J. A. KNOTTNERUS EN A. VOLOVICS

Statistisch toetsen

Indien men onderzoek doet bij een aselecte steekproef uit een doelpopulatie, kunnen tot op zekere hoogte verschillende resultaten worden gevonden in verschillende steekproeven, terwijl andersom de bij één steekproef gevonden resultaten kunnen passen bij verschillende toestanden in de doelpopulatie. Statistische toetsing is een manier om na te gaan, in hoeverre toch bepaalde uitspraken gedaan kunnen worden, en met welke mate van onzekerheid daarbij rekening moet worden gehouden.

Inleiding

In de medisch wetenschappelijke literatuur heeft men aan oncontroleerbare beweringen even weinig als aan open deuren. Het gaat erom onderbouwde uitspraken te doen met betrekking tot tevoren nog onvoldoende ontgonnen terreinen. In de regel bestaat de onderbouwing uit door middel van onderzoek verzamelde gegevens, terwijl met behulp van de medische statistiek antwoord wordt gegeven op de vraag in hoeverre bepaalde uitspraken of hypothesen de toetsing aan deze onderzoeksgegevens kunnen doorstaan.

Hierbij is ook de generaliseerbaarheid van de resultaten van belang, dat wil zeggen: de geldigheid ervan, niet alleen voor de beperkte groep onderzochten (*de onderzoekspopulatie*) maar ook voor een meer algemene *doelpopulatie*, waarop men de resultaten zou willen toepassen. Hiertoe dient aan een aantal vereisten te worden voldaan:

- *Validiteit.* De onderzoeks*opzet* en *-uitvoering* moeten geschikt zijn om uitsluitsel te geven met betrekking tot de beoogde uitspraken en de onderzoekspopulatie moet een goede afspiegeling zijn van de doelpopulatie.
- De *omvang* van de onderzoekspopulatie moet zodanig zijn dat conclusies met een voldoende mate van zekerheid kunnen worden getrokken.[1]

Statistische analyse is niet zinvol, en mogelijk zelfs misleidend, indien het onderzoek niet valide is.[1] Dat kan geïllustreerd worden aan de hand van het artikel 'Symptomatologie en selectiebias'.[2] Indien men het verband tussen moeheid en bloedarmoede onderzoekt en daartoe de frequentie van moeheid bij door huisartsen herkende anemiepatiënten (69 procent) vergelijkt met de overeenkomstige frequentie bij een even grote steekproef uit de overige spreekuurbezoekers (18 procent), dan is dit verschil groot en statistisch sterk 'significant' ($p < 0.01$) (*tabel 1*). Het verschil is echter louter een produkt van selectiebias, veroorzaakt door het feit dat bij moeheidsklachten veel vaker bloedonderzoek plaatsvindt.[2]

Over de interne validiteit van de onderzoeksopzet komen we later nog te spreken; in deze aflevering gaan we hierop niet meer dan terloops in.

Gehele populatie of een steekproef?

Maximale zekerheid met betrekking tot verbanden tussen kenmerken en verschijnselen, of met betrekking tot de effecten van interventies, is te verkrijgen door de *gehele doelpopulatie* te onderzoeken. Een praktische voorwaarde is dan wel, dat deze populatie duidelijk is afgebakend naar plaats en, niet te vergeten, naar tijd. De samenstelling van een populatie op dit moment zal met betrekking tot een aantal kenmerken niet gelijk zijn aan de samenstelling van 'dezelfde' populatie over 20 jaar of zelfs over een week. Als een huisarts op 1 januari 1988 bijvoorbeeld vaststelt dat van de 1290 mannen in zijn praktijk 1,7 procent 80 jaar of ouder is, tegen 3,1 procent van de 1316 ingeschreven vrouwen, dan staat op dat moment vast dat het percentage hoogbejaarden in de onderzochte populatie groter is bij de vrouwen dan bij de mannen. Statistisch toetsen om na te gaan of deze waarneming van toepassing mag worden verklaard op de doelpopulatie, is dan niet zinvol, of zelfs misplaatst.[3]

Ook *Van Eijk e.a.* zien af van statistische toetsing in hun onderzoek naar het verschil in morbiditeitspatroon tussen nabestaanden van patiënten die zijn overleden aan een chronische en aan een acute ziekte. De reden hiervoor is dat de totale praktijkpopulatie van dat moment in het onderzoek werd betrokken.[4] Als men deze populatie zou willen (en mogen) zien als een steekproef uit bijvoorbeeld de Nederlandse bevolking, zou statistische toetsing wél in aanmerking komen.

