12 Kruistabellen

In dit hoofdstuk bespreken we frequentieverdelingen van twee of meer variabelen. Ook deze kunnen worden weergegeven via (kruis)tabellen en grafische voorstellingen. We gaan uitvoerig in op het begrip 'statistisch verband', wat het basisprincipe is en hoe de sterkte van een verband wordt vastgesteld. Hier komen maten als de epsilon, de odds ratio en de correlatiecoëfficiënt aan de orde. In het laatste gedeelte van dit hoofdstuk bespreken we het voorspellen van de score van een eenheid op een gevolgvariabele uitgaande van zijn score op een oorzaakvariabele, met andere woorden: bivariate regressie.

Wat is een kruistabel?

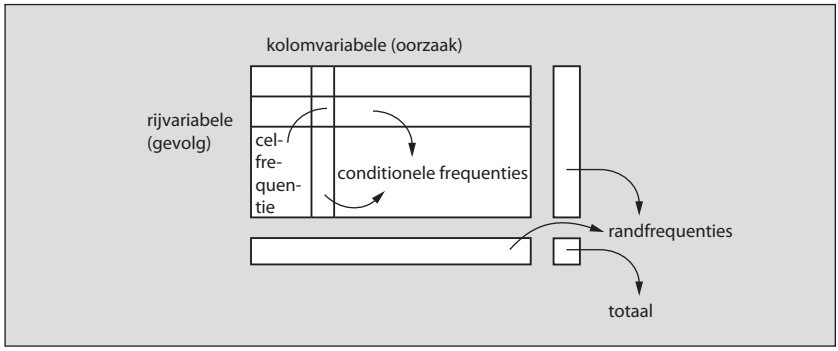
Een kruistabel is een staatje van alle logisch mogelijke combinaties van de waarden van twee (of eventueel meer) variabelen, waarbij voor elke combinatie de frequentie is aangegeven. In SPSS krijgen we, via het aanklikken van achtereenvolgens ANALYZE, DESCRIPTIVE STATISTICS en CROSSTABS, een scherm te zien waarop we een variabele als rijelement kunnen invoeren, en een andere variabele als kolomelement. Kiezen we in het GSS91-bestand voor 'exciting' als rijvariabele en voor 'sex' als kolomvariabele, dan krijgen we ongeveer tabel 12.1a te zien.

Tabel 12.1 Waardering van het eigen leven naar sekse

a. (absoluut)				b. (in %)			
	Man	**Vrouw**	**Totaal**		**Man**	**Vrouw**	**Totaal**
Opwindend	213	221	434	Opwindend	50,1	39,8	44,3
Tamelijk gewoon	200	305	505	Tamelijk gewoon	47,1	55,0	51,5
Saai	12	29	41	Saai	2,8	5,2	4,2
Totaal	425	555	980	Totaal	100,0	100,0	100,0
				N	425	555	980

Gemakkelijker in het gebruik – we willen immers weten of er verschillen zijn tussen mannen en vrouwen – is tabel 12.1b, waarin de totalen voor de mannen respectievelijk de vrouwen op 100 zijn gesteld, en de celvullingen in percentages zijn uitgedrukt. In het CROSSTABS-venster klikken we onderaan de optie CELLS aan. In het dan verschijnende venster CELL DISPLAY vinken we COLUMN PERCENTAGES aan. Je krijgt dan een tabel met zowel de absolute aantallen als de kolompercentages. Neem zo'n tabel liever niet in een rapport op, maar onderdruk een van beide, in dit geval de absolute aantallen. We gaan terug naar het voorgaande, en, nadat we weer COLUMN PERCENTAGES hebben aangeklikt, klikken we de OBSERVED COUNTS weg. Je krijgt dan de tabel met alleen de percentages. In de onderste 'totaalrij' hebben we – anders dan SPSS dat doet – de absolute totalen vermeld. Dat is handig om eventueel alle absolute aantallen te kunnen terugberekenen.

Bij een kruistabel noemen we de frequentieverdelingen van elke variabele afzonderlijk, en die rechts en onder buiten de eigenlijke tabel staan afgedrukt de *randverdelingen*. Ze worden ook wel de randfrequenties (*marginals*) genoemd. De frequentieverdeling van een variabele binnen een afzonderlijke waarde (of conditie) van de andere variabele noemen we een *conditionele frequentieverdeling* (Engels: *conditional*). *Elke* rij en *elke* kolom vormen dus een conditionele verdeling. De vullingen van de cellen noemen we de celvullingen of *celfrequenties*.



Figuur 12.1 De structuur van een kruistabel

Een kruistabel over twee dichotome variabelen (variabelen met maar twee waarden) is de eenvoudigste kruistabel die er bestaat. Tabel 12.1 heeft betrekking op een dichotome variabele (sekse) en een variabele met drie waarden, en is dus iets ingewikkelder. We duiden het formaat van een tabel aan met r x k, waarin r = het aantal rijen, en k = het aantal kolommen. Tabel 12.1 is dus een 3x2-tabel.

Voor de kop van een kruistabel gelden dezelfde spelregels als voor de frequentietabel bij één variabele (zie hoofdstuk 10). Omdat we hier te maken hebben met twee of meer variabelen, voegen we nog een spelregel toe die te maken heeft met de volgorde van de variabelen. Voor die volgorde geldt: noem de (rij)variabele (dat is de variabele die *links* van de tabel is vermeld) het *eerst*, de (kolom)variabele (dat is de variabele die *boven* de tabel is vermeld) het *laatst*. Maar welke variabele wordt nu in een kruistabel als rijvariabele opgenomen, en welke als kolomvariabele? Immers, als we 'sex' als rijvariabele en 'life' als kolomvariabele hadden opgegeven, hadden we een tabel gekregen die '90° gekanteld' is, maar die precies dezelfde informatie geeft. Hebben we een voorkeur voor de ene boven de andere weergave? Ja! Als de variabelen kunnen worden onderscheiden naar oorzaak en gevolg, komt de *gevolg*variabele *links* en de *oorzaak*variabele *boven* de tabel.

De reden hiervoor is dat we in de percentagetabellen altijd percenteren per categorie van de oorzaakvariabele, dus, in het voorbeeld, apart voor de mannen en apart voor de vrouwen. Percentages nu laten zich veel gemakkelijker controleren door optellen tot 100 *als ze onder elkaar staan*. Is er echter qua oorzaak/gevolg geen verschil te maken tussen de variabelen, dan is de keuze willekeurig. In de praktijk wordt het dan door de lay-out bepaald: hoe ziet de tabel er op papier het handigst uit?

Op basis van tabel 12.1b kunnen we nu een uitspraak doen als: 'Van de mannen vindt 50% het eigen leven opwindend, van de vrouwen ongeveer 40%.' Anders dan in tabel 12.1a – met de absolute aantallen – blijkt nu dat de mannen hun eigen leven vaker als spannend beoordelen dan de vrouwen. *Het feit dat deze percentages van elkaar verschillen, betekent dat er statistisch een verband is tussen de variabele sekse en de variabele 'waardering van het eigen leven'. Als er geen verband zou zijn, is de kolom met percentages voor de mannen precies gelijk aan die van de vrouwen, en ook precies gelijk – uiteraard – aan de kolom percentages rechts van de tabel.* Het beantwoorden van de vraag of er wel of geen verband is tussen variabelen, is een van de meest voorkomende vragen in het sociaalwetenschappelijk onderzoek. Daarom is het heel belangrijk dat we een kruistabel primair leren beoordelen in termen van: is er verschil tussen de gepercenteerde kolommen, waar zit dat verschil vooral, is het groot of is het klein? De termen '(statistisch) verband', 'samenhang', 'statistische afhankelijkheid', 'associatie' en '(cor)relatie' tussen variabelen betekenen alle hetzelfde. Voorbeelden van vragen naar het verband tussen twee variabelen zijn:

- Drinken Belgen gemiddeld meer bier dan Nederlanders?
- Zijn er in de beroepsbevolking onder de jongere leeftijdsgroepen meer vrouwen dan onder de oudere leeftijdsgroepen?

– Bestaat er onder de lager opgeleiden meer of minder werkloosheid dan onder de hoger opgeleiden?
– Is er een correlatie tussen roken en longkanker?

Verticaal percenteren en horizontaal vergelijken

Veronderstel dat een enquête ons tabel 12.2 oplevert.

Tabel 12.2 Rookgedrag naar sekse

	a. (Absoluut)				b. (%)		
	Man	**Vrouw**	**Totaal**		**Man**	**Vrouw**	**Totaal**
Rookt	104	98	202	Rookt	38	31	34
Rookt niet	168	216	384	Rookt niet	62	69	66
Totaal	272	314	586		100	100	100

Op basis van de procentuele tabel kunnen we nu uitspraken doen als 'van de mannen rookt een hoger percentage dan van de vrouwen', of – iets minder duidelijk – 'mannen zijn vaker rokers dan vrouwen'. Zoals gebruikelijk hebben we verticaal gepercenteerd. Om te kunnen zien of er een statistisch verband is, moeten we daarom horizontaal vergelijken (bijvoorbeeld 38% met 31%). Als we verticaal hadden vergeleken, bijvoorbeeld 38 met 62, dan beperken we ons in feite tot een univariate (conditionele) vergelijking: we kijken alleen naar de mannen, en vergelijken daarbinnen de rokers en de niet-rokers. Als we echter de bivariate verdeling nemen, wordt het pas interessant, en dit gebeurt alleen als we een percentage uit een reeks die optelt tot 100%, vergelijken met een percentage uit een andere kolom. *Je vergelijkt dus altijd dwars op de richting waarin je percenteert*. Conclusies moeten daarop gebaseerd zijn, en niet op vergelijkingen binnen één kolom. Maar als we nu eens *horizontaal* gepercenteerd hadden, en verticaal hadden vergeleken? Dus:

Tabel 12.2c Rookgedrag naar sekse (in %)

	Man	**Vrouw**	**Totaal**
Rookt	51	49	100
Rookt niet	44	56	100
Totaal	46	54	100

Wat voor uitspraak kunnen we doen bij vergelijking van '51' met '44'? Bijvoorbeeld deze: 'van de rokers is een hoger percentage man dan van de niet-rokers', of: 'rokers zijn vaker mannen dan niet-rokers'. Er is niets tegen deze wijze van vergelijking of tegen de gemaakte uitspraak in te brengen. Maar het is niet een erg zinvolle uitspraak, gezien onze waarschijnlijke bedoeling: het rookgedrag van mannen met dat van vrouwen vergelijken. Wat we hier, daarentegen, doen is zoiets als het vergelijken van de sekseverhouding van de rokers met die van de niet-rokers. Dat hoeft geen onzin te zijn, maar meestal willen we toch verschillende groepen die op de oorzaakvariabele onderscheiden worden, vergelijken voor wat betreft hun gedrag, hun opinie of attitude, dus op de gevolgvariabele. Denk bij het opstellen van een tabel aan de soort uitspraak die verwacht wordt!

Eerder merkten we op dat we in een kruistabel niet percentages binnen één conditionele verdeling met elkaar gaan vergelijken. Als we dat zouden doen, zeggen we niets over *het verband* tussen de variabelen: daarvoor moeten we immers juist de conditionele verdelingen met elkaar vergelijken.
Opvallend is dat in het dagelijks leven heel wat beweringen worden gedaan waarmee men de aanwezigheid van een verband wil aantonen, terwijl die beweringen alleen op één conditionele verdeling zijn gebaseerd.

