De toepasbaarheid van statistische toetsen op populatiegegevens

R. Adriaanse

Inleiding

Het is al bijna 20 jaar geleden dat H. Selvin een discussie startte over het gebruik van statistische toetsen bij survey-onderzoek (Selvin 1957). Hij was tegen dit gebruik, voornamelijk omdat een survey geen wetenschappelijk experiment is en statistische toetsen naar zijn mening alleen bij experimenten van toepassing zijn. De daarop volgende discussie heeft een aantal zaken wel opgehelderd, zie daarvoor Gold (1958), Selvin (1958a), Beshers (1958), Selvin (1958b), McGinnis (1958), Freedman (1958), Brouwer en Mokken (1958), maar er is tenminste één door Selvin gesignaleerd probleem waarover nog algemene verwarring bestaat: de toepassing van statistische toetsen op populatiegegevens of, meer algemeen, op gegevens die niet opgevat kunnen worden als een aselekte steekproef uit een bestaande, duidelijk omschreven populatie. Hoe kan deze toetsing geïnterpreteerd worden, als generalisatie naar een universum waaruit de gegevens door aselecte trekking verkregen zijn, onmogelijk is?
Het doel van dit artikel is, duidelijk te maken welke statistisch verantwoorde interpretatie in dit geval mogelijk is.

Een aantal standpunten

Om te beginnen zal ik een overzicht geven van wat een aantal verschillende auteurs over dit probleem heeft gezegd. De keuze van deze literatuur is betrekkelijk willekeurig, maar geeft naar mijn mening wel een goede indruk van de verschillende opvattingen in deze kwestie. Selvin zegt dat bij populatiegegevens de optredende verschillen klaarblijkelijk niet het gevolg kunnen zijn van 'steekproef-fouten'. Vervolgens noemt hij drie andere mogelijkheden om de optredende verschillen aan 'toevals-fouten' te wijten. In de eerste plaats, het opvatten van onderzochte populaties als steekproeven uit nog grotere hypothetische universa van mogelijkheden. Hij vindt dit concept intuïtief moeilijk te vatten en onnodig indien men onderkent dat het trekken van steek-

proeven niet de enige bron van toevals-fouten is. De twee andere door hem genoemde mogelijkheden zijn: toevallige fouten in respectievelijk de respons van de respondent en de verwerking van het materiaal (codering en kaarten ponsen). Maar deze fouten kunnen naar zijn mening geminimaliseerd worden en zijn dan minder belangrijk als mogelijke verklaringen van waargenomen verschillen. Hij besluit met de opmerking dat als deze bronnen van toevallige fouten niet werkzaam kunnen zijn het betekenisloos is om de kans te berekenen dat zij zouden zijn opgetreden (Selvin 1957: 525).

Bij Brouwer en Mokken komt de kwestie, zij het indirect, weer ter sprake: 'In het sociaal-wetenschappelijk onderzoek heeft men echter dikwijls te maken met vage, hypothetische en/of onbegrensde populaties. Het leeuwendeel van de groepspsychologische onderzoekingen valt bijvoorbeeld onder dit type. Het gevolg is dan ook dat er vele discussies plegen te worden gevoerd over de generaliseerbaarheid van dit soort studies, maar dat neemt niet weg dat men hier toch – en terecht – gebruik maakt van statistische schattings- en toetsingstechnieken. De modellen van de waarschijnlijkheidsrekening betreffen zelf dikwijls onbegrensde populaties, bijvoorbeeld het dobbelmodel (Brouwer en Mokken 1959: 86)'. Brouwer en Mokken wijzen erop dat ook het tijdstip van onderzoek behoort tot de kenmerken van de populatie: 'Uit de resultaten van een toetsing of schatting kunnen wij weinig of niets afleiden over dezelfde populatie een jaar eerder of later (ibid: 86)'.

