8. Analyseren van samenhang tussen categorische variabelen

Er bestaat een samenhang tussen twee variabelen als de verdeling van de respons (afhankelijke) variabele verandert op het moment dat de waarde van de verklarende (onafhankelijke) variabele verandert. We gaan nu kijken hoe je zo'n verband kunt vaststellen bij twee categorische, dus ordinale of nominale, variabelen.

Terminologie voor categorische data-analyse en statistische onafhankelijkheid.

Marginale verdeling

Categorische data worden vaak weergegeven in een kruistabel. Bijvoorbeeld deze:

Partijvoorkeur				
Geslacht	Democraten	Onafhankelijk	Republikeins	Totaal
Vrouw	573	516	422	1511
Man	386	475	399	1260
Totaal	959	991	821	2771

Deze tabel heeft twee rijen (man-vrouw) en drie kolommen (D-O-R). De rij- en kolomtotalen noemen we de marginale verdeling. Wil je van deze tabel bijvoorbeeld de marginale verdeling van partijvoorkeur weergeven, dan schrijf je (959, 991, 821).

Het maken van zo'n kruistabel is de eerste stap bij het doen van data-analyse met categorische variabelen. De tweede stap is om de absolute getallen om te zetten in percentages.

Conditionele verdeling

We willen weten of er een samenhang bestaat tussen geslacht en stemgedrag. We willen dus weten of het stemgedrag anders is voor mannen en voor vrouwen. We moeten dan kijken naar het stemgedrag van de mannen, en dat van de vrouwen. Daarom gebruiken we bij het berekenen van de percentages niet de totale groep.

De tabel met de percentages ziet er zo uit:

Partijvoorkeur					
Geslacht	Democraten	Onafhankelijk	Republikeins	Totaal	n
Vrouw	38%	34%	28%	100%	1511
Man	31%	38%	32%	100%	1260

*Toelichting Tabel: Vrouw-Democraten : 573/1511*100 = 38%. Man-Democraten : 959/1260*100 = 31%. Etc.*

Nu heb je de relatieve data verdeling van partijvoorkeur, afhankelijk van geslacht. Deze sets van percentages noemen we de conditionele distributie van partijvoorkeur. De conditionele verdeling van de vrouwen is (38, 34 28) voor (D,O,R). De conditionele distributie van de mannen is (31, 38, 32) voor (D,O,R).

Je kunt natuurlijk ook conditionele distributies maken voor geslacht per partijvoorkeur. Dan kreeg je bij de vrouwen (573/959*100) 60% en voor mannen (386/959*100) 40%. Meestal maak je echter een conditionele distributie voor de afhankelijke variabele.

Dus wanneer er wordt gevraagd om een conditionele distributie te maken van de response variabelen, binnen de categorieën van de verklarende variabele, doe je dat zoals bovenstaand.

Joint distribution (simultane verdeling)

Je kunt de percentages ook op een andere manier weergeven. Je berekent dan de percentages ten opzichte van de totale steekproef. Dan zou het er zo uit zien:

Partijvoorkeur			
Geslacht	Democraten	Onafhankelijk	Republikeins
Vrouw	21%	19%	15%
Man	14%	17%	14%

*Toelichting: Vrouw-Democraten : 573/2771*100 = 21%. Man-Democraten : 386/2771*100 = 14%. Etc.*

Deze verdelingen noemen we de simultane verdelingen. Maar wanneer je kijkt naar een respons (afhankelijke) en verklarende (onafhankelijke) variabele is het zinniger om te kijken naar conditionele verdelingen dan naar simultane verdelingen.

Statistisch (on)afhankelijk

Twee categorische variabelen zijn statistisch onafhankelijk wanneer de kans op het voorkomen van de ene ‘gebeurtenis’ los staat van de kans dat de andere ‘gebeurtenis’ voor komt. Anders gezegd: ze zijn statistisch onafhankelijk wanneer de kansverdeling van de mogelijke uitkomsten van de ene variabele niet wordt beïnvloedt door de uitkomsten van de andere variabele. Gebeurt dat wel, dan zijn ze statistisch afhankelijk.

