HOOFDSTUK V

BIVARIATE STATISTIEK

1. DOELSTELLINGEN

Op het eind van dit hoofdstuk zijn studenten in staat op basis van het meetniveau van twee kenmerken te kiezen voor de meeste geschikte associatiemaat om het verband tussen twee variabelen te beschrijven. Studenten kunnen de toegepaste associatiematen correct interpreteren. Bij het toepassen van de bivariate beschrijvende statistiek is het van belang een onderscheid te maken tussen symmetrische en asymmetrische relaties en bijgevolg tussen symmetrische en asymmetrische associatiematen

2. BIVARIATE ASSOCIATIEMATEN TUSSEN CATEGORISCHE VARIABELEN

2.1. Nominale variabelen

2.1.1. Te onthouden kernbegrippen

Chi²	Associatiemaat voor kenmerken op het nominale niveau. Chi² heeft een moeilijke interpretatie en geen absolute begrenzing. Wordt in hoofdzaak gebruikt in de inferentiële statistiek om na te gaan of een verband al dan niet op toeval berust. Indien Chi² een waarde van 0 aanneemt, is er geen samenhang tussen de variabelen.
Cramer's V	Associatiemaat gebaseerd op chi². Varieert net als phi van 0 tot 1 maar kan ook gebruikt worden bij grotere tabellen dan 2*2. Bij een 2*2 tabel is de waarde van Cramer's V gelijk aan de waarde van phi.
Kruistabel of contingentietabel	Tabel waarin de categorieën van twee variabelen tegenover elkaar worden gezet en waarin de waargenomen frequentie van elke combinatie van categorieën vermeld staat.
Odds (ratio)	Een maat om de verhouding tussen het voorkomen van een gebeurtenis en het niet voorkomen van een gebeurtenis te beschrijven. De odds ratio is de verhouding tussen twee odds. De odds ratio neemt de waarde aan van 1 bij afwezigheid van een verband en wijkt af van 1 naarmate het verband sterker wordt. De afwijking gebeurt naar 0 of naar + oneindig.
Percentageverschil	Bivariate associatiemaat voor kenmerken op het nominale niveau. De percentages op 1 categorie van de afhankelijke variabele worden vergeleken voor de verschillende categorieën

	van de onafhankelijke variabele. Het verschil wordt weergegeven in percentagepunten.
Phi	Associatiemaat gebaseerd op chi². Heeft de waarde 0 bij afwezigheid van associatie en waarde 1 bij perfecte statistische associatie. Wordt gebruikt in 2*2 tabellen.

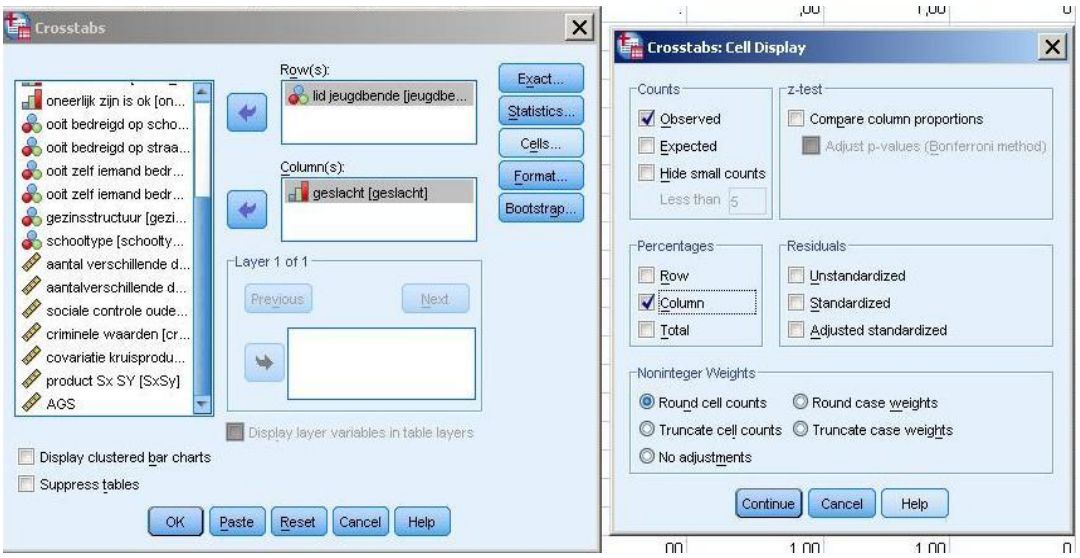
2.1.2. Statistische symbolen en formules

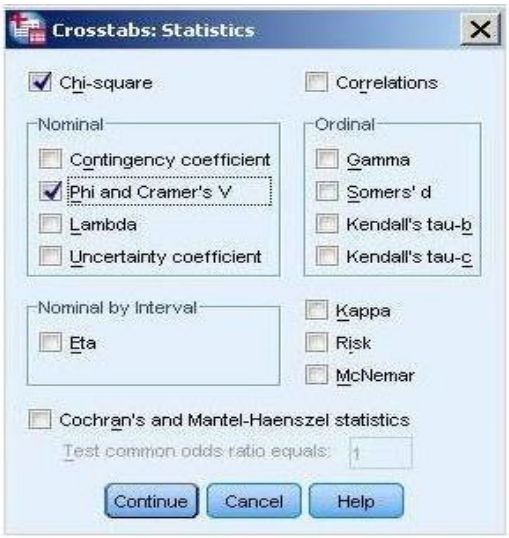
Chi²	$\chi^2 = \sum \frac{(geobserveerd - verwacht)^2}{verwacht}$

2.1.3. Associatiematen op nominaal niveau: SPSS-output en interpretatie

Zijn jongens meer betrokken bij een jeugdbende dan meisjes ?

We hebben gezien dat associatiematen op nominaal niveau worden gebruikt voor de analyse van contingentietabellen. We tonen je eerst even voor hoe je een contingentietabel in SPSS kan maken. Kies onder "Analyze" → "cross-tabs" en bepaal de afhankelijke en onafhankelijke variabele! Denk logisch na! Door te klikken op "statistics" krijgen we de associatiematen te zien. We vragen enkel de associatiematen op die passen bij het meetniveau. We maken de kruistabel door te klikken op "cells" en de juiste percentages op te vragen. Je moet verstandig kiezen tussen kolompercentages en rijpercentages. Als je de afhankelijke in de rijen plaatst, dan bereken je de kolompercentages.





Je klikt vervolgens op de associatiematen die je wil kennen. Wij kiezen voor chi-kwadraat, Phi en Cramer's V als belangrijke nominale associatiematen.

Zo ziet de output van de statistische analyse van de hierboven beschreven contingentietabel in SPSS er uit:

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
lid jeugdbende * geslacht	1548	99,6%	6	,4%	1554	100,0%

lid jeugdbende * geslacht Crosstabulation

			geslacht		
			meisje	jongen	Total
lid jeugdbende	geen lid jeugdbende	Count	786	661	1447
		% within geslacht	96,6%	90,1%	93,5%
	lid jeugdbende	Count	28	73	101
		% within geslacht	3,4%	9,9%	6,5%
Total		Count	814	734	1548
		% within geslacht	100,0%	100,0%	100,0%

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	26,785[b]	1	,000		
Continuity Correction[a]	25,729	1	,000		
Likelihood Ratio	27,447	1	,000		
Fisher's Exact Test				,000	,000
Linear-by-Linear Association	26,768	1	,000		
N of Valid Cases	1548				

a. Computed only for a 2x2 table

b. 0 cells (,0%) have expected count less than 5. The minimum expected count is 47,89.

Symmetric Measures

		Value	Approx. Sig.
Nominal by Nominal	Phi	,132	,000
	Cramer's V	,132	,000
N of Valid Cases		1548	

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

Eerst krijgen we de informatie te zien over de observaties (respondenten) en de ontbrekende waarden. Daarna volgt de uiteindelijke kruistabel. Tot slot volgen associatiematen, de chi-kwadraat waarde en automatisch verkrijgen we de bijbehorende significantietoetsen.[1] We krijgen ook de associatiematen op nominaal niveau te zien: Phi en Cramer's V. Is er nu een verband tussen geslacht en het behoren tot een jeugdbende? We lezen de informatie af uit de kruistabel en interpreteren de resultaten. We bespreken de resultaten zoals het hoort in een rapport dat je maakt in het kader van de bachelorproef of masterproef.

We rapporteren als volgt. *"Deze bivariate analyse is gebaseerd op 1548 respondenten. Dit is de effectieve steekproefgrootte. De item-nonrespons bedraagt slechts 0.4% van het totaal aantal respondenten (1554). 6.5% (101) van de respondenten geeft aan lid te zijn van een jeugdbende. Jongens (9.9%) zijn vaker lid dan meisjes (3.4%). Dit geeft een verschil van 6.5 percentagepunten."* De associatiemaat Cramer's V en Phi zijn hier aan elkaar gelijk omdat we een 2*2 tabel hebben. Er zijn twee rijen en twee kolommen omdat de beide variabelen bestaan uit twee categorieën. In de theoretische lessen hebben we gezien dat in zulke situaties de beide coëfficiënten gelijk zijn. Onthou dat je een keuze dient te maken bij de rapportage en dat je steeds dient te motiveren waarom je een bepaalde associatiemaat hanteert. Het verband is eerder aan de zwakke kant. Cramer's V bedraagt 0.132. Chi-kwadraat bedraagt 26.78. Dit betekent dat er een verschil is tussen de geobserveerde celfrequenties en de verwachte celfrequenties die men zou vinden indien er geen statistisch verband bestaat. Let op! Chi-kwadraat is geen zuivere maat. Had onze steekproef twee keer zo groot geweest, dan was de waarde voor chi-kwadraat ook twee keer zo groot.

