18. Categorische gegevens

Categorische gegevens

Bij categorische data is bestaat de uitkomstvariabele uit verschillende categorieën. Een observatie valt in één van deze categorieën, iemand is zwanger of niet zwanger, het kan niet allebei.

De theorie

Het gemiddelde uitrekenen van een categorische variabele heeft geen nut omdat de cijfers die zijn toegekend aan de categorieën arbitrair zijn. Bij categorische variabelen wordt er gekeken naar de frequenties, hoeveel observaties er zijn in een bepaalde categorie. Een tabel met de frequenties van alle categorieën wordt een *contingency tabel* genoemd.

De chi-square

Met Pearson's *chi-square test* wordt gekeken of er een verband is tussen twee categorische variabelen, bijvoorbeeld of het type training (beloning met eten of affectie) verschil maakt in of katten leren linedancen. Deze test vergelijkt de geobserveerde frequenties in de categorieën met de frequenties die je in die categorieën zou verwachten op basis van toeval. Zoals in hoofdstuk 2 al is besproken, wordt de fit (of error) van een model berekend door te kijken naar het gekwadrateerde verschil tussen de werkelijke data en het model:

$$\text{Error} = \sum(\text{geobserveerd} - \text{model})^2$$

Pearson's chi-square is ook hierop gebaseerd, alleen deel je nog door het model:

$$\chi^2 = \sum \frac{(\text{geobserveerd}_{ij} - \text{model}_{ij})^2}{\text{model}_{ij}}$$

De i staat hier voor de rijen in de contingency tabel en de j staat voor de kolommen. De geobserveerde waardes zijn de frequenties in de contingency tabel en het model zijn de verwachte frequenties. Voor elke cel in de tabel kan de verwachte frequentie berekend worden met deze formule:

$$\text{Model}_{ij} = E_{ij} = \frac{\text{rij totaal}_i \times \text{kolom totaal}_j}{n}$$

De n staat hier voor het totale aantal observaties. Met deze formule krijg je de verwachte frequentie voor elke cel in de tabel.

Dit getal vul je in bij ‘model’ in de chi-square formule, en je telt de uitkomst van elke cel bij elkaar op. Dit is de chi-square statistiek, waarbij je die waarde kunt vergelijken met de kritieke waarden in de chi-square distributie. Het aantal vrijheidsgraden is (r-1)(c-1), waarbij r het aantal rijen is en c het aantal kolommen.

Fisher's exact test
De chi-square test is een benadering van de chi-square distributie, waardoor de test bij een kleine steekproeven niet altijd nauwkeurig is. De verwachte frequentie in de cellen van de contingency tabel moeten minstens vijf zijn, dan is de benadering goed genoeg. Met kleine steekproeven kun je *Fisher's exact test* gebruiken om de exacte p-waarde van de chi-square statistiek te berekenen. Deze test is niet nodig bij grote steekproeven.

De likelihood ratio
De likelihood ratio is een alternatief voor de chi-square statistiek. Het idee hierachter is dat je een model creëert waarbij de kans om de geobserveerde data te verkrijgen maximaal is, en je dit model vergelijkt met de kans om de geobserveerde data te verkrijgen als de nulhypothese waar is. De formule voor deze ratio is:

$$L\chi^2 = 2\sum \text{geobserveerd}_{ij} \ln\left(\frac{\text{geobserveerd}_{ij}}{\text{model}_{ij}}\right)$$

i en j staan weer voor de rijen en kolommen van de contingency tabel en ln staat voor het natuurlijke logaritme, ln of log op je rekenmachine. Deze statistiek heeft ook een chi-square distributie en dezelfde vrijheidsgraden. In grote steekproeven is er weinig verschil tussen de twee statistieken, maar in kleine steekproeven kun je beter deze likelihood ratio gebruiken.