In 1963 verscheen het onderzoek van Lammers waarbij statistische toetsen op populatiegegevens werden toegepast. Lammers doet daarbij een prijzenswaardige poging het gebruik van statistische toetsen te rechtvaardigen, wat een subtiel stukje proza over deze kwestie oplevert, waarin de verschillende in omloop zijnde argumenten gecombineerd worden. Helaas moet ik mij hier beperken tot de hoofdpunten omdat een volledig citaat te veel ruimte zou vergen. Lammers begint met de volgende opmerking: 'Geen conclusies worden aan het materiaal ontleend, tenzij middels statistische toetsing is aangetoond, dat de kans, dat een zodanig resultaat op toevalsschommelingen berust, gering is. Aan deze beslissing ligt ten grondslag de gedachte dat bij deze enquêtes een steekproef en geen populatie is ondervraagd. Daar in dit onderzoek in eerste instantie niet naar generalisaties gestreefd wordt, zou men kunnen menen, dat althans de eerste enquête, waaraan bijna alle toentertijd op het KIM aanwezige adelborsten en ARO's deelnamen, op een populatie betrekking heeft'. Hij wil de bij de eerste enquête onderzochte populatie toch opvatten als een steekproef en wel op de volgende vier verschillende wijzen.

1 – 'Zo zou men de ondervraagde populatie van aspirant-officieren kunnen opvatten als een steekproef uit een aantal vergelijkbare 'populaties' bijvoorbeeld van adspirant-officieren op het KIM in andere jaren, op andere militaire academies of zelfs van leerlingen op andere met het KIM in relevante

opzichten overeenkomstige opleidingsinstituten'.
2 — Het tijdsaspect: 'Altijd wil men zijn gegevens toch wel als representatief beschouwen voor een periode, die zich uitstrekt over een langere tijd dan de tijd, gedurende welke de enquête werd afgenomen'.
3 — Toevalsschommelingen in opinies en houdingen van de respondenten.
4 — Toevalsinvloeden op het aangeven van de 'echte' opinie of houding bij de beantwoording van een enquêtevraag.
In zijn slotopmerking kiest Lammers kennelijk voor de twee laatste zienswijzen: 'De toepassing van statistische toetsen bij het vergelijken van diverse subgroepen ondervraagden heeft ten doel te voorkomen, dat men belang hecht aan eventueel tussen deze subgroepen te constateren verschillen in vraagbeantwoording, welke eenvoudig een gevolg kunnen zijn van de hierboven besproken fluctuaties van betrekkelijk toevallige aard in mening, houding of weergave daarvan bij de ondervraagden (Lammers 1963: 288-290)'.
Een ander onderzoek waarbij sprake was van toepassing van statistische toetsen op populatiegegevens is dat van Nooij (1969). In zijn boek gaat Nooij niet op de kwestie in, hij vermeldt de toetsingsresultaten zonder commentaar. Bij de boekbespreking komt de zaak wel aan de orde in: Van den Ban, Swanborn, Nooij (1969). Het grootste gedeelte van Nooij's gegevens is verzameld in de Krimpenerwaard, waar in vier gemeenten alle boeren met een bedrijf van meer dan 5 ha zijn geënquêteerd. Over de statistische toetsing van deze gegevens merkt Van den Ban op: 'Een veronderstelling hierbij is uiteraard dat de steekproef volgens toeval is gekozen. Doordat de schrijver in zijn onderzoeksgemeenten in de Krimpenerwaard alle boeren heeft getracht te ondervragen, is de waarde van een dergelijke toetsing twijfelachtig (Van der Ban e.a. 1969: 320)'.
Het commentaar van Swanborn hierop is zeer kort: 'Waarom Nooij van statistische toetsing gebruik maakt, is niet duidelijk. Bij de tabellen vermeldt hij naast elkaar gamma (in feite Q) en chikwadraat, hetgeen wat wonderlijk aandoet (ibid: 324)'. In zijn antwoord zegt Nooij 'Strikt genomen mag men alleen toetsen wanneer — behoudens het kwantitatieve aspect — de definities van populatie en steekproef identiek zijn. Deze identiteit wordt gewoonlijk nagestreefd door het trekken van een aselecte steekproef uit een nauwkeurig afgebakende populatie. Het omgekeerde is echter ook mogelijk! als de 'steekproef' gegeven is, kan men vanuit de steekproef een definitie opstellen voor de populatie. Op grond van deze redenering heb ik tenslotte besloten tot statistische toetsing. De populatie bestaat dan uit alle boeren die in soortgelijke omstandigheden verkeren als degenen die in het onderzoek zijn betrokken. Een exacte definiëring van deze populatie waaruit ondubbelzinnig blijkt wie er wél en wie er niet toe behoren, is niet te geven. Dit betekent tevens, dat het niet mogelijk is om exact aan te geven wat de betekenis is van de toetsings-

