Hoe kies ik een effect size bij het plannen van een dierexperiment en welke effectgrootte is nog interessant om te meten?

Anton FJ de Haan[1], Peter HM Klaren[2], Steven Teerenstra[3]

1 Afdeling Health Evidence, sectie Biostatistiek, Radboud Institute for Health Sciences, Radboudumc, contact: ton.dehaan@radboudumc.nl

2 Afdeling Animal Ecology & Physiology, Radboud Institute for Biological and Environmental Sciences (RIBES), Radboud Universiteit

3 Afdeling Health Evidence, sectie Biostatistiek, Radboud Institute for Health Sciences, Radboudumc

De manier waarop een effectgrootte (effect size) wordt uitgedrukt kan verschillende vormen aannemen. Bij een interventiestudie is het de uitkomst van een interventie. Bijvoorbeeld bij het behandelen van een tumor kan dit het verschil in succespercentage zijn van de behandeling met een actieve stof vergeleken met een placebobehandeling. Bij een onderzoek naar de samenhang van cholesterol in bloed en de bloeddruk kan dit de correlatiecoëfficiënt zijn tussen beide gemeten grootheden. Kortom, effect size hangt af van de onderzoeksvraag. Voor aanvang van een experiment moet je eerst een keuze maken hoe je het effect van je experiment gaat vaststellen (meten) en motiveren waarom die keuze de relevante informatie geeft die je uit het experiment wilt verkrijgen.

Hoe concreter de effectmaat is, hoe makkelijker de interpretatie van het resultaat van de studie wordt. Bijvoorbeeld een correlatiecoëfficiënt zegt alleen maar dàt uitkomsten samenhangen, maar een richtingscoëfficiënt (helling) van een regressielijn vertelt je hoe ze samenhangen. Net zo geeft een gestandaardiseerd verschil in gemiddelden weer hoeveel standaarddeviaties het verschil is, maar daarmee weet je niet of dit een relevant verschil is. Het zou een groot verschil kunnen zijn (in absolute zin) als de standaarddeviatie groot is en het zou klein kunnen zijn als de standaarddeviatie klein is. Enkele veelgebruikte effectmaten zijn:

- Verschil tussen twee steekproef- of groepsgemiddelden van een numerieke variabele als je twee (of meerdere) condities bestudeert.
- Verschil in succespercentages als je twee (of meerdere) condities bestudeert.
- Gestandaardiseerd verschil in gemiddelden: het verschil in gemiddelden wordt dan gedeeld door een spreiding. Deze spreiding kan dan zijn: spreiding in de controle groep of de gepoolde spreiding van alle groepen. Kortom, ook hier zijn meerdere keuzes mogelijk en daarom is het goed om dit vooraf vast te leggen.
- Een risicoreductie kan uitgedrukt worden als de ratio van risico's ('relative risk') of het verschil in risico's ('risk difference').
- Grootte van de correlatiecoëfficiënt als je de samenhang tussen twee metingen (van dezelfde of verschillende uitkomsten) wilt bestuderen. Dit kan uitgebreid worden als de samenhang tussen een 'afhankelijke' variabele met meerdere 'onafhankelijke' variabelen wordt bestudeerd: multipele correlatiecoëfficiënt (of verklaarde variantie: R2)
- Bij het vergelijken van meerdere groepen zijn ook andere, minder intuïtieve effectmaten gangbaar. Bijvoorbeeld de ratio van de variantie van de gemiddelden en de gepoolde variantie binnen de groepen (zoals bijvoorbeeld in ANOVA).

Hoe groot moet het effect zijn?

