $Q=0,5$ is $B(\bar{z}_{PS})$ gelijk aan $B(\bar{z})$, de vertekening zonder stratificatie. Het is duidelijk dat een sterke samenhang tussen $\bar{y}$ en $\bar{x}$ (kleine waarden van P) een grote vertekening kan veroorzaken. In dit speciale geval zal post-stratificatie altijd tot vermindering van de vertekening leiden.

In figuur 3 is voor een aantal waarden van P $mV(\bar{z}_{PS})$ getekend als functie van Q . Voor $Q=0,5$ is $V(\bar{z}_{PS})$ gelijk aan $V(\bar{z})$, de variantie zonder post-stratificatie. Indien er non-respons optreedt zal post-stratificatie niet altijd leiden tot kleinere varianties. Zoals al eerder uitgelegd, kan een sterke samenhang tussen $\bar{y}$ en $\bar{x}$ juist ook tot grotere varianties leiden. Indien de samenhang wat minder extreem sterk is, zal het effect van de homogeniteit de overhand krijgen. Voor $P=0,3$ en $P=0,4$ is de variantie bij post-stratificatie altijd kleiner dan de variantie zonder post-stratificatie.

Figuur 2. De vertekening van $\bar{z}_{PS}$ in een populatie met $p(\underline{y}=0) = p(\underline{y}=1) = 0,5$.
De responskans is P voor $\underline{y}=0$ en $1-P$ voor $\underline{y}=1$.



Figuur 3. Product van steekproefomvang en variantie van $\bar{z}_{PS}$ voor een populatie met $P(\underline{y}=0) = P(\underline{y}=1) = 0,5$. De responskans is P voor $\underline{y}=0$ en $1-P$ voor $\underline{y}=1$.

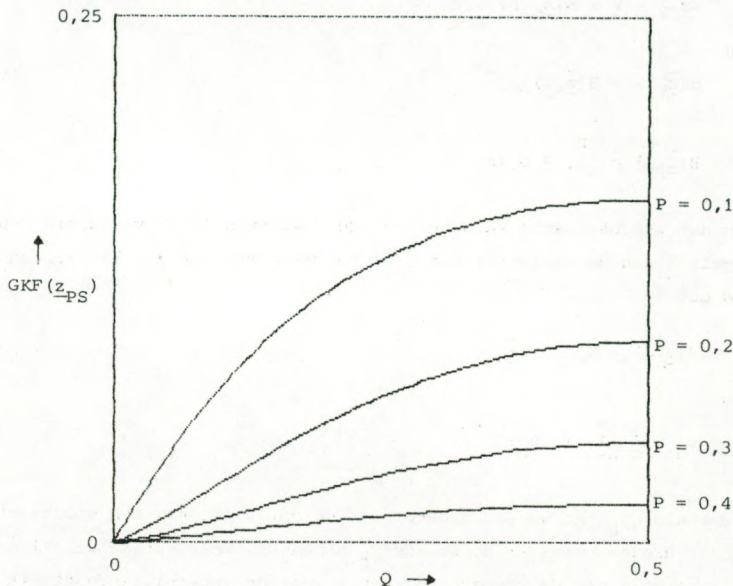


De omvang van de GKF wordt bepaald door zowel de variantie als de vertekening, maar bij een grote steekproefomvang zal de invloed van de vertekening de overhand krijgen. In figuur 4 is voor diverse waarden van P en $n=100$ de grafiek getekend van de GKF van $\bar{z}_{PS}$ als functie van Q . De waarde van de GKF voor $Q=0,5$ geeft tevens de waarde van de GKF zonder post-stratificatie. Ook voor een steekproefomvang van 100 is de invloed van de variantie al behoorlijk afgenomen. De GKF met post-stratificatie is in deze situatie altijd kleiner dan de GKF zonder post-stratificatie.

5. Geschatte gewichten

Post-stratificatie kan alleen worden uitgevoerd indien gewichten w_h bekend zijn. Als deze gewichten niet bekend zijn, dan kunnen ze toch vaak

Figuur 4. De gemiddelde kwadratische fout van $\bar{z}_{PS}$ voor een populatie met $P(\underline{y}=0) = P(\underline{y}=1) = 0,5$. De responskans is P voor $\underline{y}=0$ en $1-P$ voor $\underline{y}=1$. De steekproefomvang is 100.



geschat worden op grond van metingen die voor non-respondenten aan achtergrondvariabelen verricht zijn:

$$\bar{w}_h = \frac{n_h}{n}. \quad (80)$$

De grootheid $\bar{w}_h$ is een zuivere schatter voor w_h . Indien er geen non-respons optreedt, is dit een omslachtige procedure die niets nieuws oplevert. Immers, vervangen van w_h door $\bar{w}_h$ in definitie (53) levert weer de schatter $\bar{y}$ voor het geval er geen post-stratificatie wordt toegepast.

