Tijdschr Soc Gezondheidsz 1996; 74: 171-176

Het schatten van populatiegroottes

Toepassingen en een voorbeeld

*Filip Smit, Wim Brunenberg, Peter van der Heijden **

Het aantal psychologen dat in de individuele gezondheidszorg werkt is onbekend. Met het oog op het formuleren van beleid is deze onzekerheid hinderlijk. Er zijn weliswaar verschillende registraties van psychologen, maar die hebben geen betrekking op deze specifieke populatie, overlappen elkaar gedeeltelijk en zijn niet compleet. Toch is het op basis van dit soort registraties en door het maken van bepaalde veronderstellingen mogelijk het onbekende aantal te schatten. Dit kan onder meer met behulp van zogenaamde vangst-hervangst/capture-recapture-methoden. We voeren op drie verschillende manieren zo'n schatting uit. De resultaten zijn met elkaar in overeenstemming en wijzen op een aantal van circa 3263 psychologen van de bedoelde categorie. Onze studie illustreert dat omvangschattingen gebaseerd op capture-recapture-data succesvol gebruikt kunnen worden binnen de sociale wetenschappen. Wij verwachten dat het gebruik van dit soort schattingen zal toenemen. Er zijn tal van toepassingsgebieden. Gedacht kan worden aan het aantal daklozen, verslaafden, pathologische gokkers, prostituée's, HIV-geïnfecteerden, verwaarloosde ouderen en de omvang van criminele populaties. Deze en andere toepassingen rechtvaardigen een brede belangstelling voor vangst-hervangstmethoden.

Trefwoorden: verborgen populaties, omvangsschattingen, vangst-hervangstmethoden, incomplete registraties

Het schatten van de onbekende omvang van een populatie wordt in vele takken van de wetenschap toegepast. Voorbeelden uit sociaal-wetenschappelijk onderzoek ten behoeve van planningsdoeleinden zijn schattingen van het aantal autodieven[1], het aantal daklozen[2,3], het aantal prostituées[4,5], het aantal druggebruikers[4,6] en het aantal bezitters van illegale vuurwapens in Nederland.[7]

Omvangsschattingen kunnen ook dienst doen als controle op de compleetheid van gegevens die op hun beurt van belang zijn voor wetenschappelijk onderzoek. Zo gebruikt men omvangsschattingen om uit te rekenen hoeveel artikelen (nog) gemist zijn door reviewers die op basis van een verzameling artikelen een meta-analyse willen uitvoeren. In de medische epidemiologie wil men weten in hoeverre registraties van kankergevallen compleet zijn[8] of hoeveel gevallen van het Down's syndroom niet opgemerkt zijn. Demografen controleren soms de mate van dekking van een census met behulp van omvangsschattingen.

Met name biologen die zich bezighouden met wildstand en visserij hebben inmiddels vele statistische technieken ontwikkeld voor het uitvoeren van omvangsschattingen op basis van vangst-hervangstfrequenties. Voor enkele overzichtswerken verwijzen wij naar publicaties van Seber, Pollock, Wilson en Collins, en El-Khorazaty e.a.[9-14] Eerder publiceerden wij een overzicht van vangst-hervangstmethoden met het oog op toepassingen binnen de criminologie.[15]

Een belangrijke voorwaarde voor de toepassing van omvangsschattingen is de beschikbaarheid van vangst-hervangstdata. Dit type data kan op verschillende manieren verkregen worden. Biologen verkeren wel eens in de positie waarin het mogelijk is om op opeenvolgende momenten exemplaren uit een populatie vangen, bijvoorbeeld aan het begin en het einde van de winter wanneer bepaalde vogels aan hun trek beginnen. Op die manier worden vangstfrequenties verkregen van vogels die één maal of twee maal worden gevangen. Zo verkrijgt men twee (of meer) incomplete en elkaar gedeeltelijk overlappende registraties. De overlap tussen registraties correspondeert met de hervangst: de vogels die meermaals zijn gevangen. Andere dataverzamelingsmethoden, zoals de sneeuwbalmethode[16] en het tellen van het aantal observaties binnen een bepaalde periode, kunnen eveneens vangst-hervangstdata opleveren.[17-20] Per methode waarop vangst-hervangstdata worden verkregen, zijn weer meerdere schattingstechnieken beschikbaar. Inmiddels zijn al die schattingstechnieken uitgegroeid tot een substantieel deelgebied binnen de statistiek.

