2. Alles wat je absoluut niet wilde weten over statistiek

Statistische modellen maken

Voor het testen van hypothesen over bepaalde fenomenen is het maken van een statistisch model nodig. Wetenschappers proberen om een model te maken dat het werkelijke fenomeen zo goed mogelijk benadert. Hoe goed het model overeenkomt met de verzamelde data, wordt de *fit* van het model genoemd. In de sociale wetenschap worden *lineaire modellen* het meest toegepast. Een lineair model is een model dat gebaseerd is op een rechte lijn. In deze modellen wordt een rechte lijn gezien als best passend bij de data. De meeste statistische toetsen, zoals ANOVA en regressie, zijn gebaseerd op lineaire modellen.

Populaties en steekproeven

Wetenschappers willen meestal resultaten die gegeneraliseerd kunnen worden naar een hele populatie. Omdat populaties vaak groot zijn, is het meestal onmogelijk om de hele populatie te onderzoeken. Om toch een indruk van de populatie te krijgen, worden er *steekproeven* getrokken. Hier worden de gegevens van een deel van de populatie verzameld voor een inzicht in de hele populatie.

Statistische modellen

Alle statistische modellen komen in feite neer op deze formule:

$\text{Uitkomst}_i = (\text{model}) + \text{error}_i$

De i staat hierbij voor de i^{ste} score, dus van een bepaald persoon. In plaats van de i kun je dus ook een participantnummer of een naam lezen. Deze formule betekent dat we onze data kunnen voorspellen vanuit het model plus een bepaalde foutmarge.

Statistische modellen bestaan uit variabelen en parameters. Parameters worden niet gemeten, maar worden geschat aan de hand van de data en zijn meestal constanten. Voorbeelden van parameters zijn het gemiddelde en correlatiecoëfficiënten. Verschillende parameters hebben verschillende statistische symbolen. De b staat voor een regressiecoëfficiënt. Vaak willen we een uitkomst voorspellen aan de hand van een predictor (aangegeven met een X). Het model wordt dan:

$\text{Uitkomst}_i = (bX_i) + \text{error}_i$

Hiermee voorspel je de uitkomst van i vanuit de score van i op de onafhankelijke variabele (X_i). De b zegt iets over de relatie tussen de predictor en de uitkomst, zoals de sterkte en richting van het verband. Je kunt ook meerdere predictors gebruiken om de uitkomst te voorspellen. Het model wordt dan:

$\text{Uitkomst}_i = (b1X1_i + b2X2_i) + \text{error}_i$

In dit model voorspellen de scores van i op predictor 1 (X1) en predictor 2 (X2) samen de uitkomst. Hoe het model eruit ziet, hangt af van de waarde van de parameter b. Deze waarde is een schatting van de parameters in de populatie, omdat ze berekend zijn op basis van de steekproef. Het heet dus een geschatte parameter omdat we een steekproef hebben gemeten en op basis daarvan aannames doen over de populatie.

Het gemiddelde is het eerste eenvoudige statistische model en heeft een hypothetische waarde. Dit betekent dat het een waarde kan aannemen dat niet in de data voor hoeft te komen. Het gemiddelde aantal vrienden kan bijvoorbeeld 2.6 zijn, terwijl niemand daadwerkelijk 2.6 vrienden heeft. Bij dit model probeer je de uitkomst niet te voorspellen, maar is het gemiddelde een samenvatting van de uitkomst. Het model is in dit geval:

Uitkomst = gemiddelde + error

Je wil weten hoe goed de fit is van je model, hoe goed het model werkelijk bij de data past. Daarvoor kan bij elke score worden berekend hoeveel die afwijkt van het model. Het verschil tussen de score en het model is de *deviantie* of *error* (meetfout in het Nederlands) in het model. Dus, deviantie is de score van een persoon (de uitkomst) – het model.

