, 22-25,
28-36

willen minimaliseren (zoals in de "decision theoretic approach to statistics"). Dan is immers vrij nauwkeurig gespecificeerd wat 'optimaal' is althans wat 'toelaatbaar' is. Maar als we een eenmalig experiment hebben, en er is geen vooraf gegeven verliesfunctie dan hebben we ook veel minder duidelijke criteria om wel of niet conditioneren tegen elkaar af te wegen. Voor zover wij begrijpen blijft de beslissing dan tamelijk subjectief en afhankelijk van de interpretatie van de omstandigheden. Na een heldere uiteenzetting van dit onderscheid tussen beslissingsproblemen en "summarizing the evidence available in data" geeft Cox(1958, p. 360-361) het volgende voorbeeld:

"Suppose that we are interested in the mean d of a normal population and that, by an objective randomization device, we draw either (i) with probability ½, one observation, x, from a normal population of mean 0 and variance σ_1^2 or (ii) with probability ½, one observation x, from a normal population of mean θ and variance σ_2^2, where σ_1^2, σ_2^2 are known, $\sigma_1^2 > \sigma_2^2$ and where we know in any particular instance which population has been sampled.
The sample space formed by indefinite repitition of the experiment is clearly defined and consists of two real lines Σ_1, Σ_2, each having probability ½, and conditionally on Σ_1 there is a normal distribution of mean 6 and variance σ_1^2.
Now suppose that we ask, accepting for the moment the conventional formulation, for a test of the null hypothesis $6 = 0$, with size say 0.05, and with maximum power against the alternative $6'$, where $0' \approx \sigma_1 \gg \sigma_2$.
Consider two tests. First, there is what we may call the conditional test, in which calculations of power and size are made conditionally within the particular distribution that is known to have been sampled. This leads to the critical regions $x > 1.64\sigma_1$ *or* $x > 1.64\sigma_2$, depending on which distribution has been sampled.
This is not, however, the most powerfull procedure over the whole sample space. An application of the Neyman-Pearson lemma shows that the best test depends slightly on 0', σ_1, σ_2, but is very nearly of the following form. Take as the critical region

$x > 1.28\sigma_1$, if the first population has been sampled,

$x > 5\sigma_2$, if the second population has been sampled.

Qualitatively, we can achievé almost complete discrimination between $\theta = 0$ and $\theta = \theta'$ when our observation is from ω_2, and therefore we can allow the error rate to rise to very nearly 10% under Σ_1. It is intuitively clear, and can easily be verified by calculation, that this increases the power, in the region of interest, as compared with the conditional test.
Now if the object of the analysis is to make statements by a rule with certain specified long-run properties, the unconditional test just given is in order, although it may be doubted whether the specification of desired properties is in this case very sensible. If, however, our object is to say 'what we can learn from the data that we have', the unconditional test is surely no good."

De hamvraag kan ook zo geformuleerd worden: Willen, kunnen en mogen we de gevolgen van een vergroting Tan de kans op ten onrechte verwerpen bij trekken uit Σ_2 afwegen tegen de gevolgen van een verkleining van de kans op ten onrechte verwerpen bij trekking uit Σ_1. Nog anders geformuleerd:

1980

VVS Bulletin, 13(3), 22-2

28-

UIT DE STATISTISCHE CONSULTATIE

Onder redactie van A. Heyting
Philips Duphar B.V.
Postbus 2
1380 AA Weesp

A. Verbeek
Sociologisch Instituut
Heidelberglaan 2
3584 CS Utrecht

Auteurs worden uitgenodigd bijdragen te zenden aan een van de redactieleden.
De richtlijnen voor auteurs zijn opgenomen in het juninummer op blz. 37.

De χ^2-toetsen in kruistabellen: conditioneren of niet? - vervolg.

Door Albert Verbeek en Pieter Kroonenberg.

De reactie van Tom Snijders in dit Bulletin (1980) en een mondelinge reactie van Wachtel Sekhuis op ons artikel in het novembernummer van dit bulletin (1979) betreffen de volgende punten, corresponderend met de paragrafen van dit artikel.

A. Principiële argumenten voor het conditioneren. De titanenstrijd die hierover gestreden is (en wordt), is een fraai voorbeeld van de ruimte die wiskundige modellen laten voor interpretatie en meningsverschillen. Voor ons persoonlijk schijnen de argumenten niet dwingend of doorslaggevend.

B. Volgens Tom Snijders is het onderscheidingsvermogen van de exacte, conditionele ongerandomiseerde toets onnodig klein, en is ook de bijbehorende overschrijdingskans onnodig groot. Wij zijn het daar voor sommige gevallen wel mee eens, maar lang niet voor alle gevallen. In de?e paragraaf zetten wij een aantal criteria voor optimaliteit en een aantal toetsen op onafhankelijkheid in een kruistabel nog eens op een rij en concluderen o.a. dat van veel approximaties nog nauwelijks bekend is hoe goed of grof ze zijn. Tabel 2 geeft een overzicht van on/e bevindingen.

C. Onze paragraaf 5 van november bevat enkele storende slordigheden. Nadat we eerst deze balken uit eigen oog gepeuterd hebben, gaan we nog wat splinters en misverstanden langs die ons bij de literatuurstudie bij anderen opvielen.