Het corresponderende rekenwerk kan variëren van eenvoudige berekeningen tot complexe statistische modellen. Er is een computerprogramma *Capture* genaamd, waarmee een aantal schattingstechnieken uitgevoerd kan worden.[21] Verder wordt op internet (ftp site: ftp.cs.umanitoba.ca /pub/poplan; web site: http://www.cs.umanitoba.ca /~poplan) het computerprogramma *Poplan* aangeboden waarmee eveneens schattingen gemaakt kunnen worden van de omvang van een populatie. Daarnaast is bestaande software (zoals bijvoorbeeld BMDP, GLIM, SAS en SPSS) waarmee loglineaire analyses uitgevoerd kunnen worden, geschikt voor bepaalde vormen van capture-recapture-analyse. Hiervan geven we later een voorbeeld (zie ook de bijlage voor een SPSS-aansturing).

Hoewel het aantal mogelijke toepassingen groot is, worden capture-recapturemethoden nog zelden toegepast

* F. Smit[1], W. Brunenberg[1], P. van der Heijden[2]

1 Nederlands Centrum Geestelijke Volksgezondheid, Utrecht

2 FSW, Vakgroep Methodenleer & Statistiek, Universiteit Utrecht

in Nederland. Aan de hand van onderzoeksgegevens zullen we het gebruik van drie technieken illustreren.

ONDERWERP EN AANPAK

Ten behoeve van een onderzoek naar het noodzakelijke aantal postdoctorale opleidingsplaatsen voor psychologen hebben we onderzocht hoeveel psychologen - in de functie van psycholoog - werkzaam zijn in instellingen voor de gezondheidszorg in Nederland. Ondanks de aanwezigheid van enkele registraties is dit aantal namelijk onbekend. Bestaande registraties zijn namelijk niet compleet en overlappen elkaar gedeeltelijk.

Als eerste moesten we de populatie definiëren. De definitie luidt: "Psychologen die in de hoedanigheid van psycholoog werken in de individuele gezondheidszorg, met uitzondering van psychologen die als psychotherapeut geregistreerd staan en uitsluitend als psychotherapeut werken". Van de populatie zijn bijvoorbeeld psychologen uitgesloten die als onderzoeker werken en dus een andere functie hebben. Ook zijn psychologen uitgesloten die buiten de gezondheidszorg werken, zoals bij justitie, in de welzijnszorg, het onderwijs of het bedrijfsleven. Vanzelfsprekend is het moeilijk de omvang van een populatie te schatten wanneer niet goed aan te geven is wie wel en wie niet tot de bedoelde populatie behoort.

Voor het maken van een schatting van het totale aantal psychologen in de individuele gezondheidszorg maken we gebruik van lijsten met namen en adressen van psychologen. De eerste lijst is door ons gemaakt op basis van een enquête onder gezondheidszorginstellingen. De tweede lijst met namen is afkomstig van het ledenbestand van de beroepsvereniging van psychologen (het NIP) en de derde lijst is gebaseerd op gegevens van werkgevers in de Riaggs. De lijst van psychologen werkzaam in gezondheidszorginstellingen duiden we aan met I (van instelling), die van het NIP met N, en de lijst van Riaggpsychologen met R. Van deze lijsten zijn steeds alleen die psychologen geselecteerd die aan de definitie van de populatie beantwoorden.

De schatting van het totale aantal psychologen in de individuele gezondheidszorg zullen we op drie manieren uitvoeren. Eerst zullen we op basis van de Instellingslijst (I) een schatting uitvoeren, daarna maken we gebruik van de Petersen's methode om op basis van twee incomplete en gedeeltelijk overlappende lijsten (N en R) tot een schatting te komen, en ten slotte maken we gebruik van Fienberg's methode om op basis van de drie lijsten (I, N en R) tot een schatting te komen. Deze laatste twee methoden staan in de literatuur bekend als vormen van de *capture-recapture*/vangst-hervangstmethode. Methoden die van minder informatie gebruik maken, doen het mogelijkerwijs minder goed dan methoden die van meer informatie gebruik maken. Onder bepaalde condities leidt dit tot een voorkeur voor meer ingewikkelde modellen. We komen hier later op terug.