Een al te gemakkelijke conclusie

Het volgende krantenberichtje verscheen ooit eens in de dagbladpers: 'Twee Deense studenten in de sociologie hebben maandag in een rapport aanbevolen dat aan mannen geen rijbewijs moest worden verstrekt vóór hun 26e jaar "omdat ze voor die leeftijd te agressief zijn om een auto aan toe te vertrouwen". Meisjes kan men dat op hun 18e jaar wel verstrekken. De opstellers van het rapport hebben 38 verkeersongevallen bestudeerd waarbij voetgangers het leven verloren. Die ongelukken waren gebeurd in Aarhus in Jutland, en hadden zich afgespeeld in de afgelopen vier jaar. In 22 van de 38 gevallen was de bestuurder een man geweest in de leeftijd van 22 tot 26 jaar, aldus was gebleken (...).'

De fout die hier gemaakt wordt, komt nogal eens voor. Om dergelijke beleidsconclusies, waarmee een malafide onderzoeker de goegemeente kan oplichten, te doorzien is het heel nuttig om even een kruistabelletje te schetsen. In het voorbeeld gaat het over sekse, leeftijd en het al dan niet maken van verkeersongevallen waarbij voetgangers dodelijk verongelukken. Sekse en leeftijd voegen we daarbij voor het gemak samen tot één tweedeling: mannen van 22-26 jaar versus de rest.

Box 12.1

Vervolg

Tabel 12.3 *Het al dan niet veroorzaken van een dodelijk ongeval naar bevolkingsgroep*

	Mannen 22-26 jaar	Alle anderen	Totaal
Veroorzaakte dodelijk ongeval	22	16	38
Veroorzaakte geen dodelijk ongeval	?	?	?
Totaal	22+?	16+?	alle autorijders

Box 12.1

Omdat niet gezegd wordt (en kennelijk ook niet onderzocht is) hoeveel mannen van 22-26 jaar er in totaal als chauffeur aan het verkeer deelnemen, kan over het verband tussen beide variabelen helemaal niets gezegd worden, laat staan dat er een conclusie over rijbevoegdheid aan verbonden mag worden. Het percentage ongelukveroorzakers in de subgroep van jonge mannen zou moeten worden vergeleken met datzelfde percentage bij alle anderen, of met de verticale randverdeling. Misschien is de verhouding 22/16 óók ongeveer in de onderste rij aanwezig (dan is er dus geen enkel verband); misschien ook is deze verhouding nog schever in de onderste rij (dan zijn jonge mannen dus voorzichtiger dan de anderen); misschien ook hebben de onderzoekers gelijk met hun suggestie. Maar aangezien de hele tweede rij ontbreekt, is daarover niets te zeggen.

Kruistabellen over drie of meer variabelen

Het is mogelijk om een kruistabel over drie of meer variabelen op te stellen, of, zoals we ook wel zeggen, een kruistabel met drie ingangen; zie tabel 12.4.

Tabel 12.4 *Zelfrespect naar geslacht en sociale klasse*[*]

	Boys			Girls		
	SOCIAL CLASS					
Self-esteem	**Upper**	**Middle**	**Lower**	**Upper**	**Middle**	**Lower**
High	55%	47%	36%	47%	46%	41%
Medium	17	25	26	28	25	27
Low	28	28	39	24	29	32
Total per cent	100	100	100	100	100	100
(Number)	(89)	(1383)	(168)	(106)	(1311)	(172)

* Rosenberg, M. (1965). *Society and the adolescent self-image*. Princeton: Princeton University Press, p. 41, table 2.

Kruistabellen over drie of meer variabelen bevatten veel informatie, maar ze worden al gauw onoverzichtelijk. In SPSS worden ze eenvoudig gesplitst in kruistabellen van twee variabelen. Door in CROSSTABS een derde variabele in te voegen verschijnt een aparte inkomen/opleiding-tabel voor de mannen, en óók voor de vrouwen. In de rapportage kun je desgewenst tabellen met wel vijf ingangen opnemen om ruimte te besparen, zoals wanneer je meningen over euthanasie tegelijkertijd in verband wilt brengen met sekse, leeftijd, opleiding en godsdienst. Maar erg aanbevelenswaardig is dit niet.

Als er een duidelijke gevolgvariabele is, en wanneer deze slechts twee categorieën heeft, lossen we het probleem van de onoverzichtelijkheid van een tabel met drie ingangen vaak op zoals in tabel 12.5. Een tabel zoals deze is aanzienlijk efficiënter dan een tabel met een rij voor 'college plans' en een aparte rij voor 'no college plans'. De tweede rij zou geheel overbodig zijn, en wordt daarom weggelaten in tabel 12.5. Maar let goed op de kop van de tabel. Het is niet een kruistabel van intelligentie en sociale klasse met geslacht (zoals je misschien zou denken als je alleen op de namen van de variabelen let), maar de percentages corresponderen met de percentages van studenten die van plan zijn naar een college te gaan. De percentages van hen die niet naar een college willen gaan, zijn eenvoudig weggelaten.

Tabel 12.5 *Het percentage dat naar het hoger onderwijs wil van alle mannelijke en vrouwelijke laatste-jaarleerlingen van de Milwaukee Public High School, naar intelligentie en sociaaleconomische status*[*]

	Males		Females		Total	
Intelligence						
Low	24.0	(516)	13.1	(863)	17.2	(1379)
Middle	42.1	(620)	34.6	(741)	38.0	(1361)
High	63.4	(618)	49.0	(641)	56.1	(1259)
Socio-economic Status						
Low	24.6	(561)	14.1	(799)	18.5	(1360)
Middle	41.2	(610)	28.2	(791)	33.8	(1401)
High	66.6	(583)	53.0	(655)	59.4	(1238)
Total	44.3	(1754)	30.4	(2245)	59.5	(3999)

* W.H. Sewell & J. Michael Armer (1972). Neighbourhood Context and College Plans. In P.F. Lazarsfeld, A.K. Pasanella & Morris Rosenberg (Eds.), *Continuities in the Language of Social Research* (p. 285). New York: Free Press.

Deze wijze van opstelling van een tabel kan ook worden gebruikt als de afhankelijke variabele meer dan twee waarden heeft, en slechts een van die waarden interessant is. Wanneer je bijvoorbeeld te maken hebt met vijf antwoordcategorieën voor 'houding ten opzichte van abortus' is het denkbaar dat je alleen geïnteresseerd in 'onder alle omstandigheden toegestaan'. We gebruiken dan een tabel zoals tabel 12.5, met nu als kop: 'Percentage van de onderzochten dat onder alle omstandigheden met abortus instemt'.

De sterkte van een statistisch verband

Nominaal meetniveau

Het gaat natuurlijk niet alleen om de vraag of er wel of geen verband is, maar meteen ook om de vraag hoe sterk dat verband is. Hoe kunnen we die sterkte uitdrukken? Daar zijn vele maten voor, maar de eenvoudigste is de zogeheten epsilon: het verschil tussen de twee percentages in de bovenste rij van tabel 12.2: 50% – 40% = 10%, of 0,10.[1] Maar met een tabel die groter is dan een 2x2-tabel – zoals hier – ben je er dan nog niet, want in de middelste rij vind je (afgerond): -.08, en in de onderste rij -.02. In een tabel die meer dan vier cellen telt, zijn er dus verschillende epsilons te berekenen. Voor het gemak worden tabellen daarom nogal eens vereenvoudigd tot 2x2-tabellen, zoals hierna in tabel 12.6.

Tabel 12.6 Waardering van het eigen leven naar sekse

	(Absoluut)			(In %)		
	Man	**Vrouw**	**Totaal**	**Man**	**Vrouw**	**Totaal**
Opwindend	213	221	434	50,1	39,8	44,3
Anders	212	334	546	49,9	60,2	55,7
Totaal	425	555	980	100,0	100,0	100,0

We gebruiken hierna de Engelse manier van noteren van getallen met decimalen, dus met een punt in plaats van een komma, en we laten de nul weg als het getal voor de punt nul is.

1 In de Angelsaksische wereld gebruikt men voorafgaande aan de decimalen niet een komma, maar een punt. Een nul voor de punt wordt weggelaten. Een getal als 0,10 wordt dan ook uitgesproken als 'point ten'. Nederlandse onderzoekers hebben dat overgenomen, en spreken van 'punt tien'. Het getal 23,45 wordt geschreven als 23.45 en uitgesproken als 23 punt 45.

De (afgeronde) epsilon is nu eenduidig .10 (of -.10 als je naar de onderste rij kijkt). Het plus- of minteken wijst ons erop dat we de richting van het verband er altijd bij moeten vermelden. Hier is dat: 'mannen vinden het leven vaker opwindend dan dat vrouwen dat doen'.

Als de effectvariabele was gemeten op interval- of rationiveau, zoals lengte en gewicht, zouden we de sterkte van het verband kunnen uitdrukken als de grootte van het verschil in lengte tussen mannen en vrouwen. Indien beide variabelen gemeten zouden zijn op interval- of rationiveau (zoals lengte én gewicht) zouden we waarschijnlijk de zogeheten correlatiecoëfficiënt hebben gebruikt om de sterkte van het verband vast te leggen. We komen daar later op terug. Hier willen we alleen aangeven dat de sterkte van een statistisch verband op verschillende manieren kan worden uitgedrukt.

Een wat ingewikkelder manier om samenhang in een 2x2-tabel vast te stellen wordt meestal in het Engels aangeduid als de *odds ratio*. In het Nederlands zou het tot de lange term 'verhouding van kansverhoudingen' leiden. Daarom nemen we de Engelse term maar over. Omdat je, als je verder komt op het terrein van de data-analyse, deze term nog vaak zult ontmoeten, bespreken we deze manier hier ook. De redenering erachter is als volgt:

- Eerst kijken we naar de verhouding tussen de kans dat een man het leven opwindend vindt, ten opzichte van de kans dat hij het niet opwindend vindt. Die verhouding is (zie tabel 12.6): 213/212.
- Dan kijken we naar diezelfde verhouding voor een vrouw. Die verhouding is 221/334.
- Hoe verhouden zich nu deze twee verhoudingen? Welnu, het antwoord daarop is: 213/212 / 221/334 = 213/212 x 334/221, of 213x334 / 212x221 = 1.52. Als je dezelfde bewerking uitvoert op de procentuele tabel is de uitkomst eveneens 1.52.

Als we in een 2x2-tabel de cellen benoemen als in de figuur, komt de formule neer op (a/c) / (b/d) = ad / bc. Daarom wordt het ook wel de kruisproductenratio genoemd.

a	b
c	d

Ga na dat:
- de *odds ratio* gelijk blijft als we de verhoudingen horizontaal bepalen;
- de *odds ratio* gelijk is aan 1 ('gelijke kansen') indien en alleen indien de percentages in de kolommen gelijk zijn binnen elke rij, dus wanneer de variabelen niet gecorreleerd zijn; het 'omslagpunt' is dus 1;
- verwisselen van rijen en kolommen leidt tot, bijvoorbeeld, bc/ad. De ene waarde is de omgekeerde waarde van de ander. Beide geven de sterkte van het verband aan, maar in omgekeerde richting. Het verwisselen van rijen en kolommen doet er dus niet toe;
- de ondergrens van de *odds ratio* is 0, de bovengrens is onbepaald.

