Schatten en toetsen

ROB JAMIN

Jamin R. Schatten en toetsen. Huisarts Wet 1994; 37(8): 366-71.

Het CWO-weekend 1994 vond plaats op 18 en 19 maart in Mierlo (NB). Met behulp van cijfer-opdrachten werd praktisch geoefend in het schatten van de betrouwbaarheid van steekproeven en het toetsen van onderzoeksresultaten en hypothesen. Voor een aantal veel voorkomende onderzoekssituaties werd nagegaan, welke statistische bewerkingen het meest geschikt waren. Deze werden vervolgens uitgevoerd en de uitkomsten ervan werden geanalyseerd. Het computer-programma Statistix verleende hierbij onmisbare hulp.

R. Jamin, huisarts, Meloenstraat 19, 2564 TA Den Haag.

Inleiding

De belangstelling voor het CWO-weekend nieuwe stijl dwong de organisatoren tot twee maal toe om het maximale aantal deelnemers te verhogen. Een poging het enthousiasme te temperen door een syllabus uit te brengen van het kaliber tweedefasetentamen mocht niet baten. Mede dankzij het feit dat een der 36 aanwezigen genoegen nam met een toehoordersrol, kon het interactieve programma toch probleemloos worden afgewerkt. De syllabus, een doorwrochte coproduktie van Marijke van Daelen en Wim van Geldrop, zorgde voor een pittige voorbereiding; naast helder geformuleerde theorie bood hij breinbelastende opgaven. Het weekendprogramma was uiterst praktisch. De deelnemers werden verdeeld rond zes computers en moesten een serie opdrachten uitwerken. De samenspraak in de groepjes werd begeleid en waar nodig gestuurd door ervaren collega's.

Statistix

De eerste sessie was gewijd aan het huiswerk. Mompelend bekenden de deelnemers elkaar dat ze de finish niet hadden bereikt. Van Geldrop wekte het weerbarstige materiaal onvermoeibaar tot leven: 'Dit moet iets bij je doen', en bij een andere som: 'Hier moet je tranen van in je ogen krijgen'. Vervolgens leidde Klaas Groenier ons met de nonchalance van de ware expert door de straatjes van het programma Statistix, een elegant, menu-gestuurd programma, dat eenvoud paart aan kracht. Met behulp van dit programma begonnen we aan de eerste, redelijk simpele praktijkopdracht.

In een steekproef van 12 personen wordt de bloeddruk gemeten. Voor de diastolische druk zijn de volgende waarden gevonden: 71, 73, 84, 70, 76, 71, 92, 70, 88, 85, 70 en 86. Op welke wijze wilt u de gegevens weergeven?

Het overzicht dient ten minste de centrale tendentie en de spreiding te bevatten. Lang niet altijd zijn kenmerken of waarnemingen normaal verdeeld. Er kunnen bijvoorbeeld meer hoge dan lage waarden worden aangetroffen, of een flink aantal extreem lage waarden. In zo'n geval spreken we van een scheve verdeling. Betrekken we slechts gemiddelde en standaardafwijking van zo'n verdeling in onze beschouwing, dan krijgen we een vertekend beeld. *Andersen* geeft als voorbeeld: 'Sperm count in 10 men born between 1940 and 1945 and examined in 1975-1978 was 66.0 ± 78.5 [mean ± SD]. (Andrologica 1984; 16: 175).' Hij concludeert dan dat hier toch wel sprake is van 'a very effective contraception'.[1]

Voor het beschrijven van een scheve verdeling nemen we onze toevlucht tot een indeling in klassen van waarnemingen, de percentielen. De twaalfde percentiel (P_{12}) bijvoorbeeld is de waarde waaronder zich 12% van de gevonden waarden bevindt; de overige 88% van de gevonden waarden zijn groter dan de 12e percentiel. De vijftigste percentiel is dus de middelste waarneming: 50% van de waarnemingen is kleiner en 50% is groter. Deze vijftigste percentiel vormt een goede centrale maat en wordt gemeenlijk als mediaan aangeduid. Een scheve verdeling laat zich aldus redelijk beschrijven X : N (P_{25}, mediaan, P_{75}).