Bij steekproefgrootteberekeningen is de keuze van de effectgrootte het lastigst. Soms is er uit vorige experimenten al een indruk hoe groot het effect zou kunnen zijn en als dit een betekenisvol effect is om aan te tonen, dan kan het verwachte effect genomen worden. Het voordeel is dan dat je van tevoren weet dat deze grootte van effect ook realistisch is om te verwachten.
In discussies over de keuze van effectgrootte geven onderzoekers vaak aan dat ze niet weten hoe groot het effect is en dat ze daarom geen steekproefgrootte kunnen berekenen. Dat is echter geen steekhoudend argument. Bij de keuze van de effectgrootte gaat het erom dat je studie relevante informatie oplevert. Als je bijvoorbeeld een percentage gaat schatten (bijvoorbeeld het succes van een tumorbehandeling of de mate waarin corona onder jongeren voorkomt) en je steekproefomvang is zo klein dat het betrouwbaarheidsinterval +/- 50 % rondom je schatting ligt, dan weet je nog steeds niet of het een hoog of laag percentage is. Er zijn twee mogelijke opties. De eerste optie is de sample size zo groot nemen dat je een redelijke precisie krijgt in het effect dat je gaat schatten (ongeacht waar het ligt). Het gaat er dus om de standaardfout (standard error) van het effect voldoende klein te krijgen. De andere optie is om je af te vragen welk effect de moeite waard is om gevonden te worden. Bedenk dat de keuze van een groot effect weliswaar het voordeel heeft dat een kleine steekproef nodig is, maar je moet je afvragen of je daarmee verder geholpen bent. Immers, als het effect in je studie kleiner is dan welke je gebruikt heb voor een powerberekening, ben je dan niet teleurgesteld als dit niet statistisch significant is? Kortom, wat is SESOI, de smallest effect size of interest. Ook dit hangt af van de context: wil je een effect groter dan de best bestaande behandeling? Een effect even groot? Of, als het over risico's gaat, een effect van zoveel procentpunten minder risico? Belangrijk is dat de onderzoeker hier zijn vakinhoudelijke expertise/kennis moet aanwenden en ondanks dat dit vaak moeilijk is, moet de onderzoeker toch een keuze maken, want het is de verantwoordelijkheid van de onderzoeker om te onderbouwen dat zijn onderzoek met voldoende zekerheid zinvolle informatie zal opleveren. Vergeleken met het ongerief van de proefdieren is dit klein leed voor de onderzoeker.
De bovenstaande voorbeelden gaan vooral uit van de situatie dat de schaal waarop je meet goed te interpreteren is. Er kunnen ook situaties zijn waarin verschillen op de schaal (uitkomst/meting/variabele) waarop je meet (nog) niet goed te interpreteren zijn. Denk bijvoorbeeld aan een pas ontdekt eiwit. Dan is het een optie om effecten te vergelijken met 'biologische variabiliteit', bijvoorbeeld het effect uitgedrukt in standaarddeviaties. Men kan gedurende een project steeds meer inzicht krijgen of effecten die groot zijn in termen van de biologische variabiliteit ook (biologisch) relevant/van betekenis zijn voor je vraagstelling. En aan het einde van een project kan men zelfs al een idee hebben van de grootte van het effect van een behandeling. Dat verwachte effect kan men dan in verder (klinisch) onderzoek proberen te repliceren.

Voorbeeld

Het dierenlab wil een indruk krijgen hoe veel ratten er een besmetting met het Kilham Rat-virus hebben. Daartoe willen ze bij een aantal ratten kijken of ze met dit virus in aanraking zijn geweest. Het percentage ratten dat in aanraking is geweest is de effectmaat. Bij dit onderzoek is het dus van belang dat dit percentage voldoende nauwkeurig wordt vastgesteld. Hier moeten we dus een keuze maken voor de standaard fout (standard error). Bij een standard error van 10 % zal het betrouwbaarheidsinterval een breedte hebben van 40 %. Dat wil zeggen: het 95 %-CI zal lopen van het gevonden percentage minus 20 % tot het gevonden percentage plus 20 %. Deze 20 % = 2 x SE en 2x SE wordt ook wel de margin of error genoemd. Dit is een tamelijk vage schatting. Een keuze voor de standaard fout van 5 % is dus heel verdedigbaar. Voor deze nauwkeurigheid zijn tenminste 96 ratten nodig. Dit bereken je bijvoorbeeld in Rus Lenth's software (https://homepage.divms.uiowa.edu/~rlenth/Power/index.html) met de keuze 'CI (confidence interval) for one proportion' en omdat we niet weten waar het percentage ligt, kiezen we de optie 'worst case'. Indien je beperkt bent in het aantal ratten dat je kunt gebruiken voor deze vraag, dan ben je ook beperkt in je nauwkeurigheid. Dit is ook te onderzoeken met genoemd programma. Bijvoorbeeld: als maar tien ratten haalbaar zijn, dan wordt de margin of error 30 % en moet je je gaan afvragen of dit wel voldoende informatie geeft om de vraag te beantwoorden.