Stem-, Spraak- en Taalpathologie
1999, Vol. 8, No.2, pp. 86-93

32.8310/99/0802-086

Hoe groot moet de steekproef zijn?

T. Rietveld

Afdeling Taal en Spraak, Faculteit der Letteren, Katholieke Universiteit Nijmegen

In deze bijdrage wordt aandacht besteed aan de wijze waarop kan worden bepaald hoe groot een of meer steekproeven moeten zijn om effecten van behandelingen of condities in een of meer (sub)populatie(s) vast te stellen. Gewezen wordt op de omstandigheid dat in de spraak- en taalpathologie vaak niet met gestandaardiseerde meetprocedures en gestandaardiseerde afhankelijke variabelen wordt gewerkt; het ontbreken van die standaardisatie beperkt in ernstige mate de mogelijkheid om op een statistisch verantwoorde wijze de benodigde omvang van een steekproef te bepalen.

Inleiding

Stel, we willen een onderzoek doen naar de effecten van Presurgical Orthodontic Treatment (PSOT) bij jonge kinderen met een gehemeltespleet; de behandeling bestaat uit het aanbrengen van een gehemelteplaatje. In een enigszins acceptabele onderzoeksopzet zullen schisiskinderen at random aan twee behandelingen worden toegewezen: met en zonder PSOT. Vervolgens zal worden nagegaan of de behandeling effect heeft op relevante variabelen die bijvoorbeeld betrekking hebben op de ontwikkeling van de spleet of de spraakontwikkeling. De vraag is, hoeveel kinderen zijn in zo'n experiment nodig? Deze vraag is met name in (para-)medisch onderzoek van grote relevantie, omdat daar meer dan in andere disciplines rekening moet worden gehouden met onprettige en/of kostbare behandelingen. Het spreekt dan vanzelf dat men het aantal proefpersonen tot een minimum wil beperken. Stelt men de vraag naar het aantal benodigde proefpersonen aan een foneticus of spraakpatholoog (de auteur was vroeger alleen foneticus) dan is het antwoord vaak geformuleerd in termen van hele getallen; bij de formulering van dat antwoord wordt wat gefronst gekeken, en gemompeld "10 tot 20 kinderen lijkt me wel genoeg"; bij navraag wordt dan bedoeld "10 tot 20 kinderen in elke groep". Men zij gewaarschuwd: zo'n antwoord is óf op niets gebaseerd, óf op een grote ervaring met soortgelijke vragen en soortgelijke proefpersonen. Is dat laatste niet het geval, dan kan men net zo goed at random een getal trekken uit een lijst van 1 tot 1000. De statisticus daarentegen zal zeker niet direct een antwoord geven. Hij/zij zal de volgende vragen stellen:

Correspondentieadres: Toni Rietveld, Afdeling Taal en Spraak, Faculteit der Letteren, Katholieke Universiteit Nijmegen, Postbus 9103, 6500 HD Nijmegen.

- Welk verschil tussen de populatie(s) wilt U meten? (in het Engels heet dat verschil de *effect size* Δ);
- Wat is de variatie op de afhankelijke variabele binnen de subpopulaties? (de *standaarddeviatie* σ, geschat door s);
- Hoe erg is het als U een tussen de populaties aanwezig verschil in de steekproeven niet meet? (de toegelaten kans op het maken van een dergelijke fout noemt men β);
- Hoe erg is het als U op basis van een tussen de steekproeven gemeten verschil ten onrechte aanneemt dat dat verschil ook tussen de subpopulaties aanwezig is? (de toegelaten kans op het maken van een dergelijke fout heet α: het welbekende 'significantieniveau');
- Hoe is de proefopzet? In ons voorbeeld worden de kinderen at random aan de groepen toegewezen, maar er zijn ook opzetten denkbaar waarbij proefpersonen herhaaldelijk worden gemeten, in meerdere condities, of op verschillende tijdstippen.

