STATISTIEK VOOR HET SECUNDAIR ONDERWIJS

Kruistabellen: exploratieve methoden

Werktekst voor de leerling

Prof. dr. Herman Callaert

Hans Bekaert
Cecile Goethals
Lies Provoost
Marc Vancaudenberg

Inhoudstafel

DEEL 1. Basisbegrippen

Kruistabellen gebruik je om het verband tussen categorische veranderlijken te bestuderen. Hoe zo'n veranderlijken eruit zien, bekijk je even vooraf.

1. Veranderlijken

Een statistische studie kan gaan over personen (tieners, voetballers …) of dieren (katten, eenden …) of planten (rozen, beuken …) of zaken (kerktorens, postzegels …). De dingen die je bestudeert, zijn de **elementen** in je studie.

Bij elk element ben je geïnteresseerd in bepaalde eigenschappen. Dat zijn de **veranderlijken**. Een geneeskundig onderzoek kan bij patiënten vragen naar het geslacht, de bloedgroep en het aantal gezonde tanden. Bij elk **element** (elke patiënt) worden hier 3 **veranderlijken** opgemeten.

Voor elke veranderlijke noteer je haar **waarde**.

- De **veranderlijke** "geslacht" heeft maar twee **waarden**: mannelijk / vrouwelijk.
- De **veranderlijke** "bloedgroep" heeft vier **waarden**: O, A, B, AB.
- De **veranderlijke** "aantal gezonde tanden" heeft als **waarden** een geheel getal tussen 0 en 32.

De **waarden** van de veranderlijken "geslacht" en "bloedgroep" omschrijf je **met woorden** (of afkortingen). De **waarden** van de veranderlijke "aantal gezonde tanden" zijn **getallen** (numeriek) die uit elkaar liggen (discreet).

De uitkomsten van elke veranderlijke in dit medisch onderzoek komen terecht in een beperkt aantal categorieën. Daarom zijn dit **categorische** veranderlijken. In deze tekst werk je met veranderlijken die terechtkomen in categorieën die elkaar niet overlappen. Elke opmeting komt terecht in één en slechts één categorie.

Voorbeeld. Bij zakjes chocolade M&M-snoepjes (Choco M&M's van 45 g) kan je de kleur bestuderen. In die zakjes zitten alleen rode, groene, gele, oranje, bruine en blauwe snoepjes. De kleur is hier de **naam** van de veranderlijke. Elk snoepje komt (qua kleur) terecht in één van de 6 mogelijke categorieën: rood, groen, geel, oranje, bruin, blauw. Dat zijn de **waarden** van de veranderlijke. De veranderlijke "kleur" is een voorbeeld van een **categorische** veranderlijke.

Opdracht 1

Geef een voorbeeld van een onderzoek waar je een eigenschap (van mensen, dieren of dingen) bestudeert waarbij de opgemeten veranderlijke een **categorische** veranderlijke is. Geef de **naam** van de veranderlijke en haar **waarden.**

2. Bivariate categorische gegevens

Hebben jongens en meisjes een verschillende voorkeur voor een eerste festivalervaring?

Aan leerlingen werd gevraagd welk festival de voorkeur verdient voor iemand die voor de eerste keer naar een festival gaat: Rock Werchter (W), Pukkelpop (P) of het Dour Festival (D). Ook werd genoteerd of de ondervraagde leerling een jongen (J) of een meisje (M) was.

Hier heb je een voorbeeld van een studie die werkt met bivariate categorische gegevens. Bij elke ondervraagde leerling (= elk element in de studie) zijn **twee** veranderlijken (**bi**variaat) opgemeten. De namen van die veranderlijken zijn "festivalvoorkeur" en "geslacht". Elk van die 2 veranderlijken is **categorisch**:

- de veranderlijke festivalvoorkeur heeft 3 waarden: W (Werchter), P (Pukkelpop) en D (Dour)
- de veranderlijke geslacht heeft 2 waarden: J (jongen) en M (meisje)

Het resultaat van deze studie was als volgt:

	Geslacht	Festivalvoorkeur
1ste leerling	J	P
2de leerling	J	W
3de leerling	M	P
4de leerling	J	P
5de leerling	M	W
6de leerling	M	P
7de leerling	M	D
8ste leerling	J	P
9de leerling	M	W
10de leerling	M	P

Opdracht 2

- Hoeveel jongens hebben een voorkeur voor Rock Werchter? ……..
- Hoeveel meisjes hebben een voorkeur voor Pukkelpop? ……..
- Hoeveel jongens hebben een voorkeur voor Dour? ……..

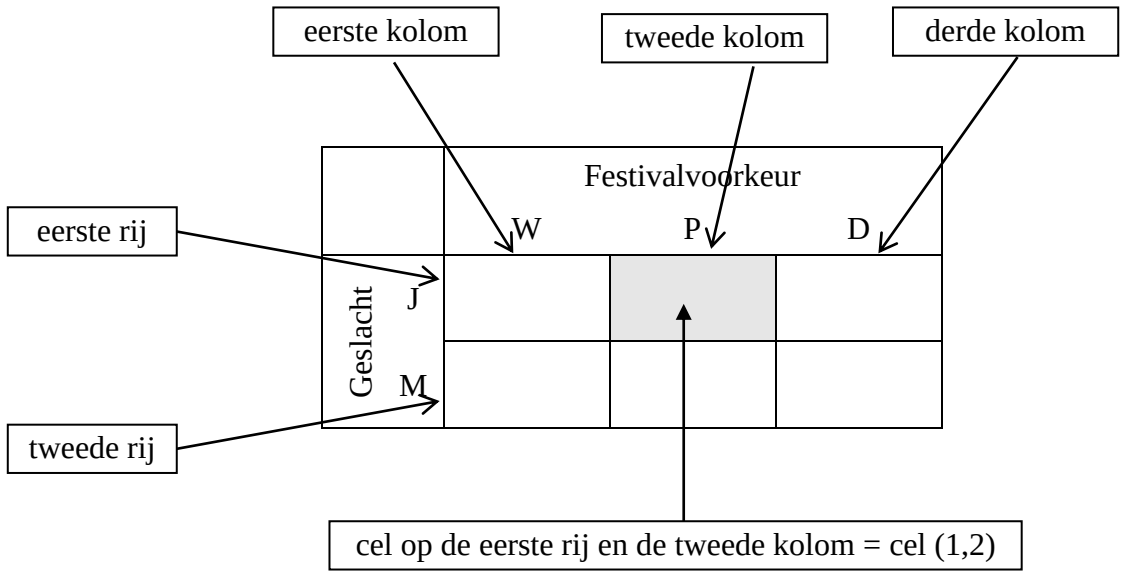
Je merkt dat "ruwe" gegevens, zoals ze bij het opmeten voluit worden genoteerd, niet handig zijn. De huidige opdracht was niet moeilijk omdat het maar om 10 leerlingen gaat. Als je op de bovenstaande vragen moet antwoorden in een onderzoek bij 150 leerlingen, dan verstuik je je ogen (en maak je telfouten). Daarom vat je de "ruwe" gegevens samen in een overzichtelijke tabel, zoals hieronder is uitgelegd.

