Hoofdstuk 2: Verbanden

Inleiding
In het gebruik van statistiek komen we vaak relaties tussen variabelen tegen. De focus van dit hoofdstuk ligt op het leren hoe deze relaties op grafische en numerieke wijze beschreven kunnen worden. Er wordt gekeken naar grafische beschrijvingen, zoals de scatterplot. Deze geeft de relatie weer tussen twee kwantitatieve variabelen. Daarna kijken we naar numerieke samenvattingen voor deze relaties en grafische en numerieke methoden voor eht beschrijven van de relatie tussen twee categorische variabelen. Tot slot wordt nog aandacht besteed aan het onderscheid tussen associatie en causatie.

2.1 Relaties
We gebruiken de term *associatie* om de relatie tussen twee variabelen te beschrijven. Een voorbeeld is de relatie tussen gewicht en lengte.

- Twee variabelen zijn *geassocieerd* wanneer een waarde op de eerste variabele iets zegt over de waarde op de andere variabele.

Twee variabelen
Bij het bekijken van de relatie tussen twee variabelen is het doel van de onderzoeker van belang. Probeert de onderzoeker de relatie alleen maar bloot te leggen of hoopt hij of zij te ontdekken dat één van de variabelen variantie in de andere variabele verklaart? In het laatste geval is het handig om onderscheid te maken tussen *verklarende (explanatory variables)* en *responsvariabelen.*

- Een *responsvariabele* is gerelateerd aan de uitkomsten van een onderzoek. Een onderzoeker wil bijvoorbeeld weten of lengte invloed heeft op gewicht. In dit geval is gewicht de responsvariabele.
- Een *verklarende variabele* verklaart of veroorzaakt veranderingen in de responsvariabelen. In ons voorbeeld is lengte de verklarende variabele.

Een beschrijving van de belangrijkste eigenschappen van een dataset die gebruikt wordt om de relatie tussen twee variabelen moet in ieder geval de volgende punten bevatten:

- *Cases*. Identificeer de cases an hoeveel er zijn in de dataset.
- *Label*. Identificeer welke variabele als label-variabele gebruikt wordt (als er één is).
- *Categorisch of kwantitatief*. Classificeer elke variabele als categorisch of kwantitatief.
- *Waarden*. Identificeer de mogelijke waarden voor elke variabele.
- *Verklarend of respons*. Wanneer toepasbaar, classificeer elke variabele als verklarende of als responsvariabele.

Causaliteit
Veel onderzoekers zijn geïnteresseerd in hoe verklarende variabelen veranderingen in de responsvariabelen *veroorzaken.* Veel relaties tussen verklarende- en responsvariabelen gaan echter niet over een directe vorm van causaliteit. Een motivatietest voor een sollicitant *voorspelt* misschien wel in welke mate deze persoon gemotiveerd zou zijn als hij of zij aangenomen wordt, maar een motivatietest *veroorzaakt* niet de motivatie om goed te presteren.

Vaak worden verklarende variabelen ook wel *onafhankelijke (independent)* variabelen genoemd. Responsvariabelen worden ook wel *afhankelijke (dependent)* variabelen genoemd. Wanneer dit gebeurt, beschrijven deze termen wiskundige ideeën, het zijn geen statistische termen. De principes die het werk onderbouwen blijven hetzelfde:

1. Begin met een grafische weergave van de data.
2. Kijk naar algemene patronen en afwijkingen van deze patronen.
3. Gebaseerd op wat je ziet, kun je numerieke smenvattingen gebruiken om specifieke aspecten van de data te beschrijven.

2.2 Puntgrafieken

Puntgrafiek (scatterplot)
Grafisch wordt de relatie tussen twee kwantitatieve variabelen vaak in een *puntgrafiek* verwerkt. De twee variabelen moeten wel bij dezelfde individuen gemeten worden.

- De waarden van de ene variabele worden op de X-as gezet, terwijl de waarden van de andere variabele op de Y-as staan. Elk individu in de data wordt als een punt in de grafiek verwerkt op basis van de scores die de persoon op de X-as en de Y-as heeft behaald.
- De verklarende variabele hoort bij de X-as. Om deze reden wordt de verklarende variabele ook wel de X-variabele genoemd. De responsvariabele wordt op de Y-as gezet. We noemen zo een variabele daarom ook wel een Y-variabele.
- Als er geen onderscheid is tussen verklarende- en responsvariabelen, dan maakt het niet uit welke variabele op de X-as belandt en welke variabele op de Y-as belandt.

