Power-analyse, of de bepaling van de steekproefgrootte*

Bert Garssen[1,2] en Hellen Hornsveld[1,3]

Inleiding

Bij het opzetten van kwantitatief onderzoek kan en moet van tevoren veel overdacht en gepland worden. Eén onderdeel daarvan is de steekproefgrootte. Vanuit het oogpunt van de representativiteit van de gegevens kan een steekproef niet groot genoeg zijn. Hoe groter de steekproef, hoe groter de kans dat de gevonden gegevens representatief zijn voor de populatie waarop het onderzoek betrekking heeft, of, met andere woorden, hoe kleiner de kans op fouten als gevolg van de steekproeftrekking (sampling error). Vaak echter wordt het aantal proefpersonen in een onderzoek beperkt door praktische factoren. Zelfs als deze beperkingen geen groot aantal toelaten, zal men graag willen weten of dit aantal nog nét genoeg is, en zo niet, hoeveel proefpersonen eigenlijk nodig zouden zijn. Veel psychologische studies leiden aan het manco van een te gering aantal proefpersonen, waardoor alleen grote effecten zijn aan te tonen. Treden die effecten op dan is er geen probleem, maar blijven ze achterwege dan laat het onderzoek geen enkele conclusie toe; het is daarmee tot een nutteloze inspanning verklaard.

Al enige tijd geleden zijn methoden beschreven om de gewenste steekproefgrootte te bepalen, maar die zijn klaarblijkelijk nog onbekend of onbemind bij onderzoekers. Er lijkt sprake te zijn van enige huiver om zich te verdiepen in een onderwerp

[1] Medische Psychologie, Academisch Medisch Centrum, Meibergdreef 15, 1105 AZ Amsterdam.
[2] Hellen Dowling Instituut, Rotterdam.
[3] Slotervaartziekenhuis, Amsterdam.
Correspondentie te richten aan H. Hornsveld.
* Met dank aan dr. P. Koele, vakgroep Psychologische Methodenleer, UvA, voor zijn deskundig commentaar op een eerdere versie van dit artikel.

waarvan men veronderstelt dat het te ingewikkeld is. Die houding zou vroeger begrijpelijk zijn, maar sinds het verschijnen van het boek over power-analyse van Cohen in 1969 is die houding eigenlijk niet meer terecht. Met enig algemeen inzicht is aan de hand van de tabellen van Cohen gemakkelijk de steekproefgrootte te bepalen. Dit artikel is een poging eventuele huiver voor het onderwerp weg te nemen, en praktische aanwijzingen te geven hoe men te werk moet gaan. We hopen daarmee het toepassen van de beschreven methoden te stimuleren.

Een voorbeeld

Een onderzoeker houdt zich bezig met de behandeling van agorafobische patiënten. Op dit gebied is nog steeds een discussie gaande over de vraag of bij patiënten met frequente paniekaanvallen een medicamenteuze behandeling nuttig of noodzakelijk is, als aanvulling op een cognitief-gedragstherapeutische behandeling. De onderzoeker twijfelt aan het nut van de toevoeging. Er zijn op dit gebied al vele studies verricht, maar opvallend weinig waarin gebruik is gemaakt van een behoorlijke registratie van de paniekaanvallen. De onderzoeker besluit in dit manco te voorzien en vergelijkt een behandeling waarin alleen cognitief-gedragstherapeutische methoden worden toegepast, met een combinatietherapie. Er worden 50 patiënten geselecteerd die random aan één van de twee behandelingsvormen worden toegewezen, dat wil zeggen 25 patiënten per conditie. Na de behandeling wordt de paniekfrequentie over een periode van vijf weken geregistreerd, en deze paniekfrequentie wordt als uitkomstmaat gekozen. Als nulhypothese geldt, dat de uitkomstmaat geen verschil zal vertonen tussen de twee groepen. Onder de alternatieve hypothese wordt op grond van eerder verricht onderzoek een hooguit middelmatig verschil verwacht ten gunste van de combinatietherapie. Gekozen wordt voor een t-toets met $\alpha_{\text{eenzijdig}} = 0{,}05$. Nadat de gegevens zijn verzameld, blijkt er een licht effect te bestaan in het voordeel van de combinatietherapie, maar het verschil is niet significant; de nulhypothese mag niet worden verworpen. Kan de onderzoeker nu tevreden constateren dat zijn scepsis is bevestigd, of was de toets niet gevoelig genoeg om eventueel bestaande verschillen te kunnen aantonen? Het laatste is inderdaad het geval. In het navolgende stuk wordt uitgelegd waarom dat zo is, en welke steekproefgrootte wél adequaat zou zijn geweest.