In SPSS kan je trouwens zelf uitrekenen hoe de verdeling er zou uitzien indien er geen statistische associatie was geweest tussen beide kenmerken. Dit kan gedaan worden door de "expected counts" op te vragen.

[1] Op de significantietoetsen wordt nu niet verder ingegaan. We hernemen dit in het deel over de inferentiële statistiek (deel V uit deze toegepaste syllabus).

lid jeugdbende * geslacht Crosstabulation

			geslacht		
			meisje	jongen	Total
lid jeugdbende	geen lid jeugdbende	Count	786	661	1447
		Expected Count	760,9	686,1	1447,0
		% within geslacht	96,6%	90,1%	93,5%
	lid jeugdbende	Count	28	73	101
		Expected Count	53,1	47,9	101,0
		% within geslacht	3,4%	9,9%	6,5%
Total		Count	814	734	1548
		Expected Count	814,0	734,0	1548,0
		% within geslacht	100,0%	100,0%	100,0%

Mocht er geen statistische associatie bestaan tussen geslacht en het al of niet lid zijn van een jeugdbende, dan zouden er 760.9 meisjes en 686.1 jongens geen lid zijn van een jeugdbende en zouden er 53.1 meisjes en 47.9 jongens lid zijn van een jeugdbende. De cijfers na de komma hebben hier geen betekenis, maar zijn het resultaat van de toepassing van de formule. Wat je moet onthouden is het volgende: bij afwezigheid van een statistische relatie, zijn de conditionele frequentieverdelingen identiek. Als 6.5% van de observaties lid is van een jeugdbende, dan moeten de kolompercentages voor meisjes en jongens ook allebei 6.5% bedragen. We controleren even door de “expected counts” te delen door de kolomtotalen en we zien inderdaad dat deze beide 6.5% zijn, voor jongens is dat 47.9/734 en voor meisjes is dat 53.1/814. Chi-kwadraat is echter gevoelig aan het aantal rijen en kolommen en aan de marginale verdelingen. Daarom is het belangrijk ook de andere associatiematen te bekijken.

De odds en oddsratio moet je zelf kunnen berekenen op basis van een contingentietabel. De odds voor het lid zijn van een jeugdbende (versus geen lid zijn) bedraagt voor jongens 73/661 en voor meisjes 28/786. De odds ratio is dus (73/661)/(28/786). Anders gesteld: de kans dat jongens lid zijn van een jeugdbende is 3.1 keer groter dan de kans dat meisjes lid zijn van een jeugdbende.

2.2. Ordinale variabelen

2.2.1. Te onthouden kernbegrippen

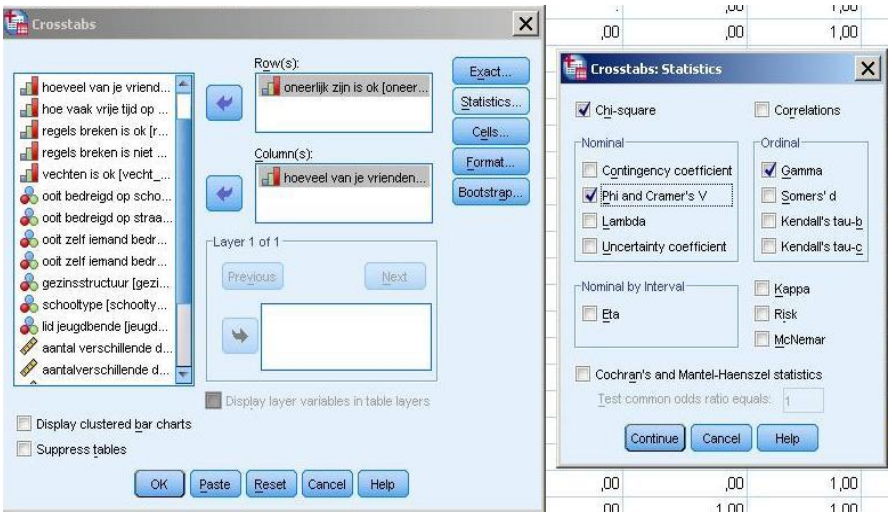
gamma	Associatiemaat voor kenmerken op het ordinale niveau. Neemt een waarde van -1 aan bij perfecte negatieve samenhang en +1 bij perfecte positieve samenhang en 0 bij afwezigheid van samenhang.
Rangcorrelatiecoëfficënt van Kendall (Kendall's Tau-b)	Ordinale maat van samenhang en neemt waarden aan van -1 tot +1 en neemt de waarde 0 aan bij afwezigheid van een lineaire samenhang.
Rangcorrelatiecoëfficiënt van Spearman (Spearman's rho)	Associatiemaat voor kenmerken op ordinaal niveau die afgeleid is van de Pearson productmomentcorrelatiecoëfficiënt. Neemt waarden aan tussen -1 en +1 en neemt de waarde 0 aan bij afwezigheid van een lineaire samenhang

2.2.2. Associatiematen op ordinaal niveau: SPSS-output en interpretatie

Relatie tussen delinquente vrienden en oneerlijk zijn

De associatie tussen ordinale kenmerken kan gebeuren aan de hand van de associatiematen Cramer's V, Gamma en de rangcorrelatiecoëfficiënt van Spearman, nl. Spearman's rho. We gebruiken opnieuw de dataset "Oefendataset1statcrim". We gaan de samenhang tussen twee uitspraken met elkaar vergelijken. Deze zijn: *"oneerlijk zijn is ok"* (de antwoordcategorieën gaan van helemaal niet akkoord tot helemaal akkoord) en *"hoeveel van je vrienden hebben al iets gestolen?"*(de antwoordcategorieën gaan van geen enkele tot bijna allemaal). Deze variabelen zijn ordinaal want ze bestaan uit ordenbare antwoordcategorieën. Bijna allemaal is meer dan geen enkele, maar de afstand daartussen is niet metrisch uit te drukken.

We tonen je eerst even hoe je de voorbeeldoefening in SPSS. We plaatsen één variabele in een rij, de andere in een kolom.



Zo ziet de output van de analyse van de contingentietabel er uit in SPSS:

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
oneerlijk zijn is ok * hoeveel van je vrienden hebben al iets gestolen?	1540	99,1%	14	,9%	1554	100,0%

oneerlijk zijn is ok * hoeveel van je vrienden hebben al iets gestolen? Crosstabulation

			hoeveel van je vrienden hebben al iets gestolen?				
			geen enkele	sommige	de meeste	bijna allemaal	Total
oneerlijk zijn is ok	helemaal oneens	Count	720	41	4	0	765
		% within hoeveel van je vrienden hebben al iets gestolen?	51,9%	28,9%	50,0%	,0%	49,7%
	oneens	Count	288	34	0	0	322
		% within hoeveel van je vrienden hebben al iets gestolen?	20,7%	23,9%	,0%	,0%	20,9%
	noch eens, noch oneens	Count	189	24	1	1	215
		% within hoeveel van je vrienden hebben al iets gestolen?	13,6%	16,9%	12,5%	50,0%	14,0%
	mee eens	Count	129	24	2	0	155
		% within hoeveel van je vrienden hebben al iets gestolen?	9,3%	16,9%	25,0%	,0%	10,1%
	helemaal mee eens	Count	62	19	1	1	83
		% within hoeveel van je vrienden hebben al iets gestolen?	4,5%	13,4%	12,5%	50,0%	5,4%
Total		Count	1388	142	8	2	1540
		% within hoeveel van je vrienden hebben al iets gestolen?	100,0%	100,0%	100,0%	100,0%	100,0%

Symmetric Measures

		Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Nominal by Nominal	Phi	,192			,000
	Cramer's V	,111			,000
Ordinal by Ordinal	Gamma	,375	,055	5,639	,000
N of Valid Cases		1540			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

Opnieuw krijgen we informatie over het aantal geldige observaties en het aantal ontbrekende waarden. We hebben in deze kruistabel een inhoudelijke overweging gemaakt: hoewel we de associatie tussen beide variabelen als symmetrisch zouden kunnen beschouwen, hebben we dit hier niet gedaan. We hebben ervoor geopteerd om "het oneerlijk zijn" te zien als een afhankelijke variabele, dit wil zeggen dat de waarden erop "worden beïnvloed" door een ander kenmerk, met name "het hebben van criminele vrienden", een statement uit de theoretische criminologie, meer bepaald de theorie van Sutherland. We konden er evengoed vanuit gaan dat het oneerlijk zijn leidt tot het hebben van criminele vrienden. En in dat laatste geval hadden we niet de kolompercentages, maar de rijpercentages moeten berekenen. Laat ons nu eens trachten de diagonaal te bekijken van links boven tot rechts onder. We zien toch wel een zeker patroon, al is het niet superduidelijk. Links boven en rechts onder zijn de hoogste percentages vast te stellen: eerder lage waarden op X hangen samen met eerder lage waarden op Y, wie matig scoort op X scoort matig op Y en wie hoog scoort op X scoort ook hoog op Y. In de theoretische syllabus werd gesteld dat de analyse van consistente en inconsistente paren kan leiden tot het ontdekken van een ordinaal lineair patroon. Hoewel we de kolompercentages zouden kunnen bespreken, is het duidelijk dat dit al snel heel veel werk vereist in een grote r*k tabel. Hierin ligt duidelijk het nut van een associatiemaat. Deze vat in één maat samen wat we hier te zien krijgen in deze 5*4 tabel.

