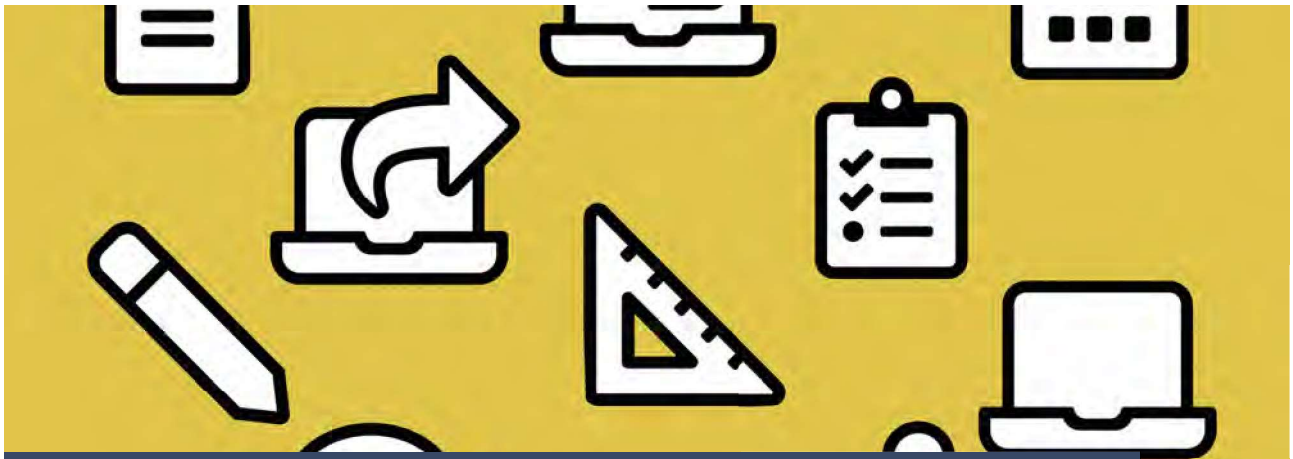

B·E·A·N·S: de E is van effectgrootte

"Of the four factors that determine statistical power, effect size is the most problematic."
Mark W. Lipsey, 1990, in: Design Sensitivity. Statistical Power for Experimental Research.

Peter H.M. Klaren
Afdeling Animal Ecology & Physiology, Radboud Institute for Biological and Environmental Sciences (RIBES), Radboud Universiteit, contact: peter.klaren@ru.nl

Inleiding

Artikel 1d, tweede lid van de Wet op de dierproeven (Wod, geldend sinds 18 december 2014) stelt: "Het aantal dieren dat in projecten wordt gebruikt, wordt tot het minimum beperkt zonder dat de doelstellingen van het project in gedrang komen." Artikel 10, tweede lid van dezelfde wet herhaalt nog eens dat uit de verschillende mogelijkheden díe dierproef moet worden gekozen waarin "een zo gering mogelijk aantal dieren" wordt gebruikt. Niet voor niets, dus, dat het ontwerp van experimenten, procedures en projecten deel uitmaakt van onze proefdierkundecursussen, in navolging van EU-Richtlijn 2010/63.

Het minimum aantal dieren waarover de Wod rept is niet een absoluut of een vaststaand minimum, maar een minimum zonder dat de doelstellingen van het project in het gedrang komen. Deze toevoeging maakt het noodzakelijk dat een onderzoeker weet heeft van deze projectdoelstellingen.
In de praktijk betekent het dat de onderzoeker kan aangeven wat de verwachte of gewenste concrete uitkomsten van een dierproef zijn. In dierexperimenteel onderzoek zijn dat vrijwel altijd de gehoopte of gewenste resultaten van een of andere interventie.

Het zijn deze gewenste resultaten die, met enkele andere aannames, een steekproefgrootteberekening mogelijk maken. Het is deze steekproefgrootte ook die, naar mijn mening, het wettelijk bedoelde minimum aantal dieren is. De gewenste resultaten vormen ook de effectgrootte die een onderzoeker klinisch of wetenschappelijk relevant en interessant vindt. Zonder vooraf uitgesproken effectgrootte kan geen zinnige steekproefgroottebereken gedaan worden [1].
Steeds meer wetenschappelijke tijdschriften vragen hun auteurs om een steekproefgroottebereke-ning, en het is opvallend dat (nog steeds) heel weinigen daarin slagen [2]. Hopelijk levert dit artikel een aanzet tot verbetering.