Yates's correctie
Yates´s continuïteitscorrectie is een correctie op de formule van Pearson en zorgt voor minder kans op een type I fout. Als je de deviatie van het model berekend, moet je hierbij 0.5 aftrekken van de absolute waarde van de deviatie, voordat je het kwadrateert. Absolute waarde betekent dat je het plus- of minteken negeert. De formule van Pearson gaat er dan als volgt uitzien.

$$\chi^2 = \sum \frac{(|\text{geobserveerd}_{ij} - \text{model}_{ij}| - 0.5)^2}{\text{model}_{ij}}$$

Yates's correctie corrigeert echter iets te veel, waardoor het te lage chi-square waardes geeft. Het is daarom beter om het niet te gebruiken.

Andere statistieken

Er zijn statistieken die de chi-square aanpassen zodat de steekproefgrootte en de vrijheidsgraden erin meegenomen worden. Ze proberen de reikwijdte van de statistiek te beperken zodat het tussen 0 en 1 valt.

De eerste is phi, een statistiek voor 2x2 contingency tabellen. Voor grotere tabellen ligt phi niet altijd tussen 0 en 1, dus wordt de contingency coefficient aangeraden. Deze heeft wel altijd een waarde tussen 0 en 1, maar bereikt de 1 eigenlijk nooit.

Het alternatief is daarom Cramér's V. Bij een 2x2 tabel zijn V en phi hetzelfde, maar wanneer een variabele meer dan twee categorieën heeft, kan V ook de maximale waarde bereiken. Deze is daarom het meest nuttig.

Meerdere categorische variabelen

Loglineaire analyse is een analyse voor wanneer je meer dan 2 categorische variabelen hebt.

Chi-square als regressie

De standaard formule voor regressie is:

Uitkomst = model + error

Het model bestaat uit de intercept, b_0, en de b-waardes voor de predictorvariabelen. Bij twee categorische variabelen met twee categorieën, kun je dit als een dummy variabele coderen. De ene categorie krijgt een code 0 en de andere een code 1. Het model wordt dan:

Uitkomst = $(b_0+b_1\text{Variabele1}+b_2\text{Variabele2}+b_3\text{Interactie})+\text{error}$

Om een model met categorische variabelen lineair te maken, moet je echter logaritmes gebruiken. Het model wordt dan:

$\text{Ln}(O_i) = \ln(\text{model}) + \ln(\varepsilon_i)$

$\text{Ln}(O_{ij}) = (b_0 + b_1\text{Variabele1} + b_2\ \text{Variabele2} + b_3\ \text{Interactie}) + \ln(\varepsilon_i)$

Als je de errorterm negeert, kun je de b_0 berekenen door de code 0 in te vullen voor de beide variabelen. De variabelen en de interactieterm zijn dan nul, waardoor $\text{Ln}(O_{ij}) = b_0$ overblijft.

De geobserveerde frequenties worden als uitkomst gebruikt, in plaats van geobserveerde gemiddelden zoals bij ANOVA. Je krijgt dan:

$\text{Ln}(O_{ij}) = b_0 + b_1x0+b_2x0+b_3x0$

$\text{Ln}(O_{ij}) = b_0$

$\text{Ln}(10) = b_0$

$b_0 = 2.303$

b_0 is dus de logaritme van de geobserveerde waarde bij de 0-codes van beide variabelen. De andere b-waardes kunnen berekend worden door de andere codes in te vullen in het model. Het model is hetzelfde als bij factor ANOVA, alleen worden hier nog log transformaties aan toegevoegd.

Een *verzadigd (saturated) model* is een model waarbij de gecodeerde variabelen de geobserveerde waardes volledig verklaren en er geen meetfout is. De standaardafwijkingen zijn dus allemaal 0.