METHODE 1: EXTRAPOLATIE

De eerste methode om het totale aantal psychologen van de bedoelde categorie te schatten berust op extrapolatie vanuit één enkele steekproef. Daartoe is aan 1203 instellingen in de gezondheidszorg een enquête gestuurd. Gevraagd is het aantal en, indien de werknemers daarmee instemden, ook de namen te noemen van de bij hen werkzame psychologen. Van de aangeschreven instellingen hebben er 626 (52%) gerespondeerd. Het totaal aantal psychologen dat in deze instellingen werkt, is gelijk aan 1773. Het vermoedelijke aantal psychologen in alle instellingen is daarom 1773 gedeeld door de steekproeffractie, ofwel: 1773 / 0,5204 = 3407.

Omdat per instelling gemiddeld 2,83 psychologen werkzaam zijn (M=2,832) met een corresponderende standaard fout van SE_M =0,206 kan het 95% betrouwbaarheidsinterval worden berekend. We vinden voor de ondergrens en de bovengrens respectievelijk 2921 en 3893 psychologen.

Volgens hetzelfde principe hebben we elders[22] de berekening nogmaals uitgevoerd, maar ditmaal is er per type instelling (zoals: Riagg, Algemeen Ziekenhuis, Verpleeghuis, en dergelijke) geëxtrapoleerd. We hebben ons namelijk afgevraagd of er tussen de verschillende typen instellingen meer of minder respons heeft plaatsgevonden en er gemiddeld meer of minder psychologen werken. Aan de hand van die meer nauwkeurige berekeningen komen we uit op een aantal van 3310 psychologen van de bedoelde categorie. Beide uitkomsten (respectievelijk 3407 en 3310) lijken dus op elkaar.

Het extrapoleren van steekproefgegevens naar de populatie is niet zonder risico. Er dient immers vertrouwen te bestaan in de representativiteit van de steekproef van instellingen en de daarbinnen aangetroffen aantallen psychologen. Bij een respons van 52% is niet zonder meer duidelijk of de responderende instellingen gelijk zijn aan de niet-responderende instellingen.[23] Dit is dus een zwakke plek van deze aanpak.

METHODE 2: PETERSEN'S SCHATTER

De Petersen's schatter is een simpele schatter waarbij gebruik wordt gemaakt van twee lijsten met namen. Deze lijsten dienen onafhankelijk van elkaar te zijn, zodat het staan op de ene lijst geen invloed heeft op de kans om op de andere lijst te staan.[9-14] We lichten deze veronderstelling met betrekking tot de onafhankelijkheid tussen beide lijsten later toe.

Op basis van onze kennis van de lijsten vermoeden we onafhankelijkheid tussen de lijsten van de Riagg en het NIP. We kunnen daarom, met voorzichtigheid, de Petersen's schatter gebruiken. De data kunnen beschreven worden door een (incomplete) kruistabel, tabel 1.

Er zijn dus 263 + 1095 = 1358 ($n_{11} + n_{21} = n_{+1}$) personen die op de NIP-lijst voorkomen, waarvan 263 (n_{11}) personen ook op de Riagg-lijst voorkomen en waarvan 1095 (n_{21}) alleen op de NIP-lijst staan. Er zijn $n_{11} + n_{12} = n_{1+}$, ofwel, 263 + 398 = 661 personen die op de Riagg-lijst voorkomen, waarvan weer 263 eveneens op de NIP-lijst staan. Het aantal dat op geen van beide lijsten voorkomt (n_{22}) is uiteraard onbekend. Daarom is de kruistabel onvolledig en is tevens het totale aantal (n_{++}) onbekend. Toch kan op basis van een simpele redenering het onbekende aantal (n_{++}) geschat worden. Bij onafhankelijkheid van de verschillende lijsten geldt namelijk

$$n_{11} = \frac{n_{+1} \times n_{1+}}{n_{++}}$$

R	N	Aantal
1	1	263
1	2	398
2	1	1095
2	2	?