B·E·A·N·S

Resultaten van experimenten worden vaak statistisch getest tegen een zogenaamde nulhypothese. Deze nulhypothese is bijna altijd 'de hypothese van geen effect'. De nulhypothese is de hypothese die we na een statistische test al of niet verwerpen ten gunste van een alternatieve hypothese die het tegendeel van de nul stelt. Het is de klassieke statistiek die deze null hypothesis significance tests (NHST) uitvoert[1]. Met de uitkomst van een significantietest wordt hier de p-waarde bedoeld. Deze p-waarde is de kans (p staat voor probability) op het vóórkomen van de waargenomen data, of extremere data, onder de aanname dat de nulhypothese waar is [3].
Er zijn vijf factoren die bepalen of een significantietest een statistisch significant resultaat oplevert of niet. Ze bepalen daarmee de gevoeligheid of statistical power van een statistische analyse. Deze factoren komen samen in het acroniem BEANS.[2] De steekproefgrootte is één van die factoren. De crux van een steekproefberekening is nu dat wanneer vier van de vijf factoren bekend zijn (of dat er een redelijke aanname voor bestaat), de vijfde (hier: steekproefgrootte, N in BEANS) kan worden berekend. We gaan die vijf factoren hieronder in niet-alfabetische volgorde langs.

A – alfa, αAlfa (α)

is het significantieniveau van een significantietest en wordt conventioneel en traditioneel (en enigszins arbitrair) gesteld op 0.05 (5 %), soms lager. Dat wil zeggen dat bij $p < \alpha$ de nulhypothese (er is geen effect) wordt verworpen en een resultaat tot statistisch significant wordt verklaard.
Een statistisch significant resultaat kan dus worden gezien als een succesvolle detectie van een bepaald effect van een bepaalde interventie. Alfa wordt ook wel als de kans op een zogenaamde Type-1-fout geïnterpreteerd. Een Type-1-fout wordt (onbewust) begaan wanneer een nulhypothese ten onrechte wordt verworpen. Een vals-positieve waarneming dus.
De onderzoeker denkt iets gedetecteerd te hebben dat er in werkelijkheid (in de populatie waar de steekproeven uit afkomstig zijn) niet is [5].

B – bèta, β

Naast een Type-1-fout kan een Type-2-fout worden begaan. De kans daarop wordt aangegeven met bèta, β. Een Type-2-fout is een vals-negatief: het onterecht accepteren van een (foute) nulhypothese. Omdat de statistische significantietest een niet-significant resultaat heeft, mist de onderzoeker het effect dat er in werkelijkheid wel is. In het algemeen en conventioneel (en arbitrair) wordt $\beta = 0.20 = 20\ \%$ nog als maximaal acceptabel beschouwd.
Uit β kan de power: $1 - \beta$, van een significantietest worden berekend. Met power wordt de gevoeligheid of het onderscheidingsvermogen van een test bedoeld. Uit het voorgaande blijkt dat een gevoeligheid van minimaal 80 % gebruikelijk is. Dat wil zeggen dat, áls er een effect in de populatie aanwezig is (dus áls een interventie een effect heeft), er een kans van 80 % is dat dat als een statistisch significant resultaat door de onderzoeker wordt gedetecteerd en opgepikt [6-8].

S – variabiliteit (standaardafwijking)

De biologische variabiliteit in een populatie én de analytische variabiliteit in metingen is de achtergrondruis waarin een effect van een interventie moet worden gemeten. Het zal intuïtief duidelijk zijn dat hoe groter de variabiliteit (hoe meer ongewenste ruis), hoe lastiger het is een bepaald effect (signaal) te detecteren. Variabiliteit verkleint de signaal-ruisverhouding. In die experimenten waarin een numeriek verschil tussen twee steekproefgemiddelden, of waarin een lineaire regressie tussen twee numerieke variabelen wordt geanalyseerd, is de standaardafwijking (Engels: standard deviation) een veelgebruikte maat voor de variabiliteit in een populatie [4].
Wetenschappelijke literatuur, klinische referentiewaarden, een pilotexperiment, of expert knowledge zijn goede bronnen waaruit vaak een realistische schatting van variabiliteit en een standaardafwijking kan worden gehaald.