3. Kruistabellen

De bivariate categorische gegevens van de bovenstaande studie vat je samen in een tabel met rijen en kolommen. Dat gaat als volgt:

- De waarden van een eerste categorische veranderlijke (bijvoorbeeld het geslacht) vormen de rijen. Je hebt hier 2 rijen nodig, een rij voor J (jongen) en een rij voor M (meisje).
- De waarden van de tweede categorische veranderlijke (festivalvoorkeur) vormen de kolommen. Je hebt 3 kolommen nodig: W (Werchter), P (Pukkelpop) en D (Dour).

Alles samen ziet de lay-out van de tabel er zo uit:



Een tabel zoals hierboven heet een **kruistabel** of **contingentietabel**.

In dit voorbeeld heeft de tabel 2 rijen en 3 kolommen. Dat is een (**2**×**3**) **kruistabel**. Je vermeldt altijd eerst de rij en dan de kolom. Zo is de cel (1,2) de plaats waar de eerste rij snijdt met de tweede kolom. Dat is de plaats waar je het aantal **jongens** met voorkeur voor **Pukkelpop** noteert. In algemene notatie stel je dat aantal voor door $\mathbf{n_{ij}}$. Ook hier is de volgorde van belang. Met n_{ij} bedoel je het **aantal** elementen in je studie die tegelijkertijd in de i^{de} rij en de j^{de} kolom terechtkomen. In dit voorbeeld is $n_{12} = 3$ want er zijn juist 3 **jongens** die voor **Pukkelpop** kiezen. De volledige kruistabel zie je hiernaast.

		Festivalvoorkeur		
		W	P	D
Geslacht	J	1	3	0
	M	2	3	1

Als, in het algemeen, de eerste categorische veranderlijke "r" categorieën heeft en de tweede heeft er "c", dan krijg je een (**r**×**c**) **kruistabel**, met r rijen (**r**ows) en c kolommen (**c**olumns).

Opdracht 3

- In de vorige opdracht heb je 3 aantallen bepaald. In welke cellen staan die aantallen?
- In welke cel is de 9^{de} leerling terechtgekomen? Waarom?

Opdracht 4

Hiernaast zie je bivariate categorische gegevens van een studie zoals hierboven. De opmetingen zijn enerzijds voluit uitgeschreven en anderzijds samengevat in een kruistabel. De uitgeschreven opmetingen zijn niet volledig, maar je kan die aanvullen met wat je ziet in de kruistabel. Ook de kruistabel is niet volledig, maar die kan je vervolledigen met de uitgeschreven opmetingen. Doe dat nu.

Geslacht	Festivalvoorkeur
J	W
J	W
M	P
M	P
J	D
M	P
M	P
M	P
M	P

		Festivalvoorkeur		
		W	P	D
Geslacht	J		3	
	M	4		2

Opdracht 5

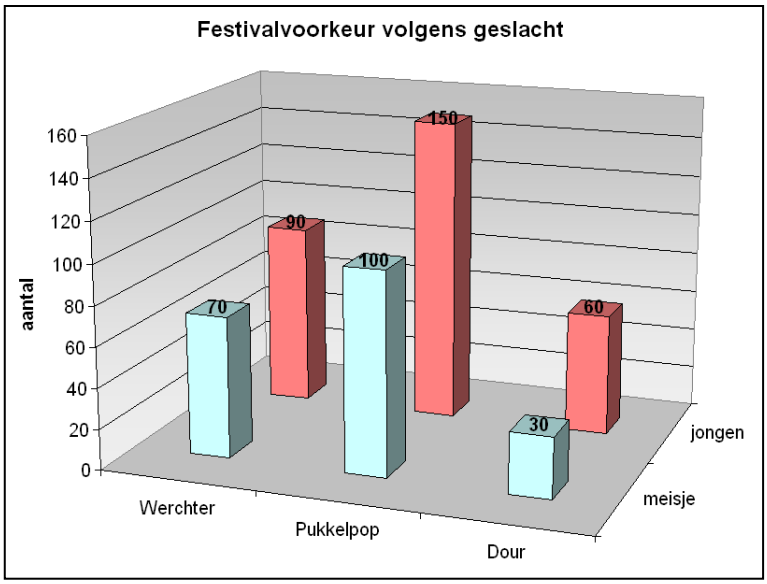
Een studie onderzoekt het rookgedrag van 200 tieners. Bij de rokers zijn er 26 jongens en 24 meisjes. Er zijn ook 54 jongens die niet roken. De rest zijn meisjes die niet roken. Stel deze informatie voor in een kruistabel. Gebruik de rijen voor het geslacht en de kolommen voor het rookgedrag. Werk met de afkortingen: J = jongen, M = meisje, R = roker, NR = niet-roker.

4. Informatie in een kruistabel

4.1. Bivariate informatie

Een studie over de samenhang tussen geslacht en festivalvoorkeur leverde bij 500 leerlingen het volgende resultaat:

		Festivalvoorkeur		
		W	P	D
Geslacht	J	90	150	60
	M	70	100	30



Een kruistabel bevat heel veel informatie. Je kan zo'n tabel beschouwen als een tweedimensionale frequentietabel. In elke cel staat het aantal (= de frequentie) leerlingen die in die cel zijn terechtgekomen. Om in een cel terecht te komen moet je aan twee voorwaarden tegelijkertijd voldoen. Zo zijn er 90 leerlingen in cel (1,1) terechtgekomen omdat er 90 jongens voor Werchter kiezen.

Sommige software kan een tweedimensionale frequentietabel grafisch voorstellen als een staafdiagram in 3D. Op het kruispunt van "meisje" en "Pukkelpop" staat een staafje van hoogte 100. In deze studie zijn er 100 meisjes met een voorkeur voor Pukkelpop.

Opdracht 6

- Hoe groot is het percent jongens met een voorkeur voor Dour in de bovenstaande studie? Verklaar je redenering.

- Hoe groot is het percent jongens in de bovenstaande studie? Verklaar je redenering.

4.2. Marginale informatie

Uit de gezamenlijke (bivariate) informatie over geslacht en festivalvoorkeur kan je informatie halen over elk van de veranderlijken afzonderlijk.

In de vorige opdracht heb je gebruik gemaakt van het feit dat er 300 jongens in de studie zitten. Hoe wist je dat? Dat getal staat toch nergens in die kruistabel?

Als je alleen het geslacht wil bestuderen, dan is dat heel eenvoudig. Voor elke waarde van de veranderlijke geslacht sommeer je over alle waarden van de veranderlijke festivalkeuze. Die som zet je in de rand (of in de marge) van de kruistabel. In het Engels zeg je: “in the margin” en daarom spreek je hier over **marginale** informatie.

Je kan de informatie over het geslacht samenvatten in een afzonderlijke frequentietabel. Je kan daarbij zowel de frequentie als de relatieve frequentie vermelden.