Twee soorten fouten bij statistische toetsing

Statistische toetsing geeft geen absolute zekerheid, het schept slechts de mogelijkheid aan te geven hoe groot de waarschijnlijkheid is dat een conclusie juist is. De conclusie die getrokken wordt kan in twee gevallen fout zijn (zie tabel 1):

Gedragstherapie, jrg. 25, nr. 2, pp. 107-121, juni 1992

Tabel 1. Twee soorten fouten bij statistische toetsing.

	De nulhypothese (er is geen verschil) is in feite:	
	juist	onjuist
We concluderen tot:		
wél een verschil	α	1 – ß
geen verschil		ß

1. Er is in werkelijkheid geen verschil, doch we concluderen ten onrechte tot wél een verschil. We noemen dit een fout van de eerste soort, of een α-fout.
2. Er kan in werkelijkheid wél een verschil zijn, maar dit verschil wordt niet gevonden en dus concluderen we ten onrechte tot geen verschil. Dit heet een fout van de tweede soort, of een ß-fout. De kans dat men deze fout maakt is ß. Stel dat ß = 0,10, dan zal men – gegeven dat er in werkelijkheid wél een verschil is – gemiddeld in 10% van de testen ten onrechte de nulhypothese accepteren. In gemiddeld 90% van de gevallen ((1-ß) x100%) zal men dan terecht de nulhypothese verwerpen. De waarde 1-ß is dus de kans dat men terecht tot een verschil besluit en dit wordt de *power* of het onderscheidingsvermogen genoemd.

De meeste statistische toetsen zijn gebaseerd op het berekenen van kansen onder de aanname dat de nulhypothese (er is geen verschil) juist is. Door het vaststellen van de α wordt de kans op het maken van een fout van de eerste soort bepaald. Wanneer bijvoorbeeld wordt gekozen voor een α van 0,05, accepteren we dat van alle toetsen waarbij de nulhypothese in feite juist is, in gemiddeld 1 op de 20 gevallen de nulhypothse ten onrechte wordt verworpen. Het bepalen van de ß is gebaseerd op het berekenen van kansen onder de aanname dat de alternatieve hypothese juist is. Hiertoe moet de alternatieve hypothese worden gespecificeerd. Dit geschiedt met behulp van effectgroottes. De waarden α en ß hangen sterk samen. Als men α kleiner maakt – door bijvoorbeeld 0,01 in plaats van 0,05 als grens te kiezen – verkleint men weliswaar de kans dat tot een verschil zal worden besloten dat er in feite niet is, maar tevens vergroot men de kans dat bestaande verschillen niet zullen worden gevonden.

We hebben ons leren wapenen tegen de α-fout, doch nauwelijks aandacht besteed aan de ß-fout. Er zijn overigens wel redenen aan te geven om ons in de eerste plaats zorgen te maken

over de α-fout, want deze kan tot grotere problemen leiden. Als er tot een verschil wordt geconcludeerd dat er in feite niet is, kan bijvoorbeeld een behandeling worden aangeraden die in werkelijkheid ineffectief is. Behalve klinische consequenties heeft een fout van de eerste soort ook implicaties voor later onderzoek. De kans is groot dat onnodig geïnvesteerd wordt in onderzoek dat voortborduurt op resultaten en hypothesen die ten onrechte zijn geaccepteerd.

Aan de andere kant zijn er ook situaties waarbij het wellicht belangrijker is om een fout van de tweede soort te vermijden. Dit geldt bijvoorbeeld bij exploratief onderzoek. In het bovenstaande voorbeeld zou de onderzoeker bijvoorbeeld kunnen nagaan of een aantal demografische gegevens en vragenlijstscores samenhangen met de therapie-uitkomst (reductie in paniekaanvallen). De onderzoeker kan dan kiezen voor een wat lagere betrouwbaarheid (grotere α) om de kans op het vinden van verschillen te vergroten. Ieder dusdanig gevonden verschil kan dan als basis dienen voor een meer gedetailleerde studie, waarbij sommige verschillen waarschijnlijk zullen verdwijnen en andere blijven.

Het vermijden van een fout van de tweede soort is ook belangrijk bij replicatie-onderzoek dat bedoeld is om na te gaan of een eerder gevonden verschil betrouwbaar optreedt. Indien het vorige onderzoek methodologisch goed in elkaar stak, zijn er al aanwijzingen gevonden voor een verschil. Het is de onderzoeker dan in de eerste plaats aan te rekenen als door een verkeerde keuze nu ten onrechte tot géén verschil wordt geconcludeerd.

