Standaardisatiemethoden

10

Abby Israëls

Statistische Methoden (10003)

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

© Centraal Bureau voor de Statistiek, Den Haag/Heerlen, 2010.
Verveelvoudiging is toegestaan, mits het CBS als bron wordt vermeld.

6016510003 X-37

Inhoudsopgave

1. Inleiding op het thema 4
2. Directe standaardisatie 9
3. Indirecte standaardisatie 17
4. Regressie-analyse 23
5. Vergelijking van de methoden 29
6. Literatuur 30
7. Bijlage 1. Toets op gelijkheid van leeftijdsspecifieke sterftekansen 32
8. Bijlage 2. Toepassing standaardisatie op sterftecijfers 34

1. Inleiding op het thema

1.1 Algemene beschrijving en leeswijzer

Vooral in de demografie en in de epidemiologie is een veel voorkomend probleem het vergelijken van uitkomsten van populaties die qua opbouw verschillen ten aanzien van achtergrondkenmerken. Een voorbeeld hiervan is het vergelijken van sterftecijfers ten gevolge van hart- en vaatziekten voor populaties met een verschillende leeftijdsopbouw. Landen met een jonge bevolking zullen bij een vergelijkbare gezondheidszorg doorgaans een lager sterftecijfer hebben dan landen met een veel oudere bevolking. Het (bruto-)sterftecijfer is dan geen goede indicator voor de gezondheid van de burgers. Pas wanneer men controleert voor leeftijdseffecten, door individuen slechts per leeftijdsklasse met elkaar te vergelijken, is een eerlijke vergelijking mogelijk. Men kan daartoe voor iedere populatie leeftijdsspecifieke sterftecijfers bepalen. Maar kan men ook gemiddelden bepalen die voor leeftijd gecorrigeerd zijn, de zogenaamde *gestandaardiseerde sterftecijfers*[1] of *gestandaardiseerde gemiddelden*. Hiervoor zijn standaardisatie-methoden ontwikkeld, waarbij men onderscheidt:

1. een doelvariabele (*Y*);
2. populaties, nl.
   a) de met elkaar te vergelijken populaties, en
   b) een 'standaardpopulatie' (referentiepopulatie);
3. variabelen waar men naar standaardiseert, de zogenaamde 'verstorende kenmerken';
4. een doelfunctie (gemiddelde, $\overline{Y}$) of doelparameter (verwachting, $\mu = E(\overline{Y})$).

In het genoemde voorbeeld is *Y* het wel/niet sterven aan een hart- of vaatziekte in een bepaalde periode en populatie, en μ de achterliggende sterftekans. Verder zijn de onderscheiden landen de populaties, waarbij bijvoorbeeld één specifiek land als standaard kan worden genomen. Ten slotte is Leeftijd (in klassen) de variabele waarnaar wordt gestandaardiseerd.

Doelvariabele *Y* kan een binaire (0/1) variabele zijn, zoals geen/wel sterfte, maar ook een meer algemene kwantitatieve variabele, zoals jaarloon, of een frequentie-variabele (bijv. het aantal zwangerschappen). Het type variabele kan van belang zijn bij het bepalen van betrouwbaarheidsmarges of bij het modelleren. Het is ook mogelijk (alle scores uit) een frequentieverdeling te standaardiseren.

---

[1] Sterftecijfers worden op het CBS gedefinieerd als aantallen overledenen in een bepaalde periode per 1 000 of 10 000 inwoners. In dit rapport definiëren we een sterftecijfer als het gemiddeld aantal sterfgevallen per inwoner.

De standaardpopulatie kan één van de onderscheiden populaties zijn. Vaak wordt echter de vereniging van de beschouwde populaties als standaard genomen (‘sompopulatie’). Ook kan men een hypothetische referentiegroep als standaard nemen. Wil men verschillen tussen populaties in de tijd volgen dan kan men een standaardjaar introduceren. Zo heeft men bijvoorbeeld de hypothetische, d.w.z. vereenvoudigde, populatie voor Europa 1950 en Europa 2000 geconstrueerd.

Gestandaardiseerd wordt naar ‘verstorende variabelen’, die een eerlijke vergelijking van de doelfunctie voor verschillende populaties in de weg staan. Mortaliteitscijfers en morbiditeitcijfers (prevalenties, of gemiddeld aantal zorgcontacten) worden bij het CBS vrijwel altijd gestandaardiseerd naar leeftijd en geslacht, of naar leeftijd per geslacht. Maar voor het vergelijken van verzuimpercentages tussen verschillende groepen medewerkers, zal ook het schaalniveau van de medewerkers als een verstoring kunnen worden opgevat, waarvoor gecorrigeerd moet worden. Verstorende variabelen zijn variabelen waarvan het effect op de doelvariabele (al lang) bekend is, en waarbij men dit effect wil ‘wegrekenen’ om overblijvende effecten of veranderingen zichtbaar te maken.

Omdat het beschrijven van de standaardisatie in algemene termen moeizame formuleringen oplevert, wordt de beschrijving zoveel mogelijk gedaan aan de hand van het voorbeeld van sterftecijfers die naar leeftijd gestandaardiseerd moeten worden. Door het gebruik van formules wordt vanzelf de benodigde algemene uitleg bereikt.

Standaardisatie is al voor actuariële berekeningen in gebruik sinds het midden van de 18e eeuw (Keiding, 1987), toen de ‘zakjapanner’ en zelfs mechanische rekenhulpen nog niet ter beschikking stonden. Andere toepassingen van standaardisatie zijn sterftecijfers naar doodsoorzaak, zoals reeds genoemd, aantal ziekenhuisopnamen, vruchtbaarheidscijfers, besteedbaar inkomen voor verschillende doelgroepen (gecorrigeerd voor bijvoorbeeld verschillen in grootte en samenstelling van het huishouden), etc.

Er bestaan vanouds twee methoden om te standaardiseren: directe en indirecte standaardisatie. Bij *directe standaardisatie* (hoofdstuk 2) wordt voor iedere populatie de verdeling van de verstorende kenmerken bij de standaardpopulatie gebruikt. Bij *indirecte standaardisatie* wordt voor iedere populatie het sterftecijfer vergeleken met het sterftecijfer dat zou zijn verkregen wanneer de leeftijdsspecifieke sterftecijfers gelijk zouden zijn aan die van de standaardpopulatie (hoofdstuk 3). Ook met lineaire regressie is het mogelijk sterftecijfers te corrigeren voor verstorende kenmerken. Bij sterftecijfers ligt het voor de hand om dan logistische regressie toe te passen, omdat Sterfte een binaire variabele is. Deze en andere vormen van *regressie-analyse* worden in hoofdstuk 4 besproken. Ook wordt de link gelegd met directe en indirecte standaardisatie. Vanwege de eenvoud en het feit dat er voor standaardisatie minder gegevens nodig zijn, werden oorspronkelijk vooral directe en indirecte standaardisatie gebruikt. In hoofdstuk 5 komen ten slotte enkele verbanden tussen de verschillende methoden aan bod. Vanzelfsprekend wordt in het rapport ingegaan op voor- en nadelen van de methoden in diverse situaties.

1.2 Afbakening en relatie met andere thema's

Zoals beschreven, worden standaardisatiemethoden veel gebruikt om sterftecijfers van populaties met elkaar te vergelijken, waarbij wordt gecorrigeerd voor verschillen in leeftijdsopbouw. Men kan overlevingstafels (zie thema 'Overlevingstafels' uit de Methodenreeks; Van der Meulen, 2009) als basismateriaal gebruiken voor directe en indirecte standaardisatie.

Standaardisatiemethoden hebben een sterke gelijkenis met samengestelde index-cijfers (zie thema 'Indexcijfers' uit de Methodenreeks; Van der Grient en De Haan, 2008). Bij beide onderwerpen gaat het om de presentatie van samenvattingsmaten, waarbij een weging plaatsvindt over de categorieën van de 'verstorende' kenmerken. We bespreken de gelijkenis in de deelparagrafen 2.5.1 en 3.5.1.

1.3 Plaats in het statistisch proces

Men kan standaardisatie als een nadere analyse van de data beschouwen. Op het CBS worden echter bij veel gezondheidsstatistieken en sommige bevolkingsstatistieken standaard zowel ongestandaardiseerde als gestandaardiseerde resultaten gepubliceerd, omdat presentatie van alleen ongestandaardiseerde cijfers gauw tot verkeerde interpretaties kan leiden. Berekenen van gestandaardiseerde cijfers is daardoor vaak een vast onderdeel van de output.

1.4 Definities

Begrip	Omschrijving
bruto-sterftecijfer	aantal overledenen in een bepaalde periode per x aantal van de bevolking. Voor x neemt men vaak 1, 1000 of 10000. In dit rapport nemen we x=1.
sterfteratio	quotiënt van bruto-sterftecijfer van onderzochte en standaardpopulatie.
standaardiseren, standaardisatie	corrigeren van gemiddelden of totaalcijfers voor de invloed van verstorende kenmerken
gestandaardiseerd gemiddelde	gemiddelde na correctie voor de invloed van verstorende kenmerken
gestandaardiseerd sterftecijfer	aangepast bruto-sterftecijfer, door correctie voor de invloed van verstorende kenmerken (voorbeeld van een gestandaardiseerd gemiddelde)
directe standaardisatie	methode waarbij (sterfte)cijfers van een populatie worden toegepast op een standaardbevolking.
indirecte standaardisatie	methode waarbij een geobserveerd (bruto-sterfte)cijfer wordt vergeleken met het (sterfte)cijfer dat wordt verkregen door uit te gaan van leeftijdsspecifieke sterftecijfers van een externe populatie.
CMF	Comparative Mortality Figure
SMR	Standard Mortality Ratio

1.5 Algemene notatie

We gebruiken het subscript i voor de klassen van de variabele (verstorend kenmerk) waar men naar standaardiseert (i=1,…,I), j voor de beschouwde populaties (j=1,…,J) en s voor de standaard- of referentiepopulatie.

Verder is in een bepaalde periode
N_{ij} = aantal personen in leeftijdsklasse *i*, populatie *j*,
D_{ij} = sterfte (aantal doden) in leeftijdsklasse *i*, populatie *j*.

Bovenstaande gegevens zijn de basisgegevens van waaruit de onderstaande kunnen worden afgeleid:

$N_{+j} = \sum_i N_{ij}$ = omvang populatie *j*,

$q_{i|j} = N_{ij}/N_{+j}$ = leeftijdsverdeling voor populatie *j*,

$D_{+j} = \sum_i D_{ij}$ = sterfte (aantal doden) in populatie *j*,

$Y_{ij} = D_{ij}/N_{ij} = \overline{D}_{ij}$ = leeftijdsspecifieke sterftecijfer in (i,j) = gemiddelde sterfte per inwoner in (*i*,*j*),

$Y_{\cdot j} = D_{+j}/N_{+j} = \overline{D}_{\cdot j}$ = bruto-sterftecijfer = gemiddelde sterfte per inwoner in *j*.

Achter een sterftefractie Y_{ij} kunnen we een sterftekans μ_{ij} denken zó dat $\hat{\mu}_{ij} = Y_{ij}$.