A. Principiële argumenten rond het conditioneren op marginalen

Fishers standpunt komt er globaal op neer dat men hoort te conditioneren omdat de marginalen geen relevante informatie bevatten over interacties (= kruisproducten) in een kruistabel, maar uitsluitend over de nauwkeurigheid van de schatters van de interacties.
Bij de formalisering van het standpunt rijzen twee problemen die nog tamelijk onopgelost zijn. Het eerste is het formaliseren van het begrip "geen relevante informatie". Er zijn namelijk verschillende definities voor ondergeschikte (ancillary) grootheden in de handel, elk met andere gewenste en ongewenste eigenschappen. Het tweede probleem betreft de vraag waarom zouden we moeten conditioneren op ondergeschikte grootheden. Er is meer hoop op een duidelijk antwoord, als we te maken hebben met een vaak herhaald experiment en een gegeven verliesfunctie waarvan we de verwachte waarde

22-25, 28-36

willen minimaliseren (zoals in de "decision theoretic approach to statistics"). Dan is immers vrij nauwkeurig gespecificeerd wat 'optimaal' is althans wat 'toelaatbaar' is. Maar als we een eenmalig experiment hebben, en er is geen vooraf gegeven verliesfunctie dan hebben we ook veel minder duidelijke criteria om wel of niet conditioneren tegen elkaar af te wegen. Voor zover wij begrijpen blijft de beslissing dan tamelijk subjectief en afhankelijk van de interpretatie van de omstandigheden. Na een heldere uiteenzetting van dit onderscheid tussen beslissingsproblemen en "summarizing the evidence available in data" geeft Cox(1958, p. 360-361) het volgende voorbeeld:

"Suppose that we are interested in the mean θ of a normal population and that, by an objective randomization device, we draw either (i) with probability ½, one observation, x, from a normal population of mean G and variance *of* or (ii) with probability ½, one observation x, from a normal population of mean θ and variance σ_2^2, where σ_1^2, σ_2^2 are known, *of* $> \sigma_2^2$ and where we know in any particular instance which population has been sampled.
The sample space formed by indefinite repitition of the experiment is clearly defined and consists of two real lines Σ_1, Σ_2, each having probability ½, and conditionally on Σ_1 there is a normal distribution of mean 6 and variance σ_1^2.
Now suppose that we ask, accepting for the moment the conventional formulation, for a test of the null hypothesis $6 = 0$, with size say 0.05, and with maximum power against the alternative θ', where $6' \approx \sigma_1 \gg \sigma_2$.
Consider two tests. First, there is what we may call the conditional test, in which calculations of power and size are made conditionally within the particular distribution that is known to have been sampled. This leads to the critical regions $x > 1.64\,\sigma_1$ or $x > 1.64\,\sigma_2$, depending on which distribution has been sampled.
This is not, however, the most powerfull procedure over the whole sample space. An application of the Neyman-Pearson lemma shows that the best test depends slightly on $\theta', \sigma_1, \sigma_2$, but *is* very nearly of the following form. Take as the critical region

$x > 1.28\,\sigma_1$, if the first population has been sampled,

$x > 5\,\sigma_2$, if the second population has been sampled.

Qualitatively, we can achieve almost complete discrimination between $\theta = 0$ and $G = \theta'$ when our observation is from ω_2, and therefore we can allow the error rate to rise to very nearly 10% under Σ_1. It is intuitively clear, and can easily be verified by calculation, that this increases the power, in the region of interest, as compared with the conditional test.
Now if the object of the analysis is to make statements by a rule with certain specified long-run properties, the unconditional test just given is in order, although it may be doubted whether the specification of desired properties is in this case very sensible. If, however, our object is to say 'what we can learn from the data that we have', the unconditional test is surely no good."

De hamvraag kan ook zo geformuleerd worden: Willen, kunnen en mogen we de gevolgen van een vergroting Tan de kans op ten onrechte verwerpen bij trekken uit Σ_2 afwegen tegen de gevolgen van een verkleining van de kans op ten onrechte verwerpen *bij* trekking uit Σ_1. Nog anders geformuleerd:

Als we een waarneming uit Σ_1 hebben, nemen we dan als uitkomstenruimte Z alleen Σ_1 (conditionele inferentie) of $\Sigma_1 \cup \Sigma_2$ (niet-conditionele aanpak)? Als het doel is 'to say what we can learn from the data' dan kiest de Titaan Cox onomwonden (en wij aarzelend) voor conditioneren op ondergeschikte grootheden, zoals *hi)* even verderop zegt:

"... Z should be taken to consist, so far as is possible, of observations similar to the observed set S, in all respects which do not give a basis for discrimination between the possible values of the unknown parameter *6* of interest. Thus, in the example, information as to whether it was Σ_1 or Σ_2 that we sampled tells us nothing about 0, and hence we make our inference conditionally on Σ_1 or Σ_2." Dit betekent nog niet dat Cox ook bij toetsen op onafhankelijkheid in een kruistabel aannemelijk heeft gemaakt (of vindt) dat de marginalen ons niets zeggen over de (on)afhankelijkheid.

Laten we dit voorbeeld besluiten met twee tamelijk extreme voorbeelden. Het eerste betreft éénmalig onderzoek waarvan de conclusie voor een breed publiek bestemd is. Eén van Cox' argumenten is dan dat het afwegen van de genoemde gevolgen voor verschillende toepassingen op verschillende wijze moet geschieden, dat dit niets te maken heeft met 'what we learn from the data' en dat we daarom eerst de conclusies moeten formuleren die niet van de wijze van afwegen afhangen. Die conclusies moeten dan conditioneel zijn. Een sterk verschillende situatie is bijvoorbeeld kwaliteitscontrole waarbij elke week een monster uit de grondstoffen getrokken wordt, uit Z, u Σ_2 (en elke week met een mogelijk verschillende 0) en waarvan de conclusies (en risico's) alleen voor het eigen bedrijf zijn. Bij een verantwoorde beslissing kan en moet men dan niet alleen de data beschouwen maar zeker ook de verwachte verliezen bij verkeerde beslissingen en wellicht ervaring met de verdeling van 0 van week tot week. Er is dan weinig reden om te conditioneren als we niet-conditioneel een kleiner verwacht verlies denken te kunnen realiseren.

