Hoorcollege: Experimenteel onderzoek & experimentele controle, z-scores en Peason *r* (SPSS)

Verband tussen twee variabelen

Er zijn twee vormen van samenhang tussen twee variabelen: samenhang en afhankelijkheid. Bij afhankelijkheid is er spraken van causaliteit: een onafhankelijke variabele veroorzaakt een afhankelijke variabele. Ook is er bij afhankelijkheid sprake van een voorspelling (regressie); je kunt voorspellen wat de onafhankelijke variabele voor invloed heeft op de afhankelijke variabele. Bij samenhang gaat het puur om correlatie; je kunt geen uitspraak doen over wat de correlatie nou veroorzaakt.

Om de samenhang tussen twee variabelen vast te stellen is het belangrijk om dezelfde n cases (de totale populatie/alle observaties) te nemen voor beide variabelen. Het is niet nuttig om de samenhang tussen lengte en gewicht te bepalen, waarbij je bij 20 mensen de lengte hebt en bij 16 mensen het gewicht hebt gemeten. Je kunt variabelen met verschillende meetniveaus met elkaar vergelijken. Je kunt de samenhang tussen twee numerieke variabelen, de samenhang tussen een categorische en een numerieke variabele en de samenhang tussen twee categorische variabelen. vergelijken Hierbij hebben de numerieke variabelen een meetniveau van interval of hoger en hebben de categorische variabelen een nominaal of ordinaal meetniveau. Je kunt stellen dat twee waarden 1 en 2 met elkaar zijn geassocieerd als waarde 1 vaker optreedt met een waarde van 2 dan met andere waarden.

Scatterplot

Je kunt de samenhang tussen variabelen globaal aflezen in een scatterplot. Je kunt de relatie tussen variabelen beschrijven met behulp van een scatterplot met een drietal dingen:

Richting

Als er een positieve relatie is tussen twee variabelen zal een hogere van variabele 1 samengaan met een hogere variabele van 2 en andersom.

Als er een negatieve relatie is tussen variabelen zal een hogere score van variabele 1 samengaan met een lagere score van variabele 2. Bij een positieve relatie zullen de punten op de scatterplot (vaak heel globaal) een opgaande lijn voorstellen en bij een negatieve relatie juist een dalende lijn.

Vorm

Lineair verband: bij een lineair verband volgen de punten op een scatterplot min of meer een rechte lijn.

Niet-lineair verband: bij een niet-lineair verband volgen de punten op een scatterplot juist geen rechte lijn, maar is er een ander verband aanwezig.

Daarnaast valt er onderscheid te maken tussen hetero –en homogeniteit. Homogeniteit betekent dat alle punten op de scatterplot min of meer in hetzelfde cluster liggen. Heterogeniteit betekent dat er meerdere clusters met punten zijn.

Sterkte

Als je een volledig rechte lijn door de punten van je scatterplot kan trekken is er een perfect verband tussen de variabelen. Er is vrijwel altijd spreiding in de observaties (er zal vrijwel nooit een perfecte lijn in de scatterplot te trekken zijn). Als de punten op de scatterplot compleet willekeurig zijn verdeeld betekent dit dat er geen verband is tussen de variabelen.

Ten slotte heb je ook met scatterplots te maken met uitbijters: observaties die ver afwijken van het algemene patroon. Uitbijters kunnen de sterkte van een verband beïnvloeden (ze kunnen de denkbeeldige lijn die je kunt trekken door de punten op de scatterplot steiler of minder steil maken). Je kunt je data plotten zonder uitbijters om dan het verband tussen de variabelen te bekijken in een scatterplot, maar je mag niet zonder reden uitbijters volledig uit de dataset verwijderen.

Covariantie

Een maat die gebruikt kan woorden om de sterkte van de samenhang tussen twee variabelen uit te drukken is de covariantie. De formule voor de covariantie is:

$$s_{yx}^2 = \frac{\sum(Y_i - Y)(X_i - X)}{N - 1}$$

De formule lijkt sterk op de formule van de variantie. De variantie is echter niet geschikt voor bivariate data (twee variabelen). De covariantie voegt de variantie van twee variabele samen in één formule. Bij de covariantie in beeld op een scatterplot zie je in de tabel de afwijking van het gemiddelde van x en y aangegeven met pijltjes. Let op: kruisproducten en covarianties kunnen ook een negatieve waarde aannemen; dit wijst op een negatief verband tussen de variabelen. Een nadeel van de covariantie is dat het slecht interpreteerbaar is. Al bereken je bijvoorbeeld lengte in centimeters zal de covariantie veel groter zijn dan als je meters gebruikt als meeteenheid. Hierdoor gaat de voorkeur vaak uit naar een gestandaardiseerde maat die de sterkte van samenhang tussen twee variabelen meet, bijvoorbeeld de Pearson (r).

Pearson (r)

De pearson (r) is een maat voor samenhang tussen twee intervalvariabelen. De correlatie kan tussen -1 en 1 liggen. -1 betekent een perfecte negatieve correlatie, 1 betekent een perfecte positieve correlatie en 0 betekent geen correlatie. De pearson (r) is gestandaardiseerd en verandert dus niet als je de meeteenheid verandert (bijv. Lengte in meters -> lengte in centimeters). De pearson (r) kan alleen gebruikt worden voor lineaire verbanden; als je hem toepast bij andere verbanden is de maat niet bruikbaar. Je kunt een scatterplot maken om te checken of er een lineair verband is. De pearson (r) is gevoelig voor uitbijters. De formule van de pearson (r) is:

rxy = (Sxy)/(SxSy)

De vuistregel voor de sterkte van de correlatie is:

small	medium	large
0.1	0.3	0.4

Kanttekeningen bij het vaststellen van correlatie. Er zijn vier dingen waar je rekening mee moet houden bij het vaststellen van correlatie:

- Niet-lineaire verbanden: correlatie betekent alleen iets bij een lineair verband.
- Uitbijters: uitbijters kunnen de richting van de regressielijn op de scatterplot sterk veranderen.
- Heterogene subgroepen: als je twee groepen met een verschillende gemiddelde samenvoegt zal dit de totale correlatie beïnvloeden.
- Restriction of range: als je niet de gehele range van alle punten op de scatterplot bekijkt kan er een ander verband tussen de variabelen waargenomen worden dan wanneer je wel de gehele range had genomen.

Correlatie is niet hetzelfde als causaliteit. Er zijn drie criteria voor causaliteit.

- variabelen moeten correleren
- Oorzaak komt voor gevolg in de tijd
- Alternatieve verklaringen zijn uitgesloten