1: komt voor op de lijst; 2: komt niet voor op de lijst;
?: aantal onbekend

Tabel 1 Frequentie van voorkomen op beide lijsten

De onbekende term is n_{++}, maar kan nu opgelost worden,

$$n_{++} = \frac{n_{+1} \times n_{1+}}{n_{11}}$$

wanneer althans onafhankelijkheid tussen beide lijsten verondersteld mag worden. In woorden staat hier dat de populatiegrootte gelijk is aan het totale aantal namen op de ene lijst vermenigvuldigd met het totale aantal namen op de andere lijst gedeeld door het aantal dat op beide lijsten voorkomt. De hierboven gegeven vergelijking is de Petersen's schatter voor de onbekende grootte van een populatie gebaseerd op twee onafhankelijke en incomplete lijsten. Het totale aantal psychologen wordt nu geschat met

$$\hat{n}_{++} = \frac{1358 \times 661}{263} = 3413$$

Het 95% betrouwbaarheidsinterval, waarvoor extra rekenwerk uitgevoerd moet worden[24], heeft een ondergrens van 3128 en een bovengrens van 3699 psychologen. Dit is in overeenstemming met de eerder verkregen schattingen van 3407 en 3310 psychologen. De eerder verkregen aantallen vallen immers binnen het betrouwbaarheidsinterval van de Petersen's schatter.

Uit het voorgaande zal duidelijk zijn dat de Petersen's methode leunt op de veronderstelling dat beide lijsten onafhankelijk zijn. Ter illustratie laten we zien wat er gebeurt wanneer er sprake is van afhankelijkheid tussen de lijsten en de onafhankelijkheidsassumptie dus wordt geschonden. We weten op basis van onze kennis van het veld dat er afhankelijkheid bestaat tussen de Instellingslijst en de lijst van de Riagg (zie tabel 2 voor de data). De corresponderende Petersen's schatter is nu 1596 psychologen (95% BI: ondergrens=1428, bovengrens=1753). Dit is een stuk lager dan de schatting die we zojuist hebben verkregen en is vrijwel zeker een onderschatting van het werkelijke aantal. In geval van afhankelijkheid dienen we dus een andere methode te gebruiken dan de Petersen's schatter, bijvoorbeeld de Fienberg's schatter.

METHODE 3: FIENBERG'S SCHATTER

In zekere zin is Fienberg's schatter een uitbreiding van de methode van Petersen. Ook bij de methode van Fienberg wordt namelijk gebruik gemaakt van een incomplete kruistabel, met dit verschil dat er ten minste drie lijsten met namen van psychologen beschikbaar dienen te zijn. Zoals gezegd, kunnen nu ook eventuele afhankelijkheidsrelaties tussen de lijsten betrokken worden in de schatting. Het idee is een meerwegskruistabel te maken zoals tabel 2. Omdat een aantal personen nergens geregistreerd staat (de 2,2,2-cel), is de tabel incompleet. Op basis van het geobserveerde deel van de tabel kan een scala aan loglineaire modellen geschat worden. We kiezen voor het meest spaarzame, voldoende passende model en schatten vervolgens op basis van dat model de frequentie van de 2,2,2-cel. Deze methode is ontwikkeld door Fienberg.[25] Een goede beschrijving is te vinden in Bishop e.a.[26] Deze methode werd o.a. door Frischer & Leyland[27] toegepast om het onbekende aantal intraveneuze druggebruikers in Glasgow te bepalen. Schouten e.a. hebben de methode toegepast om na te gaan in hoeverre een kanker-registratie compleet is.[8] Wij passen de methode nu toe op het aantal psychologen dat werkzaam is in de gezondheidszorg in Nederland. De aantallen voor zo ver bekend zijn weergegeven in tabel 2.

I	N	R	Aantal
1	1	1	50
1	1	2	72
1	2	1	119
1	2	2	167
2	1	1	213
2	1	2	1023
2	2	1	279
2	2	2	?

1: komt voor op de lijst; 2: komt niet voor de lijst;
?: aantal onbekend

Tabel 2 Frequentie van voorkomen op de drie lijsten

We geven een korte toelichting op tabel 2. Op de Instellingslijst vinden we 167 (n_{122}) namen van psychologen die op geen andere lijst staan. Het aantal van deze 167 namen staat in de 1,2,2-cel van de tabel. Er zijn 119 (n_{121}) namen die zowel voorkomen op de Instellingslijst als op de lijst van de Riagg. Zo komen er 50 (n_{111}) namen voor op de drie lijsten (de 1,1,1-cel). Ten slotte zien we bij de 2,2,2-cel vraagtekens bij het aantal psychologen dat op geen van de drie lijsten voorkomt. Dit is het onbekende aantal. Met behulp van een loglineair model kan dit onbekende aantal en daarmee het totale aantal geschat worden.

