

Statistiek en meten: wat moet je daarover weten?

Onder redactie van Diana Dolmans, Cees van der Vleuten, Albert Scherpbier en Ineke Wolfhagen

Bij het lezen van publikaties over onderzoek van onderwijs worden lezers regelmatig geconfronteerd met allerlei statistische begrippen, waarvan de betekenis niet geheel duidelijk is. Daarom heeft de redactie besloten om een reeks artikelen te publiceren waarin belangrijke onderwerpen op het gebied van statistiek en meten aan de orde worden gesteld. De redactie van deze reeks wordt gevormd door twee gastredacteuren (Diana Dolmans en Cees van der Vleuten) en twee leden van de BMO-redactie (Albert Scherpbier en Ineke Wolfhagen).

De keuze van onderwerpen is gebaseerd op veel gebruikte statistische technieken en begrippen. De nadruk ligt op de betekenis en interpretatie hiervan. Op deze manier wordt getracht een bijdrage te leveren aan de verdere professionalisering van docenten binnen het medisch onderwijs. In het derde artikel van deze reeks staat associatie en correlatie centraal.

Associatie en correlatie

J.A. Smal

Termen die aan bod komen:

puntenwolk of scattergram, curvilineair, uitbijter of outlier, correlatiecoëfficiënt, regressie-analyse, multiple regressie-analyse, causaliteit, significant, percentage verklaarde variantie, associatiematen

Onderzoek richt zich dikwijls op de vraag of er een verband bestaat tussen verschillende variabelen. Studenten die goed zijn in het ene vak, behalen vaak ook hoge cijfers in andere vakken, en omgekeerd. Deze verschillen in cijfers kunnen weer samenhangen met verschillen in intelligentie, frequentie van collegebezoek, aantal uren slapen of studeren. De mate waarin twee variabelen samenhangen kan worden uitgedrukt in een maat voor associatie of correlatie.

