Stem-, Spraak- en Taalpathologie
Vol. 15, 2007, No. 4, pp. 276-282
32.8310/04/1507-276

De toetsing van verschillen op achtergrondvariabelen

Toni Rietveld

Opleiding Taal- en Spraakpathologie, RU Nijmegen

In dit artikel wordt betoogd dat het uitvoeren van statistische toetsen op achtergrondvariabelen zoals leeftijd, IQ en SES van groepen proefpersonen die aan clinical trials deelnemen in de meeste gevallen niet zinvol is. Covariantie-analyse is in vele, maar niet in alle gevallen een beter alternatief.

Inleiding

Er bestaat een grote neiging om een ritueel uit te voeren zodra twee gemiddelden ter beschikking komen: er wordt een *t* test uitgevoerd (of het niet-parametrische equivalent daarvan). Van gemiddelden wil we weten of ze 'significant' van elkaar verschillen. In vele gevallen is dat zeer terecht: halen de proefpersonen ('deelnemers') gelijke scores als ze aan verschillende therapieën zijn onderworpen? Heel correct willen we ook nog vaststellen of de deelnemers verschil(l)(de)n op relevant geachte achtergrondvariabelen, zoals Sociaal Economische Status, IQ of leeftijd. Zo zal men bij onderzoek naar de effecten van leesmethoden geneigd zijn heel nauwkeurig te kijken of de deelnemers aan de verschillende leesmethoden wel gelijke leeftijden hadden. Zeer terecht. Een paar maanden verschil kan bij de leeftijd waarop lezen gewoonlijk wordt geleerd, veel uitmaken. Wat doet 'men' dan? Stel dat de kinderen van groep 3a methode A hebben gebruikt en die van groep 3b methode B. Er wordt een *t* test uitgevoerd op de leeftijden van de kinderen die in groep 3a en groep 3b zitten. Laten we aannemen dat het verschil in leeftijd tussen beide groepen 1.3 maanden is (groep 3b is 1.3 maanden ouder dan groep 3a) en dat het verschil significant is op het 0.05 niveau (2-zijdig). Wat betekent dat? Voordat we proberen te zeggen wat dat betekent moeten we ons eerst even voorstellen wat inductieve statistiek ook weer inhoudt. Met inductieve statistiek proberen we iets te vertellen over parameters van populaties op basis van steekproeven die at random uit die populaties zijn getrokken. We induceren op basis van beperkte informatie (uit de steekproeven) naar populaties die niet in hun geheel beschikbaar zijn. Als de effecten van leesmethode A en leesmethode B met elkaar vergeleken worden op basis van twee groepen – zeg groep 3a en groep

Correspondentieadres: Toni Rietveld, Faculteit Letteren afd. Taalwetenschap, Erasmusplein 1, 6525 HT Nijmegen. E-mail: A.Rietveld@let.ru.nl

3b – dan kunnen we zeggen dat de ene methode beter is dan de andere methode als de steekproefgegevens, bijvoorbeeld gemiddelden van leesscores, daartoe aanleiding geven ('significant van elkaar verschillen') EN als de beide groepen niet op relevante variabelen van elkaar verschillen.

In die laatste eis zit een probleem. Wat is een relevant verschil, en hoe wordt dat getoetst, of moeten we dat wel toetsen? Even terug naar de groepen 3a en 3b. Het significante verschil dat is gevonden betekent, in de strikte zin van de taak van de inductieve statistiek, dat groepen met het label '3b' ouder zijn dan groepen met het label '3a'; een bizarre en irrelevante conclusie. Dezelfde potentieel eigenaardige conclusies vinden we bij de zeer vaak voorkomende testen op relevante variabelen tussen controle- en experimentele groepen in Randomized Clinical Trials (cf. Senn, 1994). We gaan nog even door met de variabele leeftijd, in het kader van RCT's. Is het interessant te weten dat mensen die het medicijn (therapie) niet toegediend krijgen ouder/jonger zijn dan mensen die dat medicijn wel toegediend kregen? We komen hier bij de mogelijke kern van de onbedoelde zin van het significantieritueel. Stel dat we wel een verschil in leeftijd vinden tussen de experimentele en de controlegroep? De inductiestap is zinloos, hebben we inmiddels gezien. Immers, willen we echt zeggen dat mensen die in een controlegroep zijn ondergebracht - wanneer, hoe en waarom – ouder of jonger zijn dan mensen die in een experimentele groep zitten? Zouden we dat echt willen weten, dan zouden we de onderzoeksvraag moeten herformuleren: zijn mensen die bereid zijn om aan een therapie deel te nemen ouder/jonger dan de gemiddelde andere mens (de controlemens)? Een dergelijke vraag is een moeilijke en vaak ook niet zo'n interessante. Hoogstens is die vraag van belang voor degene die de inclusie van proefpersonen voor zijn rekening neemt en is de statistische test een toets of de randomisatie geslaagd is. Altman & Doré (1990) schrijven dan ook in de Lancet: "The similarity of baseline characteristics (en dus ook van relevante achtergrondvariabelen) must be established, but not by hypothesis tests".