Het gebruik van *odds ratios* heeft echter nadelen, vooral in de handen van weinig deskundige schrijvers, maar vooral ook lezers. We ontlenen het volgende voorbeeld aan Van Maanen (2009, p. 38).

> Kinderen die naar muziek op hun MP3-speler luisteren, zetten het ding vaak te hard. Uit een onderzoek bleek ook dat 'minder jongens oordoptelefoons gebruikten dan meisjes' (*odds ratio* 0.51). De argeloze lezer denkt wellicht dat half zoveel jongens als meisjes oordopjes hebben, maar dat is niet zo: 95.3% van de meisjes gebruikte oordopjes, tegen 90.8% van de jongens. Bij de jongens is de odds 90.8/0.2 = 9.87, bij de meisjes 95.3/4.7 = 20.3, hetgeen een *odds ratio* oplevert van 9.87/20.3 = .49. Een juiste, en in elk geval belangrijkere, conclusie was geweest dat het gebruik van oordopjes bij jongens en bij meisjes weinig verschilt.

Box 12.2

Gebruik van logaritmen

Dat er geen vaste bovengrens is, is dus een bezwaar tegen de *odds ratio*. Bovendien is de *odds ratio* erg gevoelig voor kleine veranderingen in de celfrequenties, en voor heel kleine cellen (de *odds ratio* kan dan zeer hoog worden), zoals we in het voorbeeld zagen. We kunnen deze bezwaren omzeilen door logaritmen te gebruiken: (log a – log b) – (log c – log d) = (log a+ log d) – (log b + log c). Als we de natuurlijke logaritme gebruiken (e = ongeveer 2.7) als basis, noemen we dit een *logit transformatie*, en de uitkomsten *logits*. De *log-odds ratio* heeft een gemiddelde van 0 (geen samenhang) en is symmetrisch tussen - en + , met een standaardafwijking van 1.83.

Er is nog een andere manier om de onbepaalde bovengrens te omzeilen. De maat Kendall's Q is een goede maat voor samenhang in 2x2-tabellen:

Q = (odds ratio – 1) / (odds ratio + 1).

Ga na dat de bovengrens van Q nadert tot 1 (als de *odds ratio* heel groot is), en de ondergrens nadert tot -1 (als de *odds ratio* heel klein is). Deze grenzen worden in feite bereikt als een cel van de tabel leeg is. Ga ook na dat Q gelijk is aan 0 als de *odds ratio* gelijk aan 1 is (de 'verhoudingen zijn gelijk', geen verband!).
Nu zou je je kunnen afvragen waarom we niet altijd Q gebruiken en de *odds ratios* maar verder vergeten. Van Maanen (2009) stelt zelfs een algemeen verbod op *odds ratios* in de media voor. Het antwoord is dat Q een grootheid betreft die gemakkelijk berekend, maar niet in woorden uitgelegd kan worden. Als je het hebt over percentages die je vergelijkt, of over kansverhoudingen die je vergelijkt, is dat wel iets dat je in woorden kunt uitdrukken ... maar dat moet je dan wel heel goed doen om misverstanden te vermijden!

Nog een tip: je kunt Q sneller berekenen, op de absolute aantallen of op de percentages, via de formule Q = (ad-bc) / (ad+bc), waarin a = de vulling van de cel linksboven, en d = de vulling van de cel linksonder; ad en bc zijn weer de bekende kruisproducten.
Bij een tabel die groter is dan twee rijen en twee kolommen, stuit je op hetzelfde probleem als bij de epsilon: er zijn diverse vergelijkingen te maken, dus diverse *odds ratios* te berekenen. Maar voor een 2x2-tabel is er maar één.

Oefening

Doe nu eerst de volgende oefening om te checken of het voorgaande duidelijk is. Vul de tabel volledig in en bereken de epsilon, de *odds ratio* en de Q.

140	?	300
?	?	260
180	380	560

In de tabel hierna is de frequentieverdeling van dezelfde 560 eenheden op *elk van de afzonderlijke variabelen* gelijk gebleven, maar de *bivariate* frequentieverdeling is iets veranderd; de sterkte van het verband zal dus anders zijn. Vul nu ook deze tabel volledig in. Schat tevoren of het verband sterker of zwakker is dan in de tabel hiervoor. Bereken vervolgens de epsilon, de *odds ratio* en de Q.

120	?	300
?	?	260
180	380	560

Box 12.3

Is er verband of is er geen verband?

We zagen al dat we spreken over een (statistisch) verband als, in een procentuele tabel, de kolommen niet gelijk zijn aan elkaar. Automatisch wil dit zeggen dat dan ook de rijen niet aan elkaar gelijk zijn; als het een het geval is, is ook het ander het geval. Als er samenhang is tussen sekse en roken, betekent dit dat er onder de mannen een hoger, of een lager, in ieder geval een *ander* percentage rokers is dan onder de vrouwen.

Om meer inzicht te krijgen in het verschijnsel samenhang bekijken we tabel 12.7a, waarin dezelfde drie (univariate) randverdelingen zijn gebruikt als in tabel 12.1, maar waarin we de cellen nu zodanig ingevuld hebben dat er *geen* statistisch verband is. Dat betekent dat we de verticale conditionele verdelingen nu hetzelfde hebben gemaakt als de randverdeling. Zouden we nu deze tabel in absolute aantallen willen hebben, dan passen we de percentages eenvoudig toe op de absolute aantallen op de onderste regel, en komen dan tot tabel 12.7b (reken na!):

Tabel 12.7 Waardering van het eigen leven naar sekse (geen verband; fictief!)

	a. (In %)				b. (Absoluut)		
	Man	**Vrouw**	**Totaal**		**Man**	**Vrouw**	**Totaal**
Opwindend	44.3	44.3	44.3	Opwindend	188	246	434
Tamelijk gewoon	51.5	51.5	51.5	Tamelijk gewoon	219	286	505
Saai	4.2	4.2	4.2	Saai	18	23	41
Totaal	100.0	100.,	100.0	N	425	555	980

De verkregen celvullingen worden de 'verwachte waarden onder statistische *on*afhankelijkheid' genoemd, of kortweg de 'verwachte waarden' (*expected counts*).

Dus nog eens samenvattend: *statistische onafhankelijkheid betekent dat je de procentuele frequentieverdeling rechts van de tabel precies zo terugvindt in elke kolom van de tabel.*

Kunnen we voor elke cel de 'verwachte' vulling berekenen als we alleen de randverdelingen van de beide variabelen kennen? Ja, dat is niet zo moeilijk. Weten we van tabel 12.8 alleen de randverdelingen en willen we bijvoorbeeld de celfrequentie linksboven weten in de situatie dat de variabelen statistisch geen verband vertonen, dan redeneren we als volgt:

- De kans op een man is 425/980.
- De kans op 'opwindend' is 434/980.
- Nu moeten we nog de kans op de combinatie 'man-opwindend' weten. Welnu, bij statistische onafhankelijkheid mogen we de productregel voor kansen toepassen. De kans op de combinatie 'man-opwindend' is dus 425/980 x 434/980.
- En in absolute aantallen: 425/980 x 434/980 x 980, oftewel 425 x 434/980 = 188.

Met andere woorden: *we berekenen elke celvulling bij statistische onafhankelijkheid door de desbetreffende randfrequenties met elkaar te vermenigvuldigen en het product te delen door N.*
Naarmate er een groter verschil is tussen enerzijds de feitelijke celvullingen en anderzijds de verwachte vullingen (bij onafhankelijkheid), is de samenhang tussen de betrokken variabelen sterker. Daarom is het van belang de 'verwachte vullingen' te berekenen. SPSS rekent als je dat wilt zowel de verwachte vulling voor elke cel uit als het verschil tussen de feitelijke en de verwachte vulling: kies CROSSTABS, CELLS, COLUMN PERCENTAGES, OBSERVED COUNTS en EXPECTED COUNTS.

Het is zaak de verkregen tabel cel voor cel goed te bestuderen, vooral als het grote tabellen (grote r en grote k) betreft. De cellen waar werkelijk 'iets aan de hand is', verraden zich door een groot verschil, een grote *residuwaarde* zoals we dat noemen. Als in alle cellen de residuen erg klein zijn, zullen we al gauw zeggen dat er geen verband is. De situatie dat alle residuen precies gelijk aan 0 zijn, zullen we niet zo gauw vinden; immers, toevalsfouten in de metingen zijn te verwachten, waardoor ook als er geen verband is de feitelijke en de verwachte waarden niet overal precies gelijk zullen zijn.

We komen nog eens terug op het begrip samenhang. We zagen al dat de sterkte van de samenhang uitgedrukt kan worden in epsilon, *odds ratio* of Q. Een bezwaar van deze maten is dat ze alleen berekend kunnen worden voor 2x2-tabellen, of liever gezegd alleen maar handig zijn voor 2x2-tabellen. We kunnen, nu we het begrip 'residu' kennen, goed beargumenteren waarom het berekenen van de sterkte van een samenhang zo belangrijk is, en we kunnen ook tot goede maten komen voor grotere dan 2x2-tabellen. De redenering loopt als volgt.

Veronderstel dat we volgens toeval een van de 980 personen eruit kiezen zonder dat we iets van deze persoon af weten, en dat we zouden willen voorspellen wat diens score is op 'waardering van het eigen leven'. *Als we alleen de randverdeling van die variabele weten*, is de beste gok: 'tamelijk gewoon', omdat

deze nu eenmaal het meest voorkomt. We hebben dan een kans van .48 dat we fout voorspellen (1 -.52; zie de randverdeling bij tabel 12.7).

Maar als we nu eens weten dat die persoon een man is? We voorspellen nu, kijkend naar alleen de linkerkolom, opnieuw 'tamelijk gewoon', omdat dat het vaakst juist is. We hebben dan weer een kans van .50 dat we fout zitten. En als we weten dat die persoon een vrouw is, van hetzelfde laken een pak. Met onze kennis over de sekse van die persoon wordt onze voorspelling er dus niet beter op.

Tot zover gingen we uit van tabel 12.7, waarin er geen verband is tussen de variabelen 'sekse' en 'waardering'. Maar ... als er wel een verband is, zoals in tabel 12.1, verandert de situatie. Als we nu weten dat de willekeurig getrokken persoon een man is, hoeven we alleen te kijken naar de conditionele verdeling voor 'mannen', en we voorspellen (zie tabel 12.1b) natuurlijk de waarde: 'opwindend' (omdat deze waarde bij de mannen het meest voorkomt). We hebben dan een kans van .50 dat we fout voorspellen. En als het een vrouw is, voorspellen we 'tamelijk gewoon' en hebben dan een kans van .45 dat we fout zitten.

Als er een verband is tussen twee variabelen helpt dus kennis van de score van een eenheid op de ene variabele ons bij het goed voorspellen van de score van die eenheid op de andere variabele. En als er geen verband is, helpt die kennis ons niets.