Het bruto-sterftecijfer is een gewogen som van de leeftijdsspecifieke sterftecijfers.

$$Y_{\cdot j} = \sum_i q_{i|j} Y_{ij} \quad . \tag{1.1}$$

Voor de standaardpopulatie wordt subscript '*j*' in deze formules vervangen door '*s*'. Vaak kiest men voor de standaardpopulatie de vereniging van alle populaties *j* (=1,..,*J*). In dat geval geldt

$$D_{is} = \sum_j D_{ij}, \qquad D_{+s} = \sum_j D_{+j}, \qquad N_{is} = \sum_j N_{ij}, \qquad N_{+s} = \sum_j N_{+j} \quad . \tag{1.2}$$

Wanneer de sterftecijfers op een bepaalde periode slaan, kan voor de populatie-omvang N_{ij} bijvoorbeeld het gemiddelde over een aantal datums uit die periode worden genomen (bijvoorbeeld het gemiddelde van de bevolking aan het begin van de periode en die aan het eind van de periode), of de populatie-omvang op de middelste datum. Praktisch gezien is het zelden nodig om preciezer te werken, maar men kan N_{ij} ook definiëren als het aantal persoonsjaren in een bepaalde periode, d.w.z. de som van de individuele risicoperioden voor alle personen, uitgedrukt in jaren.

D_{ij} is een geaggregeerde grootheid. We kunnen de achterliggende variabele *D* (Sterfte) definiëren met de persoon als objecttype. *D* is dan een binaire variabele met waarden 1 (overleden) en 0 (niet overleden), en D_{ijk} de score op variabele *D* van individu *k* uit leeftijdsklasse *i* en populatie *j* (k=1,…,N_{ij}).

Meer algemeen kunnen we uitgaan van een kwantitatieve variabele *Y* met individuele scores Y_{ijk} met k = 1,…,N_{ij}. Y_{ij} is dus gedefinieerd als het *gemiddelde* van deze individuele scores. Evenzo kunnen we Y_{ij} zien als een aggregaat van individuele scores Y_{ijk}. Als *Y* binair is, geldt $Y_{ijk} = D_{ijk}$, omdat $N_{ijk} \equiv 1$. Standaardisatiemethoden worden afgeleid in termen van *Y* en *N*, maar omdat ze vaak

worden toegepast op binaire variabelen, zullen we de meeste formules ook weergeven in termen van *D* en *N*.

Ten slotte definiëren we sterfteratio's. De *leeftijdsspecifieke sterfteratio* voor populatie *j* ten opzichte van de standaardpopulatie *s* is

$$R_i = \frac{D_{ij} / N_{ij}}{D_{is} / N_{is}} = \frac{Y_{ij}}{Y_{is}} \quad (i=1,\ldots,I) \; . \tag{1.3}$$

De *bruto-sterfteratio* voor populatie *j* t.o.v. de standaardpopulatie *s* is

$$R = \frac{D_{+j} / N_{+j}}{D_{+s} / N_{+s}} = \frac{Y_{.j}}{Y_{.s}} . \tag{1.4}$$

Gemakshalve laten we bij deze ratio's de subscripten *j* en *s* weg.

Als gevolg van verschillende leeftijdsverdelingen ligt de bruto-sterfteratio *R* niet per se tussen de maximum- en minimum-R_i in. Gestandaardiseerde sterfteratio's voldoen wel aan die eis.

Om wat meer vertrouwd met de notatie te geraken geven we in tabel A (en B) van bijlage 2 een getallenvoorbeeld; zie kolommen (1)-(10). Op deze tabel zullen in de volgende hoofdstukken een aantal standaardisatiemethoden worden toegepast. In paragraaf 2.4 wordt het oorspronkelijke doel van de analyse uitgelegd. De tabel bestrijkt de periode 1979-1986. De aantallen overledenen (D_{ij} en D_{is}) hebben dus betrekking op een periode van acht jaar. De populatie-aantallen (N_{ij} etc.) in deze tabel zijn sommen van de jaartotalen over deze acht jaren, als benadering van het aantal persoonsjaren; we hadden in plaats hiervan de gemiddelden over de jaren kunnen nemen. Het jaartotaal in jaar *t* is berekend als het gemiddelde van de bevolkingsomvang op 1 januari *t* en 1 januari *t+1*.

2. Directe standaardisatie

2.1 Korte beschrijving

Bij directe standaardisatie worden voor iedere populatie *j* de leeftijdsspecifieke sterftecijfers Y_{ij} gewogen met een standaard-leeftijdsverdeling (leeftijdsverdeling in de standaardpopulatie) i.p.v. met de eigen leeftijdsverdeling zoals bij het bruto-sterftecijfer van formule (1.1). Dit leidt tot het *direct gestandaardiseerde sterftecijfer* voor populatie *j*:

$$Y_j^{DIR} = \sum_i q_{i|s} Y_{ij} \quad . \tag{2.1}$$

Delen door het bruto-sterftecijfer in de standaardpopulatie leidt tot de zogenaamde 'Comparative Mortality Figure'[2]:

$$CMF = \frac{\sum_i q_{i|s} Y_{ij}}{\sum_i q_{i|s} Y_{is}} = \frac{\sum_i q_{i|s} Y_{ij}}{Y_{*s}} \quad . \tag{2.2}$$

De CMF is dus een maat voor de verhouding van de sterfte bij populaties *j* en *s*, gecorrigeerd voor leeftijd. Zowel het berekenen van formule (2.1) als (2.2) wordt *directe standaardisatie* genoemd.

2.2 Toepasbaarheid

1. De CMF stelt ons instaat de sterfte in een populatie *j* te vergelijken met de sterfte in de standaardpopulatie. Omdat er een vaste standaard wordt gehanteerd, laten de CMF's ook toe om de sterftecijfers in meerdere populaties *j* onderling te vergelijken; de noemers zijn namelijk gelijk. Bij indirecte standaardisatie ligt dit vergelijken van sterftecijfers van verschillende populaties moeilijker (hoofdstuk 3). Daarom heeft directe standaardisatie in het algemeen de voorkeur wanneer men de sterfte in bijvoorbeeld meerdere landen of regio's met elkaar wil vergelijken, of voor meerdere jaren of diverse herkomstgroepen.
2. In formule (2.4) in paragraaf 2.3.1 zal worden getoond dat de CMF te schrijven is als een gewogen gemiddelde van de leeftijdsspecifieke sterfteratio's R_i uit (1.3), waarbij wordt gewogen met de fractie doden in de standaardpopulatie. Er bestaat een discussie over het toepassen van directe standaardisatie wanneer de R_i onderling sterk verschillen. Sommige auteurs vinden dat men zich in dat geval moet beperken tot het publiceren van leeftijdsspecifieke sterftecijfers en dat slechts bij redelijk homogene R_i de CMF een zinvolle samenvattingsmaat van de R_i's is. Anderzijds zijn we gewend om gemiddelde scores te publiceren

[2] Verwarrend is dat sommigen de term 'Comparative Mortality Rate (CMR)' gebruiken voor het gemiddelde van het bruto- en het direct gestandaardiseerde sterftecijfer.

(bijvoorbeeld een bruto-sterftecijfer), zonder dat dit vereist dat iedereen (of iedere leeftijdsklasse) dat gemiddelde scoort. Dit geeft voor andere auteurs een rechtvaardiging voor het berekenen van een CMF, ook wanneer de R_i's onderling aanzienlijk verschillen. Maar men moet zich dan wel bewust zijn van de invloed die de gewichten van R_i op de uitkomst hebben, en het is dan toch verstandig om ook de R_i te presenteren. Meer over voor- en nadelen van het bepalen van gestandaardiseerde cijfers vindt men in Fleiss (1973, hoofdstuk 13).

3. Het kiezen van een vaste (niet-stochastische) of zeer grote standaardpopulatie vergroot de nauwkeurigheid van de CMF en vergemakkelijkt de berekening van standaardfouten (paragraaf 2.3.2). Vaak neemt men als standaard de vereniging over alle populaties j=1,…,J, d.w.z. over alle te beschouwen landen, regio's of perioden. Soms moet internationale overeenstemming bereikt worden over de keuze van een geschikte standaardpopulatie.

   Op het CBS wordt directe standaardisatie ook toegepast bij lange tijdreeksen van mortaliteits- en morbiditeitscijfers, waarbij naast verschillende jaren ook verschillende populaties (bijvoorbeeld herkomstgroepen) onderling worden vergeleken, door te standaardiseren naar leeftijd (apart per geslacht, of samen); j is dan dus een combinatie van populatie en jaar. Er zijn dan meer keuzen voor de standaardpopulatie. Soms wordt gestandaardiseerd naar de sompopulatie (bijvoorbeeld alle herkomstgroepen) in een te selecteren basisjaar. Maar men kan ook kiezen voor alleen standaardiseren binnen het jaar (met elk jaar de sompopulatie van de herkomstgroepen van dat jaar als standaard), zoals wordt gedaan bij de standaardisatie van huisartscontacten, waar nog geen lange tijdreeks is.

4. Toepassen van directe standaardisatie is af te raden wanneer voor één of meer leeftijdsklassen $q_{i|j}$ heel klein is en de bijbehorende $q_{i|s}$ veel groter. $Y_{ij} = D_{ij}/N_{ij}$ is dan op weinig waarnemingen gebaseerd en telt toch zwaar mee in (2.1). De varianties van Y_j^{DIR} en CMF, die in paragraaf (2.3.2) worden gegeven, worden daardoor zeer groot. In tabel B van bijlage 2 zouden we hier al last mee krijgen wanneer we de leeftijdsklasse 65+ meenemen, maar zeker wanneer we deze klassen zouden uitsplitsen. De verhouding $q_{i|s}/q_{i|j}$ is hier gelijk aan 9,95/0,16 = 63, hetgeen betekent dat de bijdrage voor deze leeftijdsklasse aan de standaardfout bij directe standaardisatie 63 keer zo groot is als bij het bruto-sterftecijfer $Y_{.s}$. Dit wordt nauwelijks door de andere leeftijdsklassen gecompenseerd. Vanzelfsprekend heeft naast de q-ratio ook het aantal waarnemingen, N_{ij}, invloed op de standaardfout.

   Het probleem kan zich ook voordoen voor landen met weinig ouderen bij een internationale vergelijking van sterftecijfers. Vanwege het risico op grote varianties moet men dan bij directe standaardisatie zuinig zijn met uitsplitsingen voor de hogere leeftijdscategorieën, wanneer de q-ratio daardoor sterk zou toenemen en het aantal waarnemingen (N_{ij}) gering is. Om dezelfde reden moet men zuinig zijn met het aantal verstorende kenmerken waarnaar men

standaardiseert. In principe worden namelijk alle interacties tussen deze kenmerken meegenomen; zie hierover paragraaf 4.5.1.

Een CBS-praktijkvoorbeeld is de LMR-statistiek 'Landelijke Medische Registratie, waar aantallen ziekenhuisopnamen naar herkomstland direct gestandaardiseerd worden naar de leeftijdsverdeling van de totale Nederlandse bevolking (per geslacht en totaal). Om toch betrouwbare gestandaardiseerde cijfers te verkrijgen, beperkte men aanvankelijk de populatie tot 0-50-jarigen, en later (toen de allochtonenpopulatie in oudere leeftijdsgroepen gegroeid was) tot 60 jaar.