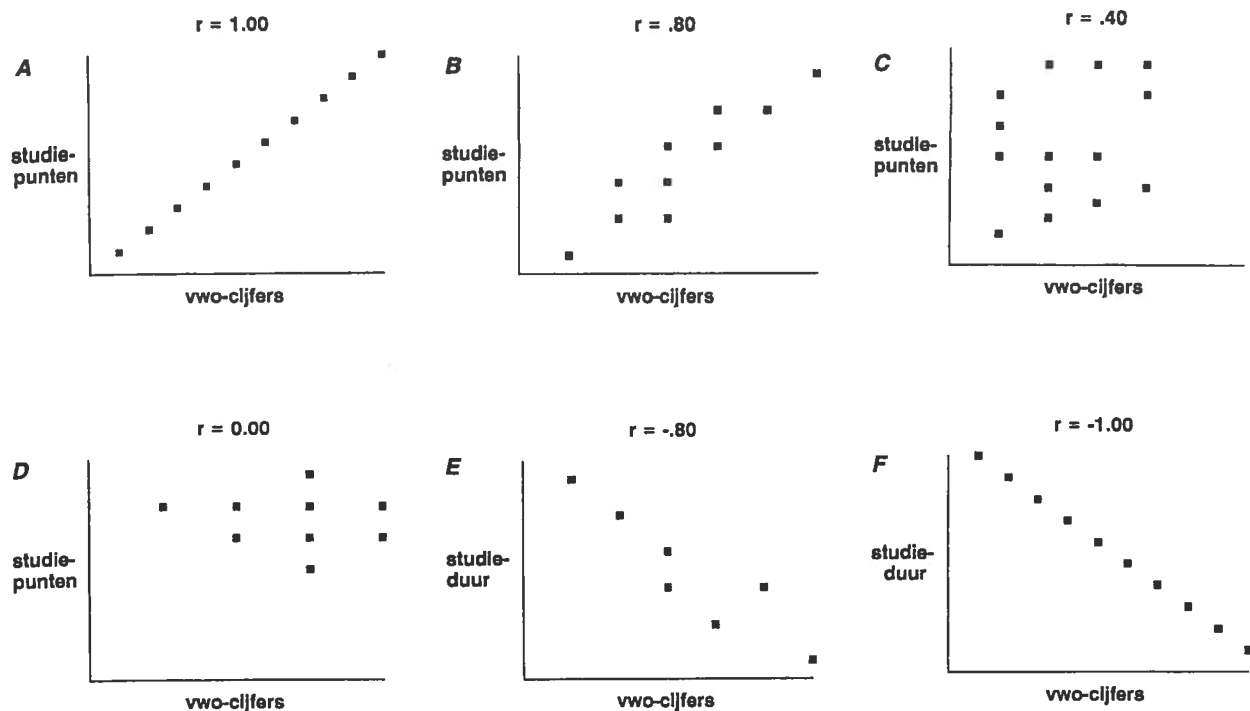
Correlatie

Een propaedeusecommissie wil uitzoeken of er een verband bestaat tussen de cijfers behaald

op het Voorbereidend Wetenschappelijk Onderwijs (VWO) en de behaalde studiepunten na een jaar. Om hiervan een beeld te krijgen kan men de resultaten van studenten uitzetten in een grafiek. Elke student wordt weergegeven met een punt. De plaats op de X-as correspondeert met het VWO-cijfer en de plaats op de Y-as correspondeert met de studiepunten. Als de resultaten van een grote groep studenten op deze wijze worden weergegeven ontstaat een verzameling punten, meestal aangeduid als *puntenwolk* of *scattergram*.

Deze puntenwolk kan verschillende vormen hebben. Het kan een perfecte rechte lijn zijn, een meer of minder langgerekte ellips of een cirkel (figuur 1).

De vorm van deze puntenwolk kan worden samengevat in een getal, dat met de daarvoor ontwikkelde formules wordt berekend: de *correlatiecoëfficiënt*. De meest gebruikte maat is de Pearson produkt-moment correlatiecoëfficiënt, uitgedrukt als r . Deze kan variëren van + 1.00 tot -1.00. Als de punten precies op een



Figuur 1. Het verband tussen twee variabelen, weergegeven als puntenwolken met bijbehorende correlatiecoëfficiënten

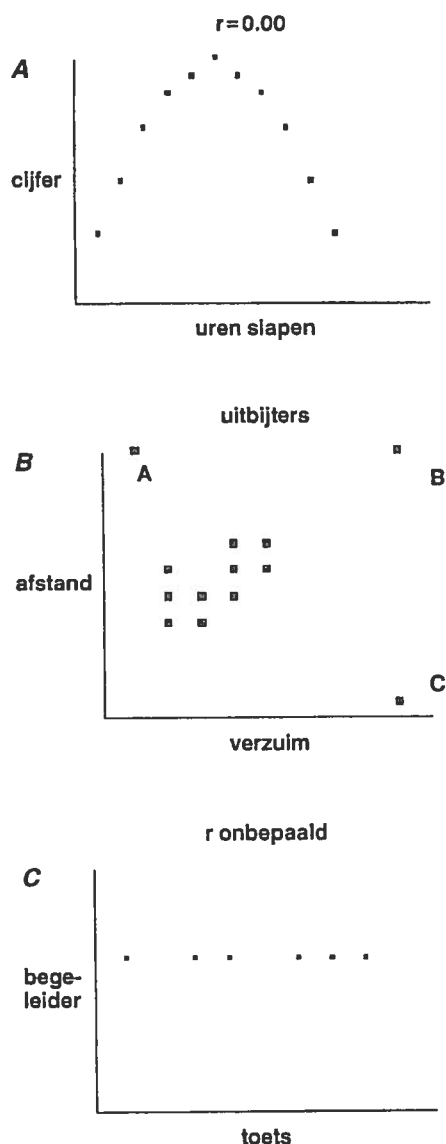
diagonale rechte lijn liggen is de correlatie $+1.00$ (figuur 1a) of -1.00 (figuur 1f). In genoemd voorbeeld zou er dan een perfect verband bestaan tussen de hoogte van de VWO-cijfers en de studiepunten in de propaedeuse. Dergelijke hoge correlaties komen in onderwijskundig onderzoek weinig voor. Wel bijvoorbeeld bij het ijken van instrumenten. Een laboratorium behoort een correlatie van $r=+1.00$ te vinden tussen de temperatuur gemeten met een digitale thermometer en een klasieke kwikthermometer.