De interpretatie van puntgrafieken
Om een eerste indruk van een puntgrafiek te krijgen, is het handig om:

- Het *algemene patroon* en *afwijkingen* te bekijken.
- De *vorm, richting* en *de sterkte* van de relatie te beschrijven.
- Oog te hebben voor *uitbijters.* Dit zijn individuele waarden die buiten het algemene patroon vallen.

Het is mogelijk dat er *clusters* in de puntgrafiek waar te nemen zijn. Dit betekent dat de data verschillende soorten individuen beschrijven.

Soorten verbanden
De relatie tussen twee variabelen kan positief of negatief zijn.

- Twee variabelen zijn *positief geassocieerd* wanneer hoge scores op de ene variabele samengaan met hoge scores op de andere variabele. Een voorbeeld is dat een hoge score op lengte vaak samengaat met een hoge score op gewicht.
- Twee variabelen zijn *negatief geassocieerd* wanneer hoge scores op de ene variabele gepaard gaan met lage scores op de andere variabele. Er is bijvoorbeeld een negatief verband tussen faalangst en prestatie op een tentamen. Hoe meer faalangst iemand heeft, hoe lager hij of zij zal scoren op een tentamen.

Wanneer er verschillende clusters in een puntgrafiek waar te nemen zijn, is het vaak handig om het patroon van elk cluster te beschrijven. In puntgrafieken zijn vaak *lineaire relaties* te ontdekken; de punten liggen dan ongeveer op een rechte lijn. De sterkte van een relatie wordt bepaald door te kijken naar de mate waarin punten in de grafiek bij elkaar in de buurt liggen. Veel spreiding gaat dus samen met een zwakke samenhang. Als je een categorische variabele aan de puntgrafiek wilt toevoegen, dan is het handig om verschillende kleuren of symbolen voor elke categorie te gebruiken. Om een duidelijke relatie in de punten te ontdekken, is het mogelijk om de grafiek als het ware *vloeiend te maken (smoothing).* Dit kan door middel van software gedaan worden. Er wordt dan een lijn door de punten getrokken. Deze lijn past het beste bij de gevonden x- en y-waarden.

2.3 Correlatie

Correlatie

Kort samengevat kan dus gezegd worden dat een puntgrafiek de vorm, richting en de sterkte van een relatie tussen twee kwantitatieve variabelen beschrijft. Het kan soms misleidend zijn om met het blote oog uitspraken te doen over de sterkte van een relatie. Door het veranderen van de getallen op de assen kan het namelijk lijken alsof er een zeer sterke samenhang is, terwijl dat niet zo hoeft te zijn. Het omgekeerde is overigens ook mogelijk. Om deze reden gebruiken we de *correlatiemaat.*

- De correlatie meet de richting en de sterke van een lineaire relatie tussen twee kwantitatieve variabelen. Vaak wordt de letter r gebruikt om de correlatie te beschrijven.
- Stel: we hebben data verzameld voor variabelen X en Y voor n aantal personen. De gemiddelden en standaarddeviatie van de twee variabelen zijn dan $\bar{x}$ en s_x voor de x-waarden en y (met een streepje erboven) en s_y voor de y-waarden.
- De correlatie r tussen X en Y is:

$$r = \frac{1}{n-1}\sum\left(\frac{x-\bar{x}}{s_x}\right)\left(\frac{y-\bar{y}}{s_y}\right)$$

  Eerst wordt van elke individuele score dus het gemiddelde van de variabele afgetrokken. Daarna wordt dit getal door de bijbehorende standaarddeviatie gedeeld. In feite worden alle scores op X en Y dus gestandaardiseerd.

Kenmerken van de correlatie

- r is negatief wanneer er sprake is van een negatieve associatie en r is positief wanneer er sprake is van een positieve samenhang.
- Correlatie maakt geen onderscheid tussen verklarende- en responsvariabelen. Het maakt, voor het berekenen van de correlatie dus niet uit welke variabele je X en welke variabele je Y noemt.
- Om een correlatie uit te rekenen moeten allebei de variabelen *kwantitatief* zijn.
- Omdat r gebruik maakt van gestandaardiseerde waarden, verandert de correlatie niet als de meeteenheden van X, Y of beide worden veranderd. Het meten van lengte in centimeters of meters en het meten van gewicht in kilogram of pond, verandert dus niets aan de correlatie. De correlatie r heeft zelf geen meeteenheid; het is alleen een getal.