resultaten. In feite komt het er op neer, dat aan geconstateerde relatieve verschillen meer waarde wordt gehecht naarmate de subsamples waartussen de verschillen zijn waargenomen, groter van omvang zijn (ibid: 327)'.
Een artikel dat de zaak eindelijk eens van een andere kant bekijkt is dat van Gold (1969). Hij wijst er op dat een significantie-toets op zijn hoogst een door tijd, plaats en mensen beperkte index van betrouwbaarheid kan geven, maar dat het dan in feite gaat om eenmalige historische kennis. Gold geeft dan het voorbeeld van een chi-kwadraat toets voor een 2 x 2 tabel met populatiegegevens en merkt daarbij op: 'It becomes clear that lack of a statistically significant chi square is especially revealing, for this tells us that an association is no larger than would ordinarily expected by randomly pairing the values of two variables (keeping the marginals constant). When lack of statistical significance by any test is found in an universe or a given set of data (keep in mind, not a sample), we can say that in the empirical world the association produced by nature is no greater than that produced by chance (e.g. random pairing). And it would seem a faire rule of thumb that, given our present state of knowledge about associations among sociological variables, we cannot with any confidence attribute substantive importance to associations of such magnitude (Gold 1969: 44)'. Gold ziet de significantietoetsing dus als een middel om vast te stellen of er sprake is van een empirische samenhang van enige betekenis in de onderzoeksgegevens, zonder te generaliseren naar een of ander universum. Hij vraagt zich slechts af of de gevonden samenhang sterker is dan men volgens toeval zou mogen verwachten.
De opvatting van Gold wordt bestreden door Morrison en Henkel, die menen dat statistische toetsen alleen mogen worden gebruikt om op grond van een waarschijnlijkheidsmodel fouten te schatten die werden gemaakt bij het trekken van de steekproef. Voorts menen zij dat, indien er bij het verzamelen van de gegevens geen sprake is van toeval, de vraag irrelevant is of het resultaat het gevolg kan zijn van een aselecte verdeling (Morrison en Henkel 1969: 134).
Tenslotte zullen we nog even nagaan wat enige handboeken over deze kwestie te melden hebben. Selltiz e.a. beginnen met op te merken dat: 'Many studies in behavioral science are carried out on accidental samples of subjects. The data are treated, however, in a manner that is appropriate only to propability samples. For example, statistical tests of significance which presuppose random sampling are applied to the data'. Zij noemen twee gebruikelijke rechtvaardigingen van statistische toetsing in dit geval. Ten eerste het argument dat men niet geïnteresseerd is in het schatten van populatiewaarden, maar in het onderzoeken van samenhangen tussen variabelen zonder te refereren aan een of andere populatie. De auteurs achten dit een vals argument, daar ook samenhangen aan steekproeffluctaties onderhevig zijn. Als tweede rechtvaar-

diging noemen zij het postuleren van een hypothetische populatie waaruit de onderzoeksgegevens een aselecte steekproef zouden zijn. Bij statistische toetsing generaliseren we dan naar deze hypothetische populatie, wat inhoudt dat we niet weten voor wie de gevonden samenhangen gelden. Daarop zeggen zij: 'The point we are making is that the scientific quest does not end with the statement tha a finding is or is nog statistically significant. We still have the task of specifying the populations for which it is or not significant'. Vervolgens geven zij echter te kennen in bepaalde gevallen deze wijze van toetsing niet af te keuren, ondermeer omdat men ook ingeval van een aselecte steekproef vaak wil generaliseren naar een andere dan de onderzochte populatie (Selltiz e.e. 1967: 541-545). Hun betoog op dit punt munt niet uit door duidelijkheid. Galtung (1967) besteed veel aandacht aan de kwestie. Hij hecht veel waarde aan het onderscheid tussen een substantieve hypothese, dat is een veronderstelling over de variabelen in de steekproef en een generalisatie hypothese, welke uitspraken over de steekproef generaliseert naar de populatie. Galtung gaat er vanuit dat statistische toetsen alleen zinvol zijn voor het toetsen van gegeneralisatie hypothesen. Over het geval dat men over populatiegegevens beschikt zegt hij: 'in this case substantive hypotheses can be tested directly on the data; test of statistical hypothese are out of order, since there is nothing to generalize (Galtung 1967: 363)'. Even verder stelt hij: 'Thus, our main thesis is that statistical tests are out of order if we do not have a sample (ibid: 364)'. Vervolgens bespreekt Galtung de gebruikelijke argumenten. In de eerste plaats het argument dat de gegevens wel een populatie vormen maar meet- en waarnemingsfouten bevatten. Daarover zegt hij: 'But there is no reason a priori to assume that models that may function well for sampling fluctations also function well for observation fluctations ... Hence, if a model for the generation of observation and measurement errors exists, and a reasonable number of the conditions are satisfied, then statistical inference techniques should be used to test the null hypothesis that findings are due to imperfection; but uncritical borrowing from models appropriate for sampling fluctuations is illegitimate (ibid: 364)'. Het andere argument, dat de populatie opgevat kan worden als een steekproef uit een groter (hypothetisch) universum, wordt door Galtung uitvoerig besproken en ongeldig verklaard. Hij besluit als volgt: 'Thus to summarize: we think the idea of the hypothetical universe with the hypothetical sampling is misplaced, when we already have data about the universe we want at least as long as no one has presented a good case for an operational definition. And that concludes our discussion of the application of statistical tests to universes: such use of statistic makes sense, but only in some cases and only when some conditions are fullfilled (ibid: 369)'. Dit standpunt is wat genuanceerder dan zijn oorspronkelijke 'main thesis'.

