Dwalingen in de methodologie. XXX. De vangst-hervangstmethode

P.G.M.VAN DER HEIJDEN EN R.A.HIRASING

Wat zijn de incidentie en de prevalentie van een bepaalde aandoening? Voor sommige aandoeningen bestaan er registraties die aangeven wie een aandoening hebben (bruikbaar voor de bepaling van de prevalentie) en soms wanneer een aandoening bij een patiënt is begonnen (bruikbaar voor de bepaling van de incidentie). Men kan dan door het aantal patiënten te tellen tot een prevalentie- of incidentieschatting komen. Een dergelijke aanpak geeft echter een onderschatting: ten eerste doordat de deelname van potentiële melders aan de registratie nooit 100% is en ten tweede doordat zij vaak vergeten patiënten te melden. Het nastreven van een respons van 100% is zeer arbeidsintensief (en dus kostbaar). Bovendien geeft de hoogte van de respons geen informatie over de volledigheid van de melders. Wil men een indruk krijgen van de niet-deelname en de vergeten meldingen, dan is onderzoek via een tweede bron noodzakelijk. Hiertoe dient de vangst-hervangstmethode.

De aanduiding 'vangst-hervangst' komt oorspronkelijk uit de biologie, waar deze methode wordt gebruikt bij het schatten van de grootte van dierenpopulaties. Hierbij wordt gekeken in welke mate dieren herhaaldelijk worden gevangen. Een van de eerste voorbeelden van de vangst-hervangstmethode behelst het schatten van het aantal Noord-Amerikaanse eenden;[1] daarbij werden grote aantallen eenden gevangen en geringd, en vervolgens uitgezet op broedplaatsen. Het aantal geringde dieren in een latere vangst maakt een schatting mogelijk van de grootte van de populatie. Hoe gaat de berekening in zijn werk?

Zie ook de artikelen op bl. 164 en 175.

SAMENVATTING

– De vangst-hervangstmethode is bruikbaar voor het schatten van de incidentie en prevalentie van een aandoening indien er verschillende registraties bestaan die kunnen worden gekoppeld.
– Er moet echter aan enkele voorwaarden zijn voldaan, wil de vangst-hervangstschatting adequaat zijn.
– Zo moet voor elke registratie gelden dat elk individu dezelfde kans heeft om in de registratie opgenomen te zijn, en de registraties dienen onafhankelijk te zijn.
– In het geval van 2 registraties dient men te veronderstellen dat deze registraties onafhankelijk zijn om tot een schatting te komen. Bij 3 of meer registraties kan deze strenge veronderstelling versoepeld worden.

Berekening. Om de omvang N van een bepaalde populatie dieren te schatten, neemt men een willekeurige steekproef van n_1 dieren uit deze populatie van N die-

Universiteit Utrecht, Faculteit Sociale Wetenschappen, Capaciteitsgroep Methodenleer en Statistiek, Postbus 80.140, 3508 TC Utrecht.
Prof.dr.P.G.M.van der Heijden, statisticus.
TNO Preventie en Gezondheid, Leiden.
Dr.R.A.Hirasing, kinderarts-jeugdarts.
Correspondentieadres: prof.dr.P.G.M.van der Heijden (p.vanderheijden@fss.uu.nl).

ren. Deze n_1 dieren worden van een merkteken voorzien en weer uitgezet. Na verloop van tijd (zodat de gemerkte dieren zich goed hebben kunnen mengen met de niet-gemerkte dieren) neemt men een tweede willekeurige steekproef van n_2 dieren. Hiervan blijken er m_2 een merkteken te hebben. Op basis hiervan kan N geschat worden: als we aannemen dat de proportie gemerkten in de tweede steekproef (m_2/n_2) een perfect beeld geeft van de proportie gemerkten in de populatie (n_1/N), dan geldt dat $m_2/n_2 = n_1/N$, en dus schatten we N met $(n_1 \times n_2)/m_2$.

Deze aanpak is typisch voor de biologie, waar men meer dan één steekproef trekt uit dezelfde populatie. In de epidemiologie trekt men zelden specifiek voor dit doeleinde meerdere steekproeven; men maakt eerder gebruik van bestaande registraties en onderzoekt de overlap hiertussen. Wij laten aan de hand van het tweede onderzoek naar de incidentie van diabetes mellitus bij 0-19-jarigen in Nederland, gepubliceerd in dit tijdschrift,[2] zien welke oplossing de vangst-hervangstmethode levert, en hoe vooronderstellingen hierin een rol spelen.