Het kan ook voorkomen dat Y_{ij} onbekend is, doordat sterftecijfers niet in iedere populatie per leeftijdsklasse zijn gegeven. In dat geval is directe standaardisatie onmogelijk en besluit men vaak tot indirecte standaardisatie.

5. Standaardisatie is niet alleen toepasbaar op dummy-variabelen zoals wel/geen sterfte, maar ook op kwantitatieve *Y*-variabelen. Zo is in Israëls en De Ree (1981) standaardisatie toegepast voor een vergelijking van lonen tussen verschillende economische bedrijfstakken, waarbij is gestandaardiseerd naar leeftijd en opleiding van de werknemers.

2.3 Uitgebreide beschrijving

2.3.1 Bepaling van de CMF

Door vermenigvuldiging van teller en noemer van formule (2.2) met N_{+s} kunnen we de CMF schrijven als

$$CMF = \frac{\sum_i N_{is} Y_{ij}}{\sum_i N_{is} Y_{is}} = \frac{\sum_i (N_{is} / N_{ij}) D_{ij}}{D_{+s}} \ . \qquad (2.3)$$

De noemer is nu het aantal doden in de standaardpopulatie, en de teller het *direct gestandaardiseerde aantal doden* in populatie *j*, d.i. het aantal personen dat zou zijn gestorven wanneer populatie *j* de leeftijdsverdeling van de standaardbevolking zou hebben.

De CMF is ook te schrijven als een gewogen som van de leeftijdsspecifieke sterfteratio's R_i met gewichten $w_{is} = D_{is} / D_{+s}$:

$$CMF = \frac{\sum_i N_{is} Y_{ij}}{\sum_i N_{is} Y_{is}} = \frac{\sum_i N_{is} Y_{is} R_i}{\sum_i N_{is} Y_{is}} = \frac{\sum_i D_{is} R_i}{\sum_i D_{is}} = \sum_i w_{is} R_i \ . \qquad (2.4)$$

We kunnen de CMF dus zien als een samenvattingsmaat voor de leeftijdsspecifieke sterfteratio's R_i, met gewichten w_{is} evenredig aan D_{is}. In paragraaf 2.2, item 2, hebben we al vraagtekens gezet bij het presenteren van de CMF wanneer de R_i te heterogeen zijn. In paragraaf 3.5 tonen we de gelijkenis van formule (2.4) met het Laspeyres-prijsindexcijfer.

2.3.2 *Standaardfout van Y_j^{DIR} en CMF*

Bij het bepalen van de *standaardfout* van Y_j^{DIR} of CMF zijn N_{ij} en N_{is} doorgaans bekende populatie-aantallen en hebben dus variantie nul. Ook wanneer het geschatte populatie-aantallen of steekproefaantallen zouden zijn, is het verdedigbaar om conditioneel op deze aantallen te werken. Het gaat immers om vergelijking van sterftekansen. Ook de aantallen doden D_{ij} en D_{+s} (i=1,…,I) zijn populatiecijfers. Maar sterftecijfers D_{ij} worden in het algemeen toch als stochastisch behandeld. Het al dan niet sterven wordt dan gezien als het resultaat van een kansmechanisme dat ook net anders had kunnen uitvallen. Zo wordt aangenomen dat het aantal doden $\underline{D}_{ij}$ binomiaal verdeeld is met parameters N_{ij} en sterftekans p_{ij}, waardoor de variantie van $\underline{D}_{ij}$ gelijk is aan[3]

$$Var(\underline{D}_{ij}) = N_{ij}\, p_{ij}(1-p_{ij}) \tag{2.5}$$

en de geschatte variantie aan

$$\mathrm{var}(\underline{D}_{ij}) = N_{ij}\,\frac{D_{ij}}{N_{ij}}\left(1-\frac{D_{ij}}{N_{ij}}\right) = D_{ij}\left(1-\frac{D_{ij}}{N_{ij}}\right)\ . \tag{2.6}$$

Wanneer de sterftekans klein is, d.w.z. D_{ij}«N_{ij} (ofwel Y_{ij}«1), is $\underline{D}_{ij}$ bij benadering Poisson-verdeeld, waardoor $Var(\underline{D}_{ij}) = N_{ij}\, p_{ij}$ en $\mathrm{var}(\underline{D}_{ij}) = D_{ij}$.

Uitgaande van de binomiale verdeling van $\underline{D}_{ij}$ is de geschatte variantie van het direct gestandaardiseerde sterftecijfer

$$\mathrm{var}(Y_j^{dir}) = \mathrm{var}\left(\sum_i q_{i|s}\,\frac{\underline{D}_{ij}}{N_{ij}}\right) = \sum_i q_{i|s}^2\,\frac{1}{N_{ij}^2}\,\mathrm{var}(\underline{D}_{ij}) = \sum_i q_{i|s}^2\,\frac{1}{N_{ij}^2}\,D_{ij}\left(1-\frac{D_{ij}}{N_{ij}}\right)\ . \tag{2.7}$$

Hierbij is aangenomen dat de geschatte sterftecijfers voor verschillende leeftijdsklassen onafhankelijk zijn. Verkeersongelukken of epidemieën verstoren deze aanname, maar deze verstoring zal doorgaans relatief klein zijn. Voor de variantie van het gestandaardiseerd *aantal* doden moeten we de variantie uit (2.7) met N_{+s}^2 vermenigvuldigen. De 95%-betrouwbaarheidsmarge van het direct gestandaardiseerde sterftecijfer is 1,96 keer de wortel uit (2.7), uitgaande van de normale verdeling.

Voor de variantie van de CMF hebben we strikt genomen ook te maken met de stochastiek van D_{+s}; zie formule (2.3). Deze stochastiek wordt in de literatuur verwaarloosd, omdat de standaardpopulatie (veelal de sompopulatie) haast altijd zeer groot is. Chiang (1984) stelt zelfs dat alleen D_{ij} als stochastisch dient te worden beschouwd. De variantieschatter van de CMF volgens formule (2.2) wordt dus

---

[3] We onderstrepen de stochastische grootheid $\underline{D}_{ij}$ in de variantieformules, ter onderscheid van de realisaties D_{ij}.

$$\mathrm{var}(CMF) = \frac{1}{Y_{\cdot s}^2}\mathrm{var}(Y_j^{dir}) = \frac{1}{D_{+s}^2}\sum_i \frac{N_{is}^2}{N_{ij}^2} D_{ij}\left(1-\frac{D_{ij}}{N_{ij}}\right) \ . \qquad (2.8)$$

Chiang (1961, 1984) gaat bij de variantieberekeningen uit van een iets andere situatie, namelijk van overlevingstafels (Van der Meulen, 2009) waar geen sprake is van jaarlijkse sterfte, maar van het overlijden in een bepaalde leeftijdsklasse.

Als *Y* een kwantitatieve variabele is, moeten de variantieformules vanzelfsprekend worden aangepast. Voor $\underline{Y}_{ij}$ wordt dan een theoretische verdeling aangenomen, of wordt de variantie geschat op grond van de waargenomen verdeling. Een voorbeeld hiervan is te vinden in De Heij e.a. (2010).

Voor het bepalen van het 95%-betrouwbaarheidsinterval van de CMF, kan men uitgaan van normaliteit van de CMF en $1{,}96SE(CMF) \equiv 1{,}96\sqrt{\mathrm{var}(CMF)}$ als marge hanteren. Omdat ratio's asymmetrisch zijn, raden Breslow en Day (1987) een log-transformatie aan. Dit geeft $1{,}96SE\{\ln(CMF)\} = 1{,}96\{SE(CMF)\}/CMF$ als 95%-marge voor de natuurlijke logaritme van CMF, waarna via de e-macht het interval kan worden teruggetransformeerd. Dezelfde transformatie kan worden gebruikt voor het toetsen van 'CMF = 1'.

De toets of de CMF's van twee verschillende populaties *j* en *j'* ten opzichte van dezelfde standaardpopulatie aan elkaar gelijk zijn is hier eenvoudig uit af te leiden (Breslow en Day, 1987). Deze toets komt er namelijk op neer dat het quotiënt van de twee gestandaardiseerde sterftecijfers (ofwel van de twee CMF's) gelijk aan 1 is.

Wanneer populatie *j* een deel is van de standaardpopulatie, zoals het geval is wanneer de standaardpopulatie de vereniging is van alle beschouwde populaties *j*, dan kan men iets nauwkeuriger toetsen door de sterfte in populatie *j* te vergelijken met die in de vereniging van de overige populaties, *s\j*; zie Yule (1934).

2.4 Voorbeeld

Voorbeeld 1. *Sterftecijfers Turkse en Nederlandse mannen, 0-44 jaar: directe standaardisatie*

In Hoogenboezem en Israëls (1990) zijn analyses gedaan naar verschillen in sterftecijfers tussen Turkse, Marokkaanse en Nederlandse ingezetenen naar diverse doodsoorzaken in de jaren 1979-1988. Aanleiding daartoe waren vragen in de Tweede Kamer over de hoge sterfte onder Turkse en Marokkaanse kinderen in Nederland, vergeleken met autochtone leeftijdsgenoten. In Hoogenboezem en Israëls (1990) is gekozen voor indirecte standaardisatie. In dit voorbeeld geven we ter vergelijking de resultaten van directe standaardisatie op de data van tabel A van bijlage 2, terwijl we in paragraaf 3.4 de resultaten van indirecte standaardisatie bespreken. De data uit tabel A wijken overigens iets af van die in Hoogenboezem en Israëls (1990); we beperken ons hier tot de jaren 1979-1986 en tot 'mannen < 45 jaar'.

Het direct gestandaardiseerde sterftecijfer volgens formule (2.1) is gelijk aan 0,00137, dus 13,7 per 10.000 mensen; zie kolom (11) in tabel A van bijlage 2. De CMF volgens formule (2.2) is derhalve gelijk aan 13,71/8,71 = 1,575; zie kolom (12). Vermenigvuldiging van teller en noemer met N_{is} / 10.000 = 38 287 704 / 10.000 toont dat het direct gestandaardiseerd *aantal* Turkse overledenen in de periode 1979-1986, de teller van formule (2.3), gelijk is aan 52 501, wat 1,575 keer het aantal overleden Nederlandse mannen van 33 336 is. Conclusie is dat de sterfte onder Turkse mannen tot 45 jaar ruim 1½ keer zo groot is als bij de autochtonen. Dat de extra sterfte niet constant over de leeftijdsklassen is, tonen de waarden van R_i in tabel A. Bij kinderen is de sterfteratio veel groter dan 1½.

Zouden we de 'omgekeerde standaardisatie' hebben toegepast, d.w.z. de sterfte van Nederlanders (j) vergelijken met de Turken als standaard, dan geeft dit een CMF van 0,619. Dat verschilt slechts weinig met de reciproque van 1,575, maar dat is geen algemeenheid, omdat verschillende standaards worden gebruikt.