Phi en Cramer's V zijn symmetrische maten. Dit betekent dat geen causale richting wordt verondersteld. Zij houden ook geen rekening met de ordenbaarheid in de data. Aangezien we te maken hebben met een r*k tabel, kijken we niet naar Phi maar naar Cramer's V. Gamma daarentegen houdt rekening met de ordenbaarheid van de data: als variabele X een hogere waarde heeft, heeft variabele Y dan ook een hogere waarde?

Hoe rapporteren we deze beschrijvende bevindingen? *"1540 respondenten werden betrokken in deze bivariate analyse. Cramer's V bedraagt 0.111. Er is een zwakke samenhang tussen de beide variabelen."*

Gamma geeft ons meer informatie dan Cramer's V. Omdat beide variabelen ordinaal zijn, verkiezen we gamma als associatiemaat boven Cramers V. Immers, Cramer's V houdt geen rekening met de ordening in de data. Gamma zegt ons dat de associatie tussen beide kenmerken matig en positief is (Gamma= 0.375). Hoe meer vrienden men heeft die al eens iets gestolen hebben, hoe meer men de neiging heeft om oneerlijk te zijn. Merk op dat gamma een hogere waarde heeft dan Cramer's V.

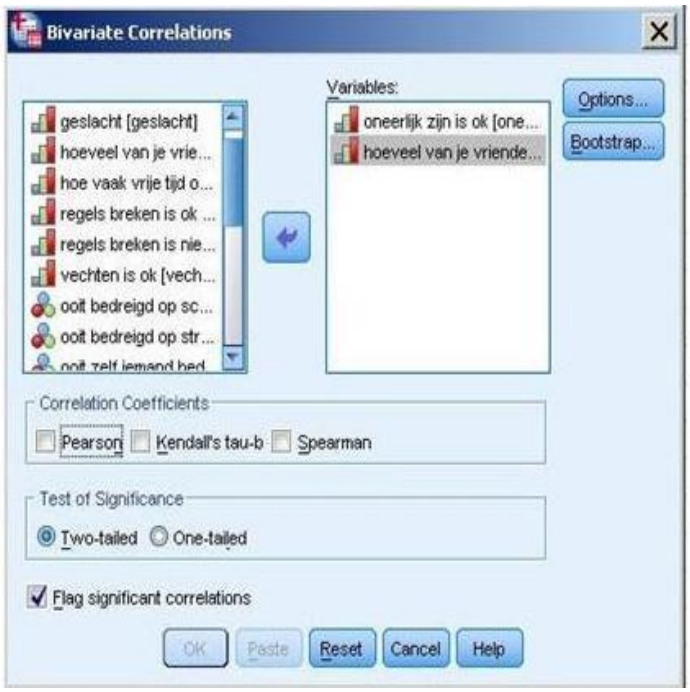
Opgelet! Gamma gaat uit van een monotone rechtlijnige samenhang tussen de beide kenmerken. In de realiteit komt het voor dat er een verband bestaat tussen beide kenmerken,

maar dat dit verband niet eenduidig rechtlijnig is. In dat verband is er sprake van een niet-lineair verband.

Let op! De relatie tussen ordinale en nominale kenmerken is gevoelig voor het kiezen van breekpunten. Stel je voor dat we deze r*k tabel zouden herleiden tot een 2*2 tabel door categorieën samen te gooien, dan wordt de uitkomst door dit keuzeproces beïnvloed. Als je ooit zelf een analyse maakt in het kader van je bachelorproef of masterproef, moet je hiermee rekening houden!!!!

De rangcorrelatiecoëfficiënt Spearman's Rho als alternatieve symmetrische ordinale associatiemaat

We kunnen de relatie tussen twee ordinale kenmerken tenslotte ook via Spearman's Rho berekenen. We tonen je eerst even hoe je deze analyse zelf kan uitvoeren.



In SPSS voeren we een rangcorrelatie-analyse uit als volgt:

Onder analyze → bivariate correlations en vink aan Spearman!

De output ziet er zo uit:

Correlations

			oneerlijk zijn is ok	hoeveel van je vrienden hebben al iets gestolen?
Spearman's rho	oneerlijk zijn is ok	Correlation Coefficient	1,000	,158**
		Sig. (2-tailed)	.	,000
		N	1548	1540
	hoeveel van je vrienden hebben al iets gestolen?	Correlation Coefficient	,158**	1,000
		Sig. (2-tailed)	,000	.
		N	1540	1546

**. Correlation is significant at the 0.01 level (2-tailed).

Spearman's Rho, de rangcorrelatiecoëfficiënt, is afgeleid van de Pearson's product-moment correlatiecoëfficiënt voor interval- en ratio variabelen. De observaties worden eerst in een gewone rangorde (1ste, 2de, 3de, …) geplaatst op de beide variabelen. Daarna past men de formule voor de productmoment correlatiecoëfficiënt toe. Bij ex-aequo's moet een aangepaste formule gebruikt worden. Uit de tabel blijkt dat de rangcorrelatiecoëfficiënt 0.158 is. Dit wijst op een eerder zwakke samenhang. Mogelijks wordt deze zwakkere samenhang veroorzaakt door ex-aequo's. Wij raden aan voor de analyse van ordinale variabelen vooral Gamma te gebruiken. Deze geeft de minst vertekende samenhang tussen beide ordinale kenmerken in het geval van lineaire associatie.

3. BIVARIATE ASSOCIATIEMATEN TUSSEN METRISCHE VARIABELEN

3.1. Symmetrische associatiematen: correlatie analyse

3.1.1. Te onthouden kernbegrippen

Covariatie (SSxy)	Maat van samenhang op het metrische niveau. De covariatie stelt de mate voor waarin twee kenmerken samen variëren. Het is de som van de kruisproducten van de deviatiescores van X en van Y. Ook: kruisproductensom of sum of squares
Covariantie (Sxy)	Maat van samenhang op het metrische niveau. Het is de som van de kruisproducten van de deviatiescores van X en Y gedeeld door n-1.
Correlatiecoëfficiënt van Pearson (Rxy)	Ook wel de productmomentcorrelatiecoëfficiënt van Pearson genoemd. De meest gebruikte bivariate associatiemaat van het gezamenlijk variëren of samenhang voor kenmerken van het metrische niveau. Geeft een indruk van de sterkte en de richting van de lineaire samenhang tussen X en Y. R varieert tussen -1 en +1 met 0 = afwezigheid van lineair verband. Het verband tussen de twee variabelen wordt verondersteld rechtlijnig te zijn (is het verband gekromd dan geeft de correlatiecoëfficiënt een vertekend beeld).
Lineaire samenhang of correlatie	Samenhang tussen waarden op twee variabelen in die zin dat waarden van de beide variabelen dezelfde of een tegengestelde tendens vertonen. Lineair betekent dat het gaat over 'samenhang ten opzichte van een rechte'. De correlatiecoëfficiënt is een maat voor de sterkte van de lineaire samenhang tussen X en Y.

Scatterplot of puntenwolk of spreidingsdiagram	Verzameling van alle elementen uit de steekproef waarbij elk punt (coördinaat) een statistische eenheid weergeeft met informatie op de X-variabele en de Y-variabele.
Symmetrische associatiemaat	Associatiemaat die de sterkte van een symmetrisch verband tussen variabelen weergeeft.
Symmetrische relatie	Geen expliciete veronderstelling over de causale relatie tussen variabelen. Er wordt enkel nagegaan in welke mate er samenhang bestaat tussen de variabelen. Deze samenhang geldt in beide richtingen. Vb. Als hoge waarden op A samenhang met hoge waarden op B, dan gaan hoge waarden op B ook samen met hoge waarden op A.

3.1.2. Statistische symbolen en formules

Covariatie of kruisproductensom of sum of squares (SSxy)	$SS_{xy} = \sum (X - \overline{X})(Y - \overline{Y})$
Covariantie (Sxy)	$s_{xy} = \frac{\sum (X - \overline{X})(Y - \overline{Y})}{n-1}$
Correlatiecoëfficiënt van Pearson (Rxy)	$r_{xy} = \frac{s_{xy}}{s_x \cdot s_y}$

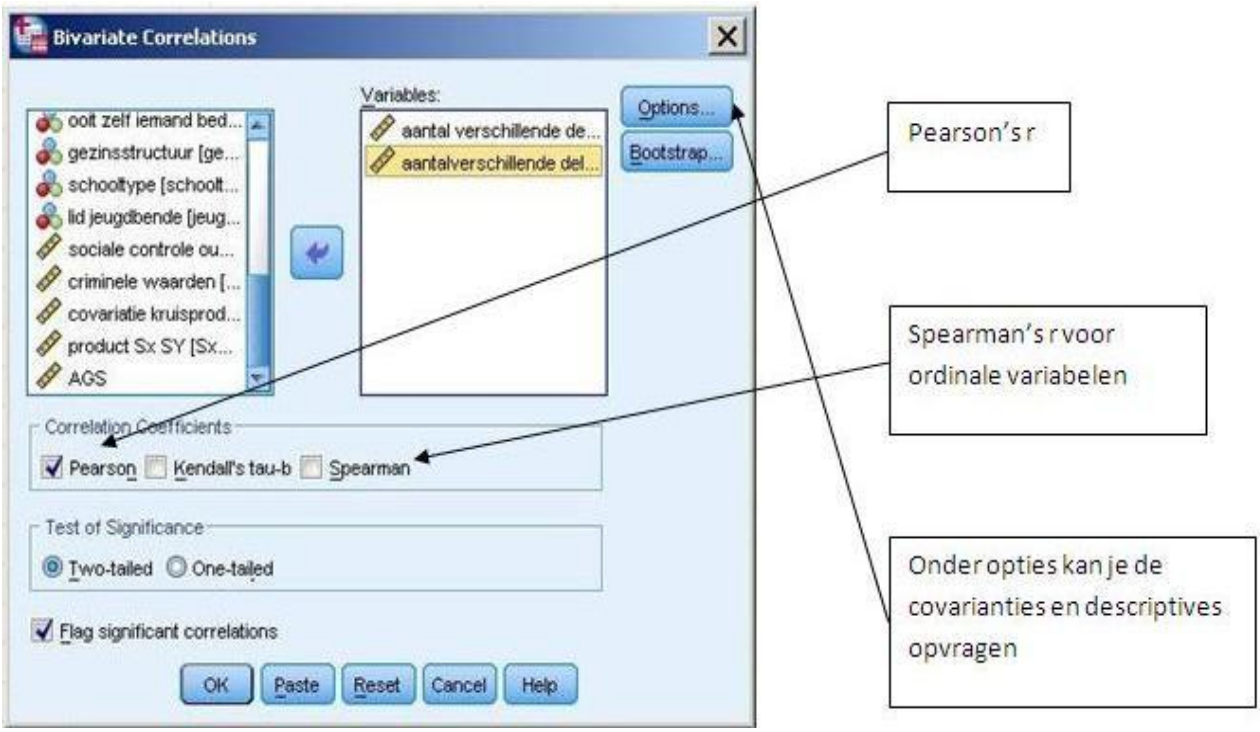
3.1.3. Symmetrische associatiematen op metrisch niveau: SPSS-output en interpretatie

