Instantie voor Dierenwelzijn Utrecht

Reviewed by department of statistics
Info: statistics@Lumc.nl
Leids Universitair Medisch Centrum

Toelichting proefopzet en statistiek

Nelleke Verhave
Paul Westers
Januari 2017

I. Inleiding

Bij het inzetten van dieren voor onderzoek is statistiek een belangrijk onderdeel om tot een kwalitatief goed experiment te komen. Statistiek biedt instrumenten om te bepalen hoeveel dieren je minimaal nodig hebt om een bepaald verwacht effect met voldoende bewijskracht aan te kunnen tonen (poweranalyse), maar ook instrumenten om de metingen te analyseren (beschrijvende en inferentiële statistiek), zodat je verantwoord conclusies kunt trekken.

Daarmee is statistiek meestal zelfs een essentieel onderdeel bij het doen van verantwoord onderzoek met aandacht voor de 3 V's. Statistiek zal geen antwoord geven op de vraag of een onderzoek zinvol en/of ethisch verantwoord is, maar met de onderzoeksvraag in het achterhoofd kan statistiek wel bijdragen aan een optimaal onderzoeksdesign.

Het is verstandig om bij het opzetten van nieuw onderzoek met proefdieren altijd een statisticus te raadplegen. Bedenk wel dat de statisticus meestal geen expert is op het onderzoeksgebied zelf. Het contact zal een wisselwerking zijn tussen de kennis van de onderzoeker en van de statisticus. Goede informatie-uitwisseling is dan ook cruciaal. Het uiteindelijke resultaat zal een balans zijn tussen een goed, verantwoord onderzoek met een optimaal design, waarbij ook al gedacht is aan de uiteindelijke analyses. De onderzoeksvraag en het welzijn van de proefdieren zullen daarbij altijd centraal staan.

Om de wisselwerking vlotter te laten verlopen worden in deze notitie de basale statistische begrippen kort uitgelegd. Aan het eind van de notitie worden referenties gegeven voor meer informatie.

II. Aandachtspunten

Om een dierproef goed op te zetten en uit te voeren, en de resultaten daarvan te analyseren en te presenteren, zijn er verschillende aandachtspunten die voor een groot deel met elkaar samenhangen:

1. Wat zijn de onderzoeksvragen? Maak daarbij onderscheid tussen primaire en secundaire onderzoeksvragen.
2. Wat is de opzet van de dierproef? Bedenk daarbij wat ga je meten, hoe ga je het meten en welke dieren je nodig hebt. Maar ook: wordt het een gepaard of ongepaard design, hoe organiseer je de randomisatie en blindering, hoe wil je later de data analyseren, welke praktische zaken zijn vereist? Etc.
3. Hoe worden de meetgegevens geanalyseerd? Of: wat is het statistisch analyseplan? Bedenk welke resultaten je wilt beschrijven, toetsen of modeleren, welke hypothesen je wilt toetsen, welke statistische analyses je wilt toepassen. Ga je één- of tweezijdig toetsen? Etc.
4. Hoeveel dieren heb je minimaal nodig? Bepaal dat op basis van het design en het statistisch analyseplan. Bedenk of de bepaalde steekproefgrootte leidt tot praktische problemen. Pas zo nodig het design, het statistisch analyseplan of de poweranalyse aan.
5. In hoeverre kan rekening worden gehouden met de 3 V's?
6. Ben je voorbereid op onverwachte zaken? Wat ga je doen met extreme metingen? Wat ga je doen als dieren uitvallen? Wat doe je als je niet voldoet aan de voorwaarden van de statistische analyses? Etc.
7. Wordt er gewerkt volgens de ARRIVE- of GSP-richtlijnen? Alle aspecten van de dierproef, van de opzet tot het schrijven van het artikel, moeten voldoen aan de richtlijnen. Houd daar rekening mee.
8. Hoe worden de meetgegevens uiteindelijk weergegeven in een databestand?. Ieder dier krijgt zijn eigen unieke code. Die code moet terugkomen in het databestand. Privacy-gebonden informatie van een dier moet in een apart bestand staan dat alleen toegankelijk is voor een zeer beperkte groep van betrokkenen. Alle data, ook als ze niet worden gebruikt in de statistische analyse, worden opgeslagen in het databestand. De structuur van het databestand wordt bepaald door de statistische analyse. De standaard is per dier 1 regel met in alle kolommen de gegevens van het dier die verzameld zijn. Dus niet alleen de persoonlijke gegevens van het dier (ras, gewicht, leeftijd etc.) maar ook de

experimentele conditie, het tijdstip van meting, de metingen zelf, etc. Houd een codeboek bij van het databestand, zodat iedereen precies kan weten hoe data ingevoerd moet worden.

9. Welke resultaten wil je presenteren in het artikel en op welke manier? Het kan erg frustrerend zijn als je achteraf bepaalde belangrijke of interessante resultaten niet kunt presenteren, omdat je daar bij de opzet van de dierproef niet aan gedacht hebt, of geen metingen voor gedaan hebt.
10. Heb je voldoende statistische bagage om de analyses te kunnen begrijpen en te verdedigen? Raak vertrouwd met de basisprincipes en -concepten van de statistische analyses die je gaat gebruiken. Het leren omgaan met de statistische software hoort daarbij.