We hadden ook de leeftijdsklassen 45-65 en 65+ kunnen meenemen (tabel B). In dat geval hadden we een CMF gekregen van 0,611 i.p.v. 1,575! Niet alleen is de sterfte in de hogere leeftijdsklassen bij Turken lager dan bij Nederlanders, het zijn ook de klassen die veruit het grootste gewicht krijgen, omdat daarin de meeste autochtonen overlijden. Het aantal Turkse 65+-ers is zelfs zo gering dat een verdere uitsplitsing naar doodsoorzaak niet mogelijk is, omdat de variantie van de CMF per doodsoorzaak te zeer zou toenemen. Hierom is in Hoogenboezem en Israëls (1990) indirecte standaardisatie toegepast (paragraaf 3.4)

Uitgaande van normaliteit van de CMF, dan wordt het 95%-betrouwbaarheids-interval voor CMF (1,462; 1,687), symmetrisch om 1,575. Gaan we uit van normaliteit van de ℓn(CMF), wat beter is, dan krijgen we het asymmetrische betrouwbaarheidsinterval (1,466; 1,692). Het verschil is gering. Vanwege de kleine sterftekansen mochten we aannemen dat de variabele Sterfte Poisson is verdeeld.

2.5 Eigenschappen

2.5.1 Relatie met het Laspeyres-prijsindexcijfer

In het themarapport 'Indexcijfers' in de Methodenreeks (Van der Grient en De Haan, 2008) komt de volgende formule voor:

$$P_L^{t,0} = \frac{\sum_i q_i^0 p_i^t}{\sum_i q_i^0 p_i^0} = \sum_i w_i^0 \frac{p_i^t}{p_i^0} = \sum_i w_i^0 I_i^{t,0} \quad . \tag{2.9}$$

Hierbij is $P_L^{t,0}$ het Laspeyres-prijsindexcijfer in verslagperiode *t* ten opzichte van basisperiode 0, p_i^t de gemiddelde prijs van artikel *i* in verslagperiode *t*, q_i^0 de geconsumeerde 'hoeveelheid' van artikel *i* in basisperiode 0 en

$w_i^0 = q_i^0 p_i^0 / \sum_i q_i^0 p_i^0$ het gewicht van het enkelvoudige prijsindexcijfer $I_i^{t,0} = p_i^t / p_i^0$ van artikel *i* in het Laspeyres-prijsindexcijfer.

Evenals bij ons zijn de q_i relatieve bijdragen (consumptiepatronen voor artikelen i.p.v. leeftijdsverdelingen), en de w_i gewichten. Gemiddelde prijzen p_i nemen de plaats in van leeftijdsspecifieke sterftecijfers Y_i, en index $I_i^{t,0}$ de plaats van ratio $R_i^{j,s} \circ R_i$.

De interpretatie van formules (2.9) en (2.4) is wel enigszins verschillend. Bij indexcijfers gaat het er altijd om gemiddelde prijzen in twee perioden met elkaar te vergelijken, waarbij de prijzen worden gewogen met hoeveelheden *q*. De populaties zijn artikelen in twee perioden, waartussen een gemiddelde prijstoename wordt gedefinieerd. Bij demografische en gezondheidsstatistieken gaat het bij standaardisatie meestal om verschillen tussen populaties op hetzelfde tijdstip, maar bij lange tijdreeksen wordt op het CBS juist wel in de tijd gestandaardiseerd, met de populatie van een basisjaar als standaard. Wanneer er meer dan twee populaties bij de analyse zijn betrokken, is bij prijsindices altijd sprake van een in de tijd *geordende* serie prijsindexcijfers; bij standaardisatie van populaties op hetzelfde tijdstip neemt men vaak de vereniging van alle populaties als standaardpopulatie.

Bij prijsindexcijfers zijn zowel de gemiddelde prijzen als de hoeveelheden stochastisch, tenzij er sprake is van volledige waarneming van prijzen en/of transacties. Individueel gemeten prijzen zijn realisaties van een kwantitatieve variabele. Bij standaardisatie voor sterfte of morbiditeit zijn de aantallen (*N*) meestal vast en is alleen sterfte een stochastische variabele, die bovendien binair is, wat de berekening van betrouwbaarheidsintervallen vereenvoudigt.

Prijzen kunnen ook in de ruimte (tussen landen) worden vergeleken i.p.v. in de tijd; meer hiervoor in de laatste alinea van paragraaf 3.5.1.

2.5.2 Standaardiseren van nominale variabelen

Tot nu toe was variabele D de binaire variabele Sterfte. Men kan gestandaardiseerde gemiddelden meer algemeen berekenen voor iedere categorie van een nominale variabele met een multinomiale verdeling; zie De Ree en Israëls (1982). Per categorie gelden dan dezelfde formules als voor ‘wel/geen sterfte’. Zo zijn in Hoogenboezem en Israëls (1990) de formules ook toegepast op sterfte naar doodsoorzaak, al is daar uiteindelijk gekozen voor indirecte standaardisatie (zie paragraaf 3.5.2 voor indirecte standaardisatie voor nominale variabelen). Het is eenvoudig aan te tonen dat de direct gestandaardiseerde sterftecijfers per doodsoorzaak optellen tot het direct gestandaardiseerde sterftecijfer voor alle doodsoorzaken tezamen, en dat de CMF’s per doodsoorzaak gewogen optellen, met het aantal doden in de standaardpopulatie als gewichten. Immers ook de noemers van de CMF’s per doodsoorzaak tellen op tot de noemer van de CMF voor de totale sterfte.

2.6 Kwaliteitsindicatoren

- Naast het berekenen van standaardfouten en toetsen (paragraaf 2.3.2), kan men de stabiliteit van de oplossing onderzoeken door een gevoeligheidsanalyse uit te voeren. Zo kan men kijken of de gestandaardiseerde cijfers bestand zijn tegen het samenvoegen van leeftijdsklassen. Wanneer dit tot grote verschillen leidt heeft men een probleem: de oplossing is dan kennelijk instabiel. Een oplossing met meer klassen zal een grotere variantie hebben, maar minder vertekening. De vertekening is alleen meetbaar wanneer men van een bepaald model uitgaat, bijvoorbeeld een lineair verband tussen leeftijd en de doelvariabele. Het is daarom niet altijd te bepalen of de mean square error (gemiddelde kwadratische afwijking) door samenvoegen van klassen toe- of afneemt. In de regel zal men besluiten klassen samen te voegen wanneer de variantie daardoor sterk afneemt.
- Het is verstandig om niet alleen op geaggregeerd niveau de (gestandaardiseerde) sterfte in populatie *j* met die in de standaardpopulatie te vergelijken, door de CMF en bijbehorende marge te bepalen, maar om ook te toetsen of leeftijdsspecifieke sterftekansen gelijk zijn, d.w.z. $\mu_{ij} = \mu_{is}$ voor i=1,…,I. Zie voor deze toets bijlage 1.

3. Indirecte standaardisatie

3.1 Korte beschrijving

Bij *indirecte standaardisatie* berekent men de 'Standard Mortality Ratio (SMR)', ook wel 'Standard Morbidity Ratio' genoemd,

$$SMR = \frac{\sum_i q_{i|j} Y_{ij}}{\sum_i q_{i|j} Y_{is}} = \frac{Y_{\bullet j}}{\sum_i q_{i|j} Y_{is}} \ . \qquad (3.1)$$

Het verschil met de CMF is dat in teller en noemer de gewichten $q_{i|s}$ zijn vervangen door $q_{i|j}$. De teller is het bruto-sterftecijfer van populatie *j*; de noemer is het sterftecijfer in populatie *j* wanneer de leeftijdsspecifieke sterftecijfers gelijk zouden zijn aan die van de standaardpopulatie. De SMR geeft dus aan hoeveel doden er naar verhouding meer of minder in populatie *j* zijn dan in de standaardpopulatie wanneer deze de leeftijdsverdeling van populatie *j* had gehad.

3.2 Toepasbaarheid

We hebben gezien in paragraaf 2.2, item 4, dat directe standaardisatie tot grote standaardfouten leidt, wanneer één of meer leeftijdsspecifieke sterftecijfers op kleine aantallen berusten en desondanks sterk meewegen. Indirect standaardiseren is hier niet gevoelig voor en verdient dan de voorkeur.

Een tweede reden om indirecte standaardisatie toe te passen is wanneer directe standaardisatie niet mogelijk is omdat er benodigde gegevens ontbreken. Vaak zijn de leeftijdsspecifieke sterftecijfers Y_{ij} (ofwel het aantal overledenen D_{ij}) niet voor alle populaties *j* bekend en kan men daarom de Y_j^{DIR} niet berekenen. Bij indirecte standaardisatie zijn de leeftijdsspecifieke sterftecijfers alleen nodig voor de standaardpopulatie, en die zijn dan vaak wel gegeven.

Samenvattend zien we dat het onbetrouwbaar zijn of ontbreken van de leeftijdsspecifieke sterftecijfers Y_{ij} reden is geen directe standaardisatie te gebruiken.

Bij indirecte standaardisatie wordt de sterfte per populatie *j* vergeleken met die in de standaardpopulatie. Het *onderling* vergelijken van meerdere populaties *j* kan echter interpretatieproblemen geven. Directe standaardisatie is daarvoor geschikter, omdat daarbij een vaste weging wordt gebruikt, namelijk die van de standaardpopulatie. Meer hierover in paragraaf 3.5.

3.3 Uitgebreide beschrijving

3.3.1 Bepaling van de CMF

We kunnen de SMR ook schrijven als

$$SMR = \frac{\sum_i N_{ij} Y_{ij}}{\sum_i N_{ij} Y_{is}} = \frac{D_{+j}}{\sum_i N_{ij} D_{is} / N_{is}} = \frac{D_{+j}}{\sum_i (N_{ij} / N_{is}) D_{is}} \ . \qquad (3.2a)$$

Nu is de teller het *waargenomen* aantal doden in populatie *j*, en de noemer het *verwachte* aantal doden in populatie *j* wanneer de sterftekans per leeftijdsklasse gelijk is aan die van de standaardpopulatie. De SMR wordt daarom ook wel weergegeven als

$$SMR = O_j / E_j \ , \qquad (3.2b)$$

waarbij $O_j = D_{+j}$ staat voor 'observed count' en $E_j = \sum_i E_{ij} = \sum_i N_{ij} Y_{is}$ voor 'expected count', een meer gebruikte notatie in de statistiek. Feitelijk is dit een uitwerking van een modelmatige aanpak, waarbij een Poisson-model wordt aangenomen voor het aantal doden $\underline{D}_{ij}$ met als verwachting

$$E_{ij} \ ^{o} \ E(\underline{D}_{ij}) = N_{ij} E(\underline{Y}_{ij}) = N_{ij} \mu_{ij} = N_{ij} Y_{is} \ . \qquad (3.3a)$$

De $\mu_{ij} = E(\underline{Y}_{ij})$ zijn parameters voor de leeftijdsspecifieke sterfte, d.w.z. dat μ_{ij} de sterftekans is voor personen uit cel (*i*,*j*), met Y_{ij} als realisatie. Breslow en Day (1975) veronderstellen een multiplicatief model voor deze sterftekansen,

$$E(\underline{Y}_{ij}) = \mu_{ij} = \varphi_i \theta_j \ , \qquad (3.3b)$$

waarbij φ_i het effect is van leeftijd op de sterftekans, en θ_j het effect van de populatie. Dit multiplicatieve model gaat er dus vanuit dat de sterftekansen geen interactie 'Leeftijd x Populatie' bevatten, wat betekent dat de waargenomen leeftijdsspecifieke sterftecijfers R_i redelijk homogeen zijn; zie hieromtrent opmerking 2 in paragraaf 2.2.