Dat is de uitleg van het begrip 'samenhang'. In het voorgaande geval – tabel 12.7 versus tabel 12.1 – is die verbetering erg klein. Maar als we – fictief – nu eens die percentagekolom voor de mannen flink hadden laten verschillen van die voor de vrouwen (we krijgen dan een heel sterk verband), dan is de verbetering van de voorspelling ten opzichte van een nulverband zeer aanmerkelijk; zie tabel 12.8:

Tabel 12.8 Waardering van het eigen leven naar sekse (abs.) (fictief: maximaal verband)

	Man	Vrouw	Totaal
Opwindend	425	9	434
Tamelijk gewoon	-	505	505
Saai	-	41	41
Totaal	425	555	980

Immers, als we weten dat iemand 'man' is, voorspellen we nu steeds 'opwindend' en maken dan geen enkele fout; voor de vrouwen voorspellen we 'tamelijk gewoon' en we hebben dan een kans van slechts (9+41)/555 = .09 om een fout te maken. Door de bank genomen maken we dus heel wat minder fouten dan bij tabel 12.1! Merk op dat we steeds dezelfde randverdelingen aanhouden, maar de sterkte van het verband tussen de beide variabelen variëren door de celvullingen te veranderen.
De vraag die nog overblijft, is: hoe komen we aan de celvullingen van tabel 12.8? We beginnen met het skelet van de tabel aan te geven en de randfrequenties in te vullen. In principe moeten we nu zorgen dat linksboven en rechtsonder de celvullingen maximaal zijn, teneinde de verhouding tussen de kolommen zo scheef mogelijk te trekken. Hoeveel eenheden kunnen er maximaal in de cel linksboven? Het kleinste aantal van de twee randfrequenties 425 en 434, dus hier komen 425 eenheden. Evenzo rechtsonder: 41. De rest van de tabel laat zich nu automatisch vullen. Bij een maximaal verband is er altijd minstens één cel die *leeg* is.

Een vraag: waarom maximaliseren we de cellen linksboven en rechtsonder, waarom niet die linksonder en rechtsboven? Welnu, dat laatste kan natuurlijk ook. De volgende tabel ontstaat dan.

Tabel 12.9 Een maximaal verband 'de andere kant op'

	Man	Vrouw	Totaal
Opwindend	-	434	434
Tamelijk gewoon	384	121	704
Saai	41	-	41
Totaal	425	555	980

Ook hier is er een maximaal verband, maar dan 'de andere kant op'; het zijn nu de mannen die hun leven wel eens saai vinden, en de vrouwen die het als opwindend beoordelen.

We hebben nu, steeds van variabelen met dezelfde univariate frequentieverdelingen uitgaand, verschillende tabellen geconstrueerd:
– met een zwak verband (tabel 12.1a en b);
– met een nulverband (tabel 12.7a en b);
– met een maximaal sterk verband de ene kant op (tabel 12.8);
– met een maximaal sterk verband de andere kant op (tabel 12.9).

In plaats van 'de ene kant op' en 'de andere kant op' spreken we meestal over een positief of een negatief verband. Maar wat we positief, en wat we negatief noemen, is in het geval van vorenstaande tabellen volstrekt nietszeggend omdat, als we de kolommen van mannen en vrouwen verwisselen zonder verder iets te veranderen, een positief verband in een negatief omslaat, of andersom. Ga dit na!

Enig nadenken over tabel 12.8 en tabel 12.9 leert ons dat het spreken in termen van 'de ene kant op' en 'de andere kant op' alleen zinvol is als de categorieën van beide variabelen in een zinvolle rangorde staan, met andere woorden bij variabelen van een ordinaal of hoger meetniveau. Wanneer we de drie categorieën links van de tabel naar hartenlust haasje-over zouden kunnen laten springen (dus 'tamelijk gewoon' als uiterste categorie), zouden nog weer andere 'maximaal'-tabellen mogelijk zijn, en verliest dit begrip zijn zin. Bij nominale variabelen is het zinloos te spreken over een positief of negatief verband. Pas als beide variabelen van ordinaal of hoger meetniveau zijn (bijvoorbeeld lengte en gewicht van mensen), is het zinvol om af te spreken wat we met een positief en wat we met een negatief verband bedoelen. De afspraak is: als hoge waarden op de ene variabele samengaan met hoge op de andere (en lage met lage), noemen we het verband positief. De samenhang tussen lengte en gewicht van mensen is dus zeker positief; de samenhang tussen opleiding en uren televisiekijken waarschijnlijk negatief.

Kunnen we nu ook bij 2x3-tabellen, of nog grotere tabellen, bijvoorbeeld door het optellen van de residuen van een tabel, komen tot één getal, één maat voor de sterkte van de samenhang tussen twee variabelen? Oppervlakkig gezien zouden we kunnen zeggen: hoe groter de som van (de absolute waarde van) deze residuen, des te sterker het verband. Maar zo eenvoudig is het nu ook weer niet. Immers, als onze steekproef tweemaal zo groot zou zijn, dus als alle getallen in een tabel tweemaal zo groot zouden worden, zou de 'som van de residuen' óók tweemaal zo groot worden, terwijl we met de sterkte van het verband niets uitgehaald hebben. En ook: hoe meer rijen en kolommen er zijn, hoe groter de som van de residuen zal zijn. Je kunt dus niet op deze eenvoudige manier de sterkte van de samenhang in de ene tabel vergelijken met die in een andere tabel waarin N groter is of die meer rijen en/of kolommen heeft.

Er bestaat een maat, de zogeheten χ^2 (spreek uit: chi-kwadraat; Engelse uitspraak: *kai-square*), die in SPSS veel aandacht krijgt, en die gebaseerd is op het verschil tussen feitelijke en verwachte waarden. De maat is als volgt:

$$\chi^2 = \Sigma (f_o - f_e)^2 / f_e$$

waar f_o = feitelijke frequentie in een cel (o van *observed*),
f_e = verwachte frequentie in een cel (e van *expected*).

De χ^2 is echter niet primair een maat voor de sterkte van een samenhang, maar wordt berekend om bij een aselecte steekproef van een bepaalde grootte na te kunnen gaan of het in de steekproef gevonden verband zo sterk is dat met een bepaalde mate van betrouwbaarheid kan worden aangenomen dat er in de populatie óók een verband tussen die twee variabelen zal zijn. Als dat het geval is, spreken we van een *statistisch significant verband*.

Het berekenen van χ^2 is bij een grote steekproef nauwelijks van belang. Bij een grote steekproef is een heel zwak statistisch verband al gauw significant. Dus statistisch significant wil helemaal niet zeggen dat het een sterk, een wetenschappelijk belangrijk of een voor het beleid interessant verband is. Dat kan zo zijn, maar het hoeft niet.

Als we daarentegen met een kleine steekproef te maken hebben, kan het berekenen van χ^2 heel nuttig zijn. We kunnen het dan zien als een eerste zeef: als een verband niet significant is, is het best mogelijk dat het toevallig is; dat het tot stand gekomen is door bijvoorbeeld fouten in de metingen, waardoor 'toevallig' in de tabel wat afwijkingen van de verwachte waarden optraden. Daar komt bij dat onder meer door het optreden van selecte non-respons een vraagteken gepast is met betrekking tot de toepasbaarheid van wiskundige statistiek. Als een verband dus niet significant is, zullen we er verder geen aandacht aan besteden. Is het wel significant, dan pas gaan we verder kijken waar in de tabel opvallende residuen zijn.

Deze overwegingen leren ons dat we niet te gauw moeten zeggen bij het bestuderen van een tabel: kijk maar, de χ^2 is significant, *dus* hebben we iets belangrijks gevonden. Een verschil van slechts enkele percentagepunten is sowieso meestal van geen enkel belang. Als van de vrouwen 41% en van de mannen 45% een of ander gedrag vertoont, wordt hier door sommigen een geweldig interessant verschijnsel van gemaakt, zeker als het significant blijkt te zijn. Maar beleidsinstanties zullen in zo'n gering verschil geen aanleiding zien om voor mannen een ander beleid te voeren dan voor vrouwen, en een 'theoretische interpretatie' van een dergelijk verschil is niet iets waar velen zich aan zullen wagen. Het constateren van een significant verschil is dus niet het antwoord op alle vragen!

Daar komt nog iets bij. Als we, en dat geldt zeker voor nominale variabelen, de sterkte van de samenhang in één getal proberen te vatten, verliezen we wel heel veel informatie. Het valt nog wel mee als we een kruistabel van een klein formaat hebben, bijvoorbeeld van 'al dan niet stemmen naar sekse'. Maar als we een tabel van veel groter formaat hebben, zoals voor 'studievoorkeur van universiteitsstudenten naar provincie van herkomst', met een formaat van 65x12, heeft het natuurlijk geen zin de eventuele samenhang in één getal uit te drukken. We zijn hier bij uitstek geïnteresseerd in de vraag of, en zo ja, waar precies de kolommen met percentages van provincie tot provincie uiteenlopen, en we zullen zo'n tabel dan ook liever 'met een timmermansoog' bestuderen, en wellicht hier en daar de grootste afwijkingen noteren. Dat gaat nog efficiënter als we de residuen bestuderen. Waar we hier in geïnteresseerd zijn, is bijvoorbeeld de constatering dat Friese studenten vaak theologie kiezen, en Limburgers 'boven kans' diergeneeskunde.

Bovendien houdt de χ^2 geen rekening met het feit dat sommige variabelen een hoger meetniveau hebben dan nominaal. Immers, de χ^2 blijft hetzelfde, ook wanneer we rijen en/of kolommen haasje-over laten springen. En dat is jammer, want daarmee wordt veel informatie niet gebruikt.

Het is daarom veel beter om, zoals al gezegd, een kruistabel voor nominale variabelen cel voor cel rustig op de relatieve grootte van de residuen na te lopen en te interpreteren. Onthoud: een residu is het verschil tussen de werkelijke celvulling en de verwachte celvulling als er geen statistische samenhang is tussen de variabelen.

Een bijzonderheid: het is nog verstandiger om naar de *standardized residuals* te kijken. Deze zijn gecorrigeerd voor verschillen tussen de eraan ten grondslag liggende aantallen. Ze kunnen worden geïnterpreteerd als standaardnormale z-waarden.

Ordinaal niveau

Als een van de beide variabelen wordt gemeten, of als beide variabelen worden gemeten op ordinaal niveau, staan er verschillende statistische maten klaar om de sterkte van het verband mee vast te leggen. Heel bekend zijn bijvoorbeeld Goodman en Kruskals 'gamma'- en Kendalls 'tau'-maten. Hier gaan we niet verder op deze maten in. Uitstekende toelichtingen op deze maten zijn te vinden in vele statistiekboeken, en zijn bovendien heel gemakkelijk via internet te

vinden. Omdat er in de onderzoekspraktijk weinig ordinale variabelen te vinden zijn die *niet* behandeld kunnen worden als variabelen gemeten op intervalniveau, laten we dit onderwerp over aan de geïnteresseerde lezer.

Interval- en rationiveau

Verschillen tussen variabelen van interval- en van rationiveau zijn vanuit statistisch oogpunt van weinig belang. Daarom nemen we die niveaus hier samen. We beginnen met de situatie waarin de oorzaakvariabele van nominaal niveau is, en de gevolgvariabele van interval- of rationiveau.

