11. Meerdere gemiddelden vergelijken, ANOVA

Analyse van variantie (ANOVA) wordt gebruikt wanneer er situaties zijn waarbij er meer dan twee condities vergeleken worden. In dit hoofdstuk wordt de *onafhankelijke ANOVA* uitgelegd. Dit gebruik je als er verschillende proefpersonen in de verschillende condities zijn.

De theorie

Hoewel ANOVA en regressie vaak los van elkaar worden besproken, kun je de ANOVA het best opvatten als een lineair model, zoals die al het hele boek besproken wordt.

ANOVA laat zien of drie of meer gemiddelden gelijk zijn aan elkaar. Dit wordt gedaan met de F-statistiek. Dit vergelijkt de systematische variantie met de niet-systematische variantie.

De F-ratio wordt ook bij regressie gebruikt. Bij regressie laat het zien hoe goed het model de uitkomst vergeleken met de meetfout kan voorspellen. Bij ANOVA testen we het verschil tussen gemiddelden door een regressiemodel erop te passen, en gebruiken we F om te kijken hoe goed de gegevens in het model passen.

ANOVA kan als een multiple regressievergelijking gezien worden, waarbij het aantal voorspellers 1 minder is dan het aantal categorieën van de onafhankelijke variabele. De categorieën krijgen dan een dummycodering. Als je de invloed van verschillende doses (placebo, lage dosis en hoge dosis) Viagra op het libido wil meten, wordt het model:

$\text{Libido} = b_0 + b_1\text{Laag} + b_2\text{Hoog} + \text{error}$.

Als je predictor uit groepen bestaat, is de beste gok voor de voorspelde waarde het gemiddelde van die groep. De placebogroep heeft dus als voorspelde waarde het gemiddelde van de placebogroep. Deze basisgroep is gecodeerd met 0 op alle dummy variabelen. Zonder de errorterm wordt het model voor de placebogroep dus als volgt:

$\text{Libido} = b_0 + (b_1 \times 0) + (b_2 \times 0)$

$\text{Libido} = b_0$

$\bar{X}_{\text{placebo}} = b_0$

De intercept van het regressiemodel (b_0) is dus altijd het gemiddelde van de baseline.

Bij de hoge dosis groep is deze groep met een 1 gecodeerd, de andere groepen met 0. Het model voor deze groep is:

$\text{Libido} = b_0 + (b_1 \times 0) + (b_2 \times 1)$

$\text{Libido} = b_0 + b_2$

De intercept was het gemiddelde van de placebogroep, en de voorspelde waarde voor de hoge dosis groep is het groepsgemiddelde.

$\bar{X}_{\text{hoge dosis}} = \bar{X}_{\text{placebo}} + b_2$

$b_2 = \bar{X}_{\text{hoge dosis}} - \bar{X}_{\text{placebo}}$

Het model voor de lage dosis groep:

Libido = b_0 +(b_1 x 1) + (b_2 x 0)

Libido = $b_0 + b_1$

$\bar{X}_{\text{hoge dosis}} = \bar{X}_{\text{placebo}} + b_1$

$b_1 = \bar{X}_{\text{hoge dosis}} - \bar{X}_{\text{placebo}}$

Dus, b_1 staat in het model voor het verschil tussen het gemiddelde van de lage dosis groep en de placebogroep en b_2 staat voor het verschil tussen het gemiddelde van de hoge dosis groep en de placebogroep.

F-ratio

Je test de verschillen tussen groepsgemiddelden met de F-ratio. Wanneer alle gemiddelden gelijk zijn, zijn de b coëfficiënten allemaal nul. Wat we weten van regressie is:

- Het simpelste model is het overkoepelende gemiddelde (grand mean).
- Er kan een ander model worden toegepast die de hypotheses volgt. Als dit model goed bij de data past, is het beter dan het overkoepelende gemiddelde als model.
- De regressiecoëfficiënten beschrijven het model.
- De regressiecoëfficiënten bepalen de vorm van het model. Hoe groter de coëfficiënten, hoe groter het verschil is tussen het model en het overkoepelende gemiddelde.
- Bij experimenteel onderzoek laten de b coëfficiënten het verschil zien tussen de groepsgemiddelden.
- Als de verschillen groot genoeg zijn is het model beter dan het overkoepelende gemiddelde.
- In dat geval zijn de groepsgemiddelden significant verschillend.