		Festivalvoorkeur			Rand-
		W	P	D	totaal
Geslacht	J	90	150	60	300
	M	70	100	30	200

Geslacht	Frequentie	Rel. freq.
J	300	60 %
M	200	40 %
Totaal	500	100 %

Opdracht 7

Gebruik de gegeven kruistabel om de veranderlijke festivalvoorkeur te bestuderen. Vul de kruistabel aan met randtotalen en maak een frequentietabel voor de festivalvoorkeur. Hoeveel percent van de leerlingen heeft een voorkeur voor Werchter?

		Festivalvoorkeur		
		W	P	D
Geslacht	J	90	150	60
	M	70	100	30
Rand-totaal				

Festival-voorkeur	Frequentie	Rel. frequ.
Totaal		

4.3. Conditionele informatie

Opdracht 8

Als je nu eens alleen naar de jongens kijkt, hoeveel percent heeft er dan een voorkeur voor Pukkelpop? Verklaar je redenering.

Bij kruistabellen kan je een voorwaarde opleggen aan een veranderlijke (bv. het geslacht). Je beperkt de studie dan tot één waarde van die veranderlijke (bv. jongens) en je kijkt welke informatie je nu hebt voor de waarden van de andere veranderlijke (festivalvoorkeur).

Je werkt nu "voorwaardelijk" of "**conditioneel**".

Als je alleen naar jongens kijkt, dan heb je genoeg aan de eerste rij van de kruistabel. Je hebt al gevonden dat de proportie jongens die voor Pukkelpop kiest, gelijk is aan 50 %. Op analoge manier vind je dat de proportie jongens die voor Werchter kiest, gelijk is aan 90/300 = 30 % en dat de proportie die voor Dour kiest 60/300 = 20 % is. Eigenlijk doe je niets anders dan de getallen op de eerste rij in de kruistabel delen door het rijtotaal. Het nieuwe rijtotaal is dan 100 % want je werkt "conditioneel op jongens".

	Festivalvoorkeur			
	W	P	D	Totaal
gegeven J	$\frac{90}{300} = 30\,\%$	$\frac{150}{300} = 50\,\%$	$\frac{60}{300} = 20\,\%$	$\frac{300}{300} = 100\,\%$

Opdracht 9

Zoek de conditionele proporties voor de festivalvoorkeur, als je alleen met de meisjes werkt.

	Festivalvoorkeur			
	W	P	D	Totaal

Je hebt nu conditioneel op elke waarde van het geslacht gewerkt. Dat betekent dat je elke rij als een afzonderlijke studie bekijkt. Je kan natuurlijk ook elke kolom als een afzonderlijke studie bekijken. Je werkt dan kolom per kolom, conditioneel op de waarden van de veranderlijke "festivalvoorkeur".

Opdracht 10

Verander de oorspronkelijke kruistabel in een tabel waar je, per festivalvoorkeur, de conditionele proportie meisjes en jongens invult. Uit die tabel kan je rechtstreeks aflezen hoe groot de proportie meisjes is bij de jongeren die een voorkeur hebben voor Pukkelpop. Hoe groot is die?

	gegeven W	gegeven P	gegeven D
Geslacht J			
Geslacht M			
Totaal			

4.4. Samenvatting

Wanneer je een huistaak maakt of studeert voor een toets, dan heb je misschien graag dat het rondom jou volledig stil is. Andere leerlingen vinden het fijn om dan naar hun favoriete muziek te luisteren op hun mp3-speler. Nog anderen zetten de radio aan, waar muziek afgewisseld wordt met interviews en spelletjes.

Op school krijg je allerlei vakken zoals Nederlands, Frans, Engels, geschiedenis, wiskunde, fysica, chemie, biologie, enz. Sommige leerlingen hebben een voorkeur voor alles wat met wetenschappen en wiskunde te maken heeft (we/wi). Bij andere leerlingen gaat de voorkeur helemaal in de andere richting (geen we/wi). Er zijn ook leerlingen bij wie het om het even is. Zij doen alle vakken ongeveer even graag.

Een navraag bij 64 leerlingen leverde het resultaat hiernaast.
"ID" is het identificatienummer van de leerling in deze studie.

ID	***Geluid***	***Voorkeur***
1	stilte	geen we/wi
2	mp3	om het even
3	radio	we/wi
4	mp3	geen we/wi
..	….	….

Opdracht 11

1. Welke veranderlijken worden in bovenstaand onderzoek bestudeerd? Geef hun naam en zeg ook welke soort veranderlijke het zijn. Wat zijn hun waarden?

2. De resultaten van de volledige studie zijn samengevat in de (3×3) kruistabel.

		Voorkeur we/wi	Voorkeur om het even	Voorkeur geen we/wi
Geluid	stilte	3	12	9
	mp3	4	16	12
	radio	1	4	3

- In welke cel staat het getal 16? Wat betekent het getal 16?

- In welke cel is de derde leerling (met ID = 3) terechtgekomen? Waarom?

- Hoeveel leerlingen luisteren naar muziek op hun mp3-speler? Hoe weet je dat?

3. Voeg aan de gegeven kruistabel alle randtotalen toe en maak een frequentietabel (met absolute en relatieve frequenties) voor de veranderlijke "voorkeur". Hoeveel percent van de onderzochte leerlingen heeft een voorkeur voor we/wi?

Voorkeur	Frequentie	Rel. frequ.
Totaal		

4. Zoek de conditionele proportie van de studievoorkeuren bij leerlingen die in stilte willen studeren. Gebruik daarvoor de onderstaande lay-out en toon je berekeningen.
   Hoeveel percent van die leerlingen (die in stilte studeren) verkiest andere vakken boven we/wi?

				Totaal

5. Onafhankelijkheid

5.1. Een voorbeeld

De getallen in de kruistabel over geluidsvoorkeur en studievoorkeur zijn niet het resultaat van een realistisch onderzoek. Zij zijn artificieel gekozen om te illustreren wat er bedoeld wordt als men zegt dat twee categorische veranderlijken onafhankelijk zijn van elkaar.

De marginale informatie voor de veranderlijke “geluid” zie je hiernaast.

Van de volledige groep onderzochte leerlingen werkt er 37.5 % in stilte, 50 % luistert naar hun mp3-speler en 12.5 % zet de radio aan.

Geluid	Frequentie	Rel. frequ.
stilte	24	37.5 %
mp3	32	50 %
radio	8	12.5 %
Totaal	64	100 %

Als “voorkeur voor geluid tijdens het studeren” helemaal niet samenhangt met studievoorkeur, dan verwacht je bij elke studievoorkeur dezelfde proporties voor de geluidsvoorkeur terug te vinden. Om dat na te gaan bereken je de conditionele proporties, per studievoorkeur.

		gegeven we/wi	gegeven om het even	gegeven geen we/wi
Geluid	stilte	$\frac{3}{8}=37.5\,\%$	$\frac{12}{32}=37.5\,\%$	$\frac{9}{24}=37.5\,\%$
	mp3	$\frac{4}{8}=50\,\%$	$\frac{16}{32}=50\,\%$	$\frac{12}{24}=50\,\%$
	radio	$\frac{1}{8}=12.5\,\%$	$\frac{4}{32}=12.5\,\%$	$\frac{3}{24}=12.5\,\%$
Totaal		$\frac{8}{8}=100\,\%$	$\frac{32}{32}=100\,\%$	$\frac{24}{24}=100\,\%$

Je merkt dat de geluidsvoorkeur op dezelfde manier verdeeld is over de drie geluidscategorieën, zowel bij de groep leerlingen die graag wetenschappen/wiskunde studeert, als bij de groep die dat niet zo graag doet, als bij wie het om het even is.