- De correlatie r is altijd een getal tussen de -1 en de 1. Waarden rond de 0 laten zien dat er sprake is van een zeer zwakke relatie. De sterkte van een relatie neemt toe naarmate r zich richting de -1 of 1 ontwikkelt. Dit betekent namelijk dat de waarden steeds meer op een rechte lijn gaan liggen en dat er weinig spreiding waar te nemen is. Een correlatie van -1 of 1 komt zelden voor en is extreem. In die gevallen liggen alle punten precies op een rechte lijn.
- Correlatie meet alleen de sterkte van de lineaire relatie tussen twee variabelen. Correlatie beschrijft dus niet de *gebogen (curved) relatie* tussen variabelen, hoe sterk deze ook is.
- Net zoals het gemiddelde en de standaarddeviatie, is ook de correlatie *niet robuust*: r wordt sterk beïnvloed door slechts een paar afwijkende scores.
- Correlatie is nooit een complete beschrijving van data waarbij twee variabelen voorkomen. Er moet onder andere ook gekeken worden naar de gemiddelden en standaarddeviaties.

2.4 Regressie

Regressielijnen

Als uit een puntgrafiek blijkt dat er sprake is van een lineaire relatie, dan willen we een zo goed mogelijk passende regressielijn ontwerpen die deze relatie beschrijft.

- Een *regressielijn* is een rechte lijn die beschrijft hoe een responsvariabele Y verandert als een verklarende variabele X verandert.
- We gebruiken een regressielijn vaak om de waarde van Y te *voorspellen* voor een gegeven waarde van X. Voor regressie is, in tegenstelling tot correlatie, wel van belang dat we een verklarende- en een responsvariabele hebben.

Een passende regressielijn vinden

Natuurlijk is er geen enkele rechte lijn die precies door alle punten van de puntgrafiek gaat. Een *lijn passend maken (fitting a line)* betekent dat we op zoek gaan naar een lijn die het beste in de buurt komt van alle punten. Stel dat Y een responsvariabele op de verticale as is en dat X een verklarende variabele op de horizontale as is. Een rechte lijn die Y aan X verbindt heeft dan de vorm van:

- $Y= b_0+ b_1x$.
- In deze formule is b_1 de *regressiecoëfficiënt (slope)* en is b_0 het *intercept.* Het intercept is de waarde van Y wanneer X nul is.

Extrapolatie (extrapolation)

Extrapoleren is het gebruik van een regressielijn om voorspellingen te doen die ver buiten de onderzochte waarden liggen. Je kunt bijvoorbeeld een puntgrafiek maken op basis van de lengte- en gewichtscores van een groep mensen. De langste persoon kan bijvoorbeeld 1.80 zijn. Als je wilt extrapoleren probeer je te voorspellen hoeveel iemand van bijvoorbeeld 1.95 weegt. Vaak leidt extrapolatie echter tot onbetrouwbare voorspellingen.

Minste-kwadraten-regressie (least-squares regression)

We willen dus een lijn vinden waarmee we waarden van Y zo goed mogelijk kunnen voorspellen op basis van waarden van X. De lijn moet zo goed mogelijk bij de punten liggen, maar wel in een *verticale* richting. Onze voorspellingen ($\hat{Y}$) zijn echter nooit perfect, er is altijd een mate van error.

- Error = geobserveerde score – voorspelde score. Fouten zijn positief als een geobserveerde respons (Y) boven de regressielijn ligt en negatief als een geobserveerde respons (X) onder de lijn ligt. We willen een lijn vinden die deze voorspellingsfouten zo klein mogelijk maakt. De meest gebruikte manier is de *minste-kwadraten-regressie (least-squares regression).*
- De *minste-kwadraten-regressielijn van Y op X* is de lijn die de som van kwadraten van de verticale afstanden (van de datapunten) zo klein mogelijk maakt. Om deze regressielijn te maken, moeten we eerst de waarden van b_0 en b_1 vinden, die samengaan met zo min mogelijk voorspellingsfouten: $\Sigma (\text{error})^2 = \Sigma(y_i - b_0 - b_1x_i)^2$.
- Vaak kan deze lijn door middel van computerprogramma's gevonden worden. Het is echter ook mogelijk om de regressielijn zelf te berekenen: $\hat{y}= b_0+ b_1x$. De waarde van b_1 wordt gevonden met de formule = $r \cdot \frac{s_y}{s_x}$. De waarde van b_0 wordt gevonden met de formule: $\bar{y} - b_1\bar{x}$.

Proportie verklaarde variantie (r^2)
Het kwadraat van de correlatie zegt ons hoeveel van de variantie in Y wordt verklaard door de regressielijn die hoort bij Y. Als een correlatie -1 of 1 is, dan is de proportie verklaarde variantie precies 1. Dit omdat dan alle variantie in Y perfect wordt verklaard door de bijbehorende regressielijn.