Er is één speciaal geval waarin het zeer belangrijk is een fout van de tweede soort te vermijden, namelijk als de nulhypothese (= er is geen verschil) tevens als onderzoekshypothese gehanteerd wordt. Dit is onder andere het geval in het bovenstaande voorbeeld waarin de onderzoeker veronderstelt dat er geen verschil is tussen twee vormen van behandeling. Het niet vinden van een significant verschil wordt in zo'n geval bijna altijd geïnterpreteerd als een bevestiging van de nulhypothese. Om deze conclusie te rechtvaardigen zou de power bijzonder hoog moeten zijn, bijvoorbeeld 0,95. We accepteren dan dat in 5% van de gevallen de nulhypothese ten onrechte wordt aangenomen.

Het is duidelijk dat per onderzoek een afweging gemaakt zal moeten worden ten aanzien van de soort fout die men het meest riskant vindt.

Gedragstherapie, jrg. 25, nr. 2, pp. 107-121, juni 1992

Steekproefomvang, spreiding en afstand tussen gemiddelden

Wat kan men doen, behalve het vergroten van α, om de kans op een ß-fout te verkleinen? Drie mogelijkheden doen zich voor: 1) kies een zo groot mogelijke steekproef; 2) tracht de spreiding binnen de steekproeven zo klein mogelijk te maken, bijvoorbeeld door storingsbronnen zoveel mogelijk uit te schakelen (andere adviezen volgen later); en 3) maak de afstand tussen de gemiddelden zo groot mogelijk door een zo krachtig mogelijke manipulatie toe te passen. De spreiding en de afstand tussen de gemiddelden worden samengevat onder de noemer effectgrootte (d). Er zijn dus vier grootheden die elkaar beïnvloeden: α, ß, N en d. Leggen we er drie vast, dan is de vierde tevens bepaald.

Effectgrootte

Wat effectgrootte in algemene termen voorstelt volgt al uit de naam: het gaat om de grootte van het verschil tussen de gemiddelden van twee steekproeven (M1 en M2). Daarbij wordt rekening gehouden met de spreiding. De effectgrootte (d) is een maat voor de afstand tussen twee steekproefgemiddelden in verhouding tot de spreiding van de scores binnen de steekproeven. In formule:

$$d = \frac{M1-M2}{sd_{gez}} \quad (1)$$

sd_{gez} = de gezamenlijke spreiding van de twee steekproeven.[1]

Uit de formule volgt dat bij een effectgroote van d = 1 de gemiddelden precies één standaarddeviatie van elkaar verwijderd liggen.

De te verwachten effectgrootte kan worden afgeleid uit literatuurgegevens of uit de resultaten van een vooronderzoek. Maar

[1] De formule voor de gezamenlijke spreiding luidt:

$$sd_{gez} = \sqrt{\frac{(N_1 - 1)\, sd_1^2 + (N_2 - 1)\, sd_2^2}{N_1 + N_2 - 2}}$$

Wanneer de steekproefgroottes N1 en N2 aan elkaar gelijk worden, gaat deze formule over in:

$$sd_{gez} = \sqrt{\frac{sd_1^2 + sd_2^2}{2}}$$

men kan ook uitgaan van een minimaal gewenste effectgrootte, door bijvoorbeeld van tevoren te stellen niet geïnteresseerd te zijn in kleine verschillen ook al treden deze systematisch op, bijvoorbeeld omdat ze als niet klinisch relevant zijn te beschouwen. Maar, wat is een klein verschil, en hoe is dat te vertalen in een bepaalde effectgrootte?

Cohen (1977) spreekt bij d = 0,20 van een gering effect, bij d = 0,50 van een middelmatig effect en bij d = 0,80 van een groot effect. Hierna wordt op drie manieren getracht een indruk te geven wat deze drie effectgroottes betekenen. Men kan allereerst de verdelingen van twee populaties voorstellen met een effectgrootte van respectievelijk d = 0,20, 0,50 en 0,80, en dan bekijken hoe groot de overlap is van de twee verdelingen. Bij een effectgrootte van 0,20 overlappen de twee populaties elkaar bijna geheel, namelijk voor 85,3%. Bij een d = 0,50 overlappen ze nog grotendeels, namelijk voor 67,0% en bij d = 0,80 voor ongeveer de helft (53,6%).