Je kan ook conditioneren op de verschillende soorten geluid. Bij onafhankelijkheid verwacht je dat de conditionele proporties van de studievoorkeuren niet veranderen als je van het ene geluid naar een ander overstapt.

In de vorige opdracht vond je dat 12.5 % van alle onderzochte leerlingen een voorkeur heeft voor we/wi, bij 50 % is het om het even en 37.5 % heeft liever andere vakken. Die proporties blijven dezelfde als je voorwaardelijk werkt, op de deelgroepen opgesplitst volgens geluidsvoorkeur. Dat zie je hieronder.

	Voorkeur			
	we/wi	om het even	geen we/wi	Totaal
gegeven stilte	$\frac{3}{24}=12.5\,\%$	$\frac{12}{24}=50\,\%$	$\frac{9}{24}=37.5\,\%$	$\frac{24}{24}=100\,\%$
gegeven mp3	$\frac{4}{32}=12.5\,\%$	$\frac{16}{32}=50\,\%$	$\frac{12}{32}=37.5\,\%$	$\frac{32}{32}=100\,\%$
gegeven radio	$\frac{1}{8}=12.5\,\%$	$\frac{4}{8}=50\,\%$	$\frac{3}{8}=37.5\,\%$	$\frac{8}{8}=100\,\%$

Je merkt dat de studievoorkeur op dezelfde manier verdeeld is over we/wi, om het even, geen we/wi, zowel bij de leerlingen die in stilte willen studeren als bij hen die luisteren naar hun mp3-speler of naar de radio.

Opdracht 12

Hebben jongens en meisjes dezelfde voorkeur voor festivals? Controleer of je bij de bestudeerde groep van 500 leerlingen kan zeggen dat festivalvoorkeur en geslacht onafhankelijk zijn.
In opdracht 7 heb je gevonden dat 32 % van die 500 leerlingen kiezen voor Werchter, 50 % voor Pukkelpop en 18 % voor Dour. Is dat ook zo voor de jongens en de meisjes afzonderlijk? Vul onderstaande tabel in met conditionele proporties en trek je besluit.

		Festivalvoorkeur		
		W	P	D
Geslacht	J	90	150	60
	M	70	100	30

	Festivalvoorkeur			
	W	P	D	Totaal
gegeven J				
gegeven M				

Opdracht 13

Een studie bij 5000 jongeren onderzoekt of er een verband is tussen hun sterrenbeeld en hun houding tegenover "body-art". Aan elk van de jongeren is gevraagd of zij zowel een piercing als een tatoeage hebben, ofwel alleen maar één van beide, ofwel geen van beide. Tegelijkertijd is ook hun sterrenbeeld genoteerd. Al die informatie staat samengevat in een (12×3) kruistabel, waarvan je hier alleen de randtotalen ziet. Vul de tabel aan zodat, voor deze studie, "body-art" en "sterrenbeeld" onafhankelijk zijn. Toon je redenering en je berekeningen.

		Body-art			
		zowel piercing als tatoeage	ofwel piercing ofwel tatoeage	geen van beide	Totaal
sterrenbeeld	Waterman				210
	Vissen				300
	Ram				560
	Stier				490
	Tweelingen				520
	Kreeft				480
	Leeuw				490
	Maagd				450
	Weegschaal				390
	Schorpioen				400
	Boogschutter				450
	Steenbok				260
Totaal		500	1500	3000	5000

5.2. Structuur van een kruistabel bij onafhankelijkheid

In het voorbeeld over geluidsvoorkeur en studievoorkeur kon je op 2 manieren redeneren om aan te tonen dat die twee veranderlijken onafhankelijk zijn:

- als "voorkeur voor geluid tijdens het studeren" helemaal niet samenhangt met "studievoorkeur", dan verwacht je dat de conditionele proportie van de geluidsvoorkeuren niet verandert als je van de ene studievoorkeur overstapt naar een andere

- als "voorkeur voor geluid tijdens het studeren" helemaal niet samenhangt met "studievoorkeur", dan verwacht je dat de conditionele proportie van de studievoorkeuren niet verandert als je van het ene geluid overstapt naar een ander.

Je kan nu één van beide redeneringen in detail bekijken (de andere werkt volledig analoog).

Van de volledige groep leerlingen werkt er 37.5 % in stilte, 50 % luistert naar hun mp3-speler en 12.5 % zet de radio aan. Dat heb je vroeger gevonden. Die proporties moeten nu dezelfde zijn bij elke studievoorkeur. Dat betekent bijvoorbeeld dat bij de 8 leerlingen die een voorkeur voor we/wi hebben, er 37.5 % in stilte werkt, 50 % naar hun mp3-speler luistert en 12.5 % de radio aanzet. Het aantal leerlingen dat in cel (1,1) terechtkomt moet dus $(37.5\ \%) \times 8 = 3$ zijn. Bemerk dat 37.5 % niets anders is dan het totaal van de eerste rij (24) gedeeld door het algemeen totaal van de tabel (64). Bovendien is het totaal van de eerste kolom gelijk aan 8. Alles samen heb je:

cel (1,1) = (rijtotaal 1^{ste} rij) × (kolomtotaal 1^{ste} kolom) / (tabeltotaal).

		Voorkeur we/wi	Voorkeur om het even	Voorkeur geen we/wi	Totaal
Geluid	stilte	$24 \times 8/64 = 3$			24
	mp3				
	radio				
Totaal		8			64

Een kruistabel waarbij deze eigenschap geldt voor elke cel weerspiegelt onafhankelijkheid.

<u>Modeleigenschap</u>

de categorische veranderlijken zijn onafhankelijk

⇕

cel (i,j) = (rijtotaal i^{de} rij) × (kolomtotaal j^{de} kolom) / (tabeltotaal)
voor **elke** cel (i,j) van de kruistabel

Opdracht 14

Hoe zou de perfecte kruistabel eruit zien als er helemaal geen samenhang is tussen “festivalvoorkeur” en “geslacht”? Maak gebruik van de modeleigenschap voor onafhankelijkheid bij kruistabellen.

		Festivalvoorkeur			Totaal
		W	P	D	
Geslacht	J				300
	M				200
Totaal		160	250	90	500

6. De paradox van Simpson

Kruistabellen kunnen een aanduiding geven dat twee categorische veranderlijken onafhankelijk zijn. Zij kunnen ook wijzen op een samenhang in een of andere richting. In bepaalde studiegebieden bijvoorbeeld kunnen meisjes betere resultaten halen dan jongens. Als je meerdere van dergelijke tabellen samenvoegt, dan kan het gebeuren dat die samenhang verdwijnt of zelfs van richting verandert. De reden hiervoor kan een onderliggende verdoken veranderlijke zijn. Het is niet altijd eenvoudig om die op het spoor te komen.