Tip: houd een logboek bij waarin je niet alleen alle gemaakte keuzes en afspraken terug kunt vinden maar ook kunt verantwoorden. Schroom niet om op tijd bij de proefdierdeskundige en/of statisticus langs te gaan om advies te vragen, al was het maar ter controle.

III. Opzet van dierproeven

Onderzoeksvraag

De basis van iedere dierproef of serie dierproeven vormen één of meerdere onderzoeksvragen. Deze onderzoeksvragen zijn geformuleerd op basis van bestaande of nieuwe theorieën, of als vervolg op eerder onderzoek. Meestal zijn er 1 of 2 hoofdonderzoeksvragen (primaire onderzoeksvragen), en zijn er daarnaast minder belangrijke, maar interessante onderzoeksvragen (secundaire onderzoeksvragen). Een goede formulering van de onderzoeksvragen is essentieel voor het slagen van een dierproef. Zo moet de onderzoeksvraag niet te vaag zijn, maar ook niet te gedetailleerd. Beter een serie van eenvoudige onderzoeksvragen dan een ingewikkelde onderzoeksvraag met veel toeters en bellen.

De primaire onderzoeksvraag bevat vaak de kern van de onderzoeksvraag, en heeft betrekking op de belangrijkste meting. De secundaire onderzoeksvragen hebben betrekking op de andere metingen, of zijn verfijningen van de primaire onderzoeksvraag. De primaire onderzoeksvraag vormt de basis voor de poweranalyse.

Proefopzet (Design)

Met design wordt de manier waarop de proef wordt opgezet bedoeld. Het design wordt voornamelijk bepaald door de primaire onderzoeksvraag, maar er wordt ook rekening gehouden met de secundaire onderzoeksvragen. Bij het design wordt de onderzoeksvraag verder geoperationaliseerd. Het is dus essentieel dat de onderzoeksvragen duidelijk en eenduidig zijn geformuleerd.

Na het goed nadenken over de onderzoeksvragen is het nu dus belangrijk om goed na te denken over de opzet van de dierproeven. Welke experimentele condities ga je met elkaar vergelijken? Ga je gebruik maken van 2 of meer onafhankelijke groepen (parallel of ongepaard design) of juist afhankelijke groepen (gepaard of gematcht design)? Wat voor type data worden gemeten en hoe ga je die meten? Wat is praktisch mogelijk? Hoeveel dieren heb je nodig en aan welke eisen moeten de dieren voldoen? Zullen de dieren de dierproef overleven, en zo ja, zijn ze dan nog geschikt voor andere dierproeven?

Een belangrijk onderdeel bij het opzetten van het design is alvast nadenken over de statistische analyse. Een kleine wijziging aan het design kan de statistische analyse verbeteren. Aan de andere kant stelt een statistische analyse ook eisen aan het design.

Gepaard of ongepaard

Bij ongepaarde design worden de dieren *at random* gekoppeld aan een experimentele conditie. Bij een gepaard design kun je denken aan dieren afkomstig uit hetzelfde nest die *at random* verdeeld worden over de experimentele condities, of dieren waarbij op verschillende momenten in de tijd wordt gemeten (herhaalde metingen). De minst simpele vorm van dit laatste is een meting voor en na een behandeling. Een bijzondere

vorm van gepaarde data is als dieren op basis van kenmerken aan elkaar gematcht zijn. Dieren die aan elkaar gepaard of gematcht zijn noemen we ook wel een paar.

Het voordeel van gepaarde groepen is dat bij de statistische analyse gecorrigeerd kan worden voor de verschillen tussen de dieren (biologische variabiliteit), en je dus minder dieren nodig hebt. Het moet echter praktisch mogelijk zijn en uiteraard ethisch verantwoord.

Controlegroepen

Om te bepalen of interventie(s), meestal (een) specifieke experimentele conditie(s), een effect heeft(hebben), vergelijk je de resultaten met een referentie-experimentele conditie, meestal een controlegroep of een controlemeting (voor de interventie). Het is essentieel dat de referentie of controlegroep in samenstelling en behandeling vergelijkbaar is met de experimentele groep(en). Daarom zijn randomisatie en geblindeerd werken van groot belang.

Blindering en Randomisatie

Met blinderen zorg je ervoor dat informatie die verwijst naar de toegewezen experimentele condities geheim blijft voor alle belanghebbenden van de dierproef, die anders, bewust of onbewust, beïnvloed zouden kunnen worden door deze informatie.

Randomisatie zorgt ervoor dat dieren volstrekt willekeurig worden toegewezen aan een experimentele conditie. Een gebruikelijke gecontroleerde manier van randomisatie is een aselecte loting. Ieder dier heeft dezelfde kans om bij een experimentele conditie ingedeeld te worden. Bij een gepaard design randomiseer je binnen ieder paar, zodat alle experimentele condities vertegenwoordigd zijn in een paar. Je maakt op voorhand een randomisatieplan, zodat tijdens de dierproef vaststaat bij welke experimentele conditie een volgend dier wordt ingedeeld. Dit voorkomt dat bepaalde dieren bij een specifieke experimentele conditie worden ingedeeld. Het beste is als de toewijzing van de dieren aan een specifieke conditie niet door de onderzoeker zelf gebeurt, maar blind, door een onafhankelijke persoon. Het ligt voor de hand om dat te laten doen door de verzorger van de dieren. Controle achteraf is nodig om te toetsen of de toewijzing conform het randomisatieplan is uitgevoerd.