Is er een samenhang tussen daderschap en slachtofferschap ?

Symmetrische associatiematen voor kenmerken gemeten op metrisch niveau zijn de covariatie, de covariantie en de correlatie. We geven het voorbeeld hoe je de correlatiecoëfficient uitrekent tussen twee metrische kenmerken. We gebruiken het aantal keer dat men slachtoffer is geweest en dader is geweest als voorbeeld. Beide concepten zijn gemeten op het metrische niveau. Het berekenen van de correlatiecoëfficiënt van Pearson gebeurt aan de hand van de formule:

$$r_{xy} = \frac{s_{xy}}{s_x \cdot s_y}$$

SPSS geeft de waarde van de correlatiecoëfficiënt zelf via de procedure “correlations”



De output van een correlatie-analyse ziet er als volgt uit:

Correlations

		aantal verschillende delicten slachtoffer	aantalverschillende delicten gepleegd
aantal verschillende delicten slachtoffer	Pearson Correlation	1	,444**
	Sig. (2-tailed)		,000
	N	1491	1453
aantalverschillende delicten gepleegd	Pearson Correlation	,444**	1
	Sig. (2-tailed)	,000	
	N	1453	1504

**. Correlation is significant at the 0.01 level (2-tailed).

We geven enkele richtlijnen met betrekking tot de sterke van het gevonden verband:

Deze associatiemaat gaat van -1 tot +1 en is 0 bij afwezigheid van een lineair verband. *

- waarde tussen 0 en 0.10 (in absolute waarden): verwaarloosbaar verband
- waarde tussen 0.10-0.30 (in absolute waarden) : zwak tot matig bivariaat verband
- waarde tussen 0.30-0.60 (in absolute waarden): matig tot sterk bivariaat verband
- waarde hoger dan 0.60 (in absolute waarden): heel sterk bivariaat verband

Er is in het voorbeeld tussen het plegen van delicten en het slachtoffer een matige positieve (lineaire) samenhang vast te stellen met een waarde van 0.44.

3.2. Asymmetrische associatiematen: bivariate regressie analyse

3.2.1. Te onthouden kernbegrippen

Afhankelijke variabele	Responsvariabele of explanandum
Determinatie coëfficiënt (R^2)	Een goodness of fit maat die weergeeft hoeveel procent van de geobserveerde verschillen of varia(n)tie in de afhankelijke variabele kan verklaard worden op basis van de onafhankelijke variabele. = proportie van de totale variatie in Y die door X kan verklaard worden. In de bivariate regressieanalyse is R^2 gelijk aan het kwadraat van de correlatiecoëfficiënt tussen X en Y.
Gestandaardiseerde regressiecoëfficiënt (β)	Geeft de sterkte weer van een effect. Is in de bivariate regressie analyse gelijk aan de correlatiecoëfficiënt. Berekening: covariantie tussen X en Y delen door het product van standaardafwijking van X met standaardafwijking van Y. Neemt een waarde aan tussen -1 en +1 waarbij -1 en +1 een perfecte relatie aanduiden (dwz de onafhankelijk variabele kan de afhankelijke perfekt voorspellen, dus geen residuen. Bij een waarde 0 heeft de onafhankelijke geen effect op de afhankelijke variabele.
Foutenterm of error	Foutenterm of residu of residuele term is het verschil tussen de werkelijk (geobserveerde) waarde van de afhankelijke variabele en de voorspelde waarde van de afhankelijke variabele.
Intercept	Zie: regressieconstante
Onafhankelijke variabele	Predictor variabele of explanans

Ongestandaardiseerde regressiecoëfficiënt (b)	Geeft aan met hoeveel eenheden Y toeneemt als x met één eenheid toeneemt. Berekening: covariantie tussen X en Y gedeeld door de variantie in X.
Regressie analyse	Het, met een zekere mate van precisie, voorspellen van de score op een afhankelijke variabele (Y) op basis van één (bivariate regressie) of meerdere onafhankelijke (multiple regressie) X-variabelen.
Regressie coëfficiënten	Regressieconstante of het intercept en het regressiegewicht dat de hellingshoek van de regressielijn aanduidt.
Regressieconstante of het intercept (a)	De verwachte of voorspelde waarde van de afhankelijke variabele Y wanneer de waarde op de onafhankelijke variabele 0 bedraagt.
Regression sum of squares	Geeft de variatie in de afhankelijke variabele weer die voorspeld kan worden op basis van de onafhankelijke variabele. Wordt berekend door de afwijkingen van de voorspelde waarden van de afhankelijke variabele ten opzichte van het rekenkundig gemiddelde van de afhankelijke variabele te kwadrateren en vervolgens deze kwadraten te sommeren.
Residual sum of squares	De residuele variatie of residual sum of squares wordt berekend door de afwijkingen van de werkelijke waarde van de afhankelijke variabele ten opzichte van de voorspelde waarde van de afhankelijke variabele te kwadrateren en op te tellen.

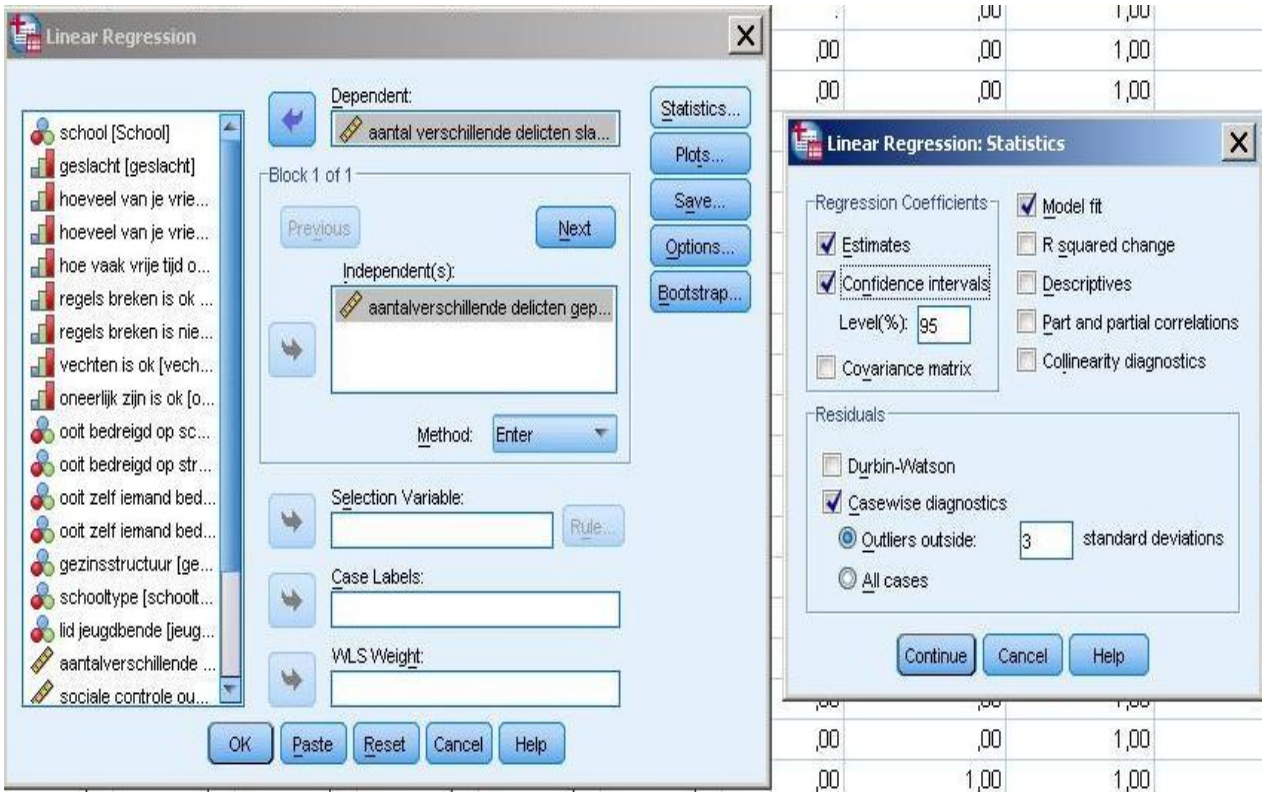
3.2.2. Statistische symbolen en formules

Determinatie coëfficiënt (R^2)	$R^2 = \frac{\sum_{i=1}^{n}\left[\hat{y}_i - \overline{Y}\right]^2}{\sum_{i=1}^{n}\left[y_i - \overline{Y}\right]^2}$ ← Teller: Regression sum of squares ← Noemer: Total sum of squares
Enkelvoudige lineaire regressievergelijking	$Y = a + b_1X + e$
Ongestandaardiseerde regressiecoëfficiënt (b)	$b_1 = r(X,Y)\, S_Y/S_x$
Regressieconstante of intercept (a)	$a = \overline{Y} - b_1 \overline{X}$

3.2.3. Asymmetrische associatiematen op metrisch niveau: SPSS-output en interpretatie (bivariate regressie analyse metrische variabelen)

Wie kwaad doet, kwaad ontmoet ?