Tabel 12.10 Rapportcijfer voor lezen naar sekse

Rapportcijfer	Meisjes	Jongens	Totaal
3	2	3	5
4	7	10	17
5	9	11	20
6	21	26	47
7	29	29	58
8	12	9	21
9	6	4	10
Totaal	86	92	178

De vraag luidt: is er een samenhang tussen sekse en rapportcijfer? In de taal van ons dagelijks leven zouden we waarschijnlijk zeggen: is er verschil tussen jongens en meisjes wat het rapportcijfer betreft? We hebben al gezien dat de vraag of er verschil is tussen twee of meer groepen op een variabele X, *dezelfde betekenis* heeft als de vraag 'of er samenhang is tussen de betrokken variabelen'. Dat wil zeggen tussen de variabele waarop de groepen zijn onderscheiden (bijvoorbeeld sekse), en variabele X.

Het geven van een antwoord op de vraag is simpel: we berekenen het gemiddelde voor meisjes en dat voor jongens, en kijken of er een verschil tussen die twee is. En hoe groter dat verschil is, des te sterker de samenhang. Blijft alleen de vraag of dat verschil zo groot is, dat het niet aan toeval toegeschreven kan worden. Die vraag kunnen we – met de genoemde reserves – door middel van statistische toetsing beantwoorden: we kijken of het verschil 'significant' is. Zo niet, dan besteden we er verder geen aandacht aan.

Bedenk dat we ook heel goed de tabel in procenten hadden kunnen omzetten, en vervolgens hadden kunnen inspecteren of voor achtereenvolgens het cijfer 3 bij de meisjes en de jongens in de tabel hetzelfde percentage verschijnt; dan voor het cijfer 4, enzovoort (analoog aan tabel 12.1b). Maar bij een variabele op intervalniveau hebben de volgorde van de categorieën en de grootte van de afstand betekenis, vandaar dat we van het gemiddelde gebruikmaakten. We zien overigens aan de tabel in één oogopslag (nu ja ...) dat meisjes wat frequenter in de hogere cijfers en wat minder frequent dan de jongens in de lagere cijfers zitten, en daaraan kunnen we al voorspellen dat het gemiddelde voor de meisjes wat hoger zal uitvallen dan voor de jongens. *Probeer een dergelijke schatting altijd tevoren te maken als je een tabel in handen hebt!*

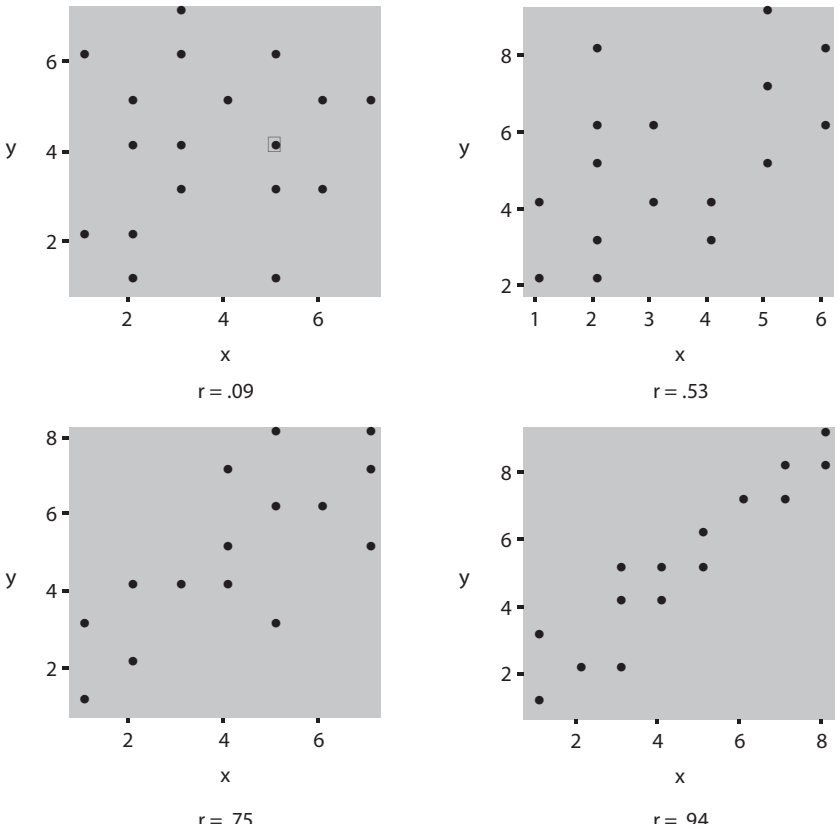
Als de gevolgvariabele ook op nominaal niveau was geweest, hadden we de rijen naar willekeur mogen verwisselen, en had het begrip hoge of lage cijfers geen betekenis gehad. Als de gevolgvariabele op interval- of rationiveau ligt, is het berekenen van gemiddelden, apart voor elke waarde van de oorzaakvariabele, *veel efficiënter*.

Is het ook mogelijk dat de oorzaakvariabele op intervalniveau ligt, en de gevolgvariabele op nominaal meetniveau? Dat komt inderdaad, al betreft het uitzonderingsgevallen, wel eens voor. Denk aan leeftijd als oorzaak en bijvoorbeeld keuze van politieke partij als effect. In principe staan de leeftijdscategorieën dan in de kolommen en VVD, PvdA, CDA enzovoort als rijen in de tabel, en wordt het aantal in elke leeftijdscategorie op 100% gesteld. Het is denkbaar dat voor het CDA het percentage van links naar rechts steeds groter wordt, wat erop duidt dat het percentage CDA-stemmers toeneemt naarmate we naar een oudere leeftijdscategorie kijken. Het ligt in dit geval echter eerder voor de hand om horizontaal te percenteren. Of nog liever: om, in afwijking van de gewoonte, de politieke-partijkeuze in de kolommen en de leeftijdscategorieën als rijvariabele op te nemen, en dan zoals gewoonlijk verticaal te percenteren. We kunnen dan de gemiddelde leeftijd per politieke partij berekenen, waarmee we dus duidelijk gebruikmaken van het rationiveau van de variabele 'leeftijd'. Nu kunnen we deze gemiddelden van partij tot partij vergelijken. Daarmee is de werkwijze vergelijkbaar met die in het sekse/rapportcijfers-voorbeeld. Maak zelf een schets van de verschillende tabellen die hier aan de orde zijn.

En hoe is de situatie als beide variabelen op interval- of rationiveau liggen? We kunnen nu echt de scores van elke eenheid op twee variabelen afzetten door de variabelen op een assenstelsel weer te geven; we krijgen dan een puntenwolk (*scattergram*, *scatter plot* of *scatter diagram*). Als je een puntenwolk bekijkt, zijn er drie aspecten van belang:

1. de richting van de statistische samenhang (positief, dat wil zeggen van linksonder naar rechtsboven);
2. de vorm (min of meer langs een rechte lijn, of langs een gekromde lijn);
3. de spreiding (*scatter*) over het plaatje (hoe smaller de puntenwolk is, des te sterker is de samenhang).

Omdat het aantal onderscheiden waarden (in de ruwe data) vaak heel groot is, is het zelfs praktischer om een puntenwolk te maken dan een tabel. Zouden we een leesbare tabel willen, dan zouden we de vele waarden eerst moeten groeperen, wat weer verlies aan informatie betekent.
Figuur 12.2 laat enkele mogelijke puntenwolken zien van twee variabelen die beide op interval- of rationiveau liggen. Uit de erbij afgedrukte waarde van de correlatiecoëfficiënt (waarvoor we het symbool r gebruiken; r_{xy} is de correlatie tussen de variabelen X en Y) blijkt dat deze coëfficiënt hoger wordt naarmate de puntenwolk smaller is, en in de buurt van 0 komt naarmate de puntenwolk ronder is.

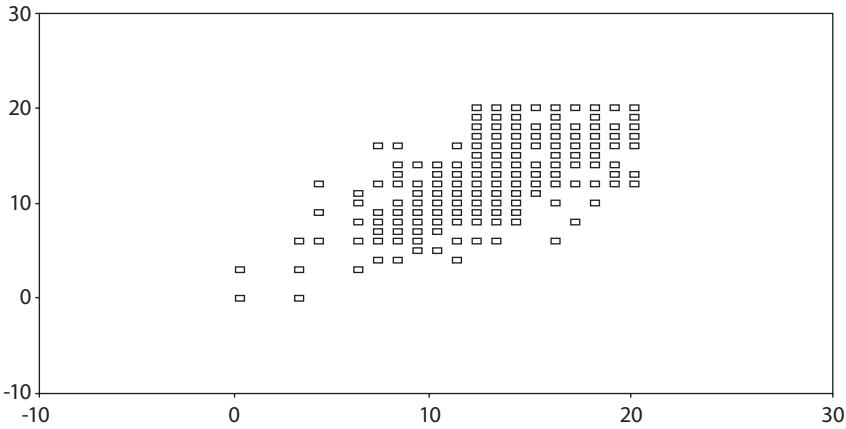


Figuur 12.2 Verschillende puntenwolken

Wanneer we het verband tussen twee variabelen op intervalniveau bestuderen, is het erg nuttig om niet alleen de r te berekenen, maar om bovendien de puntenwolk te zien. Dit is via SPSS mogelijk:

- Klik op GRAPHS.
- Klik op SCATTER.
- Klik op SIMPLE.
- Klik op DEFINE.

Nu kun je de variabelen kiezen. Kies bijvoorbeeld binnen de GSS-file de eigen opleiding in aantal jaren op de x-as, en de opleiding van de partner (*spouse*) op de y-as. Je krijgt nu een puntenwolk te zien, waarin elk groepje waarnemingen door een rood vierkantje gerepresenteerd wordt. Helaas is in de vraagstelling '20 of meer jaren opleiding' kennelijk niet verder uitgesplitst, want zowel bij de respondent als bij de partner zijn er erg veel die in die categorie vallen. Maar de vorm van de puntenwolk is wel duidelijk: er is een sterke positieve correlatie (r = .62). Met andere woorden: soort zoekt soort!

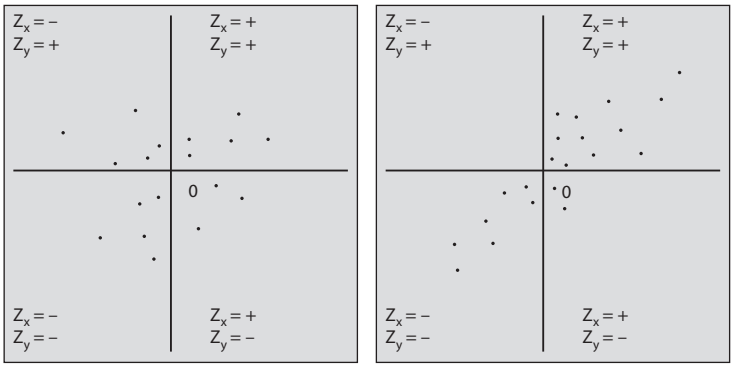


Figuur 12.3 Opleiding partner naar opleiding respondent (in aantal jaren scholing)

We leggen hierna het principe van de correlatiecoëfficiënt r uit, omdat dit de belangrijkste maat voor samenhang tussen twee variabelen is, en omdat deze in allerlei wetenschappen gebruikt wordt. In hoofdstuk 10 lieten we zien dat we ruwe scores zonder verlies aan informatie kunnen omzetten in *standaardscores*, en wel als volgt:

$$z_x = \frac{X_i - \overline{X}}{s_x}$$

Wanneer we nu de puntenwolk tekenen van twee variabelen, die beide uitgedrukt zijn in standaardscores, zouden we bij een zwakke correlatie tussen de twee zoiets als de figuur links kunnen krijgen, en bij een sterke correlatie de figuur rechts (zie figuur 12.4).