Error = geobserveerde uitkomst – model

De totale error is dan gelijk aan de som van alle devianties. Net als in het vorige hoofdstuk, is ook hier de som van de devianties altijd gelijk aan nul, omdat bij sommige personen de score boven het gemiddelde ligt, en bij anderen onder het gemiddelde. Samen zijn ze uiteraard precies het gemiddelde. Daarom worden, net als in het eerste hoofdstuk, de deviaties gekwadrateerd. Je krijgt dan de *som van de gekwadrateerde meetfouten (SS)*.

Sum of squared errors (SS) = $\sum$(uitkomst – model)²

Dit is dezelfde formule als die ook in het vorige hoofdstuk staat, maar in plaats van de staat hier ‘model’, zodat het niet alleen voor het gemiddelde kan staan, maar ook voor meer ingewikkelde modellen.

Zoals eerder al is gezegd, is de waarde van de SS afhankelijk van de steekproefgrootte. Dit is onhandig, omdat we niet zozeer geïnteresseerd zijn in de steekproef, maar in de populatie. De oplossing hiervoor is het berekenen van de mean squared error (gemiddelde meetfout). Dit doe je door de SS te delen door het aantal *vrijheidsgraden* (df).

Vrijheidsgraden staat voor het aantal observaties dat vrij is om te variëren. Als je het gemiddelde van 4 mensen weet, kun je voor de eerste drie een getal verzinnen, maar heb je geen keus voor het vierde getal als je het gemiddelde constant wil houden. Dat ligt vast. Het aantal *vrijheidsgraden* is dus het aantal observaties (de steekproefgrootte) min 1. De variantie is een speciale naam voor de mean squared error die wordt gebruikt als het gemiddelde wordt gebruikt als model. De formule voor de variantie wordt dan:

Variantie (s^2) = $\frac{SS}{df}$ = $\frac{\sum(\text{uitkomst} - \bar{x})^2}{N-1}$

Bij een kleine standaardafwijking (s, de wortel van de variantie) liggen de data allemaal dicht bij het gemiddelde. Het gemiddelde is in dat geval een goed model voor voorspellingen, het heeft een goede fit. Bij een grote standaardafwijking heeft het model van het gemiddelde een slechte fit.

De parameters in het model worden geschat op basis van de verzamelde data. Als je niet zou weten hoe je het gemiddelde moet berekenen, zou je die kunnen schatten. Vervolgens zou je op basis van deze schattingen de SS van je data kunnen berekenen om te bekijken hoe goed de fit van je model is. Schattingen ver van het werkelijke gemiddelde, leveren een grote SS op, je data wijkt dan veel af van je model. De waarde van de parameter is de waarde waarbij de minste meetfouten optreden, dus met de kleinste SS. De parameter kan nog steeds een grote SS opleveren, dat betekent dat het een model is met een slechte fit, maar de gekozen parameter is altijd degene die de minste error oplevert. Deze methode wordt de *method of least squares* genoemd.

Meer dan alleen gegevens

De standaardafwijking geeft aan hoe goed het gemiddelde past bij de steekproefgegevens. Maar als we op basis hiervan willen schatten wat de parameters in de populatie zijn, moeten we weten hoe representatief de steekproefgegevens zijn voor de populatie. Dit is belangrijk omdat de uitkomsten van verschillende steekproeven uit dezelfde populatie lichtelijk kunnen verschillen.

Van elke steekproef die je neemt, kun je het *steekproefgemiddelde* berekenen. Het gemiddelde van alle steekproeven is een schatting van het populatie gemiddelde μ. Tussen de steekproeven is een *steekproefvariatie*. Er zijn kleine verschillen tussen de steekproeven omdat men bijvoorbeeld toevallig een aantal slimme mensen of domme mensen bij elkaar heeft gezet. Een *steekproefverdeling* is een frequentieverdeling van alle steekproefgemiddelden. Je kunt deze verdeling gebruiken om te kijken hoe representatief een steekproef is voor de populatie.