Blindering en randomisatie zijn essentieel omdat je zo (on)bewuste storende invloeden op je experiment kunt voorkomen.

Bias

Bij een bias zijn je resultaten vertekend doordat bijvoorbeeld een structureel een te hoge of te lage waarde wordt gemeten. Dit heeft vooral effect op subjectieve parameters, maar mag zeker niet genegeerd worden bij vermeende objectieve metingen. Het is dus verstandig om vooraf alle meetinstrumenten te controleren op mogelijke bias. Trainingen en duidelijke afspraken over de subjectieve parameters met de betrokkenen die de metingen gaan uitvoeren, horen er ook zeker bij.

IV. Statistisch Analyseplan

Bij het opzetten van dierproeven moet je ook alvast nadenken hoe je de data uiteindelijk gaat analyseren. Dit schrijf je op in het statistisch analyseplan. Niet alleen bepalen de onderzoeksvraag en het design de statistische analyse, maar de statistische analyse stelt ook eisen aan het design. Elke statistische analyse geeft niet alleen schattingen van de effecten maar ook het resultaat van een statistische toets.

Uitkomstvariabele

De uitkomstvariabele is de variabele die de uitkomst van een meting of waarneming weergeeft. Hiermee worden de resultaten van de dierproef beschreven en geanalyseerd. Je kunt onderscheid maken tussen primaire en secundaire uitkomstvariabelen.

Continue variabele

Als de meting of waarneming een reëel getal is (bijv. lengte, gewicht, bloeddruk), dan zeggen we dat de uitkomstvariabele continue is. In de meeste basis-statistische analyses (bijv. ANOVA, t-toets, correlatie/regressie) wordt verondersteld dat een continue variabele normaal verdeeld is.

Discrete variabele

Als de data bestaan uit categorale waarden, dan spreken we ook wel van een discrete variabele. Een discrete variabele kan binair zijn (wel of niet drachtig, dood of levend), nominaal (bloedgroep, kleur van de vacht) of ordinaal (kaal, lichte beharing, matige beharing, mooie vacht, dikke vacht). In tegenstelling tot een nominale variabele, hebben de mogelijke uitkomsten van een ordinale variabele een logische rangorde.

Tellingen zijn een speciale soort van uitkomstvariabele. In principe zijn ze niet continue maar als de range van de mogelijke tellingen groot is, dan worden ze toch wel als continue beschouwd. De grens waarbij tellingen als discrete of continue variabele wordt beschouwd is arbitrair, en hangt af van de dierproef.

Om verschillende redenen bestaat er wel eens de wens om continue variabelen te transformeren naar categorale variabelen (bijv. “leeftijd in jaren” wordt “jong, middelbaar of oud”). Op zich is dat niet erg, als je bij het formuleren van je onderzoeksvraag en het opzetten van de dierproef daar maar rekening mee hebt gehouden. In ieder geval zal informatieverlies optreden en dus verlies aan power.

Normale verdeling

De standaard parametrische statistische toetsen (bijv. ANOVA, t-toets, correlatie/regressie) gaan uit van een continue variabele die normaal verdeeld is. Met een normaalverdeling wordt bedoeld dat alle meetwaarden verdeeld zijn volgens een klokvorm. De meeste metingen vallen rond de gemiddelde meting, en hoe verder je van het gemiddelde afgaat des te minder waarden je meet.

Het is een misverstand dat de data normaal verdeeld moet zijn. Bij de meeste standaard statistische analyses is juist de voorwaarde dat de residuen normaal verdeeld moeten zijn. In de dagelijkse praktijk wordt bij de t-toets en ANOVA gekeken of de data per groep normaal verdeeld zijn. Het mag duidelijk zijn dat dit minder power heeft dan kijken of de overall residuen van alle groepen samen normaal verdeeld zijn.

De meeste statistische programma’s geven bij een parametrische statistische toets een optie om te bepalen of de residuen normaal verdeeld zijn. Parametrische statistische toetsen zijn robuust. Dat wil zeggen dat kleine schendingen van de normale verdelingen geen gevolgen hebben voor de toets.

Mocht een continue variabele niet normaal verdeeld zijn, dan kan eventueel een transformatie op de meting worden uitgevoerd (bijv. het logaritme of de wortel) of gebruik worden gemaakt van non-parametrische statistische technieken (bijv. Mann-Whitney, Kruskal-Wallis, Spearman rangcorrelatie). Het is een misverstand om te veronderstellen dat non-parametrische geen andere voorwaarden heeft. Zo hebben de Mann-Whitney en de Kruskal-Wallis toets ook nog de voorwaarde dat de vorm van de verdelingen bij de verschillende groepen gelijk zijn. En dat de verdeling van de verschillen bij de Wilcoxon rangtekentoets symmetrisch is.

Naast de voorwaarde van de normale verdeling, heeft elke statistische toets ook nog andere voorwaarden, waarvan de meest belangrijke zijn: onafhankelijkheid en gelijkheid van spreiding.