De variabiliteit kan een onderzoeker controleren en beheersen door bijvoorbeeld, en als het mogelijk is, dieren van één geslacht, één leeftijd, hetzelfde gewicht, afkomstig uit hetzelfde nest, dezelfde faciliteit, etc. te gebruiken. Consistentie in laboratoriumwerk is eveneens belangrijk: gebruik dezelfde machines en apparatuur (en lees de handleidingen), werk volgens expliciete protocollen (kamertemperatuur en overnacht incuberen zijn bijvoorbeeld geen universele constanten), houd de kwaliteit van chemicaliën in de gaten (vaak hebben die een houdbaarheids- of kwaliteitsgarantiedatum, herhaaldelijk invriezen en ontdooien bevordert de kwaliteit niet). In het geval van categoriale variabelen hebben we te maken met tellingen in categorieën waarvoor geen standaardafwijking kan worden berekend. Hier wordt de standaardfout (een andere maat voor de variabiliteit) van een waarneming of telling bepaald door het totaal aantal waarnemingen in een experiment.

N – steekproefgrootte

Tenzij er bijvoorbeeld met kostbare dieren wordt gewerkt en (financiële) middelen beperkt zijn, of wanneer een studie wordt uitgevoerd aan patiënten met een zeldzame ziekte, hoeft de factor steekproefgrootte in principe weinig hoofdbrekens te kosten. Wanneer een onderzoeker de conventionele waarden $\alpha = 0.05$ en $\beta = 0.20$ (dus power = $1 - 0.20 = 0.80$) aanneemt, een reële schatting heeft van de variabiliteit, dan is met de gewenste effectgrootte de steekproef-grootte vrij gemakkelijk te berekenen [9-11]. De software G*Power stelt de onderzoeker in staat om voor meer dan 40 experimentele ontwerpen deze steekproefgrootteberekening uit te voeren.[3] Daarbij biedt de software opties voor poweranalyse en andere analyses.[4]

E – effectgrootte

Zoals in het citaat bovenaan dit artikel al is aangegeven: dit is de meest problematische van de vijf factoren. Problematisch omdat, anders dan voor α en β, er geen conventionele constante waarden zijn voor de effectgrootte. Problematisch ook, wellicht, omdat hier een beroep wordt gedaan op de expertise van de onderzoeker en zijn/haar inzicht in de doelstellingen van het onderzoek.
Om een voorbeeld te geven: in een onderzoek dat is gericht op de ontwikkeling van een nieuwe statine is het niet gepast een statisticus te vragen naar een wenselijk effect van een dergelijk medicijn. Hoe groot is een 'klinisch interessante en relevante' daling van het cholesterolgehalte in het bloed? Een medicus kan dit uiteraard beter inschatten dan een niet ter zake kundige statisticus!

De schrijver dezes is óók geen deskundige op het gebied van statines[5], maar suggesties voor een reëel, minimaal interessant effect van een nieuwe statine kunnen zijn:

- Een daling in serumcholesterol die minstens even groot is als die van de beste huidige therapie.
- Een daling in cholesterol die x % beter is dan die van de beste huidige statine.
- Een daling in cholesterol die x % minder kans geeft op de ontwikkeling van hart- en vaatziekten.
- X % minder bijwerkingen dan die van de beste huidige statine.
- Een daling in serum cholesterolgehalte met een 95 % betrouwbaarheidsinterval van 1 mmol/L.
- Een samenhang tussen serum cholesterol en statine-dosis met een correlatiecoëfficiënt van
- tenminste $r = 0.70$.

Afhankelijk van welke variabele wordt gemeten dient de onderzoeker een gepaste maat voor de effectgrootte te kiezen. Deze zou eigenlijk al a priori, in het project omschreven moeten zijn. Wanneer de nieuwe statine aan een groep patiënten wordt toegediend en de werking ervan wordt vergeleken met die van een placebogroep is een simpel verschil tussen gemiddelde cholesterolgehaltes in het bloed een voor de hand liggende effectgrootte. Wanneer in een prospectieve studie de incidentie van bijwerkingen wordt geteld zijn bijvoorbeeld relative risk of odds ratio toepasselijke maten voor de effectgrootte.