Stel dat bij ons voorbeeld de twee variabelen geslacht en partijvoorkeur onafhankelijk van elkaar zouden zijn, dan zouden de percentages zo zijn verdeeld dat je bij Democraten een even groot percentage mannen als vrouwen hebt, en bij Onafhankelijke en Republikeinen ook. Maar dat is niet het geval.

Chi-kwadraat toets

Wanneer we zeggen dat twee variabelen onafhankelijk zijn, hebben we het over variabelen in de populatie. We verwachten wel dat de verdeling in de steekproef min of meer gelijk is aan die in de populatie, maar dat is die nooit helemaal. We willen dus kijken of het waarschijnlijk is dat we deze verschillen bij toeval tegenkomen in de steekproef. Dus dat de variabelen in de populatie wel onafhankelijk zijn, maar vanwege de steekproeffout de verdeling toch niet helemaal gelijk is. We toetsen dan

H0: de variabelen zijn statistisch onafhankelijk

Ha: de variabelen zijn statistisch afhankelijk

Chi-kwadraat

We berekenen dit met de Chi-kwadraat. De Chi-kwadraat toets vergelijkt de geobserveerde frequenties met de frequenties die voldoen aan H0 (de verwachte frequenties). Je kunt deze het beste ook in een tabel zetten. Hierbij blijven de rij- en kolomtotalen hetzelfde, maar voldoen de getallen aan onafhankelijkheid. De tabel ziet er zo uit:

Partijvoorkeur				
Geslacht	Democraten	Onafhankelijk	Republikeins	Totaal
Vrouw	573 (522,9)	516 (540,4)	422 (447,7)	1511
Man	386 (436,1)	475 (450,6)	399 (373,3)	1260
Totaal	959	991	821	2771

*Toelichting: De getallen in de tabel zonder haakjes zijn de geobserveerde frequenties. De getallen tussen de haakjes zijn de verwachte frequenties als H0 waar is. Vrouw-Democraten : 1511/2771*959 = 522,9. Man-Democraten : 1260/2771*959 = 436,1. Etc. Je kunt ook rij-totaal*kolomtotaal/steekproeftotaal. Dus voor Vrouw-Democraten: (1511*959)/2771 = 522,9.*

Geobserveerde frequenties noteren we met 'f_o'. De verwachte frequenties noteren we met 'f_e'.

Deze gebruik je voor het berekenen van de Chi-kwadraat, dat we weergeven met het symbool X^2.

De formule voor X^2 is: $X^2 = \sum \frac{(f_o - f_e)^2}{f_e}$.

Dus je berekent de verschillen tussen de geobserveerde en de verwachte frequentie. Deze kwadrateer je. Deze deel je door de verwachte frequentie. En dat doe je voor elke cel (= optellen).

Interpretatie van X^2

Wanneer H0 waar is, dan zullen de geobserveerde frequenties (f_o) dicht liggen bij de verwachte frequenties (f_e), dan zal X^2 klein zijn.

Wanneer H0 niet waar is, dan zullen f_o en f_e niet dicht bij elkaar liggen, waardoor de X^2 groot zal zijn.

Hoe groter X^2, hoe groter de kans wordt dat je H0 kunt gaan verwerpen. Het wordt dan onwaarschijnlijker dat de verschillen die je hebt gevonden toevallig zijn.

Kenmerken van X^2

De Chi-kwadraat verdeling geeft aan hoe groot X^2 moet zijn voordat je H0 kunt verwerpen. De verdeling heeft een aantal kenmerken:

- De verdeling is altijd positief. X^2 kan nooit negatief zijn.
- De verdeling is rechts scheef (lange staart rechts)
- De precieze vorm van de verdeling hangt af van het aantal vrijheidsgraden (df). Voor de Chi-kwadraat verdeling geldt : $\mu = df$; $\sigma = 2df$. De verdeling is rechts scheef, en wordt 'platter' naarmate df groter wordt.
- Hoe groter de kruistabellen zijn, hoe meer vrijheidsgraden je hebt, hoe groter je X^2 is.
- Hoe groter X^2, hoe groter de kans dat je H0 kunt gaan verwerpen.