Albinski merkt op: 'De inductieve statistiek heeft niets anders tot doel dan het bestuderen van mogelijkheden van generalisatie van steekproefuitkomsten. Wanneer men een bepaalde populatie bestudeert en de conclusies uitsluitend voor die populatie wil laten gelden, dan heeft het volstrekt geen zin betrouwbaarheidsmarges te bepalen of significantietoetsen te hanteren'. Hij wijst er echter op dat men zich soms niet beperkt tot de onderzochte populatie maar wil generaliseren naar een groter universum en gaat dan als volgt verder: 'Kan de inductieve statistiek hier diensten bewijzen of niet? Deze vraag kan niet met een ja of nee worden beantwoord . . . In dat geval heeft ze waarschijnlijk de functie van een zeef, die aangeeft welke conclusies de onderzoeker wel zal trekken en welke niet. De berekende kansen mogen hier in ieder geval niet als kansen worden geïnterpreteerd. Waar men geen beter kriterium voor het trekken van conclusies kan vinden, worden soms allerlei significantietoetsen in deze oneigenlijke zin gebruikt en zij kunnen toch ook in deze gevallen nog een nuttige functie vervullen. Toch is de eventuele generaliseerbaarheid van de conclusies hier statistisch niet te rechtvaardigen. Zij moet op andere gronden worden verdedigd (Albinski 1971: 175-176)'.
Swanborn (1972) zegt slechts zijdelings iets over de kwestie. In verband met noodzakelijke voorwaarden voor een causale afhankelijkheid tussen variabelen merkt hij op: 'Een minimum eis is in ieder geval, dat het verband sterker is dan het verband dat we zouden verkrijgen indien de eenheden zuiver bij toeval over de beide variabelen verdeeld zouden zijn. Er zijn nu statistische toetsen (zgn. randomization tests) om te onderzoeken of de gevonden samenhang voldoende afwijkt van de te verwachten samenhang 'bij toeval' (Swanborn 1972: 141-142)'. Deze formulering lijkt veel op die van Gold (1969), maar even verder blijkt dat Swanborn zich toch beperkt tot het geval van een steekproef uit een populatie: '. . . ., omdat het bij statistische toetsing gebruikte mathematische model gebaseerd is op de gedachte dat wanneer we slechts enkele eenheden trekken uit en populatie en we verdelen deze over twee groepjes, door het toeval vrij aanzienlijke verschillen kunnen bestaan (ibid: 142)'.
Over het generalisatieprobleem zegt hij elders: 'Wij zijn geen aanhanger van het puristische standpunt dat de resultaten van een steekproefonderzoek op geen enkele wijze gegeneraliseerd mogen worden naar andere populaties dan die waaruit de steekproef op aselecte wijze is getrokken . . . Indien een groep verstandige wetenschappelijke onderzoekers geen enkel belangrijk argument kan bedenken waarom de Rotterdamse bevolking bijvoorbeeld qua opvattingen over geboorteregeling, zou afwijken van de nederlandse bevolking, dan mogen we met de nodige voorzichtigheid de Rotterdamse resultaten doortrekken naar de nederlandse bevolking (ibid: 161'. Over het gebruik van statistische toetsen bij populatiegegevens zegt hij echter niets.