Tabel 3 laat zien welke loglineaire modellen geschat zijn op het observeerbare deel van tabel 2. In de bijlage hebben we voor de geïnteresseerde lezer de corresponderende aansturing voor SPSS afgedrukt. We merken op dat de 2,2,2-cel gehanteerd is als een 'structurele 0'. Hiermee wordt aangegeven dat hier geen observaties hebben kunnen plaatsvinden - wat iets anders is dan geen (0) observaties.

Het meest spaarzame, voldoende passende model is 3c. De keuze voor model 3c motiveren we als volgt. Statistisch gesproken past het model goed bij de data. Dit model, met interacties I*N en I*R, is bovendien in overeenstemming met onze kennis van de drie lijsten. De Instellingslijst is namelijk samengesteld met kennis van de NIP-lijst. Dat verklaart de gevonden samenhang. Een samenhang tussen de Instellingslijst en de Riagg-lijst is eveneens aannemelijk. De interacties in model 3c reflecteren de 'maatschappelijke (on)zichtbaarheid' van verschillende subgroepen binnen de populatie.[26]

Omdat we in tabel 2 slechts zeven cellen hebben met geobserveerde aantallen, kunnen we niet meer dan zeven parameters schatten. Dit houdt in dat de 3-wegsinteractie niet berekend kan worden en daarom zijn we gedwongen afwezigheid van zo'n 3-wegsinteractie veronderstellen. Wel zouden we een model kunnen schatten met drie 2-wegsinteracties (N*R, N*I, R*I), maar dan zijn er geen vrijheidsgraden meer om te toetsen.

Model	Hoofdtermen			Interactietermen		$\chi 2$	(df)	p
1	I	N	R			137,8	(3)	0,000
2a	I	N	R	N*R		132,0	(2)	0,000
2b	I	N	R	I*N		92,9	(2)	0,000
2c	I	N	R	I*R		17,3	(2)	0,000
3a	I	N	R	N*R	I*N	33,7	(1)	0,000
3b	I	N	R	I*R	N*R	10,1	(1)	0,001
3c	I	N	R	I*N	I*R	0,0	(1)	0,907

χ^2 : de likelihood ratio

Tabel 3 Passendheid van de loglineaire modellen

Parameter	Schatter	Standaard fout
Const	7,2004	–
I	–2,0792	0,1204
N	–0,2699	0,0910
R	–1,5692	0,0753
I*N	–0,5821	0,1413
I*R	1,2226	0,1256

Alle schatters hebben | z-waarden | >3,00

Tabel 4 Parameterschattingen van model 3c

Om redenen van efficiëntie wordt in het algemeen gekozen voor het meest spaarzame en toch voldoende passende model. Spaarzame modellen genieten de voorkeur, want naarmate er meer parameters geschat worden, wordt het betrouwbaarheidsinterval rondom de geschatte waarde ($\hat{n}_{222}$) groter en wordt de schatting dus minder nauwkeurig.[26]

De schatters van de parameters van model 3c staan in tabel 4.

Het verwachte aantal voor de 2,2,2-cel is gelijk aan

$$\hat{n}_{222} = \exp(Const) = \exp(7{,}2004) = 1340$$

Uitgaande van het loglineaire model komt het totale aantal op 1340 (=onbekend) + 1923 (=bekend) = 3263 psychologen die werkzaam zijn in de gezondheidszorg. De ondergrens en de bovengrens van het 95% betrouwbaarheidsinterval zijn respectievelijk 3000 en 3526. Voor de berekening van het 95% betrouwbaarheidsinterval verwijzen we naar Bishop e.a.[26], hoofdstuk 6. Dit betrouwbaarheidsinterval heeft een aanzienlijke overlap met de eerder verkregen betrouwbaarheidsintervallen en we concluderen daaruit dat de schattingen met elkaar in overeenstemming zijn.

VERONDERSTELLINGEN

Bij de gepresenteerde schattingen plaatsen we enkele kanttekeningen. De schattingen stoelen niet alleen op geobserveerde data, maar ook op enkele veronderstellingen. Het niet voldoen aan deze veronderstellingen kan leiden tot minder realistische schattingen. Behalve met de eerder genoemde onafhankelijkheid tussen de registraties, dient met drie andere veronderstellingen rekening gehouden te worden: 1) homogeniteit van de populatie, 2) gesloten-heid van de populatie en 3) perfecte 'record linkage'. We bespreken deze veronderstellingen achtereenvolgens.