Als deze vragen zijn beantwoord kan de statisticus een antwoord geven, en zeggen hoe groot de steekproeven minimaal moeten zijn. In het geval van het onderzoek naar de effecten van PSOT was het antwoord: 23 kinderen in elke behandeling. Dit antwoord kon worden gegeven op basis van de volgende gegevens:
De *afhankelijke variabele* is de cephalometrische hoek SNA (sella-nasion-point A);
De *effect size* in SNA-hoek bedraagt 3°;
De *standaarddeviatie* van de SNA-hoek is 4,1°;
De kans op het *ten onrechte* aannemen dat er een verschil is tussen beide groepen mag maximaal 5% bedragen: α=0,05;
De *kans op het ten onrechte aannemen* dat er geen verschil is tussen beide groepen mag maximaal 20% bedragen: $\beta = 0{,}20$

Het vermogen van een toets

De bovengenoemde kans β staat in direct verband met een belangrijk concept, nl. dat van het *Vermogen van een toets* (Engels: Power). Het vermogen van een toets is gelijk te stellen aan de sterkte van het 'vergrootglas' bij visuele waarneming. Hoe kleiner het te detecteren verschil is, hoe sterker het vergrootglas dient te zijn. Als we een klein verschil tussen twee gemiddelde waarden van twee populaties willen vaststellen op basis van twee steekproeven, en de variatie in de populaties is vrij groot, dan mogen we aannemen dat het verschil tussen de twee steekproefgemiddelden eveneens vrij sterke variaties zal vertonen, vooral als de steekproeven relatief klein zijn. De kans dat we dan op basis van de steekproeven een verschil tussen de populaties kunnen detecteren, is dan relatief klein. Dat dat zo is, kunnen we alleen goed begrijpen als we eerst naar een belangrijk begrip van de statistiek kijken, nl. de Standard Error (S.E.), de standaard deviatie van een steekproefgrootheid ('statistic'). Voor een eenvoudige vorm van de S.E. kijken we naar figuur 1. Hier is afgebeeld de verdeling van een steekproefgemiddelde.

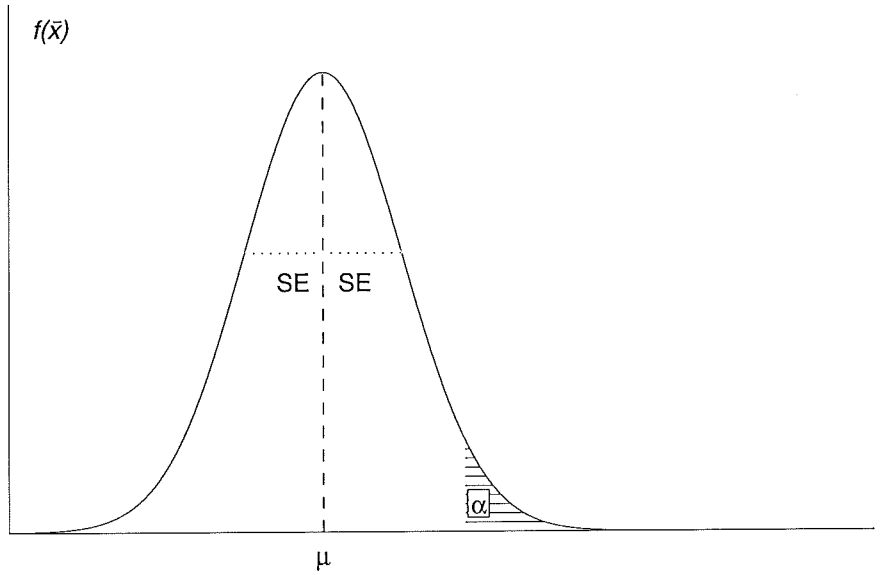

Fig. 1. De verdeling van steekproefgemiddelden $\bar{x}$ rondom het populatiegemiddelde μ