Nu gaan we terug naar ons voorbeeld. We willen toetsen of er een relatie is tussen partijvoorkeur en geslacht. H0 : geslacht en partijvoorkeur zijn statistische onafhankelijk. Ha : geslacht en partijvoorkeur zijn statistisch afhankelijk. Wanneer je de X^2 uitrekent dan is deze 16,2. Dat kun je zelf narekenen. Of deze significant is moet je opzoeken in de tabel met alle p-waarden van de Chi-kwadraat verdeling. We hebben df = 2. Nu moet je in de tabel kijken, in de rij van df = 2, tussen welke twee getallen jouw Chi-waarde ligt. Kies de laagste. Als je dan omhoog gaat, heb je je p-waarde.

In ons voorbeeld zien we dat X^2 groter is dan 13,82. Dan omhoog zien we dat die hoort bij een p-waarde van 0.001. We verwerpen H0 en concluderen dat het erg onwaarschijnlijk is dat we deze verschillen per toeval zouden tegenkomen, en dat er dus wel degelijk een verband bestaat tussen geslacht en partijvoorkeur.

Je kunt alleen een Chi-kwadraat toets doen wanneer de verwachte frequenties (f_e) in elke cel groter zijn dan 5.

Vrijheidsgraden bij een Chi-kwadraattoets

Het aantal vrijheidsgraden bij een Chi-kwadraattoets (df) bereken je door : (r – 1) * (c – 1). Dit betekent dat je het aantal rijen neemt -1. Dat vermenigvuldig je door het aantal kolommen – 1. In onze tabel hadden we twee rijen en drie kolommen. Dus 1 * 2 = 2.

Residuen

De Chi-kwadraat toets zegt echter niks over de richting of de sterkte van de samenhang. Deze toets geeft alleen aan of de variabelen een verband hebben met elkaar. Er kan geen uitspraak worden gedaan over significante verschillen.

Daarom kijken we naar residuen. Een residu is het verschil tussen de geobserveerde en verwachte frequentie in een cel: f_o - f_e. Kijken we bijvoorbeeld naar het residu van vrouw-democraten, dan is dat de geobserveerde frequentie (573) minus de verwachte frequentie (522,9) = 50,1.

Maar hoe weet je nu of een residu groot genoeg is, dat het onwaarschijnlijk is dat het toeval is dat je die tegenkomt? Daarvoor gebruik je de gestandaardiseerde residuen. Je krijgt dit gestandaardiseerde residu (z) door het residu te delen door de standaardfout. Deze standaardfout is de fout die je verwacht wanneer H0 waar zou zijn.

Nu kunnen we de formule voor gestandaardiseerde residuen weergeven:

$$Z = \frac{f_0 - f_e}{se} = \frac{f_0 - f_e}{\sqrt{f_e(1 - \text{rijproportie})(1 - \text{kolomproportie})}}$$

Van deze gestandaardiseerde residuen weten we dat deze normaal verdeeld zijn, met een gemiddelde van 0, met een standaarddeviatie van 1. Wanneer het gestandaardiseerde residu boven of onder de 3 komt, is dat genoeg bewijs voor een bestaand effect in die cel.

Associatiematen voor ordinale variabelen

Nu kijken we naar associatiematen voor kruistabellen met ordinale variabelen. Bij ordinale variabelen kan zich een positief of een negatief verband voordoen. Een positief verband is wanneer iemand hoog op x scoort, ook hoog op y scoort, en wie laag op x scoort, laag op y scoort. Een negatief verband is wanneer iemand hoog op x scoort en laag op y scoort, en wie laag op x scoort, hoog op y scoort.

Concordante en discordante paren

We gaan dit onderzoeken aan de hand van concordante en discordante paren. Een paar van observaties is concordant wanneer de persoon die, ten opzichte van de persoon in een lagere klasse, hoger scoort op de ene variabele ook hoger scoort op de andere variabele. Een paar van observaties is discordant wanneer de persoon die hoger scoort op de ene variabele lager scoort op de andere variabele. Hieronder staat de berekening van het aantal concordante en het aantal discordante paren.

Om deze berekening te begrijpen, moet je goed kijken welke getallen uit de tabel worden vermenigvuldigd. De vermenigvuldigingen worden uiteindelijk bij elkaar opgeteld. Als je het eenmaal ziet, kun je het makkelijk zelf doen.