Een zeer belangrijke tool in de criminologie is de lineaire regressieanalyse waar de onderzoeker geïnteresseerd is in het voorspellen van de waarden op de afhankelijke variabele op basis van een onafhankelijke variabele. We hanteren hier een populair vraagstuk uit de criminologie: er wordt wel eens beweerd, en we hebben het hierboven reeds aangetoond, dat er een verband bestaat tussen enerzijds het plegen van delicten en anderzijds het slachtoffer worden van delicten. Bij de berekening van de correlatiecoëfficiënt hebben we vastgesteld dat de samenhang in onze dataset inderdaad positief is: 0.44. De bivariate regressieanalyse is een asymmetrische analysetechniek die ons toelaat voorspellingen te maken voor de waarden op de Y-variabele (de afhankelijke variabele) op basis van de onafhankelijke predictor X. We moeten hier dus een expliciete keuze maken vanuit theoretische gronden. Welke variabele beschouwen we als afhankelijke en welke als onafhankelijke? We gaan er hier theoretisch van uit dat we het plegen van delicten ("kwaad doen") zien als onafhankelijke variabele. Wie kwaad doet, zou dan meer slachtoffer kunnen worden. Dit is een eenzijdige oorzakelijke interpretatie van een gekend verband uit criminologisch onderzoek, maar om didactische redenen houden we ons aan deze interpretatie. In SPSS voeren we een regressie-analyse uit door onder "analyze" te kiezen voor "linear regression".



De afhankelijke variabele brengen we via de overbrengknop onder "Dependent" en de onafhankelijke variabele brengen we via de overbrengknop onder "Independent". Door op "options" te klikken, kunnen we de regressieparameters opvragen. We vragen systematisch de "estimates" (regressiecoëfficiënten), de betrouwbaarheidsintervallen[2], de "model fit" (determinatiecoëfficiënt), de beschrijvende statistieken en de residuele termen op.

De output van een regressie-analyse ziet er in SPSS als volgt uit:

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	aantalverschillende delicten gepleegd[a]	.	Enter

a. All requested variables entered.

b. Dependent Variable: aantal verschillende delicten slachtoffer

[2] Hierop komen we verder terug.

SPSS informeert onder "Variables entered" welke de afhankelijke en onafhankelijke variabele is. Dit is belangrijk. Zo zien we of we geen vergissingen gemaakt hebben.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,444[a]	,197	,196	1,84028

a. Predictors: (Constant), aantalverschillende delicten gepleegd

b. Dependent Variable: aantal verschillende delicten slachtoffer

Onder "model summary" staan achtereenvolgens de correlatiecoëfficiënt, de determinatiecoëfficiënt, de aangepaste determinatiecoëfficiënt en de standaardfout van de schatter. Deze twee laatste waarden worden later behandeld. De Anova tabel is zeer belangrijk bij de inferentiële statistiek en dit vormt een onderdeel van een volgend hoofdstuk. Een aantal zaken zullen we hier reeds bespreken.

Uit de *"model summary"* of de samenvatting van de fit van het statistische model, zien we de volgende coëfficiënten:

R: deze is de samenhang tussen de geobserveerde waarde en de voorspelde waarde op de Y-variabele. Deze waarde is hier gelijk aan de bivariate correlatiecoëficiënt tussen X en Y.

R Square: deze is de determinatiecoëfficiënt. Het is de verklaarde variantie: 19.7% van de geobserveerde verschillen op de afhankelijke variabele (slachtofferschap) kan verklaard worden vanuit het plegen van delicten.

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1203,779	1	1203,779	355,451	,000[a]
	Residual	4913,997	1451	3,387		
	Total	6117,776	1452			

a. Predictors: (Constant), aantalverschillende delicten gepleegd

b. Dependent Variable: aantal verschillende delicten slachtoffer

Uit de ANOVA tabel kunnen we hetzelfde aflezen als wat erboven stond:

We krijgen zicht op de "Regression sum of squares", de "Residual sum of squares" en de "Total sum of squares".

De formule van de determinatiecoëfficiënt zegt:

R square is $$\frac{\sum_{i=1}^{n}\left[\hat{y}_i - \overline{Y}\right]^2}{\sum_{i=1}^{n}\left[y_i - \overline{Y}\right]^2}$$

Ofwel: "*Neem de som van het kwadraat van het verschil tussen elke voorspelde waarde van Y op basis van X (Ŷ of Y-hat) minus het gemiddelde op Y (="regression sum of square")."*

"En deel deze som vervolgens door de som van het kwadraat van het verschil tussen alle geobserveerde waarden van Y minus het gemiddelde op Y (= "total sum of squares" of de variatie in Y)."

Dus: 1203.779/6117.776= 0.197 of 19.7% van de waargenomen verschillen in slachtofferschap kan verklaard worden door de waargenomen verschillen in delicten plegen.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	1,449	,061		23,825	,000	1,330	1,568
	aantalverschillende delicten gepleegd	,475	,025	,444	18,853	,000	,426	,525

a. Dependent Variable: aantal verschillende delicten slachtoffer

schattingen intercept en richtingscoef.

geschatte regressielijn is =

Y-hat= 1.449 + 0.475 X

De belangrijkste informatie staat in de output onder *"coefficients".* Hier lezen we de regressiecoëfficiënten af. We zien dat SPSS standaard het intercept en de hellingshoek (richtingscoëfficiënt) weergeeft. SPSS geeft zowel de ongestandaardiseerde coëfficiënten als de gestandaardiseerde coëfficiënten.

De constante (a of b_0) of het intercept bedraagt 1.449. Dit is het verwacht aantal keer dat iemand slachtoffer wordt voor iemand die geen enkele keer een crimineel feit heeft gepleegd.

De ongestandaardiseerde richtingscoëfficiënt (b_1) bedraagt 0.475. SPSS geeft deze weer met de benaming B. Dit is de verwachte toename in Y als X met een eenheid stijgt, dus:

$\hat{Y} = b_0 + b_1 X$

Of: $\hat{Y}$ = 1.449 + 0.475 (aantal verschillende delicten gepleegd). We kunnen dus door een waarde op X (aantal delicten gepleegd) in te vullen, te weten komen hoeveel keer iemand verwacht wordt slachtoffer te worden. Voor iemand die 3 delicten heeft gepleegd, wordt deze verwachting: 1.449 + 0.475*3.

De gestandaardiseerde richtingscoëfficiënt komt in een bivariate analyse overeen met de bivariate correlatiecoëfficiënt. De overige parameters worden in een later hoofdstuk beschreven.

Interessant is dat SPSS ons de mogelijkheid biedt om "outliers" of "uitbijters" te bekijken. Dit zijn observaties waar we een erg slechte voorspelling hebben gedaan. Met erg slecht bedoelen we dat we er met de voorspelling zeer ver naast zaten: minstens drie standaardafwijkingen (zie ook het hoofdstuk over de normale verdeling en standaardnormale scores). Dit is interessant omdat we deze cases afzonderlijk in de diepte kunnen bestuderen.

Casewise Diagnostics[a]

Case Number	Std. Residual	aantal verschillende delicten slachtoffer	Predicted Value	Residual
65	4,130	10,00	2,3998	7,60020
154	3,301	8,00	1,9245	6,07554
159	3,560	8,00	1,4491	6,55088
210	4,647	10,00	1,4491	8,55088
216	3,043	8,00	2,3998	5,60020
295	3,016	7,00	1,4491	5,55088
358	3,845	9,00	1,9245	7,07554
431	3,043	8,00	2,3998	5,60020
640	4,388	10,00	1,9245	8,07554
700	4,130	10,00	2,3998	7,60020
926	-3,370	,00	6,2025	-6,20252
1142	3,016	7,00	1,4491	5,55088
1145	3,560	8,00	1,4491	6,55088
1351	3,613	10,00	3,3505	6,64952
1379	-3,112	,00	5,7272	-5,72718
1523	3,016	7,00	1,4491	5,55088
1553	3,355	10,00	3,8258	6,17418

a. Dependent Variable: aantal verschillende delicten slachtoffer

Residuals Statistics[a]