Invullen van (3.3b) in (3.3a) betekent dat wordt aangenomen dat de kansvariabele $\underline{D}_{ij}$ een Poisson-verdeling heeft met parameter $(N_{ij}\varphi_i\theta_j)$. Model (3.3) kan worden gezien als een Poisson-regressie (McCullagh en Nelder, 1989). Een specifieke schatting van de parameters leidt tot de SMR als schatter voor θ_j; zie Breslow and Day (1975). We komen hierop terug in deelparagraaf 4.5.2.

Een alternatieve schrijfwijze voor formule (3.2a), vergelijkbaar met formule (2.4) voor de CMF, is

$$SMR = \frac{\sum_i N_{ij} Y_{ij}}{\sum_i N_{ij} Y_{is}} = \frac{\sum_i N_{ij} Y_{is} R_i}{\sum_i N_{ij} Y_{is}} = \sum_i w_i^* R_i \ . \qquad (3.4)$$

met $w_i^* = N_{ij} Y_{is} \Big/ \sum_i N_{ij} Y_{is}$. Dus evenals de CMF is de SMR een gewogen gemiddelde van de leeftijdsspecifieke ratio's R_i, zij het dat de gewichten ingewikkelder te interpreteren zijn. Het is dan ook niet verwonderlijk dat een aantal

prettige eigenschappen die bij directe standaardisatie gelden, bij indirecte standaardisatie niet gelden. We komen hier in paragraaf 3.5 op terug.

Een derde representatie van de SMR is:

$$SMR = \frac{\sum_i N_{ij} Y_{ij}}{\sum_i N_{ij} Y_{is}} = \frac{\sum_i N_{ij} Y_{ij}}{\sum_i N_{ij} Y_{ij} / R_i} = \left(\frac{\sum_i N_{ij} Y_{ij} R_i^{-1}}{\sum_i N_{ij} Y_{ij}} \right)^{-1} = \left(\sum_i w_{ij} R_i^{-1} \right)^{-1} \quad (3.5)$$

met $w_{ij} = D_{ij} / D_{+j}$. In paragraaf 3.5 tonen we de gelijkenis met het Paasche-prijsindexcijfer.

Evenals bij directe standaardisatie kan de ratio (SMR i.p.v. CMF) op het niveau van sterftecijfers worden gebracht door vermenigvuldiging met $Y_{.s}$, het sterftecijfer in de standaardpopulatie:

$$Y_j^{INDIR} = SMR \times Y_{.s} = \frac{Y_{.j}}{\sum_i q_{i|j} Y_{is}} Y_{.s} \ . \quad (3.6)$$

Y_j^{INDIR} wordt het *indirect gestandaardiseerde sterftecijfer* genoemd. Evenzo leidt vermenigvuldiging van de SMR met D_{+s} tot het *indirect gestandaardiseerde aantal doden* in populatie *j*. Meestal beperkt men zich echter tot de SMR. De indirect gestandaardiseerde cijfers leveren geen extra interpretatie op en worden wel ontraden. Fleiss (1973, p. 169) laat zien dat Y_j^{INDIR} groter (of kleiner) kan zijn dan alle Y_{ij}, hetgeen onwenselijk is.

3.3.2 Standaardfout van SMR

We zouden kunnen aannemen dat $\underline{D}_{+j}$ binomiaal verdeeld is met parameters N_{+j} en sterftekans $p_{.j}$, met $\hat{p}_{.j} = Y_{.j}$. Sterftekansen verschillen echter sterk tussen leeftijdsklassen. Daarom gaan we evenals in paragraaf 2.3.2 uit van een binomiale verdeling van $\underline{D}_{ij}$ met parameters N_{ij} en p_{ij} (i=1,…,I). Binnen iedere leeftijdsklasse heeft iedereen dan wel dezelfde sterftekans. Gebruikmakend van formule (2.5) is de variantie van $\underline{D}_{+j}$

$$Var(\underline{D}_{+j}) = Var(\sum_i \underline{D}_{ij}) = \sum_i Var(\underline{D}_{ij}) = \sum_i N_{ij} p_{ij} (1 - p_{ij}) \ . \quad (3.7)$$

De schatter hiervan is

$$\text{var}(\underline{D}_{+j}) = \sum_i N_{ij} Y_{ij} (1 - Y_{ij}) = \sum_i D_{ij} (1 - \frac{D_{ij}}{N_{ij}}) \ , \quad (3.8)$$

Dit leidt volgens formule (3.2a) tot

$$\text{var}(SMR) = \text{var}(\frac{\underline{D}_{+j}}{E_j}) = \frac{1}{E_j^2} \sum_i D_{ij} (1 - \frac{D_{ij}}{N_{ij}}) \ , \quad (3.9)$$

met E_j het verwachte (expected) aantal doden. Aangenomen wordt hier dat E_j niet stochastisch is, wat inhoudt dat D_{is} als niet-stochastisch wordt opgevat. Dit is te verdedigen wanneer men de zaak modelmatig beziet, zoals Breslow en Day doen (zie formule (3.3)), of wanneer de D_{is} zo groot zijn dat hun effect op de variantie van de SMR te verwaarlozen is. Zonder de model-aanname zal de variantie van SMR groter zijn, omdat dan ook D_{is} als stochastisch wordt beschouwd.

Evenals bij de directe standaardisatie worden de formules eenvoudiger wanneer we aannemen dat $\underline{D}_{ij}$ Poisson is verdeeld. Uit formule (3.9) volgt dat de variantie van SMR dan gelijk is aan SMR^2/D_{+j}. Voor betrouwbaarheidsintervallen en voor het toetsen of SMR = 1 behoort men eerst een log-transformatie op de SMR uit te voeren, net als bij de CMF in paragraaf 2.4 (Breslow en Day, 1987).

3.4 Voorbeeld

Sterftecijfers Turkse en Nederlandse mannen, 0-44 jaar: indirecte standaardisatie

We passen nu indirecte standaardisatie toe op tabel A van bijlage 2. De resultaten staan in de kolommen (13) en (14).

Het aantal doden per 10.000 Turkse mannen, Y_j, was in de periode 1979-1986 gelijk aan 14,6 (kolom 8). Dit is de teller van formule (3.1). De noemer, het sterftecijfer per 10.000 Turkse mannen wanneer de leeftijdsspecifieke sterftecijfers gelijk zouden zijn aan die van de Nederlandse mannen, is gelijk aan 9,05 (kolom 13). De SMR voor Turkse mannen t.o.v. Nederlandse mannen is dus gelijk aan 14,6/9,05 = 1,616. Dat is iets hoger dan de CMF van 1,575 uit paragraaf 2.4.[4] Dat komt doordat er relatief meer Turken dan Nederlanders zitten in de leeftijdsklassen met hoge sterfteratio's R_i, namelijk de 1-14-jarigen; vergelijk formules (2.2) en (3.1). (Y_j^{INDIR}, het indirect gestandaardiseerde sterftecijfer volgens formule (3.6), is gelijk aan 1,616 x 8,7 x 10^{-4} = 0,00141, oftewel 14,1 doden per 10.000.) Het verschil met directe standaardisatie voor de populatie tot 45 jaar blijkt dus niet erg groot. Bij beperking tot deze leeftijd had dus ook directe standaardisatie kunnen worden toegepast. In Hoogenboezem en Israëls (1990) is voor indirecte standaardisatie gekozen, omdat directe standaardisatie tot grote standaardfouten zou hebben geleid voor andere herkomstgroepen, met name voor Turkse en Marokkaanse vrouwen. Bovendien ging het om een vergelijking van buitenlanders met Nederlanders, en niet tussen buitenlanders onderling, zoals in paragraaf 2.4 is uitgelegd.

Wanneer je alle leeftijden beschouwt, volgens de indeling van tabel B van bijlage 2, dan wijkt directe standaardisatie wel sterk af van indirecte standaardisatie: CMF = 0,611 en SMR = 0,920. Dit verschil wordt veroorzaakt door het veel kleinere aandeel van de leeftijdscategorieën 45-64 en 65+ bij Turkse t.o.v. Nederlandse mannen. In die categorieën overleden relatief weinig Turken.

---

[4] Merk op dat beide gestandaardiseerde sterftecijfers kleiner zijn dan het bruto-sterftecijfer van 1,68 (kolom 10). Leeftijd verklaart dus een deel van de hogere sterfte bij Turkse mannen.

De ‘omgekeerde standaardisatie’, d.w.z. de sterfte van Nederlandse mannen (*j*) ten opzichte van de Turkse mannen als standaard, geeft een SMR van 0,635. Dat verschilt slechts weinig met de reciproque van 1,616, maar dat is geen algemeenheid, omdat verschillende standaards worden gebruikt.

Uitgaande van normaliteit van de SMR, is het 95%-betrouwbaarheidsinterval voor SMR (1,507; 1,726), symmetrisch om 1,616. Gaan we uit van normaliteit van de ℓn(SMR), wat beter is, dan krijgen we het asymmetrische betrouwbaarheidsinterval (1,510; 1,730). Het verschil is gering. Vanwege de kleine sterftekansen hebben we hierbij aangenomen dat de variabele Sterfte Poisson is verdeeld.

3.5 Eigenschappen

3.5.1 Relatie met het Paasche-prijsindexcijfer

In het themarapport ‘Indexcijfers’ in de Methodenreeks (Van der Grient en De Haan, 2008) komt de volgende formule voor:

$$P_P^{t,0} = \frac{\sum_i q_i^t p_i^t}{\sum_i q_i^t p_i^0} = \left[\sum_i w_i^t \left(\frac{p_i^t}{p_i^0} \right)^{-1} \right]^{-1} = \left[\sum_i w_i^t \left(I_i^{t,0} \right)^{-1} \right]^{-1} \qquad (3.10)$$

met $P_P^{t,0}$ het Paasche-prijsindexcijfer in verslagperiode *t* ten opzichte van basisperiode 0, p_i^t de gemiddelde prijs van artikel *i* in verslagperiode *t*, q_i^t de geconsumeerde hoeveelheid van artikel *i* in verslagperiode *t* en $w_i^t = q_i^t p_i^t / \sum_i q_i^t p_i^t$ het gewicht van het enkelvoudige prijsindexcijfer $I_i^{t,0} = p_i^t / p_i^0$ van artikel *i* in het Paasche-prijsindexcijfer. We zien hier de gelijkenis tussen de formules (3.10) en (3.5), analoog aan de in paragraaf 2.5 beschreven gelijkenis tussen de CMF en het Laspeyres-prijsindexcijfer.

Prijzen kunnen ook in de ruimte (tussen landen) worden vergeleken i.p.v. in de tijd. Dergelijke (internationale) ‘koopkrachtpariteiten’ worden zowel met directe als indirecte standaardisatie bepaald, waarna een gemiddelde van beide wordt genomen, iets wat in het algemeen bij standaardisatiemethoden weinig wordt toegepast, vermoedelijk omdat het de interpretatie bemoeilijkt.