De chi-square test kijkt of twee variabelen onafhankelijk van elkaar zijn. Het heeft geen interesse in het gecombineerde effect van de variabelen, dus de interactieterm wordt in het model weggelaten. Het model wordt dan:

$$\text{Ln (model)} = b_0 + b_1 \text{ Variabele1} + b_2 \text{ Variabele2}$$

Je kunt nu de geobserveerde waardes niet voorspellen zoals in het verzadigde model, omdat er informatie verloren is gegaan. Daardoor verandert de uitkomst en de b-waardes. De chi-square test is gebaseerd op verwachte frequenties, dus de b-waardes worden ook berekend op basis van deze verwachte frequenties (E). b_0 is dan de log van de verwachte waarden als alle categorieën 0 zijn. De b-waarden laten het verschil in verwachte frequenties zien tussen de variabelen.

Loglineaire analyse
Het lineaire model voor loglineaire analyse is hetzelfde als voor multiple regressie en ANOVA, behalve dat het logaritmes bevat. Een extra variabele in de vergelijking betekent een extra b-waarde en een extra interactieterm. Bij drie voorspellers (A, B, C) heb je dus ook de interactietermen AB, BC en AB en de drieweg interactie ABC, met bijbehorende parameters (b).

Loglineaire analyse werkt op basis van deze principes, maar dan andersom, zoals bij backwards regressie. Dat betekent dat je begint met het verzadigde model, waarbij de predictors de uitkomst perfect voorspellen. Bij een loglineaire analyse probeer je een eenvoudiger model op de data te passen zonder dat je enorm veel voorspellende power verliest. Als het eenvoudigere model niet erg verschilt van de complexe, behoud je het eenvoudige nieuwe model.

Het weghalen van de voorspellers gaat hiërarchisch, de interactie van de hoogste orde (ABC) wordt als eerst weggehaald. Met het nieuwe model worden de verwachte frequenties berekend en die verwachte frequenties worden dan vergeleken met de geobserveerde frequenties. Als het verwijderen van een bepaalde interactie geen significant effect heeft op de likelihood ratio, kan de term uit het model worden weggelaten. Zo ga je door tot je een effect vindt dat wel een effect heeft op het model.

Om te kijken of het nieuwe model de likelihood ratio verandert, wordt het volgende berekend.

$$LX2_{\text{Verandering}} = LX^2_{\text{Huidig model}} - LX^2_{\text{Vorige model}}$$

De assumpties voor categorische data

De chi-square test heeft twee belangrijke assumpties, namelijk onafhankelijkheid en verwachte frequenties. De onafhankelijkheid van de gegevens bij de chi-square betekent dat elke persoon slechts in 1 cel van de contingency tabel kan voorkomen. Iemand kan niet in twee categorieën vallen. Bij herhaalde metingen kan dus geen chi-square test gebruikt worden.

De tweede assumptie is dat bij een 2x2 contingency tabel (twee variabelen met twee categorieën) de verwachte frequenties in elke cel groter dan 5 moeten zijn. Bij grotere tabellen en met meer dan 2 variabelen (loglineaire analyse) is de regel dat de alle verwachte frequenties groter dan 1 moeten zijn, en niet meer dan 20% van de verwachte frequenties mag minder dan 5 zijn. Schending van deze assumptie betekent een gigantisch verlies van power.

Bij een 2x2 model kan Fisher's exacte test een uitkomst bieden bij schending van de assumptie van verwachte frequenties.

Bij loglineaire analyse heb je vier opties: het laten vervallen van een variabele (de variabele waarvan je het minste effect verwacht), het laten vervallen van een van de categorieën van de variabelen, meer data verzamelen of het accepteren van het powerverlies.

Als je data van een van de variabelen wil laten vervallen, kan dit als de hoogste orde interactie niet-significant is en ten minste één van de lagere orde interactietermen met de variabele die je wil verwijderen moet niet-significant zijn.

Je kunt ook categorieën van een variabele laten vervallen als je weinig observaties hebt in één van die categorieën. Je combineert die categorie dan met een andere. Dit kan alleen als die combinatie theoretisch logisch is.