De onderhavige steekproef is klein en kan alleen daarom al niet een normale verdeling hebben. Als maat voor de centrale tendentie verdient de mediaan dus de voorkeur; als maten voor de spreiding geven percentielklassen een goed beeld, bijvoorbeeld de kwartielen met de mediaan (0, 25, 50, 75 en 100%). Samen met de onder- en bovengrenzen van het 95%-betrouwbaarheidsinterval ontstaat een tamelijk volledig en duidelijk beeld van de steekproef. Het programma levert deze gegevens snel en geruisloos nadat via het scherm enkele kruisjes zijn gezet (*tabel 1*).

Deze tamelijk omslachtige omschrijving is in één oogopslag te vatten met een grafische weergave. In het voorbeeld zijn de klassen te weinig gevuld voor een histogram, maar levert de box-and-whisker-plot, ook wel snorrendoos geheten, een fraai beeld (*figuur 1*). De blokjes worden aan onder- en bovenzijde begrensd door respectievelijk de 25e en de 75e percentiel; de dwarslijn midden in de blokjes is

de mediaan en de verticaal uitlopende lijntjes (whiskers) eindigen ter plaatse van de 5e en 95e percentiel.

Gepaarde waarnemingen

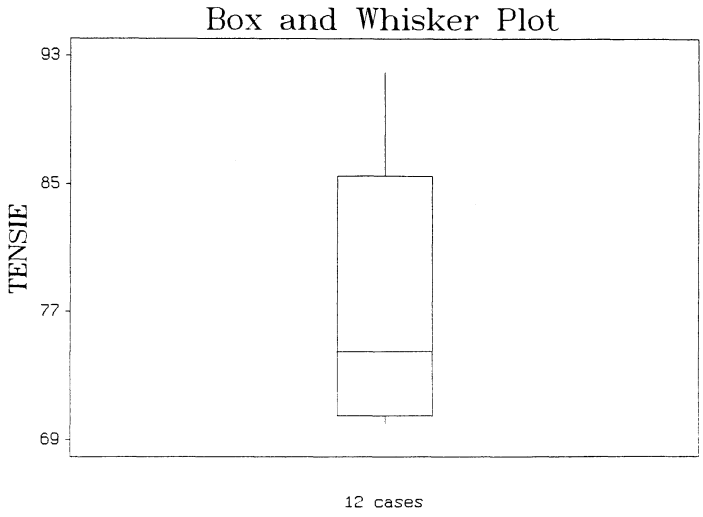
De tweede opdracht voegde een dimensie toe: er worden twee series waarnemingen met elkaar vergeleken. Tien personen met essentiële hypertensie krijgen eerst antihypertensivum A en na een wash-outperiode middel B. De bloeddrukdaling (mm Hg) in beide groepen is weergegeven in *tabel 2*. Bekijk de twee groepen met behulp van beschrijvende statistiek, kies een parametrische toets voor het verschil in gemiddelden en vergelijk het resultaat met dat van enkele non-parametrische toetsen.

Luttele tellen na het inbrengen van de waarden in Statistix verschenen de eerste overzichten op het scherm. Omdat de twee series metingen worden verricht aan objecten die in feite uit één steekproef afkomstig zijn, spreken we van gepaarde waarnemingen. Wanneer de kenmerken in een populatie normaal zijn verdeeld, zijn de parameters (gemiddelde, standaarddeviatie, e.d.) toegankelijk voor wiskundige bewerkingen. De statistische toetsen die dat doen, heten daarom parametrisch. Bij scheve of andere verdelingen moeten we ons behelpen met afgeleide kenmerken, zoals frequenties, rangordes en percentielen. De toetsen die hiermee uit de voeten kunnen, heten non-parametrisch.