Deze verdeling laat zien dat wanneer we uit een populatie een heleboel steekproeven trekken van een bepaalde grootte (n = 10, of n = 100), de gemiddelden van die steekproeven zullen varieren rondom het populatiegemiddelde. Het zal duidelijk zijn dat de variatie van die steekproefgemiddelden zal toenemen naarmate de populatie waaruit getrokken is meer variatie vertoont (bij een grote σ) en de steekproef (n) klein is. De variatie van die steekproefgemiddelden heet de Standard Error, en voor $\bar{x}$ bedraagt deze:

$$S.E.(\bar{x}) = \frac{\sigma}{\sqrt{n}}$$

Er zijn ook verdelingen van ingewikkelder steekproefgrootheden, zoals bijvoorbeeld het verschil tussen twee steekproefgemiddelden $\bar{x}_2 - \bar{x}_1$; de formules voor de bijbehorende standard errors zijn ook wat ingewikkelder. Maar in al die formules komen steeds weer twee belangrijke bepalende elementen voor: de standaarddeviaties van de populaties waaruit de steekproeven worden getrokken (σ), en de grootte van de steekproeven (n). Kortom, de S.E. is een functie van twee elementen:

$$S.E. = F(\sigma, 1/n)$$

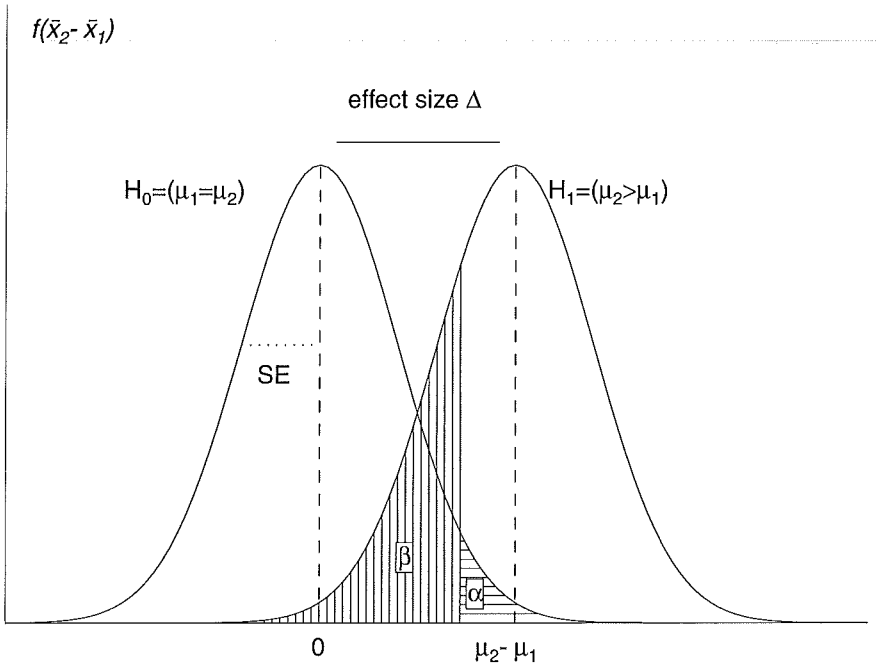

Fig. 2. De verdelingen van de verschillen tussen steekproefgemiddelden $\bar{x}_2 - \bar{x}_1$ rondom twee hypothetische verschilscores, nl.: $\mu_2 - \mu_1 = 0$ en $\mu_2 - \mu_1 > 0$

Dus: hoe groter de variatie in de populatie(s), hoe groter de Standard Error, en hoe groter de steekproef, hoe kleiner de Standard Error. Het eerste element hebben we vaak niet onder controle, het tweede element, de grootte van de steekproef, vaak wel (maar natuurlijk ook niet altijd).