Een laatste punt is nog het volgende. Behalve de door ons in par. 4 (1979) beschreven onderzoeksopzetten zijn er uiteraard ook andere. Ook zijn er bijvoorbeeld opzetten waarbij de beide marginale verdelingen door de onderzoeksopzet vastgelegd zijn, althans onder H_0. Twee karakteristieke voorbeelden hiervan zijn:

i) Een proefpersoon moet van tien monsters boter proberen vast te stellen welke verse boter zijn en welke koelhuisboter. Hij weet dat er van beide soorten vijf monsters zijn. De nulhypothese is dat hij geen verschil kan proeven.

ii) Van twintig patiënten worden er aselect tien toegewezen aan behandeling A, terwijl de overigen behandeling B krijgen. De nulhypothese is dat beide behandelingen even effectief zijn voor het wel of niet herstellen van de patiënten. Als na afloop van het experiment veertien patiënten hersteld zijn, dan zouden dit bij iedere andere uitkomst van de aselecte toewijzing ook veertien opgeleverd hebben, omdat onder de nulhypothese het al dan niet herstellen juist onafhankelijk is van de behandeling. Met andere woorden onder H_0 is ook deze marginaal vast. Men kan de marginaal bijvoorbeeld wel als stochastisch opvatten als men kan of wil aannemen dat ook de patiënten aselect gekozen zijn uit een welomschreven populatie.

Men kan zich afvragen of Snijders' voorbeeld in dit opzicht wel gelukkig gekozen is. Met name bij patiënten is er doorgaans geen sprake van een expliciet randomisatie mechanisme. Zij zijn immers meestal een zelfselectie uit

een uiterst onduidelijke populatie. Daarnaast zijn ze vaak niet aselect maar tamelijk willekeurig aan A of B toegewezen. Men zou dus kunnen verdedigen dat hier helemaal geen kansmodel aan ten grondslag ligt en dat inferentie dus onmogelijk (of: misleidend) is.

B. Criteria bij de keuze van een toetsgrootheid

Als men bij een bepaalde toepassing Fishers standpunt niet deelt, dan moet er dus een keus gemaakt worden uit een aantal beschikbare toetsen. Overwegingen ten aanzien van de volgende aspecten geven voor ons daarbij de doorslag:

a. Het onderscheidingsvermogen.
b. Het gemak waarmee de toetsingsgrootheid en overschrijdingskans (of een benadering ervan) bepaald kunnen worden.
c. De nauwkeurigheid van de gebruikte benaderingen.
d. Het gebruik van randomisatie bij de toets.
e. Beperkingen ten gevolge van de steek proef opzet.
f. Het doel van het onderzoek en onze interpretatie van de populatie waaraan de waarnemingen ontleend zijn.

De eerste vijf aspecten zullen verderop als toetssteen gebruikt worden voor verschillende mogelijke manieren van toetsen. Het laatste aspect is zozeer verweven met de inhoudelijke achtergrond van de data dat het niet losstaand als toetssteen kan dienen. Twee manieren waarop het de keus kan beïnvloeden zijn deze:

- Het in A bij het voorbeeld van Cox genoemde onderscheid tussen éénmalig onderzoek en onderzoek dat opgevat wordt als één uit een lange reeks over eenkomstige onderzoeken, waarbij we de kans op verkeerde beslissingen uit het ene onderzoek mogen afwegen tegen de kans op verschillende beslissingen uit elk ander.
- Het op dezelfde plaats genoemde onderscheid tussen onderzoek dat gericht is op één bepaalde beslissing en onderzoek dat feiten moet opleveren die door een breed publiek op zeer verschillende wijze toegepast moet kunnen worden. We kunnen dit laatste geval karakteriseren als het gebruik van statistiek als communicatiemiddel. Hierbij is het (vaak) verstandig een methode te kiezen die met-aanvechtbaar is. Zo biedt bijvoorbeeld de populatie waaraan de waarnemingen ontleend zijn bij sociologisch onderzoek vaak ruimte voor controverse ("Is een steekproef van Utrechtse studenten representatief voor Nederlandse studenten?"). Als dit zo is, dan is het veiliger om in eerste instantie de populatie 'klein' te houden. Voor onze kruistabellen is dit dan vaak een argument om te toetsen bij de waargenomen marginalen. Als we het over popuplatie en steekproefopzet eens zijn geworden, dan geldt eenzelfde argument voor de verschillende mogelijke toelaatbare toetsen. Als zij toch allemaal hetzelfde resultaat opleveren dan heeft het vermelden van de meest conservatieve toets meer overtuigingskracht dan het vermelden van een 'optimale' maar controversiële toets. En omgekeerd, als verschillende min of meer acceptabele toetsen elkaar tegenspreken, dan lijkt de enige onaanvechtbare conclusie: er zijn nog onvoldoende gegevens voor éénduidige conclusies.