Statistische analyse

Bij een statistische analyse kun je denken aan het toetsen van verwachtte effecten, maar ook aan het beschrijven van mogelijke associaties. In het eerste geval kan je denken aan het vergelijken van gemiddelden en proporties. Als we spreken over associaties dan denken we meestal aan samenhangen tussen 2 of meer categorale variabelen (bijv. chi-kwadraattoets of loglinear model) of tussen 2 of meer continue variabelen (

bijv. correlatie en (multiple) lineaire regressie). Overlevingsstatistiek en (multiple) logistische regressie zijn voorbeelden waarbij je te maken hebt met een mix van continue en categorale variabelen.

Zowel bij het toetsen van effecten als bij het modelleren van associaties is het toetsen een belangrijk onderdeel.

Nulhypothese en alternatieve hypothese bij een statistische toets

In termen van statistiek geeft de nulhypothese weer, in exacte en kwantitatieve vorm, wat je **niet verwacht** als uitkomst van je experiment. De nulhypothese moet dus zo zijn geformuleerd dat je deze met de parameters uit je experiment wel of niet kunt verwerpen. Kortweg houdt de nulhypothese in dat er geen effect is. De alternatieve hypothese houdt dan in dat er wel een effect is.

De toetsingsprocedure is dat je, ervan uitgaande dat de nulhypothese waar is, op basis van de data tracht aannemelijk te maken dat de nulhypothese niet waar kan zijn. Je kan het vergelijken met een rechtszaak waarbij het uitgangspunt is dat de gedaagde onschuldig is en vervolgens met bewijsmateriaal getracht wordt om te bewijzen dat de gedaagde schuldig is.

Wanneer de nulhypothese niet verworpen kan worden, kun je concluderen dat het verwachtte effect niet is gevonden of althans niet aannemelijk is gemaakt. In een rechtszaak zou dat betekenen dat er onvoldoende bewijsmateriaal was om de gedaagde schuldig te verklaren. Dat wil echter niet zeggen de nulhypothese bevestigd is en dat er geen effect is: een gedaagde die niet schuldig is verklaard hoeft nog niet onschuldig te zijn.

Wanneer de nulhypothese wel verworpen kan worden, kun je veronderstellen dat de alternatieve hypothese (‘wel effect’) waar is.

Het effect kan afhankelijk van de onderzoeksvraag betrekking hebben op het verschil tussen 2 of meer gemiddelden of proporties, maar bijvoorbeeld ook over de samenhang tussen 2 of meer gemeten uitkomst variabelen. Het resultaat van het toetsen van de effecten wordt vaak uitgedrukt als: wel (nulhypothese verwerpen) of niet (nulhypothese niet verwerpen) statistisch significant. Naast de statistische significantie is er ook een ander aspect dat vaak genegeerd of onderschat wordt, nl. (klinische) relevantie. Een gevonden effect kan wel statistisch significant zijn, maar voor de dagelijkse praktijk niet relevant. Aan de andere kant kan een effect statistisch niet significant zijn, maar gezien de grootte van het effect wel (klinisch) relevant.

Bij het publiceren van de resultaten van de dierproeven is het dan ook essentieel om naast het vermelden van de p-waarde (t.b.v. statistische significantie) ook de grootte van het effect te vermelden met een betrouwbaarheidsinterval (t.b.v. (klinische) relevantie).

Tot zover zijn we ervan uitgegaan dat het doel van de dierproef was: aantonen dat er wel een effect is. Echter, het is ook mogelijk dat je aan wilt tonen dat er géén effect is. Het principe blijft verder gelijk, alleen wordt nu de nulhypothese dat er wél een effect is, en moeten de data het bewijs leveren dat het effect te verwaarlozen is. Dit soort designs worden ook wel bio- equivalentiestudies genoemd. Hiervoor zijn in de regel veel meer proefdieren nodig.

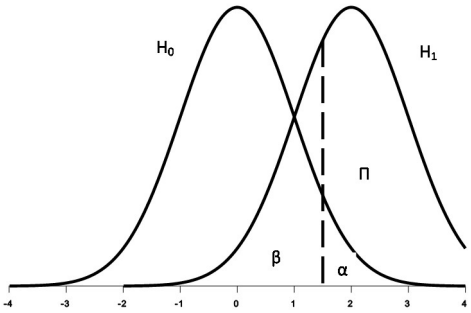
Inleiding toetstheorie

Het principe van toetsen is het nagaan of er voldoende bewijs is dat een bepaalde veronderstelling (‘er is een effect’) aannemelijk is. Het uitgangspunt (de nulhypothese) is juist datgene wat je niet hoopt te vinden: ‘er is geen effect’. Echter, er is onzekerheid over de werkelijkheid, omdat maar een beperkt aantal gegevens verzameld wordt (steekproef). Toeval kan dus een oorzaak zijn van het verschil tussen de verzamelde gegevens en de nulhypothese. Als de kans op toeval te klein wordt, wordt toeval als onwaarschijnlijk gezien en is er sprake van een significant resultaat: de nulhypothese wordt verworpen.

Bij het toetsen kunnen twee soorten fouten worden gemaakt:

1. De nulhypothese wordt ten onrechte verworpen (een onschuldige wordt schuldig verklaard) en dat wordt een type I fout genoemd, De kans op een type I fout wordt aangegeven met α. Deze fout wordt ook wel de onbetrouwbaarheid van de toets genoemd.
2. De nulhypothese wordt ten onrechte niet verworpen (een schuldige wordt onschuldig verklaard) en dat wordt een type II fout genoemde en de kans daarop wordt aangegeven met ß.