Uit de parametrische groep komt de gepaarde T-toets het meest in aanmerking (*tabel 3*). Het gemiddelde verschil blijkt 4,3 te zijn, de standaardfout 1,39 en de

Tabel 1 *Beschrijvende statistiek van 12 bloeddrukwaarden*

N	12
Ondergrens 95%-betrouwbaarheidsinterval	72,705
Gemiddelde	78,000
Bovengrens 95%-betrouwbaarheidsinterval	83,295
Standaardafwijking	8,3339
Minimum	70,000
Eerste kwartiel	70,250
Mediaan	74,500
Derde kwartiel	85,750
Maximum	92,000



Figuur 1 *De belangrijkste waarden van tabel 1 in een 'box and whisker plot'* →

Tabel 2 *Bloeddrukdaling in mm Hg*

Patiënt	A	B
1	10	5
2	5	0
3	6	2
4	8	8
5	12	0
6	12	2
7	5	5
8	10	3
9	5	5
10	0	0

Tabel 3 *Keuzediagram voor toetsen*

Aard van de steekproeven	Parametrisch	Non-parametrisch
Gegevens op numeriek niveau		
Twee, onafhankelijke	• ongepaarde T-toets • ANOVA enkelvoudig*	• twee-steekproeventoets van Willcoxon (alias Mann-Whitney) • toets van Kruskal-Wallis *
Twee, afhankelijke	• gepaarde T-toets • ANOVA repeated measure*	• teken-toets • rang-tekentoets • toets van Friedman*
Gegevens op nominaal niveau		
Twee, onafhankelijke	• z-toets	• χ^2-toets*
Twee, afhankelijke		• Mc Nemar-toets

* Ook voor meer dan twee steekproeven.

p-waarde 0,0129. Dit laatste cijfer geeft de kans aan dat de gevonden verschillen op toeval berusten; in dit geval iets meer dan 1 procent, ruim onder de 5 procent die meestal als afkappunt wordt gekozen.

Voor de non-parametrische benadering zijn aan te bevelen de tekentoets, de rangtekentoets van Wilcoxon en de tweezijdige toets van Friedman. De tekentoets telt het aantal malen dat het verschil tussen de waarden der paren positief dan wel negatief is; de waarde 0, geen verschil dus, wordt genegeerd. De rang-tekentoets van Wilcoxon berekent de verschillen tussen de waarden per paar, zet deze verschillen op volgorde van absolute waarde en telt de rangnummers die aldus ontstaan bij elkaar op voor de positieve en de negatieve kant. De toets van Friedman geeft per paar een rangnummer aan elke variabele en berekent vervolgens het gemiddelde van de rangnummers per variabele. De p-waarden zijn weergegeven in *tabel 4*.

In de statistiek is er na heftige discussie over de verdiensten van parametrische en non-parametrische toetsen een soort consensus ontstaan. Omdat mag worden aangenomen dat een steekproef van 30 of meer objecten een normale verdeling heeft, kan er vanaf dat aantal met parametrische toetsen worden gewerkt. Zelfs in onze (te) kleine steekproef van 10 objecten liggen alle gevonden p-waarden in dezelfde orde van grootte. Alleen bij twijfel is aan te raden om een non-parametrische toets te gebruiken. De lezer uitmaken of de vertraging waarmee deze discussie alsnog in medische kringen wordt gevoerd, berust op prudentie dan wel conservatisme. Eenzijdig toetsen – verleidelijk vanwege de kleinere p-waarden die het oplevert en de kortere bewerkingstijd – is in beginsel alleen aangewezen wanneer een verband in één der richtingen logisch is uitgesloten.

Analyse van drie groepen

De volgende opdracht betrof een dieetprobleem: vergelijk de mate waarin patiënten zijn afgevallen bij drie verschillende regimes met energiebeperking (*tabel 5*). De resultaten in de drie groepen van elk tien patiënten dienden te worden vergeleken met behulp van beschrijvende statistiek. Vervolgens moest worden nagegaan of er een significant verschil in gemiddelden bestond ($\alpha = 0{,}05$) tussen de drie regimes; eerst met behulp van een parametrische toets, daarna met een non-parametrische toets.