In de praktijk wordt er meestal van uitgegaan dat er sprake is van een – in omvang en in de tijd – beperkte steekproef uit een doelpopulatie. Dit gebeurt deels op praktische gronden: men heeft zelden de middelen om de gehele doelpopulatie te onderzoeken. Bovendien worden vaak variabelen bestudeerd die, zoals de bloeddruk, nogal wat biologische variatie in de tijd kennen en waarbij de toestand op één bepaald moment de resultante van een min of meer toevallige samenloop van omstandigheden is. De belangrijkste reden is echter, dat het bij wetenschappelijk onderzoek in de regel niet gaat om de situatie in een bepaalde, aan tijd en plaats gebonden populatie; men streeft een meer algemene geldigheid na. In feite is dan iedere denkbare onderzoekspopulatie een aselecte steekproef uit een bepaalde, deels zelfs toekomstige 'superpopulatie' van menselijke ervaring. Bij generalisatie van het onderzoeksresultaat spelen dan, naast het toetsen, ook de wetenschappelijke voorkennis en beoordeling een rol. Resultaten van een wetenschappelijk onderzoek staan immers nooit op zichzelf, maar moeten gezien worden in het grotere verband van resultaten van andere onderzoekingen op hetzelfde gebied.

Variabiliteit

Indien men onderzoek doet bij een aselecte *steekproef* uit een al dan niet scherp afgebakende doelpopulatie, blijft er met betrekking tot de generaliseerbaarheid van de resultaten een bepaalde mate van onzekerheid bestaan. De bronnen voor deze onzekerheid zijn:

- *Het werken met steekproeven.* De relaties of effecten die men in een bepaalde aselecte steekproef vindt, zullen

Rijksuniversiteit Limburg, Postbus 616, 6200 MD Maastricht.
Prof. dr. J.A. Knottnerus, hoogleraar huisartsgeneeskunde, vakgroep Huisartsgeneeskunde; A. Volovics, statisticus, vakgroep Medische Informatica en Statistiek.
Correspondentie: Prof. dr. J.A. Knottnerus.

niet precies dezelfde zijn als in andere mogelijke aselecte steekproeven uit dezelfde populatie. Soms zijn er zelfs grote verschillen te vinden.

- *Meetonnauwkeurigheden*. Veel kenmerken, zoals geslacht, leeftijd, een Collesfractuur of de bloedgroep, kunnen zeer nauwkeurig worden bepaald. Andere daarentegen, zoals de bloeddruk en de duur van de vruchtbare periode, worden vaak met een meer of minder grote meetfout bepaald.
- *Biologische variabiliteit*. Bijna alle biologische en gedragswetenschappelijke parameters vertonen – soms grote -verschillen tussen (ook gezonde) personen. Daarnaast zijn er vaak aanzienlijke schommelingen in de tijd te constateren bij eenzelfde individu.

Ten aanzien van deze drie aspecten bestaat er een bepaalde mate van *variabiliteit*, onafhankelijk van het bestaan van een eventuele systematische vertekening (bias) bijvoorbeeld ten gevolge van een verkeerde instelling of ijking van het meetinstrument. Het gevolg van variabiliteit is dat tot op zekere hoogte verschillende onderzoeksresultaten gevonden kunnen worden in verschillende aselecte steekproeven (onderzoekspopulaties) uit dezelfde doelpopulatie, en andersom, dat de bij één steekproef waargenomen resultaten kunnen passen bij verschillende toestanden in de doelpopulatie.

Statistische toetsing

Statistische toetsing is een manier om na te gaan, in hoeverre, gegeven de genoemde variabiliteit, toch bepaalde uitspraken gedaan kunnen worden, en met welke mate van onzekerheid daarbij rekening moet worden gehouden.

In het onderzoek van *Smits e.a.*[5] werd in een steekproef van spreekuurbezoekers onder meer bestudeerd, of patiënten die door anderen uit hun omgeving naar de dokter waren gestuurd, vervolgens vaker naar de specialist werden verwezen. Van de patiënten die door anderen waren gestuurd, bleek 14 procent te worden verwezen, tegen 8 procent van degenen die zelfstandig hadden besloten de dokter te raadplegen. Voor dit verschil wordt een p-waarde van 0.06 vermeld (chikwadraattoets).

De p-waarde – die kan variëren van 0 tot 1 – geeft aan in welke mate een gevonden verschil (hier dus het verschil tussen de percentages 8 en 14 procent) in overeenstemming is met de 'nulhypothese'. In dit geval luidt de nulhypothese dat de kansen op verwijzing naar de specialist in beide groepen *even groot* zijn. Hoe groter de p-waarde, des te groter de kans dat het gevonden resultaat kan worden aangetroffen als de nulhypothese in werkelijkheid waar is. En hoe kleiner de p-waarde, des te onwaarschijnlijker wordt het dat het betreffende resultaat wordt gevonden als er in werkelijkheid geen verschil is.

Het is dus begrijpelijk dat de p-waarde kleiner is naarmate een groter verschil tussen de percentages wordt gevonden: het wordt immers steeds minder waarschijnlijk dat dit verenigbaar is met de nulhypothese 'geen verschil', en er is dus meer reden om aan te nemen dat een andere ('alternatieve') hypothese beter bij de gegevens past ('er is wél verschil'). Bij een zeer gering verschil tussen de gevonden percentages is dus een hogere p-waarde te verwachten.