5.1. Non-respons

In geval er wel non-respons optreedt, kan deze procedure wel zinvol zijn. Definieer daartoe de schatter $\bar{z}_{EW}$ door

$$\bar{z}_{EW} = \sum_{h=1}^H w_h \bar{z}_h, \quad (81)$$

waarin $\bar{z}_h$ als gedefinieerd in (57). Aangezien n_h en $\bar{z}_h$ ongecorreleerd zijn is de onzuiverheid van $\bar{z}_{EW}$ even groot als die van $\bar{z}_{PS}$:

$$E\bar{z}_{EW} = \bar{y} + B(\bar{z}_{EW}), \quad (82)$$

waarin

$$B(\bar{z}_{EW}) = B(\bar{z}_{PS}) \quad (83)$$

en

$$B(\bar{z}_{PS}) = \sum_{h=1}^H w_h B_h(\bar{z}). \quad (84)$$

Gezien het stochastische karakter van de gewichten is de variantie van $\bar{z}_{EW}$ niet gelijk aan de variantie van $\bar{z}_{PS}$. De variantie van $\bar{z}_{EW}$ kan worden benaderd met

$$V(\bar{z}_{EW}) = V_W + V(\bar{z}_{PS}), \quad (85)$$

waarin

$$V_W = \frac{1}{n} \sum_{h=1}^H w_h (\bar{z}_h - \bar{z})^2. \quad (86)$$

De grootheid nV_W kan worden opgevat als de variantie van een stochastische variabele die met kans w_h de waarde $\bar{z}_h$ aanneemt. Deze variantie zal met name groot zijn als de verwachting van $\bar{z}$ voor de verschillende strata nogal uiteenloopt.

Aangezien V_W kleiner wordt met toenemende n , zal voor voldoende grote n de winst in nauwkeurigheid die met post-stratificatie behaald wordt, niet ongedaan worden gemaakt door de extra variantie-component die ontstaat door het schatten van de gewichten.

5.2. Speciaal geval

In het speciale geval is V_W gelijk aan

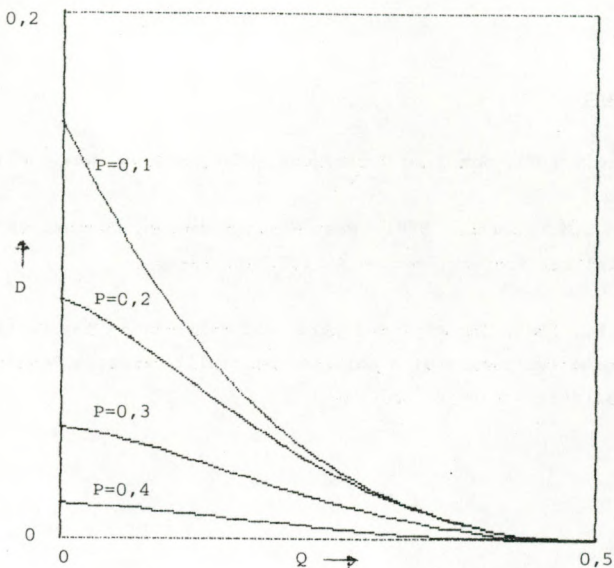
$$V_W = \frac{1}{n} \left[\frac{(1-P)(1-Q)/2}{PQ + (1-P)(1-Q)} - \frac{(1-P)Q/2}{P(1-Q) + (1-P)Q} \right]^2. \quad (87)$$

Teneinde het effect van het schatten van de gewichten in dit geval wat nader te kunnen bekijken is in figuur 5 voor $n=100$ en een aantal waarden van P de grafiek getekend van

$$D = \text{GKF}(\bar{z}) - \text{GKF}(\bar{z}_{EW}) \quad (88)$$

als functie van Q .

Figuur 5. De toename in de nauwkeurigheid door post-stratificatie met geschatte gewichten in een populatie met $P(\underline{y}=0) = P(\underline{y}=1) = 0,5$. De responskans is P voor $\underline{y}=0$ en $1-P$ voor $\underline{y}=1$. De steekproefomvang is 100.



Positieve waarden van D duiden erop dat het in die situatie de moeite waard is om gewichten te schatten. Uit de grafiek blijkt dat vrijwel altijd winst kan worden behaald. Alleen bij een kleine samenhang of een slechte stratificatie is het minder zinvol.

6. Conclusie

Indien door het optreden van non-respons schattingen voor populatieparameters onzuiver zijn, dan is het zeker de moeite waard om te overwegen om post-stratificatie toe te passen. Wil post-stratificatie zinvol zijn, dan is het wel noodzakelijk om zorgvuldig de stratificatievariabele(n) te selecteren. Met name stratificatievariabelen die leiden tot strata die homogeen zijn met betrekking tot de responskansen zullen de vertekening reduceren en daarmee de nauwkeurigheid verhogen. Zijn de voor post-stratificatie benodigde gewichten niet bekend, maar kunnen ze wel uit de steekproef geschat worden, dan is dit de moeite waard om toe te passen, mits de steekproef niet te klein is.

7. Literatuur

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

Holt, D. en T.M.F. Smith, 1979, Post Stratification, Journal of the Royal Statistical Society, series A, 142, pp. 33-46.

Stephan, F.F., 1945, The expected value and variance of the reciprocal and other negative powers of a positive Bernoulli variate, Annals of Mathematical Statistics 16, pp. 50-61.