7. De correlatie

De relatie tussen twee variabelen kan aangegeven worden met de correlatie. Twee variabelen kunnen op drie manieren gerelateerd zijn aan elkaar: (1) positief, (2) niet gerelateerd en (3) negatief. Een positief verband betekent dat een toename in de ene variabele samenhangt met een toename in de andere variabele. Een negatief verband betekent dat een toename in de ene variabele samenhangt met een afname in de andere variabele. Niet gerelateerd betekent dat er geen samenhang is tussen de variabelen.

Aan het begin van de correlatieanalyse maak je eerst een scatterplot om de relatie tussen de variabelen te bekijken. Zoals in hoofdstuk 2 al is aangegeven, komt veel in de statistiek neer op één model, het general linear model, met de formule:

Uitkomst = model + error.

Wat dit model inhoudt, hangt af van wat je wil onderzoeken. Bij een model dat de relatie tussen twee variabelen weergeeft, zal je de ene variabele (de uitkomst) willen voorspellen met de andere variabele (de predictor). Het model wordt dan:

Uitkomst = bX + error.

Predictorvariabelen worden aangegeven met de letter X. De parameter b staat voor de relatie tussen de twee variabelen. De data uit de steekproef wordt gebruikt om de b te schatten. Als je slechts één predictorvariabele hebt, is b de Pearson correlatie coëfficiënt (r).

Hoe meet je de relatie?

Covariantie

De makkelijkste manier om te kijken of er een relatie is tussen twee variabelen, is te kijken naar de covariantie. Als twee variabelen met elkaar covariëren zijn ze geassocieerd aan elkaar. Om te weten wat *covariantie* is moeten we terug naar de variantie.

Als er een relatie is tussen twee variabelen zijn veranderingen in de ene variabele merkbaar in de andere variabele. Dat betekent dat als de ene variabele afwijkt van zijn gemiddelde, de andere variabele dat ook doet.

Om de gelijkheid van de patronen te berekenen, kunnen we de totale deviaties berekenen. Maar dit geeft het inmiddels bekende probleem waarbij positieve en negatieve deviaties elkaar uitwissen. Als je slechts een onafhankelijke variabele hebt, kun je de deviaties kwadrateren en daarmee de SS berekenen. Bij meerdere predictors kun je de deviaties van de ene variabele vermenigvuldigen met de deviaties van de andere variabele. Dan krijgen we de *kruisproduct deviaties*. Als dit dan nog door het aantal observaties -1 wordt gedeeld krijgen we de covariantie. In formulevorm:

Covariantie (x,y) = $\frac{\sum (X - \overline{X})(Y - \overline{Y})}{N - 1}$

Dit is dezelfde formule als waarmee je de variantie berekent, behalve dat je nu de deviaties van de twee variabelen vermenigvuldigt.

De covariantie geeft aan of variabelen samenhangen, een positieve covariantie betekent een positieve relatie en een negatieve covariantie betekent een negatieve relatie.

Een probleem is wel dat de covariantie afhankelijk is van de schaal waarop de variabelen gemeten zijn, het is niet gestandaardiseerd. Hierdoor kunnen we niet zeggen of het getal groot of klein is in vergelijking met andere metingen.

De correlatie coëfficiënt en standaardisatie

Om dit probleem met de covariantie op te lossen, wordt gebruik gemaakt van *standaardisatie*, de covariantie wordt omgezet in een standaardmeting. Dit doe je met de standaardafwijking. Omdat je twee standaardafwijkingen hebt (van beide variabelen), moet je deze met elkaar vermenigvuldigen. De gestandaardiseerde covariantie is de correlatie coëfficiënt (r). De formule hiervoor is te vinden op pagina 266.

Deze r wordt de *Pearson correlatie coëfficiënt* genoemd. De uitkomst moet tussen de -1 en de +1 vallen. Een score van +1 betekent dat de variabelen perfect positief gecorreleerd zijn aan elkaar. Een score van -1 betekent dat de variabelen perfect negatief aan elkaar gecorreleerd zijn. Een score van 0 betekent dat de variabelen niet gecorreleerd zijn en dat er geen relatie is.

r is dus een correlatie, maar het wordt ook gebruikt voor het meten van de effectgrootte, omdat een correlatie zegt hoe sterk het verband is tussen twee variabelen. r=0.1 is een klein effect, r=0.3 is een gemiddeld effect en r=0.5 is een groot effect.

Een correlatie tussen twee variabelen heet een *bivariate correlatie.*

Significantie

Wetenschappers toetsen de hypothesen meestal met kansen. Je test de hypothese dat je r significant afwijkt van 0 (=geen relatie). We kunnen op twee manieren de hypothesen toetsen. De eerste manier is de z-score berekenen. De Pearson r is niet normaal verdeeld, maar Fisher heeft een formule gevonden waarin r als normaal beschouwd kan worden. Deze formule is te vinden op pagina 268, maar het opzoeken is niet de moeite waard als je alleen een hypothese wil toetsen, omdat we deze methode vrijwel niet gebruiken.