Een tweede manier om de effectgroottes voor te stellen, is ze te vertalen in verklaarde variantie. Wanneer de gemiddelden van twee populaties ver uiteen liggen en hun spreiding klein is, wordt de variantie in de gegevens in sterke mate bepaald door het groepslidmaatschap (= of een element afkomstig is uit de ene of de andere populatie). Bij een effectgrootte van 0,2 wordt slechts 1% van de variantie verklaard door het groepslidmaatschap, bij 0,5 is dat 6% en bij een d = 0,8 is de verklaarde variantie 14%. Overigens gelden deze gegevens, alsmede die voor de mate van overlap, alleen als aangenomen mag worden dat de twee populaties ongeveer normaal verdeeld zijn en een ongeveer gelijke spreiding vertonen.

Een derde manier om zich de betekenis van de drie effectgroottes voor te stellen is zich op concrete voorbeelden te richten. Cohen (1977) geeft de volgende voorbeelden: Hoewel een effectgrootte van 0,2 klein lijkt, wordt het gevonden bij het verschil in lengte tussen 15- en 16-jarige meisjes. Een effectgrootte van 0,5 werd aangetoond voor het IQ-verschil tussen administratieve en gedeeltelijk geschoolde werkers ('clerical and semiskilled workers'). Ten slotte geldt een effectgrootte van 0,8 voor het IQ-verschil tussen gepromoveerden en eerstejaarsstudenten.

In klinisch onderzoek zal men vaak zwaardere normen hanteren. Als voorbeeld: Bij de behandeling van obesitaspatiënten is een therapie-effect niet erg groot te noemen als het gewicht voor de behandeling 125 kg (sd = 8 kg) en na de behandeling 117 kg (sd = 8 kg) is, ook al is de effectgrootte + 1,0. In het algemeen kan men beter afgaan op eerder verricht onderzoek en de

Gedragstherapie, jrg. 25, nr. 2, pp. 107-121, juni 1992

eigen eisen die men stelt aan de klinische relevantie, dan zich te richten op de algemene en wat lage normen van Cohen.

Power-analyse

Power-analyse betreft het berekenen van de vereiste steekproefgrootte, uitgaande van de gewenste power, of het berekenen van de power als de steekproefgrootte is gegeven. Er bestaat een onderlinge samenhang tussen de vier-eenheid α, ß, d, en N. Vatten we het bovenstaande samen, dan neemt de vereiste steekproefgrootte toe naarmate we een kleinere α en ß kiezen, en naarmate we een kleinere effectgrootte willen kunnen aantonen.

De keuze van de α en ß maakt men onafhankelijk van de soort statistische toets. Wat de α betreft, bestaat de conventie om 0,05 te gebruiken. Bij het toepassen van de power-analyse wordt uitgegaan van kennis of verwachtingen over de richting van het effect, en dan ligt het voor de hand een eenzijdige a priori kans van 0,05 ($\alpha_1 = 0,05$) te kiezen, overeenkomend met een tweezijdige kans van 0,10 ($\alpha_2 = 0,10$). Cohen (1977) stelt voor de ß 4x zo groot te nemen als de α. De conventie volgend, wordt ß = 0,20, en de power = 0,80. Bij deze waarde voor ß, wordt genoegen genomen met de kans dat bij gemiddeld één op de vijf toetsen ten onrechte tot geen verschil wordt besloten, dus een werkelijk bestaand verschil over het hoofd wordt gezien. Wil men voorzichtiger zijn, dan kan men een α van 0,01 kiezen, en een power van 0,96. Men moet er dan wel op voorbereid zijn dat men grote aantallen proefpersonen nodig heeft.

Er is wat voor te zeggen om de eisen ten aanzien van de beide soorten fouten op deze manier aan elkaar te koppelen. Het heeft meestal geen zin de zekerheid alleen te verhogen met betrekking tot de kans op één soort fout. Anderzijds is in de inleiding reeds betoogd, dat zich ook situaties kunnen voordoen waarin het vermijden van één soort fout meer aandacht dan gewoonlijk zou moeten krijgen. Bij een replicatie-onderzoek is bijvoorbeeld de keuze voor een $\alpha = 0,10$ en een power = 0,90 goed te verdedigen.

Het bepalen van de steekproefgrootte of de power kan in de praktijk geschieden aan de hand van de tabellen uit Cohen's boek, of aan de hand van formules, die onder andere zijn te vinden in een artikel van Singer, Lovie en Lovie (1986). De tabellen van Cohen zijn onderverdeeld in power-tabellen (de berekening van de power op grond van N, α en het gevonden of verwachte effect) en steekproefgrootte-tabellen (de berekening van N op grond van de gewenste α, ß en de effectgrootte). Voor waarden

van α en ß die niet in Cohen te vinden zijn kan men de formules van Singer et al. (1986) gebruiken.