Nadat we hierboven enkele aspecten hebben genoemd om verschillende toetsen mee te beoordelen, zullen we nu eerst aangeven wat de voornaamste verschillen zijn tussen de toetsen op onafhankelijkheid in een kruistabel.
Uit par. 6 van ons vorig artikel (1979) zou men de indruk kunnen krijgen

vervolg op pag. 28

dat er nooit verschil is tussen conditioneel en niet-conditioneel toetsen, wat betreft significantieniveau en p-waarde (of overschrijdingskans). Immers als het kritiekgebied onder H_0 onder elke conditionering juist zeg 5% groot is, dan is het totale kritieke gebied ook van deze grootte. Wat wij daarbij niet vermeldden is het volgende. Omdat de uitkomsten hier discreet zijn kan men natuurlijk (bijna) nooit een a van precies 5% realiseren (tenzij men randomiseerd, d.w.z. dat men bij een of meer uitkomsten loot met een vooraf gegeven kans of men wel of niet zal verwerpen. Voor éénmalige onderzoeken en bij het vermelden van overschrijdingskansen is randomiseren echter weinig relevant.). Hierdoor krijgt men bij elk stel marginalen m (= bij elke conditionering) steeds een iets andere kans, zeg $p(a,m)$, die steeds zo dicht mogelijk onder de a gekozen wordt. Bij niet-conditionele toetsen kan men het kritieke gebied zó uitbreiden dat bij sommige marginalen de conditionele kans op ten onrechte verwerpen 'iets' boven de 5% uitkomt, terwijl de totale (= niet-conditionele) kans op ten onrechte verwerpen toch onder de 5% blijft. Deze uitbreiding van het kritieke gebied geeft natuurlijk 'gratis' ook een groter onderscheidend vermogen, en dat is, zoals Snijders terecht opmerkt, een belangrijk argument vóór niet-conditionele toetsen.
Een probleempje hierbij is dat de niet-conditionele kans op ten onrechte verwerpen afhangt van de ware verdeling van de marginalen. Als Pr_θ de ware verdeling is en als O in H_0 ligt. dan is de kans op ten onrechte verwerpen

$$Pr_\theta(\text{verwerpen}) = \sum_m Pr(\text{verwerpen} \mid m)\, Pr_\theta(m) \qquad (\bullet)$$

en terwijl Pr(verwerpen | m) *niet* van *O* afhangt (hetgeen conditioneel toetsen zo aantrekkelijk maakt) hangt $Pr_\theta(m)$ *wel* van O af. Dit kan bijvoorbeeld opgelost worden door als niveau van het kritieke gebied (of als overschrijdingskans) het supremum van $Pr_\theta(\text{verwerpen})$ te nemen over alle $\theta \in H_0$.
Een andere oplossing is het gebruik van één benadering voor alle 0. Dit doet Snijders in feite, omdat hij $\underline{X}^2$ als toetsingsgrootheid neemt, en (voor alle 0) de verdeling van X^2 benadert met een χ^2-verdeling.
Een andere verfijning is nog dat men onafhankelijk van de steekproefopzet hetzij één hetzij beide marginalen kan 'de-conditioneren' (oftewel middelen zoals in (*). Zo houdt Snijders nog steeds één marginaal vast, omdat steekproefopzet 4b, die hij hanteert, nu eenmaal één marginaal vastlegt.
Bij opzet 4a kan men over beide marginalen de-conditioneren. Ter onderscheiding zullen we toetsen meestal noemen naar de verdeling waarop ze gebaseerd zijn:

twee marginalen vast	: hypergeometrisch of conditioneel	
een marginaal vast	: productmultinomiaal	niet-conditioneel
alleen het totaal aantal waarnemingen vast	multinomiaal	niet-conditioneel

Met een niet-conditionele toets bedoelen we een toets waarbij niet geconditioneerd wordt. Bij opzet 4a is dit dus multinomiaal en bij 4b productmultinomiaal.
(Toegegeven bij opzet 4a zou een productmultinomiale toets ook conditioneel zijn, maar dit lijkt op ZIJD minst een zeer weinig voorkomende variant. En in A bijvoorbeeld i) en ii) zou je alleen van hypergeometrisch mogen spreken en strict genomen niet van conditioneel.)

Dat het verschil tussen de conditionele grootte en het supremum van de niet-conditionele grootten van het kritiek gebied bij 2 x 2 tabellen vaak aanzienlijk is (en veel groter dan wij verwacht hadden) blijkt o.a. uit de literatuur die Snijders opgeeft. McDonald, Davis & Milliken (1977) zeggen dat de niet-conditionele grootte 'gewoonlijk' tussen ½ en ¼ van de conditionele grootte ligt. Garside & Mack (1976) geven (o.a.) dat bij een conditionele $a = 0.05$ en 70 waarnemingen de niet-conditionele kans steeds kleiner dan 0.03 is (bij één stel vaste marginalen met 20 en 50 waarnemingen). Zie verder ook Boschloo (1970). Overigens behandelen alle drie artikelen alleen product-multinomiale toetsen. Waarom het multinomiale geval herhaaldelijk volledig genegeerd wordt is ons niet duidelijk. Voor kruistabellen met meer vrijheidsgraden kennen wij hierover geen literatuur. Wel rapporteren Roscoe & Byars (1971) en Margolin & Light (1974) overschrijdingskansen voor (alweer) product-nomiale toetsen op onafhankelijkheid voor 2×2 tot 5×5, respectievelijk voor 2×3 tabellen. Zij vergelijken echter alleen de niet-conditionele verdeling met de χ^2-benadering en niet de niet-conditionele en de conditionele verdeling. Bovendien behandelen ze alleen enkelvoudige nulhypothesen, d.w.z. één gegeven vaste marginale kansverdeling en één gegeven vaste marginaal. In tabel 1 geven wij voor een 3 x 3-tabel een aantal overschrijdingskansen van de veel gebruikte drempelwaarde $9.49 = \chi^2_{.05,vg=4}$ en wel voor verschillende hypergeometrische, product-multinomiale en multinomiale verdelingen (d.w.z. wederom alleen voor enkelvoudige nulhypothesen). Het eerste deel van de tabel betreft $N = 9$: en dan zijn er zeven mogelijke marginalen zonder nullen. Het tweede deel betreft $N = 21$; en dan zijn er teveel mogelijke marginalen om ze afzonderlijk te vermelden. We geven een histogram in stengel-en-blad diagram ('stem and leave display' van J.W. Tukey - Zie bijvoorbeeld zijn boek Exploratory Data Analysis, Addison Wesley, 1977), plus de marginalen van de meest extreme tabellen. Zo betekent de eerste regel dat er vier hypergeometrische verdelingen zijn waarvan de verwerpingskans boven de vijf procent ligt, namelijk 5.1, 5.3, 5.5 en 5.6 procent. Dit zijn de verdelingen waarvan de ene marginaal steeds (7,7,7) is en de andere respectievelijk (12,6,3), (8,7,6), (9,6,6) en (15,3,3). Een conclusie uit deze tabel is dat bij een 3 x 3-tabel met 21 waarnemingen de χ^2• benadering voor de product-multinomiale toets (dit wordt straks T3b) in een met-verwaarloosbaar aantal gevallen zeer grof is, zij het dat de grove fouten steeds naar de veilige kant zijn. Voor de multinomiale toets is de benadering heel goed (bij deze specifieke HO).

