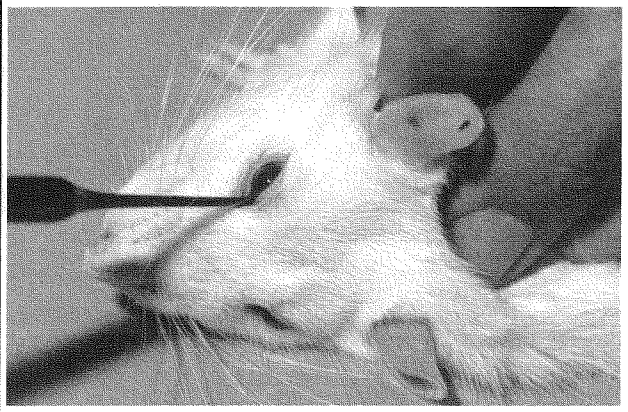
Hein A. Van Lith

Afdeling Proefdierkunde, Faculteit der Diergeneeskunde, Universiteit Utrecht, Postbus 80.166, 3508 TD Utrecht

Inleiding

In de voorgaande afleveringen van *(Basale) Statistiek en Dierexperimenten* hebben we ons beziggehouden met het bestuderen van één kenmerk (variabele), waarvoor we gegevens hebben verzameld en waarvoor eventueel ook kengetallen waren uitgerekend. In de praktijk gebeurt het dikwijls dat twee of meer variabelen gelijktijdig worden waargenomen. Vanaf het ogenblik dat we met meer dan één variabele te maken hebben komt het probleem van samenhang aan de orde.

Afbeelding 1. Orbitapunctie bij een rat onder ethernarcose.



Het bestuderen van de samenhang tussen variabelen is een zeer belangrijk onderdeel van de statistiek. Bij het (statistisch) onderzoek naar de samenhang tussen twee variabelen spreken we van *bivariate technieken*, waarbij *'bi'* verwijst naar *'twee'* en *'variaat'* naar het Engelse *'variate'* wat *'veranderlijke die aan het toeval onderhevig is'* betekent. Van diverse technieken bestaan uitbreidingen voor de analyse van verbanden tussen meer dan twee variabelen: *multivariate technieken*. In deze serie beperken we ons tot het beschrijven van waarnemingsresultaten van twee kenmerken.

Men spreekt van *associatie* bij variabelen (kenmerken) met een nominaal meetniveau of een beperkt ordinaal meetniveau. *Rangcorrelatie* wordt gehanteerd bij variabelen met minimaal ordinaal niveau en (bijna) volledige ordening en *correlatie* bij een interval/ratio meetniveau (Biotechniek 2003-1). Er zijn een onnoemelijk aantal maten voor samenhang opgesteld, afhankelijk van het meetniveau van de variabelen. In de volgende paragrafen wordt de samenhang tussen twee nominale variabelen besproken.

De kruistabel

Voor het analyseren van de samenhang tussen twee (of meer) variabelen op nominaal niveau wordt gebruik gemaakt van *kruistabellen* (ook wel *contingentietabellen* genoemd), althans als het aantal mogelijke waarden van beide nominale variabelen niet al te groot is. Het gaat hier om een tabel waarbij twee variabelen gekruist worden. Vandaar de benaming. Men plaatst de categorieën van de ene variabele in de *rijen* en de categorieën van de andere variabele in de *kolommen*. Wanneer er een onderscheid kan worden gemaakt tussen een *afhankelijke* en *onafhankelijke* variabele bestaat er een lichte voorkeur om de categorieën van de afhankelijke variabele in de rijen te plaatsen. Hierbij is de afhankelijke variabele die variabele, waarvan de waarden bepaald worden door (afhankelijk zijn van) de waarden van de onafhankelijke variabele.

De effecten van een éénmalige orbitapunctie bij ratten onder ethernarcose (afbeelding 1) zou wel eens af kunnen hangen van de uitvoerder. Tabel 1 geeft de frequentieverdeling van 474 oogbollen, verdeeld naar biotechnicus en naar de stand van de oogbol. De totalen aan de rand (voor de rijen en kolommen) heten de *randtotalen* of ook wel *randfrequenties*. Elke combinatie van waarden van de beide variabelen vormt een cel. Het aantal waarnemingen in iedere cel geeft aan hoe vaak een combinatie van beide variabelen voorkomt en wordt de *celfrequentie* genoemd.

Tabel 1. Waargenomen frequenties. Vier biotechnici voeren een éénmalige orbitapunctie bij ratten onder ethernarcose uit.