Tenslotte, geen assumptie, maar wel belangrijk, is het vermelden dat kleine verschillen in celfrequenties al statistisch significant kunnen zijn in erg grote steekproeven. Daarom moet je kijken naar rij en kolom percentages om de effecten te interpreteren.

De chi-square in SPSS

Er zijn twee manieren waarop je categorische data in SPSS kun invoeren, via de ruwe data of via gewogen groepen. Als je de ruwe scores in SPSS invoert stelt elke rij een deelnemer voor. Je hebt dan net zoveel rijen als dat je deelnemers hebt, en je hebt zoveel kolommen als variabelen.

De gegevens kunnen ook zo ingevoerd worden dat er een extra variabele wordt aangemaakt om het aantal deelnemers per categorie weer te geven, de frequentie. In dit geval stelt elke rij de combinatie van de categorieën voor. Hierbij heb je net zoveel rijen als dat er combinaties van categorieën zijn. Twee variabelen met twee categorieën hebben dus vier rijen in SPSS. Deze methode scheelt veel werk.

In SPSS ga je voor de tweede methode naar Data – Weight Cases en dan selecteer je Weight cases by. De variabele waarop geselecteerd moet worden kan dan naar Frequency variabele gesleept worden.

Als eerste maak je in SPSS een contingency tabel bij Crosstabs, dan bekijk je de verwachte frequenties en voer je de chi-square test uit. Hiervoor ga je naar Analyze – Descriptive Statistics – Crosstabs. Sleep de ene variabele naar de rijen en de andere variabele naar de kolommen. Als er nog een derde categorie is, kan die ingevoerd worden bij de layer, waardoor de tabel opsplitst voor deze extra categorie.

Bij statistics vind je de verschillende test statistieken. In het voorbeeld worden de chi-square test, Phi, Cramer’s V en Lambda gebruikt.

Bij Cells kun je kiezen welke data je in de tabel wil hebben. Het is belangrijk om de verwachte frequenties (Expected Counts) te selecteren om de assumptie te controleren dat die niet onder de 5 zijn. Het is ook handig de rij, kolom en totaal percentages aan te vinken omdat die makkelijker te interpreteren zijn dan de werkelijke frequenties. Compare column proporties geeft een z-test die de frequenties in de cellen van de kolommen van de tabel vergelijkt. Als je deze gebruikt, moet je ook de Bonferroni correctie toepassen, bij Adjust p-values. Gestandaardiseerde residuen zijn ook handig om aan te vinken, mocht je een significant effect vinden.

De Exact knop geeft Fisher’s exact test, welke je kan gebruiken bij een kleine steekproef.

Output
Het eerste deel van de output is de contingency tabel van de categorische variabelen met de frequenties en percentages te zien. Bij de rij Expected count kun je controleren of aan de assumptie van meer dan 5 verwachte frequenties voldaan is.

Als je Compare column proporties hebt aangevinkt, staan er letters in het subscript bij de rij Count. Als dit verschillende letters zijn, betekent dit dat de proporties van de kolomvariabele significant van elkaar verschillen. Het gaat dus niet om de frequenties zelf, maar om de percentages.

Het volgende deel van de output laat de chi-square statistiek zien. De Pearson chi-square kijkt of er een relatie is tussen de twee variabelen. Als deze test significant is, is er inderdaad een relatie. In het voorbeeld betekent het dat er een significant verschil is tussen het type training of katten leren linedancen. Tenslotte, wanneer erom gevraagd is, wordt nog de tabel met de Phi, Cramér's V en de Contingency Coefficient weergegeven.