Men dient te beseffen dat men bij de berekening van de p-waarde bepaalde veronderstellingen moet doen op grond van een theoretisch model, afhankelijk van de gebruikte statistische toets. Dat is nodig omdat men de 'werkelijkheid', bijvoorbeeld met betrekking tot de vorm van de verdelingen van de bestudeerde variabelen in de doelpopulatie, niet met zekerheid kent. In het algemeen echter is, bij keuze van een geschikte toets, de p-waarde een goede schatting van de kans dat, gegeven het betreffende onderzoek, een bepaald verschil of een nog groter verschil wordt aangetroffen, indien in werkelijkheid de nulhypothese ('er is geen verschil') waar is. Als men dus op grond van de onderzoeksgegevens van *Smits e.a.* concludeert dat er een verschil bestaat, loopt men een risico van 6 procent, dat indien de nulhypothese waar is, deze toch wordt verworpen.

In de praktijk hanteert men vaak een bepaalde vooraf vastgestelde drempelwaarde, meestal een p-waarde van 0.05. Hieraan wordt dan de term 'statistische significantie' verbonden: als de p-waarde lager dan of gelijk aan 5 procent is ($p \leq 0.05$), is er sprake van een statistisch significant verschil op het 5-procentniveau. Deze drempelwaarde kan natuurlijk ook strenger (bijvoorbeeld op 0.01) of minder streng (bijvoorbeeld op 0.10) gesteld worden.

Omdat echter elke drempel arbitrair is, valt er veel voor te zeggen om de gevonden p-waarde zelf te vermelden, zoals *Smits e.a.* in dit geval ook gedaan hebben. Het is dan aan de lezer om uit te maken hoeveel gewicht hij hieraan toekent. Bovendien wordt dan ook voorkomen, dat de rapporteur zich in allerlei bochten gaat wringen, bijvoorbeeld door te stellen dat het gevonden verschil bijna, of net niet significant was op het 5-procent-niveau. Het gaat hier om een glijdende schaal en het is niet nodig een p-waarde van 0.43 en een p-waarde van 0.06 over één kam te scheren. In het tweede geval is er duidelijk meer reden om aan de nulhypothese te twijfelen, ook al zijn beide p-waarden groter dan 0.05. Evenzo is op deze manier direct duidelijk, dat het verschil tussen een 'significante' p-waarde van 0.04 en een 'niet-significante' p-waarde van 0.06 niet groot is, in tegenstelling tot het verschil tussen de p-waarden 0.04 en 0.43.

Vaak kan men er echter niet omheen om drempelwaarden te formuleren, bijvoorbeeld als besloten moet worden of een gevonden verschil moet leiden tot een vervolgonderzoek, of als men vóór het onderzoek van start gaat, een schatting wil maken van de benodigde grootte van de onderzoekspopulatie (waarover later meer).

Toetsen

De p-waarde kan met behulp van verschillende statistische toetsen worden berekend, afhankelijk van de variabelen waarmee men te maken heeft. Meestal gebeurt dit tegenwoordig door de computer, die men dan wel de juiste toets moet opgeven. Bepaalde eenvoudige toetsen zijn ook uit te voeren met behulp van een zakrekenmachine.

De *chikwadraat-toets* wordt vaak gebruikt bij twee dichotome variabelen, waarbij de waarnemingen in de vorm van een vierveldentabel kunnen worden samengevat. Dit is bijvoorbeeld het geval als het te toetsen verband als relatief risico of odds-ratio wordt vermeld, maar ook bij het onderzoek van *Smits e.a.*, waarbij het gaat om het verschil in percentages verwijzingen tussen twee groepen.

Als men te maken heeft met een numerieke variabele, zoals de bloeddruk, en de verdeling hiervan in twee groepen wil vergelijken, kan de zoge-

naamde *Student-toets* (ook wel *t-toets* genoemd) worden gebruikt. Als het verband tussen twee numerieke variabelen wordt bestudeerd, kan men toetsen in welke mate de waargenomen regressiecoëfficiënt (zie de vorige aflevering van deze serie[6]) overeenstemt met de nulhypothese 'geen verband'.

Behalve de aard van de variabelen zijn er ook andere factoren die de keuze van de te gebruiken toets bepalen, en voor iedere toets moet aan bepaalde voorwaarden worden voldaan. Er bestaat een groot aantal toetsen en het consulteren van een statisticus is vaak nuttig om tot een juiste keuze te komen.

