

Onderzoekspracticum 2, aantekeningen college 5

Toetsingsschema enkelvoudige ANOVA (variantieanalyse)

1. Formuleer je onderzoeksvraag.
2. Ga de aannamen af en ga na of eraan is voldaan.
 - De afhankelijke variabele moet normaal verdeeld zijn. Of dit het geval is, kan worden bepaald met een normaalkwantielplot. Bij schending zijn de resultaten nog wel betrouwbaar zolang de totale steekproef groot genoeg is. In het boek *Introduction to the Practice of Statistics* van Moore, McCabe en Craig wordt echter niet aangegeven hoeveel proefpersonen de groepen bij enkelvoudige variantieanalyse in totaal moeten bevatten.
 - Proefpersonen moeten aselekt getrokken zijn. Als dit niet het geval is, dan zijn de resultaten beperkt generaliseerbaar.
 - Proefpersonen moeten onafhankelijk van elkaar zijn getrokken. Als dit niet het geval is, dan gebruik je een ander soort variantieanalyse.
 - Er moeten gelijke populatievarianties zijn. Dit kan worden gecontroleerd met de vuistregel: er zijn gelijke varianties als de grootste standaarddeviatie minder dan twee keer de kleinste standaarddeviatie is. Als hier niet aan voldaan is, zijn de resultaten nog wel betrouwbaar bij (vrijwel) gelijke steekproefgroottes.
3. Stel je hypothesen op.
 - $H_0: \mu_1 = \mu_2 = \mu_3 = \dots = \mu_i$
 - H_a : niet alle μ_i 's zijn gelijk
4. Kies de juiste toets, in dit geval de F-toets, en bepaal α (vaak 0.05)
5. Voer je berekening uit. $F = MSG / MSE$ (zie college 4).
6. Zoek de bijbehorende p-waarde op in Tabel E (*Introduction to the Practice of Statistics* van Moore, McCabe en Craig).
7. Neem een beslissing: vergelijk p met α . Als $p < \alpha$, dan verwerp je de nulhypothese.
8. Bereken de effectgrootte door middel van R^2 . $R^2 = SSG / SST$ (zie college 4).
9. Trek je conclusie en beschrijf deze inhoudelijk. Plaats kanttekeningen over de aannamen.

Met behulp van dit toetsingsschema kan worden bepaald of gemiddelden van elkaar verschillen, maar het wordt niet duidelijk welke gemiddelden dat zijn. Om te bepalen welke gemiddelden verschillen, gebruiken we multiple vergelijkingen. Op deze manier kunnen gemiddelden in paren vergeleken worden.

Multiple vergelijkingen

Als we een experiment uitvoeren met drie groepen, dan moeten er drie toetsen uitgevoerd worden om de gemiddelden paarsgewijs met elkaar te vergelijken: een toets voor groep 1 en 2, een toets voor groep 2 en 3 en een toets voor groep 1 en 3. Een toets voor een vergelijking tussen bijvoorbeeld groep 1 en 2 ziet er als volgt uit:

$$t_{12} = \frac{(\bar{x}_1 - \bar{x}_2)}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Let op: s_p is hierin niet de samengestelde standaarddeviatie van groep 1 en 2, maar de samengestelde standaarddeviatie van groep 1, 2 en 3 (zie college 4 voor het berekenen van de samengestelde standaarddeviatie). Hierdoor wordt de standaarddeviatie van de populatie stabiel geschat en dit vergroot de power, de kans dat verschillen worden ontdekt. De algemene formule voor het uitvoeren van multiple vergelijkingen is:

$$t_{ij} = \frac{(\bar{x}_i - \bar{x}_j)}{s_p \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}}$$

Hierin geven de letters i en j de groepen aan. We concluderen dat gemiddelden van elkaar verschillen als $t_{ij} > t^{**}$. Als dit namelijk het geval is, dan is $p < \alpha$ en verwerpen we de nulhypothese. De waarde van t^{**} is afhankelijk van de methode die we voor multi-pele vergelijkingen gebruiken. Er zijn twee methoden: de LSD-methode en de Bonferroni-methode.

LSD-methode

De LSD-methode (Least Significant Difference) is de standaardmethode.

- De kritieke waarde t^{**} is hierbij de waarde die hoort bij $\alpha/2$ van een t (DFE)-verdeling. Dat je de alpha deelt door twee heeft te maken met het feit dat je tweezijdig toetst.
- Het aantal vrijheidsgraden is DFE: het totaal aantal proefpersonen min het aantal groepen ($N - I$).
- In tabel D (*Introduction to the Practice of Statistics* van Moore, McCabe en Craig) kan de bijbehorende waarde van t^{**} worden gevonden.

Bijvoorbeeld: we hebben drie groepen met in elke groep tien proefpersonen en we toetsen bij een alpha van 0.05. Om de kritieke waarde te bepalen, delen we eerst alpha door twee. De juiste waarde wordt dan 0.025. Het aantal vrijheidsgraden is $N - I$, dus in dit geval $30 - 3 = 27$. In tabel D zoeken we de t -waarde op bij $df = 27$ en $p = 0.025$. De waarde die we dan vinden is 2.052 en dit is de kritieke waarde t^{**} .