3.5.2 Standaardiseren van nominale variabelen

In deelparagraaf 2.5.2 hebben we de situatie besproken waarin de totale sterfte gesplitst wordt naar doodsoorzaak. Per doodsoorzaak gelden dan dezelfde formules als voor ‘sterfte totaal’. Omdat bij de SMR, net als bij de CMF, zowel de tellers als de noemers over de doodsoorzaken optellen tot resp. teller en noemer voor alle doodsoorzaken tezamen, is ook de totale SMR een gewogen som van de SMR’s per doodsoorzaak. De verwachte (expected) aantallen doden per doodsoorzaak vormen de gewichten.

3.5.3 Vergelijking SMR's van twee populaties met de standaard

In paragraaf 3.2 is al vermeld dat het onderling vergelijken van twee populaties door middel van indirecte standaardisatie problematisch is. Fleiss (1973, p. 161) geeft een voorbeeld van twee populaties *j* en *j'* met exact dezelfde leeftijdsspecifieke sterftecijfers, dus $Y_{ij}=Y_{ij'}$, maar met verschillende leeftijdsverdelingen. Dat leidt tot SMR{*j*:*s*} ≠ SMR{*j'*:*s*}; zie bijvoorbeeld formule (3.4). Er is bij indirecte standaardisatie dus geen volledige correctie voor leeftijdsverschillen tussen de populaties (Rothman, 1986). Maar in de praktijk zijn de verschillen doorgaans klein. De twee SMR's zijn wel gelijk wanneer *s* de vereniging is van de populaties *j* en *j'*. Dan is immers $Y_{ij}=Y_{ij'}=Y_{is}$ en zijn beide SMR's gelijk aan 1. Algemener geldt dat wanneer $Y_{ij}/Y_{ij'}$ constant is, de verhouding van de SMR's gelijk is aan die constante. Formule (2.4) toont dat wanneer $Y_{ij}=Y_{ij'}$ wel altijd geldt dat CMF{*j*:*s*} = CMF{*j*:*s*}, omdat een vaste standaard wordt gebruikt.

3.6 Kwaliteitsindicatoren

Zie paragraaf 2.6.

4. Regressie-analyse

4.1 Korte beschrijving

Bij regressie-analyse wordt een afhankelijke variabele *Y* geschreven als functie van één of meer verklarende variabelen. Veelal gebruikt men het lineaire (additieve) regressiemodel $Y=X\beta+\varepsilon$. Voor iedere in het model opgenomen *X*-variabele geeft de bijbehorende *ß*-parameter de invloed op *Y*, na correctie voor de invloed van de overige *X*-variabelen op *Y*. Nominale verklarende variabelen kunnen worden meegenomen door het creëren van dummy-variabelen. Met regressie-analyse kan men dus effecten voor elkaar corrigeren, dus ook het effect van een populatie *j* op het sterftecijfer voor leeftijdsinvloeden. Men kan ook gemiddelden voor de populaties berekenen, gecorrigeerd voor dergelijke leeftijdsinvloeden. Wanneer Leeftijd en/of andere verstorende kenmerken in klassen worden ingedeeld, zouden we dit 'regressie-standaardisatie' kunnen noemen. Omdat de verklarende variabelen in klassen zijn ingedeeld, kunnen we ook spreken van variantie-analyse i.p.v. regressie-analyse. Het gaat dan om een variantie-analyse van Sterfte op Leeftijd (klassen *i*) en Populatie (klassen *j*).

4.2 Toepasbaarheid

In het geval van meerdere verstorende kenmerken worden bij directe en indirecte standaardisatie automatisch alle interacties tussen deze kenmerken meegenomen. Combinaties van deze kenmerken kunnen als één (product)variabele worden opgevat. Regressie-analyse kan ook modellen zonder interacties aan. Deze modellen zijn veel zuiniger in het aantal parameters. Ook zouden kwantitatieve *X*-variabelen (covariaten) in de regressievergelijking kunnen worden meegenomen. In die zin kan regressie-analyse meer dan standaardisatie. Maar standaardisatiemethoden zijn voor ons doel (paragraaf 1.1) inzichtelijker. Dit komt vooral doordat bij directe en indirecte standaardisatie iedere populatie *j* rechtstreeks (paarsgewijs) met populatie *s* wordt vergeleken en omdat daarbij één leeftijdsverdeling, N_{ij} of N_{is}, wordt gehanteerd. Bij 'regressie-standaardisatie' worden in het geval van een sompopulatie alle (twee of meer) beschouwde populaties gezamenlijk geanalyseerd en zijn de leeftijdsverdelingen van alle populaties van invloed op het eindresultaat. Dat dit minder inzichtelijk is, volgt al uit het feit dat er een inverse moet worden berekend om resultaten te verkrijgen. Intuïtief lijkt regressie-analyse dus op standaardisatie, zoals uit paragraaf 4.1 blijkt. In paragraaf 4.5 gaan we verder in op de vergelijking tussen regressie-analyse en standaardisatie. Dan zien we dat directe standaardisatie dan overeenkomt met een gewogen vorm van regressie-analyse (paragraaf 4.5.1)

Als alle interacties worden meegenomen, kunnen de regressie-analyses geaggregeerd plaatsvinden, net als bij directe en indirecte standaardisatie. Het maakt dan dus niet uit of de doelvariabele *Y* met scores Y_{ijk} kwantitatief is of binair

(dummy-variabele). In het laatste geval heeft men logistische regressie als alternatief; zie paragraaf 4.5.2.

4.3 Uitgebreide beschrijving

We kunnen bij de regressie-analyse enerzijds uitgaan van de situatie waarin de standaardpopulatie *s* de vereniging is van alle beschouwde populaties *j*, zodat formule (1.2) van toepassing is. Dit betekent dat alle populaties *j* onderling worden vergeleken, maar impliciet ook worden vergeleken met hun complement *s*\j. Anderzijds kan populatie *s* een externe populatie zijn waar populatie *j* geen onderdeel van is. In dat geval kan een regressie-analyse worden uitgevoerd waarbij alleen populatie *j* en *s* worden meegenomen, met $j \cup s$ als sompopulatie. In deze paragraaf gaan we qua notatie uit van de eerste situatie, waardoor bijvoorbeeld N_{is} de populatie-omvang is in leeftijdsklasse *i* over alle populaties *j*. In paragraaf 4.4 tonen we een voorbeeld met de tweede situatie.

Omdat regressie-analyse een individueel model is, gebruiken we hier de notatie op individueel niveau, zoals reeds is ingevoerd in paragraaf 1.5. D_{ij} is dus een dummy-variabele (Sterfte) met score D_{ijk}=1 als individu *k* uit leeftijdsklasse *i* en populatie *j* in de onderzoeksperiode is overleden, en anders D_{ijk}=0 (*i*=1,…,*I*; *j*=1,…,*J*). We kunnen de notatie Y_{ijk} gebruiken in plaats van D_{ijk} om de theorie te generaliseren tot kwantitatieve variabelen. Zoals in paragraaf 1.5 is vermeld, geldt voor dummy variabelen dat $Y_{ijk} = D_{ijk}$, omdat N_{ijk}=1.

Variantie-analyse van *Y* op de kwalitatieve variabelen Leeftijd en Populatie kan worden weergeven als

$$Y_{ijk} = D_{ijk} = \beta_0 + \sum_i \beta_i^L X_{ik}^L + \sum_j \beta_j^P X_{jk}^P + \varepsilon_{ijk} \quad . \tag{4.1}$$

Hierbij is X_i^L de dummy-variabele voor leeftijdsklasse *i* (of meer algemeen voor de *i*e klasse van het verstorende kenmerk) en X_j^P voor de *j*e populatie. Dus $X_{ik}^L = 1$ als individu *k* in de *i*e leeftijdsklasse zit (en anders $X_{ik}^L = 0$) en $X_{jk}^P = 1$ als *k* tot de *j*e populatie hoort (en anders $X_{jk}^P = 0$). Verder zijn de ß's regressiecoëfficiënten en de ε_{ijk} storingen.

Ter identificatie van de parameters leggen we restricties op. We nemen daarvoor de restricties die bij een multipele-classificatie-analyse[5] (MCA) worden toegepast:

[5] Multipele-classificatie-analyse (MCA) is een procedure van SPSS die alleen in de syntax-mode kan worden uitgevoerd, via het ANOVA-commando. De MCA-parameters conform restricties (4.2) en (4.3) kunnen echter ook achteraf eenvoudig worden herleid uit een variantie-analyse waarbij andere restricties zijn gehanteerd. In paragraaf 4.4 wordt dit aan de hand van een voorbeeld toegelicht.

$$\sum_i \sum_j \sum_k b_i^L = \sum_i \sum_j N_{ij} b_i^L = \sum_i N_{is} b_i^L = 0 \qquad (4.2)$$

en

$$\sum_i \sum_j \sum_k b_j^P = \sum_i \sum_j N_{ij} b_j^P = \sum_j N_{+j} b_j^P = 0 \ . \qquad (4.3)$$

Er is dus geen specifieke referentiecategorie, maar de parameters zijn voor iedere variabele (verstorend kenmerk en populatie) gewogen gemiddeld nul over de categorieën, waarbij wordt gewogen met de frequenties.

Door deze restricties wordt $ß_0$ gelijk aan μ, de verwachting van het aantal doden, en zijn de ß-parameters afwijkingen hiervan. We kunnen nu $\mu + b_j^P$ zien als de verwachte (gemiddelde) score voor populatie j na correctie voor de verstorende kenmerken. De schatter hiervan kunnen we beschouwen als het via regressie-analyse gestandaardiseerde gemiddelde (sterftecijfer).

Minimalisatie van $\sum_i \sum_j \sum_k e_{ijk}^2 = \sum_i \sum_j N_{ij} e_{ij}^2$, de kwadratensom van de residuen, leidt tot restricties voor de parameterschatters b_i^L en b_j^P die analoog zijn aan (4.2) en (4.3). Hieruit volgt dat het algemeen gemiddelde gelijk is aan

$$\hat{\mu} = \bar{Y}_{\bullet} = \bar{D}_s = \frac{1}{N_s} \sum_i \sum_j \sum_k D_{ijk} \ . \qquad (4.4)$$

Het regressie-gestandaardiseerd gemiddelde sterftecijfer kan nu worden gedefinieerd als

$$Y_j^{REGR} = \hat{\mu} + b_j^P = \bar{Y}_{\bullet} + b_j^P \ . \qquad (4.5)$$

In plaats van uit te gaan van formule (4.1), die gebaseerd is op individuele waarnemingen, kunnen we de parameters ook schatten op geaggregeerd niveau,

$$Y_{ij} = \bar{D}_{ij} = \mu + \sum_i b_i^L X_i^L + \sum_j b_j^P X_j^P + e_{ij} \ . \qquad (4.6)$$

waarbij $e_{ij} = \frac{1}{N_{ij}} \sum_k e_{ijk}$.

Minimalisatie van $\sum_i \sum_j N_{ij} e_{ij}^2$ naar de parameters leidt onder dezelfde restricties tot dezelfde parameterschatters.