Een belangrijk punt is dat bij de meeste statistische toetsen (waaronder de chikwadraat- en de t-toets) ook de grootte van de onderzoekspopulatie wordt betrokken. Omdat bij een grote steekproef nauwkeuriger schattingen mogelijk zijn dan bij kleine, zal eenzelfde verschil in percentages in het eerste geval een lagere p-waarde bereiken dan in het tweede geval. Dat kan tweeërlei interpretatieproblemen opleveren. Bij zeer grote populaties kan bijvoorbeeld een miniem en irrelevant verschil in gemiddelde bloeddruk (1 mm Hg) nog statistisch 'significant' zijn. Anderzijds zal in zeer kleine steekproeven zelfs een groot en op zichzelf relevant verschil (bijvoorbeeld 15 mm Hg) nog niet de gekozen significantiedrempel overschrijden. Zo'n bevinding betekent dan niet zozeer dat er geen relevant verschil is, maar dat er in het onderzoek gewoon te weinig bewijsmateriaal voorhanden was om tot een eventueel verschil te kunnen concluderen. Het besef hiervan ligt ten grondslag aan de gewoonte om vóór het uitvoeren van een onderzoek te bepalen welk verschil men nog *relevant* genoeg vindt om door middel van statistische toetsing te kunnen 'aantonen'. Dit nog relevant geachte verschil is direct van invloed op de minimaal benodigde steekproefomvang.

Fouten van de eerste en tweede soort

Het is inmiddels duidelijk dat men bij het doen van uitspraken op grond van onderzoeksresultaten twee soorten fouten kan maken:

- als in werkelijkheid de nulhypothese (bijvoorbeeld: er is geen verschil) waar is, kan men toch concluderen tot het bestaan van een verschil; dit noemt men een *fout van de eerste soort*;
- als in werkelijkheid niet de nulhypothese, maar een alternatieve hypothese (bijvoorbeeld: er is een relevant verschil) waar is, kan men toch concluderen tot het niet bestaan van een verschil; dit is een *fout van de tweede soort* (*tabel 2*).

Het spreekt vanzelf dat men de kans op deze twee fouten binnen de perken wil houden. Geheel uit te sluiten zijn ze nooit, maar men kan wel bepaalde grenzen afspreken. We hebben gezien dat 5 procent (0.05) vaak nog wordt geaccepteerd als kans dat een geldige nulhypothese bij statistische toetsing ten onrechte wordt verworpen (fout van de eerste soort). Deze of een andere gekozen grenswaarde wordt ook wel α genoemd. Alleen als men een p-waarde lager of gelijk aan α vindt ($p \leq \alpha$) – bij een significant resultaat dus – zal men de nulhypothese verwerpen.

De kans dat bij een statistische toetsing op een bepaald significantie niveau een *niet*-geldige nulhypothese toch *niet* wordt verworpen (fout van de tweede soort), noemt men β. Dikwijls worden hiervoor waarden van 10 procent (0.1) of 20 procent (0.2) aanvaardbaar geacht. Het komt erop neer, dat men bij het vinden van een p-waarde groter dan α (een 'niet-significant' resultaat dus) pas aanneemt dat de nulhypothese waar is, als de kans dat de nulhypothese ten onrechte *niet* wordt verworpen, kleiner dan of gelijk aan een bepaalde β is. Is de laatstgenoemde kans groter dan β, dan zit men met het probleem dat noch de nulhypothese, noch de alternatieve hypothese mag worden verworpen.

In de statistiek wordt in dit verband ook gesproken van de *power* van een onderzoek: de kans dat een bepaald (klinisch relevant) verschil, indien aanwezig, ook inderdaad een significante p-waarde oplevert. Deze *power* is gelijk aan 100 procent-β. Meer intuïtief geformuleerd, kan de power opgevat worden als het vermogen van een bepaald onderzoek om een relevant verschil op te sporen.

Steekproefomvang

De kans op een fout van de tweede soort is afhankelijk van de drempelwaarde die men voor de fout van de eerste soort kiest: hoe stringenter het significantieniveau dat men hanteert, des te hoger wordt de kans op een fout van de tweede soort en des te lager dus de *power*. Dit ligt voor de hand: hoe hoger de eisen zijn die ten aanzien van het verwerpen van de nulhypothese worden gesteld, des te moeilijker is het om hieraan te voldoen.

Als de grootte van de onderzoekspopulatie reeds vastligt, zal deze 'concurrentie' van beide fouten onvermijdelijk zijn, en kunnen ze niet beide tegelijk worden teruggedrongen. Het is daarom verstandig zich vóór het definitief plannen van een onderzoek rekenschap te geven van de *zekerheid* waarmee men op grond van de resultaten bepaalde uitspraken wil kunnen doen. Die gewenste zekerheid is het uitgangspunt voor de schatting van de benodigde omvang van de onderzoekspopulatie op basis van:

- vooraf gespecificeerde kansen op fouten van de eerste (α) en de tweede soort (β) die nog aanvaardbaar zijn;
- een vooraf gespecificeerd minimaal (relevant) verschil.