Commentaar

In deze literatuur komen we steeds een bepaalde opvatting van statistische toetsen tegen, waarvan de belangrijkste elementen zijn:

1 — Statistische toetsen hebben als enig doel het toetsen van generalisatie hypothesen, d.w.z. na te gaan of steekproefuitkomsten ook gelden voor de populatie waaruit de steekproef getrokken is.

2 — Statistische toetsen zijn slechts bruikbaar als er sprake is van 'toevallige fouten', te onderscheiden in twee soorten:

a — steekproef-fouten, het niet-representatief zijn van de aselecte steekproef, voor de populatie waaruit hij getrokken is,

b — niet-systematische meetfouten, alle toevallige variatie waardoor de onderzoeksresultaten afwijken van wat men aan de waarnemingseenheden wilde meten.

De eerste uitspraak wordt door velen als vanzelfsprekend gezien, wat dan, in de meest uitgebreide vorm, tot uitspraak 2 leidt. Ook bij 2b kan namelijk een steekproefpopulatie model gehanteerd worden, de populatie is dan een abstracte populatie bestaande uit alle mogelijke onafhankelijke herhalingen van het onderzoek met dezelfde waarnemingseenheden.[1] Deze opvatting is begrijpelijk gezien de nadruk in het traditionele statistiekonderwijs op steekproefverdelingen en parametrische toetsen, de zgn. generaliserende statistiek. Daarbij ziet men echter over het hoofd dat niet-parametrische (verdelingsvrije) toetsen voor waarnemingsparen of twee of meer waarnemingsreeksen, zoals bijvoorbeeld de chi-kwadraat-, de rangcorrelatie- en de Wilcoxontoets, in feite zgn. randomisatie toetsen zijn welke een andere theoretische achtergrond hebben dan de parametrische toetsen. Met randomisatie toetsen wordt in eerste instantie getoetst of bepaalde subcategorieën van de onderzochte waarnemingseenheden opgevat kunnen worden als aselecte steekproeven uit het totaal der onderzochte eenheden. Indien de waarnemingseenheden een aselecte steekproef uit een populatie vormen kan het toetsingsresultaat gegeneraliseerd worden naar die populatie maar deze tweede stap is niet noodzakelijk!

Uitspraak 1 kan echter weerlegd worden zonder dat we in hoeven te gaan op de technische details van randomisatietoetsen. Het is voldoende te bedenken dat een statistische toets is opgebouwd uit drie elementen: een populatie, een steekproef en een selectieprincipe (de wijze waarop de steekproef uit de populatie verkregen is). Nu is statistische toetsing of voorspelling mogelijk indien van deze drie elementen er twee bekend zijn. We spreken van *generaliserende statistiek* als steekproef en selectieprincipe (aselecte trekking) bekend zijn en op basis daarvan uitspraken worden gedaan over de onbekende populatie. Een ander geval is dat populatie en steekproef (subpopulatie) bekend

zijn maar het selectieprincipe niet, men kan dan toetsen of de steekproef (subpopulatie) aselect uit de populatie getrokken is wat ik *toetsing op aselectheid* zal noemen. De derde mogelijkheid, populatie bekend en als selectieprincipe aselecte trekking, stelt ons in staat voorspellingen te doen over een nog onbekende steekproef. Ik zal me beperken tot de eerste twee gevallen en die ter verduidelijking schematisch weergeven.

	generaliserende statistiek	toetsing op aselectheid
populatie-gegevens	*onbekend*	bekend
selectie-principe	bekend: aselect	*onbekend*
steekproef (subpopulatie)	bekend	bekend
nulhypothese	aselecte steekproef is afkomstig uit *populatie met bepaalde veronderstelde kenmerken*	de steekproef is een *aselecte steekproef* uit de populatie

Nu is toetsing op aselectheid zinnig indien men van een veronderstelde aselecte steekproef na wil gaan of deze voor die benaming in aanmerking komt, zodat uitspraak 1 reeds op grond daarvan verworpen kan worden. Het wordt interessanter wanneer we opmerken dat een aselecte steekproef een steekproef is waarvan het selectieprincipe onafhankelijk is van de te onderzoeken variabelen, met andere woorden: toetsing op aselectheid is in feite toetsing op onafhankelijkheid tussen selectieprincipe en onderzoeksvariabelen. In de volgende paragraaf zal ik aantonen dat ook voor de onderzoeksvariabelen onderling geldt dat toetsing op onafhankelijkheid in feite neerkomt op een toetsing op aselectheid met behulp van de gebruikelijke randomisatietoetsen. Van de door mij aangehaalde literatuur benadrukt alleen Gold (1969) het gebruik van randomisatietoetsen voor populatiegegevens, zonder echter op te merken dat dergelijke toetsen specifieke toetsen op afhankelijkheid zijn in plaats van redelijke vuistregels om het belang van gevonden associaties te beoordelen. Op dit zwakke punt richt zich de door mij vermelde kritiek van Morrison en Henkel (1969: 134) waarvan ik de onjuistheid in de volgende paragraaf zal laten zien. Voor het overige geeft Galtung (1967) een uitstekende samenvatting van en kritiek op de gebruikelijke opvattingen, maar blijft zelf helaas ook steken in de veronderstelling dat statistische toetsen slechts bruikbaar zijn voor het toetsen van generalisatie hypothesen.