Voorbeeld. Voor de berekening van de incidentie van diabetes beschikte een van ons (R.A.H.) over 2 registraties, namelijk die uit een enquête onder alle kinderartsen en internisten in Nederland (de KIN-registratie) en die uit een enquête onder leden van de Diabetes Vereniging Nederland (de DVN-registratie).[2] Wij concentreren ons hier op de schatting van de incidentie bij de subgroep 0-4-jarigen van 1988-1990; wij duiden deze groep aan als 'de populatie'. Er waren 160 kinderen bekend bij de KIN-registratie en 110 bij de DVN-registratie. De overlap tussen beide registraties was 99; dat wil zeggen dat 99 kinderen in beide registraties voorkwamen. Gebruik van de DVN-registratie leverde dus 11 nieuwe kinderen op vergeleken met de KIN-registratie (tabel).

Als we de KIN-registratie beschouwen als de eerste steekproef en de DVN-registratie als de tweede, is $n_1 = 160$, $n_2 = 110$, $m_2 = 99$, en $N = 171 + ?$. De beschreven methode levert als schatting van de populatieomvang N van 0-4-jarigen met diabetes mellitus $n_1n_2/m_2 = 177{,}8$ op, en het vraagteken (het aantal kinderen dat in geen van beide registraties zit) moet dus 6,8 zijn (95%-betrouwbaarheidsinterval: 3,5-13,2). De proportie van de 177,8 kinderen die in minimaal 1 registratie bekend is, ook bekend onder de naam 'ascertainment', is gelijk aan $171/177{,}8 \times 100 = 96\%$.

Vooronderstellingen. Wat zijn de vooronderstellingen van de vangst-hervangstmethode bij dit voorbeeld?[3]
– De populatie waarvan de omvang bepaald wordt, is 'gesloten', dat wil zeggen verandert niet van omvang tijdens de periode van onderzoek.
– Individuen kunnen gekoppeld worden in de 2 registraties, dat wil zeggen dat men van een individu kan vaststellen of het in de beide registraties voorkomt.
– Voor elke registratie afzonderlijk heeft elk individu dezelfde kans om in de registratie terecht te komen.
– De 2 registraties zijn onafhankelijk.

Laten we nu de vooronderstellingen nader bekijken waaronder de schatting van 177,8 juist is. Voor de incidentie van diabetes lijkt de eerste vooronderstelling (gesloten populatie) geen probleem.

Voor de tweede vooronderstelling (koppeling) is het van belang te weten op basis van welke variabelen de koppeling heeft plaatsgevonden. Kunnen er ten onrechte koppelingen wel of niet tot stand zijn gebracht? De koppeling vond in dit geval plaats op basis van het geslacht, de geboortedatum, de datum van de eerste insuline-injectie en de toenmalige woonplaats. Door de gedetailleerdheid van deze variabelen lijkt het vrijwel onmogelijk dat er hier iets is misgegaan. Echter, men kan zich voorstellen dat indien alleen de geboortedatum beschikbaar zou zijn geweest, er vaker koppelingen konden mislopen. Wat is hiervan de invloed? Stel bijvoorbeeld eens dat 1 kind ten onrechte niet gekoppeld is. Dit leidt dan tot een dubbeltelling, die nog eens extra doorwerkt in de schatting: de totale schatting wordt 172 + 7,0 = 179,0. Een schrikbarende toename is 1,2 niet, maar bij registraties die moeilijk te koppelen zijn, kan dit verschil natuurlijk behoorlijk oplopen.