Concordante paren

Stel we hebben de volgende tabel, met betrekking tot ‘geluk’ en ‘inkomen’. In totaal hebben 67 mensen een beneden gemiddeld inkomen, 68 mensen hebben een gemiddeld inkomen en 22 mensen hebben een boven gemiddeld inkomen. We gaan kijken of mensen gelukkiger worden met naarmate het inkomen stijgt en ongelukkiger naarmate het inkomen daalt. Dit bekijken we door middel van het aantal concordante paren.

	Geluk			
Inkomen	Niet erg gelukkig	Redelijk gelukkig	Heel erg gelukkig	**Totaal**
Beneden gemiddeld	16 (24%)	36 (54%)	15 (22%)	**67 (100%)**
Gemiddeld	11 (16%)	36 (53%)	21 (31%)	**68 (100%)**
Boven gemiddeld	2 (9%)	12 (55%)	8 (36%)	**22 (100%)**
Totaal	**29**	**84**	**44**	**157**

Je berekent de concordante paren (C) als volgt. Je begint linksboven in de hoek (Beneden gemiddeld, niet erg gelukkig). Je streept alles weg dat in dezelfde rij staat, en alles weg dat in dezelfde kolom staat:

	Geluk			
Inkomen	Niet erg gelukkig	Redelijk gelukkig	Heel erg gelukkig	**Totaal**
Beneden gemiddeld	16 (24%)			**67 (100%)**
Gemiddeld		36 (53%)	21 (31%)	**68 (100%)**
Boven gemiddeld		12 (55%)	8 (36%)	**22 (100%)**
Totaal	**29**	**84**	**44**	**157**

Vervolgens gaan we wat er in de cel linksboven staat (16) vermenigvuldigen met alles in de overgebleven cellen. De overgebleven cellen bevatten namelijk allemaal personen die hoger scoren op de ene variabele én hoger scoren op de andere variabele, ten opzichte van alle personen in de cel linksboven (16). De vermenigvuldiging van deze cellen met de cel linksboven, is het vormen van concordante paren in de klasse 'beneden gemiddeld' en 'niet erg gelukkig' (de cellen 'totaal' worden hierbij genegeerd). Die paren bereken je als volgt:

16 * (36 + 21 + 12 + 8) = 1232

Nu gaan we een cel naar rechts (beneden gemiddeld, redelijk gelukkig), en doen we hetzelfde: strepen alles in dezelfde rij weg, en alles in dezelfde kolom weg. De overgebleven cellen bevatten namelijk allemaal personen die hoger scoren op de ene variabele én hoger scoren op de andere variabele, ten opzichte van alle personen in de tweede cel van links (36).

	Geluk			
Inkomen	Niet erg gelukkig	Redelijk gelukkig	Heel erg gelukkig	**Totaal**
Beneden gemiddeld		36 (54%)		**67 (100%)**
Gemiddeld			21 (31%)	**68 (100%)**
Boven gemiddeld			8 (36%)	**22 (100%)**
Totaal	**29**	**84**	**44**	**157**

Het is belangrijk dat je alleen naar de getallen aan de rechterkant van je cel kijkt. Nu doen we weer het getal uit onze cel, maal de overgebleven getallen:

36 * (21 + 8) = 1044.

Als we nog een cel naar rechts gaan (beneden gemiddeld, heel erg gelukkig), dan zien we dat er niks overblijft aan de rechterkant, dus deze cel heeft geen concordante paren. Dus gaan we een rij naar beneden en beginnen we weer aan de linkerkant, bij cel (gemiddeld, niet erg gelukkig). We doen weer hetzelfde: doorstrepen van alles in dezelfde rij en kolom, en we kijken alleen naar wat er rechts (en onder) overblijft:

	Geluk			
Inkomen	Niet erg gelukkig	Redelijk gelukkig	Heel erg gelukkig	**Totaal**
Beneden gemiddeld				**67 (100%)**
Gemiddeld	11 (16%)			**68 (100%)**
Boven gemiddeld		12 (55%)	8 (36%)	**22 (100%)**
Totaal	**29**	**84**	**44**	**157**

We gaan deze cel (11) vermenigvuldigen met de overgebleven cellen. De overgebleven cellen bevatten namelijk allemaal personen die hoger scoren op de ene variabele én hoger scoren op de andere variabele, ten opzichte van deze cel (11).