Bij regressie-analyse is het vrij ongebruikelijk om het gemiddelde $\hat{\mu}$ bij een geschatte parameter op te tellen, terwijl dat bij gestandaardiseerde gemiddelden juist wel gebeurt. Binnen SPSS – General Linear Model bestaat een met (4.5) vergelijkbare mogelijkheid, nl. de procedure ‘Estimated Marginal Means’, maar dat is een weinig bruikbare optie.

4.4 Voorbeeld

Sterftecijfers Turkse en Nederlandse mannen, 0-44 jaar: regressie-standaardisatie
We passen regressie-analyse toe op twee populaties: de Turkse mannen (*j*) en de Nederlandse mannen (*s*). De sompopulatie is derhalve $j \cup s$ met totale omvang $N_{ij}+N_{is}$ in leeftijdsklasse *i*. We gebruiken hiervoor de aantallen uit de kolommen (2) en (3) van tabel A; de kolommen (6) en (7) uit die tabel geven aan hoeveel van hen een score 1 hebben op de doelvariabele Sterfte; de anderen hebben score 0.

Schatting van formule (4.1) onder de MCA-restricties (4.2) en (4.3) geeft $\hat{\mu}$ = 0,000879 en $\beta^P_{Turk} = 0{,}000548$, zoals blijkt uit kolom 1 van tabel 1. Het met regressie gestandaardiseerde sterftecijfer voor de Turkse mannen is dus gelijk aan $\hat{\mu} + \beta^P_{Turk} = 0{,}00143$. Bij directe en indirecte standaardisatie vonden we respectievelijk $Y_j^{DIR} = 0{,}00137$ en $Y_j^{INDIR} = 0{,}00141$, wat nauwelijks scheelt. Merk op dat bij indirecte en regressie-analyse dezelfde gewichten N_{ij} worden gebruikt, d.w.z. dat alle individuen uit populatie *j* even zwaar meetellen (zie paragraaf 4.5.2). Alleen hebben we bij de regressie-standaardisatie in dit voorbeeld noodgedwongen iets andere overall-gewichten, omdat we hier te maken hebben met de vereniging van Turkse en Nederlandse mannen $N_{ij}+N_{is}$. Omdat N_{is} groot is ten opzichte van N_{ij} maakt dat in dit voorbeeld weinig uit.

Tabel 1. Parameterschatters voor regressie-standaardisatie met als populaties de Turkse mannen (j) en Nederlandse mannen (s)

Variabele	Parameterschatters	
	$\sum_{i=1}^{6} N_{ij\cup s}\hat{\beta}_i^L = \sum_{j=1}^{2} N_{+j}\hat{\beta}_j^P = 0$ (MCA-restricties)	$\hat{\beta}_6^L = \hat{\beta}_2^P = 0$ (referentiecategorieën)
Constante term	0,000879	0,00162
Leeftijd		
0	0,003464	0,002720
1-4	-0,000358	-0,001103
5-14	-0,000610	-0,001355
15-24	-0,000151	-0,000896
25-34	-0,000052	-0,000797
35-44	0,000745	0
Populatie		
Turken	0,000548	0,000556
Nederlanders	-0,000008	0

De laatste kolom van tabel 1 geeft de parameterschatters van een variantie-analyse waarbij de laatste categorie van Leeftijd en Populatie worden weggelaten. Zoals in voetnoot 4 is vermeld, kunnen we de parameterschatters van de MCA-oplossing hier eenvoudig uit berekenen. Dit kan door per variabele het met de aantallen personen

gewogen gemiddelde van de parameters af te trekken, en de afgetrokken waarden vervolgens bij de constante term weer op te tellen. Merk op dat de verschillen tussen geschatte parameters behorende bij dezelfde variabelen in kolom 1 en 2 gelijk zijn.

4.5 Eigenschappen

4.5.1 Directe standaardisatie en regressie

In Israëls en De Ree (1981) wordt getoond dat het direct gestandaardiseerde sterftecijfer Y_j^{DIR} kan worden verkregen als resultaat van een *gewogen* lineaire regressie voor het geval we een aantal populaties *j* onderling en met hun sompopulatie *s* vergelijken. Directe standaardisatie hanteert hetzelfde additieve model (4.6) als regressie-analyse, maar een andere verliesfunctie voor het schatten van de ß-parameters. Geminimaliseerd wordt $\sum_i \sum_j N_{is} e_{ij}^2$ in plaats van $\sum_i \sum_j N_{ij} e_{ij}^2$. Dit betekent dat de gekwadrateerde afwijkingen e_{ij}^2 niet worden gewogen met hun werkelijke celfrequenties N_{ij}, maar met de frequenties N_{is} uit de standaardpopulatie. De restricties uit (4.2) en (4.3) worden hierdoor aangepast. Restrictie (4.3) wordt nu $\sum_j b_j^P = 0$, hetgeen eenvoudig is te zien door N_{ij} te vervangen door N_{is}; de gewichten hangen nu niet meer van *j* af. Met deze verliesfunctie en restricties wordt de schatter voor $m + b_j^P$ gelijk aan het direct gestandaardiseerd gemiddelde.

Omdat bij (ongewogen) regressie-analyse in het geval van een sompopulatie *s* alle individuen even zwaar meetellen, heeft 'regressie-standaardisatie' kleinere standaardfouten dan directe standaardisatie en leidt dus tot efficiëntere schatters voor ß, indien het model geldt. Waar bij regressie-standaardisatie de parameters voor alle populaties *j* simultaan moeten worden geschat, kan dit bij directe standaardisatie ook per populatie *j* geschieden. Dit maakt directe standaardisatie eenvoudiger en doorzichtiger. In het geval van een externe populatie komt regressie-analyse met de bijbehorende gewichten overeen met directe standaardisatie.

4.5.2 Indirecte standaardisatie en regressie

In tegenstelling tot directe standaardisatie hanteert indirecte standaardisatie N_{ij} als weging van de sterftecijfers Y_{ij}, zoals te zien is in de formules van hoofdstuk 3. Alle individuen tellen dus even zwaar mee. In dit opzicht lijkt indirecte standaardisatie dus meer dan directe standaardisatie op regressie-standaardisatie. Anderzijds is indirecte standaardisatie niet gebaseerd op een additief model voor de sterftekans, maar op een multiplicatief model, zoals we hebben gezien in formule (3.3).

We lieten daar zien dat het model achter de SMR kan worden beschouwd als een multiplicatieve regressie van Y_{ij} op de variabelen Leeftijd en Populatie. We kunnen formule (3.3b), $E(Y_{ij}) = m_j = j_i q_j$, analoog aan formule (4.1) en (4.6) noteren met

parameters μ, β_i^L en β_j^P in plaats van φ_i en θ_j. Als we de standaardpopulatie als referentiegroep *j'* (= *s**j*) nemen, wordt β_j^P de verhouding in sterftequotiënt tussen populatie *j* en de referentiegroep, gecorrigeerd voor leeftijd. In het algemeen (ook voor kwantitatieve *Y*-variabelen) zouden we de parameters kunnen schatten door links en rechts van het gelijkteken de logaritmen te nemen en de kleinste-kwadratenmethode toe te passen. Maar als de aantallen doden $\underline{D}_{ij}$ Poisson-verdeeld zijn met verwachtingsparameters $N_{ij}\mu_{ij} = N_{ij}\varphi_i\theta_j$, dan is Poisson-regressie met maximum-likelihood-schating mogelijk. Deze ML-schatter wijkt overigens wat af van SMR{*j*:*s*}, zoals Breslow en Day (1975) aantonen. Hetzelfde verschijnsel doet zich voor bij toepassing van logistische regressie.

Als *Y* binair is, zoals bij de variabele Sterfte (*D*), is logistische regressie op Leeftijd en Populatie een alternatief, waarbij we de populaties *j* en *s**j* onderscheiden. We hebben dat uitgevoerd voor het voorbeeld uit paragraaf 3.4. Het logistische-regressiemodel luidt: $\ell n\{p/(1-p)\} = X\beta$, ofwel $p/(1-p) = \exp(X\beta)$ waarbij $\exp(\beta_j^P)$ de invloed op de kansverhouding wel/geen sterfte weergeeft van populatie *j* t.o.v. de referentiepopulatie *j'*, na correctie voor leeftijd. Als (1-*p*) bijna 1 is, is dit een redelijke benadering voor de SMR. ML-schatting geeft $\hat{\beta}_j^P = 1{,}616$, wat afgerond identiek is aan de SMR in voorbeeld 3.4, en leidt tot vrijwel hetzelfde betrouwbaarheidsinterval. Wanneer de standaardpopulatie groot is en *p* heel klein blijkt dit dus een goede benadering. Zou de standaardpopulatie veel kleiner zijn, dan verandert $\hat{\beta}_j^P$ van waarde, wat ongewenst is wanneer we echt willen standaardiseren. De SMR blijft in dat geval vanzelfsprekend gelijk aan 1,616.

5. Vergelijking van de methoden

We hebben in dit themarapport de twee bekendste standaardisatiemethoden getoond, directe en indirecte standaardisatie. Als index voor de verhouding van de sterfte tussen beschouwde en standaardpopulatie betreft dit respectievelijk de CMF en SRM. De indices zijn theoretisch gezien gelijk wanneer de verhouding van de leeftijdsspecifieke (sterfte)cijfers in de beschouwde populatie *j* en standaardpopulatie *s* constant is, d.w.z. $Y_{ij}=c.Y_{is}$ voor $i=1,\ldots,I$, oftewel $R_i=c$. De weging van deze leeftijdsspecifieke sterfteratio's is dan irrelevant. Grote verschillen tussen de CMF en SRM ontstaan pas wanneer de R_i sterk heterogeen zijn. Maar in dat geval ligt het meer voor de hand om leeftijdsspecifieke sterftecijfers te publiceren dan gestandaardiseerde cijfers. Tot deze laatste zal men dan overgaan wanneer de cijfers per se geleverd moeten worden of wanneer er voor vele verschillende populaties gestandaardiseerde cijfers moeten worden geleverd (bijvoorbeeld voor alle 20 doodsoorzaken uit een bepaalde classificatie of voor alle Europese landen).

CMF en SMR zijn vanzelfsprekend ook gelijk wanneer *j* en *s* dezelfde leeftijdsverdeling hebben ($q_{i|j}=q_{i|s}$). Er hoeft dan niet voor leeftijd te worden gecorrigeerd. Tenslotte zijn ze in verwachting aan elkaar gelijk wanneer de verschillen in sterfteratio en leeftijdsverdeling ongecorreleerd zijn.

Kilpatrick (1962) ziet de CMF en SMR als schatters voor dezelfde parameter, uitgaande van perfecte homogeniteit van sterfteratio's. In dat geval is aan te geven in welke situatie de CMF efficiënter is dan de SMR, en omgekeerd; zie ook Van der Maas en Habbema (1981). Het kan dan efficiënt zijn om een, al dan niet gewogen, gemiddelde van beide te nemen. Praktisch gezien zal de homogeniteit echter nooit exact gelden en is zo'n gecombineerde index voor beschrijvende statistiek minder doorzichtig.