In de verslaglegging van verricht medisch onderzoek krijgt de fout van de tweede soort veel minder aandacht dan de fout van de eerste soort: men besteedt meestal alleen aandacht aan statistisch significante resultaten ($p \leq \alpha$). Voor het beoordelen van de betekenis van het *niet* kunnen aantonen van een relevant verband in een bepaald onderzoek, kan het echter nuttig zijn ook de waarde van β te vermelden. Indien deze hoog is, bevat de studie eenvoudigweg te weinig informatie om een zinvolle uitspraak te doen over het al dan niet bestaan van een relevant verband. De oorzaak hiervan is vaak dat de onderzoekspopulatie te klein was: meestal heeft men om een bepaald verband met een redelijke mate van betrouwbaarheid *uit te sluiten* een grote onderzoekspopulatie nodig. Als men bijvoorbeeld een verschil in verwijspercentages tussen twee groepen van 8 procent versus 14 procent een relevante bevinding zou vinden, dan zou men, uitgaande van een α van 0.05 en een β van 0.10, in elke groep 566 spreekuurbezoekers nodig hebben om een uitspraak te kunnen doen over het wel of niet bestaan van zo'n verschil.

In *tabel 3* is het verband weergegeven tussen verschillende (gebruikelijke) waarden van α, β en de minimaal gewenste steekproefomvang, indien men ervan uitgaat dat bij statistische toetsing

Tabel 1 *Het voorkomen van moeheid bij door de huisarts herkende anemiepatiënten, vergeleken met een even grote steekproef uit de overige spreekuurbezoekers.*

	Anemiepatiënten		Steekproef uit overige patiënten	
	n	%	n	%
Moeheidsklachten	80	69%	21	18%
Geen moeheidsklachten	36	31%	95	82%
Totaal	116	100%	116	100%

Volgens chikwadraat-toets, tweezijdig: $p < 0.001$. Hier wordt een groot, statistisch significant verschil gevonden, geheel ten gevolge van selectiebias.[2]

Tabel 2 *De fout van de eerste soort en de fout van de tweede soort op grond van de uitkomst van een onderzoek naar het verband tussen twee variabelen X en Y (bijvoorbeeld X: al dan niet door anderen naar de dokter zijn gestuurd; Y: het verwijspercentage).*

		Werkelijke toestand in de doelpopulatie	
		Er is geen verband tussen X en Y (nulhypothese is waar)	Er is geen verband tussen X en Y (alternatieve hypothese is waar)
Conclusie op grond van de bevindingen bij de onderzoekspopulatie	Er is geen verband tussen X en Y	Correcte conclusie	Fout van de eerste soort
	Er is een verband tussen X en Y	Fout van de eerste soort	Correcte conclusie

Tabel 3 *Minimaal benodigde steekproefomvang per subgroep, indien men een eventueel verschil tussen twee groepen van 8% versus 14% nog statistisch 'significant' ($p \leq \alpha$) wil kunnen aantonen ('tweezijdige' toetsing en gelijke groepsgrootte).*

β	α		
	0.01	0.05	0.10
0.05	960	701	582
0.10	803	566	464
0.20	630	426	334

Hoe kleiner α en/of β worden gekozen, des te groter de benodigde steekproef.

Tabel 4 *Minimaal benodigde steekproefomvang per subgroep, wanneer men een bepaald minimaal verschil in percentages nog 'significant' ($p \leq \alpha$) wil kunnen aantonen. Uitgegaan wordt van een α van 0.05 en β van 0.10.*

Nog aan te tonen minimaal verschil in procenten		Benodigde steekproefomvang per subgroep
groep 1	groep 2	
8	10	4295
8	14	566
8	20	171
8	25	95
8	35	44
8	45	25

een verschil tussen de percentages 8 procent en 14 procent nog als 'significant' (d.w.z. $p \leq \alpha$) moet kunnen worden aangetoond (uitgaande ook van gelijke grootte van de subgroepen). We zien hierin het volgende geïllustreerd:
- naarmate α en β stringenter (dus lager) worden gesteld, wordt de gewenste steekproefomvang groter;
- bij een gelijkblijvende α kan alleen een kleinere β worden bereikt door een grotere steekproef te nemen.

Uit *tabel 4* blijkt voorts dat, gegeven een bepaalde α en β, de steekproefomvang sterk toeneemt naarmate men een kleiner verschil nog wil kunnen aantonen. De relevantie van een verschil tussen 8 procent en 10 procent is echter hoogst twijfelachtig.

Bij deze berekeningen is overigens uitgegaan van zogenaamde tweezijdige toetsing, en daarover gaat de volgende paragraaf.

Eenzijdig of tweezijdig toetsen

Bij statistische toetsing van een verschil of verband kan men *vooraf* kiezen tussen 'eenzijdig' en 'tweezijdig' toetsen. Dit komt erop neer dat men in het eerste geval alleen geïnteresseerd is in één bepaalde richting van het verband, bijvoorbeeld: percentage P1 is hoger dan percentage P2, gemiddelde M1 is hoger dan gemiddelde M2, het relatief risico R is groter dan 1. In het tweede geval gaat het om de meer open vraag of er tussen twee percentages of gemiddelden verschillen bestaan, en of het relatief risico een andere waarde heeft dan 1.0.