We moeten hier nog een complicerend aspect aan de orde stellen, nl. het verschil tussen achtergrondvariabelen (relevante variabelen zoals leeftijd en SES) en baselinemetingen (metingen op de afhankelijke variabele – ook wel uitkomstvariabele genoemd - voordat de behandeling wordt gestart) en ons zelfs de vraag stellen of deze twee variabelen wel zo verschillend zijn. Eén verschil is onbetwistbaar: een achtergrondvariabele is geen afhankelijke variabele, en baselinemetingen zijn wel metingen op de afhankelijke variabele, uitgevoerd op een tijdstip voorafgaand aan de therapie. Echter, buiten dit op eerste gezicht fundamentele verschil, leidt de detectie van verschillen tussen twee groepen deelnemers (al-dan-niet statistisch getoetst) over het algemeen wel tot bezorgdheid. Immers, als we in groep B een hogere score op een relevante achtergrondvariabele vaststellen dan in groep A, kan dat leiden tot een hogere score op de afhankelijke variabele. Een hogere score op de baseline kan ook effect hebben op de afhankelijke variabele ná therapie. Een bekend verschijnsel is dat mensen

die vóór de therapie al een hoge score vertonen op de afhankelijke variabele, niet zoveel verbetering vertonen als mensen die een lage score hadden vóór de therapie. Dit effect is ook vaak zichtbaar in onderzoek naar de effecten van taalmethoden (c.f. Neri, 2007). Ofschoon zowel verschillen tussen groepen bij baselinemetingen als op achtergrondvariabelen tot een gelijke mate van bezorgdheid leiden, zullen we ons in deze bijdrage beperken tot achtergrondvariabelen.

Wat kunnen of moeten we doen als er verschillen zijn tussen groepen op relevante variabelen?

Wat kunnen we zeggen als we toch aan het significantieritueel meedoen (waarvoor ik overigens niet pleit)? De volgende mogelijkheden kunnen worden onderscheiden:

1. Er is een significant verschil op een relevant geachte variabele tussen twee groepen die bestaan uit at random toegewezen deelnemers. Indien het verschil significant is bevonden, is de random toekenning mogelijk niet zo random geweest als men had gewenst. De vraag is of het significante verschil relevant is. Twee maanden bij aanvankelijk leesonderwijs is iets anders dan twee maanden bij een afasietest van mensen rond de 55 jaar.

2. Er is geen significant verschil gevonden op een relevant geachte variabele tussen twee groepen die samengesteld zijn uit at random toegewezen deelnemers. Dan zijn er ook weer een aantal mogelijkheden: a) de test had onvoldoende vermogen (teveel variatie binnen de groepen, te klein effect), waardoor een verschil tussen de populaties niet kon worden gedetecteerd; b) de twee groepen verschillen inderdaad niet op de relevante variabelen.

Deze twee mogelijkheden en submogelijkheden zouden de onderzoeker niet enthousiast moeten maken voor significantietests voor achtergrondvariabelen. De indruk bestaat verder dat een *t* test als een soort maat voor de grootte van een effect wordt gebruikt, in de zin dat als *t* significant is, het verschil tussen de twee betrokken gemiddelden wel groot zal zijn, en als hij niet significant is, het effect wel klein zal zijn en niet relevant. Dat een dergelijke interpretatie niet correct is, leert ons de formule van de *t* test (t = (gemiddelde van A – gemiddelde vanB)/[$\sqrt{}$(variantie van A+variantie van B)]) en wordt hieronder aan de hand van een heel eenvoudig getallenvoorbeeld met twee datasets geillustreerd. De datasets stellen scores op een willekeurige achtergrondvariabele voor (b.v. leeftijd, of schoolopleiding). T-tests voor onafhankelijke steekproeven zijn uitgevoerd, en de toetsing was tweezijdig.

Tabel 1. Twee gefingeerde datasets, 1 en 2, met scores op een achtergrondvariabele, ieder met twee groepen, A en B; n = 5 in elke groep; significantieniveau = 0.05 (tweezijdig). Gem. = Gemiddelde.