De kans dat de nulhypothese terecht wordt verworpen (een schuldige wordt schuldig verklaard) wordt de power of het onderscheidend vermogen genoemd. De power wordt meestal aangegeven met Π (= 1- ß)



Figuur 1: Overzicht type I/II fouten en power

Bij het uitvoeren van een toets worden de volgende stappen ondernomen

1. Formuleer de nulhypothese en de alternatieve hypothese.
2. Kies de gewenste de onbetrouwbaarheid van de toets. Meestal α is 5%.
3. Het is mogelijk om een andere α te kiezen, bijv. bij risicovol onderzoek α = 1% of bij exploratief onderzoek α = 10%.
4. Bepaal de toetsingsgrootheid en de verdeling ervan onder de nulhypothese.
5. Bereken de uitkomst van de toetsingsgrootheid.
6. Bepaal de bijbehorende overschrijdingskans (p-waarde) of de kritieke waarde of het (1-α)*100% betrouwbaarheidsinterval.
7. Verwerp de nulhypothese als
   a. de p-waarde kleiner is dan α, of
   b. de toetsingsgrootheid groter is dan de kritieke waarde, of
   c. de waarde onder de nulhypothese niet in het betrouwbaarheidsinterval ligt.
8. Formuleer je conclusie, maar altijd in termen van de context van het onderzoek en nooit met statistisch jargon.

Bedenk dat bij het toetsen van de nulhypothese de gewenste power niet van belang is.

één- of tweezijdig toetsen

Wanneer je met grote stelligheid kan verwachten dat het effect maar één richting uit kan gaan (bijv. door pijnstillers kan de pijn niet erger worden) dan kun je overwegen om het effect éénzijdig te gaan toetsen. Dit kan ervoor zorgen dat je uiteindelijk minder dieren nodig hebt om je alternatieve hypothese te kunnen bevestigen.

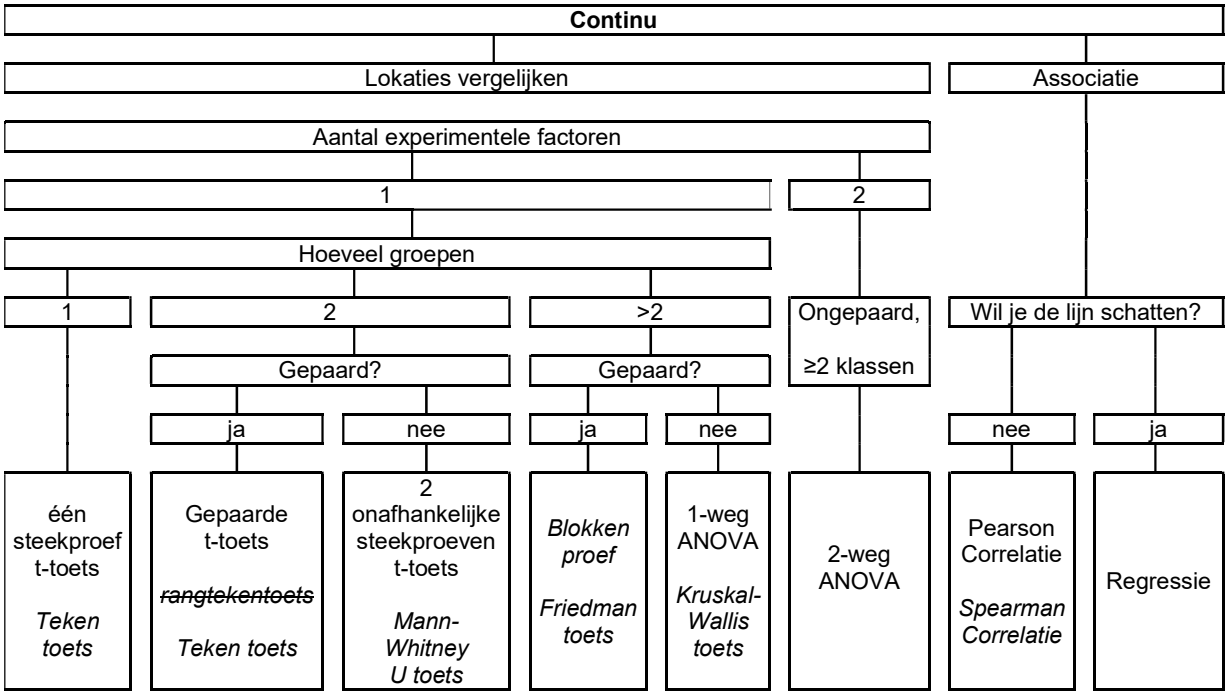
Echter als in de praktijk blijkt dat het effect juist wel de andere richting op gaat, heb je enerzijds vooraf niet goed nagedacht over de dierproef en anderzijds moet je dan ongeacht de grootte van het effect concluderen dat de nulhypothese niet kan worden verworpen. Gebruik dus alleen een éénzijdige toets als je er absoluut zeker van bent dat het effect alleen maar één richting uit kan gaan. Achteraf besluiten om toch maar een tweezijdige toets uit te voeren of, nog erger, een éénzijdige toets in de gevonden richting, is niet toegestaan.

