7. Het vergelijken van twee groepen

Het vergelijken van groepen

Vaak worden twee groepen met elkaar vergeleken. Bij kwantitatieve variabelen kijk je dan naar gemiddelden, en bij categoriale variabelen kijk je dan naar proporties. Wanneer je twee groepen met elkaar vergelijkt, creëer je een binaire variabele: een variabele met twee categorieën (soms ook wel dichotoom genoemd). Stel bijvoorbeeld dat je mannen en vrouwen vergelijkt, dan creëer je een binaire variabele 'geslacht' met de categorieën mannen en vrouwen. Het vergelijken van deze groepen is een voorbeeld van een bivariate statistische methode.

Twee groepen kunnen afhankelijk en onafhankelijk van elkaar zijn. De groepen zijn afhankelijk wanneer de respondenten van nature 'matchen' met elkaar, bijvoorbeeld wanneer je dezelfde groep gebruikt voor en na een meting. Stel dat je wilt weten of studenten betere resultaten hebben na een bepaald lesprogramma, dan is de kans groot dat de studenten die al beter presteerden voor het lesprogramma ook beter presteren na het programma. De twee resultaten zijn dus afhankelijk van elkaar. Groepen zijn onafhankelijk wanneer er geen sprake is van 'matching' tussen de groepen, bijvoorbeeld wanneer je gebruik maakt van randomisatie.

Standaardfout van groepsverschil

Stel dat we twee groepen met elkaar vergelijken: mannen en vrouwen en hun tijdsbesteding aan koken. Mannen en vrouwen zijn twee groepen, met allebei een ander populatiegemiddelde en een andere schatting daarvan. Je hebt dan ook twee standaardfouten. De standaardfout geeft namelijk aan hoe precies je schatting van de parameter is. Omdat we het verschil tussen mannen en vrouwen in de populatie willen weten, heeft ook dit verschil een standaardfout (want je schat het populatieverschil met je steekproefverschil).

De formule voor de geschatte standaardfout (van het verschil) is:
$\sqrt{(se_1)^2 + (se_2)^2}$ met $se = \frac{s}{\sqrt{n}}$. Hierbij is se1 de standaardfout van groep 1 (mannen) en se2 de standaardfout van groep 2 (vrouwen). Omdat je hier met twee groepen werkt, heb je ook twee keer een 'n', namelijk het aantal mannen en het aantal vrouwen. Dit geven we aan met n1 en n2. De 's' staat voor de standaard deviatie van de groep, en daar heb je er hier twee van. Deze geven we weer met s1 en s2. De formule voor de geschatte standaardfout kun je dan ook zo schrijven:
$\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$

Betrouwbaarheidsinterval van groepsverschil

Het betrouwbaarheidsinterval bestaat uit je puntschatting van het verschil ± de t-score maal de standaardfout. De formule ziet er voor het groepsverschil zo uit:

$(\bar{y}_2 - \bar{y}_1) \pm t\,(se)$ waarbij $se = \frac{s}{\sqrt{n}}$

Wanneer het betrouwbaarheidsinterval positieve waarden aangeeft, dan betekent dat dat μ_2 - μ_1 positief is, en dus dat μ_2 groter is dan μ_1. Wanneer het betrouwbaarheidsinterval negatieve waarden heeft, betekent het dan ook dat μ_2 kleiner is dan μ_1.

Significantie toets van groepsverschil

We kunnen testen of de twee groepen significant van elkaar verschillen. Normaal wordt toetsingsgrootheid t berekend door de geschatte parameter min de nulhypothese te doen en die te delen door de standaardfout van de schatting. De geschatte parameter is hier het verschil tussen de twee groepen (dus y2-y1). Je nulhypothese stelt dat er 'niets' aan de hand is en dat er geen verschil is tussen mannen en vrouwen: het verschil is 0. De standaardfout werd berekend met de formule: $\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$

De formule voor toetsingsgrootheid t ziet er dan zo uit:

$$t = \frac{(\bar{y}_2 - \bar{y}_1) - 0}{se}$$

Voorbeeld

We gaan met een voorbeeld de standaardfout berekenen, het betrouwbaarheidsinterval en vervolgens het verschil toetsen. Stel dat we het verschil tussen mannen en vrouwen in de tijd (aantal minuten per dag) die zij besteden aan huishoudelijk werk bekijken. Dit zijn de gegevens:

Geslacht	Steekproefgrootte	Gemiddelde	Standaarddeviatie
Mannen	1219	23	32
Vrouwen	733	37	16

Standaardfout:

De formule voor de standaardfout (se) is: $\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$

Als we dit invullen met onze gegevens: $(32)^2/1219 + (16)^2/733 = 1{,}09$

Betrouwbaarheidsinterval:

De formule voor de betrouwbaarheidsinterval is: $(\bar{y}_2 - \bar{y}_1) \pm t\,(se)$. Bij een alpha van .05 moeten we gebruik maken van de t-waarde ±1,96. We hebben berekend dat de standaardfout 1,09 is. De gemiddeldes van de mannen en vrouwen zijn gegeven. We vullen de formule in: (37 – 23) ± 1,96 (1,09) = (12 ; 16).

Significantie toets

De formule voor de t-toets is $t = \frac{(\bar{y}_2 - \bar{y}_1) - 0}{se}$.

We kunnen de formule weer gewoon invullen: $\bar{y}_2 - \bar{y}_1$ = 37 – 23 = 14. Dus 14 – 0 / 1,09 = 12,8. Wanneer we de t-waarde 12,8 opzoeken in de tabel van de t-distributie, dan zien we dat deze een p-waarde heeft van <.000. Het is dus een significant verschil.