Figuur 12.4 Het principe van de correlatiecoëfficiënt

Omdat het gemiddelde van een in standaardscores uitgedrukte variabele gelijk is aan 0, gaat het midden van de puntenwolk door de oorsprong (het snijpunt van de horizontale en de verticale as).

In elk van de beide tekeningen staan *rechtsboven* de eenheden die groter zijn dan het gemiddelde op X, en ook groter zijn dan het gemiddelde op Y. Aangezien het gemiddelde van beide gelijk aan 0 is, betekent dit dat we rechtsboven de eenheden zien die positief zijn op zowel X als Y. *Linksboven* staan de eenheden die negatief zijn op X, en positief op Y, en *linksonder* de eenheden die op beide variabelen onder het gemiddelde liggen, dus negatief zijn. *Rechtsonder*, ten slotte, zijn de eenheden gesitueerd die op X positief zijn, maar op Y negatief.

Een verband wil zeggen dat bepaalde waarden op variabele X samengaan met bepaalde waarden op variabele Y. Bij samenhangen tussen twee variabelen op interval- of rationiveau hebben we vaak te maken met die vorm van samenhang waarbij hoge scores op X samengaan met hoge scores op Y, en lage scores op X met lage scores op Y. Een sterk verband wil dus zeggen dat de eenheden

een behoorlijk smalle puntenwolk vormen. Immers, bij een ronde puntenwolk gaan met hoge scores op X zowel hoge als lage scores op Y samen (ga na).

Welnu, een smalle puntenwolk komt erop neer dat er veel eenheden rechtsboven en linksonder liggen, en heel weinig linksboven en rechtsonder. Aan een dergelijke situatie willen we dus een hoge waarde van de correlatiecoëfficiënt toekennen. Daarom vermenigvuldigen we voor elke eenheid de z_x- met de z_y-score; we tellen de N productjes bij elkaar op en delen door N. Immers, rechtsboven en linksonder zijn de productjes positief (plus maal plus is plus, en min maal min is ook plus), en in de beide andere kwadranten zijn ze negatief (plus maal min is min). Zolang er meer eenheden worden gevonden in het kwadrant rechtsboven en linksonder samengenomen dan in het kwadrant linksboven en rechtsonder samengenomen, zijn er meer positieve productjes dan negatieve productjes (immers + x + en – x – is positief, en + x – en – x + is negatief).[2]

$$r = \frac{\sum z_x z_y}{N} \quad (-1 < r < 1)^3$$

Merk op dat de correlatiecoëfficiënt niet verandert als je een constante aan een of aan beide variabelen toevoegt (bijvoorbeeld door de scores 1, 2, 3, 4 en t te vervangen door 0, 1, 2, 3 en 4, of door alle scores met een constante te vermenigvuldigen (door bijvoorbeeld lengte in inches te vervangen door lengte in centimeters).
Als de positieve productjes tezamen geteld precies opwegen tegen de som van de negatieve productjes (bij een 'ronde' puntenwolk) is r = 0. Als alle punten daarentegen keurig op een rechte lijn van linksonder naar rechtsboven liggen, is r = 1. Maar meestal vertonen de punten enige spreiding rondom de lijn en ligt r tussen 0 en 1 in. En waar we het nog niet over hadden: als de puntenwolk van linksboven naar rechtsonder loopt, is de correlatie negatief; met als ondergrens -1.
In SPSS vind je r door STATISTICS aan te klikken.

Als je een klein aantal data-'punten' hebt, is het gemakkelijk om de correlatiecoëfficiënt met de hand uit te rekenen. Veronderstel dat we de volgende data hebben:

2 Deze opmerking is niet helemaal juist; het komt natuurlijk ook aan op de grootte van de productjes.

3 Sommige auteurs gebruiken (n-1) in de noemer in plaats van n, om statistische redenen die we hier niet verder uitleggen (zie hoofdstuk 11, noot 1). Bij een grote n is het verschil natuurlijk verwaarloosbaar, maar bij een n kleiner dan – laten we zeggen – 20 maakt het verschil.

X	Y	$(X_i - \bar{X})$	$(Y_i - \bar{Y})$	$(X_i - \bar{X})(Y_i - \bar{Y})$
3	5	-8	-3	24
5	7	-6	-1	6
10	4	-1	-4	4
14	11	3	3	9
17	8	6	0	0
17	13	6	5	30
SX = 66	SY = 48			S = 73

$\bar{X} = 66/6 = 11$ $\bar{Y} = 48 / 6 = 8$

$S_x^2 = \{(3-11)^2 + (5-11)^2 + \ldots\} / 6 = 30.3$ $S_x = \sqrt{30.3} = 5.5$

$S_y^2 = \{(5-8)^2 + (7-8)^2 + \ldots\} / 8 = 60$ $S_y = \sqrt{60} = 7.7$

$r_{xy} = 73 / (6 \times 5.5 \times 7.7) = .29$

Ga zorgvuldig na of deze berekeningen overeenkomen met de formules voor r en s.

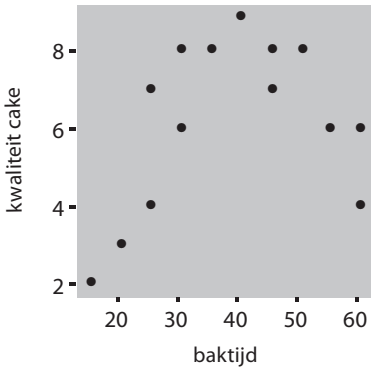
Het is mogelijk om een hele reeks van correlatiecoëfficiënten te berekenen in één SPSS-job. Als je alle onderlinge correlaties tussen vijf variabelen wilt weten, bestaat de output uit een 5x5-matrix waarin alle tien correlaties vermeld zijn,[4] plus het aantal eenheden waarop ze zijn berekend, plus de significantie van deze correlaties.

Is het berekenen van de r bij een kruistabel van twee intervalniveauvariabelen altijd zinvol? Nee; de r is gebaseerd op, zoals we zeiden, een specifieke samenhang, namelijk die samenhang waarbij de punten op en rondom een *rechte lijn* zijn geplaatst. Als er een *kromlijnig* verband is (zie figuur 12.5), is de r niet geschikt. Zouden we in zo'n geval ten onrechte de r berekenen, dan zal deze nooit erg hoog zijn, zelfs als alle waarnemingen keurig op de (kromme) lijn liggen. Ook hierom is het laten zien van de puntenwolk erg nuttig.

Een curvilineair verband zou kunnen ontstaan als je de kwaliteit van veertien cakes op een tienpuntsschaal laat beoordelen, afgezet tegen de baktijd in minuten. Als de baktijd te kort of te lang is, kun je verwachten dat de beoordeling laag is, terwijl als de baktijd het goede midden houdt, de beoordeling van de cakes heel goed uitvalt. Als we de correlatie zouden berekenen (hier is die .35), dan zien we de bijzondere vorm van de samenhang over het hoofd. Ook om deze reden is het altijd inspecteren van de plots van een bivariate samenhang een goede gewoonte. Een oplossing voor curvilineariteit is het transformeren

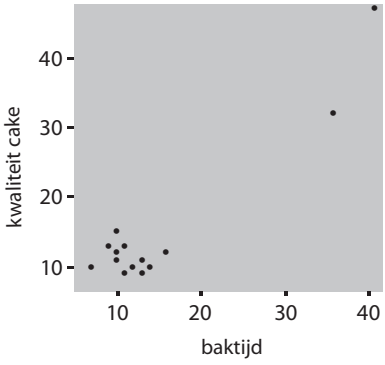
4 Een 5x5-matrix bevat natuurlijk 25 correlaties, maar die op de diagonaal zijn niet relevant, omdat de correlatie van een variabele met zichzelf altijd gelijk aan 1 is. Verder zijn de correlaties in de bovendriehoek uiteraard gelijk aan die in de onderdriehoek. Het aantal interessante correlaties is dus 1/2k(k-1) bij k variabelen.

van een van de variabelen. Je kunt verschillende transformaties, bijvoorbeeld machtsfuncties, uitproberen om een mooi lineair plaatje te bereiken.



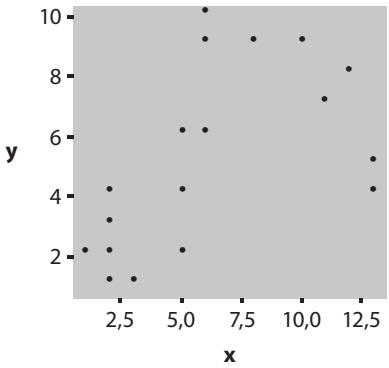
Figuur 12.5 Kromlijnige verbanden

Een *scatter plot* kan ons er door nog andere eigenaardigheden tegen waarschuwen om zonder verder nadenken een correlatiecoëfficiënt te berekenen. Figuur 12.6 laat een bijna ronde puntenwolk zien met echter een paar uitschieters. De aanwezigheid van deze uitschieters maakt dat de correlatiecoëfficiënt tamelijk hoog is, terwijl voor het overgrote deel van de waarnemingspunten de correlatie in de buurt van 0 ligt. Zonder de puntenwolk te zien zouden we dit niet weten. Een uitschieter kan van een lage correlatie een hoge maken, en andersom. Het kan zelfs voor een negatieve correlatie zorgen, terwijl er voor het merendeel van de punten een positieve correlatie is (of andersom). Reden te meer om een puntenwolk te tonen. Bij uitschieters is een mogelijke oplossing om ze weg te laten (maar er wel een 'mentale notitie' van te maken), en de correlatie over de overgebleven punten te berekenen. Figuur 12.6 toont een correlatie van .94; zonder de twee uitschieters is deze correlatie echter negatief (r = -.19).



Figuur 12.6 De invloed van uitschieters op een correlatiecoëfficiënt

En ten slotte kan uit een puntenwolk blijken dat er eigenlijk niet één wolk is, maar twee min of meer van elkaar gescheiden wolken, die elk apart ook nog een eigen verband laten zien. Dit zou erop kunnen wijzen dat de steekproef uit verschillende populaties, elk met een eigen karakter, is getrokken. Je zou de zaak in tweeën moeten splitsen, en de twee deelverzamelingen apart moeten behandelen (figuur 12.7).



Figuur 12.7 Verschillende subgroepen

De 'totaal'-correlatie is hier .59. Maar de puntenwolk toont dat er een subgroep is met een bijna nulcorrelatie, en een andere subgroep met een negatieve samenhang.

Regressie: een inleiding

Een *scatter plot* geeft ons een idee van de vorm van de puntenwolk van bivariate scores, en met correlatiecoëfficiënten kunnen we de sterkte van het verband op een precieze manier uitdrukken (tenminste als de puntenwolk min of meer lineair is). Het onderzoek van bivariate samenhangen dient nog een ander belangrijk doel: beter te kunnen voorspellen. Een *scatter plot* maakt het ons mogelijk om de score van een eenheid op variabele Y te voorspellen vanuit de score van die eenheid op variabele X, althans tot op zekere hoogte.