Tabel 4 *p-waarden van diverse toetsen*

toets	p-waarde
gepaarde T	0,0129
teken	0,0156
Wilcoxon	0,0156
Friedman	0,0143

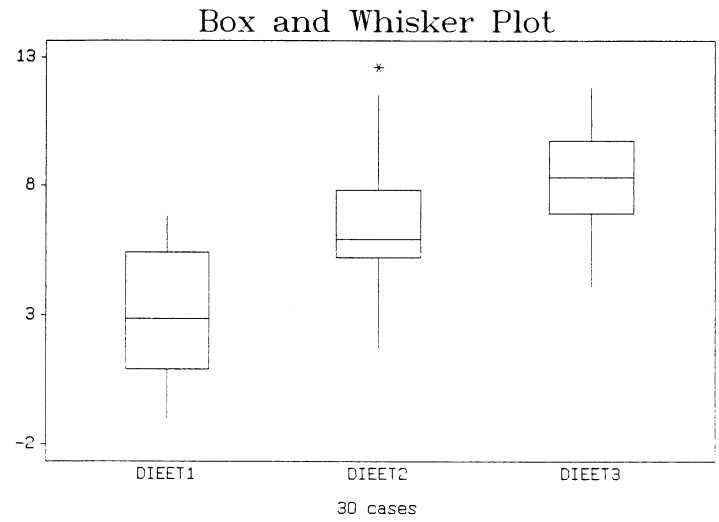
Tabel 5 *Gewichtsveranderingen in kg*

1	2	3
–1	1,7	4,1
0,3	3	6,1
0,9	5,2	6,9
2,1	5,6	7,4
2,6	5,7	8
3,1	6,1	8,6
3,3	6,6	8,8
5,4	7,8	9,7
5,7	11,5	10,6
6,8	12,6	11,8

Gemiddelde en mediaan liggen in de groepen 1 en 3 dicht bij elkaar, hetgeen wijst op een min of meer normale verdeling; in groep 2 liggen eerste en derde kwartiel dichter bijeen, maar is de spreiding tussen de uiterste waarden groter. De 95%-betrouwbaarheidsintervallen van de groepen 1 en 3 overlappen in het geheel niet, die van 1 en 2 nauwelijks, en tussen 2 en 3 is er een aanmerkelijke overlap. Omdat groep 1 ook de waarde 0 en negatieve waarden omvat, is de kans groot dat dieet 1 in feite geen gewichtsdaling bewerkstelligt. Bij de groepen 2 en 3 is de conclusie gerechtvaardigd dat er wel een gewichtsdaling optreedt.

De box-and-whiskerplot maakt deze woordenvloed snel aanschouwelijk (*figuur 2*). De plot markeert uit eigen beweging met een sterretje een extreme waarde in dieetgroep 2. Zo'n ver-weggelegen waarde wordt aangeduid met de term uitbijter. In de wereld van het toeval zijn uitbijters te allen tijde mogelijk, maar er moet steeds ernstig rekening worden gehouden met de mogelijkheid dat er een meet- of verwerkingsfout in het spel is.

Als parametrische toets komt de ANOVA (ANalysis Of VAriance) in aanmerking; deze toetst de veronderstelling dat



Figuur 2 *Box and whisker plot op basis van tabel 5*

alle groepen een steekproef uit dezelfde populatie zijn; met andere woorden dat de gemiddelde waarden van de groepen niet van elkaar afwijken. Als maat hiervoor wordt het quotiënt genomen van de variantie tussen de groepen en de interne variantie binnen de groepen, de zogenoemde F-waarde. Hoe meer F in de buurt van 1 blijft, des te kleiner is het vermoedelijke verschil tussen de groepen.

De F-waarde blijkt 9,70 te zijn bij een significantie-niveau (p-waarde) van 0,0007; er is dus met grote waarschijnlijk een aanzienlijk verschil tussen de groepen. Wanneer het aantal groepen toeneemt, wordt de kans groter dat er significantie wordt gevonden, terwijl de nulhypothese (geen verschil tussen de groepen) toch geldig is. Bij vijf groepen worden tien vergelijkingen uitgewerkt, een voor de externe en een voor de interne variantie; de kans dat op grond van toeval één p-waarde onder 0,05 wordt gevonden, is dan al 0,29.