SS_T

Om de totale variantie te vinden gebruiken we de volgende formule:

$SS_T = \sum(x - \bar{x}_{grand})^2$ of

$SS_T = s^2_{grand}\,(n-1)$

De *grote variantie* is de variantie van alle scores ongeacht uit welke conditie ze komen. Het aantal vrijheidsgraden is hier 1 minder dan de totale steekproefgrootte (df = N-1).

SS_M

De kwadratensom van het model laat zien hoeveel van de totale variantie verklaard kan worden door het feit dat verschillende gegevens uit verschillende groepen komen. Het laat het verschil zien tussen de voorspelde waarde door het model en het grote gemiddelde.

SS_M bereken je door eerst het verschil tussen het gemiddelde van elke groep ($\overline{X_k}$) en het grote gemiddelde te berekenen. Daarna kwadrateer je alle gemiddelden. Vervolgens vermenigvuldig je elk resultaat met het aantal deelnemers in de groep (n_k). Ten slotte tel je de waardes van elke groep op. In formulevorm:

$$SS_M = \sum n_k(\overline{X}_k - \overline{X}_{grand})^2$$

De vrijheidsgraden zijn hier altijd 1 minder dan het aantal groepen (df = k-1).

SS_R

De kwadratensom van de residuen vertelt ons hoeveel variantie er niet verklaard kan worden door het model. Het is het verschil tussen de werkelijke data en wat er voorspeld was. Het laat het verschil zien tussen de score van een deelnemer en het gemiddelde van de groep waar de deelnemer in zit. De SS_R is de SS van elke groep bij elkaar opgeteld. Omdat de SS en de variantie nauw samenhangen, kan het met de volgende formules berekend worden:

$$SS_R = \sum (X_k - \overline{X_k})^2 \quad \text{of}$$

$$SS_R = \sum s_k^2 (n_k - 1)$$

Het aantal vrijheidsgraden is bij de kwadratensom van de residuen het totale aantal vrijheidsgraden min de vrijheidsgraden van het model ($df_R = df_T - df_M$). Ook wel N-k.

De gemiddelde kwadratensom

Omdat de SS afhankelijk is van het aantal scores waarmee het berekend wordt, wordt de gemiddelde kwadratensom berekend. Dit heet de *Mean Square* (MS). Het wordt als volgt berekend:

$$MS_M = \frac{SS_M}{df_M}$$

$$MS_R = \frac{SS_R}{df_R}$$

MS_M laat de verklaarde variantie zien en MS_R de variantie die verklaard wordt door externe factoren.

Vanuit deze twee statistieken kan de F-ratio berekend worden:

$$F = \frac{MS_M}{MS_R}$$

Wanneer deze ratio lager dan 1 is laat het een niet-significant effect zien. Een F-ratio groter dan 1 hoeft niet per se significant te zijn. Bij de F-distributie in de Appendix kun je opzoeken wat de maximale F-waarde is die bij een bepaald aantal vrijheidsgraden per toeval nog gevonden kan worden, ofwel de kritieke waarde. Als de kritieke waarde wordt overschreden, is de F-ratio significant en kun je concluderen dat de gemiddelden niet gelijk zijn.

De F-ratio vertelt je alleen of er verschillen in gemiddelden zitten, maar niet welke groepen precies van elkaar verschillen. Daarom wordt een ANOVA ook wel een *omnibus test* genoemd. Met follow-up tests kun je erachter komen welke groepen van elkaar verschillen.