Gestandaardiseerde residuen
Met de gestandaardiseerde residuen kunnen we een significante chi-square test in stukjes hakken. Bij een 2x2 contingency tabel is het vrij duidelijk wat het verband precies is, maar met grotere tabellen kun je hiervoor gebruik maken van de gestandaardiseerde residuen. Het is vergelijkbaar met de testen na een ANOVA.

$$\text{Gestandaardiseerd residu} = \frac{\text{geobserveerd}_{ij} - \text{model}_{ij}}{\sqrt{\text{model}_{ij}}}$$

Deze formule lijkt op de formule voor de chi-square, behalve dat het residu niet wordt gekwadrateerd. Dat is alleen nodig om de deviaties positief te maken als je ze bij elkaar optelt. De gestandaardiseerde residuen hebben dus een directe relatie met de teststatistiek, omdat de chi-square ongeveer opgetelde gestandaardiseerde residuen zijn. Elk gestandaardiseerd residu is eigenlijk een z-score, wat betekent dat voor elke score gekeken kan worden naar de significantie.

De effectgrootte
Voor categorische gegevens kan de *odds ratio* berekend worden. Het kan het beste uitgerekend worden bij een 2X2 tabel. De odds ratio bereken je als volgt. In het voorbeeld over linedancende katten bereken je eerst de kans dat een kat danst als voedsel de beloning is:

$$\text{Odds}_{\text{dansen na voedsel}} = \frac{\text{aantal dansende katten beloond met voedsel}}{\text{aantal niet dansende katten beloond met voedsel}}$$

Op dezelfde manier bereken je de kans dat katten dansen na affectie.

$$\text{Odds}_{\text{dansen na affectie}} = \frac{\text{aantal dansende katten beloond met affectie}}{\text{aantal niet dansende katten beloond met affectie}}$$

De odds ratio is dan:

$$\text{Odds Ratio} = \frac{\text{Odds}_{\text{dansen na voedsel}}}{\text{Odds}_{\text{dansen na affectie}}}$$

De odds ratio van het voorbeeld is 6.65, wat betekent dat als een kat voedsel kreeg als beloning, de kans dat ze gingen dansen 6.65 keer groter was dan wanneer ze affectie kregen als beloning.

De resultaten rapporteren
Bij de chi-square wordt de statistiek, de significantie en het aantal vrijheidsgraden vermeld. Het is ook handig de contingency tabel weer te geven.

SPSS en loglineaire analyse

De gegevens voor loglineaire analyse worden op dezelfde manier ingevoerd als bij de chi-square test. Als eerste controleer je de verwachte frequenties in de contingency tabel. Dit doe je net zoals de chi-square test bij Crosstabs. De assumptie bij loglineaire analyse is dat geen enkele verwachte frequentie minder dan 1 mag zijn en dat maximaal 20% een verwachte frequentie van minder dan 5 heeft. Bij de rij Expected count zie je de verwachte waarden en kun je deze assumptie dus controleren.

Voor de analyse ga je naar Analyze – Loglinear – Model Selection. De variabelen die je in de analyse wilt hebben kan je slepen naar Factor(s). Bij Define range moet je aangeven welke codes je gebruikt hebt om de categorische variabelen te coderen.

In het basisscherm staat standaard de backward elimination methode aangevinkt. De andere optie is Enter, wat hetzelfde is als forced entry bij regressie. Bij loglineaire analyse is de backward methode echter het beste. Ook de knop Model kan het beste met rust gelaten worden.

Bij Opties kan je voor Parameter estimates kiezen, wat een tabel met geschatte parameters voor elk effect produceert. Met de optie Association table krijg je chi-square statistieken voor alle effecten in het model. Echter, als hogere orde interacties significant zijn, heeft het weinig zin om te kijken naar lagere orde effecten.

De output

Het eerste tabel van de output bevat een overzicht van het aantal deelnemers en het aantal categorieën van de variabelen. De tweede tabel bevat de verwachte en geobserveerde frequenties van alle combinaties van categorieën.

Daarna staat er een tabel met de Goodness of fit tests, met daarin Pearson's chi-square en de Likelihood ratio. Deze statistieken testen of de verwachte frequenties significant verschillen van de geobserveerde frequenties. We willen graag dat we een goed model hebben, dus geen significant verschil tussen de verwachte en geobserveerde frequenties. In de initiële output zijn de waardes allemaal 0 en kan de significantie niet worden berekend, omdat dit het verzadigde model is. Het model heeft een perfecte fit met de data.