Cohen geeft tabellen voor t-toetsen (voor gepaarde en ongepaarde gegevens) correlatiecoëfficiënten, verschillen tussen correlatiecoëfficiënten, tekentoetsen, verschillen tussen proporties, chi-kwadraat-toetsen en F-toetsen bij variantie-analyse, covariantie-analyse en multipele regressie.

In de editie uit 1969 van Cohen komen fouten voor die in de latere editie (1977) zijn verwijderd. Volgens Hoogstraten en Koele (1981) komt in deze herziene versie nog steeds een fout voor in het gedeelte over de variantie-analyse. Deze fout resulteert in een geringe overschatting van de steekproefgrootte, en een onderschatting van het onderscheidingsvermogen. Men raadplege Hoogstraten en Koele (1981) voor een meer precieze berekening van de power bij variantie-analyse.

Als toelichting wordt gestart met de meest klassieke situatie van twee steekproeven uit normaal verdeelde populaties; de situatie waarin een t-toets voor niet-gepaarde gegevens mag worden toegepast.

T-toets voor onafhankelijke steekproeven

In tabel 2 is het aantal benodigde proefpersonen weergegeven bij drie verschillende niveaus van de power (0,80, 0,90 en 0,95), drie verschillende niveaus van α (eenzijdig), namelijk $\alpha = 0{,}01$, $\alpha = 0{,}05$ en $\alpha = 0{,}10$ en bij drie verschillende effectgroottes. Het is belangrijk te weten dat het aantal proefpersonen per groep (conditie) is berekend. Aangezien bij de t-toets twee groepen worden vergeleken, is het totale aantal benodigde proefpersonen tweemaal zo hoog als in tabel 2 is weergegeven.

Uit de voorbeelden in tabel 2 blijkt dat in het gunstigste geval ($\alpha_1 = 0{,}10$, ß = 0,80 en een groot verwacht effect) toch nog 14 proefpersonen per groep nodig zijn. Bij het meer realistische uitgangspunt van een middelmatige effectgrootte en $\alpha_1 = 0{,}05$, zijn 50 proefpersonen nodig. Wil men voorzichtiger zijn, door $\alpha = 0{,}01$ en ß = 0,95 te kiezen, dan zijn bij een groot verwacht effect 51 en bij een middelmatig effect 128 proefpersonen per groep nodig.

Gedragstherapie, jrg. 25, nr. 2, pp. 107-121, juni 1992

Tabel 2. Minimaal benodigde aantal proefpersonen bij vier soorten statistische toetsen, voor drie niveaus van de power, drie niveaus van de eenzijdige overschrijdingskans (α_1) en drie niveaus van de effectgrootte (klein, middelmatig en groot volgens Cohen, 1977).

power	0,80			0,90			0,95		
α_1	0,10	0,05	0,01	0,10	0,05	0,01	0,10	0,05	0,01
	t-toets onafhankelijke steekproeven								
klein	226	310	503	329	429	652	428	542	790
middelmatig	36	50	82	53	69	105	69	87	128
groot	14	20	33	21	27	42	27	35	51
	t-toets afhankelijke steekproeven (r = 0,50)								
klein	114	155	253	165	215	327	215	270	396
middelmatig	20	26	42	28	36	54	36	45	65
groot	8	11	17	12	15	22	15	18	26
	chi-kwadraat (2 x 2)								
klein	618	785	1168	856	1051	1488	1082	1300	1781
middelmatig	69	87	130	95	117	165	120	144	198
groot	25	31	47	34	42	60	43	52	71
	correlatiecoëfficiënt								
klein	450	618	998	655	864	1296	864	1105	1585
middelmatig	49	68	107	71	93	138	93	118	168
groot	17	22	36	24	31	45	31	39	55

T-toets voor gepaarde gegevens

De te volgen procedure voor de power-an-am-analyse is in principe voor iedere toets gelijk. Bij het vaststellen lenllen van α en ß gelden dezelfde overwegingen als hierboven genoe-noe-noemd. Alleen voor het vaststellen van de effectgrootte is enige aæ aje aanpassing vereist. Bij een t-toets voor gepaarde gegevens corrigrriprrigeert men de effectgrootte als volgt:

$$d = \sqrt{\frac{d'}{\sqrt{1-r}}} \quad (2)$$

d1 = effectgrootte gebaseerd op niet-gepaarde rdeirde gegevens; d = definitieve (gecorrigeerde) effectgrootte; r = correlatiecoëffoëfcoëfficiënt voor het verband tussen de gepaarde gegevens.