We kunnen de hypothese namelijk ook toetsen met een t-toets. De t-toets heeft N-2 vrijheidsgraden en r kan er direct in toegepast worden.

$$t_r = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}}$$

Betrouwbaarheidsintervallen

SPSS kan geen betrouwbaarheidsintervallen voor r uitrekenen, maar wel bootstrap betrouwbaarheidsintervallen. Dit betrouwbaarheidsinterval is ook accuraat als de steekproefverdeling niet normaal verdeeld is. Als je een gewoon betrouwbaarheidsinterval wil, kun je het handmatig uitrekenen. Je krijgt dan een betrouwbaarheidsinterval in z-scores. Dit kun je weer omzetten in correlaties. De formules die je hiervoor nodig hebt, zijn te vinden op bladzijde 269.

Interpretatie (voorzichtig met causaliteit)
Een correlatie betekent niet per se causaliteit!

Causaliteit bij correlaties kan voor twee redenen niet aangenomen worden:

- Het derde variabele probleem. Causaliteit tussen twee variabelen kan niet aangenomen worden omdat ook nog andere variabelen invloed kunnen hebben op de correlatie.
- De richting van de causaliteit. Een correlatie zegt niets over welke variabele de verandering bij de andere variabele teweegbrengt.

Correlatie analyse in SPSS

Voor elke variabele die je gemeten hebt, maak je een aparte kolom in SPSS. Elke rij stelt een proefpersoon voor.

Bivariate correlatie

Voordat je begint met het uitvoeren van analyses, moet je checken of aan je assumpties (specifiek die van lineariteit en normaliteit) is voldaan. Zo ja, dan kun je een Pearson correlatie uitvoeren. Als er uitschieters zijn, of het is niet normaal verdeeld, kun je kiezen voor een bootstrap betrouwbaarheidsinterval, de Spearman correlatie of Kendall's tau (zie verderop).

Voor de bivariate correlatie in SPSS ga je naar analyze - correlate - bivariate. De Pearson correlatie is automatisch door SPSS al aangevinkt. Op bladzijde 273 staat de dialog box van SPSS met alle opties verder uitgelegd.

Van alle variabelen die je naar variables sleept, wordt een combinatie gemaakt en alle correlaties zijn te zien in een tabel. De tabel wordt een correlatiematrix genoemd. De correlaties staan allemaal dubbel in de matrix, dus de ene helft (diagonaal) kun je negeren. Significante correlaties geeft SPSS aan met een (dubbele) asterisk (*).

In de output staat ook het bootstrap betrouwbaarheidsinterval. Als je variabelen niet normaal verdeeld zijn, kun je hiernaar kijken in plaats van naar de significantie. Als de 0 (geen effect) in het betrouwbaarheidsinterval zit, is er geen significante correlatie. Valt de 0 niet in het interval, heb je wel een significante correlatie.

Het gebruik van R^2

Het kwadraat van de correlatie is de *coëfficiënt van determinatie* R^2. Het meet het percentage gedeelde variantie. Het is een maat voor hoeveel variantie de gecorreleerde variabelen delen. R^2 wordt vaak gerapporteerd als dat een bepaald percentage variantie verklaard wordt door de andere variabele. Dit impliceert echter causaliteit, terwijl dat niet het geval hoeft te zijn bij een correlatie.

Spearman 's rho

Een non-parametrische versie van de Pearson correlatie is de *Spearman correlatie coëfficiënt,* r_s.

Zoals de eerder besproken non-parameterische toetsen is deze statistiek gebaseerd op rangscores, en is daarom bestand tegen schendingen van de assumpties. Bij deze toets ken je eerst rangscores toe aan de gegevens en pas je daarna de Pearson vergelijking toe.

In SPSS ga je naar hetzelfde scherm als bij de Pearson correlatie. Alleen zet je nu het vinkje bij Pearson uit, en vink je Spearman aan. Voor een voorbeeld van de SPSS output zie blz. 277. Om een robuust betrouwbaarheidsinterval te genereren, gebruik je ook hier de optie bootstrap.

Kendall's tau

Kendall's tau τ is handig wanneer je een kleine steekproef hebt met veel gelijke scores, dus tied ranks. Ook dit is een non-parametrische statistiek en is dus bestand tegen schendingen van assumpties. Voor deze statistiek vink je in het correlatie dialog scherm de optie Kendall's tau-b aan in plaats van Pearson of Spearman. Een voorbeeld van de SPSS output is op blz. 278 te vinden.

Biseriële en punt biseriële correlaties

Deze correlaties worden gebruikt als één van de twee variabelen dichotoom is. Een variabele kan op twee manieren dichotoom zijn. Het kan discreet zijn, waarbij er maar twee categorieën mogelijk zijn en er geen andere opties tussen liggen. De andere dichotome manier is continu waarbij er twee categorieën in de test zijn maar in werkelijkheid nog meer tussen opties zijn.