Het toetsen scala

De toetsen die tot nu toe genoemd zijn, zullen we nog even op een rijtje zetten en aan een globaal vergelijkend consumentenonderzoek onderwerpen aan de hand van de criteria a-e. Het verschil tussen Tl t/m T4 zit hem in het wel of niet conditioneren en het exact bepalen dan wel benaderen van de verdeling van de toetsingsgrootheid. In elk van de vier gevallen zijn weer verschillende toetsingsgrootheden mogelijk. De vier meest gebruikte zijn:

1) Pearsons $X^2 = \Sigma(\text{expected} - \text{observed})^2 / \text{expected}$
2) de likelihood ratio $LR = 2 \text{ I obs } \ln(\text{obs}/\text{exp})$
3) Freeman-Tukeys grootheid $FT = \Sigma(\sqrt{\text{obs}} + \sqrt{\text{obs} + 1} - \sqrt{4\,\text{exp} + 1})^2$.
4) In 2 x 2-tabellen: de waargenomen celfrequentie X_{11} in de linkerbovenhoek.

T1. Fishers exacte toets + generalisaties

Hypergeometrische toetsen, conditioneel op de waargenomen marginalen en

Tabel 1

Enkele exacte overschrijdingskansen voor Pearson's K^2 bij een nominaal overschrijdingsniveau van 5% in 3x3 tabellen met 9 en 21 waarnemingen

N=9

	twee marginalen vast (rijmarginaal is (3,3,3) voor alle tabellen)	één marginaal vast (kansvector voor dr rij marginaal is (1/3,1/3,1/3))
kolommarginaal	$Pr_{HG}(X^2 \geq \chi^2_{.05,vg=4})^{++}$	$Pr_{PM}(X^2 \geq \chi^2_{.05,vg=4})^{++}$
(1,3.1)	10.00	4.42
(5,2,2)	7.14	3.32
(4,3,2)	5.71	3.44
(5,3,1)	3.57	3.4«
(4,4,1)	0.00	3.69
(6,2,1)	0.00	3.93
(7,1,1)	0.00	6.01

$Pr_M(X^2 \geq \chi^2_{.05,vg=4}) = 0.0333$

N=21

stengel-en-blad diagram voor $Pr_{HG}(X^2 \geq \chi^2_{.05,vg=4})^{++}$

twee marginalen vast

(rijmarginaal (7,7,7) voor alle tabellen)

	kolommarginaal	
5«	1356	(12,6,3);(8,7,6);(9,6,6);(15,3,3)
4	001235588889	
3	0236778	
2	577888	
1	77*888	
0«	00	(18,2,1);(19,1,1)

stengel-en-blad diagram voor $Pr_{PM}(X^2 \geq \chi^2_{.05,vg=4})^{++}$

één marginaal vaat

(kansvector voor dr rijmarginaal is (1/3,1/3,1/3) voor alle tabellen)

	kolommarginaal	
5«		
4	011222233344445555666	
3	00166788	
2	2669988	
1		
0«	8	(19,1,1)

$Pr_M(X^2 \geq \chi^2_{.05,vg=4}) = 0.0446$

NB. HG • Hypergeometrisch; PM > Product-multinomiaali *n* - Multinomiaal
»g • vrijheidsgraden; »• - waarden zijn »ft 100 vermenigvuldigd

wel zonder randomisatie. De verdeling wordt exact bepaald (vandaar de naam 'exacte toets'). Voor éénzijdige hypothesen in 2 x 2-tabellen zijn de eenzijdige versies van bovengenoemde vier toetsgrootheden equivalent (en UMP). Voor tweezijdige hypothesen in 2 x 2-tabellen zijn ze slechts asymptotisch equivalent. Voor tabellen met meer dan een vrijheidsgraad heeft 4) geen simpel eenzijdig equivalent en zijn 1), 2) en 3) weer alleen asymptotisch equivalent.

T2. De Boschloo, McDonald, Davis & Milliken toets
(+ generalisatie naar r x c).
Niet-conditionele toetsen met ab toetsingsgrootheid de conditionele overschrijdingskans.
We moeten nu drie 'onbetrouwbaarheidsdrempels' onderscheiden:
- de vooraf gegeven 'toegestane onbetrouwbaarheid' α_t, zeg 5%;
- de 'gerealiseerde niet-conditionele onbetrouwbaarheid' α_g. Deze moet dus kleiner gelijk α_t zijn en dus liefst zo dicht mogelijk er bij om zo veel mogelijk onderscheidingsvermogen te realiseren;
- de toegestane 'conditionele onbetrouwbaarheid' α_c.