6.1. Domme meisjes

Uit alle eerstejaarsstudenten aan de Vlaamse universiteiten werden lukraak 1000 jongens en 1000 meisjes geselecteerd en men noteerde of zij in de eerste examenzittijd geslaagd waren. Het resultaat van dit onderzoek is samengevat in de volgende tabel (J = jongen, M = meisje).

		Geslaagd		Totaal
		Ja	Neen	
Geslacht	J	502	498	1000
	M	464	536	1000
Totaal		966	1034	2000

De marginale informatie zegt dat er in totaal $\frac{966}{2000} = 0.483 = 48.3\,\%$ van de studenten geslaagd is. Maar hoe zit het met de resultaten van meisjes en jongens afzonderlijk? Dat haal je uit de conditionele informatie, gegeven het geslacht:

	Geslaagd		Totaal
	Ja	Neen	
gegeven J	$\frac{502}{1000} = 50.2\,\%$	$\frac{498}{1000} = 49.8\,\%$	$\frac{1000}{1000} = 100\,\%$
gegeven M	$\frac{464}{1000} = 46.4\,\%$	$\frac{536}{1000} = 53.6\,\%$	$\frac{1000}{1000} = 100\,\%$

Bij de jongens is 50.2 percent geslaagd en bij de meisjes is dat maar 46.4 percent. Het gaat hier over grote groepen en dus toont deze studie aan dat er tussen meisjes en jongens een beduidend verschil is in studieresultaten.

Besluit: Meisjes zijn dommer dan jongens.

Maar is dat zo …. ?

Aan een universiteit kan je heel veel verschillende dingen studeren, maar alle studiegebieden kan je samenvatten in drie grote groepen:

- De groep der "exacte" wetenschappen, zoals wiskunde, chemie, burgerlijk ingenieur, …
- De groep der "humane" wetenschappen, zoals psychologie, economie, talen, rechten, …
- De groep der "medische" wetenschappen, zoals geneeskunde, biomedische, farmacie, …

Ga nu op zoek in welke groep die domme meisjes zitten. Dat is niet zo moeilijk als je weet dat bij dit onderzoek niet alleen aan de studenten gevraagd is of zij geslaagd zijn, maar ook wat zij in dat eerste jaar studeerden. De resultaten per studiegroep zien er als volgt uit. De tabellen tonen niet alleen de aantallen, maar geven ook (tussen vierkante haakjes) de conditionele slaagpercentages per geslacht. Bemerk dat het hier gaat over **dezelfde studenten** die je in de bovenstaande kruistabel hebt ontmoet.

Groep der "exacte" wetenschappen							
		Geslaagd Ja		Neen		Totaal	
Geslacht	J	215	[59.7 %]	145	[40.3 %]	360	[100 %]
	M	25	[62.5 %]	15	[37.5 %]	40	[100 %]

De domme meisjes zijn niet te vinden in de groep der exacte wetenschappen. Zij zijn daar slimmer dan de jongens want 62.5 percent van de meisjes is geslaagd en slechts 59.7 percent van de jongens.

Groep der "humane" wetenschappen							
		Geslaagd Ja		Neen		Totaal	
Geslacht	J	91	[37.9 %]	149	[62.1 %]	240	[100 %]
	M	205	[41 %]	295	[59 %]	500	[100 %]

De domme meisjes zijn ook niet te vinden in de groep der humane wetenschappen. Zij zijn daar de slimste, want 41 percent van de meisjes is geslaagd en slechts 37.9 percent van de jongens.

Groep der "medische" wetenschappen							
		Geslaagd Ja		Neen		Totaal	
Geslacht	J	196	[49 %]	204	[51 %]	400	[100 %]
	M	234	[50.9 %]	226	[49.1 %]	460	[100 %]

In de groep der medische wetenschappen zijn de meisjes ook slimmer dan de jongens, want 50.9 percent van de meisjes is geslaagd tegenover 49 percent van de jongens.

Besluit: Meisjes zijn slimmer dan jongens want zij zijn overal beter, in elk van de 3 studiegroepen.

Vraag: Op basis van **dezelfde steekproef van studenten** besluit je: "meisjes zijn dommer dan jongens" maar ook "meisjes zijn slimmer dan jongens". Wat denk je nu over de uitspraak "***cijfers liegen niet***" (er is niet geknoeid met de cijfers, ze zijn eerlijk opgemeten)?

Opdracht 15

Om de paradox over de domme meisjes te ontrafelen werk je in deze opdracht in de veronderstelling dat “slagen” onafhankelijk is van “geslacht”: meisjes zijn even slim als jongens.

1. Hoeveel percent studenten is geslaagd bij de exacte wetenschappen? Gebruik deze informatie om onderstaande kruistabel in te vullen waarbij je ervoor zorgt dat slagen onafhankelijk is van het geslacht (meisjes zijn even slim als jongens).

Groep der “exacte” wetenschappen				
		Geslaagd Ja	Neen	Totaal
Geslacht	J			360
	M			40

2. Hoeveel percent studenten is geslaagd bij de humane wetenschappen? Gebruik deze informatie om onderstaande kruistabel in te vullen waarbij je ervoor zorgt dat slagen onafhankelijk is van het geslacht (meisjes zijn even slim als jongens).

Groep der “humane” wetenschappen				
		Geslaagd Ja	Neen	Totaal
Geslacht	J			240
	M			500

3. Hoeveel percent studenten is geslaagd bij de medische wetenschappen? Gebruik deze informatie om onderstaande kruistabel in te vullen waarbij je ervoor zorgt dat slagen onafhankelijk is van het geslacht (meisjes zijn even slim als jongens).

Groep der “medische” wetenschappen				
		Geslaagd Ja	Neen	Totaal
Geslacht	J			400
	M			460

4. Tel nu alles samen en vul de onderstaande kruistabel in. Wat leert die over het percent geslaagde jongens in vergelijking met het percent geslaagde meisjes?

		Geslaagd Ja	Neen	Totaal
Geslacht	J			1000
	M			1000
Totaal				2000

5. Kan je de paradox van Simpson verklaren bij dit voorbeeld? Welke verdoken veranderlijke zie je niet in de tabel van punt 4 hierboven? Welke rol speelt die?

6.2. Een gekleurde rechtspraak

In de jaren 1976-77 stonden in Florida 326 beklaagden terecht wegens moord. Sommigen kregen de doodstraf, anderen niet.

Politieke activisten klaagden het gerecht aan, omdat het de zwarten discrimineerde. De zwarten werden meer tot de doodstraf veroordeeld dan de blanken. Om hun actie te steunen gebruikten de activisten de officiële gegevens van de rechtbanken zelf.

Ras van de dader	Ras van het slachtoffer	Doodstraf Ja	Neen
Blank	Blank	19	132
	Zwart	0	9
Zwart	Blank	11	52
	Zwart	6	97

Opdracht 16

Herschrijf de gegevens van Florida in twee afzonderlijke kruistabellen, één voor de moorden waarbij het slachtoffer blank was en één waarbij het slachtoffer zwart was. Schrijf naast de aantallen ook (tussen vierkante haakjes) de conditionele proportie uitgesproken doodstraffen per ras van de dader.