- Ook kan r^2 gezien worden als de variantie van de voorspelde scores ($\hat{Y}$) gedeeld door de variantie van de geobserveerde waarden (Y).

2.5 Beperkingen van correlatie en regressie

Residuen
Zelfs met een zo goed mogelijk passende regressielijn, liggen nooit alle punten precies op de lijn. Sommige punten worden dus niet goed voorspeld aan de hand van de regressielijn. De punten die afwijken van de regressielijn worden residuen genoemd.

- Een *residu* is het verschil tussen een geobserveerde waarde van een responsvariabele en de voorspelde waarde volgens de regressielijn: residu = y- $\hat{y}$. Het gemiddelde van alle residuen is altijd *nul.*
- Een *residu-plot* is een puntgrafiek van alle regressieresiduen ten opzichte van de verklarende variabele. Met zo een plot kan nagegaan worden in hoeverre een regressielijn goed past. Als de regressielijn past bij het algemene patroon van de data, dan zal er *geen patroon* aanwezig zijn in de residuen. Een uitbijter is een observatie die ver van het algemene patroon binnen een residu-plot ligt.
- Punten die *uitbijters* zijn in de Y-richting van een puntgrafiek hebben grote residuen, maar dat hoeft niet voor andere residuen te gelden.
- Een score is *invloedrijk (influential*) voor een rekenkundige berekening als de verwijdering ervan zou leiden tot een belangrijke verandering in de berekening. Punten die uitbijters in de X-richting zijn, hebben vaak invloed op de minste-kwadraten-regressielijn.
- De minste-kwadraten-regressielijn is, net zoals de correlatie, niet robuust.

Op de loer liggende variabelen (lurking variables)
De relatie tussen twee variabelen kan vaak het beste begrepen worden door ook naar de invloed van andere variabelen te kijken. Op de loer liggende variabelen kunnen een correlatie of een regressie misleidend maken.

- Een *op de loer liggende variabele (lurking variable)* is een variabele die niet in het onderzoek als een verklarende- of responsvariabele opgenomen is, maar toch de interpretatie van de relatie tussen deze variabelen kan beïnvloeden.

Correlatie en causaliteit
Een (sterke) relatie tussen een verklarende variabele (X) en een responsvariabele (Y) is geen bewijs voor het feit dat X veranderingen in Y *veroorzaakt.* Correlatie zegt dus niets over causaliteit. Daarnaast is het zo dat een correlatie die op de *gemiddelde scores* van individuen gebaseerd is vaak veel *hoger* is dan een correlatie die gebaseerd is op gewone scores. Ook kan er in sommige gevallen sprake zijn van het restricted-range probleem: de data bevat dan geen informatie over alle mogelijke scores op de verklarende- en de responsvariabele. In dat geval zullen de correlatie (r) en de proportie verklaarde variantie (r^2) *lager* uitvallen dan als alle mogelijke scores bij de data betrokken zouden worden. Onderzoekers maken vaak gebruik van meerdere verklarende variabelen. Een hoge score op een rekentoets (Y) kan bijvoorbeeld samenhangen met aanleg, maar ook met motivatie en opvoeding. Als een onderzoeker meerdere verklarende variabelen gebruikt, dan doet hij of zij aan *multipele regressie*. Er kan een correlatie tussen alle verklarende variabelen samen en de responsvariabele berekend worden. Deze correlatie wordt een *multipele correlatiecoëfficiënt* genoemd.

2.6 Data van tweewegtabellen (two- way tables)

Categorische data
Puntgrafieken zijn handig als er sprake is van kwantitatieve data. Bij categorische data dienen *tweewegtabellen* gebruikt te worden. Voorbeelden van categorische variabelen zijn sekse en beroep. Een tweewegtabel laat zien hoe vaak verschillende combinaties van twee categorische data voorkomen. Hoeveel mannen en hoeveel vrouwen zijn bijvoorbeeld psycholoog van beroep? En hoeveel mannen en vrouwen zijn dokter? Sekse wordt in het algemeen als *rijvariabele* in zo een tabel gebruikt, terwijl de andere variabele vaak de *kolomvariabele* is. Elke combinatie van de twee variabelen vormt een *cel.* In ons voorbeeld worden twee beroepen en twee geslachten onderzocht. Hier horen dus vier cellen bij. Om de relatie tussen twee categorische variabelen te beschrijven, berekenen we verschillende percentages, bijvoorbeeld het percentage mannen dat dokter is of het percentage vrouwen dat psycholoog is. Bij elkaar opgeteld komen de proporties precies op 1 uit. De verzameling van deze proporties maakt deel uit van de *verzamelde distributie (joint distribution)* van de twee categorische variabelen.