Effecten op de stand van de oogbol

	Stand van de oogbol* (gepunctueerde zijde)			
biotechnicus	exophthalmus	enophthalmus	normaal	totaal
A	1	2	99	**102**
B	6	7	137	**150**
C	6	10	134	**150**
D	7	16	49	**72**
totaal	**20**	**35**	**419**	**474**

**Exophthalmus = het naar voren geplaats zijn van de oogbol;*
Enophthalmus = het terugzakken van de oogbol in de oogholte

Tabel 2. Resultaten van tabel 1, maar nu procentueel weegegeven (rijpercentages)

	Stand van de oogbol* (gepunctueerde zijde)			
biotechnicus	exophthalmus	enophthalmus	normaal	totaal
A	1.0	1.9	97.1	**100.00**
B	4.0	4.7	91.3	**100.00**
C	4.0	6.7	89.3	**100.00**
D	9.7	22.2	68.1	**100.00**

**Exophthalmus = het naar voren geplaats zijn van de ogbol;*
Enophthalmus = het terugzakken van de oogbol in de oogholte

Om beter te kunnen onderscheiden wat in een kruistabel precies wordt weergegeven is het verstandig om niet naar de absolute aantallen te kijken, maar naar percentages. Op basis van de tabel met de absolute frequenties kan een tabel met diverse relatieve frequenties worden berekend:

1 rijpercentages: hierbij worden de frequenties van iedere cel gedeeld door het bijbehorende rijtotaal (in de laatste kolom) en vermenigvuldigd met 100
2 totaalpercentages: hierbij worden de frequenties van iedere cel gedeeld door het totaal generaal (in de tabel rechtsonder) en vermenigvuldigd met 100
3 kolompercentages: hierbij worden de frequenties van iedere cel gedeeld door het bijbehorende kolomtotaal (in de laatste rij) en vermenigvuldigd met 100.

Tabel 3. Verwachte frequenties. Vier biotechnici voeren een eenmalige orbitapunctie bij ratten onder ethernarcose uit.

Effecten op de stand van de oogbol

	Stand van de oogbol* (gepunctueerde zijde)			
biotechnicus	exophthalmus	enophthalmus	normaal	totaal
A	4.31	7.53	90.16	**102**
B	6.33	11.08	132.59	**150**
C	6.33	11.08	132.59	**150**
D	3.04	5.32	63.64	**72**
totaal	**20**	**35**	**419**	**474**

**Exophthalmus = het naar voren geplaats zijn van de ogbol;*
Enophthalmus = het terugzakken van de oogbol in de oogholte

Indien men nu de rijpercentages bestudeert (tabel 2) valt op, dat de percentages van exophthalmus en enophthalmus bij biotechnicus A veel lager zijn dan bij de andere biotechnici. Bij biotechnicus D komen afwijkingen van de stand van de oogbol relatief gezien het meest voor. We zouden kunnen zeggen, dat er tussen de kenmerken 'biotechnicus' en 'stand van de oogbol' wellicht een relatie (samenhang) bestaat. Of dat een kwestie is van kundigheid van de biotechnicus of van de wijze van uitvoering, is uit deze kruistabel niet op te maken; daarvoor zouden aanvullende gegevens moeten worden verzameld.

Wanneer de beide variabelen onafhankelijk zijn, kunnen we van te voren uitrekenen hoeveel waarnemingen er per cel moeten komen. Iedere celwaarde moet gelijk zijn aan het product van de randtotalen gedeeld

Tabel 4. Berekening van de Chi-kwadraat op basis van de gegevens van de tabellen 1 en 3

rij	kolom	waargenomen	verwacht	(waargenomen – verwacht)	(waargenomen – verwacht)^2	(waargenomen – verwacht)^2/verwacht
A	exophthalmus	1	4.31	-3.31	10.9561	2.542018561
A	enophthalmus	2	7.53	-5.53	30.5809	4.061208499
A	normaal	99	90.16	8.84	78.1456	0.866743567
B	exophthalmus	6	6.33	-0.33	0.1089	0.017203791
B	enophthalmus	7	11.08	-4.08	16.6464	1.502382671
B	normaal	137	132.59	4.41	19.4481	0.146678482
C	exophthalmus	6	6.33	-0.33	0.1089	0.017203791
C	enophthalmus	10	11.08	-1.08	1.1664	0.105270758
C	normaal	134	132.59	1.41	1.9881	0.014994343
D	exophthalmus	7	3.04	3.96	15.6816	5.158421053
D	enophthalmus	16	5.32	10.68	114.0624	21.440300752
D	normaal	49	63.64	-14.64	214.3296	3.367844123
som		474	474	0		**39.240270392**

door het totale aantal. Indien dit toegepast wordt op de gegevens van tabel 1 krijgen we tabel 3. Als de waargenomen en de verwachte celfrequenties in iedere cel hetzelfde zijn, berust de verdeling blijkbaar op toeval en zijn de indelingscriteria van de variabelen dus *onafhankelijk* van elkaar. Uiteraard worden in de praktijk nooit exact de verwachte frequenties als waargenomen frequenties geregistreerd, zodat er altijd wel een verschil zal zijn. Bestaat er daarentegen een groot verschil tussen waargenomen en verwachte celfrequenties dan berusten de waargenomen frequenties waarschijnlijk niet op toeval en is er wel sprake van een samenhang. Met behulp van een statistische toets kan overigens worden nagegaan of er een *significant* (= verantwoorde conclusies toelatend) verband bestaat tussen twee nominale variabelen in een kruistabel.