Indien een langwerpige ellips gevonden wordt, bijvoorbeeld $r=0.80$ (figuur 1b), bestaat er weliswaar geen perfect, maar nog wel een sterk verband tussen VWO-cijfers en behaalde studiepunten; studenten met hoge VWO-cijfers halen in deze situatie meestal ook meer studiepunten. Dit verband is minder sterk bij een correlatie van $r=0.40$ (figuur 1c) en verdwijnt geheel als de puntenwolk een bolvorm vertoont. In het laatste geval is er geen enkel verband tussen VWO-cijfers en studiepunten: de correlatie is 0.00 (figuur 1d). De correlatie kan ook negatief zijn. Een negatieve correlatie

is plausibel bij een vergelijking van de VWO-cijfers met de studieduur: hogere VWO-cijfers corresponderen met een korte studieduur en lage VWO-cijfers met een lange studieduur (figuur 1e en 1f). De berekening van een correlatiecoëfficiënt is op zichzelf niet ingewikkeld, maar wel tijdrovend. Daarom gebruikt men hiervoor meestal een computer.

Bijzondere puntenwolken

Het is zeer inzichtelijk en nuttig om bij een reeks gegevens waarvan men de correlatie wil uitrekenen eerst een puntenwolk of scattergram te (laten) tekenen. Zo'n tekening heeft twee functies. Allereerst krijgt men een idee of het lineair model waarop de correlatie-analyse gebaseerd is, wel van toepassing is. De correlatiecoëfficiënt is namelijk te beschouwen als een indicatie voor de mate waarin de vorm van de puntenwolk een rechte lijn benadert. Uit het scattergram kan echter blijken dat een *curvilinear* model, bijvoorbeeld een parabool of hyperbool, misschien veel adequater is om het verband tussen de variabelen te beschrijven.



Figuur 2. Voorbeelden van bijzondere puntenwolken: (a) een curvilineair verband, (b) de invloed van uitbijters, (c) een situatie waarin slechts bij één variabele spreiding is

Een dergelijk kromlijinig verband - een omgekeerde U - vonden wij bijvoorbeeld tussen het aantal uren slapen en de behaalde cijfers van een groep studenten (figuur 2a). De studenten die gewoonlijk zo'n 7,5 uur per etmaal slapen, bleken de hoogste cijfers te behalen.¹ Zowel de studenten die minder uren slapen als de studenten die meer uren slapen, behaalden lagere cijfers. Hoewel er wel degelijk een verband bestond tussen slapen en cijfers, zou de correlatiecoëfficiënt laag uitvallen. Deze is dus geen

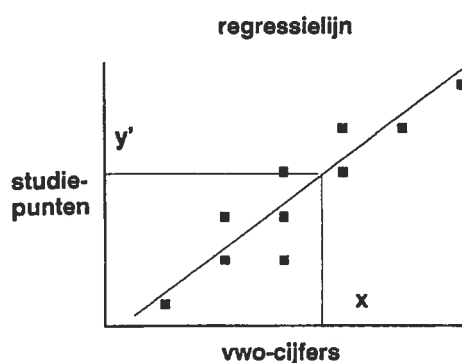
geschikte maat wanneer het gezochte verband niet lineair is.

