Hoofdstuk 9: Analyse van tweewegtabellen

Inleiding
In dit deel zal uitgelegd worden hoe twee of meer populaties vergeleken moeten worden wanneer de responsvariabele twee of meer categorieën heeft. Ook zal uitgelegd worden hoe onderzocht kan worden of twee categorische variabelen onafhankelijk zijn.

9.1 Gevolgtrekkingen uit tweewegtabellen

Samenhang
Het is mogelijk om de samenhang tussen twee categorische variabelen uit te zoeken. Denk in dit verband maar eens aan geslacht (man/vrouw) en het hebben van een rijbewijs (ja/nee). De variabelen kunnen in een tweewegtabel gezet worden met twee rijen en twee kolommen. De kolommen staan voor onafhankelijke steekproeven uit verschillende populaties. Er zijn c aantal distributies voor de rijvariabele, één voor elke populatie.

- De nulhypothese stelt dat er *geen samenhang* tussen twee categorische variabelen bestaat. Deze hypothese zegt dus eigenlijk dat de c-distributies van elke rijvariabele hetzelfde zijn.
- De alternatieve hypothese stelt dat er sprake is van samenhang tussen de twee variabelen. In de alternatieve hypothese wordt echter geen richting aangegeven. Deze hypothese zegt dus dat de distributies niet allemaal gelijk zijn.

De chi-kwadraat toets
Om de nulhypothese over de rxc (rij x kolom) informatie te toetsen, vergelijken we de geobserveerde celtellingen met de *verwachte celtellingen*. Omdat het om een tweewegtabel gaat, zijn er in totaal vier cellen.

- De verwachte celtelling: (rijtotaal x kolomtotaal) / n.

Om de nulhypothese te toetsen moet een *chi-kwadraattoets* berekend worden.

- Eerst moet het verschil tussen elke geobserveerde telling en de bijbehorende verwachte telling berekend worden. Alle verschillen moeten gekwadrateerd worden, zodat alle uitkomsten positief zijn.
- Vervolgens moet elk gekwadrateerde verschil door de bijbehorende verwachte telling gedeeld worden. Dit is een methode om de verschillen te standaardiseren.
- Tot slot moeten alle resultaten opgeteld worden. Het resultaat is de *chi-kwadraat toets (X^2)*. De bijbehorende formule is: $X^2 = \Sigma$ (geobserveerde telling - verwachte telling)^2 / verwachte telling.

Chi-kwadraatdistributie
Als de verwachte tellingen en de geobserveerde tellingen erg verschillend zijn, zal er een grote chi-kwadraat toets gevonden worden. Grote waarden van X^2 geven bewijs tegen de nulhypothese. Om een p-waarde te vinden gaan we aan de gang met de *chi-kwadraatdistributie*. Zoals de t-distributies zijn vrijheidsgraden ook belangrijk voor chi-kwadraatdistributies. Er kunnen alleen maar positieve chi-kwadraat toetsen verkregen worden op basis van onderzoeksdata. De chi-kwadraatdistributie heeft een afwijking naar rechts.

- Als de nulhypothese waar is, dan heeft X^2 een distributie met (r-1)(c-1) vrijheidsgraden. De p-waarde kan gevonden worden door de berekende chi-kwadraattoets op te zoeken in de chi-tabel en te kijken tussen welke p-waarden deze ligt.

Berekeningen

De chi-kwadraattoets kan dus in het kort als volgt uitgevoerd worden:

1. Bekijk eerst de rij- en kolompercentages.
2. Bereken vervolgens de verwachte tellingen en gebruik deze om de chi-kwadraattoets te berekenen.
3. Gebruik de kritische waarden uit de chi-tabel om een p-waarde vast te stellen.
4. Trek tot slot een conclusie over de samenhang tussen de rij- en kolomvariabelen.

De z-toets en de chi-kwadraattoets

Een z-toets uitvoeren op basis van dezelfde onderzoeksdata geeft dezelfde resultaten als een chi-kwadraat toets. Het voordeel van een z-toets is echter dat we zowel eenzijdig als tweezijdig kunnen toetsen, terwijl we met de chi-kwadraat toets alleen tweezijdig kunnen toetsen. Het voordeel van de chi-kwadraat toets is dat er meer dan twee populaties met elkaar vergeleken kunnen worden.

Modellen voor tweewegtabellen

De chi-kwadraattoets kan in twee situaties uitgevoerd worden: (1) als meerdere populaties vergeleken moeten worden en (2) als onafhankelijkheid getoetst moet worden.

1. In de eerste situatie kun je bijvoorbeeld de wijnverkoop in drie omgevingen testen. Je kunt dan een tabel maken met twee categorische variabelen (wijn en muziek) met drie mogelijkheden per variabele: Frans, Italiaans en Anders. In dit geval ben je op zoek naar de samenhang tussen soorten wijn en het soort muziek dat in een restaurant wordt gedraaid. De nulhypothese is dan dat er geen samenhang is tussen soort wijn en soort muziek. De proporties zijn volgens deze hypothese dus hetzelfde in alle *populaties.*
2. In de tweede situatie worden de scores op twee variabelen van één populatie onderzocht. Een voorbeeld is dat studenten van een universiteit (man/vrouw) wordt gevraagd naar hun mening over abortus (voor/tegen). In dat geval wordt er dus een random steekproef getrokken en worden per individu de waarden voor de twee variabelen genoteerd. De nulhypothese zegt in dit geval dat de rij- en kolomvariabelen onafhankelijk zijn. Sekse en mening over abortus zouden dus niet samenhangen.

Bij het onafhankelijkheidsmodel is er dus sprake van een enkele steekproef. De kolomtotalen en rijtotalen zijn random variabelen. De totale steekproefgrootte *n* wordt door de onderzoeker gekozen, de colom- en rijsommen zijn pas bekend nadat de data zijn verzameld.
Voor het vergelijken-van-populatiesmodel daarentegen, is er een steekrpef voor elk van twee of meer populaties. De kolomsommen zijn de steekproefgroottes die geselecteerd zijn tijdens het ontwerpen van het onderzoek.

De nulhypothese in beide modellen stelt dat er geen relatie is tussen de kolomvariabele en de rijvariabele. Gelukkig is de test voor de hypothese van ‘geen relatie’ hetzelfde voor beide modellen: de Chi-kwadraattoets. Er zijn ook statistische modellen die gerelateerd zijn aan de Chi-kwadraattoets, die het mogelijk maken om drieweg- of meerwegtabellen te analyseren.

9.2 Goodness of fit

Data voor n aantal observaties van een categorische variabele met k aantal mogelijke uitkomsten worden genoteerd als $n_1, n_2, n_2 \ldots n_k$ observaties in k aantal cellen. De bijbehorende nulhypothese gaat over de kansen $p_1, p_2, p_3 \ldots p_k$ voor alle mogelijke uitkomsten. Voor elke cel moet het totale aantal observaties (n) vermenigvuldigd worden met de kans die gebruikt wordt om de verwachte tellingen te berekenen:

- Verwachte telling = np_i.
- De chi-kwadraat toets meet hoeveel de geobserveerde celtellingen verschillen van de verwachte celtellingen. De formule voor deze toets is:
- X^2 =Σ(geobserveerde telling-verwachte telling)^2 /verwachte telling.
- De bijbehorende vrijheidsgraden zijn k-1 en de p-waarden kunnen teruggevonden worden in de chi-tabel.