Onafhankelijkheid en randomisatie

Alle randomisatietoetsen kunnen worden opgevat als toetsen op afhankelijkheid tussen variabelen. Bij toetsen voor twee of meer waarnemingsreeksen, zoals bijvoorbeeld de toets van Wilcoxon, is er namelijk in feite sprake van een toetsing op afhankelijkheid tussen de subcategorieën (waarnemingsreeksen) en een onderzoeksvariabele.Daarom zal ik me in het nu volgende beperken tot het bespreken van toetsing op afhankelijkheid voor populatiegegevens. Als voorbeeld neem ik een (kleine) populatie waarvan alle gegevens over twee onderzoeksvariabelen X en Y bekend zijn. Stel dat de variabele X bijvoorbeeld vier mogelijke waarden x_1, x_2, x_3 en x_4 heeft en Y drie mogelijke waarden y_1, y_2 en y_3. Schematisch voorgesteld zien de resultaten er dan als volgt uit:

variabele Y:

variabele X:	y_1	y_2	y_3	
x_1				m_1
x_2				m_2
x_3				m_3
x_4				m_4
	n_1	n_2	n_3	N

De randtotalen voor X en Y zijn resp. m_1 t/m m_4 en n_1 t/m n_3 terwijl de cellen van het schema geacht worden de frequenties van de verschillen x en y combinaties te bevatten. De vraag is nu of er voor de betreffende populatie sprake is van afhankelijkheid, direct of indirect, tussen de variabelen X en Y. Daarbij is het van belang eerst duidelijk vast te stellen wanneer we spreken van afhankelijkheid tussen variabelen. We zeggen dat twee variabelen onderling afhankelijk zijn (binnen een bepaalde populatie) indien van de betreffende populatie-elementen de waarde van ene variabele direkt of indirekt beinvloed is door de waarde van de andere variabele. Is dit niet het geval dan spreken we van onafhankelijkheid tussen de variabelen.
Hiermee zijn we er echter nog niet omdat deze definitie alleen spreekt over het elkaar wel of niet beïnvloed hebben van beide variabelen, maar niet over de wijze waarop zich dat empirisch manifesteert. Om iets te kunnen zeggen

over het wel of niet afhankelijk zijn van de variabelen in de onderzochte populatie is een meer operationele definitie van onafhankelijkheid vereist. Een dergelijke definitie zit impliciet al vervat in de defnitie van een aselecte steekproef:

Een *aselecte steekproef* is een steekproef (subpopulatie) uit een populatie welke is samengesteld door middel van een selectieprincipe dat onafhankelijk is van de te onderzoeken variabelen.

Het belang van deze definitie zit hem in het feit dat daardoor de statistische eigenschappen van aselecte steekproeven in verband worden gebracht met het begrip onafhankelijkheid tussen variabelen. Immers het selectieprincipe kan worden opgevat als een selectievariabele die, onafhankelijk van de onderzoeksvariabelen, twee mogelijke waarden aanneemt zodanig dat daardoor de populatie in twee gedeelten gesplitst wordt: de steekproef en de rest. De definitie van onafhankelijkheid kan daardoor als volgt geformuleerd worden: De variabelen X en Y zijn in de onderzochte populatie onderling *onafhankelijk* dan en alleen dan wanneer de subpopulaties gevormd door de elementen met eenzelfde waarde van Y aselecte steekproeven (zonder teruglegging) zijn uit de populatie van X-waarden.

Voor ons voorbeeld betekent dat dus dat X en Y onafhankelijk zijn als de kolommen van het schema aselecte steekproeven zijn uit de randverdeling voor X. De rol van X en Y (rijen en kolommen) mag hierbij uiteraard verwisseld worden.