Een nadeel van de LSD-methode is dat het de kans vergroot op significante effecten die enkel door toeval worden veroorzaakt. Dit wordt kanskapitalisatie genoemd. Als er drie groepen paarsgewijs met elkaar vergeleken worden, dan is de kans op een type-1 fout (onterecht de nulhypothese verwerpen) 5% bij een alpha van 0.05. De kans op minstens één type-1 fout is over de drie toetsen heen drie keer die 5%. Dit is drie keer zo groot als gewenst. Om dit te voorkomen, kan de Bonferroni-methode gebruikt worden.

Bonferroni-methode

De Bonferroni-methode zorgt ervoor dat de totale kans op een onterechte verwerping van de nulhypothese onder *alle* vergelijkingen gelijk blijft aan een alpha van 0.05.

- Als je drie keer paarsgewijs toetst, zorg je ervoor dat je ook drie keer zo streng toetst.
- Deel $\alpha/2$ door het aantal toetsen dat je uitvoert.
- Verder werkt de procedure hetzelfde als bij de LSD-methode.

Bijvoorbeeld: we hebben weer drie groepen met in elke groep tien proefpersonen en we toetsen bij een alpha van 0.05. Om de kritieke waarde t^{**} te vinden, delen we $\alpha/2$ door het aantal toetsen dat we uitvoeren. Bij drie groepen kunnen we drie toetsen uitvoeren, dus we delen 0.025 door 3. $0.025/3 = 0.0083$, oftewel 0.83%. Deze waarde is echter niet in Tabel D te vinden. Wel kan SPSS de bijbehorende t -waarde uitrekenen. Bij een waarde van 0.0083 hoort een t^{**} -waarde van 2.552.

Simultane betrouwbaarheidsintervallen

De betrouwbaarheidsintervallen van $\mu_i - \mu_j$ (het verschil) tussen populatiegemiddelden hebben de volgende vorm:

$$(\bar{x}_i - \bar{x}_j) \pm t^{**} s_p \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}$$

De kritieke waarde t^{**} hangt af van de methode die gebruikt is om deze waarde te bepalen. S_p is weer de samengestelde standaarddeviatie van *alle* groepen, niet alleen van de twee groepen waarvan het betrouwbaarheidsinterval wordt bepaald. Bij drie groepen en een 95%-betrouwbaarheidsinterval geven de

drie intervallen van $\mu_1 - \mu_2$, $\mu_2 - \mu_3$ en $\mu_1 - \mu_3$ samen 95% zekerheid dat ze *alle drie tegelijkertijd* de waarden bevatten van $\mu_1 - \mu_2$, $\mu_2 - \mu_3$ en $\mu_1 - \mu_3$. Simultaan betekent gelijktijdig en daarom spreken we van simultane betrouwbaarheidsintervallen.

SPSS

In SPSS kunnen de LSD-methode en de Bonferroni-methode worden uitgevoerd. Via Analyze → Compare Means → One-way ANOVA wordt de variantieanalyse uitgevoerd. In het One-Way ANOVA scherm kan worden gekozen voor de optie 'post hoc', waarbij vervolgens de opties 'LSD' en 'Bonferroni' kunnen worden aangevinkt. Bij de knop 'options' kan verder nog worden gekozen voor 'means plot'. In de uitvoer wordt dan een grafiek weergegeven met daarin de gemiddelde waarden van alle groepen.

Als SPSS aangeeft dat er een p-waarde is van 0.000 (onder Sig. in de tabel), dan schrijven we dit op als $p < 0.001$. Als we dit zouden noteren als $p = 0.000$, dan zou dit namelijk suggereren dat de kans dat iets op toeval berust gelijk is aan nul en dit is niet het geval. Dat SPSS wel een waarde van 0.000 geeft, heeft te maken met afronden op drie decimalen.

Wat betreft de p-waarden in SPSS is er bij Bonferroni iets raars aan de hand. SPSS vermenigvuldigt de kansen namelijk met het aantal toetsen dat is uitgevoerd. Hierdoor zijn de p-waarden bij Bonferroni (onder Sig. in de tabel) anders dan de p-waarden bij LSD. Gebruik in SPSS altijd de p-waarden die bij LSD staan. Let op: de waarden van de betrouwbaarheidsintervallen zijn bij Bonferroni wel juist en kun je gebruiken.

Het kan wel eens voorkomen dat er volgens SPSS geen significant verschil is tussen de gemiddelden van groep 1 en 2 en ook niet tussen de gemiddelden van groep 2 en 3, maar wel tussen de gemiddelden van groep 1 en 3. Logisch gezien lijkt dit paradoxaal. Het kan echter zo zijn dat de steekproeven te klein waren, waardoor een effect te klein was om gedetecteerd te worden. Dat een effect te klein is om gevonden te worden, betekent niet dat het effect niet bestaat.