Net zoals de standaardafwijking iets zegt over hoe representatief het gemiddelde is voor de data, heb je ook een standaarddeviatie die iets zegt over hoe representatief de steekproefgemiddelden zijn voor het populatiegemiddelde. Dit wordt de *standaard meetfout van het gemiddelde* of *standard error (SE)* genoemd.

Deze standard error is te berekenen met grofweg dezelfde formule als de standaardafwijking. Bij de standard error neem je echter het verschil tussen het steekproefgemiddelde en het gemiddelde van alle steekproeven als basis voor de SS, in plaats van het verschil tussen een score en het gemiddelde. In werkelijkheid zal je echter nooit een steekproefverdeling kunnen maken, want je gaat niet tientallen steekproeven trekken. Bij grote steekproeven (n>30) kan de *centrale limiet stelling* worden gebruikt, die zegt dat een steekproefverdeling normaal verdeeld is, waarbij het gemiddelde gelijk is aan het populatiegemiddelde en de standaardafwijking berekend wordt met deze formule:

$$\sigma_{\bar{x}} = \frac{s}{\sqrt{N}}$$

Betrouwbaarheidsintervallen
Een steekproef geeft dus een schatting van de parameters van de populatie en met de standard error krijg je een idee hoe deze schattingen verschillen bij verschillende steekproeven. Met deze informatie kun je grenzen berekenen waartussen je denkt dat het werkelijke populatiegemiddelde valt. Dit heet een *betrouwbaarheidsinterval.* Een betrouwbaarheidsinterval van 95% voor het gemiddelde betekent dat het populatiegemiddelde bij 95% van de steekproeven binnen deze berekende grenzen valt.

Om dit betrouwbaarheidsinterval te berekenen, moet je weten wat de grenzen zijn waarbinnen 95% van de steekproefgemiddelden zal vallen. Dit doe je met de standaard normaalverdeling, de z-verdeling, die in het vorige hoofdstuk is besproken. 1.96 is de z-score bij een 95% betrouwbaarheidsinterval. Dit betekent dat 95% van de z-scores tussen de -1.96 en 1.96 liggen.

Z-scores zijn gestandaardiseerde scores, waarbij de ruwe data is omgezet in een standaard normaalverdeling, waarbij het gemiddelde 0 is en de standaardafwijking 1. Andersom kun je z-scores terugvertalen naar de ruwe data, door in de formule de z-score, de standaardafwijking en het gemiddelde in te vullen (zie hoofdstuk 1 voor de formule voor het berekenen van z-scores). Omdat het bij betrouwbaarheidsintervallen gaat om de variatie in de steekproefgemiddelden, gebruik je hierbij de standaard error in plaats van de standaardafwijking.

Onderste grens van het 95% betrouwbaarheidsinterval =– (1.96 x SE)

Bovenste grens van het 95% betrouwbaarheidsinterval = + (1.96 x SE)

Andere betrouwbaarheidsintervallen hebben niet 1.96 in de formule maar andere getallen, zoals -2.58 en 2.58 voor een 99% betrouwbaarheidsinterval. Voor het juiste getal, zie de tabel met z-scores in de appendix. Onthoud dat de normale verdeling symmetrisch is, dus wanneer 99% in het midden zit, zit er in beide staarten de helft van 1%, dus 0.005. Bij dit getal moet je in de tabel kijken om de z-score voor een 99% betrouwbaarheidsinterval te vinden.

Bij kleine steekproeven kan de centrale limietstelling niet gebruikt worden. Hiervoor gebruik je de t-verdeling en wordt de t-score gebruikt in de formules voor het betrouwbaarheidsinterval in plaats van de z-score. De t-verdeling is een verdeling die van vorm verandert als de steekproefgrootte toeneemt. Daarom hangt de t-score af van de vrijheidsgraden.