Een voorbeeld

Stel we willen weten of het gevoel van controlerolcrole bij patiënten toeneemt door het aanbieden van een stress-managemermermentcursus. De scores voor en na de cursus worden vergeleken. Dit is een sen sen situatie waarin een t-toets voor gepaarde gegevens gebruikt wordt. We zije zije zijn alleen tevreden als de scores op een 10-puntsschaal minimaal 2 puntenuntenunten verschuift, en weten uit vorig onderzoek dat de gezamenlijke standaardtardharddeviatie 4 is.

Als het zou gaan om niet-gepaarde gegevens, ens, ens, zou de effectgrootte gelijk zijn aan 2/4 = 0,5. Bij gepaarde gegevens mo ms moet de aldus berekende effectgrootte worden gecorrigeerd door te delen denlen door $\sqrt{(1-r)}$. In dit voorbeeld geeft r het verband aan tussen de scores resres voor en na de behandeling. Stel dat r = 0,50, dan wordt de werkelijke eke eke effectgrootte 0,71.

Als de r niet bekend is, is 0,50 een niet al tal t al te gewaagde keuze. Kennis over de feitelijke r is echter altijd ted tjd te prefereren, omdat bij een wat grotere r de vereiste steekproefgrrfgrefgrootte afneemt.

De reden dat de correlatiecoëfficiënt wowo wordt gebruikt bij de correctie is als volgt te begrijpen. Bij gepaarderdarde gegevens wordt een deel van de variantie van de gegevens bij bij bij de postmeting verklaard uit de waarden bij de premeting. Wi Wi. Wanneer bij pre- en postmeting twee verschillende groepen zon zon zouden zijn gebruikt, is er over het geheel genomen geen verbandanoand tussen een bepaalde waarde voor en na de behandeling (r = 0). 0)- 0).

De benodigde minimale steekproefgrootoorootte bij gepaarde gegevens kan nu worden opgezocht of berekenkenkend. Er is bij de correlatiecoëfficiënt van 0,50 ongeveer de helft vft vft van het aantal proefpersonen nodig in vergelijking met een t-toetsoetoets voor niet-gepaarde gegevens. In feite is de situatie nog gunstinstinstiger, en is slechts 1/4 van het aantal proefpersonen nodig, omdandmdat de steekproefgrootte per cel geldt, en bij herhaalde metingen wn wn worden dezelfde proef-

personen tweemaal gebruikt. Er zijn uiteraard ook situaties waarin de t-toets voor gepaarde gegevens wordt gebruikt en het laatste niet geldt, bijvoorbeeld wanneer het IQ-verschil tussen de seksen wordt bepaald door telkens een zuster en broer uit hetzelfde gezin te nemen.

Chi-kwadraat-toets

Vaak worden gegevens in de vorm van frequenties weergegeven in zogenaamde contingentietabellen, met k kolommen en r rijen. Een voorbeeld is het weergeven van de resultaten van een onderzoek naar sekseverschillen (2 kolommen) bij angststoornissen (bijvoorbeeld beperkt tot paniekstoornis mét en zonder agorafobie; 2 rijen). Toetsing geschiedt hier met de chi-kwadraat-toets. Voor deze toets wordt een aparte effectgrootte berekend, die niet is te herleiden tot de effectgrootte d. Cohen gebruikt de effectgrootte-maat W, die vergelijkbaar is met de chi-kwadraat zélf. De chi-kwadraat is, zoals bekend, een maat voor het verschil tussen de waargenomen frequenties en de frequenties verwacht volgens de nulhypothese. De W is een maat voor het verschil tussen de proporties (= frequentie/N) verwacht op grond van de nulhypothese, en de proporties verwacht volgens de alternatieve hypothese. Het verband is als volgt:

$$W = \sqrt{\frac{\chi^2}{N}} \quad (3)$$

De W is afhankelijk van het aantal vrijheidsgraden (k-1)*(r-1). In tabel 2 zijn de vereiste steekproefgroottes weergegeven voor de meest simpele situatie: de 2 x 2 tabel (df = 1). Voor de effectgrootte-maat W komt een klein, middelmatig en groot effect overeen met respectievelijk W = 0,10, W = 0,30 en W = 0,50. Belangrijk is op te merken, dat voor de chi-kwadraat-toetsen de tabellen van Cohen, en ook onze tabel 2, betrekking hebben op *het totale aantal proefpersonen,* en niet op het aantal proefpersonen per cel.