De laatste twee vooronderstellingen (elk individu heeft dezelfde registratiekans respectievelijk de 2 registraties zijn onafhankelijk) hangen nauw samen, want de eerste geeft aan dat elk individu dezelfde kans heeft op aanwezigheid in de tweede registratie, of men nu in de eerste registratie zit of niet, en daaruit volgt de laatste vooronderstelling. Toch hebben wij gemeend deze vooronderstellingen niet samen te moeten voegen, omdat hierover in de literatuur geen consensus bestaat; men tracht momenteel deze twee vooronderstellingen verder uit elkaar te trekken met behulp van de begrippen 'afhankelijkheid' en 'schijnbare afhankelijkheid'.[3]

De vooronderstelling van onafhankelijkheid is meestal de lastigste, omdat de juistheid ervan bij 2 bronnen niet te verifiëren is. Bij veel aandoeningen kan ook de ernst van de ziekte leiden tot schending van deze vooronderstelling. Bij astma is het bijvoorbeeld waarschijnlijk dat juist kinderen die hier meer last van hebben zowel vaker bekend zijn bij een specialist als vaker lid zijn van een patiëntenvereniging. Door de aard van de ziekte diabetes zal ernst waarschijnlijk geen rol spelen in de kans op registratie.

In potentie kan een variabele die invloed op beide kansen heeft echter een enorm effect hebben. Hier staat

Verdeling van kinderen met diabetes in 2 registraties,[2] die als 2 steekproeven uit dezelfde populatie zijn te beschouwen; de vraagtekens geven onbekende aantallen weer, de cursieve cijfers zijn de bekende gegevens waarop de schatting van de totale populatiegrootte werd gebaseerd

aantal kinderen	*aantal kinderen*		
	in DVN-registratie	*niet in DVN-registratie*	*totaal*
in KIN-registratie	*99**	61	*160*
niet in KIN-registratie	11	?	11 + ?
totaal	*110*	61 + ?	171 + ?

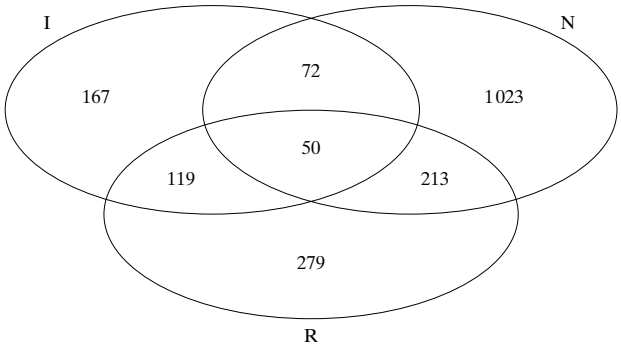
DVN-registratie = registratie uit een enquête onder leden van de Diabetes Vereniging Nederland; KIN-registratie = registratie uit een enquête onder alle kinderartsen en internisten in Nederland.
*Gekoppelde gegevens.

tegenover dat als een variabele invloed heeft op slechts één van beide kansen, dit niet bezwaarlijk is. Dit laatste wordt zelden in de literatuur gerapporteerd en is dus bij velen onbekend. Dit leidt tot de volgende conclusie: als men vermoedt dat de derde vooronderstelling geschonden is doordat voor beide registraties de kansen variëren, moet men pogen de variabele die deze vooronderstelling schendt te meten en deze te gebruiken om deelsteekproeven te creëren waarvoor de vooronderstelling dan weer opgaat. Indien slechts de kans van 1 registratie varieert, kan een dergelijke variabele genegeerd worden.

Meerdere registraties. Een andere oplossing is dat men gebruikmaakt van meer dan 2 registraties. Bij 3 registraties heeft men de beschikking over $2 \times 2 \times 2 - 1 = 7$ frequenties, waarop loglineaire modellen geschat kunnen worden. In deze loglineaire modellen kunnen interacties tussen registraties worden opgenomen. Deze interacties staan toe dat registraties afhankelijk zijn. De laatste twee vooronderstellingen worden nu vervangen door de vooronderstelling dat de kans om in elke combinatie van 2 registraties terecht te komen voor elk individu gelijk is, en dat in ieder geval de interactie tussen 3 registraties tegelijkertijd afwezig is. Dit maakt de toepassing van de vangst-hervangstmethode aanzienlijk realistischer.