Doordat men bij eenzijdige toetsing niet ook nog te maken heeft met de kans op een foutieve conclusie in de andere richting, wordt eerder een significant resultaat bereikt. Een lage p-waarde of een hoge *power* wordt al gehaald bij een minder sterk verband of een minder grote onderzoekspopulatie dan bij tweezijdige toetsing. De p-waarde is bij eenzijdige toetsing dan twee keer zo klein bij dezelfde gegevens, en dat is aantrekkelijk.

Over de vraag of en wanneer men eenzijdig mag toetsen, lopen de meningen uiteen. Volgens sommigen impliceert het doen van een onderzoek, dat niet met zekerheid bekend is in welke richting een verband zal gaan.[7] Ook zijn er auteurs die een gerandomiseerd experiment niet ethisch verantwoord vinden, als er al bij voorbaat van wordt

uitgegaan dat de behandeling van de ene groep patiënten hoogstens gelijkwaardig en mogelijk slechter zou zijn dan die van de andere groep. *Miettinen* daarentegen stelt dat er altijd voorkennis is en dat men zelden 'blanco' aan een onderzoek begint.[8] Een hypothese, en dus ook de onderzoeksbelangstelling, is volgens hem per definitie eenzijdig gericht.

Men kan in deze een eigen positie bepalen. Gezien de discussie over dit onderwerp doet men er echter verstandig aan in de regel tweezijdige toetsing toe te passen. Men heeft dan aan strengere eisen voldaan, en de conclusies zullen als meer overtuigend worden ervaren. Er is dan wel een grotere onderzoekspopulatie nodig. Verder dient men vanzelfsprekend zijn keuze te verantwoorden in de rapportage over het onderzoek.

Betrouwbaarheidsintervallen

Het onderwerp van de statistische significantie is zeer inzichtelijk te benaderen via zogenaamde betrouwbaarheidsintervallen (meestal 95 procent). De gedachtengang voor de interpretatie van een 95-procent-betrouwbaarheidsinterval is dan als volgt: voor doelpopulaties waarvoor de onderzoekspopulatie representatief is, mag men ervan uitgaan dat de kans 95 procent is, dat de te meten parameter binnen de grenzen van het te vinden betrouwbaarheidsinterval zal liggen. In het besproken voorbeeld van het artikel van *Smits e.a.* is de 'puntschatting' van het verschil in percentages 14 procent – 8 procent = 6 procent, en zouden de grenzen van het betrouwbaarheidsinterval op respectievelijk –0,1 procent en 12,1 procent kunnen worden gesteld. Het feit dat dit interval de waarde 0 bevat, geeft aan dat het verschil niet significant van 0 verschilt als men zou toetsen op het 5-procent-significantieniveau; met andere woorden: de p-waarde is groter dan 0.05.

In het algemeen: hoe smaller het betrouwbaarheidsinterval, des te nauwkeuriger is de 'schatting'. Natuurlijk maakt men ook bij de berekening van betrouwbaarheidsintervallen gebruik van bepaalde vooronderstellingen, die beter opgaan naarmate de onderzoekspopulatie groter is. Betrouwbaarheidsintervallen kunnen worden afgeleid voor alle denkbare parameters, zoals (een verschil tussen) gemiddelden, (een verschil tussen) percentages, relatief risico en regressiecoëfficiënt. In het algemeen worden ze rechtstreeks berekend uit de in de onderzoekspopulatie vastgestelde waarde van de parameter (bijvoorbeeld het verschil tussen twee percentages) en de maat voor precisie van die schatting: de zogenaamde standaard fout (standard error, SE). Het 95 procent betrouwbaarheidsinterval wordt dan in de regel bepaald door: gemeten waarde $\pm$ 1.96 SE.

Uiteraard kan men ook andere betrouwbaarheidsgrenzen hanteren, bijvoorbeeld 90 procent of 99 procent. De genoemde vermenigvuldigingsfactor 1.96 voor de SE moet dan worden vervangen door respectievelijk 1.64 of 2.58.

Het grote voordeel van een betrouwbaarheidsinterval ten opzichte van louter de vermelding van een verschil met de bijbehorende p-waarde is, dat dit inzichtelijker is en meer informatie bevat over de mate van nauwkeurigheid van de schatting van het verschil. Men heeft bovendien een indruk van alle mogelijke hypothesen die nog met de waarnemingen in overeenstemming zijn bij toetsing op het gehanteerde significantieniveau.

Ook voor de interpretatie van betrouwbaarheidsintervallen geldt dat rekening moet worden gehouden met de *power* van de studie. Naarmate de studieomvang kleiner is, groeit de kans dat een betrouwbaarheidsinterval breed uitvalt, en de 'nul-waarde' (= 'geen verband') omvat. Dit houdt in dat het risico op het 'missen' van verbanden dan groter is.