Het slachtoffer is blank				
		Doodstraf Ja	Neen	Totaal
Ras van de dader	Blank			
	Zwart			

Het slachtoffer is zwart				
		Doodstraf Ja	Neen	Totaal
Ras van de dader	Blank			
	Zwart			

De politieke activisten argumenteerden als volgt (*vul in*):

- van de blanken die een blanke hebben vermoord, kreeg …..…% de doodstraf, maar van de zwarten die een blanke hebben vermoord, is …….. % ter dood veroordeeld
- van de blanken die een zwarte hebben vermoord, kreeg ……. de doodstraf, maar bij de zwarten die een zwarte hebben vermoord, is …….. % ter dood veroordeeld.

Dat kan toch allemaal geen toeval zijn. De zwarten kunnen niet rekenen op een eerlijk proces in Florida. De cijfers zijn toch overduidelijk!

Opdracht 17

Start met de gegevens van Florida en gebruik die om één kruistabel te maken waaruit blijkt dat niet tegen de zwarten maar tegen de blanken de grootste proportie doodstraffen is uitgesproken. Kan je ook een verklaring geven voor deze paradox?

6.3. De andere kant van het gelijk

Een bedrijf heeft 200 werknemers, 100 arbeiders en 100 bedienden. Op het einde van het jaar wordt er een extra premie toegekend, maar alleen aan die werknemers van wie de manager vindt dat zij zich het voorbije jaar bijzonder goed hebben ingezet.

Op een TV show verklaart de manager fier dat hij in zijn bedrijf een vrouwvriendelijke politiek voert. Hij beweert dat het percent vrouwen dat van hem zo'n extra premie heeft gekregen groter is dan het percent mannen.

In diezelfde show verklaart een vrouwelijke vakbondsafgevaardigde dat de zogenaamde vrouwvriendelijke politiek van de manager een flagrante leugen is. Sterker nog, zij beweert over cijfers te beschikken die aantonen dat, zowel bij de arbeiders als bij de bedienden, het percent mannen dat een extra premie heeft gekregen groter is dan het percent vrouwen.

Als je denkt dat zowel de manager als de vakbondsafgevaardigde cijfermateriaal kan tonen om hun standpunt te verdedigen, bedenk dan een concrete situatie die beide uitspraken ondersteunt. Stel hiervoor ook de nodige kruistabellen op, bespreek die, en verklaar de paradox.

DEEL 2. Uitbreiding

7. Samenhang

Hoe de structuur van een kruistabel eruit ziet als je te maken hebt met onafhankelijke categorische veranderlijken heb je hierboven geleerd. Maar wat gebeurt er als ze niet onafhankelijk zijn?

De studie van samenhang bij categorische veranderlijken is zeer uitgebreid (in het programma van "Master in statistiek" aan de universiteit heb je zelfs een afzonderlijk vak: "Analyse van categorische data").

In deze tekst beperken we ons tot twee aspecten:

- samenhang in een algemene $(r \times c)$ kruistabel
- samenhang in een (2×2) kruistabel

7.1. De (r× c) kruistabel

7.1.1. Geobserveerd versus verwacht

Een Amerikaanse studie onderzocht bij 17-jarigen het verband tussen het globale geluksgevoel in het voorbije jaar (ongelukkig, gelukkig, zeer gelukkig) en het aantal verschillende sekspartners in dat jaar (0, 1, 2+).

Opdracht 18

In de onderstaande kruistabel staan alleen de randtotalen. Vervolledig de tabel zodat hij een perfecte onafhankelijkheid tussen "sekspartners" en "geluksgevoel" weerspiegelt. Gebruik de modeleigenschap voor onafhankelijkheid bij kruistabellen.

		Geluksgevoel			Totaal
		ongelukkig	gelukkig	zeer gelukkig	
Sekspartners	0				420
	1				340
	2+				240
Totaal		180	540	280	1000

De resultaten van de studie leverden de volgende kruistabel:

		Geluksgevoel			Totaal
		ongelukkig	gelukkig	zeer gelukkig	
Sekspartners	0	70	235	115	420
	1	50	222	68	340
	2+	30	143	67	240
Totaal		150	600	250	1000

Je ziet onmiddellijk dat de opgemeten waarden geen perfecte onafhankelijkheid weerspiegelen. Met die vaststelling kan je in de statistiek nog geen conclusie formuleren.

Inderdaad, op dit ogenblik ga je het kader waarin je werkt verruimen. Je kijkt niet alleen naar de opmetingen van een bepaalde studie, maar je denkt ook aan eigenschappen van een onderliggende populatie.

Veronderstel eens dat bij de volledige populatie van alle 17-jarige Amerikanen (dat zijn er meer dan een miljoen) het "aantal sekspartners vorig jaar" en "het globale geluksgevoel vorig jaar" onafhankelijk zijn van elkaar. Denk je dan dat een lukrake steekproef van 1000 Amerikanen een kruistabel oplevert die perfecte onafhankelijkheid weerspiegelt? Waarschijnlijk niet. Twee veranderlijken die **in de populatie** onafhankelijk zijn, leveren nog geen perfecte onafhankelijkheid in de kruistabel van je **steekproef**.

- Je hebt zopas een (ideale) kruistabel opgesteld die je **verwacht** te zien bij onafhankelijkheid.
- Je hebt ook een **geobserveerde** kruistabel, dat is de tabel die bij de studie is opgemeten.

Een eerste belangrijke vraag gaat over het verschil tussen wat je ziet en wat je had verwacht.

1. Wijkt de **geobserveerde** kruistabel niet te veel af van de **verwachte** kruistabel? In dat geval kan de afwijking toegeschreven worden aan het toeval van de steekproef. Je kan dan bij de veronderstelling blijven dat er *onafhankelijkheid is in de totale populatie*.

2. Is de afwijking tussen de **geobserveerde** en de **verwachte** kruistabel groot? In dat geval geloof je niet meer dat je steekproef komt uit een populatie waar er onafhankelijkheid is.

Een tweede belangrijke vraag is: hoe meet je **het verschil tussen tabellen**?

Geobserveerde aantallen

		Geluksgevoel		
		ongelukkig	gelukkig	zeer gelukkig
Sekspartners	0	70	235	115
	1	50	222	68
	2+	30	143	67

Verwachte aantallen

		Geluksgevoel		
		ongelukkig	gelukkig	zeer gelukkig
Sekspartners	0	63	252	105
	1	51	204	85
	2+	36	144	60

Om het verschil tussen kruistabellen te bepalen start je per cel.

Als je zou werken met: (geobserveerd aantal) – (verwacht aantal) dan levert dat voor cel (1,1) de waarde 70 – 63 = 7 en voor cel (1,2): 235 – 252 = – 17. Je krijgt positieve en negatieve verschillen en als je die allemaal samentelt krijg je 0, voor elke geobserveerde tabel. Dat helpt dus niet.