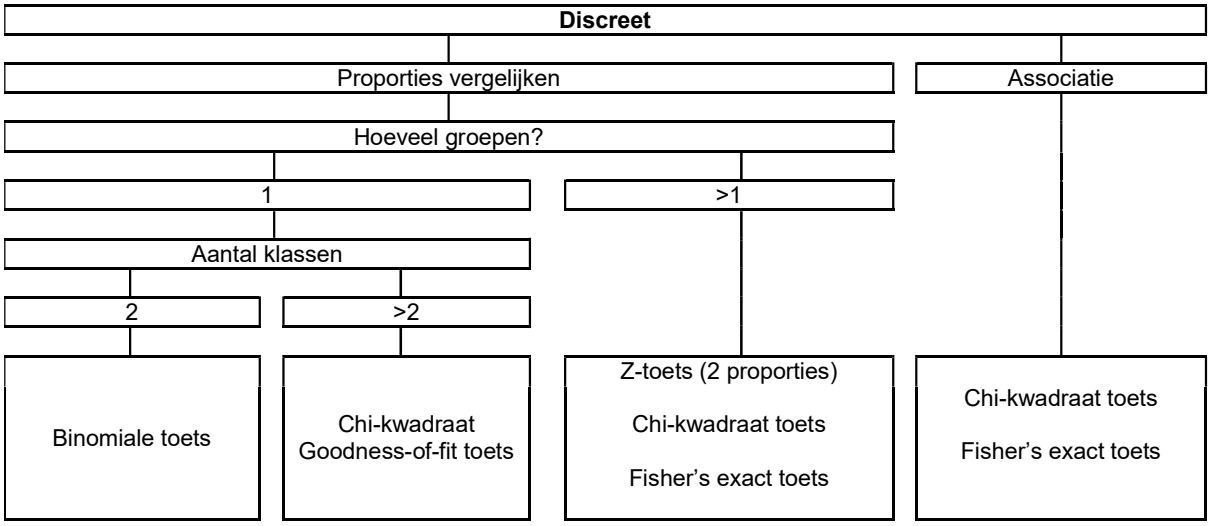
Power

Power wordt gebruikt om de kans aan te geven dat de nulhypothese terecht wordt verworpen. In het voorbeeld van de rechtszaak: een schuldige wordt inderdaad schuldig verklaart. De power wordt in een formule aangegeven met: Π=1-β, waarbij β staat voor de kans op een type II fout.

Schematisch overzicht basis statistische analyses

Afhankelijk van de onderzoeksvraag en het design van de dierproef, staat er een heel scala aan mogelijke statistische technieken tot je beschikking. In onderstaande figuur staan de standaard statistische technieken schematisch weergegeven. Bedenk echter dat het scala aan mogelijke statistische technieken veel breder is. Denk maar aan overlevingsstatistiek, herhaalde metingen, multi-way ANOVA of ANOVA voor gepaarde data, technieken voor hiërarchische structuren (Multi level), etc.





V. bepaling Steekproefgrootte

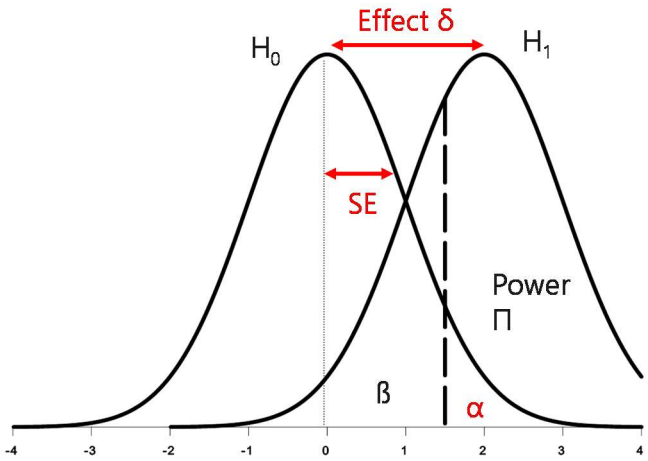
Een belangrijk onderdeel van het design is het aantal benodigde proefdieren. Het doel is om een optimum te vinden tussen niet te veel proefdieren (ethisch en economisch niet wenselijk, verplicht niet meer dieren dan noodzakelijk), maar ook weer niet te weinig (voorspellende waarde waarborgen, te weinig power). Dit optimum kun je bepalen met behulp van de poweranalyse.

Poweranalyse

Er zijn twee momenten in het onderzoek waarbij het interessant kan zijn om een poweranalyse uit te voeren:

1. Voorafgaand aan de studie. Het doel is dan om een optimum te vinden tussen een niet te grote steekproef (want dat is bijvoorbeeld ethisch en economisch niet wenselijk), en een niet te kleine steekproef (want dan zal je niet gemakkelijk een mogelijk aanwezig klinisch relevant verschil kunnen aantonen).
2. Na afloop van studie. Hier gaat het om een mogelijke te lage power herkennen en onderscheid kunnen maken tussen statistische significantie en de grootte van een (klinisch relevant) effect c.q. sterkte van samenhang. Deze is alleen interessant als er wel een klinisch relevant effect gevonden is, maar dat effect was niet statistisch significant.