Toets T2 gaat nu als volgt. Kies eerst één van de toetsingsgrootheden 1) t/m 4), of een eigen favoriet. (Boschloo en McDonald e.a. bekijken alleen 2 × 2-tabellen met éénzijdige alternatieve hypothese en dan zijn alle 4 genoemde mogelijkheden equivalent). Voer nu een conditionele toets uit op niveau α_c, maar vermeldt als niveau het niet-conditionele niveau α_g. Als α_t (of α_g) gegeven is kies je het conditionele niveau α_c zo groot als mogelijk is zonder het niet-conditionele niveau te groot (d.w.z. groter dan α_t) te laten worden. Merk op dat je altijd $\alpha_c \geqslant \alpha_t$ kunt nemen. Boschloo geeft tabellen waarmee je α_c bij gegeven α_t kunt opzoeken, en Buhrman (1979) geeft tekst en beschrijving van een computerprogramma om α_c te berekenen. Het volgende voorbeeld dat van Boschloo afkomstig is geeft aan hoe dramatisch de verschillen tussen α_c en α_t kunnen zijn. In een 2 x 2-tabel met één vaste marginaal met 15 en 10 waarnemingen hoort bij een $\alpha_t = .05$ een $\alpha_c = 09$!
Merk nog op dat de verzamelingen kritieken gebieden van Tl en T2 gelijk zijn, althans indien beide op dezelfde toetsingsgrootheid gebaseerd zijn. Het verschil zit hem echter in de wijze waarop de onbetrouwbaarheid van een kritiek gebied berekend wordt.

T3. χ^2-toetsen
Niet-conditionele toetsen met als toetsingsgrootheid één van de genoemde vier of een variant daarop. met name voor 2 x 2-tabellen zijn er allerlei 'continuiteitscorrecties' voorgesteld. De meest bekende is wellicht de correctie die naar Yates (1934) genoemd wordt. Zie verder ook Conover (1974). Schouten, Molenaar, van Strik & Boomsma (in druk) en vooral ook Garside & Mack (1976).
We onderscheiden nog:

T3a. De verdeling van de toetsingsgrootheid wordt exact bepaald en voor het niveau wordt weer het sup over alle verdelingen uit H_0 genomen.

T3b. De verdeling van de toetsingsgrootheid wordt (voor elke verdeling uit H_0) benaderd door een χ^2-verdeling. Dit is de meest gebruikte variant van T3, en natuurlijk is de bedoeling van bovengenoemde correcties om deze benadering beter te maken. (Opmerkelijk is dat bij de voorgestelde correcties in het algemeen geen onderscheid gemaakt wordt tussen

multinomiale toetsen en product-multinomiale toetsen.) Wel kan het zijn dat tengevolge van de toegepaste correctie de waarden van de toetsingsgrootheid zo veranderen, dat de tabellen anders geordend worden. Met andere woorden, door de toetsingsgrootheid te 'corrigeren' creëert men een andere toets. Vergelijk Conover (1974) ten aanzien van X^2 en Yates' correctie. Men zou voor zo'n nieuwe toets een eigen rechtvaardiging moeten vinden, maar aangezien toch alle hier beschreven toetsen asymptotisch equivalent zijn en hun rechtvaardiging meestal ook op asymptotische argumenten is gebaseerd, is dit nauwelijks mogelijk.

T4. de UMPU-toets
De toets die onder de *zuivere* toetsen in een 2 x 2-tabel uniform meest onderscheidend (d.w.z. uniformly most powerful unbiased) is voor alle in par. 4 (1979) beschreven onderzoeksopzetten, is Fishers exacte toets, maar dan met randomisatie zó dat bij alle marginalen de kans op verwerpen precies hetzelfde is.

Het vergelijkend consumentenonderzoek

Conclusies uit tabel 2:
1. In de praktijk valt toets T4 in zeer veel gevallen af, namelijk
   - bij eenmalig onderzoek
   - bij het vermelden van overschrijdingskansen (als tegenstelling tot toetsen met een vooraf gegeven niveau)
   - bij grotere tabellen dan 2×2
2. In gevallen zoals de hierboven in A gegeven voorbeelden i) en ü) blijft alleen de exacte, conditionele toets Tl over.
3. Bij steekproefopzet 4.a en 4.b (zie het novembernummer 1979) lijkt bij 2 x 2-tabellen T3b met een geschikt gecorrigeerde X^2 als toetsingsgrootheid een goede keus, die makkelijker te berekenen is dan T2. De laatste heeft echter het voordeel dat hij (bij aanzienlijk meer rekenwerk) exacte overschrijdingskansen oplevert (althans het supremum daarvan). Bij grotere tabellen verdient in de praktijk T3b meestal de voorkeur, indien tenminste bekend is dat de approximatie voldoende nauwkeurig is. Het doorslaggevend argument is hierbij de eenvoud van de berekeningen. Als de kwaliteit van de approximatie twijfelachtig is levert Tl een alternatief dat in elk geval met een computer snel te berekenen is. Voor T2 en T3a is al gauw zeg in de orde van 1000 maal zoveel computertijd nodig. Een nadeel van T2 en T3 is dat niet-conditioneel toetsen niet steeds onaanvechtbaar is.
4. Ons inziens is Snijders' afwijzing van conditionering geldig onder de volgende omstandigheden
   - voor 2 x 2-tabellen met steekproefopzet 4a of 4b (zie 1979)
   (En dan verdient het gebruik van een gecorrigeerde X^2 de voorkeur boven de gewone X^2).

Voor grotere tabellen is er echter nog nauwelijks numerieke informatie over het verschil in onderscheidingsvermogen tussen Tl enerzijds en T3b anderzijds. In het bijzonder is voor tabellen met een aantal kleine verwachte waarden niet duidelijk wanneer de fout in de approximatie bij T3b opweegt tegen de (vermoedelijke) winst aan onderscheidingsvermogen indien op een correct niveau getoetst zou worden.