Assumpties

Net als bij andere lineaire modellen, is er de assumptie dat de varianties in de groepen gelijk zijn, dus homoscedasticiteit. Deze assumptie kun je testen met Levene's test. Als Levene's test significant is, verwerp je de nulhypothese die homoscedasticiteit aanneemt, en kun je de *Brown-Forsythe F* gebruiken of de *Welch's F.*

ANOVA is een robuuste test als de F type I fouten onder controle heeft en de F genoeg power heeft. Scheve distributies lijken de power en type I fouten weinig te beïnvloeden. Kurtosis heeft echter wel invloed. Alleen wanneer groepsgroottes gelijk zijn is de F-statistiek vrij robuust voor schendingen van de assumptie van normaliteit. Ook is de ANOVA redelijk robuust voor schendingen van de assumptie van homoscedasticiteit, mits de groepsgroottes gelijk zijn.

De assumptie van onafhankelijkheid mag in elk geval niet geschonden worden. Als de scores met elkaar correleren, neemt de kans op een type I fout heel snel toe.

Als de assumptie van gelijke varianties is geschonden, kan F worden aangepast, maar bij schending van normaliteit moet de data getransformeerd worden. Je kunt ook de Kruskal-Wallis test gebruiken, die heeft geen assumptie van normaliteit. Andere robuuste metingen zijn niet beschikbaar via SPSS, maar wel via R.

Contrasten

Om erachter te komen tussen welke gemiddelden het verschil zit, moet je een follow-up test doen bij een significante ANOVA. Er zijn hiervoor twee methodes die niet zorgen voor een inflatie van de kans op een type I fout.

Ten eerste kun je de variantie in componenten opdelen. Dit kan gedaan worden door *geplande vergelijkingen* (*geplande contrasten*) en is handig als je specifieke hypothesen hebt.

Je kan ook elke groep met een t-toets vergelijken en dan een strenger significantiecriterium hanteren om de kans op een type I fout gelijk te houden. Dit worden *post hoc vergelijkingen* genoemd en is handig wanneer je geen specifieke hypothesen hebt. Eerst worden contrasten verder besproken.

Voor geplande vergelijkingen moeten de hypothesen gemaakt zijn voordat er gegevens verzameld worden. Voor deze methode wordt de variantie in kleine onafhankelijke stukken gedeeld.

Je kunt bijvoorbeeld eerst de controlegroep vergelijken met andere twee condities. Daarna kun je die twee condities weer met elkaar vergelijken. Er worden altijd twee delen variantie vergeleken. Wanneer een groep in een contrast alleen staat kan het niet meer in een volgend contrast gebruikt worden. Daarom zijn er k-1 contrasten.

Als eerst wordt de variantie opgedeeld in: de variantie verklaard door het model en de variantie die niet verklaard wordt door het model. Voor experimenteel onderzoek is het handig als de eerste splitsing de splitsing tussen de experimentele groep en de controle groep is. Splitsingen waarbij experimentele groepen met elkaar vergeleken worden, zijn gebaseerd op je hypothesen.

Bij het voorbeeld over Viagra is het eerste contrast die tussen de placebogroep en de twee dosisgroepen. Het tweede contrast is tussen de hoge dosis groep en de lage dosis groep. Als er in het tweede contrast een significant verschil is, weet je dat de lage dosis groep significant verschilt van de hoge dosis groep, maar weet je niet zeker of het verschilt van de placebogroep. Hiervoor heb je een post hoc test nodig.

Gewichten bij contrasten
Wanneer we contrasten maken geven we waardes aan de variabelen in het regressiemodel. *Gewichten* zijn de waarden die gegeven worden aan de dummy variabelen. Er zijn een aantal regels die helpen bij de gewichten.

- Kies verstandige vergelijkingen. Als een groep alleen staat, kan het immers niet meer gebruikt worden in verdere contrasten.
- Groepen met positieve waarde worden vergeleken met groepen met een negatieve waarde.
- De som van de gewichten van een contrast moet nul zijn.
- Als een groep niet meedoet in de vergelijking krijgt het een gewicht van nul. Hiermee is het verwijderd uit de berekeningen.
- Het gewicht dat wordt toebedeeld aan de groepen binnen een deel variantie is gelijk aan het aantal groepen in het andere deel van het contast.