Factoren die de power bepalen zijn 1) design van het onderzoek; 2) één- of tweezijdige toets; 3) onbetrouwbaarheid van de toets (α); 4) grootte van het effect (δ); 5) grootte van de spreiding (σ) en 6) omvang van de steekproef (n). In onderstaand plaatje staan deze factoren weergegeven (NB. SE is een functie van σ en n). In het algemeen kan gesteld worden dat de power groter wordt als het effect groter wordt, de onbetrouwbaarheid groter wordt, de spreiding kleiner wordt of de omvang van de steekproef groter (zie figuur 2).



Figuur 2: Overzicht van de factoren die de power bepalen

Voor de bepaling van het minimale benodigde aantal dieren heb je de volgende gegevens nodig:

1. welke statistische analyse wordt uitgevoerd (m.a.w. wat is het design);
2. één- of tweezijdig toetsen;
3. kans op een type I-fout (α);
4. gewenste power (Π);
5. minimaal relevante of te verwachten effect (δ) en
6. te verwachten spreiding c.q. standaarddeviatie (σ)

Keuze van grootte van minimaal relevant effect

Het grootste probleem is vaak het vaststellen van het minimaal relevante of te verwachten effect en de te verwachten spreiding. Voor het bepalen van het minimaal relevante effect kan gedacht worden aan de grens waarbij een effect niet meer praktisch interessant of relevant is (klinische relevantie). Om een indicatie te krijgen over de verwachte spreiding kan gekeken worden naar een pilotstudie, literatuurstudie of algemene kennis.

Voor het bepalen van de minimale omvang van de steekproef is het echter niet van belang om het effect en de spreiding zelf te weten, maar gaat het om hun verhouding. Die verhouding wordt de *effect size* (= ES) genoemd. Voor een steekproefgrootteberekening voor de ongepaarde t-toets is bijvoorbeeld de ES gedefinieerd als ES = $\delta/\sigma = (u_1-u_2)/sd$. In sommige literatuur (zoals het handboek van Van Zutphen) worden percentages gebruikt voor de aanduiding van verwacht effect en verwachte spreiding. Op zich maakt dat niets uit, maar met absolute waarden is het effect veel beter te kwantificeren in de oorspronkelijke meetschaal. Een effect van 10% zegt niets als niet bekend is waaraan die 10% gerelateerd is.

1. Van Zutphen: *Effect size* (ES) = effect / CV = $\{(u_1-u_0)/u_0\} / \{sd/u_0\}$ met u_0 is gemiddelde in controlegroep en u_1 gemiddelde in behandelde groep, sd is de (gepoolde) standaarddeviatie per groep en CV is variantiecoëfficiënt d.w.z. de sd uitgedrukt als percentage van het gemiddelde.
2. Standaard literatuur: Effect size (ES) = verschil in gemiddelden / sd = (u_1-u_0) / sd met u_0 is gemiddelde in controlegroep en u_1 gemiddelde in behandelde groep en sd is de (gepoolde) standaarddeviatie per groep.

Het kan wel eens gebeuren dat je totaal geen idee hebt wat je minimaal relevante effect is of wat de te verwachten spreiding kan zijn. In dat geval kan je je toevlucht nemen tot de Cohen's effectmaten. Het is echter aan te bevelen om ernaar te streven geen gebruik te maken van de Cohen's effectmaten.

Wat als minimale steekproefgrootte praktisch niet haalbaar?

Het komt voor dat het berekende minimaal benodigde proefdieren stuit op praktische zaken. Betreft het alleen het aantal proefdieren, dan kun je dat proberen op te lossen door de instellingen van de poweranalyse aan te passen of anderzijds je design aan te passen. Is het aantal proefdieren niet het probleem, maar kunnen niet alle metingen op 1 dag gedaan worden, dan zou je dat op meerdere dagen kunnen doen. Bedenk wel dat dit gevolgen zal hebben voor je design, je statistische analyse en je poweranalyse. Overleg met een statisticus en/of proefdierdeskundige is dan zeker op zijn plaats.

Hoe Bepaal je de steekproefgrootte?

Software

Er zijn verschillende programma's die een poweranalyse en steekproefgrootteberekening kunnen uitvoeren. De meest bekende zijn nQuery, PASS, G*Power en PS. De laatste 2 zijn gratis te downloaden van het internet. Daarnaast zijn er talloze sites die voor specifieke statistische technieken de power of de steekproefgrootte kunnen bepalen. Controleer dan wel eerst of de berekeningen van die sites kloppen.

De bepaling van het minimaal benodigde aantal dieren kan ook bepaald worden door de package-'pwr' van R.

Steekproefgrootte berekening met de hand

Bij eenvoudige statistische analyse kunnen de minimaal benodigde dieren ook met de hand worden berekend, waarbij β = 1-power.