Set 1	
Groep A	**Groep B**
6	6
5	7
6	7
5	6
6	7
Gem. = 5,6	Gem. = 6,6
t_8 = -2,887, p = ,020	

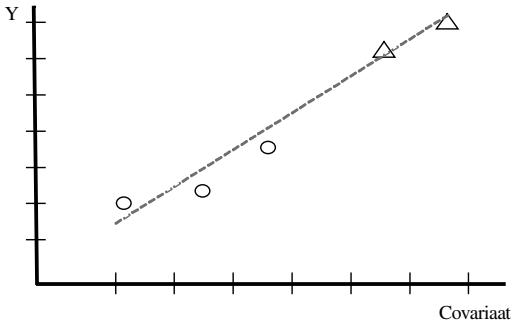
Set 2	
Groep A	**Groep B**
1	6
5	7
6	7
5	6
6	7
Gem. = 4,6	Gem. = 6,6
t_8 = -2,085, p = ,071	

We zien dat het verschil van 1 schaalpunt in Set 1 significant is, terwijl een groter verschil, nl. van 2 schaalpunten in Set 2, niet significant is op het 5% niveau (Levene's test voor homogeniteit der varianties was niet significant voor beide datasets). Wat nu? De oorzaak van het onverwachte verschil in significantie is natuurlijk het feit dat de variantie in groep A van Set 2 groter lijkt te zijn dan in Set 1 (let op de score 1 in Groep A van Set 2). Aldus komt het verschil tussen de gemiddelden in Set 2 niet boven de ruis uit, en in Set 1 wel. Betekent dit nu dat we voor Set 1 niet kunnen aannemen dat de deelnemers gelijk zijn op een relevant geachte achtergrondvariabele, maar voor Set 2 wel? Nee natuurlijk; een blik op de data maakt duidelijk dat 100% van de data in Groep A van Set 2 een lagere score heeft op de achtergrondvariabele dan groep B. Op basis van dit eenvoudige voorbeeld zien we dat het toetsritueel op deze wijze ambigue resultaten kan opleveren.

Wat moeten we dan doen? Iedereen zal het er over eens zijn dat bij het ene onderzoek een klein verschil relevant is, en bij het andere niet. De beoordeling van de relevantie van gevonden verschillen ligt bij de ervaren onderzoeker. Wat we zeker moeten doen is een transparant overzicht geven van de scores op de gemeten achtergrondvariabele(n). De welbekende Box-plots zijn hiervoor zeer geschikt; het gebruik ervan wordt in menig statistiek- en methodologieboek zeer aangeraden (zie ook Lang & Secic, 1997).

Kan men dan helemaal niets doen met verschillen tussen proefpersonen op variabelen die relevant geacht worden voor de scores op de afhankelijke variabele? Jawel, die mogelijkheid is er, in de vorm van covariantie-analyse. In deze analyse worden de verschillen tussen de proefpersonen op een relevant geachte achtergrondvariabele als het ware 'uit de analyse gehaald'. Deze benadering veronderstelt wel dat er een lineaire relatie bestaat tussen de covariaat (een achtergrondvariabele zoals bv. leeftijd of IQ) en de afhankelijke variabele. We moeten dus overschakelen van de benadering van 'toetsen van verschillen' (de '*t* test') naar het vaststellen van correlatie ('Pearson's *r*'). Als er een correlatie is tussen een covariaat en de afhankelijke variabele, dan weten we dat die covariaat gerelateerd is (niet noodzakelijkerwijze via een cau-