Je kunt nu een regressievergelijking van de contrasten opstellen:

$$\text{Libido} = b_0 + b_1\text{Contrast}_1 + b_2\text{Contrast}_2$$

Contrast 1 was placebogroep versus de twee dosisgroepen, contrast 2 was de groep met lage dosis versus de groep met hoge dosis. Wanneer alle contrasten optellen tot nul kunnen we zeggen dat de contrasten onafhankelijk of *orthogonaal* zijn. Dit is belangrijk omdat als de

b coëfficiënten onafhankelijk zijn, de p-waardes ook ongecorreleerd zijn.

Niet-orthogonale vergelijkingen
Dit komt voor wanneer je een groep die al een keer alleen is gebruikt ook nog in een ander contrast gebruikt. Wanneer dit het geval is, zijn de contrasten dus gecorreleerd, en daarmee zijn ook de p-waardes gecorreleerd. Om een inflatie van de familywise error rate te voorkomen, moet je conservatiever zijn in je significantieniveau.

Standaard contrasten
In SPSS zijn er een aantal contrasten die standaard aangeboden worden (zie tabel op blz. 456). SPSS ziet dan de laagste code als groep 1 en de hoogste code als groep 2. Sommige standaard contrasten zijn orthogonaal en andere zijn niet-orthogonaal.

Polynomiale contrasten

Een *polynomiaal contrast* kijkt naar de trend van de gegevens (hoe de lijn in de grafiek loopt). De simpelste is de *lineaire trend*, dit is een rechte lijn. Bij de *kwadratische trend* loopt er een bocht in de lijn waardoor de lijn eerst in de ene richting loopt en vervolgens de andere richting op gaat. Je hebt minimaal drie groepen nodig om een kwadratische trend te zien.

Een stap verder is de *cubic trend*. In deze lijn zijn twee bochten de zien. Er zijn minimaal vier groepen nodig om deze trend te zien. De *quartic trend* heeft nog een bocht extra, en heeft dus ook minimaal vijf groepen nodig. De trends zijn van belang als er een volgorde in de categorieën van de onafhankelijke variabele zitten. Elke trend heeft een aantal codes voor de dummyvariabelen in het regressiemodel.

Post hoc
Vaak heeft de onderzoeker geen specifieke hypothesen en wil het de gegevens verkennen. Post hoc toetsen zijn *paarsgewijze vergelijkingen* die alle verschillende combinaties toetsen. Bij deze methode wordt de familywise meetfout voorkomen door het significantieniveau voor alle testen .05 te houden. Het kan dus door de α te delen door het aantal vergelijkingen. Dit heet de *Bonferroni correctie*. Hierdoor is de kans op type II fout groter.

Drie dingen moeten bekeken worden als een post hoc test uitgevoerd gaat worden. Eerst moet bekeken worden of de type I fout onder controle is, dan wordt gekeken of de type II fout onder controle is (de toets moet genoeg statistische power hebben) en als laatst moet gekeken worden of de toets betrouwbaar is als niet aan alle assumpties voldaan is.

Type I en II fouten

Een test die type I fouten goed onder controle houdt, doet dat ten koste van de power (type II fout). Als de kans op een type II fout klein is, is de kans op een type I fout groter. De least-significant difference probeert de type I fout helemaal niet te controleren en voert voor alle combinaties een t-toets uit. De Studentized Newman-Keuls procedure is ook een liberale toets en let niet op de familywise meetfout.

Testen die wel op de type I fout letten, maar daarmee niet de statistische power hoog kunnen houden zijn de Bonferroni en Tukey toets. Wanneer je gelijke steekproefgrootte en groepsvariantie hebt kan je het beste Tukey of REGWQ gebruiken. REGWQ heeft een goede power en een strakke controle over de type I fout. Het kan echter niet gebruikt worden bij verschillende steekproefgroottes.