Eigenlijk wil je weten of een geobserveerd aantal *niet te ver afwijkt* van wat je verwacht. Je zou dus met "afstand" moeten werken wat tot "absolute waarde" leidt. Voor veel wiskundige bewerkingen is de "absolute waarde" onhandig en dus stap je over op het kwadraat. Dat is altijd positief en een groter verschil geeft ook een grotere kwadratische bijdrage dan een kleiner verschil. Zo ben je aangeland bij:

$$[(\text{geobserveerd aantal}) - (\text{verwacht aantal})]^2.$$

Ten slotte wordt nog een correctie doorgevoerd.

Als je 3 verwacht en je ziet 6 dan is dat een serieuze afwijking. Het kwadratisch verschil is $(6\text{-}3)^2 = 9$. Als je 300 verwacht en je ziet 303 dan is dat maar een kleine afwijking. Maar ook hier is het kwadratisch verschil gelijk aan $(303\text{-}300)^2 = 9$.

Om met de grootteorde van wat je verwacht rekening te houden, werk je proportioneel:

$$\frac{\left[\left(\textit{geobserveerd aantal}\right)-\left(\textit{verwacht aantal}\right)\right]^2}{\left(\textit{verwacht aantal}\right)}$$

Als je 3 verwacht en je ziet 6 dan levert de formule $(6\text{-}3)^2 / 3 = 3$.
Als je 300 verwacht en je ziet 303 dan levert de formule $(303\text{-}300)^2 / 300 = 0.03$.

7.1.2. De chi-kwadraat statistiek

Om het verschil tussen twee kruistabellen in één getal te karakteriseren, bereken je voor elke cel de chi-kwadraat bijdrage $\frac{\left[\left(geobserveerd\ aantal\right)-\left(verwacht\ aantal\right)\right]^2}{\left(verwacht\ aantal\right)}$ en dan maak je de som over alle cellen. Het getal dat je zo bekomt is een waarde van wat de chi-kwadraat statistiek genoemd wordt.

Opdracht 19

Bereken de waarde van de chi-kwadraat statistiek voor de studie bij die 1000 Amerikaanse jongeren. Schrijf in onderstaande tabel de chi-kwadraat bijdrage in elke cel en maak dan de som van al die bijdragen (werk tot op 2 decimalen nauwkeurig). Nota: χ is de Griekse letter "chi".

Chi-kwadraat bijdragen				
		Geluksgevoel		
		ongelukkig	gelukkig	zeer gelukkig
Sekspartners	0			
	1			
	2+			

De chi-kwadraat waarde is : $\chi^2 = ...$

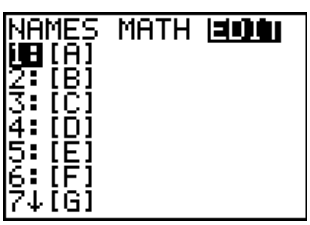
Om te weten of de gevonden chi-kwadraat waarde wijst op een klein, matig of groot verschil tussen de kruistabellen moet je een studie maken van het gedrag van de chi-kwadraat statistiek. Dat valt buiten de leerstof van het secundair onderwijs maar je begrijpt wel dat grotere chi-kwadraat waarden wijzen in de richting van afhankelijke veranderlijken in de populatie.

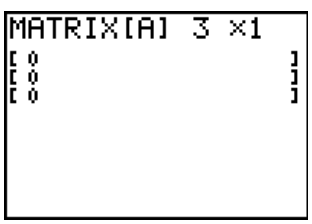
Door stap voor stap de berekeningen te doorlopen, heb je geleerd hoe de chi-kwadraat waarde een maatstaf is om het verschil tussen kruistabellen te bepalen. Nu je weet hoe dat werkt, is het handig om die berekeningen niet telkens met de hand uit te voeren. Je GRM kan je hierbij helpen.

Geobserveerde kruistabel, verwachte kruistabel en chi-kwadraat waarde met de GRM

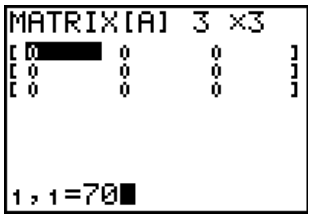
Het enige wat je moet doen is de geobserveerde kruistabel (de opmetingen uit je studie) inbrengen in de GRM. Daarvoor doorloop je de volgende stappen.

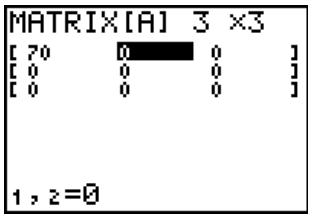
Druk [2nd][MATRIX] en loop met [▶] naar EDIT. Zorg ervoor dat je op 1:[A] staat en druk dan [ENTER]. Je moet nu eerst zeggen hoeveel rijen en hoeveel kolommen de matrix [A] heeft. Tik 3 voor de rijen, loop dan tot je achter het ×teken staat en tik dan terug 3.



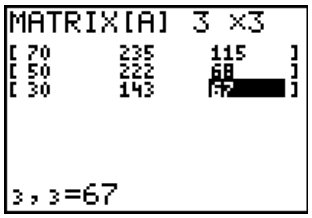


Als je nu op [ENTER] drukt verschijnt er een (3×3) matrix waar in elke cel een nul staat. Om te beginnen sta je in cel (1,1). Tik nu 70 en [ENTER]. Het getal 70 staat nu in cel (1,1) en de cursor is ondertussen versprongen naar cel (1,2).

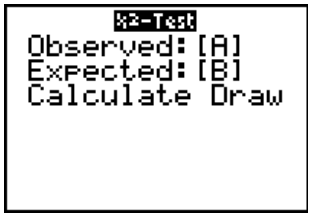




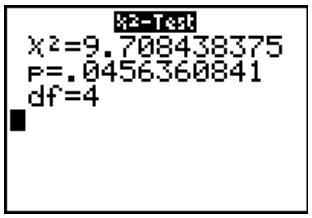
Tik nu 235 en [ENTER]. Ga zo verder tot je de matrix volledig hebt opgevuld met de **geobserveerde** aantallen. Druk dan [2nd][QUIT].



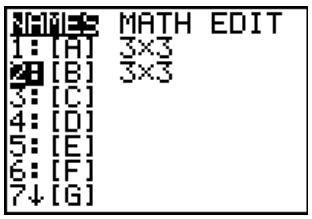
Druk [STAT], loop naar TESTS en loop dan naar beneden tot de cursor naast χ^2-Test... staat. Druk dan [ENTER].

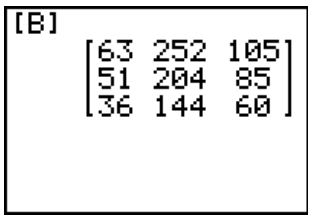


In het χ^2-Test scherm moet je invullen in welke matrix de geobserveerde (**Observed:**) kruistabel staat. Aangezien jij die in de matrix [A] hebt gezet moet je hier niets veranderen. Op de volgende regel kan je aangeven in welke matrix de verwachte (**Expected:**) kruistabel (verwacht in de veronderstelling van perfecte onafhankelijkheid) moet terechtkomen. Ook hier kan je akkoord gaan met de voorgestelde matrix [B] en hoef je niets te veranderen. Loop nu naar **Calculate** en druk [ENTER]. De gevonden χ^2 waarde is 9.71. Met de hand vond jij dat $\chi^2=9.72$. Het kleine verschil heeft te maken met afronding.