In het hiervoor genoemde voorbeeld weten we, dat bij paniekstoornis mét agorafobie een groot sekseverschil is te verwachten, terwijl bij paniekstoornis zónder geen of een gering verschil optreedt. Stel we kiezen de gebruikelijke α-waarde van 0,05, maar willen voorzichtiger zijn dan gebruikelijk met betrekking tot de ß en kiezen daarvoor 0,10. Bij een groot verwacht effect zijn dan totaal 42 patiënten nodig, en bij een middelmatig effect 117.

Pearson product moment-correlatiecoëfficiënt

Bij het bepalen van de vereiste steekproefgrootte bij correlatieberekeningen kan als effectmaat de (verwachte) correlatiecoëfficiënt zélf gebruikt worden. De grootte ervan moet echter anders geïnterpreteerd worden dan de effectmaat d voor het verschil in gemiddelden. Een klein, middelmatig en groot effect komen dan overeen, aldus Cohen, met correlatiecoëfficiënten van 0,10, 0,30 en 0,50. Een aantal voorbeelden wordt gegeven in tabel 2.

Verkleinen van de spreiding

Hoe groter de effectgrootte, des te minder proefpersonen heeft men nodig bij een gegeven α en ß. Het is daarom het overwegen waard of men door het aanpassen van experimentele factoren de effectgrootte kan doen toenemen. Dat kan door te trachten de afstand tussen de gemiddelden te vergroten door een krachtiger manipulatie toe te passen en/of door verkleining van de spreiding (zie formule 1). De spreiding kan verkleind worden door

a. het wegnemen van externe storingsbronnen;
b. het kiezen van een opzet met herhaalde metingen;
c. het werken met homogene groepen.

Externe storingsbronnen kunnen van velerlei aard zijn, en behandeling van dit punt voert hier te ver; zie daarvoor de vele teksten over onderzoeksopzetten. Door het kiezen van een opzet met herhaalde metingen wordt een belangrijke storingsbron ingeperkt, namelijk de interindividuele verschillen. Het voordeel hiervan zit al 'ingebakken' in de aangepaste berekening van de effectgrootte, zoals besproken onder het hoofd t-toetsen. Het werken met homogene groepen betekent meestal het kiezen van een 'randomized block design'. Er worden deelgroepen gevormd aan de hand van een variabele waarvan een grote invloed is te verwachten op de uitkomstvariabele. In het eerste voorbeeld is te verwachten dat het therapieresultaat afhankelijk is van de duur van de stoornis. Op grond daarvan kan men twee subgroepen formeren: één met een lange en één met een korte voorgeschiedenis. Binnen beide subgroepen worden vervolgens patiënten random verdeeld over de twee behandelingsvormen. Een andere methode is de beperking tot één homogene groep waarin men het sterkste effect verwacht; in het voorbeeld waarschijnlijk de groep met een korte ziektegeschiedenis. Beperking kan strijdig zijn met de eis tot generaliseerbaarheid van de resultaten. Als opzet voor een deelstudie of een vooronderzoek is het

echter zeer wel te verdedigen. Als men een nieuwe behandeling wil testen, waarvan men het idee heeft dat ze vooral effectief zal zijn bij niet zo ernstige patiënten, is er niets op tegen patiënten daarop te selecteren. Als het onderzoek positief uitvalt, kan daarna uitgezocht worden voor welke patiënten de behandeling wél en voor welke patiënten zij niét effectief is.

Discussie

Als een onderzoek met een betrekkelijk gering aantal proefpersonen toch een statistisch significant verschil oplevert, heeft men het goed getroffen: er is klaarblijkelijk sprake van een krachtig effect. Een dergelijke opzet houdt echter een aanzienlijk risico in. Het niet vinden van een statistisch significant verschil laat onzekerheid achter, omdat het onderscheidingsvermogen – de kans op het aantonen van een verschil – te gering zou kunnen zijn geweest.