Een tweede functie van een scattergram is dat het snel een beeld geeft van afwijkende punten op onwaarschijnlijke plaatsen. Dergelijke 'uitbijters' of 'outliers' kunnen een correlatie zeer sterk beïnvloeden. Figuur 2b kan dit effect illustreren. De oorspronkelijke puntenwolk - zonder de punten A, B of C - correspondeert met een correlatie van $r=0.70$. Door toevoeging van een enkele uitbijter, hetzij A, B of C, wordt de correlatie respectievelijk $r=0.30$, $r=0.90$ of $r=0.10$. Het is dus belangrijk om enige aandacht hieraan te besteden. Soms blijkt zo'n uitbijter een typefout te zijn, die eerst hersteld moet worden. Soms betreft het een uitzonderlijk geval, zoals in het volgende voorbeeld.

Een bedrijfsarts in het midden van het land testte de hypothese dat er een positieve correlatie bestond tussen ziekteverzuim en woon-werkafstand. Hij verwachtte dat mensen die ver weg woonden, vaker verzuimden dan mensen die vlak bij het werk woonden. In zijn scattergram kwam echter één persoon naar voren die heel ver weg woonde, maar nooit verzuimde. Dit bleek de directeur te zijn, die in Eelde woonde en dagelijks met een privévliegtuig naar het werk kwam!

In zo'n geval kan een onderzoeker een zuiverder beeld van het verband krijgen door deze uitbijters niet in de berekeningen te betrekken. Het is natuurlijk wel noodzakelijk om zo'n ingreep te verantwoorden.

Er zijn ook situaties waarin de puntenwolk de vorm van een horizontale (of verticale) lijn heeft. Dit patroon komt bijvoorbeeld tevoorschijn bij een vergelijking van de cijfers die begeleiders geven aan co-assistenten met de scores op een objectieve kennistoets. In de cijfers van de docent-begeleiders komt nauwelijks spreiding voor: zij geven hun 'toekomstige collega's' bijna altijd een 8. De toetsresultaten daarentegen variëren wél (figuur 2c). In dit geval kan er geen correlatie berekend worden. Voor het berekenen van de correlatie tus-



Figuur 3. Met behulp van de regressielijn kan y' voorspeld worden uit x

sen twee variabelen is het noodzakelijk dat beide variabelen een spreiding hebben.

De hoogste correlatie is te verwachten als de puntenwolk de vorm van een diagonale lijn benadert. De exacte richting is niet belangrijk. De richting van de puntenwolk is afhankelijk van de toevallig gekozen schaalverdeling op de x - en y -as, en heeft geen invloed op de hoogte van de correlatiecoëfficiënt. Als men bijvoorbeeld alle gegevens met hetzelfde getal vermenigvuldigt of bij alle gegevens hetzelfde getal optelt of aftrekt, verandert de hoogte van de correlatie niet.

Regressie

Als er een correlatie bestaat tussen twee variabelen, bijvoorbeeld tussen VWO-cijfers en studiesucces, kan men hiermee uit het ene gegeven het andere voorspellen. De studieadviseur kan dan bijvoorbeeld extra begeleiding gaan geven aan studenten die waarschijnlijk hun propaedeuse niet zullen halen. Hiertoe past men *regressie-analyse* toe. Men berekent welke lijn door de puntenwolk de minste fouten in de voorspelling oplevert: de regressielijn (figuur 3). De formule hiervan luidt in standaardcores $z'_y = r z_x$ (voor de betekenis van standaardcores zie de Z-toets in het artikel van Van Breukelen).² Hieruit is een formule of te leiden van de vorm $y' = ax + b$, waarin x het behaalde VWO-cijfer is en y' het voorspelde

aantal studiepunten. Als er tussen beide variabelen een hoge correlatie bestaat, zal het aantal fouten in de voorspellingen klein zijn, bij een lage correlatie zullen veel voorspellingen er ver naast zijn.