Een betrouwbaarheidsinterval wordt grafisch weergegeven door middel van een error bar. Als twee betrouwbaarheidsintervallen totaal niet overlappen met elkaar, is het waarschijnlijk dat je te maken hebt met twee verschillende populaties.

Gebruik van statistische modellen voor het testen van onderzoeksvragen

Bij de statistische toetsen worden hypothesen getoetst. Volgens Fisher weet je alleen of er een werkelijk effect is, als er slechts een kleine kans is dat het resultaat per toeval werd bereikt. De kans die in de praktijk vaak wordt aangehouden is 5%, wat hetzelfde is als een kans van 0.05.

Neyman en Pearson vonden dat wetenschappelijke stellingen moesten worden verwoord als testbare hypothesen. De hypothese die zegt dat er een bepaald effect is, heet de alternatieve hypothese (of H_1). De nulhypothese (H_0) is het tegenovergestelde, die zegt dat er geen effect is.

Onderzoek richt zich vaak op het verzamelen van bewijs dat de nulhypothese verwerpt. Als je de nulhypothese kunt verwerpen, is dat een aanwijzing dat de alternatieve hypothese juist is, maar het is slechts steun voor de hypothese, geen bewijs. Je praat niet over of de H_0 of de H_1 waar is, maar over de kans dat je bepaalde data verkrijgt, als de nulhypothese waar zou zijn.

Hypotheses kunnen een richting aangeven of dit niet doen. Een hypothese die een bepaalde richting van het effect aangeeft, is een eenzijdige hypothese. Een voorbeeld hiervan is de hypothese: Als je je inbeeldt dat je chocola eet, eet je er minder van. Een voorbeeld van een hypothese zonder richting, een tweezijdige hypothese, is: Het inbeelden van het eten van chocola heeft invloed op de hoeveelheid chocola die je eet.

De logica achter nulhypothese significantie toetsen (NHST) volgt hieronder in een aantal stappen:

1. Eerst neem je aan dat de H_0 waar is, dus dat er geen effect is.
2. Vervolgens pas je een statistisch model toe op de data dat de alternatieve hypothese volgt, en kijk je hoe goed de fit van dat model is (hoeveel variantie in de data het model verklaart).
3. Om te bepalen hoe goed de fit is, bereken je de kans (de p-waarde) dat je dat ‘model’ krijgt als de nulhypothese waar is.
4. Als die kans klein is (meestal 0.05 of kleiner), concludeer je dat het model goed past op de data. Dit betekent steun voor de alternatieve hypothese.

Zoals in hoofdstuk 1 is uitgelegd, bestaat er systematische en niet-systematische variatie. Systematische variatie wordt verklaard door het model dat je op de data toepast, dus door de hypothese die je toetst. Niet-systematische variatie kan niet door dit model (deze hypothese) worden verklaard.

Om te kijken of het model een goed model is voor de data, vergelijk je de systematische variatie met de niet-systematische variatie. Deze manier om te kijken of de hypothese een goede verklaring is voor de gegevens, is een *test statistiek.*

$$\text{Test statistiek} = \frac{\text{Variantie verklaard door het model}}{\text{Variantie niet verkl aard door het model}} = \frac{\text{Effect}}{\text{Error}}$$

Als het model goed is, verwacht je dat het meer variantie wel kan verklaren dan dat het niet kan verklaren. In dat geval verwachten we dus een test statistiek groter dan 1. Maar dat een test statistiek groter is dan 1, betekent niet per se dat het ook significant is. Met deze test statistiek kun je de kans bepalen dat die waarde wordt verkregen als de nulhypothese waar is. Hoe meer variantie het model verklaart in vergelijking met de variantie die het niet kan verklaren, hoe groter de waarde van de test statistiek, en hoe kleiner de kans is dat die waarde per toeval wordt verkregen. Als deze kans kleiner is dan 0.05, verwerpen we de nulhypothese. De test statistiek is dan *significant*.