11 * (12 + 8) = 220

We gaan weer een cel naar rechts (gemiddeld, redelijk gelukkig), en doen we hetzelfde: strepen alles in dezelfde rij weg, en alles in dezelfde kolom weg. De overgebleven cel bevat namelijk personen die hoger scoren op de ene variabele én hoger scoren op de andere variabele.

	Geluk			
Inkomen	Niet erg gelukkig	Redelijk gelukkig	Heel erg gelukkig	**Totaal**
Beneden gemiddeld				**67 (100%)**
Gemiddeld		36 (53%)		**68 (100%)**
Boven gemiddeld			8 (36%)	**22 (100%)**
Totaal	**29**	**84**	**44**	**157**

We gaan deze cel (36) weer vermenigvuldigen met de overgebleven cel:

36 * 8 = 288.

Deze vier cellen (16, 36, 11, 36) waren de enige cellen die we konden paren met cellen rechtsonder. Dat betekent dat we nu alle concordante paren hebben berekend. Deze paren moet je optellen, dat geeft het totaal aantal concordante paren:

1232 + 1044 + 220 + 288 = 2784 concordante paren. Dus C = 2784

Discordante paren

Voor het berekenen van de discordante paren (D) doen we in principe hetzelfde, maar dan begin je rechtsboven in de hoek, en kijk je naar alles wat er linksonder overblijft. De cellen 'totaal' worden ook hierbij genegeerd. Hieronder is de hele tabel gegeven, zodat je zelf kan gaan wegstrepen.

	Geluk			
Inkomen	Niet erg gelukkig	Redelijk gelukkig	Heel erg gelukkig	**Totaal**
Beneden gemiddeld	16 (24%)	36 (54%)	15 (22%)	**67 (100%)**
Gemiddeld	11 (16%)	36 (53%)	21 (31%)	**68 (100%)**
Boven gemiddeld	2 (9%)	12 (55%)	8 (36%)	**22 (100%)**
Totaal	**29**	**84**	**44**	**157**

Het eerste disconcordante paar wordt gevormd met de cel helemaal rechtsboven (15). Streep alle cellen in dezelfde rij en in dezelfde kolom weg. Vermenigvuldig daarna de cel (15) met alle cellen linksonder.

15 * (11 + 36 + 2 + 12) = 915

We schuiven een kolom naar links. Hiermee (36) wordt het volgende discordante paar gevormd. Streep alle cellen in dezelfde rij en in dezelfde kolom weg en vermenigvuldig met de overgebleven cellen:

36 * (11 + 2) = 468

We schuiven een rij naar onder. Begin weer helemaal rechts en vermenigvuldig deze cel (21) met alle cellen linksonder:

21 * (2 + 12) = 294

We schuiven een kolom naar links. Met deze cel (36) wordt het laatste discordante paar gevormd:

36 * 2 = 72

Deze vier cellen (15, 36, 21, 36) waren de enige cellen die we konden paren met cellen linksonder. Dat betekent dat we nu alle discordante paren hebben berekend. Deze paren moet je optellen, dat geeft het totaal aantal discordante paren:

915 + 468 + 294 + 72 = 1749 discordante paren. Dus D = 1749

Wanneer je een positief verschil hebt tussen de concordante en discordante paren (C – D) is er een positieve relatie tussen de twee variabelen (inkomen en geluk) . Wanneer C – D negatief is, is er een negatieve relatie tussen de twee variabelen.

Gamma

Omdat je bij grotere steekproeven ook meer paren hebt en vaker ook grotere verschillen tussen C en D, standaardiseren we dit verschil. Dit standaardiseren geeft de geschatte gamma ($\hat{\gamma}$). De formule hiervoor is $\hat{\gamma} = \frac{C-D}{C+D}$.

We weten een aantal dingen over deze gamma: de waarde ervan valt tussen -1 en 1; gamma geeft aan of de relatie positief dan wel negatief is; hoe groter gamma, hoe sterker de samenhang tussen twee variabelen.