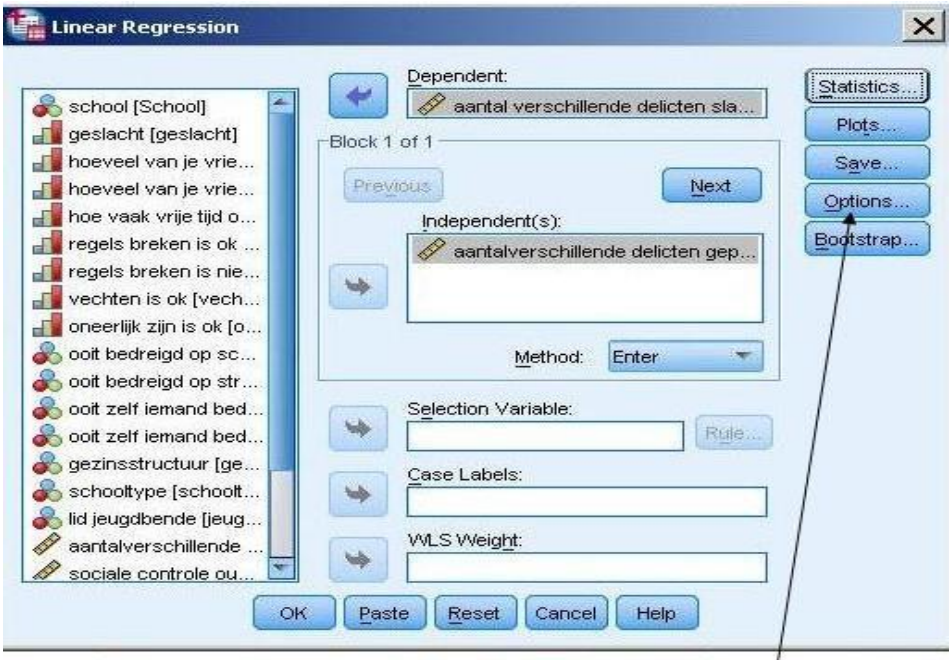
	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	1,4491	7,6285	2,1466	,91052	1453
Residual	-6,20252	8,55088	,00000	1,83965	1453
Std. Predicted Value	-,766	6,021	,000	1,000	1453
Std. Residual	-3,370	4,647	,000	1,000	1453

a. Dependent Variable: aantal verschillende delicten slachtoffer

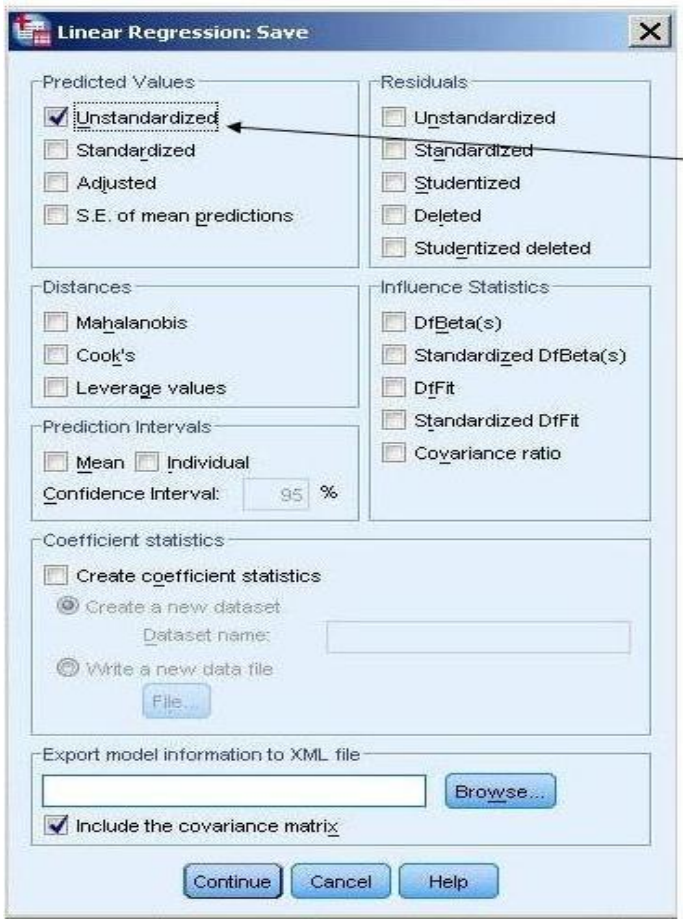
Tot slot geeft SPSS ook de residuele statistieken (mimimum, maximum, gemiddelde, standaardafwijking). We zien achtereenvolgens de voorspelde waarde (predicted value), de residuele, de gestandaardiseerde voorspelde waarde (deze heeft natuurlijk een gemiddelde van nul en std van 1), precies zoals de gestandaardiseerde residuele term. De "residual statistics" zijn de beschrijvende statistieken van de residuele termen. Een residuele term is het verschil tussen wat er geobserveerd werd als waarde op Y en wat we verwachtten als waarde op Y op basis van onze kennis over X.

Zelf stapsgewijs narekenen van de determinatiecoëfficiënt aan de hand van SPSS

De determinatiecoëfficiënt is ook het kwadraat van correlatie tussen de geobserveerde waarden op Y en de voorspelde (of verwachte) waarde voor Y op basis van X. Laten we dit even zelf narekenen. We tonen aan dat SPSS de formule inderdaad correct toepast. Zo gaan we te werk:

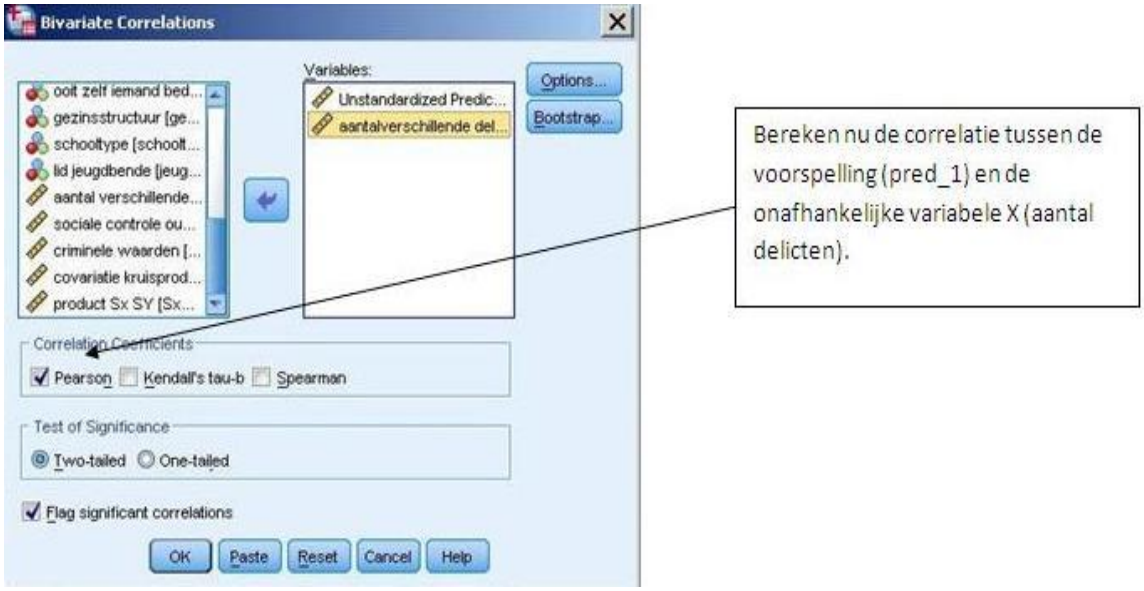


Klik op "options" en bewaar de verwachte waarde van Y op basis van X



Bewaar de ongestandaardiseerde voorspelde waarde

Opgelet! Deze voorspellingen op basis van X worden door SPSS weggeschreven (of bewaard) onder de naam Pre_1. Je kan zelf deze naam veranderen als je wil.



Lees het resultaat af uit de tabel.

Correlations

		Unstandardized Predicted Value	aantal verschillende delicten slachtoffer
Unstandardized Predicted Value	Pearson Correlation	1	,444**
	Sig. (2-tailed)		,000
	N	1504	1453
aantal verschillende delicten slachtoffer	Pearson Correlation	,444**	1
	Sig. (2-tailed)	,000	
	N	1453	1491

**. Correlation is significant at the 0.01 level (2-tailed).

Hiermee hebben we aangetoond dat R inderdaad de samenhang is tussen de voorspelling en de geobserveerde waarde. Nemen we daarvan het kwadraat, dan bekomen we inderdaad de determinatiecoëfficiënt.

4. SAMENVATTENDE TABEL

NOMINAAL MEETNIVEAU	ORDINAAL MEETNIVEAU	METRISCH MEETNIVEAU
Percentageverschil (uitgedrukt in percentagepunten)	Gamma -1 tot +1	Covariatie
Odds (ratio) 1= geen associatie Van 0 tot + ∞	Spearmans rho -1 tot +1	Covariantie
Chi² Van 0 tot + ∞	Kendall's Tau-b -1 tot +1	Correlatie-coëfficiënt
Phi (2*2 tabel) 0 tot 1		
Cramer's V (r*k tabel) 0 tot 1		

5. OEFENINGEN

1. De Chi²-toets is één van de meest gebruikte manieren om relaties tussen twee of meer categorische variabelen te bestuderen.
   Onderzoekers verzamelden gegevens over rookstatus en de diagnose longkanker bij een willekeurige steekproef van volwassenen. Elk van deze variabelen is dichotoom: een persoon rookt momenteel of niet en heeft een longkankerdiagnose of niet.

Longkanker diagnose			
Rookstatus	Diagnose	Geen diagnose	Totaal
Roker	60	300	
Niet-roker	10	390	
			N=

- Hoeveel bedraagt de marginale frequentieverdeling voor de variabele longkanker diagnose?

- Hoeveel bedraagt de marginale frequentieverdeling voor de variabele Rookstatus?

- Hoeveel procent van de respondenten met longkanker diagnose is roker?

- Hoeveel procent van de respondenten zonder longkanker diagnose is roker?

- Hoeveel bedraagt het relevante percentageverschil?