Nog een voorbeeld. Als voorbeeld bespreken wij nu de schatting van het aantal psychologen dat in de individuele gezondheidszorg werkzaam is.[4] Er zijn 3 lijsten met namen van psychologen. De eerste is gebaseerd op een enquête onder gezondheidszorginstellingen (deze geven we aan met I), de tweede is afkomstig uit het ledenbestand van de beroepsvereniging van de psychologen (N), en de derde is gebaseerd op gegevens van werkgevers in de Regionale Instellingen voor Ambulante Geestelijke Gezondheidszorg (R). Na koppeling van deze lijsten hebben we de in de figuur weergegeven 7 geobserveerde frequenties; bij optelling van de getallen is de uitkomst 1923. Nu kan onderzocht worden welk loglineaire model 7 schattingen van verwachte frequenties heeft die erg lijken op de geobserveerde frequenties (bij 2 lijsten is dit niet mogelijk, want dan zijn de 3 geobserveerde frequenties (in de tabel de frequenties 99, 61 en 11) gelijk aan de 3 verwachte frequenties). Het loglineaire model dat aanneemt dat de 3 lijsten onafhankelijk zijn, past erg slecht ($\chi^2 = 137{,}8$; df = 3; p = 0,000), dus dit model wordt verworpen. Nu worden interacties toegevoegd aan dit model, en na enig zoekwerk[3] vinden we dat het model met zowel interactie tussen I en N als tussen I en R, maar zonder interactie tussen N en R, zeer goed past ($\chi^2 = 0{,}0$; df = 1; p = 0,907). Dit model levert een schatting op van 1340 extra psychologen die werkzaam zijn in de individuele gezondheidszorg en in geen van de registraties voorkomen; de totale schatting van dit type psychologen is dus 1340 + 1923 = 3263 (95%-betrouwbaarheidsinterval: 3000-3526).

In de eerdere publicatie van dit voorbeeld wordt uitgelegd hoe de analyse kan plaatsvinden met bijvoorbeeld het computerprogramma Statistical Package for the Social Sciences (SPSS).[4] In diverse andere artikelen wordt de vangst-hervangstmethode meer in detail uit de doeken gedaan.[3 5] Recentelijk is er veel aandacht besteed aan modellen waarin men de laatste twee vooronderstellingen (elk individu heeft dezelfde registratiekans respectievelijk de registraties zijn onafhankelijk) apart modelleert door gebruik te maken van latente variabelen, met als doel de vooronderstellingen realistischer te maken.[6 7]



Verdeling van namen van psychologen over 3 registratielijsten, de eerste gebaseerd op een enquête onder gezondheidszorginstellingen (I), de tweede afkomstig uit het ledenbestand van de beroepsvereniging van de psychologen (N), en de derde gebaseerd op gegevens van werkgevers in de Regionale Instellingen voor Ambulante Geestelijke Gezondheidszorg (R).

ABSTRACT

Roaming through methodology. XXX. The capture-recapture method

– The capture-recapture method is useful for the estimation of the incidence and prevalence of a disease if there exist multiple lists that can be linked.

– Some assumptions have to be fulfilled for the estimate to be adequate.

– For example, for every list every individual should have the same probability to be on the list, and the lists should be independent.

– In case of two lists one has to assume that these are independent in order to get an estimate. When three or more lists are available this strong assumption may be mitigated.

LITERATUUR

1 Seber GAF. The estimation of animal abundance and related parameters. Londen: Griffin; 1982.

2 Hirasing RA. De incidentie van diabetes mellitus bij 0-19-jarigen in Nederland (1988-1990). Ned Tijdschr Geneeskd 1995;139:1088-91.

3 International Working Group for Disease Monitoring and Forecasting. Capture-recapture and multiple-record systems estimation I: history and theoretical development. Am J Epidemiol 1995;142: 1047-58.

4 Smit F, Brunenberg W, Heijden P van der. Het schatten van populatiegroottes: toepassingen en een voorbeeld. Tijdschr Soc Gezondheidsz 1996;74:171-6.

5 Cormack RM. Problems with using capture-recapture in epidemiology: an example of a measles epidemic. J Clin Epidemiol 1999; 52:909-14, 915-33.

6 Agresti A. Simple capture-recapture models permitting unequal catchability and variable sampling effort. Biometrics 1994;50:494-500.

7 Fienberg SE, Johnson MS, Junker BW. Classical multilevel and Bayesian approaches to size estimation using multiple lists. Series A. Journal of the Royal Statistical Society 1999;162:383-405.

Aanvaard op 17 juli 2000