Het bewijs dat deze definitie identiek is met de vroegere is eenvoudig en in het voorgaande al min of meer aangegeven. Stel dat X en Y onafhankelijk zijn, dan zijn de Y-waarden niet beïnvloed door die van X. Neem de Y-waarden als selectieprincipe, dan is dat principe onafhankelijk van X-waarden zodat de betreffende subpopulaties (kolommen) aselecte steekproeven uit de populatie zijn. Omgekeerd, als de subpopulaties bij de verschillende waarden van Y aselecte steekproeven uit de populatie van X-waarden zijn, dan is Y een selectievariabele die onafhankelijk is van X, dus X en Y zijn onafhankelijk.

Het voorgaande komt er dus op neer dat de vraag, of X en Y in de onderzochte populatie onderling onafhankelijk zijn, identiek is met de vraag of de kolommen van het schema aselecte steekproeven zijn uit de populatie van X-waarden (waarvan de frequentieverdeling door de randtotalen gegeven wordt). Toetsing op onafhankelijkheid komt dus neer op toetsing op aselectheid (randomisatie). In ons voorbeeld kunnen we toetsen op onafhankelijkheid tussen X en Y door op deze populatiegegevens een chi-kwadraattoets toe te passen. Bij significantie luidt dan de conclusie dat de kolommen (of rijen) geen aselecte steekproeven uit de populatie vormen dus dat X en Y onderling afhankelijk zijn. Dit is dus toetsing van een zogenaamde substantieve hypo-

these in plaats van een generalisatie hypothese zoals gebruikelijk bij de generaliserende statistiek.
Een dergelijke wijze van toetsen is vooral van belang voor kleine populaties waarbij relatief grote afwijkingen van de theoretisch verwachte waarden kunnen optreden zonder dat er sprake is van afhankelijkheid tussen de variabelen. (De theoretisch verwachte waarde voor een combinatie x_i, y_j in het schema is $m_i x\ n_j/N$). Ter verduidelijking van een en ander zullen we nog een simpel getallenvoorbeeld bekijken.[2]

Een voorbeeld

Bij een fabriek werken 40 werknemers welke gelijksoortige arbeid verrichten, 20 werknemers hebben echter een grotere mate van zelfstandigheid bij hun werk dan de overigen en de bedrijfsleiding is er in geïnteresseerd of er een verband bestaat tussen de mate van zelfstandigheid en de productiecijfers van de betrokkenen. Daarbij is men uitsluitend geïnteresseerd in dit ene bedrijf en wordt niet gestreefd naar generalisatie met betrekking tot andere fabrieken of werknemers. De verzamelde gegevens laten het volgende (fictieve) resultaat zien:

		zelfstandigheid: groot		klein		
produktiecijfer:	hoog	5	(6,5)	8	(6,5)	13
	laag	15	(13,5)	12	(13,5)	27
		20		20		40

De getallen tussen haakjes zijn de theoretische verwachtingen bij onafhankelijkheid tussen de twee variabelen. Uit deze gegevens mag men niet de conclusie trekken dat er binnen deze populatie op de een of andere wijze een relatie bestaat tussen beide variabelen, hoewel de cijfers suggereren dat grotere zelfstandigheid iets met lagere produktiecijfers te maken heeft. Dat deze conclusie niet gerechtvaardigd is blijkt bij toetsing van de nulhypothese dat beide variabelen onafhankelijk zijn. Onder deze nulhypothese is de eerste kolom een aselecte steekproef uit de populatie (de tweede kolom uiteraard ook). Toetsing van de hypothese met behulp van chikwadraattoets voor een 2 x 2 tabel (met continuïteitscorrectie) levert een chikwadraatwaarde kleiner dan 1 op welke verre van significant is. Bij berekening van de exactekans op 5 of minder hoge produktiecijfers in de eerste kolom (onder de nulhypothese) blijkt deze gelijk te zijn aan 0,46. Uit deze zeer grote overschrijdingskans blijkt

wel dat er geen reden is te veronderstellen dat de eerste kolom géén aselecte steekproef uit de populatie is. De nulhypothese wordt niet verworpen, uit deze gegevens mag niet de conclusie worden getrokken dat er een relatie bestaat tussen beide variabelen.