Homogeniteit

De eerste veronderstelling betreft de homogeniteit van de populatie. Dit wil zeggen dat tussen groepen van personen binnen de populatie geen ongelijke kansen zijn om geregistreerd te worden; dat bijvoorbeeld oudere psychologen niet vaker in registraties voorkomen dan jongere. In de Engelstalige literatuur wordt dit weleens 'variable catchability' genoemd.

Cormack[28] doet een suggestie aan de hand hoe men kan uitzoeken of de populatie wat pakkansen betreft homogeen is. Hij gaat uit van de redenering dat de onbekende groep die nooit geregistreerd is meer zal lijken op de groep die slechts één keer geregistreerd is en minder zal lijken op groepen die meerdere keren geregistreerd zijn. In het verlengde van deze redenering stelt Cormack voor een plaatje te maken van de gestandaardiseerde residuen van het meest spaarzame best passende loglineaire model tegen het aantal keren dat iemand op een lijst voorkomt. Op onze data toegepast laat het plaatje (hier niet getoond) geen samenhang zien tussen de grootte van de residuen en het aantal geobserveerde keren dat iemand geregistreerd is. We concluderen daaruit dat het model even goed past voor de groep die vaker geregistreerd is als voor de groep die slechts één keer geregistreerd is. We denken daarom dat het model ook goed past voor de groep die nergens geregistreerd is.

Zou er wel sprake geweest zijn van heterogeniteit, dan kan overwogen worden te stratificeren. Men maakt eerst homogene strata, bijvoorbeeld oudere versus jongere psychologen, en voert vervolgens per homogeen stratum aparte schattingen uit.

Gesloten-heid

De tweede veronderstelling betreft het al dan niet 'gesloten' karakter van de populatie. Wanneer er sprake is van een 'gesloten populatie' is de ware populatiegrootte niet onderhevig aan verandering ten gevolge van geboorte, sterfte en migratie. Bij dit onderzoek betekent dit dat de gebruikte lijsten actueel dienen te zijn en dat gedurende het onderzoek geen nieuwe leden op de lijsten opgenomen en ex-leden weer van de lijsten afgevoerd hadden moeten worden. Wat betreft het onderhavige onderzoek bestaat op dit punt bij ons weinig twijfel. De drie lijsten zijn actueel, zijn allen op hetzelfde tijdstip en in een tijdspanne van enkele weken samengesteld. Daarom verwachten we niet dat de schattingen sterk zullen afwijken van het ware aantal.

Zou er wel sprake zijn van een open populatie, dan staan er twee wegen open om hiermee om te gaan. Of men voert het onderzoek uit in een korte periode zodat men voor die korte periode de populatie als gesloten mag beschouwen, of men werkt met meer ingewikkelde modellen waarin migratie, geboorte en sterfte verwerkt zijn. Dit soort schattingen kunnen uitgevoerd worden met het genoemde computerprogramma *Poplan*.

Perfecte 'record linkage'

Wanneer er gewerkt wordt met verschillende lijsten, dan dient met zekerheid vastgesteld te kunnen worden of een persoon op lijst A dezelfde persoon is die op lijst B voorkomt. Met andere woorden, personen dienen uniek geïdentificeerd te kunnen worden. Dit kan wel eens voor complicaties zorgen.

In het huidige onderzoek hebben we op dit front geen problemen ondervonden. De ledenlijst van het NIP is

openbaar en bevat meerdere gegevens van de leden (naam, adres, en vaak waar zij werken) aan de hand waarvan zij uniek identificeerbaar zijn. Dezelfde gegevens hadden we ook tot onze beschikking voor de andere lijsten. Er is handmatig door twee personen gematched. In slechts enkele gevallen leidde dit tot beoordelingsverschillen. Deze beoordelingsverschillen konden later door consensus worden opgelost.

Samenvattend

Omvangsschattingen berusten niet alleen op geobserveerde data, maar ook op een aantal veronderstellingen. Schending van assumpties kan leiden tot minder realistische schattingen. Het is daarom altijd zaak de keuze van de schattingsmethode te laten afhangen van de onderliggende veronderstellingen en de vraag in hoeverre deze veronderstellingen worden bedreigd. Wat betreft het onderhavige onderzoek menen we dat de relevante assumpties niet geschonden zijn. Bovendien bereiken we langs meerdere wegen uitkomsten die in hoge mate met elkaar in overeenstemming zijn. Dit geeft vertrouwen in de door ons gerapporteerde uitkomsten.