De Bonferroni-toets corrigeert het significantieniveau paarsgewijs door het aantal berekende vergelijkingen in de uitkomst te verdisconteren. Bij grotere aantallen groepen valt zo'n aanvulling op de ANOVA sterk aan te bevelen. In het onderhavige geval blijken de gemiddelden van de groepen 2 en 3 niet significant van elkaar af te wijken.

Bij non-parametrische toetsen is het in feite niet mogelijk om paarsgewijs te vergelijken. De groepen, in casu de personen die één van drie diëten volgen, kunnen als drie onafhankelijke steekproeven uit eenzelfde populatie worden beschouwd, mits de deelnemers hun dieet aselect krijgen toegewezen. De verdeling van de uitkomstvariabele gewichtsdaling wordt dan bepaald door de onafhankelijke variabele dieet.

De Kruskal-Wallis toets komt hiervoor het meest in aanmerking. In deze toets worden de objecten van alle groepen samen beschouwd en krijgen ze een rangnummer naar grootte van de gevonden waarden. Deze rangnummers worden vervolgens voor elke groep gesommeerd en bewerkt tot de Kruskal-Wallis H-waarde. Bij de nulhypothese dat de groepen een gelijke verdeling hebben, heeft de H-waarde een chi-kwadraatverdeling. We vinden een hoge waarde voor H (13,17) bij een sterk significantieniveau (0,0014).

Tabel 6 *Aantallen personen naar leeftijdklasse*

Leeftijd	Onderzoek		Krimpen	
(jaren)	man	vrouw	man	vrouw
35-39	239	–	295	240
40-49	458	–	547	530
50-59	300	307	385	395
60-69	231	249	293	286
70-79	144	165	153	193
≥80	30	59	35	71

Tabel 7 *Procentuele verdeling leeftijdklassen*

Leeftijd	Nederland		Krimpen	
(jaren)	man n=7.480.000	vrouw n=7.648.700	man n=3399	vrouw n=3348
0- 9	12,7	11,9	12,9	13,0
10-19	13,0	12,1	12,4	11,7
20-29	17,6	16,3	15,7	14,8
30-39	16,6	15,5	17,0	16,5
40-49	14,9	13,9	16,1	15,8
50-59	10,4	10,0	11,3	11,8
60-69	8,2	9,0	8,6	8,5
70-79	4,9	6,9	4,8	5,8
≥80	1,8	4,1	1,0	2,1

Representativiteit van steekproeven

De lezer moet verder, doch de deelnemers van het weekend konden zich gaan verpozen. Na de avondmaaltijd kregen ze een concert voorgeschoteld door het ensemble Kanzone onder leiding van Hugo de Lil, die werken van Purcell, Bach, Händel en Elgar van een geestige introductie voorzag.

De volgende dag ging het verder. Naast fictieve cijfers werd ook materiaal van feitelijk onderzoek voorgeschoteld. Knuistingh Neven doet in de eigen praktijk een onderzoek naar de prevalentie van het slaap-apneusyndroom. Mogelijke lijders daaraan zijn met behulp van vragenlijsten geselecteerd. Op grond van een van een literatuurstudie heeft hij alle mannen van 35 jaar en ouder benaderd en alle vrouwen van 50 jaar en ouder, totaal 2182 personen. De respons hierop was 88,2 procent.

De weekendgangers kregen twee tabellen voorgelegd met leeftijdopbouw; ze dienden vast te stellen of de onderzochte populatie representatief is voor Krimpen aan de Lek (*tabel 6*) en vervolgens of de bevolkingsopbouw van het dorp representatief is voor de Nederlandse bevolking (*tabel 7*). Het volgende beperkt zich tot de vergelijking tussen Krimpen en Nederland, aangezien tussen de onderzoeksgroep en Krimpen mutatis mutandis hetzelfde geldt.