Eenzijdig en tweezijdig toetsen

Bij een *eenzijdige toets* wordt een statistisch model getoetst met een hypothese die een richting voor het effect aangeeft. Bij een *tweezijdige toets* heeft de hypothese geen richting. Bij het tweezijdige model wordt het onze beslissende waarde van 0.05 ook in tweeën gesplitst, omdat de vijf procent overschrijding in beide staarten van de normale verdeling zit. In de linker- en rechterstaart zit dus allebei 0.025. Een tweezijdige toets betekent immers dat het effect zowel negatief als positief kan zijn.

Omdat bij de eenzijdige toets de beslissende grens van 0.05 slechts in één staart zit (in de richting die de hypothese voorspelt), ligt de kritieke grens op een lagere waarde en wordt dus sneller een significant resultaat gevonden. Echter, het effect moet wel in de voorspelde richting zijn. Als je bij een eenzijdige toets een significant resultaat vindt in de verkeerde richting, moet de nulhypothese alsnog aangenomen worden.

Eenzijdige toetsen zijn zelden een goed idee. Mocht je per ongeluk een resultaat in de verkeerde richting vinden, dan moet je dit resultaat negeren. Dat is moeilijk, want juist zo'n onverwacht resultaat willen wetenschappers graag verklaren. Eenzijdige toetsen zijn alleen geschikt als een resultaat in de verkeerde richting zorgt voor dezelfde actie als een niet-significant resultaat. Tenslotte nodigen eenzijdige toetsen uit tot onethische wetenschap, waarbij resultaten die met een tweezijdige toets niet significant zijn, met een eenzijdige toets alsnog als significant worden gepresenteerd.

Er zijn twee soorten fouten die gemaakt kunnen worden. De eerste is de *type I fout*. Deze fout wordt gemaakt wanneer we denken dat er een effect in de populatie is, terwijl dat in de werkelijkheid niet zo is. De kans op deze fout is het significantieniveau, het *α-niveau* (meestal .05). De andere fout is de *type II fout*. Bij deze fout denkt men dat er geen effect is in de populatie terwijl dat in werkelijkheid wel zo is. Een acceptabel niveau voor de kans op type II fout is .2 (= *β-niveau*).

Type I en type II fouten

De twee typen fouten zijn aan elkaar gerelateerd. Als je de kans verkleint dat je een effect voor waar aanneemt (dus een kleinere α), vergroot je de kans dat je een werkelijk effect over het hoofd ziet. Minder kans op een type I fout, zorgt dus voor meer kans op een type II fout, en andersom.

Het meest gebruikte significantieniveau is 0.05, wat betekent dat de kans op een type I fout 5% is. Voor elke test is de kans op geen type I fout dus 0.95. Meestal voeren onderzoekers echter meerdere testen uit.

Bij drie testen is de kans op geen type I fout $(0.95)^3$=0.857. Dit betekent dat de kans dat er minimaal één type I fout optreedt in deze testen 14.3% is. Dit fenomeen noem je de *familywise* of *experimentwise error rate*. Om dit probleem te voorkomen, kun je het significantieniveau aanpassen zodat de kans op een type I fout 0.05 blijft. De makkelijkste manier hiervoor is de *Bonferroni correctie*. Hierbij deel je de α door het aantal vergelijkingen. Voor 10 toetsen, gebruik je dan een significantieniveau van 0.005 (0.05/10).

Een probleem met het beheersen van de familywise error rate is dat het ten koste gaat van de mate waarin een test in staat is een effect te vinden, de zogenaamde statistische *power*. De power van een test is de kans dat een in de populatie bestaand effect gevonden wordt. Dit is dus het tegenovergestelde van een type II fout (β), die zegt dat een bestaand effect niet gevonden wordt. Power is dus 1-β.