Bovenstaande figuur lijkt ingewikkeld, maar is dat niet. Wat is daar afgebeeld? Net zoals steekproefgemiddelden een verdeling vertonen – zie figuur 1 – zullen ook verschillen tussen steekproefgemiddelden een verdeling tonen. In deze figuur laten we er twee zien:

- De linkse verdeling heeft betrekking op het verschil tussen twee steekproefgemiddelden $\bar{x}_2 - \bar{x}_1$ als de populatiegemiddelden gelijk aan elkaar zijn: $\mu_2 = \mu_1$ (de situatie die vaak de 'nul-hypothese' wordt genoemd).
- De rechter verdeling heeft ook betrekking op het verschil tussen twee steekproefgemiddelden $\bar{x}_2 - \bar{x}_1$, maar nu als het populatiegemiddelde 2 groter is dan het populatiegemiddelde 1: $\mu_2 > \mu_1$ (de 'alternatieve hypothese').

De Standard Error is ook ingetekend, alleen voor de linker verdeling. We weten dat de S.E. groter wordt naarmate de standaard-deviaties in de populaties groter worden en/of de steekproefgrootte kleiner wordt.

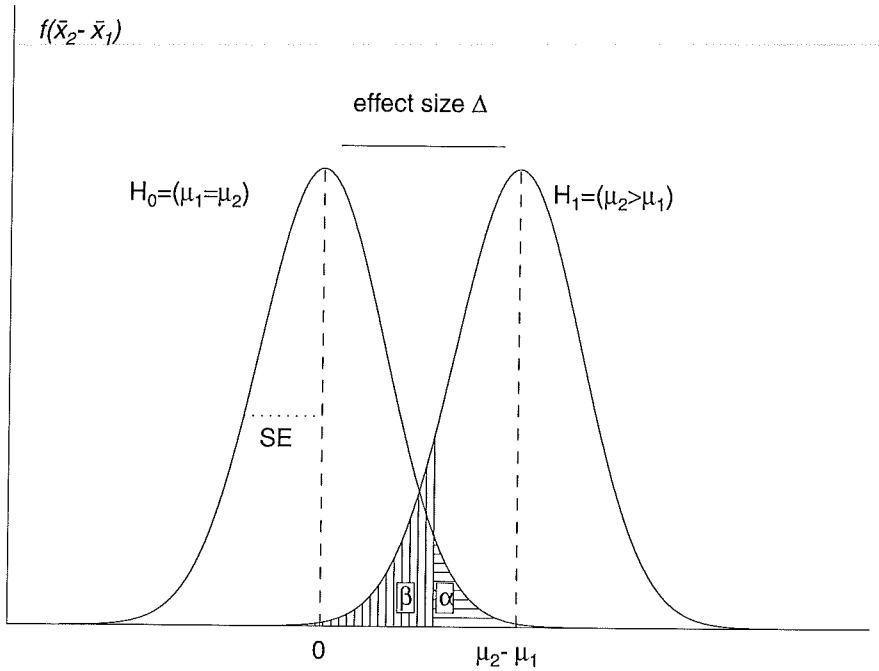

Fig. 3. Als in figuur 2, maar nu met kleinere SE

Het gaat hier om de grootte van het gebied dat met β is aangeduid. Immers, dat gebied reflecteert de kans dat een observatie (hier $\bar{x}_2 - \bar{x}_1$) als te klein wordt beschouwd om gegenereerd te kunnen zijn door de verdeling waarbij $\mu_2 > \mu_1$, en daarom maar beschouwd wordt als gegenereerd vanuit de verdeling waarbij de aanname is: $\mu_2 = \mu_1$. De onderzoeker zegt dan dat de nulhypohese niet is verworpen ("niet significant"). Als de S.E. kleiner is – zie figuur 3 – is β ook kleiner, wanneer althans de andere gegevens – de afstand tussen μ_2 en μ_1 en de waarde van α – gelijk blijven.