Wanneer we paarsgewijs voor Krimpen en Nederland het aantal personen in elke leeftijdsklasse afzetten tegen het overblijvende deel van de bevolking ontstaat een

serie 2×2-tabellen. Het aantal inwoners van Krimpen per leeftijdsklasse (observed: O) wordt berekend door het Krimpense percentage te vermenigvuldigen met het totale aantal inwoners van Krimpen; in feite het omdraaien van de berekening op grond waarvan het percentage is vastgesteld. Vervolgens wordt het Nederlandse percentage van die klasse vermenigvuldigd met het totaal aantal inwoners van Krimpen; aldus ontstaat een aantal dat de betreffende klasse zou hebben wanneer Krimpen exact dezelfde samenstelling zou hebben als de Nederlandse bevolking (expected: E). Hierop kan een chi-kwadraattoets worden uitgevoerd: $\chi^2=\Sigma(O-E)^2/E$.

Statistix kan helaas de orde van grootte van de randtotalen die aldus ontstaan (7.648.700 – 3348) niet aan. Als alternatief wordt daarom de Kolmogorov-Smirnovtoets gebruikt. De bijzonder hoge p-waarden die hierbij worden gevonden, bevestigen het beeld dat grafisch al was opgeroepen: de populaties stemmen qua leeftijdopbouw overeen. Om de onderzoekspopulatie te vergelijken met de Krimpense bevolking kunnen de absolute aantallen uit *tabel 6* worden gebruikt. Ook hier bewijst de toets van Kolmogorov-Smirnov goede diensten.

Op grond van hun antwoorden werden 193 personen aangemerkt als risicopatiënt. Bij 167 van hen werd thuis gedurende een nacht een registratie van de ademhaling uitgevoerd; bij 9 personen mislukte deze. De registratie geschiedde door middel van een zogeheten thermistor op de neus, een apparaat dat temperatuurverschillen in de lucht kan meten; het ging natuurlijk om het verschil tussen in- en uitgeademde lucht. Op grond van zo'n registratie kan een apneu-index (AI) worden vastgesteld. Bij AI ≥5 is er verdenking op het slaap-apneusyndroom. Er is sprake van het syndroom wanneer er meer dan vijf maal per uur een apneu van langer dan 10 seconden optreedt.

De 25 personen die met de thermistor een AI ≥5 scoorden, werden in het Leidse slaaplaboratorium aan een polysomnografie onderworpen. De apneu-index die daarbij wordt gevonden, geldt als gouden standaard voor het slaap-apneusyndroom. De verschillen tussen beide registraties (*tabel 8*) lijken op het eerste gezicht nogal groot te zijn. Ga na of de gevonden verschillen ook significant zijn.

Tabel 8 *Apneu-index thuis en in het laboratorium*

Patiënt	Thuis	Laboratorium
1	6	4
2	20	5
3	5	11
4	5	21
5	15	32
6	23	19
7	35	45
8	5	0
9	7	2
10	10	4
11	5	0
12	6	0
13	10	6
14	5	8
15	6	12
16	10	0
17	7	13
18	7	0
19	5	4
20	18	32
21	30	35
22	6	0
23	7	16
24	5	8
25	8	0

Het gaat in wezen om twee gepaarde steekproeven. Bij normale verdeling zou de t-toets in aanmerking komen, maar vanwege het kleine aantal waarnemingen moet hier non-parametrisch worden getoetst. De toetsen die voor de hand liggen, zijn de tekentoets, de rang-somtoets van Mann-Whitney en de rang-tekentoets van Wilcoxon. Alle leveren hoge p-waarden: 0,3450 en 0,3224 en zelfs 0,9250. De verschillen tussen de registratie thuis en met polysomnografie zijn dus vrijwel zeker niet aan het toeval te wijten, met andere woorden: de registraties komen aardig overeen. Tot opluchting van de onderzoeker.