De kruistabel met de verwachte waarden onder onafhankelijkheid zie je als volgt. Druk [2nd][MATRIX], loop naar [B] en druk twee keer [ENTER]. Bemerk dat deze kruistabel exact overeenkomt met de kruistabel die jij in opdracht 18 hebt berekend. Je GRM gebruikt, net als jij, de randtotalen van tabel A om tabel B te berekenen.





7.2. De (2×2) kruistabel

7.2.1. Verschil in proporties

Een studie bij 200 tieners onderzoekt of het rookgedrag van jongens verschilt van dat van meisjes.

De gevonden resultaten zijn samengevat in de kruistabel met afkortingen: J = jongen, M = meisje, R = roker, NR = niet-roker.

		Rookgedrag		Totaal
		R	NR	
Geslacht	J	26	54	80
	M	24	96	120
Totaal		50	150	200

Een (2×2) tabel gebruik je dikwijls om twee groepen (jongens en meisjes) te bestuderen op een karakteristiek (rookgedrag) die maar twee uitkomsten heeft (R en NR).

Je kan nu het rookgedrag bestuderen, conditioneel op het geslacht. En aangezien het rookgedrag maar twee uitkomsten heeft, heb je, per geslacht, genoeg aan het percent rokers (want dan ken je ook het percent niet-rokers).

Opdracht 20

Hoeveel percent rokers bevat de bovenstaande studie? Als je veronderstelt dat rookgedrag en geslacht onafhankelijk zijn, hoeveel percent rokers verwacht je dan bij de jongens? En bij de meisjes?
Gebruik je antwoord om een kruistabel op te stellen die je verwacht wanneer het rookgedrag helemaal niets te maken heeft met het geslacht. Toon je berekeningen.

		Rookgedrag		Totaal
		R	NR	
Geslacht	J			80
	M			120
Totaal		50	150	200

Bijkomende vraag.
Als je de verwachte kruistabel opstelt volgens de modeleigenschap voor onafhankelijkheid bij algemene (r×c) kruistabellen, krijg je dan hetzelfde resultaat?

Bij onafhankelijkheid verwacht je dezelfde proportie rokers bij jongens en bij meisjes.

In de uitgevoerde studie zijn er 26/80 = 32.5 % jongens die roken en 20 % meisjes. Dat verschil in proportie is redelijk groot en kan erop wijzen dat, in de totale populatie, het rookgedrag afhankelijk is van het geslacht. Om dit na te gaan heb je methoden van de verklarende statistiek nodig.

Bij (2×2) kruistabellen kan je **het verschil in proporties** gebruiken om een idee te krijgen over *de sterkte van de samenhang* tussen de twee categorische veranderlijken.

7.2.2. Relatief risico

Het verschil in proporties is niet altijd de beste maatstaf om de samenhang te beschrijven. Dat zie je in volgend voorbeeld.

In een experimentele fase worden geneesmiddelen uitgetest op proefdieren, onder meer om de schadelijke neveneffecten te onderzoeken. Als de proportie proefdieren die neveneffecten vertoont 0.49 is voor het ene geneesmiddel en 0.48 voor het andere, dan lijkt dat goed in elkaars buurt te liggen. Maar wanneer deze geneesmiddelen uiteindelijk op de markt gebracht worden en de neveneffecten bij mensen treden op met een proportie van 0.011 in het ene geval en 0.001 in het andere, dan lijkt dit wel belangrijk. Op 1000 mensen krijgen 11 mensen nevenverschijnselen in het ene geval en slechts 1 persoon in het andere geval.

Als je het verschil in proporties uitrekent, dan is dat telkens gelijk aan 0.01, zowel bij de proefdieren als bij de mensen. Het kan dus goed zijn om ook andere maatstaven te hanteren om de sterkte van de samenhang te bestuderen.

Voorbeeld

De manager van een kruideniersbedrijf merkt dat meerdere van zijn werknemers huiduitslag krijgen. Het zijn blijkbaar vooral werknemers die in contact komen met selder. Om dit verder te onderzoeken worden gedurende enige tijd alle 140 werknemers van het bedrijf opgevolgd en men noteert of zij huiduitslag krijgen en of zij in contact komen met selder. De resultaten zijn samengevat in de kruistabel.

		Huiduitslag		Totaal
		Ja	Neen	
Contact met selder	Ja	24	36	60
	Neen	4	76	80
Totaal		28	112	140

Van de 60 werknemers die in contact komen met selder krijgen er 24/60 = 40 % huiduitslag. Bij de werknemers die niet in contact komen met selder krijgen er slechts 4/80 = 5 % huiduitslag. Het grote verschil in proporties wijst op een sterke samenhang tussen “in contact komen met selder” en “huiduitslag krijgen”.

Je kan deze studie ook op een andere manier bekijken en een andere maat voor samenhang gebruiken: **het relatief risico**. Hierbij vergelijk je het risico dat je loopt als je wel met selder in contact komt, met het risico dat je loopt als je niet met selder in contact komt. Concreet bereken je de verhouding van twee conditionele proporties:

$$relatief\ risico = \frac{proportie\ aandoeningen\ in\ de\ blootgestelde\ groep}{proportie\ aandoeningen\ in\ de\ niet\ blootgestelde\ groep}$$

Bij de werknemers van het kruideniersbedrijf is het relatief risico voor huiduitslag ten gevolge van contact met selder gelijk aan $\frac{40\%}{5\%} = 8$. Dit betekent dat, in deze studie, de proportie werknemers die huiduitslag krijgt 8 keer zo groot is bij hen die in contact komen met selder in vergelijking met de groep die niet met selder in contact komt. Als je deze studie mag veralgemenen dan kan je zeggen dat het risico om huiduitslag te krijgen 8 keer groter is als je met selder in contact komt dan wanneer dat niet het geval is.

In de definitie van het relatief risico staat een verhouding van twee proporties. De uitkomst kan dus gelijk welk positief getal zijn. Als het relatief risico groter is dan één, dan is het risico groter in de blootgestelde groep dan in de andere. Bij een relatief risico kleiner dan één is het juist andersom.

Opdracht 21

Een relatief risico dat gelijk is aan één wijst erop dat er geen samenhang is tussen de twee categorische veranderlijken. Verklaar waarom dat zo is. Vul de kruistabel in zodat het relatief risico gelijk is aan 1.

		Huiduitslag		Totaal
		Ja	Neen	
Contact met selder	Ja			60
	Neen			80
Totaal		28	112	140

Opdracht 22

Is er een verband tussen het houden van een kat als huisdier en het lijden aan een allergie voor huisstofmijt? Bereken het relatief risico op een allergie voor huisstofmijt en verklaar je resultaat ook in woorden. Gebruik de kruistabel van een studie bij 300 personen.

		Allergie voor huisstofmijt		Totaal
		Ja	Neen	
Kat als huisdier	Ja	72	48	120
	Neen	36	144	180
Totaal		108	192	300