Er zijn situaties waarin wij beschikken over meerdere voorspellers. Het studiesucces in de propaedeuse is misschien ook, of beter, te voorspellen met een selectie uit de VWO-cijfers, of de resultaten van de tentamens in het eerste trimester, of andere variabelen zoals leeftijd en sekse. Via *multiple regressie-analyse* kan - gewoonlijk met een computer - een regressievergelijking opgesteld worden, waarin verschillende voorspellers zijn opgenomen. Voor elke voorspeller wordt een gewicht berekend. Dit gewicht is afhankelijk van de correlatie tussen deze voorspeller met de te voorspellen variabele en van de extra informatie die deze voorspeller toevoegt ten opzichte van de andere voorspellers. Het is bijvoorbeeld mogelijk dat zowel wiskunde- als natuurkundecijfers hoog correleren met propaedeuseresultaten en dus beide in aanmerking komen als voorspeller. Als de cijfers voor deze twee vakken onderling hoog correleren, voegt het tweede vak weinig toe en krijgt in de regressievergelijking een laag gewicht.

Regressie-analyse is gericht op het doen van voorspellingen. Dit houdt impliciet in dat er een volgorde is: de ene variabele (bijvoorbeeld VWO-cijfers) gaat in de tijd vooraf aan de andere (bijvoorbeeld propaedeusecijfers). Men noemt deze ook de onafhankelijke, respectievelijk afhankelijke variabele. Bij het werken met correlaties speelt dit onderscheid niet. Het begrip correlatie is louter een maat voor de samenhang van twee variabelen, zonder dat er gesproken wordt van afhankelijke of onafhankelijke variabelen.

Causaliteit

Correlaties worden gewoonlijk berekend om een verband tussen variabelen op te sporen of een hypothetisch verband te toetsen. Bijvoor-

beeld: is er verband tussen collegebezoek en tentamencijfers? Een van de grootste valkuilen bij het werken met correlaties is de aanname van *causaliteit*. De regel is simpel: met een correlatie kan nooit een causaal verband worden aangetoond. Het klassieke voorbeeld hierbij is het verhaal van de ooievaars. In Zweden - zo wordt verteld - constateerde men in het begin van deze eeuw een duidelijke correlatie tussen het aantal aanwezige ooievaars en het aantal geboren baby's. Het zou echter onjuist zijn om op grond van de correlatie te besluiten dat het kleinere aantal ooievaars de oorzaak was van de geboortedaling. Beide variabelen zijn waarschijnlijk afhankelijk van een derde variabele (interveniërende variabele), namelijk de toenemende verstedelijking, die zowel leidde tot kleinere gezinnen als tot minder ooievaars.

Ook bij een positieve correlatie tussen collegebezoek en tentamencijfers is het verleidelijk om te besluiten dat het collegebezoek de oorzaak is van betere studieprestaties. De omgekeerde verklaring is namelijk ook te verdedigen. Er is in onderzoek bijvoorbeeld gevonden dat studenten die op de middelbare school hoge cijfers behaalden, later op de universiteit vaker college liepen.³ Misschien hebben de studenten met hoge cijfers meer tijd over en kunnen zij zich de luxe permitteren om colleges te volgen. De studenten met slechte resultaten volgen misschien minder colleges omdat zij zich voorbereiden op de herkansingen die zij nog moeten doen.

De interpretatie van een correlatie

Het blijkt lastig om de hoogte van een correlatiecoëfficiënt goed te interpreteren. Dit komt omdat de hoogte van de correlatie niet proportioneel verdeeld is. Een coëfficiënt $r=0.60$ wil niet zeggen dat het gevonden verband twee keer zo sterk is als bij een coëfficiënt $r=0.30$. Ook is het verschil tussen coëfficiënten van $r=0.40$ en $r=0.50$ niet gelijk aan het verschil tussen coëfficiënten $r=0.50$ en $r=0.60$.