Vier-veldentabel

Met de laatste opdracht begaven de deelnemers zich op het moeizame pad van casustiek en intuïtie naar kennis. Het was een huisarts opgevallen dat hij bij kinderen met otitis media het trommelvlies steeds goed à vue kreeg door het ontbreken van oorsmeer. Hij meende dat mogelijk ten gevolge van de ontsteking weinig of geen cerumen wordt aangemaakt en formuleerde de volgende hypothese: bij acute otitis media (OMA) is geen oorsmeer aanwezig, en de aanwezigheid van cerumen pleit tegen de diagnose OMA. Vragen aan de deelnemers:

- hoe zou u dit onderzoek opzetten;
- welke patiënten selecteert u;
- hoe groot is de steekproef;
- welke statistische toets stelt u voor?

Gekozen wordt voor een verkennend onderzoek bij kinderen tot 5 jaar. Daartoe worden alle kinderen die zich gedurende een afgebakende periode presenteren ingesloten, ongeacht de klacht. Uitgesloten worden monauriculaire kinderen. Voor het stellen van de diagnose al of niet OMA worden de criteria van de NHG-standaard gehanteerd. De aan- of afwezigheid van oorsmeer wordt beoordeeld door een onafhankelijke waarnemer.

Aldus ontstaan vier groepen (*tabel 9*). Voor het berekenen van de steekproefomvang is het nodig om de grootte vast te stellen van α, β en de drempel voor de relevantie van het te vinden verschil. De gebruikelijke waarden voor α en β zijn 0,05 respectievelijk 0,20. Het kleinste verschil dat relevant wordt geacht, geven we uit logische overwegingen een hoge waarde: de hypothese lijkt immers niet erg plausibel en aanvaarding kan vergaande consequenties hebben.

Tabel 9 *Vierveldentabel*

	OMA –	OMA +
cerumen –	n	n
cerumen +	x_1	x_2
	$p_1 = x_1/n$	$p_2 = x_2/n$

Wanneer we aannemen dat in een pilot p_1=0,1 blijkt, zal p_2 in de orde van 0,02 dienen te zijn. De omvang van de steekproef kan nu worden berekend aan de hand van een formule waarin p_1, p_2, α en β worden gewogen. We zouden volgens deze berekening 135 patiënten moeten insluiten. Door β te vergroten tot 0,3 tillen we impliciet minder zwaar aan het risico dat we de nulhypothese ten onrechte verwerpen, en brengen we het aantal benodigde patiënten terug tot 106.

Tot slot werd gevraagd om een fictief onderzoeksresultaat te toetsen (*tabel 10*). We hebben te maken met twee ongepaarde steekproeven, zodat het in de rede ligt om de χ^2-toets te kiezen. We vinden een χ^2-waarde van 6,75 (redelijk sterk verband) bij een significantie-niveau van 0,4554 (net onder de grens van 5 procent die we meestal aanhouden).

Kleine verschillen tussen de geobserveerde en verwachte waarden kunnen hoge waarden voor χ^2 opleveren. Wanneer de steekproef klein is en de verwachte waarden laag zijn, is Fisher's toets het meest geschikt. Deze ontbreekt helaas in het menu van Statistix.

Tabel 10 *Fictief onderzoeksresultaat, waarin 0=nee en 1=ja*

Patiënt	Smeer	OMA
1	0	1
2	1	0
3	0	1
4	0	0
5	0	1
6	0	1
7	1	0
8	0	0
9	1	1
10	0	1

Conclusie

Inhoudelijk én organisatorisch had de voorbereidingscommissie, gevormd door Marijke van Daelen, Wim van Geldrop, Arie Knuistingh Neven en Leo Veehof, het programma zeer degelijk voorbereid. De deelnemers – toch al vrijwillig gekomen – raakten hierdoor te meer bevlogen. Het praktisch werken met Statistix op computers maakte er een levendig en buitenal leerzaam weekend van.

Literatuur

1 Andersen B. Methodological errors in medical research. Oxford: Blackwell Scientific Publications, 1991. ■