> Het lijkt een beetje vreemd dat het woord 'voorspellen' wordt gebruikt als we al over de precieze scores van een verzameling eenheden op zowel X als Y beschikken. Maar we moeten ons realiseren dat onze gegevens gewoonlijk berusten op een soms kleine steekproef. Voorspellen wordt relevant als we beschikken over een reeks scores op X en benieuwd zijn welke

score we dan kunnen verwachten op Y. Het gebruiken van de resultaten van een kleine steekproef op basis waarvan we een formule vinden om Y te schatten vanuit onze kennis van X, is dan aan de orde. Denk aan het voorspellen van het 'toegevoegde inkomen' bij een jaar meer onderwijs, de toename van de landbouwproductie bij gebruik van een bepaalde hoeveelheid kunstmest per hectare of aan het verminderen van relletjes door het inzetten van meer politie.

Om het principe van regressie te begrijpen kijken we naar figuur 12.4 rechts. Als we bij lage z_x-scores beginnen, zien we dat bij de laagste z_x-score verschillende Y-scores passen (althans, in principe: in de figuur is een beperkt aantal punten ingetekend). Als we vanuit deze z_x-score de bijpassende z_y -score willen voorspellen, kiezen we het gemiddelde van die z_y -scores voor dat z_x-punt, wat een acceptabel voorstel lijkt. Gaan we vervolgens meer naar rechts langs de x-as, dan komen we bij steeds hogere z_x-waarden. En als we bij elk van die z_x-waarden de bijpassende z_y-waarden bepalen, nemen we steeds het gemiddelde van de z_y-scores op het betreffende z_x-punt. We zien nu dat die gemiddelde z_y-waarden ook oplopen. Dat betekent dat kennis van de z_x-score van een eenheid een veel betere voorspelling mogelijk maakt van de z_y-waarde dan wanneer we niets van de z_x-score zouden weten. We kunnen echter niet precies voorspellen, omdat per z_x-waarde er nog verschillende eenheden zijn die een enigszins uiteenlopende z_y-score hebben. De voorspelling zou pas perfect kunnen zijn als alle punten precies op een rechte lijn zouden liggen. We zouden dan via de bekende formule Y = bX + a exact kunnen aangeven wat iemands Y-score is, gegeven een bekende X-score. En andersom, als we eens kijken naar figuur 12.4 links, dan zien we dat als je de gemiddelde z_y-scores vergelijkt over alle z_x-scores, die gemiddelden dankzij de 'ronde' puntenwolk heel weinig van elkaar verschillen. In zo'n geval – bij een lage correlatie dus – schieten we weinig op met onze kennis van de score op een variabele X om de score op Y te voorspellen.

Als we uitgaan van een lineair verband (de puntenwolk vertoont geen rare krommingen), kunnen we de vergelijking van een rechte lijn gebruiken (Y = bX + a) om Y te bepalen vanuit een bekende X-score. En bij standaardscores hebben we het dan nog gemakkelijker, omdat de lijn door de oorsprong gaat; het gemiddelde van alle gestandaardiseerde variabelen is gelijk aan 0. De lijn wordt als volgt gedefinieerd:

$$\hat{z}_y = b_{yx} z_x$$

Het dakje op de z_y betekent dat z_y slaat op een voorspelde score, die dus in het algemeen niet gelijk is aan de feitelijke z_y-score. ($\hat{Y}$ wordt in het Engels uitgesproken als *Y hat*.) De vraag is nu: wat is de waarde van b_{yx}? Zonder bewijs stellen we dat deze gelijk is aan de correlatiecoëfficiënt r_{xy}. Met andere woorden:

$$\hat{z}_y = rz_x$$

De lijn die hiermee gedefinieerd is, is de zogeheten kleinste kwadratensomlijn. Dat wil zeggen, dat de som van de gekwadrateerde afstandjes van de werkelijke Y-scores tot de gemiddelden van Y, voor elk van de betreffende X-punten, zo klein mogelijk is. En daar gaat het om: we willen onze voorspelling zo goed mogelijk maken.

We kijken nog eens nauwkeuriger naar figuur 12.4 rechts. Als we naar de hoogste z_x-scores gaan, zien we dat de bijbehorende gemiddelde z_y-scores wat lager zijn dan de z_x-scores. En als we in dezelfde figuur naar de lage z_x-scores gaan, schatten we terecht in dat de bijbehorende z_y-gemiddelden wat hoger zijn. Interessant is in dit verband de volgende historische noot, omdat daarmee verklaard wordt waar het begrip regressie vandaan komt:

> 'Sir Francis Galton related the height of sons to the heights of their fathers with a regression line. The slope of his line was less than 1. That is, sons of tall fathers were tall, but not as much above the average height as their fathers had been above their mean. Sons of short fathers were short, but generally not as far from their mean as their fathers. Galton interpreted the slope correctly as indicating a "regression" toward the mean height – and "regression" stuck as a description of the method he had used to find the line.' (De Veaux & Velleman, 2003)

En als je de hele redering nu eens zou willen omdraaien, en de z_x-score zou willen voorspellen vanuit de z_y-score? Wiskundig is dat natuurlijk ook mogelijk. Je gebruikt dan de formule $\hat{z}_x = b_{xy}z_y$. En we zouden tot een vergelijkbare conclusie komen: de voorspelde scores op X zijn wat minder extreem dan de scores op Y waarvan we uitgingen. Merk hierbij op dat de formule $\hat{z}_y = rz_x$ niet omgekeerd gebruikt kan worden om z_x te voorspellen vanuit z_y, zoals je misschien gedacht zou hebben; z_x is iets anders dan $\hat{z}_x$, en z_y is iets anders dan $\hat{z}_y$!

We zien dat er dus wiskundig altijd twee regressielijnen zijn: een voor het voorspellen van Y vanuit X, en de ander voor het voorspellen van de score op X van-

uit Y. Meestal is maar een van beide zinvol, omdat bij het berekenen van regressies we weten wat oorzaak en wat gevolg is.

Die regressielijnen (zoals we de 'voorspellingslijnen' voortaan zullen noemen) vormen een hoek met elkaar. Hoe ronder de puntenwolk is, des te groter de hoek die de regressielijnen met elkaar maken; in het uiterste geval, bij r = 0, is de een verticaal en de andere horizontaal. Alleen als r = 1 vallen de regressielijnen samen en – als we nog steeds van standaardscores uitgaan – maken ze een hoek van 45° met de x- en de y-as.

Tot zover bespraken we regressie in termen van standaardscores. Van meer praktisch belang lijkt de regressie te zijn als je uitgaat van de oorspronkelijke, ruwe scores. Zo willen we weten wat de gemiddelde inkomensgroei zal zijn in euro's als mensen een jaar meer opleiding krijgen; of we zijn geïnteresseerd in de toename van de opbrengst per hectare met 50 kg meer kunstmest.

Tabel 12.11 en de figuren 12.8 en 12.9 illustreren zulke problemen (Blalock, 1979, p. 394). Het gaat om de vraag of als je een aantal steden met elkaar vergelijkt, het inkomensverschil tussen blanke Amerikanen en Afro-Amerikanen samenhangt met, en te voorspellen valt uit, het percentage Afro-Amerikanen in die stad. De figuren 12.8 en 12.9 laten de *scattergrams* zien. In tegenstelling tot de *scattergrams* van variabelen die in standaardscores zijn uitgedrukt, laten deze zien dat de schalen langs de x-as (% Afro-Amerikanen) en de y-as (inkomensverschillen) heel verschillend zijn. Dit brengt met zich mee dat zelfs met een hoge correlatie de hoek tussen de regressielijnen niet in de buurt komt van een 45° tussen de x- en de y-as. Ook gaan de regressielijnen niet door de oorsprong (0,0), wat betekent dat we met een *intercept* (afstand tussen de oorsprong en het snijpunt van de lijn met de y-as) te maken hebben. We zien ook dat de twee lijnen een tamelijk grote hoek met elkaar vormen doordat de correlatie laag is (ongeveer .30).

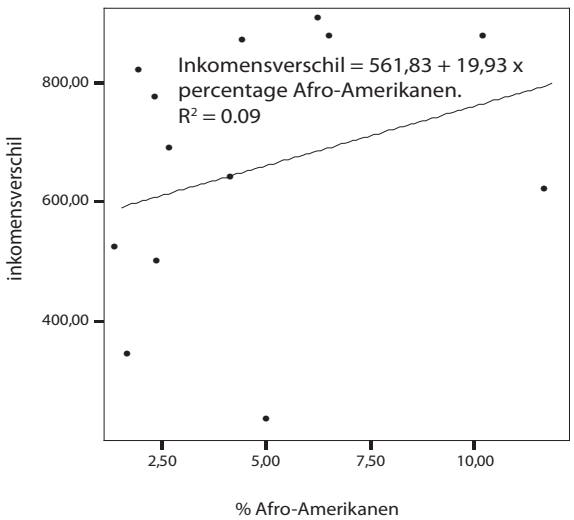
Als we nu uitgaan van de algemene vergelijking van een rechte lijn (Y = bX + a), wat is dan de best passende regressielijn, bijvoorbeeld die waarmee we Y voorspellen vanuit de score op X?
We definiëren de regressielijn (voor Y op X) weer als de lijn die de som van de gekwadrateerde verticale afstandjes van de waargenomen Y-scores tot de voorspelde punten op de lijn voor elke waarde van X minimaliseert. Weer zonder toelichting de formules:

$$a_{yx} = \bar{Y} - b_{yx}\bar{X}$$

$$b_{yx} = \frac{\Sigma (X_i - \bar{X})(Y_i - \bar{Y})}{\Sigma (X_i - \bar{X})^2}$$

Tabel 12.11 *Percentage Afro-Amerikanen, en gemiddelde inkomensverschillen tussen Afro-Amerikanen en anderen, in dertien steden in het middenwesten van de VS*

	% Afro-Amerikanen	Inkomensverschil
1	2.13	809,00
2	2.52	763,00
3	12.86	612,00
4	2.55	492,00
5	2.87	679,00
6	4.23	635,00
7	4.62	859,00
8	5.19	228,00
9	6.43	897,00
10	6.70	867,00
11	1.53	513,00
12	1.87	335,00
13	10.38	868,00



Figuur 12.8 *Inkomensverschil naar % Afro-Amerikanen (dertien steden)*

In dit voorbeeld is: b_{yx} = 19.931; a_{yx} = 561.83. Met andere woorden, de formule om 'inkomensverschil' te voorspellen vanuit het 'percentage Afro-Amerikanen in de staat' is:

$\hat{I}$ = 19.931 x Percentage Afro-Amerikanen + 561.83

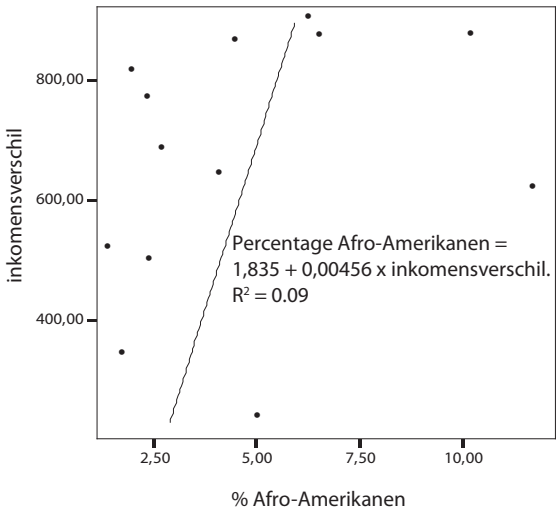
De parameter b_{yx} wordt de regressiecoëfficiënt genoemd (let er goed op dat de afhankelijke variabele, hier Y, altijd het eerst genoemd wordt in het subscript). Parameter 'a' wordt het *intercept* genoemd. 'a_{yx}' is de afstand langs de y-as tussen het snijpunt van de regressielijn met de y-as en de oorsprong (0,0). Check in figuur 12.8 dat deze waarden min of meer correct zijn.