Er zijn geen vaste regels te geven hoe hoog een correlatie moet zijn om van betekenis geacht te worden. Een correlatie van $r=0.90$ wordt gewoonlijk hoog genoemd en een waarde van $r=0.10$ laag. De betekenis van de tussenliggende waarden is sterk afhankelijk van de gebruiker. Volgens Feinsein worden in de medische literatuur correlatiecoëfficiënten van $r=0.50$ of hoger als heel goed beschouwd,⁴ maar bij epidemiologisch of sociologisch onderzoek zouden onderzoekers al juichen bij een correlatie van $r=0.13$.

Het is gebruikelijk te vermelden of de correlatie *significant* is. Hiertoe berekent men de kans dat een dergelijke of hogere coëfficiënt bij toeval gevonden zou worden, ook als er in werkelijkheid geen verband tussen twee variabelen zou bestaan en de 'werkelijke' correlatie $r=0.00$ is. De significantie van een correlatiecoëfficiënt is echter voornamelijk afhankelijk van het aantal waarnemingen en het gekozen significantieniveau. Een correlatie van $r=0.07$ is laag en nauwelijks groter dan $r=0.00$, maar bij 1000 personen kan ook deze lage coëfficiënt significant zijn, terwijl de relevantie te verwaarlozen is.

Een beter beeld van de sterkte van een verband krijgt men door de correlatiecoëfficiënt te kwadrateren. Dit getal, r^2 , kan wél opgevat worden als een proportie. Of als een percentage wanneer men r^2 met 100 vermenigvuldigt, zoals gebruikelijk is.

Wanneer men de correlatie tussen twee variabelen berekent, geeft het kwadraat van de correlatiecoëfficiënt aan welk deel van de variantie in de ene variabele te verklaren is uit de andere variabele, het zogenaamde *percentage verklaarde variantie*. Als er tussen de cijfers voor twee vakken, bijvoorbeeld anatomie en fysiologie, een correlatie bestaat van $r=.80$, dan wordt $0.80 \times 0.80 \times 100\% = 64\%$ van de variantie in de fysiologiecijfers verklaard door het anatomiecijfer. Anders geformuleerd: onze onzekerheid over de cijfers die de studenten voor fysiologie behalen, wordt met 64% gereduceerd als wij de anatomiecijfers kennen.

Tabel 1. Correlatiecoëfficiënten voor verschillende meetniveaus

Meetniveau	Coëfficiënt
Ratio/interval	Pearson produkt-moment correlatie
Ordinaal	Spearman rangcorrelatie
	Kendall rangcorrelatie
Nominaal	Contingentie coëfficiënt
	Cramers' V
	ϕ (Phi)
	Kappa

Andere correlatiematen

Er bestaan verschillende correlatiematen voor de verschillende meetniveaus: nominaal, ordinaal, interval- of ratio-niveau (voor de betekenis van de meetniveaus zie Dolmans).⁵ Bij onderzoek naar het verband tussen twee variabelen die beide op interval- of ratio-niveau gemeten zijn, gebruikt men gewoonlijk de Pearson produkt-moment correlatie (tabel 1). Bijvoorbeeld het verband tussen tentamencijfers en collegebezoek. Bij gegevens op ordinaal niveau is de rangcorrelatie van Spearman toepasbaar of de rangcorrelatie van Kendall. Bijvoorbeeld een onderzoek naar het verband tussen sociale klasse en de hoogte van de schoolopleiding die men kiest. Voor gegevens op nominaal niveau bestaan er eveneens verschillende maten, bijvoorbeeld Cramers V of coëfficiënt ϕ (phi). De meeste computerprogramma's bieden de mogelijkheid om verschillende van deze maten te berekenen. De onderzoeker moet afhankelijk van het meetniveau beslissen welke coëfficiënt het meest adequaat is. Elk van deze maten heeft overigens weer eigen beperkingen en een vergelijking van de verschillende coëfficiënten is niet goed mogelijk. Men raadplege hiervoor uitgebreide handboeken statistiek.