Marginale en conditionele distributies
Naast een verzamelde distributie is het ook mogelijk om marginale distributies weer te geven. Je kunt dan van beide variabelen afzonderlijk een proportiedistributie maken. Je kunt dus een distributie maken van sekse (met de proportie mannen en vrouwen die onderzocht zijn) en een distributie maken voor beroep (met de bijbehorende proportie voor doktor en psycholoog). Een conditionele distributie geeft echter meer informatie dan afzonderlijke marginale distributies. Je kijkt dan bijvoorbeeld naar de proportie doktoren, gegeven dat iemand een man is.

Staafdiagrammen helpen ons om de relatie tussen twee categorische variabelen te ontdekken. Geen enkele grafische weergave laat de vorm van de relatie tussen categorische variabelen zien en geen enkele numerieke waarde (zoals de correlatie) is een uiting van de sterkte van de relatie tussen dit soortvariabelen. Er kunnen ook driewegtabellen ontworpen worden.

Paradox van Simpson
Zoals bij kwantitatieve variabelen, kunnen op de loer liggende variabelen ook invloed hebben op de relatie tussen categorische variabelen.

- Een verband of vergelijking die opgaat voor alle onderzochte groepen kan van richting veranderen wanneer de data wordt gecombineerd tot een enkele groep. Deze verandering van richting wordt het *paradox van Simpson* genoemd. Dit paradox laat in extreme vorm zien dat relaties misleidend kunnen zijn wanneer er op de loer liggende variabelen aanwezig zijn.

2.7 Causaliteit

Samenhang
Correlatie zegt alleen iets over de mate waarin twee variabelen samenhangen. Met een (sterke) correlatie kan daarom niets gezegd wordt over causaliteit. Als we zien dat veel faalangst samengaat met lagere schoolcijfers, kunnen we dus (nog) niet concluderen dat faalangst de *oorzaak* van de lage cijfers is.

- Als variabele X variabele Y veroorzaakt, is er sprake van causaliteit (X→Y). Causaliteit kan door middel van experimenten ontdekt worden. In dat geval worden waarden van variabele X gevarieerd om het effect op Y te onderzoeken. Andere factoren worden constant gehouden. Dit om de invloed van op de loer liggende variabelen zo klein mogelijk te houden.
- Het is ook mogelijk dat X en Y lijken samen te hangen omdat ze allebei worden beïnvloed door een andere variabele, namelijk variabele Z. Dit wordt ook wel *algemene respons (common response)* genoemd. Variabele Z is in dit geval dus een op de loer liggende variabele. De waargenomen correlatie tussen X en Y is dus misleidend.
- Tot slot kan er sprake zijn van *confounding.* Twee variabelen zijn 'confounded' wanneer hun effecten op een responsvariabele niet van elkaar kunnen worden onderscheiden. Deze 'confounded' variabelen kunnen zowel verklarende als op de loer liggende variabelen zijn. De waargenomen correlatie tussen X en Y is dus misleidend als er sprake is van confounding.

Oorzakelijke verbanden
Soms is het niet mogelijk om causaliteit te ontdekken door middel van experimenten. Je kunt mensen bijvoorbeeld niet laten roken om te kijken of ze kanker krijgen. Zo een onderzoek zou onethisch zijn. Uit onderzoek blijkt dat rookgedrag vaak wel samengaat met kanker, maar hieruit mag nog niet geconcludeerd worden dat roken kanker veroorzaakt. Dit omdat er geen sprake is geweest van een experiment. Hoe moet causaliteit ontdekt worden als er geen experimenten gedaan kunnen worden? De onderstaande factoren zijn van belang.

- Als blijkt dat er een *sterke samenhang tussen variabele X (roken) en variabele Y (kanker)* bestaat, dan moet hier aandacht aan besteed worden.
- Ook moet de *samenhang consistent* zijn. Uit onderzoeken in verschillende landen blijkt bijvoorbeeld dat roken en kanker vaak samengaan.

- Ook moet blijken dat *hoge doseringen samengaan met sterkere reacties*. Mensen die erg veel roken krijgen bijvoorbeeld vaker kanker.
- *De vermoede oorzaak moet aan het gevolg voorafgaan*. Longkanker blijkt zich bijvoorbeeld pas na vele jaren van roken te ontwikkelen. *De vermoede oorzaak moet plausibel zijn.* Uit onderzoeken met dieren blijkt bijvoorbeeld dat sigarettenrook kanker veroorzaakt.