Introduction to the practice of statistics, Moore, McGabe & Craig, 2012, 7e druk

Hoofdstuk 15 Nonparametrische toetsen

Normaliteit

Bij het trekken van conclusies uit experimenten maken we vaak gebruik van toetsen die de aanname doen dat er een normaalverdeling is in de populatie(s). Deze toetsen zijn redelijk robuust: schending van de aanname van normaliteit levert geen grote problemen op, vooral niet wanneer de steekproeven erg groot zijn. Wanneer de populatieverdeling echter duidelijk niet normaal verdeeld is en de steekproeven klein zijn, moeten er andere methoden gebruikt worden:

- Als duidelijke niet-normaliteit het gevolg is van uitbijters, dan moeten deze uitbijters verwijderd worden als ze niet tot de populatie behoren. Als ze wel tot de populatie behoren, kunnen er andere statistische methoden gebruikt worden die geen aanname van normaliteit doen.
- Soms kunnen data worden getransformeerd, zodat de verdeling van de data meer normaal wordt. Een voorbeeld hiervan is het gebruik van logaritmen.
- Soms kunnen data beter worden beschreven door middel van een andere standaardverdeling. De parameters van zo'n verdeling kunnen beschreven worden met behulp van speciale methoden.
- Bootstrapmethoden en permutatietoetsen zijn methoden die geen normaliteit vereisen.
- Ook andere non-parametrische methoden vereisen geen normaliteit. Deze methoden maken, in tegenstelling tot bootstrapmethoden en permutatietoetsen, geen gebruik van werkelijke waarden. Voorbeelden hiervan zijn rangtoetsen, die hieronder zullen worden besproken.

Rangtoetsen vereisen dat de populaties een continue verdeling hebben. Elke verdeling moet dus kunnen worden beschreven met een dichtheidscurve. De vorm van de curve maakt bij rangtoetsen niet uit. Toetsen die de aanname van normaliteit doen, maken gebruik van populatiegemiddelden of steekproefgemiddelden. Rangtoetsen maken gebruik van medianen.

De Wilcoxon rangsomtoets

De Wilcoxon rangsomtoets wordt gebruikt wanneer er in een experiment twee onafhankelijke steekproeven met elkaar worden vergeleken en de aanname van normaliteit geschonden is. De methode is als volgt:

- Rangschik alle waarnemingen van laag naar hoog.
- Nummer deze waarnemingen. De laagste waarneming krijgt rangnummer 1.
- Kies een van de steekproeven uit als eerste steekproef en tel de rangnummers bij elkaar op. Deze rangsom wordt W genoemd en is de Wilcoxon rangsomstatistiek. Hieronder staat een voorbeeld. De dikgedrukte waarden komen uit steekproef 1, de niet-dikgedrukte waarden komen uit steekproef 2. In dit geval wordt W dus 1 + 2 + 4 + 5 = 12.

Score op de test	**5.4**	**5.8**	6.1	**6.7**	**6.9**	7.5	8.1	8.4
Rangnummer	**1**	**2**	3	**4**	**5**	6	7	8

Bij sommige experimenten worden categorische variabelen omgezet in numerieke variabelen. Dit is onder andere het geval bij stellingen. Het volledig oneens zijn met de stelling is bijvoorbeeld 1 punt, het volledig eens zijn met de stelling is 5 punten. De t-toets behandelt deze variabelen als betekenisvolle getallen, terwijl dit in werkelijkheid niet het geval is. Onderzoekers gebruiken in zo'n geval vaak liever de rangsomtoets, omdat deze gebruik maakt van rangnummers in plaats van werkelijke waarden. Een ander voordeel is het feit dat uitbijters geen invloed hebben.

Om de Wilcoxon rangsomtoets uit te voeren, moeten naast de waarde van W ook nog het gemiddelde en de standaarddeviatie van W berekend worden. Het gemiddelde van W wordt berekend door:

$$m_W = \frac{n_1(N+1)}{2}$$

N is het totaal aantal observaties en n_1 is de steekproefgrootte van de eerste steekproef. De standaarddeviatie van W wordt berekend door:

$$s_W = \sqrt{\frac{n_1 n_2 (N+1)}{12}}$$

De aanname die hierbij wordt gedaan, is dat de twee populaties dezelfde continue verdeling hebben. Wanneer de waarde van W ver afligt van zijn gemiddelde (μ_W), dan zijn de verdelingen van de populaties niet identiek. Waarden van de ene populatieverdeling zijn dan systematisch hoger dan waarden van de andere populatieverdeling. Om te toetsen of het effect significant is, oftewel om de p-waarde te kunnen vergelijken met alpha α, kan gebruik gemaakt worden van speciale tabellen of van software. Een goede benadering is echter ook het gebruik van z-scores:

$$z = \frac{W - m_W}{s_W} = \frac{W - n_1(N+1)/2}{\sqrt{n_1 n_2 (N+1)/12}}$$

Voor een meer accurate z-score wordt vaak eerst nog een continuïteitscorrectie toegepast. In werkelijkheid is er namelijk in de steekproeven geen sprake van een continue verdeling, maar van discrete waarden (weergegeven als staafjes). Een score van 15 bijvoorbeeld bezet het interval van 14.5 tot 15.5 in de verdeling. De continuïteitscorrectie werkt als volgt:

- Als W groter is dan μ_W, dan halen we 0.5 af van W. Vervolgens vullen we de gecorrigeerde W in de z-formule in.
- Als W kleiner is dan μ_W, dan tellen we 0.5 op bij W. Vervolgens vullen we de gecorrigeerde W in de z-formule in.
- Wanneer we tweezijdig in plaats van eenzijdig willen toetsen, vermenigvuldigen we de gevonden p-waarde met 2. De continuïteitscorrectie hebben we dan van tevoren al uitgevoerd.

Hypothesen van de rangsomtoets

Omdat we bij de rangsomtoets medianen vergelijken in plaats van gemiddelden, worden de hypothesen als volgt:

H_0: mediaan$_1$ = mediaan$_2$

H_a: mediaan$_2$ ≠ mediaan$_2$ (tweezijdig) of bijvoorbeeld mediaan$_1$ > mediaan$_2$ (eenzijdig)

Dit geldt echter alleen als de populatieverdelingen dezelfde vorm hebben. In praktijk is dit vaak niet het geval. Daarom worden de hypothesen vaak geformuleerd in woorden:

H_0: De twee verdelingen zijn gelijk.

H_a: De waarden van de ene verdeling zijn systematisch hoger.

Knopen bij de rangsomtoets

Het kan zijn dat meerdere proefpersonen dezelfde score hebben behaald tijdens een experiment. Bij het toekennen van rangnummers wordt dan het gemiddelde genomen van de rangen die deze waarden bezetten. Hieronder staat een voorbeeld ter verduidelijking. In dit voorbeeld bezet score 6.1 zowel rangnummer 3 als rangnummer 4. Het gemiddelde van deze rangnummers wordt dan (3+4)/2 = 3.5.

Score op de test	**5.4**	**5.8**	6.1	**6.1**	**6.5**	7.5	8.1	8.4
Rangnummer	**1**	**2**	3.5	**3.5**	**5**	6	7	8

Bij knopen verandert de exacte verdeling van de Wilcoxon rangsom W. De standaarddeviatie van W (σ_W) moet worden aangepast. Statistische software is vereist wanneer je data knopen bevatten, omdat statistische software automatisch de nodige aanpassingen doet.

Rangsomtoets, t-toets en permutatietoets

De Wilcoxon rangsomtoets vervangt als het ware de t-toets voor twee onafhankelijke steekproeven wanneer er geen sprake is van een normaalverdeling in de populaties. Wanneer de steekproeven klein zijn en er geen sprake is van normaliteit, is de Wilcoxon rangsomtoets namelijk betrouwbaarder dan de t-toets. De t-toets gaat samen met een betrouwbaarheidsinterval. De rangsommethode daarentegen legt de nadruk echt op de toets, niet op het betrouwbaarheidsinterval. Een ander verschil is het feit dat het trekken van conclusies bij de rangsomtoets beperkt blijft tot simpele settings. Met de t-toets kunnen resultaten van meer complexe experimentele designs onderzocht worden.

Een rangsomtoets en permutatietoets zijn beide non-parametrische toetsen, maar ze verschillen op bepaalde aspecten. Het berekenen van de steekproevenverdeling onder de nulhypothese is hetzelfde voor beide toetsen, maar gaat gemakkelijker bij de rangsomtoets. Software geeft daarom alleen p-waarden voor rangsomtoetsen (en andere rangtoetsen) en niet voor permutatietoetsen. Een voordeel van permutatietoetsen ten opzichte van rangsomtoetsen is flexibiliteit. Permutatietoetsen bieden een brede keuze aan statistieken die gebruikt kunnen worden om twee steekproeven met elkaar te vergelijken. Ook zijn ze bijvoorbeeld te gebruiken bij multipele regressie.