Bereken Chi²

- Bereken de verwachte waarden/frequenties voor elke cel (dit zijn de verwachte frequenties wanneer de twee variabelen onafhankelijk zijn of niet samenhangen)

Longkanker diagnose			
Rookstatus	Diagnose	Geen diagnose	Totaal
Roker	60	300	
Niet-roker	10	390	
			N=

- Hoeveel bedraagt Chi²?
- Welke associatiemaat gebaseerd op Chi² kun je bijkomend berekenen?
- Hoeveel bedraagt deze? Interpreteer.

2. *Is er een verband tussen vuurwapenbezit en biologisch geslacht?*

   In een hypothetisch onderzoek wordt aan 817 mannen en aan 1040 vrouwen gevraagd of zij een vuurwapen bezitten. De onderzoeker wil weten of er een verband bestaat tussen het bezit van een vuurwapen en biologisch geslacht. Hieronder vind je de frequenties. Maak een volledige analyse van de kruistabel.

   Opgelet! Bepaal vooraf welke variabelen je zou beschouwen als onafhankelijk en als afhankelijk. Plaats de onafhankelijke in de kolommen (=kolomvariabele) en de afhankelijke in de rijen (rijvariabele).

Ben jij in het bezit van een vuurwapen?		
Vuurwapenbezit	JA	NEEN
Man	343	474
Vrouw	260	780

- Bereken het relevante percentageverschil.
- Bepaal de H0 en Ha.
- Bereken Chi².
- Welke associatiemaat gebaseerd op Chi² kun je bijkomend berekenen?
- Hoeveel bedraagt deze? Interpreteer.

3. *Is er een verband tussen het roken van sigaretten en het drinken van alcohol bij studenten?*

   Hieronder vind je een 2*2 kruistabel. Aan 110 studenten werd de vraag gesteld of ze al dan niet rookten en of ze al dan niet dronken.

		Rook je?	
		JA	NEE
Drink je?	JA	50	15
	NEE	20	25

   - Bereken het relevante percentageverschil.
   - Bepaal de H0 en Ha.
   - Bereken Chi².
   - Welke associatiemaat gebaseerd op Chi² kun je bijkomend berekenen ?
   - Hoeveel bedraagt deze ? Interpreteer.

4. *Is er een verband tussen burgerlijke staat en drinkgedrag?*

   Een nationale enquête werd uitgevoerd om informatie te verkrijgen over de alcoholconsumptiepatronen van volwassenen in Vlaanderen op basis van hun burgerlijke staat. Een willekeurige steekproef van 1772 inwoners van 18 jaar en ouder leverde de weergegeven gegevens op:

		Aantal glazen alcohol per maand			
		0	1-60	Meer dan 60	Totaal
Burgerlijke staat	Vrijgezel	67	213	74	354
	Gehuwd	411	633	127	1173
	Weduwe weduwnaar	85	51	7	143
	Gescheiden	27	60	15	102
	Totaal	590	957	225	1772

   - Bepaal de H0 en Ha.
   - Bereken Chi².
   - Welke associatiemaat gebaseerd op Chi² kun je bijkomend berekenen ?
   - Hoeveel bedraagt deze ? Interpreteer.

5. *Angst voor jeugdbendes.*

De relatie tussen *bezorgdheid voor diefstal door jeugdbendes* en *daadwerkelijk slachtoffer worden van diefstal* werd onderzocht. Interviews met een steekproef van middelbare scholieren leverden de volgende kruistabel op:

		Bezorgdheid voor diefstal door jeugdbendes		TOTAAL
		JA	NEE	
Angst om daadwerkelijk slachtoffer te worden	JA	58	45	
	NEE	29	71	
	TOTAAL			

- Bereken Chi².
- Welke associatiemaat gebaseerd op Chi² kun je bijkomend berekenen ?
- Hoeveel bedraagt deze ? Interpreteer.

6. *Is er een verband tussen 'ontrouw' en 'geluk in de relatie'?*

   Onderstaande kruistabel geeft data over het aantal mannen dat al dan niet ontrouw is, in relatie tot de mate waarin zij aangeven gelukkig te zijn in hun relatie. De variabele 'geluk in de relatie' werd hercodeerd naar een dichotome variabele op basis van de mediaan. Scores hoger dan de mediaan werden gecodeerd als 'ongelukkig' en scores gelijk aan of lager dan de mediaan werden gecodeerd als 'gelukkig'.

		Ontrouw?		TOTAAL
		Ontrouw	Trouw	
Geluk in de relatie	Ongelukkig	56	101	157
	Gelukkig	62	287	349
	TOTAAL	118	388	506

   - Hoeveel procent van het aantal mannen die rapporteren gelukkig te zijn in hun relatie, is ontrouw?
   - Hoeveel procent van het aantal mannen die rapporteren trouw te zijn in hun relatie, is ongelukkig?
   - Is er een verband tussen ontrouw en geluk in de relatie?
   - Hoe sterk is het verband?

7. De Vlaamse minister van onderwijs start een grootschalig onderzoek op naar pestgedrag in het Vlaamse secundair onderwijs. Check-It, het ondertussen wereldvermaarde onderzoeksbureau, bezorgt binnen de kortste keren het onderzoeksrapport aan de minister. Eén van de resultaten belicht het eventuele geslachtsverschil in het pesten.

	Jongens	Meisjes	Totaal
Niet-pester	4613	5530	10143
Pester	1132	568	1700
Totaal	5745	6098	11843

- Bereken de oddsratio van jongens tegenover meisjes (pesters vs niet-pesters).

- Wat zegt de zojuist berekende oddsratio van jongens tegenover meisjes (pester vs. Niet-pesters) ?

  a. Er zijn in totaal 2,39 maal minder pesters dan niet-pesters (zowel onder de meisjes als onder de jongens)
  b. De verhouding tussen pesters en niet-pesters ligt 2,39 keer hoger bij de jongens dan bij de meisjes
  c. De verhouding tussen pesters en niet-pesters ligt 2,39 keer hoger bij de meisjes dan bij de jongens
  d. Er zijn in totaal 2,39 maal meer pesters dan niet-pesters (zowel onder de meisjes als onder de jongens)

8. Beantwoord onderstaande vragen

	Jongens	Meisjes	Totaal
Ooit alcohol gebruikt	410	472	
Nooit alcohol gebruikt	623	1263	
Totaal			

a. Bepaal voor mannen de odds op 'ooit gebruikt' t.o.v. 'nooit gebruikt' en interpreteer

b. Bepaal voor vrouwen de odds op 'ooit gebruikt' t.o.v. 'nooit gebruikt' en interpreteer

c. Bepaal vervolgens de oddsratio mannen t.o.v. vrouwen. Druk de betekenis uit in je eigen woorden.

d. Bepaal vervolgens de oddsratio vrouwen t.o.v. mannen. Druk de betekenis uit in je eigen woorden

9. Lees onderstaande uitspraken. Elk van de uitspraken bevat een blunder. Geef voor elk geval aan wat er mis is.

'Er bestaat een grote correlatie tussen het geslacht van Amerikaanse werknemers en hun inkomen.'

'We vonden een grote correlatie (r = 1.09) tussen de door studenten gegeven beoordeling over de kwaliteit van het lesgeven van stafleden en de door andere stafleden gegeven beoordeling'.

'De correlatie tussen plantdichtheid en de opbrengst van maïs bleek r = 0.23 kubieke meter te zijn'.

10. Onze emotionele reactie op sociale afwijzing wordt omschreven als 'sociale pijn'. Een onderzoek bekeek of sociale afwijzing zorgt voor activiteit in hersengebieden waarvan bekend is dat deze door fysieke pijn worden geactiveerd. Als dat het geval is, dan ervaren we werkelijk sociale en fysieke pijn op dezelfde manier. Personen in het onderzoek werden eerst bij een sociale activiteit betrokken en vervolgens opzettelijk buitengesloten terwijl de toename van de bloedcirculatie in hun hersenen werd gemeten. Na elke activiteit vulden de personen een vragenlijst in om aan te geven in welke mate zij zich buitengesloten voelden.

    Hieronder volgen de gegevens voor 13 proefpersonen. De verklarende variabele is 'sociaal leed'. Deze variabele is gemeten aan de hand van de scores in elke vragenlijst, waarbij de score na de buitensluiting wordt gerelateerd aan de score na afloop van de activiteit. Dus, waarden groter dan 1 tonen de mate van pijn veroorzaakt door het buitensluiten. De te verklaren variabele is de activiteit in de anterior cingulate cortex, een hersengebied dat door fysieke pijn wordt geactiveerd.

Proefpersoon	Sociaal leed	Hersenactiviteit
1	1.26	-0.055
2	1.85	-0.040
3	1.10	-0.026
4	2.50	-0.017
5	2.17	-0.017
6	2.67	0.017
7	2.01	0.021
8	2.18	0.025
9	2.58	0.027
10	2.75	0.033
11	2.75	0.064
12	3.33	0.077
13	3.65	0.124

OPGAVE

Beschrijf de richting en de sterkte van de relatie tussen sociaal leed en hersenactiviteit. Suggereren de gegevens dat hersenactiviteit in het 'pijn'-gebied werkelijk direct is gerelateerd aan pijn vanwege sociale uitsluiting?

11. De stofwisselingssnelheid, de snelheid waarmee het lichaam energie verbruikt, is van belang bij onderzoek naar gewichtstoename, diëten en lichaamsbeweging. In onderstaande tabel vind je de gegevens over het vetvrij lichaamsgewicht (=gewicht van een persoon zonder vet) en de stofwisselingssnelheid in rust voor 7 vrouwen en 7 mannen die als proefpersoon bij een onderzoek naar afvallen zijn betrokken. De stofwisselingssnelheid wordt gemeten als het aantal calorieën dat per 24 uur wordt verbrand. De onderzoekers denken dat het vetvrij lichaamsgewicht een belangrijke invloed op de stofwisselingssnelheid heeft.