De power van een test wordt beïnvloed door de volgende factoren:

- De sterkte van het effect. Grote effecten worden makkelijker gevonden.
- Het significantieniveau. Als we heel streng zijn in de beslissing wanneer iets significant is, dus een kleine α hebben, worden effecten minder snel gevonden. Om deze reden neemt de power af als je een Bonferroni correctie toepast.
- De steekproefgrootte. Grote steekproeven zijn representatiever voor de populatie, en hebben daardoor minder error. Test statistieken zijn de ratio tussen het effect en de error, dus bij een kleinere error wordt het effect eerder gevonden.

Er is een verband tussen statistische significantie en betrouwbaarheidsintervallen. Als twee 95% betrouwbaarheidsintervallen elkaar net aanraken bij de uiteinden, is dat een weergave van een p-waarde van ongeveer 0.01 als de nulhypothese wordt getoetst dat er geen verschil is tussen de gemiddelden. Als er ruimte zit tussen de betrouwbaarheidsintervallen, betekent dat een p-waarde kleiner dan 0.01. Bij een p-waarde van 0.05 is er een matige overlap.

Net zoals de power afhangt van de steekproefgrootte, zo is er ook een belangrijke link tussen de steekproefgrootte en significantie. Hoe groter de steekproef, hoe sneller een effect significant is. Een kleinere waarde op de test statistiek, kan dan toch al significant zijn. Dit komt doordat de standard error kleiner is bij een grote steekproef. De standard error is immers de standaardafwijking gedeeld door $\sqrt{N}$. Een kleinere error betekent dat een effect sneller gevonden wordt. Een test statistiek deelt immers het effect door de error, en minder error betekent dat een kleiner effect sneller gevonden wordt.

In erg grote steekproeven kan het gebeuren dat een er een significant effect wordt gevonden, die in werkelijkheid weinig voorstelt. Andersom kunnen belangrijke effecten gemist worden door een te kleine steekproef. Een non-significant resultaat betekent dus niet per se dat de nulhypothese waar is.

Een significant resultaat betekent echter ook niet per se dat de nulhypothese onwaar is. Het blijft namelijk een kwestie van kans rekenen, waarbij verkeerde conclusies kunnen worden getrokken.

Nog een probleem met nulhypothese significantie toetsen is dat het vaak een alles of niets kwestie is. Een p-waarde net iets boven 0.05 is niet significant, net onder 0.05 is wel significant. Dat is vreemd, omdat die 0.05 niets meer is dan een handige vuistregel.

Effectgrootte

Een test kan wel significant zijn, maar het effect heeft ook een belangrijke bijdrage. De grootte van het effect wordt ook getest (*effectgrootte*). Het is een gestandaardiseerde meting van de grootte van het geobserveerde effect. Cohen's *d,* Pearson's correlatiecoëfficiënt en de odds ratio meten de effectgrootte.

Cohen's d bereken je als volgt: $$\text{Cohen's d} = \frac{\overline{X}_1 + \overline{X}_2}{s}$$

d=0.2 komt overeen met een klein effect, d=0.5 is een gemiddeld effect en 0.8 is een groot effect. De effectgrootte wordt niet beïnvloed door de steekproefgrootte.

Ook Pearson's r kan gebruikt worden als maat voor de effectgrootte. Het geeft immers de sterkte van het verband aan. R^2 staat voor het percentage verklaarde variantie. r=.10 is een klein effect, het verklaart dan 1% van de variantie. r=.30 is een medium effect, r^2=9%. r=.50 is een groot effect, r^2 is dan 25%.

Meestal worden er meerdere onderzoeken naar hetzelfde onderwerp gedaan. Deze verschillende onderzoeken geven vaak verschillende resultaten. Door de effectgroottes te combineren uit verschillende studies krijg je betere schattingen van de populatie. Dit wordt *meta-analyse* genoemd. Hiermee wordt dan een gemiddelde effectgrootte berekend.