Slotopmerkingen

De lezer van onderzoeksverslagen kan de steekproefomvang niet plannen, maar krijgt deze voorgeschoteld. Daarom is het in het licht van het besprokene nuttig de volgende situaties te onderscheiden:

- Bij zeer kleine onderzoekspopulaties is de kans klein dat verschillen, ook al zijn ze reëel en zelfs sterk, een lage p-waarde of een smal betrouwbaarheidsinterval kunnen opleveren. De *power* is laag. Men heeft in deze gevallen weinig aan statistische toetsing, en deze kan, ter voorkoming van misverstanden, vaak beter achterwege blijven (zo het onderzoek zelf al zinvol is).
- Bij extreem grote onderzoekspopulaties zal een gering, klinisch niet relevant verschil al gauw een zeer lage p-waarde of een smal betrouwbaarheidsinterval hebben. De *power* is hoog voor minimale verschillen, soms te hoog. Ook in dit geval is statistisch toetsen dus van twijfelachtig nut, en is de onderzoekspopulatie onnodig groot. Men kan statistisch toetsen in dit geval vergelijken met het opsporen van extrasystolie met een 24-uurs ECG als men van een positief resultaat spreekt zodra één extrasystolie wordt gevonden.
- Het nut van statistisch toetsen is het grootst wanneer de onderzoekspopulatie noch zeer klein, noch extreem groot is. Er moet dan een redelijk sterk, relevant, verband zijn voordat er een zeer lage p-waarde uit de bus komt. Wat 'middelgroot' is, is niet in één getal aan te geven. Dit hangt onder andere af van hoe vaak de te bestuderen factoren voorkomen, welk verschil men relevant vindt, en welke risico's op foute conclusies men wil accepteren.[9]

Een veelvoorkomend misverstand is, dat de p-waarde de kans is dat, gegeven het gevonden onderzoeksresultaat, de nulhypothese waar is en dat er dus geen verband is. Deze gedachte berust op een verkeerd ('omgekeerd') begrip van de p-waarde. Deze is immers niet meer dan de uitdrukking van de kans dat, *gegeven de nulhypothese*, de aangetroffen resultaten zouden worden gevonden. De uiteindelijke beoordeling (van de kans) of er een verband is, is een kwestie van interpretatie door onderzoekers en lezers, en daarvoor draagt niet alleen het onderzoeksresultaat maar ook de wetenschappelijke voorkennis bouwstenen aan. De vergelijking kan getrokken worden met de huisarts die zijn diagnostische hypothesen toetst aan de uitslag van het laboratoriumonderzoek. Deze uitslagen kunnen beter passen bij de ene dan bij de andere hypothese, maar de diagnostische conclusies zal de arts zelf moeten trekken mede in het licht van a priori informatie.

In de praktijk spelen ook overwegingen met betrekking tot de haalbaarheid een rol bij het bepalen van de omvang van de onderzoekspopulatie. Een belangrijk probleem is ook, dat men lang niet altijd voldoende voorinformatie heeft om de gewenste berekeningen uit te voeren. Tenslotte kunnen er verschillen van opvatting bestaan over de vraag

welk verschil nog klinisch relevant is. Indien men evenwel de principes van toetsing en van de schatting van de benodigde steekproefomvang begrijpt, kan men de hiermee samenhangende beslissingen beter onderbouwd nemen.

[1] Knottnerus JA, Volovics A. Medisch-statistische besprekingen voor de huisarts. Huisarts Wet 1987; 30: 349-50, 356.

[2] Knottnerus JA, Knipschild PG, Sturmans F. Symptomatologie en selectiebias. Vertekening van het verband tussen klachten en diagnosen ten gevolge van selectie naar hogere echelons. Huisarts Wet 1985; 28: 325-330.

[3] Van Eijk JThM, Gubbels JW. Wetenschappelijk onderzoek in de huisartsgeneeskunde. 2e dr. Lelystad: Meditekst, 1987.

[4] Van Eijk J, Smits A, Huygen F, Van den Hoogen H. Overlijden als gevolg van chronische of acute ziekte en het morbiditeitspatroon van de nabestaanden. Huisarts Wet 1987; 30: 336-9, 341.

[5] Smits A, Van der Grinten R, Huygen F, Van den Hoogen H. 'Ze stuurden me naar de dokter'. Een gezinsgeneeskundig signaal. Huisarts Wet 1987; 30: 377-80.

[6] Knottnerus JA, Volovics A. Correlatie en regressie. Huisarts Wet 1988; 31: 18-22.

[7] Colton T. Statistics in medicine. Boston: Little, Brown and Company, 1974.

[8] Miettinen OS. Theoretical epidemiology. New York: John Wiley & Sons, 1985.

[9] Pocock SJ. Clinical trials. A practical approach. New York: John Wiley & Sons, 1983.