De SE zou verkleind kunnen zijn door grotere steekproeven te nemen. Verkleining van β betekent dat minder snel de nulhypothese zal worden geaccepteerd, die namelijk zegt dat het verschil tussen de twee steekproefgemiddelden alleen maar door 'steekproefvariatie' tot stand is gekomen, en dat de steekproeven uit dezelfde populatie afkomstig zijn. Hoe kleiner β, hoe groter het vermogen van onze toets om op basis van steekproeven te kunnen besluiten dat deze afkomstig zijn van populaties met verschillende gemiddelden. Het vermogen van een toets heeft officieel de volgende vorm:

$$\text{Vermogen} = 1 - \beta.$$

Als een verschil tussen twee steekproefgemiddelden als 'niet-significant' wordt ver-

klaard, moeten we dus wel weten hoe groot het vermogen van de toets is geweest. Als dat vermogen klein was, hoeft het geen opzien te baren als een verschil 'niet significant' blijkt te zijn. Als het vermogen daarentegen groot was, hoeven we niet te verwachten dat de steekproeven getrokken zijn uit populaties met verschillende gemiddelden. De (oude) les is dus: bij een niet significant verschil altijd het vermogen van de toets noemen!

Het vermogen van bijvoorbeeld een verschiltoets is de kans dat we een verschil tussen twee subpopulaties detecteren op basis van een steekproef, gegeven de aanname dat de twee subpopulaties inderdaad van elkaar verschillen. Het vermogen hangt, zoals we gezien hebben, af van vier factoren:

- de grootte van het effect dat we willen detecteren: de *effect size;*
- de grootte van de steekproef;
- de variatie in de populatie;
- de grootte van de steekproef.

De eerste twee factoren – de grootte van het te detecteren effect en de grootte van de steekproef – heeft de onderzoeker voor een deel zelf in de hand. Het is goed voorstelbaar dat op basis van deze gegevens het vermogen te berekenen is; het gaat daarbij om het bepalen van de oppervlakte onder een kansverdeling (in ons geval steeds de 'rechter verdeling'), waarbij de rechtergrens van het β-gebied bepaald wordt door α in de linker verdeling. In ons verhaal gaat het echter om iets anders. Wij willen niet *post-hoc* het vermogen $(1-\beta)$ van de toets bepalen, maar hebben eisen gesteld wat betreft dat vermogen (die moet bijvoorbeeld 0,80 bedragen, bij een kans van α die 0,05 bedraagt) en de effect size die we willen kunnen detecteren. Hierbij gaat het dus om de grootte van n die tot vervulling van die eisen leidt. Wij zullen niet ingaan op het rekenwerk dat daarvoor nodig is, maar wel enige aanwijzingen voor de intuïtie geven:

De grootte van de SE bepaalt de vorm van verdelingen van bijvoorbeeld $\bar{x}_2 - \bar{x}_1$. Als de effectsize en α bekend zijn, en ook de gewenste grootte van β dan moet alleen nog maar de grootte van SE worden vastgelegd. Zoals we gezien hebben is die afhankelijk van σ en n. De standaard-deviatie σ moet bekend zijn, dan wel moet daarvan een goede schatting te geven zijn; n blijft dan over. Kennis van de variatie is makkelijker te verkrijgen indien er gestandaardiseerde variabelen zijn waarop wordt gemeten. Een voorbeeld van dat type scores vormen 'nasalance'-percentages van een nasaliteitsmeter, of de eerder genoemde cephalometrische hoek SNA met zijn standaarddeviatie van 4,1°. Minder gestandaardiseerd – tot nog toe – zijn allerlei perceptieve schalen die in de spraak- en taalpathologie in gebruik zijn: 'verstaanbaarheid', 'nasalering'. Sommige onderzoekers gebruiken drie-, andere vijf- en weer anderen tienpuntsschalen. Deze verschillen zien we niet alleen tussen onderzoekers van verschillende landen, waar het aantal schaalpunten zich nog wel eens aan het schoolbeoordelingssysteem lijkt aan te passen, maar ook tussen onderzoekers binnen hetzelfde land. Dit gebrek aan standaardisering maakt het moeilijk om schattingen te maken over de variatie in de populatie. Er is dus alle reden om te streven naar gestandaardiseerde meetprocedures, omdat alleen die het mogelijk maken om een verantwoorde schatting te maken van de variatie in de populatie; die variatie is een belangrijk ingredient bij de bepaling van