Kruistabellen zijn eigenlijk alleen maar bedoeld voor nominale gegevens. Echter, wanneer we een gemengde verzameling variabelen hebben (d.w.z. een nominale en een ordinale variabele) is het ook zinvol om hiervoor een kruistabel te maken.

Chi-kwadraat

De grootheid die men op basis van de verwachte en waargenomen frequenties kan uitrekenen heet de *Chi-kwadraat* en wordt aangeduid met de Griekse letter Chi (en het kwadraatteken): χ^2. De berekening van de Chi-kwadraat is als volgt: bereken de verschillen tussen de waargenomen en verwachte frequenties, kwadrateer de verschillen, deel de kwadraten door de verwachte frequentie. De som van de uitkomsten is de Chi-kwadraat (tabel 4; vetgedrukte waarde). Wanneer de waargenomen frequenties gelijk zijn aan de verwachte, is er zoals reeds gemeld géén verband en is de Chi-kwadraat gelijk aan 0. Als er wel verband is, dan verschillen de verwachte en de waargenomen frequenties van elkaar en is de Chi-kwadraat in ieder geval groter dan 0. De Chi-kwadraat is dus een maat voor de samenhang in een kruistabel. Echter, de waarde van de Chi-kwadraat hangt af van het totaal aantal waarnemingen en van het aantal rijen en kolommen in een kruistabel. Daardoor wordt het lastig de samenhang in twee kruistabellen te vergelijken wanneer die op verschil-

lende aantallen waarnemingen zijn gebaseerd. Ofwel, de Chi-kwadraat is niet genormeerd, dat wil zeggen: er zijn geen vaste grenzen voor de hoogte of de laagte van de uitkomst. Aan de waarde van de Chi-kwadraat kun je de sterkte van de associatie dus niet aflezen.

De Chi-kwadraat zelf wordt vooral gebruikt in de toetsende statistiek. Hoe groter de Chi-kwadraat, hoe geringer de kans dat het verband toevallig is en dus hoe groter de kans dat er een systematisch verband aanwezig is. Gelukkig is de kansverdeling van de Chi-kwadraat bekend. De kans dat een gegeven waarde van Chi-kwadraat door het toeval wordt bepaald is op te zoeken in een tabel of te berekenen met een spreadsheet of statistisch programma. In het algemeen geldt dat het interpreteren van een Chi-kwadraat pas zinnig is, als voldaan is aan twee voorwaarden. Ten eerste moeten de verwachte celfrequenties groter of gelijk aan 1 zijn. Ten tweede mag het percentage cellen met een verwachte frequentie welke kleiner is dan 5, niet hoger zijn dan 20 procent. Een verdere bespreking van de Chi-kwadraat stellen we uit tot in de toetsende statistiek.

Sterkte van een samenhang

Aangezien men aan de Chi-kwadraatwaarde de sterkte van een samenhang niet kan aflezen, zijn er andere maten ontwikkeld, die dit nadeel niet hebben en bij voorkeur tussen 0 (geen samenhang) en 1 (perfecte samenhang) in liggen. Veruit de belangrijkste samenhangmaat voor nominale variabelen is de *Cramérs V* en daarom bespreken we alleen deze maat. Cramér heeft V^2 gelijk gesteld aan: χ^2 gedeeld door N(k – 1), waarbij k = aantal rijen of aantal kolommen (kleinste van de twee). *Cramérs V* zelf is hieruit de positieve wortel. Voor de waarde van V^2 kan globaal de volgende subjectieve kwalificatie worden aangehouden:

$V^2 < 0.05$: zwak/matig algemeen verband;
$0.05 \leq V^2 < 0.15$: middelmatig sterk algemeen verband;
$V^2 \geq 0.15$: sterk algemeen verband.

De *Cramérs V* voor de resultaten van tabel 1 is ≈ 0.20 en moet dus als een zwakke/matige samenhang worden betiteld.

Samenvattend

De simultane verdeling van twee nominale variabelen kan worden weergegeven door een tweedimensionale frequentietabel: een kruistabel. Zo'n tabel bevat celfrequenties en randfrequenties. Met behulp van toetsen die gebaseerd zijn op de χ^2 (Chi-kwadraat) kan onderzocht worden of er tussen de beide variabelen een statistisch verband bestaat, of dat de variabelen onafhankelijk van elkaar zijn. Aan de hand van een samenhangmaat kan een indruk verkregen worden van de sterkte van het verband. Voor nominale variabelen bestaan een groot aantal samenhangmaten. Cramér's V is de meest geschikte maat, die gebaseerd is op de grootheid χ^2.