Proefpersoon	Biologisch geslacht	Vetvrij lichaamsgewicht	Stofwisselingssnelheid
1	M	62.0	1792
2	M	62.9	1666
3	V	36.1	995
4	V	54.6	1425
5	V	48.5	1396
6	V	42.0	1418
7	M	47.4	1362
8	V	50.6	1502
9	V	42.0	1256
10	M	48.7	1614
11	V	40.3	1189
12	M	51.9	1460
13	M	51.9	1867
14	M	46.9	1439

OPGAVE:

- Is de samenhang tussen 'vetvrij lichaamsgewicht' en 'stofwisselingssnelheid' positief of negatief?
- Hoe sterk is de samenhang?
- Is de richting en de sterkte van de samenhang verschillend voor mannen en vrouwen?
- Bepaal de helling van de regressielijn voor de stofwisselingssnelheid ten opzicht van het vetvrij lichaamsgewicht.
- Bepaal de helling van de regressielijn voor het vetvrije lichaamsgewicht ten opzichte van de stofwisselingsratio.

12. Voor verschillende ontwikkelingslanden is koffie een belangrijk exportartikel. Wanneer de koffieprijzen hoog zijn, kappen de boeren vaak bossen om meer koffiebomen te planten. Hieronder vind je de gegevens van de prijzen die aan koffieverbouwers in Indonesië werden betaald en de mate van ontbossing in een nationaal park dat in een koffieproducerende regio ligt, beide over een periode van vijf jaar.

PRIJS (dollarcenten per Am. pond[3])	ONTBOSSING (percent)
29	0.49
40	1.59
54	1.69
55	1.82
72	3.10

OPGAVE

- Wat is de verklarende variabele?

- Hoeveel bedraagt Pearson's correlatie?

- De prijs van koffie werd uitgedrukt in dollars. Als de koffie in euro's was geprijsd en de dollarprijzen in de tabel hierboven waren vertaald in equivalenten euro's, zou de correlatie tussen koffieprijs en percentage ontbossing dan veranderen? Verklaar jouw antwoord.

[3] 1 Amerikaanse pond is 0.45359237 kg.

13. Een studente vraagt zich af of mensen van overeenkomstige lengte geneigd zijn met elkaar uit te gaan. Ze meet haar eigen lengte, die van haar kamergenote en die van de vrouwen in de naastgelegen kamers. Vervolgens meet ze de lengte van de eerste man waarmee elke vrouw uitgaat. Hieronder zijn de data (lengte in inches[4]):

Vrouwen	66	64	66	65	70	65
Mannen	72	68	70	68	71	65

OPGAVE

- Bepaal de correlatie tussen de lengtes van mannen en vrouwen.
- Als de lengtes in centimeters waren gemeten, in plaats van in inches, in welk opzicht verandert de correlatie dan?

[4] 1 inch is 2.54 cm.

14. WAAR of FOUT ?

Uitspraak	WAAR	FOUT
Als we de samenhang berekenen tussen twee nominale variabelen met elk twee categorieën, dan bedraagt de waarde voor Cramers V dezelfde als Phi.		
Cramers V is gebaseerd op Chi-kwadraat.		
Chi-kwadraat is een goodness-of-fit maat.		
De regressiecoëfficiënt bepaalt of je rechte daalt of stijgt. b > 0 : rechte stijgt b < 0 : rechte daalt		
Als $R^2 = 1$: is Y perfect onvoorspelbaar.		
Bij variabelen gemeten op twee verschillende parametrische meetniveaus dient men sensu stricto nominale associatiematen te gebruiken.		
De gekwadrateerde correlatiecoëfficiënt R^2 zegt welk deel van de variantie in Y verklaard wordt door X.		
Als variabele X gemeten is op het nominale niveau en variabele Y op het ordinale niveau, dan is gamma de beste oplossing.		
Spearman's rho is gebaseerd op de Pearson's r, maar dan voor ordinale variabelen.		
Bij een r* k tabel is Phi gelijk aan V.		
Het is zinloos om de product-moment correlatiecoëfficiënt r te berekenen als je een curvilineair verband ontdekt.		
De verhouding tussen de regression sum of squares en de residual sum of squares is gelijk aan de determinatiecoëfficiënt.		
De residuele term is het verschil tussen de observatie en de voorspelling.		
Positieve residuen liggen boven de regressielijn.		

15. *Blijf je slank door wiebelen en draaien?*

Sommige mensen worden niet zwaarder, zelfs als zij zich overeten. Een mogelijke verklaring is dat activiteiten als wiebelen en draaien en andere niet-sportmatige activiteiten hieraan ten grondslag liggen.

Onderzoekers gaven 16 gezonde jonge volwassenen 8 weken lang te veel voeding. Zij maten de *gewichtstoename* (in kilo's) en hanteerden als een verklarende variabele de toename in *energieverbruik* (in calorieën) uit activiteiten zoals wiebelen en draaien en andere niet-sportmatige dagelijkse activiteiten. Hieronder vind je de gegevens

ENERGIEVERBRUIK (in calorieën)	GEWICHTSTOENAME (in kilo's)
-94	4.2
-57	3.0
-29	3.7
135	2.7
143	3.2
151	3.6
245	2.4
355	1.3
392	3.8
473	1.7
486	1.6
535	2.2
571	1.0
580	0.4
620	2.3
690	1.1

OPGAVE

- Wat is de onafhankelijke variabele?
- Wat is de afhankelijke variabele?
- Hoeveel bedraagt Pearson's correlatiecoëfficiënt? Interpreteer.
- Bepaal de bivariate regressievergelijking.
- Wat is de meeteenheid van de ongestandaardiseerde richtingscoëfficiënt.
- Interpreteer de betekenis van de ongestandaardiseerde richtingscoëfficiënt.
- Hoeveel bedraagt de gewichtstoename voor een individu van wie het energieverbruik met 400 calorieën toeneemt?

16. Een universiteitskrant interviewt een psycholoog over de door studenten beoordeelde kwaliteit van het lesgeven door stafleden. De psycholoog zegt: *"De gegevens tonen aan dat de correlatie tussen de onderzoeksproductiviteit en de onderwijskundige beoordeling van een staflid dicht bij 0 ligt."* De krant maakt hier het volgende van: "*Professor McDaniel vindt dat goede onderzoekers meestal slechte leraren zijn, en andersom."*

    Leg uit waarom dit bericht onjuist is. Geef in jouw eigen woorden de bedoelingen van de psycholoog weer.

17. Controlevragen

- Waar staan in een kruistabel de afhankelijke en de onafhankelijke variabele meestal?

- Stel, je brengt de gegevens van de vakantiekeuze van 1ste Bac-Criminologie studenten in kaart (goedkoop versus duur). Ook weet je of deze studenten een studielening hebben of niet. Waar zet je deze variabelen in de kruistabel?

- Stel dat je wilt weten of er onder de studenten die een lening hebben, vaker dure dan goedkope vakantie worden geboekt. Welke manier van percenteren kies je dan (rijpercentage of kolompercentages)?

- Wanneer kun je een Chi-kwadraattoets inzetten?

- Welk meetniveau moeten de variabelen minimaal hebben?

- Stel, je wilt de samenhang onderzoeken tussen aantal jaren opleiding en inkomen. Kun je dan ook een Chi-kwadraattoets gebruiken?

- Hoe interpreteer je de Spearman-rangcorrelatie in vergelijking met de Pearson correlatiecoëfficiënt?

- Stel, je vindt een correlatiecoëfficiënt van 0.45. Hoe sterk is het gevonden verband ?

- Welke waarden kan een correlatiecoëfficiënt aannemen?

- Stel, je vindt tussen 'fietsen' en 'conditie' een verband van 0.70. Hoe zou je dit interpreteren?

- Wat is het verschil tussen 'een verband' en 'een effect'?

- Wanneer gebruik je een regressieanalyse?

- Wat betekent 'de constante' in de regressievergelijking?

- Wat is de functie van een regressiecoëfficiënt?

- Hoe kun je de verklaarde variantie definiëren?

- Wat zegt de verklaarde variantie over het regressiemodel?

- Stel, je vindt een R^2 van 60%. Hoe sterk is je model dan?

- Hoe reken je de residuele waarde uit?

- Stel dat je het *relatieve netto*-effect van een variabele wilt analyseren in relatie tot andere effecten. Welke coëfficiënt gebruik je dan?

- Wat is het verschil in interpretatie tussen de B en Beta-coëfficiënt?

18. UITSPRAKEN: JUIST OF FOUT?

Hangt het dagelijks eten van fruit samen met een betere gezondheid (gemeten aan de hand van aantal ziektedagen)?	WAAR	FOUT

EFFECT VAN FRUIT OP ZIEKTE

	Beta	b	a
FRUIT	**-0,58**	**-1,14**	**6,21**

Afgerond 34% van de variantie in ziekte kan verklaard worden door de variantie in het eten van fruit.		
Voor iemand die 0 dagen thuisblijft wegens ziekte, verwachten we dat hij/zij gemiddeld 6.21 stukken fruit eet.		
Bij toename van 1 standaardafwijking in het aantal stukken fruit dat men eet, verwachten we een afname met 1.14 standaardafwijkingen in aantal ziektedagen.		
Joris at gemiddeld 2.5 stukken fruit per dag en bleef 4 dagen thuis. Het residu bedraagt afgerond 0.64.		
Als Els gemiddeld per dag 2 stukken fruit meer eet dan Marie, dan verwachten we dat Els afgerond 2 dagen minder thuis blijft wegens ziekte dan Marie.		