saal verband!) aan de afhankelijke variabele. Zoals bekend moeten bij correlatiecoëfficiënten twee aspecten in de beschouwing worden betrokken: a) de grootte van de correlatie (r = ,80 betekent een sterker lineair verband tussen twee variabelen dan r = ,60; lineair betekent dat het verband tussen de variabelen x en y geschreven kan worden als $y = a_0 + a_1X$) en b) de significantie van de correlatie. Een correlatie van ,60 is bij 11 paren waarnemingen niet significant op het 5%-niveau (2-zijdig), bij 12 paren wel. Het lijkt erop dat we zo bij hetzelfde dilemma zijn aangekomen dat hierboven is besproken. Significant, wat nu? Niet-significant: OK? Het antwoord is redelijk eenvoudig, althans op het eerste gezicht. We kunnen gewoon de covariaat, al-dan-niet significant gerelateerd aan de afhankelijke variabele, opnemen in onze covariantie-analyse. We krijgen dan een zuiverder beeld van het effect van onze factor(en), aangezien het effect van de covariaat op de uitkomstvariabele is weggezuiverd. Over het algemeen leidt een covariantie-analyse ('ANCOVA': ANalysis of COVAriance) tot een groter vermogen om eventuele verschillen tussen gemiddelden in de populaties op te sporen, omdat uit de errorvariantie de component wordt weggenomen die samenhangt met de covariaat (zie Rietveld & van Hout, 2005). Covariantie-analyse is helaas ook weer niet geheel zonder complicaties. Zo moeten de regressielijnen die de covariaat en de afhankelijke variabele in de afzonderlijke groepen met elkaar verbinden parallel lopen (in technische termen: de β_j (hellingen, 'slopes') moeten gelijk zijn. Gelukkig - zie Harwell (2003) – is covariantie-analyse behoorlijk robuust tegen schendingen van allerlei statistische assumpties, vooral als a) we te maken hebben met gelijke aantallen in de betrokken onderzoeksgroepen ('balanced designs') en b) als het design 'randomized' is, wat erop neerkomt dat we aselect deelnemers aan de behandelingen hebben kunnen toekennen. Wel vervelend is de situatie waarbij er zowel een sterke correlatie is tussen de covariaat en de afhankelijke variabele en de betrokken groepen zeer verschillende scores hebben op de afhankelijke variabele. Een voorbeeld: als degenen die therapie B volgen bijna allemaal een hogere score hebben op de covariaat leeftijd dan degenen die therapie A volgen. Als dat zo is, kan het hele therapie-effect door de covariaat worden verklaard, zie de volgende figuur.



Figuur 1. Scores van deelnemers aan twee therapieën, A (cirkels) en B (driehoekjes) op de afhankelijke variabele Y en de covariaat X

In dit overdreven voorbeeld zien we dat de scores van de deelnemers in de twee groepen zowel op de afhankelijke variabele als op de covariaat sterk van elkaar verschillen; hier is zelfs geen overlap in de scores te zien. Verder is er van een sterk lineair verband tussen X en Y sprake. Deze situatie zegt maar een ding: vertel mij wat Uw score op de covariaat is, en ik weet Uw score op de afhankelijke variable. We hoeven dan niet te weten of therapie A of B gevolgd is. Een covariantie-analyse zou voor dit type data dan ook geen significant effect hebben opgeleverd, een 'gewone' variantie-analyse mogelijk wel (maar dan mogelijk ten onrechte).

Conclusie

In de taal- en spraakpathologie is vaak geen sprake van random toewijzing van proefpersonen aan verschillende behandelingen; zeker bij volwassenen, maar ook bij kinderen is dat om allerlei praktische redenen vaak heel moeilijk. Om die reden wordt vaak statistisch getoetst of groepen vergelijkbaar zijn op relevant geachte achtergrondvariabelen. In deze bijdrage hebben wij willen laten zien dat het toetsen van verschillen tussen scores op achtergrondvariabelen zoals IQ, leeftijd en SES, een zinloze activiteit is, en in strijd met de bedoeling van de inductieve statistiek. Het is echter wel van groot belang om na te gaan of de deelnemers aan verschillende behandelingen (groepen) verschillen op relevant geachte variabelen. Daarvoor moeten alle middelen van de beschrijvende statistiek worden gebruikt. Wanneer er een lineaire correlatie is – significant of niet – tussen een achtergrondvariabele (covariaat) en de afhankelijke variabele, is vaak de toepassing van covariantie-analyse geboden. ANCOVA is echter geen panacé voor verschillen tussen groepen op relevante achtergrondvariabelen. De onderzoeker zal altijd moeten aantonen dan wel plausibel maken dat verschillen tussen groepen proefpersonen geen effect zullen hebben gehad op eventueel gevonden verschillen op de uitkomstvariabele.

Summary

This article shows that carrying out statistical tests on background variables like age, IQ and SES of subjects participating in clinical trials is not warranted in most cases. Analysis of covariance is a good alternative in many, but not all cases

Referenties

Altman, D.G. & Doré, C.J. (1990). Randomisation and baseline comparisons in clinical trials. *The Lancet,* 335, 149-153.

Harwell, M. (2003). Summarizing Monte Carlo Results in Methodological Research: The Single-Factor, Fixed-Effects ANCOVA Case. *Journal of Educational and Behavioral Statistics*, 28, 45-70.

Lang, Th., A. & Secic, M. (1997). *How to Report Statistics in Medicine*. Philadephia: American College of Physicians

Rietveld, Toni & van Hout, Roeland. (2005). *Statistics in Language Research: Analysis of Variance*. Berlin: Mouton de Gruyter.

Senn, S. (1994). Testing for baseline balance in clinical trials. *Statistics in Medicine*, 13, 1715-1726.