Regressie-analyse is in die zin wat problematischer in het feit dat wanneer er meerdere populaties zijn met de sompopulatie als standaard, de standaardisatie ondoorzichtiger is. Directe en indirecte standaardisatie kunnen altijd paarsgewijs (*j*,*s*) worden bepaald, terwijl het regressieresultaat ook afhangt van de andere populaties, *j'*. Wanneer er maar één populatie *j* met de standaard *s* wordt vergeleken, moet er een sompopulatie worden gecreëerd om een ongewogen regressie-analyse te kunnen uitvoeren.

6. Literatuur

Breslow, N.E. en Day, N.E. (1975), Indirect standardization and multiplicative models for rates, with reference to the age adjustment of cancer incidence and relative frequency data. *Journal of Chronical disease* 28, 289-303.

Breslow, N.E. en Day, N.E. (1987), *Statistical methods in cancer research, Volume II – The design and analysis of cohort studies*. International Agency for Research on Cancer, Scientific Publications no. 82, Lyon.

Chiang, C.L. (1961), Standard errors of the age-adjusted death rate. *USGPO, Vital statistics* 47, 275-285.

Chiang C.L. (1984), *The life table and its applications*. Robert Krieger Publishing Company, Malabar, Florida.

Fleiss, J.L. (1973), *Statistical methods for rates and proportions*. Wiley, New York.

Grient, H.A. van der en Haan, J. de (2008), *Indexcijfers*. Rapport Methodenreeks, CBS, Voorburg.

Hoogenboezem, J. en Israëls, A.Z. (1990), Sterfte naar doodsoorzaak onder Turkse en Marokkaanse ingezetenen in Nederland, 1979-1988. *Maandbericht Gezondheidsstatistiek* 9 (8), 5-20.

Israëls, A.Z. en Ree, S.J.M. de (1981), *Standaardisatie, met toepassing op het Loonstructuuronderzoek 1972*. Interne nota, CBS, Voorburg.

Keiding, N. (1987), The method of expected number of deaths, 1786–1886–1986. *International Statistical Review* 55, 1-20.

Kilpatrick, S.J. (1962), Occupational mortality indices. *Population studies* 16, 175-187.

Maas, P.J. van der en Habbema, J.D.F. (1981), Standaardiseren van ziekte- en sterftecijfers: mogelijkheden en beperkingen. *Tijdschrift voor Sociale Geneeskunde* 59 (8), 259-270.

McCullagh, P. en Nelder, J.A. (1989), *Generalized linear models*. Chapman and Hall, Londen.

Meulen, A. van der (2009), *Overlevingstafels*. Rapport Methodenreeks, CBS, Den Haag.

Molenaar, W. (1973), *Simple approximations to the Poisson, binomial and hypergeometric distributions*. Mathematisch Centrum, Report SW 9/73, Amsterdam.

Ree, S.J.M. de en Israëls, A.Z. (1982), *Een noot over het gebruik van standaardisatie bij nominale variabelen*. Interne nota, CBS, Voorburg.

Rothman, K.J. (1986), *Modern epidemiology*. Little Brown and Co, Boston.

Yule, G.U. (1934), On some points relating to vital statistics, more especially statistics on occupational mortality. *Journal of the Royal Statistical Society* 97, 1-84.

Bijlage 1. Toets op gelijkheid van leeftijdsspecifieke sterftekansen

We willen toetsen of twee leeftijdsspecifieke sterftekansen, μ_{ij} en μ_{is} (of μ_{ij} en $\mu_{ij'}$) gelijk zijn voor zekere i. Deze hypothese houdt in dat hun schatters $Y_{ij} = D_{ij} / N_{ij}$ en $Y_{is} = D_{is} / N_{is}$ niet te sterk verschillen. We kunnen de hypothese ook formuleren als $\mu_{ij} / \mu_{is} = 1$, waarvoor $R_i = Y_{ij} / Y_{is}$ de toetsingsgrootheid is.

Om het algemeen te houden, gaan we er vanuit dat D_{ij} en D_{is} aan het toeval onderhevige grootheden zijn en gaan dus uit van een superpopulatie. We nemen aan dat D_{ij} en D_{is} binomiaal zijn verdeeld, met sterftekansen respectievelijk p_{ij} en p_{is}. We kunnen nu onze hypothese toetsen met de 'exacte toets van Fisher'. De hiervoor benodigde grootheden, of eigenlijk de realisaties ervan, staan in tabel A; het aantal overlevenden wordt genoteerd met '*L'*. Gegeven de totaal-rij en totaal-kolom van deze tabel is het aantal overledenen D_{ij} dan hypergeometrisch verdeeld met parameters N_{ij}, D_{i+} en N_{i+}. Als D_{ij} te sterk afwijkt van $N_{ij} D_{i+} / N_{i+}$ leidt dit tot verwerping van de hypothese $\mu_{ij} = \mu_{is}$.

Tabel A. Frequentietabel voor exacte toets van Fisher

	Aantal doden	Aantal overlevenden	Totaal
j (Turken)	D_{ij}	$L_{ij} = N_{ij}$-D_{ij}	N_{ij}
s (Nederlanders)	D_{is}	$L_{is} = N_{is}$-D_{is}	N_{is}
Totaal	D_{i+}	$L_{i+} = N_{i+}$-D_{i+}	N_{i+}

De kans dat de hypergeometrisch verdeelde D (onbekend aantal doden in populatie *j*) kleiner of gelijk is aan de realisatie D_{ij}, $P[D \le D_{ij}]$, kan worden benaderd door een Poisson-verdeling

$$P[D \le D_{ij}] \approx P[v \le D_{ij} \mid \mu = \frac{(2N_{ij} - D_{ij})(2D_{i+} - D_{ij})}{2(N_{is} + L_{i+} + 1)}]$$

met *v* een Poisson-verdeelde grootheid met parameter μ. Een tweede benadering voor $P[D \le D_{ij}]$ is

$$P[D \le D_{ij}] \approx P[u \le 2\frac{\sqrt{(D_{ij} + 1)(L_{is} + 1)} - \sqrt{L_{ij} D_{is}}}{\sqrt{N_{i+} - 1}}] \ .$$

met u een standaard-normaal verdeelde grootheid. De Poisson-benadering kan het beste worden gehanteerd als μ kleiner is dan 30. Zie Molenaar (1973) voor meer informatie over deze benaderingen. Indien de steekproefomvang groot genoeg is, kan in plaats van de hypergeometrische toets voor tabel A ook een Chi-kwadraat-toets worden gebruikt.

Voor het simultaan toetsen van $\mu_{ij} = \mu_{is}$ voor *alle* leeftijdsklassen (i=1,…,I), kan men de Chi-kwadraat-toets gebruiken voor een Ix2-tabel, met als kolommen het aantal doden en aantal overlevenden. Voor kwamtitatieve afhankelijke variabelen kunnen t- en F-toetsen worden gebruikt.

Bijlage 2. Toepassing standaardisatie op sterftecijfers van Turkse (j) t.o.v. Nederlandse (s) mannen, 1979-1986

Tabel A. Leeftijd tot 45 jaar

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
i	N_{ij}	N_{is}	$q_{i\|j}$	$q_{i\|s}$	D_{ij}	D_{is}	$Y_{ij} = D_{ij}/N_{ij}$	$Y_{is} = D_{is}/N_{is}$	$R_i = Y_{ij}/Y_{is}$	$q_{i\|s}Y_{ij}$	CMF formule (2.2)	$q_{i\|j}Y_{is}$	SMR formule (3.1)
			%	%			per 10.000	per 10.000		per 10.000		per 10.000	
0	18 298	671 538	3,20	1,75	138	2 863	75,4	42,6	1,77	1,32		1,37	
1-4	68 183	2 700 772	11,94	7,05	87	1 372	12,8	5,1	2,51	0,90		0,61	
5-14	138 922	8 270 650	24,33	21,60	103	2 168	7,4	2,6	2,83	1,60		0,64	
15-24	135 924	9 777 094	23,80	25,54	158	7 054	11,6	7,2	1,61	2,97		1,72	
25-34	87 608	9 299 700	15,34	24,29	110	7 625	12,6	8,2	1,53	3,05		1,26	
35-44	122 078	7 567 950	21,38	19,77	239	12 254	19,6	16,2	1,21	3,87		3,46	
Totaal	571 013	38287 704	100	100	835	33 336	14,6	8,7	1,68	13,71	**1,575**	9.05	**1,616**

Legenda: (1) Leeftijdsklasse; (2) Omvang populatie *j* naar leeftijd; (3) Omvang populatie *s* naar leeftijd; (4) Leeftijdsverdeling *j*; (5) Leeftijdsverdeling *s*; (6) Aantal doden *j*; (7) Aantal doden *s*; (8) Leeftijdsspecifieke sterftecijfer *j*; (9) Leeftijdsspecifieke sterftecijfer *s*; (10) Sterfteratio pop. *j* t.o.v. pop. *s*; (11) berekening teller CMF; (12) CMF; (13)) berekening noemer SMR; (14) SMR

Tabel B. Alle leeftijden

(1) I	(2) N_{ij}	(3) N_{is}	(4) $q_{i\|j}$	(5) $q_{i\|s}$	(6) D_{ij}	(7) D_{is}	(8) $Y_{ij} = D_{ij}/N_{ij}$	(9) $Y_{is} = D_{is}/N_{is}$	(10) $R_i = Y_{ij}/Y_{is}$	(11) $q_{i\|s}Y_{ij}$	(12) CMF formule (2.2)	(13) $q_{i\|j}Y_{is}$	(14) SMR formule (3.1)
			%	%			per 10.000	per 10.000		per 10.000		per 10.000	
0	18 298	671 538	2,82	1,23	138	2 863	75,4	42,6	1,77	0,93		1,20	
1-4	68 183	2 700 772	10,51	4,94	87	1 372	12,8	5,1	2,51	0,63		0,53	
5-14	138 922	8 270 650	21,41	15,12	103	2 168	7,4	2,6	2,83	1,12		0,56	
15-24	135 924	9 777 094	20,95	17,87	158	7 054	11,6	7,2	1,61	2,08		1,51	
25-34	87 608	9 299 700	13,50	17,00	110	7 625	12,6	8,2	1,53	2,13		1,11	
35-44	122 078	7 567 950	18,81	13,83	239	12 254	19,6	16,2	1,21	2,71		3,05	
45-64	76 838	10 982 309	11,84	20,07	345	105 593	44,9	96,1	0,47	9,01		11,39	
65+	1 031	5 443 711	0,16	9,95	39	364 872	378,3	670,3	0,56	37,64		1,06	
Tot.	648 882	54 713 724	100	100	1 219	503 801	18,8	92,1	0,20	56,24	**0,611**	20,41	**0,920**

Legenda: (1) Leeftijdsklasse; (2) Omvang populatie *j* naar leeftijd; (3) Omvang populatie *s* naar leeftijd; (4) Leeftijdsverdeling *j*; (5) Leeftijdsverdeling *s*; (6) Aantal doden *j*; (7) Aantal doden *s*; (8) Leeftijdsspecifieke sterftecijfer *j*; (9) Leeftijdsspecifieke sterftecijfer *s*; (10) Sterfteratio pop. *j* t.o.v. pop. *s*; (11) berekening teller CMF; (12) CMF; (13)) berekening noemer SMR; (14) SMR