Regressielijnen, en meer in het bijzonder hun vergelijkingen, zijn voorbeelden van datamodellen. Het gebruiken van een rechte lijn om scores te voorspellen die niet op die lijn liggen, maar in de buurt van die lijn, betekent dat de lijn niet meer dan een benadering is. Elke werkelijke score kan worden gesplitst in een deel dat je kunt voorspellen door gebruik te maken van die lijn, en een deel dat je niet kunt voorspellen: het residu. Daarom kan ook de totale variantie van de scores worden gesplitst in een deel dat je kunt 'verklaren' via het model, en een deel dat je niet kunt verklaren. De residuen vertonen normaliter niet een bepaald patroon. Als dat wel zo is, is een derde variabele of een combinatie van verschillende variabelen daarvoor verantwoordelijk.

Veronderstel nu dat we, omgekeerd, ook de andere regressielijn, die van het percentage Afro-Amerikanen op 'inkomensverschil' willen weten. Het gebruik van dezelfde formule leidt nu tot:

b_{xy} = .00456, en a_{xy} = 1.835

Het voorspelde 'percentage Afro-Amerikanen' = 0.00456 x Inkomensverschil + 1.835.
De factor a_{xy} stelt het intercept voor: de afstand langs de x-as tussen het snijpunt met de regressielijn en de oorsprong.



Figuur 12.9 *Percentage Afro-Amerikanen naar inkomensverschil (dertien steden)*

Ten slotte: de formules voor de regressiecoëfficiënten en de correlatiecoëfficiënt zijn nauw verwant. Om te beginnen hebben ze alle drie dezelfde teller (de zogeheten *covariantie*). In de noemers vinden we respectievelijk de variantie van X, de variantie van Y en de wortel van het product van deze twee. Daarom is de correlatiecoëfficiënt ook de wortel uit het product van de beide regressiecoëfficiënten. De formules zijn in ruwe scores de volgende:

$$b_{xy} = \frac{\Sigma (X - \bar{X})(Y - \bar{Y})}{\Sigma (X - \bar{X})^2}$$

$$b_{xy} = \frac{\Sigma (X - \bar{X})(Y - \bar{Y})}{\Sigma (Y - \bar{Y})^2}$$

$$r = \frac{\Sigma (X - \bar{X})(Y - \bar{Y})}{\sqrt{\Sigma (X - X)^2 \, \Sigma (Y - Y)^2}}$$

Terwijl het verschijnsel 'correlatie' of 'samenhang' een symmetrische zaak is, is 'regressie' asymmetrisch. Correlatie slaat op de samenhang, regressie op het voorspellen van een score op de ene variabele vanuit een bekende score op de andere variabele, de grootte van de verandering in een variabele als gevolg van

een verandering in de andere variabele. Regressie wordt dan ook met name toegepast bij de analyse van causale modellen.

Omdat regressiecoëfficiënten sterk afhankelijk zijn van de schaal waarin de metingen zijn uitgedrukt (meting van dezelfde objecten in millimeters, centimeters of inches geeft drie verschillende regressiecoëfficiënten), worden deze niet-gestandaardiseerde regressiecoëfficiënten vaak vervangen door gestandaardiseerde coëfficiënten, gebaseerd op de ruwe scores, wat betekent dat elke variabele een gemiddelde heeft van 0 en een standaardafwijking van 1. Niet-gestandaardiseerde regressiecoëfficiënten worden aangeduid met de letter b, gestandaardiseerde met de letter ß. Het gebruik van gestandaardiseerde coëfficiënten is eigenlijk alleen nuttig als je te maken hebt met multipele regressie, dat wil zeggen dat je werkt met verschillende onafhankelijke variabelen tegelijk. Het stelt je in staat om de grootte van de effecten van deze variabelen met elkaar te vergelijken.

Samenvattend: als je te maken hebt met variabelen gemeten op interval- en rationiveau, die redelijk lineair samenhangen, beschik je over een prima techniek om de score op een afhankelijke variabele te voorspellen gegeven de score op één of meer onafhankelijke variabelen. We berekenen de noodzakelijke regressievergelijkingen op basis van een steekproef, en we gebruiken deze vergelijkingen voor grotere domeinen waar de score op de ene variabele bekend is en die op de andere niet.

Oefeningen

Oefening 12.1

Hoeveel randverdelingen, hoeveel conditionele verdelingen en hoeveel cellen zijn er in een r x k-tabel?

Oefening 12.2

a. Bereken de verwachte waarden in beide tabellen hierna. Ga na dat de verhoudingen in de kolommen per tabel gelijk zijn.

		X					X			
		1	2				1	2	3	
	1	..	..	60		1	..	..	..	130
Y	2	..	..	90	Y	2	..	..	..	65
						3	..	..	..	39
		100	50	150			72	108	54	234

b. Elke wijziging in een van de cellen leidt tot een samenhang. Maar de samenhang kan variëren van heel zwak tot heel sterk. Zo kunnen we in de linkertabel een zwakke samenhang creëren door maar een klein beetje van de situatie van onafhankelijkheid af te wijken (hierna, links) en een heel sterke samenhang door te streven naar minstens één cel waarin de vulling o is (hierna, rechts).

		X				X		
		1	2			1	2	
Y	1	42	18	60	Y	10	50	60
	2	58	32	90		90	..	90
		100	50	150		100	50	150

Verander ook de vullingen in de 3x3-tabel zodanig dat (a) een zwakke samenhang ontstaat, en (b) een sterke samenhang. Let erop dat de randverdelingen steeds hetzelfde blijven. Bereken in beide gevallen de residuen.

Oefening 12.3 Gegeven de volgende kruistabel:

	Digitale tv	
	Ja	Nee
Gemeente A	100.000	10.000
Gemeente B	125.000	8.000
Gemeente C	20.000	1.000

a. Is de opstelling van de tabel in overeenstemming met de regels? Stel de tabel, inclusief de kop, beter op.
b. Indien de onderzoeker de gemeenten met elkaar wil vergelijken op het percentage dat een digitale tv-aansluiting heeft, hoe moet hij dan percenteren?
c. Is er een statistisch verband tussen beide variabelen?
d. Geef een samenvattende uitspraak op basis van de tabel.
e. Geef eveneens een samenvattende uitspraak als je de andere kant op percenteert.

Oefening 12.4 Gegeven de volgende kruistabel.

Delinquenten en niet-delinquenten in twee gemeenten A en B, 1990 (bron: fictief)

	A	B
Delinquenten	205	205
Niet-delinquenten	4.800	10.800

a. Iemand doet de uitspraak: in A zijn er evenveel delinquenten als in B. Is die uitspraak juist? Wat wordt waarschijnlijk bedoeld? Is hij dan juist?
b. Percenteer verticaal en geef een vergelijkende uitspraak.
c. Percenteer horizontaal en geef een vergelijkende uitspraak.
d. Welke van deze uitspraken vind je het nuttigst?

Oefening 12.5

Gegeven het volgende stukje tekst in een rapport:

'Met de rechter in aanraking gekomen ongehuwde vrouwen.

Leeftijd	Frequentie	Percentage
9 - 19 jaar	155	85.7
20 - 24 jaar	25	13.8
25 jaar en ouder	1	0.5
	181	100,0

Deze tabel toont aan dat de criminaliteit in de bestudeerde bevolkingsgroep omgekeerd evenredig is met de leeftijd. De grootste misdadigheid treft men aan onder de jongsten.'
Geef met behulp van de tabel je mening over cijfers en commentaar.

Oefening 12.6

Gegeven de kruistabel hierna.

Inkomensklasse naar sekse en leeftijd (fictieve data)

	Man		Vrouw	
	Leeftijd		Leeftijd	
	<30 jr	>30 jr	<30 jr	>30 jr
Meer dan € 3.000	5	45	25	40
€ 1.000 tot € 3.000	15	30	75	20
Minder dan € 1.000	40	15	25	15

a. Hoeveel ingangen heeft deze tabel?
b. Maak de kruistabellen van sekse en inkomen, van leeftijd en inkomen, en van sekse en leeftijd.
c. Geef de frequentieverdelingen van elk van de variabelen.
d. Trek een in goed Nederlands geformuleerde conclusie naar aanleiding van elk van de kruistabellen.

Oefening 12.7

In een onderzoek in textielbedrijven werd onder het personeel een enquête gehouden. Een van de vragen luidde: 'Zou u liever ander werk doen dan u nu doet?' Het antwoord op de vraag kon zijn: 'nee', 'ja' of 'weet niet'. In totaal wer-

den 1250 personen geïnterviewd. Het materiaal werd gesplitst naar leeftijdsklassen: 18-25 jaar, 26-35 jaar, 36-45 jaar en 46 jaar en ouder. De mannen waren als volgt over de klassen verdeeld: 150, 250, 200 en 400. De vrouwen: 50, 90, 50 en 60.
Van de mannen in de leeftijdsklasse 18-25 antwoordde 36% bevestigend en 60% ontkennend; van de vrouwen was dit respectievelijk 8% en 92%. In de leeftijdsklasse 26-34 antwoordden 60% van de mannen en 80% van de vrouwen ontkennend, en 30% respectievelijk 20% bevestigend. In de leeftijdsklasse 35-44 waren deze getallen respectievelijk 70%, 70%, 29% en 20%. In de hoogste leeftijdsklasse antwoordde van de vrouwen 20% bevestigend en 75% ontkennend, en van de mannen antwoordde 76% ontkennend en 23% bevestigend.
Breng deze gegevens onder in een kruistabel in absolute aantallen. Denk ook om een goede kop.
Trek uit deze opdracht ook de conclusie dat het rapporteren van de oorspronkelijke tabel vaak verstandiger is dan alleen maar een aantal conclusies geven (in kranten lezen we vaak alleen enkele conclusies)!

Literatuurtips

Regressie:

Blalock, H.M. (1979). *Social statistics* (2nd ed.). Tokyo: McGraw-Hill Kogakusha.

Veaux, R.D. de & Velleman, P.F. (2003). *Intro Stats*. Boston: Addison Wesley.

Kernbegrippen

kruistabel
samenhang
epsilon
odds ratio
log-odds ratio
Kendall's Q
onafhankelijkheid van variabelen
maximaliseren van een verband
chi-kwadraat
correlatiecoëfficiënt
tabelsplitsing
partiële correlatie
predictie in statistische zin
regressie
regressiecoëfficiënt
intercept
gestandaardiseerde regressie
ongestandaardiseerde regressie