Wanneer niet aan de assumpties wordt voldaan
Gabriel's kan gebruikt worden wanneer de steekproefgroottes licht verschillen. Bij grote verschillen kan je beter Hochberg's GT2 gebruiken, maar dan moet er wel gelijke variantie zijn. Bij onzekerheid over gelijke groepsvariantie kan je het beste de Games-Howell procedure gebruiken.

Eenweg ANOVA in SPSS

Voor eenweg ANOVA ga je naar Analyze – Compare means – One-way ANOVA. Er is een ruimte voor de afhankelijke variabelen en voor de factor (onafhankelijke variabele). Bij Contrasts kan je geplande contrasten vinden. Als je polynomial aanvinkt kan je de trend van de gegevens onderzoeken. Als je slechts drie groepen hebt, is het zinloos om een hoger niveau dan quadratic aan te klikken. Bij de tabel voer je ook in welke gemiddelden je vergeleken wilt hebben en geef je ook het gewicht bij het contrast. Het eerste gewicht dat je invoert geldt voor de eerste groep. Het wordt ingevoerd bij Coefficients.

Bij post hoc tests kan je aangeven welke procedures SPSS uit kan voeren. Bij Options is het handig om homogeneity of variance test aan te vinken. Wanneer niet aan deze assumptie voldaan kan worden, kunnen ook nog de *Brown-Forsythe F* en de *Welch's F* gekozen worden.

Bij de optie Bootstrap is het belangrijk om te weten dat de belangrijkste test, de F-ratio, niet wordt gebootstrapped. De bootstrap werkt bij de gemiddelden, als je vraagt om beschrijvende statistieken, en bij de contrasten en post hoc tests.

Output

Als eerst is het handig om een staafdiagram van de meetfouten te maken. Als ze allemaal overlappen betekent dit dat er geen tussengroep verschillen zijn.

De output geeft eerst de test van homogeniteit van varianties. Wanneer Levene's test significant is kunnen we zeggen dat de varianties significant van elkaar verschillen. De Brown-Forsythe F of de Welch's F kan dan gerapporteerd worden.

De ANOVA tabel in de output wordt onderscheiden in een tussengroep deel en een binnengroep deel. De binnengroep is de niet-systematische variatie in de gegevens. Ook staat in deze tabel de F-ratio en de p-waarde.

De tabel Contrast Coefficients laat de contrasten zien. Het is handig deze tabel na te kijken om er zeker van de zijn dat de gewichten goed ingevoerd zijn. Bij contrast tests kan je naar het bovenste gedeelte kijken wanneer de Levene's test niet significant was.

In de tabel Multiple comparisons vind je de post hoc tests. Hier worden de groepen allemaal paarsgewijs met elkaar vergeleken. Bij elk paar staat een p-waarde vermeld. Een significante p-waarde betekent een verschil tussen de twee groepen. Tukey's test en de REGWQ test maken subsets van groepen die niet significant van elkaar verschillen.

Een gemiddelde dat rekening houdt met de relatie tussen de variantie en de steekproefgrootte is het *harmonische gemiddelde.* Het vermindert de bias die er zou zijn als de steekproefgroottes ongelijk zouden zijn.

De effectgrootte

Voor ANOVA kunnen we R^2 berekenen door SS_M te gebruiken. Bij ANOVA wordt het *eta squared* genoemd. Dus:

$$R^2 = \eta^2 = \frac{SS_M}{SS_T}$$

Om de effectgrootte voor de populatie in de schatten gebruiken we *omega squared*.

$$\omega^2 = \frac{SS_M - (df_M)MS_R}{SS_T + MS_R}$$

ω^2 is meestal accurater dan R^2. Het is echter vaak interessanter om de kijken naar de effectgrootte van de contrasten.

$r_{contrast} = \sqrt{\frac{t^2}{t^2 + df}}$

Bij het rapporteren van een ANOVA vermeld je de F-ratio en de vrijheidsgraden.