Terminologie

De terminologie op het gebied van maten voor het verband tussen twee variabelen is niet altijd scherp gedefinieerd. In het dagelijkse spraak-

gebruik wordt het woord correlatie vaak gebruikt als synoniem voor 'verband' of 'relatie'. Het is niet ongewoon iemand te horen zeggen: er is een correlatie tussen geslacht en levensverwachting. Een statisticus denkt hier niet direct aan een correlatie, maar aan een verschil tussen de gemiddelde levensverwachting van mannen en vrouwen. Het woord correlatie wordt gewoonlijk gereserveerd voor het beschrijven van een lineair verband tussen variabelen van ordinaal, ratio- of intervalniveau. Op het nominale meetniveau spreekt men eerder van *associatiematen* of contingenties. Maar ook in statistische handboeken blijken termen als correlatie, associatie, regressie naast en door elkaar gebruikt te worden.

Literatuur

1. Ritskes RR, Smal JA. Slapen of studeren: verband tussen slaapgedrag en studieprestaties medische studenten onderzocht. *Medisch Contact* 1986; 29: 915-7.
2. Van Breukelen G. Statistische toetsen en betrouwbaarheidsintervallen. Reeks Statistiek en meten: wat moet je daarover weten? *Bulletin Medisch Onderwijs* 1993; 12(2): 73-83.
3. Sade RM, Stroud MR. Medical students attendance at lectures: effects on medical school performance. *Journal of Medical Education* 1982; 57: 191-2.
4. Feinstein AR. *Clinical epidemiology: the architecture of clinical research*. Philadelphia: Saunders, 1985.
5. Dolmans D. Beschrijvende statistiek. Reeks Statistiek en meten: wat moet je daarover weten? *Bulletin Medisch Onderwijs* 1993; 12(1): 27-32.

Aanbevolen literatuur

- Feinstein AR. *Clinical epidemiology: the architecture of clinical research*. Philadelphia: Saunders, 1985.
- Nijdam B, Van Buuren H. *Statistiek voor de sociale wetenschappen*. Deel 1. Beschrijvende statistiek. Alphen: Samson, 1988.

Tabel 2. Resultaten van vijf studenten

Student	VWO	Punten juni	Punten juli	Duur studie
1	6.5	41	42	260w
2	9.0	37	38	300w
3	7.5	33	34	340w
4	7.0	29	30	380w
5	7.0	25	26	420w

Opdrachten

- 1 In tabel 2 ziet u gegevens van vijf studenten. De correlatie tussen gemiddelde VWO-cijfers en studiepunten in juni is 0.16.
 - a Hoeveel procent van de variantie in de behaalde studiepunten in juni wordt verklaard door het VWO-cijfer?
 - b Maak een scattergram van deze vijf studenten. Welke student valt 'uit de toon'?
 - c Schat met het blote oog hoe hoog de correlatie zou zijn als student 1 buiten beschouwing wordt gelaten.
 - d Alle studenten deden in juli een week stage en behaalden daarvoor nog een stu-

diepunt (kolom 4: punten juli). Wat is de correlatie tussen de punten in juni en juli?

- e De uiteindelijke studieduur van deze vijf studenten (in weken) staat in kolom 5. Wat is de correlatie tussen studiepunten in juli en de uiteindelijke studieduur?

DE AUTEUR

J.A. Smal is als onderwijskundige verbonden aan de Faculteit der Geneeskunde van de Universiteit Utrecht.

Correspondentie-adres:

J.A. Smal. Stafafdeling Onderwijs en Onderzoek, Sectie Onderwijsontwikkeling, Faculteit der Geneeskunde, Bijlhouwerstraat 6, 3511 ZC Utrecht