Conclusie

Zoals ik in het voorgaande heb getracht duidelijk te maken is het wel degelijk mogelijk en zinvol om significantietoetsen toe te passen op populatiegegevens. Bij kleine populaties is het zelfs noodzakelijk om te voorkomen dat men niet bestaande verbanden signaleert! Dat over deze kwestie zoveel misverstanden bestaan zal voor een deel wel te wijten zijn aan onvoldoende inzicht in het specifieke karakter van de randomisatietoetsen. Belangrijker lijkt mij echter dat in het methoden en technieken onderwijs en de daarbij gebruikte handboeken te weinig wordt ingegaan op de vraag wat we bedoelen als we zeggen dat twee variabelen onderling onafhankelijk zijn binnen een beperkte populatie. Meestal wordt het begrip afhankelijkheid ingevoerd met behulp van associatiematen en een oneindig grote populatie in het achterhoofd, waarbij iedere afwijking van de theoretische verwachting in die populatie (uiteraard) wijst op afhankelijkheid tussen de variabelen. Dat, bij onafhankelijkheid tussen variabelen in een kleine populatie, afwijkingen van de theoretisch verwachte aantallen normaal zijn wordt dan kennelijk over het hoofd gezien.
Het zal tenslotte duidelijk zijn dat ik de door Lammers (1963) en Nooij (1969) toegepaste significantietoetsing juist acht, zij het op andere dan door de auteurs aangevoerde gronden. De door hen uitgevoerde significantietoetsen zijn geldig voor het toetsen van verbanden of verschillen binnen de populatie van onderzochte waarnemingseenheden (respondenten), maar generalisatie naar een andere populatie is op grond daarvan niet mogelijk als de onderzochte waarnemingseenheden geen aselecte steekproef uit die populatie zijn.

Noten

1 Van dergelijke abstracte populaties is sprake bij experimentele of quasi-experimentele proefopzetten. De daarbij gebruikte analyse-technieken zoals regressie-, variantie- en factoranalyse zijn gebaseerd op een additief mathematisch model dat ook een zogenaamde foutenterm ('error') bevat.
2 Aardige voorbeelden van andere dan de gebruikelijke randomisatietoetsen kan men o.a. vinden in Winch en Campbell (1969).

Literatuur

M. Albinski (1971), *Survey-research,* Aula 345, Het Spectrum, Utrecht.
A. van den Ban, P. Swanborn, A. Nooy (1969), Kommentaren (en antwoord) op

'De Boerenpartij' door A. Nooy',, *Mens en Maatschappij,* 44, p. 319-329.
J. Beshers (1958), On 'A critique of tests of significance in survey research', *American Sociological Review,* 23, p. 199.
M. Brouwer en R. Mokken (1959), 'Over de bruikbaarheid van statistische toetsen in het sociaal-wetenschappelijk onderzoek', *Sociologische Gids,* 6, p. 81-88.
R. Freedman (1958), 'Some observations on sampling tests', *Sociologische Gids,* 5, p. 114-118.
J. Galtung (1967), *Theory and Methods of Social Research,* Allen & Unwin, London, p. 358-388.
R. Mc. Ginnis (1958), 'Randomization and inference in sociological research', *American Sociologcal Review,* 23, p. 408-414.
D. Gold (1958), 'Comment on 'A critique of tests of significance', *American Sociological Review,* 23, p. 85-86.
D. Gold (1969), 'Statistical tests and substantive significance', *The American Sociologist,* 4, p. 42-46.
C. J. Lammers (1963), *Het Koninklijk Instituut voor de Marine,* Van Gorcum, Assen.
D. Morrison, R. Henkel (1969), 'Significance tests reconsidered, *The American Sociologist,* 4, p. 131.
A. Nooy (1969), *De Boerenpartij: Desoriëntatie en radicalisme onder de Boeren,* Boom en Zoon, Meppel.
C. Sellitz e.a. (1967), *Research Methods in social relations,* Revised Edition, Holt, Rinehart and Winston, New York.
H. Selvin (1957), 'A critique of tests of significance in survey research', *American Sociological Review,* 22, p. 519-527.
H. Selin (1958a), 'Reply to Gold's Comment on A Critique of tests of significance', *American Sociological Review,* 23, p. 86.
H. Selvin (1958b), 'Reply to Beshers', *American Sociological Review,* 23, p. 199-200.
P. Swanborn (1972), *Aspecten van sociologisch onderzoek,* 2e druk, Boom en Zoon, Meppel.
R. F. Winch, D. T. Campbell (1969), 'The significance of tests of significance', *The American Sociologist',* 4, p. 140-143.