De vraag is dan welke effecten uit het model verwijderd kunnen worden zonder de fit significant te verminderen. Dat staat in het volgende deel van de output. De K-way and High-Order Effects laat zien of het weghalen van interacties effect heeft. De eerste rij, K=1, vertelt of het verwijderen van de hoofdeffecten en de hogere orde effecten significante invloed zou hebben op de fit van het model. Als dit niet significant is, dan betekent dat dat het verwijderen van alles uit je model de fit niet verandert.

De volgende rij, K=2, vertelt of het verwijderen van tweeweg interacties en hogere orde interacties een significante invloed heeft op de fit van het model. Als dit significant is, zorgt het verwijderen van de tweeweg en drieweg interacties voor een significante vermindering van de fit van het model.

De derde rij, K=3, vertelt of het verwijderen van de drieweg interactie (en hogere orde interacties, als die er zijn) een significante invloed heeft op de fit van het model. Als dit significant is, zorgt het verwijderen van deze drieweg interactie voor een significante vermindering van de fit, wat betekent dat het dus niet verwijderd kan worden uit het model.

De onderste helft van de tabel, K-way effects, test hetzelfde als de bovenste helft, maar dan zonder de hogere orde effecten er bij te nemen. K=1 test hier dus alleen of het verwijderen van de hoofdeffecten de fit verminderen, K=2 test alleen de invloed van het verwijderen van de tweeweg interacties en K=3 test alleen de drieweg interactie. Als de drieweg interactie een significante invloed heeft, wordt niet verder gekeken naar de lagere orde effecten.

Bij de Association tabel (Partial Associations in de output) worden de effecten verder bekeken en kan je zien welke van de interacties significant is. Echter, ook al is een van de interacties niet significant, als de hoogste orde interactie significant is, wordt daar verder niet naar gekeken. Parameter estimates laat hetzelfde zien als de association tabel, maar dan op basis van z-scores in plaats van op basis van chi-square tests.

Het volgende deel van de output laat de backward eliminatie zien. Het laat zien hoe het model eruit ziet als de drieweg interactie weggehaald wordt. Als deze interactie een significant heeft, blijft het in het model en zijn er geen verdere stappen nodig. Tenslotte evalueert SPSS dit uiteindelijke model met de likelihood ratio. Dit moet niet significant zijn, in dat geval heeft het model een goede fit voor de data.

Na deze analyse moet deze interactie geïnterpreteerd worden. Het is handig om hiervoor de frequenties van de categorieën te plotten in een staafdiagram.

Het vervolg

Een andere manier om de drieweg interactie te interpreteren, is om chi-square analyses uit te voeren op de verschillende niveaus van de variabelen. In het voorbeeld waarbij je katten en honden vergelijkt in het effect van training voor eten of affectie op het leren dansen, kun je een aparte chi-square test uitvoeren voor katten en honden, en de resultaten met elkaar vergelijken.

De effectgrootte

Ook bij loglineaire analyse kan de effectgrootte het best met de odds ratio uitgerekend worden. Bij meer dan 2 variabelen of meer dan 2 categorieën is de beste methode het in stukjes te delen zodat je 2x2 tabellen krijgt en over die 2x2 tabellen de odds ratio te berekenen. In het voorbeeld zou je bijvoorbeeld de odds ratio voor katten en honden apart uitrekenen.

Het rapporteren van de resultaten

Vermeld bij loglineaire analyse in ieder geval de likelihood ratio statistiek (^{2}). Voor significante effecten kun je de verandering van de chi-square rapporteren. Je kan ook de z-scores met bijbehorende betrouwbaarheidsintervallen vermelden. Als je hogere orde interacties in kleinere stukjes opdeelt, vermeld dan voor elk stukje de chi-square statistiek.