We besluiten deze paragraaf met de volgende relativering. Zelfs als er informa-

Tabel 2
Globale aanduiding van enkele eigenschappen van de bovengenoemde toetsen op onafhankelijkheid in een rxc-tabel. Indien over 2x2-tabellen meer informatie beschikbaar is en indien 2x2-tabellen zich anders gedragen dan grotere tabellen dan j» de informatie voor 2x2-tabellen tussen ( ) geplaatst.

	Eigenschappen De kans p_θ op verwerpen			Berekeningen				
	Onder H_o		Onder H_a					
toets	Is $\sup\{p_\theta : \theta \in H_o\}$ kleiner dan a?	maat- .. gelijk ! zuiver	(onderscheidingsvermogen)	eenvoud [1])	exact of approximatie	randomisatie?	conditioneel?	Toepasbaar in opzet 4.a en 4.b en in vb, i 4 ii?[6])
T_1 Fisher	kleiner (veel kleiner)	nee, nee	klein (resp. zeer klein) tov T_2[2]) voor alle θ.	computer (zakrekenmachientje of tabellen)	exact	nee	j»	4.a, 4.b en vb i t ii
T_2 Boschloo	<, naar bijna =	nee, nee ongeveer als T_1	goed?[2]) (ongeveer als T_4)	zeer veel werk (computer of tabellen)	exact	nee	nee	4.a, 4.b
T_{3a} χ^2,LR,(Yates) etc., exact	<, maar bijna =	nee, nee minder dan T_1	goed?[2]) (ongeveer als T_4)	(computer)	exact	nee	nee	4.a, 4.b
T_{3b} idem met χ^2-approx	hopelijk ongeveer •	nee, nee minder dan T3a	goed?[2]) (goed[4])	eenvoudig	approx[5])	nee	nee	4.a, 4.b
(T_4 UMPU) (alleen voor 2x2)	(exact gelijk)	(ja, ja)	(optimaal onder de zuivere toetsen)	(computer of zakrekenmachientje of tabellen)	(exact)	(ja)	(ja)	(4.a, 4.b, vb i & ii)

t) 4.a. en 4.b. verwijst naar par. 5 (1979);vb i t ii staan hierboven vermeld.
5) voor. kleine aantallen is niet goed bekend vanneer deze approximatie te grof wordt.
4) voor 2x2-tabellen zijn er varianten op χ^2 ("net correctie") die heel nauwkeurig χ^2-verdeeld zijn, althans bij α=.05 er. α=.01.
3) we hanteren 4 categorieën, alsvolgt gecrdend: eenvoudig en/of tabellen; rekenmachientjes; computer; zeer veel werk.
2) voor grotere tabellen dan 2x2 is weinig numerieke informatie beschikbaar.
I) d.w.z. is p_θ voor alle $\theta \in H_o$ hetzelfde?

tie is over het verschil in onderscheidingsvermogen van de verschillende toetsen, dan nog is het bij veel toepassingen van r x c-tabellen geenszins duidelijk tegen welke alternatieven een groot onderscheidingsvermogen gewenst is. De alternatieve hypothese ('afhankelijkheid') is vaak noodzakelijk erg vaag (en minder vaak een gevolg van onvoldoende nadenken vooraf). Men moet zich in zo'n geval wel afvragen hoe zinvol het is om de overschrijdingskans van één bepaalde toetsingsgrootheid heel nauwkeurig te berekenen als we eerst een tamelijk arbitraire keus gemaakt hebben tussen verschillende toetsingsgrootheden die in verschillende overschrijdingskansen resulteren. Wel is het ons inziens belangrijk dat er meer inzicht verworven wordt in de mate waarin de toetsen van elkaar verschillen, in de nauwkeurigheid van verschillende χ^2-benaderingen en in de condities waaronder deze benaderingen 'voldoende' nauwkeurig zijn. Het nut van T1, T2 en T3a ligt onder andere in zulk theoretisch onderzoek.

C. Corrigenda

Wachtel Sekhuis maakt ons attent op enkele slordigheden in paragraaf 5 van het artikel in het novembernummer. Allereerst is de conclusie halverwege pagina 21 niet helemaal correct. Weliswaar is voor elke permutatie π de kans van de rij waarnemingen $(Y_1,Z_{\pi(1)}),\ldots,(Y_n,Z_{\pi(N)})$ wel hetzelfde, maar daaruit volgt niet dat die kans, gegeven de marginalen, dan 1/N! is. Immers, niet alle permutaties leiden tot verschillende reeksen waarnemingen, omdat (zeker in een kruistabel) niet alle Zj (hoeven te) verschillen! Zo leidt de rij waarnemingen (1,1) (2,1) (2,2) niet tot 3! =6 verschillende rijen met dezelfde marginalen, maar slechts tot drie. Deze fout heeft echter geen consequenties voor de rest van het betoog, noch voor het bovenaan pag. 20 beschreven simulatie-algoritme.

Verder is ook de rechtvaardiging voor conditionering op pag. 21 en 24 punt i en ii erg slordig geformuleerd. Er staat bijna dat alle permutatietoetsen zuiver en maatgelijk zijn. Een betere formulering is als volgt (onder de gegeven H_0 en H_a en bij onderzoeksopzet 4a of 4b):

i) Onder deze condities is elke zuivere toets maatgelijk (het argument is dat het onderscheidingsvermogen continu is, en dat elke verdeling uit H_0 de limiet is van een rij verdelingen uit H_a).

ii) Onder deze condities is elke maatgelijke toets een toets waarbij onder elke conditionering op de marginalen de kans op verwerpen hetzelfde is (Dit volgt uit Lehmann - Testing Statistical Hypotheses - stelling 4.3.2). Zo'n toets heet een (gerandomiseerde) permutatietoets.

iii) Om bij elke conditionering op de marginalen dezelfde kans op verwerpen te realiseren is in het algemeen randomiseren noodzakelijk.
Zie ook Tom Snijders, tweede helft pag. 17. Overigens is het ons niet geheel duidelijk of er bij grotere tabellen dan 2 x 2 nu (gerandomiseerde) toetsen op onafhankelijkheid bestaan die precies zuiver zijn. Kan iemand ons aan literatuur of ideeën helpen?