Een voorbeeld van een discreet dichotome variabele is of iemand dood is of leeft. Iemand kan niet erger dood zijn dan een ander en je kunt niet een beetje dood zijn. Een voorbeeld van een continu dichotome variabele is of iemand zijn statistiek tentamen heeft gehaald. Je kunt hierbij op het randje gefaald hebben, of met een heel goed cijfer geslaagd. Hoewel je het opdeelt in twee categorieën, is er een onderliggend continuüm.

De *punt-biseriële correlatie coëfficiënt* (r_{pb}) wordt gebruikt als de variabele discreet dichotoom is, de *biseriële correlatie coëfficiënt* (r_b) wordt gebruikt als de variabele continu dichotoom is. Om in SPSS de biseriële correlatie te krijgen moet je eerst de punt-biseriële correlatie berekenen.

Bij een punt-biseriële correlatie, voer je gewoon een Pearson correlatie uit, waarbij de dichotome variabele gecodeerd is met 0 en 1. Of je correlatie negatief of positief is, hangt af van welke categorie je in de dichotome variabele welk cijfer toekent.

De partiële correlatie

De partiële correlatie is de relatie tussen twee variabelen waarin de effecten van een andere (derde) variabele constant worden gehouden. Dit doe je op het moment dat je drie variabelen hebt, die allemaal met elkaar samenhangen.

Stel, je wil de prestatie op een tentamen meten, maar hiervoor heb je twee samenhangende variabelen, nakijktijd en examenangst. Deze variabelen hebben dan een unieke gedeelde variantie met de prestatie, en een gedeelde variantie die overlapt, omdat angst en nakijktijd ook samenhangen. Als je dit overlappende stuk variantie weghaalt, kun je kijken wat de unieke relatie is tussen bijvoorbeeld examenangst en prestatie. Op bladzijde 282 staat een grafische weergave van wat een partiële correlatie precies inhoudt.

Voor de partiële correlatie in SPSS ga je naar analyze - correlatie - partial. Bij options kan je aangeven dat je de zero-order correlaties wilt hebben. Dit zijn de bivariate correlaties zonder de controle van andere variabelen. Bij deze optie krijg je zowel de bivariate correlaties als de partiële correlatie. Zo kan je het verschil tussen de twee correlaties goed vergelijken.

Semi-partiële correlaties

Het verschil tussen de *semi-partiële correlatie* (of *part correlatie)* en de partiële correlatie is dat de semi-partiële correlatie controleert voor het effect dat een derde variabele heeft op één van de variabelen in de correlatie (en niet op beide variabelen, zoals bij de partiële correlatie). Partiële correlaties zijn vooral handig als je de unieke bijdrage van de variabelen wilt weten, dus kijken wat de relatie tussen X en Y is zonder Z. Semi-partiële correlatie is handig als je de variantie van een bepaalde variabele wilt verklaren.

Het vergelijken van correlaties

Het vergelijken van onafhankelijke correlaties (bijvoorbeeld die van mannen en vrouwen) kan door de correlaties om te zetten in z_r-scores. Met deze z_r kan de z-score van de verschillen uitgerekend worden.

Voor de formule voor het berekenen van deze Z-verschilscore, zie pagina 286. In de tabel in de appendix kun je kijken welke p-waarde hoort bij die z-score.

Voor afhankelijke correlaties, zoals afkomstig van dezelfde proefpersonen, kan een t-toets gebruikt worden. Hier is een formule voor, die je eventueel zou kunnen opzoeken op pagina 287.

Het berekenen van de effectgrootte

Correlaties zijn eigenlijk effectgroottes, dus hiervoor zijn geen verdere berekeningen nodig. Maar bij non-parametrische correlaties zijn er een paar verschillen. Je kunt Spearman's rho gewoon kwadrateren om het percentage gedeelde variantie te krijgen, omdat je dezelfde formule gebruikt als bij de Pearson correlatie.

Spearman's R^2 is de proportie verklaarde variantie van de rangscores.

Kendall's τ is echter niet vergelijkbaar met r, en mag niet gekwadrateerd worden om de gedeelde variantie te berekenen. De keuze van de correlatiecoëfficiënt maakt veel uit voor de effectgrootte, dus wees hierop bedacht.

Het rapporteren van correlaties

Bij een correlatie hoef je alleen te melden hoe groot de correlatiecoëfficiënt is en of het een significante waarde heeft. Bij de correlatie hoef je geen 0 te noteren voor de coëfficiënt, omdat men weet dat het toch niet boven de 1 gaat, dus een punt voor het getal is genoeg. Rapporteer de coëfficiënt met 2 of 3 decimalen achter de komma. Elke correlatie heeft een eigen letter.