In dit nummer van *Huisarts en Wetenschap* plaatst de Maastrichtse huisarts Rethans een aantal kanttekeningen bij het standaardenbeleid van het NHG. De redactiecommissie is van mening dat met het naschrift van de voorzitter van het Genootschap het laatste woord over dit onderwerp nog allerminst is gezegd; zij nodigt daarom de lezers nadrukkelijk uit te reageren op deze discussie.

Standaardenbeleid

In het oktobernummer van *Huisarts en Wetenschap* 1987 refereert Tielens als voorzitter van het NHG opnieuw aan het standaardenbeleid van het Genootschap.[1] Hij spreekt in dit verband onder meer over de wetenschappelijke taak die het NHG als organisatie heeft, en over het spanningsveld tussen het NHG en perifere huisartsen. Op de Referatendag van 1986 sprak Tielens' voorganger Bottema ook over het standaardenbeleid en daarbij waarschuwde hij voor de breuk die aan het ontstaan was tussen wetenschappelijke beoefenaren van de huisartsgeneeskunde (lees: universitaire vakgroepen en het NHG) en de mensen in het veld (lees: perifere huisartsen). Die breuk zou vooral tot stand kunnen komen doordat de twee groepen meer en meer geïsoleerd van elkaar te werk gaan.

Hoewel ik de uitgangspunten van het standaardenbeleid – zij het aarzelend – onderschrijf, wil ik toch enige kritische opmerkingen maken over dit onderwerp, met name omdat ik vrees dat er ongemerkt een nieuw paradigma in de huisartsgeneeskunde aan het binnensluipen is: datzelfde standaardenbeleid.

Allereerst is het nuttig om te weten dat op dit moment alleen in Nederland een standaardenbeleid wordt ontwikkeld. In Engeland, Denemarken en Canada, landen die qua onderzoeksniveau vergelijkbaar zijn met Nederland, speelt de hele discussie niet. McWhinney is er zelfs een tegenstander van (persoonlijke mededeling).

Al in 1976 maakte *Senior* een onderscheid tussen 'competence' en 'performance'.[2] Competence werd gedefinieerd als 'what a physician is capable of doing' en performance als 'what a physician actually does in his day-to-day practice'. Hiermee werd in de Amerikaanse toetsingskeuken duidelijk een verschil gemaakt tussen wat iemand kent en wat iemand doet. *Neufeld and Norman* waarschuwen in hun recente boek 'Assessing clinical competence' tegen diverse manieren van standaardenbepaling.[3] Zij spreken in dit verband over het gevaar van de 'armchair method'. Hiermee wordt bedoeld dat een groep experts min of meer vanuit hun kijk op het vak achterover leunend een standaard vaststelt en die vervolgens toegepast op de beroepsgroep.

Hoewel de uitdrukking 'armchair method' niet in zijn geheel opgaat voor het standaardenbeleid van het NHG, schuilt hetzelfde gevaar in de manier waarop nu standaarden worden gemaakt en de manier waarop over standaarden wordt gesproken. Met de opmerkingen van *Senior* en *Neufeld and Norman* in gedachten is het dus niet verwonderlijk dat huisartsen in het veld bij het toepassen van standaarden op hun handelen keer op keer niet hoger scoren dan 50-70 procent van de standaard.[4-5] *Norman* toonde met simulatiepatiënten overtuigend aan dat artsen die op papier een standaard hadden vastgesteld, in de praktijk slechts 68 procent op die standaard scoorden.[6]

Al eerder heeft *Wigersma* uitgesproken dat protocollen hooguit bruikbaar zouden zijn voor bepaalde vormen van onderwijs en onderzoek, met name voor kennisoverdracht.[7] Uit de lage scores die huisartsen op standaarden behalen wort vaak de conclusie getrokken dat huisartsen moeten worden bijgeschoold. In 1982 toonden *Sibley et al.* echter aan dat het toenemen van kennis via nascholing van artsen niet gepaard ging met beter handelen in de praktijk.[8] Hiermee wordt zelfs de kennisoverdracht van protocollen ter discussie gesteld.

Bij het constateren van de lage scores die huisartsen behalen op protocollen, is echter ook andere redenering mogelijk, namelijk dat de standaarden moeten worden bijgesteld naar een meer praktische benadering. Het gevalideerd zijn van standaarden betekent op dit moment vaak dat wordt gekeken of praktizerende huisartsen hun consulten kunnen registreren aan de hand van een checklist van een protocol. Dit is echter de zaak omdraaien: het zou de voorkeur verdienen om eerst eens praktisch te kijken naar wat huisartsen echt doen tijdens hun spreekuur en in aansluiting daarop de standaard vast te stellen (waarbij huisartsen ongetwijfeld nooit 100 procent hoeven te scoren). Mijns inziens zou het NHG er dan ook goed aan doen om het onderscheid 'competence-performance' eens nader te bestuderen en vervolgens te overwegen of het misschien zinvoller en effectiever zou zijn om het standaardenbeleid vanuit dat concept te presenteren. Met het