In het voorbeeld dat we aan het begin van dit artikel gaven, is duidelijk sprake van een te geringe power. Bij de gekozen α = 0,05, een steekproefomvang van 25, en een verwachte middelmatige effectgrootte van d = 0,50 is de power 0,54. Er is dus bijna 50% kans dat de nulhypothese (= er is geen verschil) ten onrechte wordt geaccepteerd. Het design is ongeschikt om middelmatige effecten mee aan te tonen, en het is zeker niet gerechtvaardigd om te concluderen dat beide behandelingsvormen niet van elkaar verschillen. Eigenlijk was het onderzoek alle geïnvesteerde moeite en geld niet waard. Als de onderzoeker uit het voorbeeld met enige zekerheid wil aantonen dat het toevoegen van een medicamenteuze behandeling aan een cognitief-gedragstherapeutische behandeling geen effect heeft, zal zij of hij moeten kiezen voor een hoge power, bijvoorbeeld 0,95. Er zijn dan 87 proefpersonen per conditie nodig. Een steekproefomvang van 25 personen per conditie zou wél voldoende zijn geweest als er een groter effect werd verwacht, en als het doel van de onderzoeker niet zo duidelijk was gericht op het aantonen van géén verschil. Een groter therapie-effect zou men kunnen verwachten als gekozen was voor een opzet waarbij één van beide behandelingsvormen zou worden vergeleken met géén behandeling. Een verwacht therapie-effect van 0,80 zou in dit voorbeeld een power gegeven van 0,87. Bij deze opzet is de kans dat er ten onrechte geconcludeerd wordt tot geen verschil, gereduceerd tot 13%. De vraagstelling bij dit laatste voorbeeld en de daaruit voortvloeiende keuze van de controlegroep is echter fundamenteel verschillend van de vraagstelling uit het eerste voorbeeld. In het eerste

voorbeeld gaat het om een vergelijking in de effectiviteit van twee behandelingsvormen, in het tweede voorbeeld wordt de vraag gesteld hoe groot het behandelingseffect is ten opzichte van géén behandeling. Het is duidelijk dat beide vraagstellingen verschillende eisen stellen aan de power en de steekproefgrootte.

Het uitvoeren van een power-analyse zou een vast onderdeel bij het opzetten van elk experiment moeten zijn. Dat het zo weinig wordt gedaan, heeft behalve met de vermeende ingewikkeldheid van de berekeningen, ook te maken met een aantal onzekerheden: welke power moet men kiezen, en hoe moet de effectgrootte worden bepaald? Wat de power betreft; iedereen die een statistische test heeft leren uitvoeren, heeft ook leren werken met het vaststellen van de α, en de achtergronden daarvan zijn niet moeilijker dan voor de power. Bovendien heeft Cohen het met zijn adviezen gemakkelijk gemaakt. Door deze te volgen slaat men geen gek figuur, al is het te overwegen de ß-fout meer gewicht te geven dan Cohen doet. Het bepalen van de effectgrootte is lastiger. Als er geen pilot study is gedaan en er zijn ook geen schattingen te maken op grond van literatuurgegevens, hetgeen zelden het geval zal zijn, dan moet men het doen met de globale aanwijzingen van Cohen omtrent kleine, middelmatige en grote effecten. Het toepassen van de meest simpele standaardregels bij de berekening van de vereiste steekproefgrootte is altijd beter dan geen power-analyse toe te passen. Deze regels houden in, dat uitgegaan wordt van $\alpha = 0{,}05$, ß = 0,20, en middelmatige effectgrootte (en $r = 0{,}50$ bij gepaarde gegevens). In tabel 2 is voor vier gangbare toetsen af te lezen welke aantallen proefpersonen daarbij horen.

Het verdient echter de voorkeur om op grond van gegevens de effectgrootte te berekenen, al was het maar om het inzicht te vergroten in de te verwachten afstand tussen de gemiddelden en de spreiding in de gegevens (bij toetsen van het type t-toets of F-toets). Dat geeft ook zicht op factoren die wellicht experimenteel zijn aan te passen: vergroting van de afstand tussen de gemiddelden door een krachtiger manipulatie toe te passen, en een verkleining van de spreiding.

Summary

Power analysis

Although standard statistical texts pay considerable attention to the problem of alpha-errors (concluding that there is a difference, while in fact

there is not) little attention has been paid to beta-errors (concluding there is no difference, while in fact there is). The probability of committing the first kind of error is established by the investigator as the level of significance, or alpha (α). The complement of ß (1 – ß) is the probability of obtaining a significant result, and is referred to as the power of a statistical test. Power analysis represents a method for determining power and/or establishing the needed sample size. Due to small sample sizes, many studies have insufficient power. In this article we discuss power-analysis in the context of several research examples.

Key words: power analysis

Literatuur

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences.* New York: Academic Press.

Hoogstraten, Joh., & Koele, P. (1981). Cohen's foute berekening van de vergeten fout. *Tijdschrift voor onderwijsresearch, 6,* (4), 174-181.

Singer, B.R., Lovie, A.D., & Lovie, P. (1986). Sample size and power. In: *New developments in statistics for psychology and the social sciences.* Londen: Methuen, pp. 129-142.