Conclusie

Steekproefgrootteberekeningen zijn met aannames voor α, β, variabiliteit, en belangrijk: *een smallest effect size of interest* [1], en met gebruikmaking van (gratis) software [12-14], goed uit te voeren. Het vraagt de onderzoeker zich bewust te zijn van de doelstellingen van een project en in staat te zijn een beroep te doen op de eigen deskundigheid. Het bepalen van een te meten effect van een interventie is niet aan een statisticus over te laten. De onderzoeker dient ook bereid te zijn zich enigszins te verdiepen in de, toegegeven, soms wat minder makkelijk toegankelijke software-handleidingen. Deze investeringen staan echter in geen verhouding tot het ongerief dat proefdieren kan worden aangedaan, en worden goedgemaakt door de kwaliteitsverbetering van het onderzoek.

Bronnen

1. Lakens D (2021) Sample size justification. Collabra: Psychology 8: 33267 (doi: 10.1525/collabra.33267)
2. Macleod MR, McLean AL, Kyriakopoulou A, et al. (2015) Risk of bias in reports of in vivo research: A focus for improvement. PLoS Biology 13: e1002273 (doi: 10.1371/journal.pbio.1002273)
3. Curran-Everett D (2009) Explorations in statistics: hypothesis tests and P values. Advances in Physiology Education 33: 81-6 (doi: 10.1152/advan.90218.2008)
4. Howard BR (2002) Control of variability. ILAR Journal 43: 194-201 (doi: 10.1093/ilar.43.4.194)
5. Miller J, Ulrich R (2019) The quest for an optimal alpha. PloS One 14: e0208631 (doi: 10.1371/journal.pone.0208631)
6. Cohen P (1982) To be or not to be: control and balancing of type I and type II errors. Evaluation and Program Planning 5: 247-53 (doi: 10.1016/0149-7189(82)90076-3)
7. Sedgwick P (2011) Sample size and power. British Medical Journal 343: (doi: 10.1136/Bmj.D5579)
8. Krzywinski M, Altman N (2013) Power and sample size. Nature Methods 10: 1139-40 (doi: 10.1038/nmeth.2738)
9. Brysbaert M (2019) How many participants do we have to include in properly powered experiments? A tutorial of power analysis with reference tables. Journal of Cognition 2: 16 (doi: 10.5334/joc.72)
10. Lazic SE, Clarke-Williams CJ, Munafò MR (2018) What exactly is 'N' in cell culture and animal experiments? PLoS Biology 16: e2005282 (doi: 0.1371/journal.pbio.2005282)
11. Lazic SE (2018) Four simple ways to increase power without increasing the sample size. Laboratory Animals 52: 621-9 (doi: 10.1177/0023677218767478)
12. Faul F, Erdfelder E, Lang AG, Buchner A (2007) G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behavior Research Methods 39: 175-91 (doi: 10.3758/Bf03193146)
13. Faul F, Erdfelder E, Buchner A, Lang A-G (2009) Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. Behavior Research Methods 41: 1149-60 (doi: 10.3758/BRM.41.4.1149)
14. Lenth RV (2007) Statistical power calculations. Journal of Animal Science 85: E24-9 (doi: 10.2527/jas.2006-449).

Noten

[1] De lezer heeft ongetwijfeld al eens kennisgemaakt met enkele significantietesten (bijvoorbeeld Students t-test, chi-kwadraat test, ANOVA of variantie-analysevariantieanalyse, etc.) en de beroemde/beruchte p-waarde als uitkomst daarvan.

[2] In het citaat bovenaan dit artikel spreekt Lipsey van 4 factoren die de uitkomst van een significatietest bepalen, niet de 5 die in BEANS voorkomen. De verklaring ligt erin dat Lipsey een zogenaamde relatieve effectgroottemaat (Cohen's d) beschouwt, waarin E (effectgrootte) en S (standaardafwijking) samenkomen.

[3] Samen met tutorials en achtergrondartikelen te downloaden van: https://www.psychologie.hhu.de/arbeitsgruppen/allgemeine-psychologie-und-arbeitspsychologie/gpower.

[4] Er is ook andere software beschikbaar: Piface van Russell V. Lenth voert poweranalyses en steekproefgrootte-berekeningen uit voor 14 experimentele scenario's (http://www.stat.uiowa.edu/~rlenth/Power). De gebruikers van de software R hebben de keuze uit verschillende packages zoals: BUCSS, MBESS, pamm, pwr, simglm, Superpower, WebPower (en meer) die exacte analyses en simulaties uitvoeren.

[5] Statines zijn een groep medicijnen die zogenaamd LDL-cholesterol ('slecht' cholesterol) in het bloed helpen verlagen. Ze worden vaak voorgeschreven ter voorkoming van hart- en vaatziekten.