Hoe verwarrend de discussie in de literatuur is, moge blijken uit de volgende voorbeelden. Zo roept Berkson (1978a, p. 41) (een tegenstander van conditioneren) ten onrechte Plackett (1977) aan ter ondersteuning van zijn standpunt. Plackett laat echter juist zien dat op basis van alleen de marginalen van een 2 x 2-tabel de meest aannemelijke schatter voor het kruisproduct gedegenereerd is (namelijk $\pm\infty$). Dit is een formeel argument ten gunste van Fishers standpunt dat de marginalen geen relevante informatie bevatten (maar zijn

er misschien andere, betere schatters? En leveren marginalen + waargenomen kruisproduct niet misschien meer informatie op dan bij conditionele inferentie gebruikt wordt?). Bovendien stelt Berkson (1978b) 'relevante informatie' expliciet gelijk aan 'informatie' Dit betekent o.i. dat hij eist dat de conditionele verdeling onafhankelijk moet zijn van de marginalen, en dat is natuurlijk niet zo, en dat was ook nooit Fishers bedoeling. Corsten en de Kroon (1979) leveren weer andere kritiek op Berkson (1978a), maar hun argumenten vóór conditioneren zijn ons nog niet duidelijk geworden.

Een ander misverstand lijkt te vinden in Boschloo (1970) die vermeldt dat Barnard eerst een met-conditionele toets voorstelde, maar later gezwicht zou zijn voor Fishers argumenten. Uit Barnards antwoord (1945b) op Fishers kritiek (1945) ten aanzien van zijn artikeltje in Nature (1945a) en uit het daarop volgende artikel (1947) blijkt niets van dit alles. In tegendeel, Barnard geeft in 1947 een uiteenzetting van de drie onderzoeksopzetten die wij vermeldden (par. 4 (1979) + A voorbeeld i en ii) en geeft nog eens aan dat voor toetsing van de gelijkheid van twee proporties zijn eigen voorstel bij uitstek geschikt is en niet de exacte toets van Fisher.

Tenslotte zijn o.i. de voorstellen van Boschloo (1970) en McDonald e.a. (1977) volstrekt gelijk, hetgeen door de laatsten in nevelen gehuld schijnt te worden (p. 146).

Literatuur

Barnard, G.A. (1945a): "A new test for 2 x 2-tables". Nature 156, 177.

Barnard, G.A. (1945b): "A new test for 2 x 2-tables". Nature 156, 783-784.

Barnard, G.A. (1947): "Significance tests for 2 x 2-tablcs". Biometrika.

Berkson, J. (1978a): "In dispraise of the exact test". J. Stat. Planning and Inference 2, 27-42.

Berkson, J. (1978b): "Do marginal totals of the 2 x 2-table contain relevant information respecting the table proportions?" J. Stat. Planning and Inference 2, 43—44.

Boschloo, R.D. (1970). "Raised conditional level of significance for the 2 x 2-table when testing the equality of two probabilities". Statistica Neerlandica 24, 1–35.

Buhrman, J.M. (1979): "The computation of an unconditional critical level when testing the equality of two unknown probabilities". (SN 10/79, Stat. Report) Math. Centrum. Amsterdam.

Conover, W.J. (1974): "Some reasons for not using the Yates continuity correction on 2 x 2 contingency tables". J. Amer. Sut. Association 69, 374-382 (with comments and rejoinder).

Corsten, L.C.A. A J.P.M. de Kroon (1979): "Comment on J. Berkson's paper "In dispraise of the exact text" ". J. Stat. Planning and Inference 3. 193-197.

Cox, D.R. (1958): "Some problems connected with statistical inference". The Annals of Mathematical Statistics 29, 357-372.

Fisher, R.A. (1945): "A new test for 2 x 2-tables". Nature 156. 388.

Garside, G.R. A C. Mack (1976): "Actual type I error probabilities for various tests in the homogeneity case of the 2 x 2 contingency table". American Statistician 30. 18-21.

Kroonenberg, P.M. A A. Verbeek (1979):"Bepaling van exacte overschrijdingskansen var X^2 in r × c-tabellen". VVS Bulletin 12 (11). 14-21, 24, 8.

McDonald, L.L., B.M. Davis & G.A. Milliken (1977): "A nonrandomized unconditional test for comparing two proportions in 2 x 2 contingency tables". Technometrics 19, 145–157.

Margolin, B.H. A RJ. Light (1974): "An analysis of variance for categorical data 11: small sample comparison with chi-square and other competitors". J. Amer. Stat. Association 69, 755-764.

Plackett, R.L. (1977): "The marginal totals of 2 x 2-table". Biometrika 64, 37-42.
Roscoe, J.T. & J.A. Byars (1971): "An investigation of the restraints with respect to sample size commonly imposed on the use of the chi-square statistic". J. Amer. Stat. Association 66, 755-759.
Schouten, H.J.A., I.W. Molenaar, R. van Strik A A. Boomsma: "Comparing two independent binomial proportions by a modified chi-square test". Biometrical Journal (Biometrisches Zeitschrift), in druk.
Snijders, T. (1980): "De χ^2-toets voor kruistabellen: konditioneren of niet?" VVS Bulletin 13 (2), 16–21, 14.
Yates, F. (1934): "Contingency table involving small numbers and the χ^2-test". J. Roy. Statist. Soc., Suppl. 1, 217–235.