Tabel 1. De grootte van de benodigde steekproeven om een verschil van 20 percentpunten te kunnen detecteren tussen twee proporties: π_1 en π_2

π_1	π_2	benodigde steekproefgrootte
0,10	0,30	67
0,20	0,40	89
0,30	0,50	101
0,40	0,60	106
0,50	0,70	101
0,60	0,80	89

de steekproefgrootte die nodig is om behandelingsmethoden zinvol met elkaar te vergelijken of effectmetingen uit te voeren.

Ter oriëntatie geven we in tabel 1 de grootte van de steekproeven die benodigd zijn om een verschil van 20 percentpunten tussen de proporties in twee populaties (π_1 en π_2) te kunnen detecteren; het gaat hier dus om een variabele die is uitgedrukt in percentages, bijvoorbeeld het percentage gestotterde woorden. De tabel maakt duidelijk dat de benodigde steekproefgrootte niet onaanzienlijk is, en ook nog varieert bij verschillende waarden van de veronderstelde proporties in de populatie (π). De standaard statistische pakketten als SPSS bieden weinig of geen mogelijkheden om steekproefgrootten te bepalen. Men moet zijn toevlucht nemen tot specifiek daarvoor ontwikkelde programmapakketten, zoals *nQuery Advisor 2.0*, of tabellen zoals gegeven in Lemeshow et al. (1993).

Conclusie

We keren terug naar het voorbeeld van de inleiding. De statisticus had bepaald hoe groot de steekproeven moesten zijn, op basis van gegevens omtrent een specifieke variabele, de SNA. Dat betekent nog niet dat die steekproeven groot genoeg waren voor andere variabelen, zoals scores op variabelen die de spraak- en taalontwikkeling moeten beschrijven. Veel minder is bekend van scores op die variabelen; bovendien zijn metingen op die variabelen vaak minder gestandaardiseerd. Het zal daarom geen verbazing wekken dat de steekproeven voor detectie van verschillen op de spraak- en taalvariabelen niet altijd groot genoeg bleken te zijn om acceptabele vermogens van de toetsen op te leveren. De gestandaardiseerde variabelen – vaak afkomstig uit de medisch-fysiologische hoek – domineren de situatie bij de bepaling van de steekproefgrootte en daarmee ook een belangrijk deel van het design van een onderzoek. Er is slechts één manier om uit te komen onder die dominantie: zelf standaardiseren. Ik denk dat hier een belangrijke taak ligt voor de beroepsorganisaties van degenen die werkzaam zijn op het terrein van de taal- en spraakpathologie.

Summary

In this contribution we discuss the way in which the sample size can be calculated which is needed for the detection of effects of conditions or treatments in one or more (sub)population(s). Special attention is paid to the situation in speech and language pathology, where quite often standardised measurement procedures and standardised dependent variables are not available. This lack of standardisation is a limiting factor when it comes to determining the needed sample size in a statistically sound way.

Noot

Ik dank Dr.Ir. B. Cranen voor hulp bij het maken van de figuren.

Literatuur

Lemeshow, S., Hosmer, D.W., Klar, J. & Lwanga, S.K. (1993). *Adequacy of Sample Size in Health Studies.* Chichester: John Wiley & Sons.

Prahl, C. (1993). *A study into the effects of presurgical orthopedic treatment in complete unilateral cleft lip and palate patients: a three centre prospective clinical trial Nijmegen, Amsterdam and Rotterdam.* Intern protocol, Academisch Ziekenhuis Nijmegen.

Rietveld, T. & van Hout, R. (1993). *Statistical Techniques for the Study of Language and Language Behaviour.* Berlin: Mouton de Gruyter.