DISCUSSIE

Aan de hand van een voorbeeld hebben we verschillende methoden van het schatten van de grootte van een populatie laten zien. Met behulp van deze drie methoden komen we uit op een schatting van tussen de 3000 en de 3526 psychologen die werkzaam zijn in de individuele gezondheidszorg.

Meer in het algemeen kan het van belang zijn te weten hoe groot een bepaalde groep is. Wanneer voor zo'n groep geen steekproefkader beschikbaar is, dan kunnen conventionele (epidemiologische) methoden niet worden gebruikt of is hun gebruik inefficiënt. Omvangsschattingen gebaseerd op vangst-hervangstdata kunnen onder zulke omstandigheden wellicht uitkomst bieden. Dat rechtvaardigt een brede belangstelling in vangst-hervangstmethoden om de omvang van een populatie te schatten.

ABSTRACT

Estimating the size of a population: applications in social science and an example

The number of psychologists working in the individual health care in the Netherlands is not known. This uncertainty hampers planning and policy making. In this article we estimate their number in three different ways. The estimates are based on single, double and triple incomplete and partially overlapping data sources. Using some assumptions and applying capture-recapture methods we finally estimate their number as 3263, give or take 263. This estimate is supported by other estimates also presented in this article. Our study illustrates that population size estimators based on capture-recapture data can be used successfully in the context of social science research. Population size estimators are and will be used to estimate the number of homeless people, drugs users, pathological gamblers, prostitutes, HIV infected persons, neglected elderly people (with regard to unmet need assessment) and the size of criminal groups. These and other applications warrant a broad interest in capture-recapture methods.

Key words: population size estimation, incomplete registries, capture-recapture methods

LITERATUUR

1 Collins MF, Wilson RM. Automobile theft: estimating the size of a criminal population. J Quantitat Criminol 1990;6:395-409.

2 Darcy L, Jones DL. The size of the homeless men population of Sydney. Australian J Social Issues 1975;10:208-15.

3 Fisher N, Turner SW, Pugh R, Taylor C. Estimating numbers of homeless and homeless mentally ill people in North East Westminster by using capture-recapture analysis. Br Med J 1994;308:27-30.

4 Bloor M, Leyland A, Barnard M, McKegeney N. Estimating hidden populations: a new method of calculation the prevalence of drug-injecting and non-injecting female street prostitution. Br J Addiction 1991;86:1477-83.

5 Rosmo DK, Routledge R. Estimating the size of criminal populations. J Quantitat Criminol 1990;6:293-314.

6 Frisher M, Goldberg D, Green S. How many drug injectors are there in the UK? Intern J Drug Policy 1993;4:190-3.

7 Heijden P van der, Smit F, Gils G van. Schattingen van het aantal slachtofferloze delicten. Politia Nova 1993;3:(geheel).

8 Schouten LJ, Straatman H, Kiemeney LALM, Gimbrère CHF, Verbeek ALM. The capture-recapture method for estimation of cancer registry completeness: a useful tool? Intern J Epidemiol 1994;23:1111-6.

9 Seber GAF. The estimation of animal abundance and related parameters. London: Charles Griffin, 1982.

10 Seber GAF. A review of estimating animal abundance. Biometrics 1986;42:267-92.

11 Seber GAF. A review of estimating animal abundance II. Intern Statist Rev 1992;60:129-66.

12 Pollock KH. Modelling capture-recapture, and removal statistics for estimation of demographic parameters of fish and wildlife populations: past, present and future. Am Statist Ass 1991;86(413):225-38.

13 Wilson RM, Collins MF. Capture-recapture estimation with samples of size one using frequency data. Biometrika 1992;79:543-53.

14 Khorazaty MN El, Imrey PB, Koch GG, Wells HB. Estimating the total number of events with data from multiple-record systems: a review of methodological strategies. Intern Statist Rev 1977;45:129-57.

15 Smit F, Heijden P van der, Gils G van. Enkele weinig gebruikte methoden om het aantal plegers van misdrijven te schatten. Tijdschr Criminologie 1994;36:96-119.