De Wilcoxon rangtekentoets

De Wilcoxon rangtekentoets wordt gebruikt wanneer er sprake is van afhankelijke steekproeven en de aanname van normaliteit geschonden is. De methode is als volgt:

- Omdat het gaat om afhankelijke steekproeven zijn de waarnemingen gerangschikt in paren. Bepaal voor elk paar wat het absolute verschil is tussen de twee metingen (bijvoorbeeld tussen de voor- en nameting). Het gaat om absolute verschillen, dus het verschil is altijd positief. Rangschik deze absolute verschillen van laag naar hoog. Wanneer het verschil nul is, verwijder je deze uit de rangorde.
- Ken rangnummers toe aan de verschillen. Het kleinste verschil krijgt rangnummer 1.
- Maak in de rangorde duidelijk welke verschillen er oorspronkelijk positief waren en welke negatief.
- Tel de rangnummers van de oorspronkelijk positieve verschillen bij elkaar op. Deze rangsom wordt W^+ genoemd en is de Wilcoxon rangtekenstatistiek

Het gemiddelde van W^+ is:

$$m_{w^+} = \frac{n(n+1)}{4}$$

Hierbij gaan we ervan uit dat de verdeling van responsen niet te wijten is aan een verschillende behandeling binnen paren. In de formule staat n voor het aantal paren. Bij herhaalde metingen vormt elke proefpersoon als het ware een paar met zichzelf, dus in dat geval is n gelijk aan het aantal proefpersonen.

De standaarddeviatie van W^+ is:

$$s_{w^+} = \sqrt{\frac{n(n+1)(2n+1)}{24}}$$

Als de waarde van W^+ ver afligt van zijn gemiddelde (μ_{W+}), dan zijn er systematische verschillen binnen paren. De verdeling van de rangtekenstatistiek wanneer de nulhypothese waar is, is bij benadering normaal in het geval van een grote steekproef. We kunnen daarom z-scores gebruiken om de p-waarde te bepalen. Dit werkt hetzelfde als bij de Wilcoxon rangsomtoets, alleen gebruiken we nu W^+, μ_{W+} en σ_{W+}. Voor een meer accurate benadering van de z-score moet er weer eerst een continuïteitscorrectie worden toegepast. Ook dit gaat volgens dezelfde procedure als bij de Wilcoxon rangsomtoets.

Knopen bij de rangtekentoets

Bij de Wilcoxon rangtekentoets kan er sprake zijn van knopen *binnen* paren en van knopen *tussen* paren. Een knoop binnen een paar houdt in dat er twee keer hetzelfde gemeten wordt: het verschil is nul. Nul is niet negatief of positief, dus daarom worden alle nul-waarden uit de rangorde verwijderd. Waarnemingen waarbij het verschil nul is, zijn echter in het voordeel van de nulhypothese. Wanneer er veel knopen binnen paren zijn, zullen de resultaten dus vertekend raken en eerder richting de alternatieve

hypothese wijzen. Hierdoor verandert ook de verdeling en zo ook de standaarddeviatie van W^+ (σ_{W+}). Statistische software doet hiervoor de juiste aanpassingen. Als er knopen zijn tussen paren, dan houdt dit in dat twee of meer paren uitkomen op hetzelfde absolute verschil. De oplossing is dan om het gemiddelde te nemen van de rangen die ze bezetten, evenals bij de Wilcoxon rangsomtoets.

De Kruskal-Wallistoets

Wanneer we meer dan twee gemiddelden met elkaar willen vergelijken, maken we gebruik van enkelvoudige variantieanalyse (ANOVA). De aanname hierbij is dat de populatieverdelingen bij benadering normaal zijn en een gelijke verdeling hebben, oftewel gelijke standaarddeviaties. Als niet aan deze eisen voldaan wordt, kan de Kruskal-Wallistoets gebruikt worden. Deze toets vervangt dan de F-toets voor ANOVA. De aanname dat de steekproeven onafhankelijk en random getrokken zijn blijft hierbij hetzelfde. Verder wordt de aanname gedaan dat er in elke populatie een continue verdeling van responsen is. Hypothesen worden geformuleerd in woorden:

- H_0: De verdelingen van alle groepen zijn gelijk.
- H_a: De waarden van de sommige verdelingen zijn systematisch hoger

De Kruskal-Wallisstatistiek duiden we aan met H en is eigenlijk hetzelfde als SSG (de kwadratensom tussen groepen). H wordt op de volgende manier berekend:

$$H = \frac{12}{N(N+1)} \sum \frac{R_i^2}{n_i} - 3(N+1)$$

In deze formule staat n_i voor de steekproefgrootte van de i'de steekproef. N staat voor het totaal aantal observaties. Alle N observaties moeten worden gerangschikt, zodat de waarden van R_i kunnen worden bepaald. R_i is namelijk de rangsom voor de i'de steekproef. Het aantal populaties geven we aan met I. H heeft bij benadering een chi-kwadraatverdeling met I – 1 vrijheidsgraden. Aan de hand van de chi-kwadraatverdeling kunnen p-waarden worden bepaald. Als H groot is, dan wordt de nulhypothese verworpen.