16 Frank Ö, Snijders T. Estimating hidden populations using snowball sampling. J Official Statist 1994;10:53-67.

17 Zelterman D. Robust estimation in truncated discrete distributions with application to capture-recapture experiments. J Statist Planning Inference 1988;18:225-37.

18 Chao A. Estimating animal abundance with capture frequency data. J Wildlife Management 1988;52:295-300.

19 Chao A. Estimating population size for sparse data in capture-recapture experiments. Biometrics 1989;45:427-38.

20 Chao A, Jeng S. Estimating population size for capture-recapture data when capture probabilities vary by time and individual animal. Biometrics 1992;48:201-16.

21 Rexstad E, Burnham K. Capture: abundance estimation of closed animal populations. Fort Collins, Colorado: Colorado Cooperative Fish and Wildlife Research Unit, Colorado State University, 1991.

22 Brunenberg W, Neijmeijer L, Hutschemaekers G. Beroep: Psycholoog/pedagoog. Een verkennend onderzoek naar persoon, werk en werkplek van pedagogen en psychologen in de gezondheidszorg. Utrecht: NcGv, 1995.
23 Berk RA, Subash SR. Selection bias in sociological data. Soc Sci Res 1982;11:352-98.
24 Seber GAF. The estimation of animal abundance and related parameters. London: Charles Griffin, 1982:59-61.
25 Fienberg S. The multiple recapture census for closed populations and incomplete 2^k contingency tables. Biometrika 1972:59:591-603.
26 Bishop Y, Fienberg S, Holland P. Discrete multivariate analysis: theory and practice. Cambridge, Mass: MIT-Press, 1975.
27 Frischer M, Leyland A. Reliability of population and prevalence estimates. Lancet 1992;339:995.
28 Cormack RM. Log-lineair models for capture-recapture. Biometrics 1989;45:395-413.

CORRESPONDENTIE-ADRES
F. Smit, NcGv, Sectie M&T, Postbus 5103, 3502 JC Utrecht, tel. 030-2971100

Voor publicatie aanvaard op 6 mei 1996

Bijlage Geannoteerde aansturing voor omvangschatting met SPSS (6.1 voor Windows)

In deze bijlage geven we een geannoteerde aansturing van SPSS (6.1 voor Windows). Voorafgaand maken we drie opmerkingen.
1 De procedure HILOGLINEAR is alleen geschikt voor model-selectie. De parameterschattingen van deze procedure worden namelijk alleen gegeven voor het verzadigde model.
2 De procedure LOGLINEAR kan gebruikt worden voor de model-selectie, maar omdat in de output de constante van het model niet gegeven wordt, is deze procedure niet geschikt voor parameterschattingen t.b.v. omvangschattingen.
3 De procedure GENLOG is zowel geschikt voor modelselectie (maar dit is omslachtig), als ook voor parameterschattingen.

Aansturing					Commentaar
Data inlezen					
data list free					'procedure "data list"
/I N R freq cw.					'Variabelen, frequentie, structurele 0
var labels I 'instellingen' /N 'nip' /R 'riaggs'.					'labels
begin data.					'begin inlezen van data
1	1	1	50	1	
1	1	2	72	1	
1	2	1	119	1	
1	2	2	167	1	'data
2	1	1	213	1	
2	1	2	1023	1	
2	2	1	279	1	
2	2	2	0	0	'NB. structurele 0!
end data.					'eind inlezen van data
weight by freq.					'wegen met de frequenties
Modelselectie					
Hiloglinear					'procedure "hierarchical loglinear analysis"
I (1,2) N (1,2) R (1,2)					'variabelen (laagste, hoogste klasse)
/cweight cw					'aangeven van structurele nul
/method backward					'achterwaartse stapsgewijze model-selectie
/print freq resid					'statistieken opvragen
/design I*N I*R N*R					'kies de drie 2-wegsinteracties als hiërarchisch hoogste model
Het meest spaarzame, voldoende passende model (3c) is I, N, R, I*N, I*R					
Parameterschattingen model 3c					
Genlog					'procedure "generalized loglinear models"
I N R					'variabelen
/cstructure=cw					'structurele 0
/model=mult					'multinominaal model
/print estim					'opvragen van de parameterschattingen
/plot none					'onderdrukken van plaatjes
/design I N R I*N I*R.					'model 3c
Exp (constante) = onbekend aantal. Geschat aantal = bekend aantal + onbekend aantal					