	Input	Berekening van steekproefgrootte (n)
1 gemiddelde	sd(σ), effect(δ)	1. $n = \frac{\sigma^2}{\delta^2}\left(z_{1-\alpha/2} + z_{1-\beta}\right)^2$ 2. rond n naar boven af 3. df = n-1 4. $n = \frac{\sigma^2}{\delta^2}\left(t_{1-\alpha/2;df} + t_{1-\beta;df}\right)^2$ 5. herhaal stap 2 t/m 4 tot dat n niet meer veranderd
1 kans	P_0 en P_1 (kansen onder de nulhypothese en alternatieve hypothese)	$n \geq \frac{\left(Z_\alpha\sqrt{p_0(1-p_0)} + Z_\beta\sqrt{p_1(1-p_1)}\right)^2}{(p_1 - p_0)^2}$
2 gemiddelden (ongepaard)	sd(σ), effect(δ)	1. $n = 2\frac{\sigma^2}{\delta^2}\left(z_{1-\alpha/2} + z_{1-\beta}\right)^2$ 2. rond n naar boven af 3. df = n-1 4. $n = 2\frac{\sigma^2}{\delta^2}\left(t_{1-\alpha/2;df} + t_{1-\beta;df}\right)^2$ 5. herhaal stap 2 t/m 4 tot dat n niet meer veranderd
2 kansen (ongepaard)	p_C = de kans In de controle groep, p_E = de kans in de behandelde groep, $\delta_0 = p_E - p_C$	$n \geq \frac{p_C(1-p_C)+p_E(1-p_E)}{\delta_0^2}\left(Z_\alpha + Z_\beta\right)^2$

Met z_p de z-waarde onder de standaard normaal verdeling waarbij Pr(Z<z-waarde)=p en analoog $t_{p;df}$ de t-waarde onder de t-verdeling met vrijheidsgraden df Pr(T<t-waarde)=p. Bijvoorbeeld als α=5% dan is $z_{1-.05/2}$ = 1.96.

Power analyse bij 1-weg-ANOVA-theorie

Voor het vergelijken van de gemiddelden van k condities is de meest optimale statistische analyse de 1-weg ANOVA. Het principe van de 1-way ANOVA is dat getoetst wordt of er überhaupt wel een verschil is tussen de k condities. De nulhypothese is dan 'er is geen effect tussen de k condities'. Als deze nulhypothese wordt verworpen kan geconcludeerd worden dat bij tenminste 2 condities de gemiddelden van elkaar verschillen. Een interessante logische vraag is dan bij welke 2? Om die vraag te beantwoorden zijn er post-hoc-toetsen.

De meeste post- hoc-toetsen zijn gericht op het paarsgewijs vergelijken van condities, maar het is ook mogelijk om gemiddelden van subsets (al dan niet met wegingsfactoren) van condities met elkaar te vergelijken. In dit laatste geval wordt er eerder gesproken over contrasten dan over post-hoc-toetsen, maar eigenlijk zijn het speciale soorten van post-hoc-toetsen. De meeste post-hoc-toetsen zijn ruwweg te beschrijven als aangepaste versies van de ongepaarde t-toets.

Bij k condities zijn er maximaal k*(k-1)/2 mogelijke paarsgewijze vergelijkingen mogelijk, die ieder een onbetrouwbaarheid hebben van 5% (α). Dat betekent dat de kans dat bij tenminste 1 van al die paarsgewijze vergelijkingen de nulhypothese ten onrechte wordt verworpen groter wordt dan 5%. Als je alle mogelijke paarsgewijze vergelijkingen uitvoert dan is die kans maximaal 5*k*(k-1)/2. De werkelijke kans hangt mede af van de afhankelijkheid tussen de posthoc toetsen en de gebruikte posthoc toets. Om toch de overall onbetrouwbaarheid op 5% te houden worden de onbetrouwbaarheden bij de post-hoc toetsen aangepast, d.w.z. verlaagt. Echter een verlaging van de kans op een type I fout zorgt er voor dat de kans op een type II fout groter wordt, en dus de power lager wordt (zie figuur 1).

In de loop der jaren zijn er tientallen post-hoc-toetsen ontwikkeld die ernaar streefden om de overall onbetrouwbaarheid op 5% te houden met zo min mogelijk verlies van power. Bijna alle post-hoc-toetsen,

behalve de LSD-toets, worden op de een of andere manier gecorrigeerd voor het aantal post-hoc-toetsen. Door toenemende kennis is echter ook gebleken dat sommige van die post-hoc-toetsen nog wel beschikbaar zijn in allerlei statistische software, maar beter niet gebruikt kunnen worden. In de dagelijkse praktijk wordt meestal de Tukey (in geval alle paarsgewijze vergelijkingen), Dunnett (alleen vergelijking t.o.v. een referentiegroep) of Bonferroni (in geval een beperkte selectie van paarsgewijze vergelijkingen en/of contrasten) gebruikt. Natuurlijk wordt de keuze ook bepaald door welke post-hoc-toets in het onderzoeksveld gebruikelijk is.

De post-hoc-toetsen bij een ANOVA-analyse kan op verschillende manieren worden benaderd:

1. Theorie: precies volgens de theorie. Zelfs de onbetrouwbaarheid wordt geheel volgens de theorie aangepast: de aangepaste onbetrouwbaarheid wordt α/(k*(k-1)7/2). Dit kan leiden tot zeer lage aangepaste onbetrouwbaarheden.
2. Praktisch: precies volgens de theorie, maar met inachtname van het feit dat de onbetrouwbaarheid aangepast moet worden, en tegelijk een ondergrens voor de onbetrouwbaarheid hanterend. Bijv.: ongeacht het aantal post-hoc-toetsen een aangepaste onbetrouwbaarheid hanteren van 1%.
3. Geen correctie: zonder aanpassing van de onbetrouwbaarheid, omdat de post-hoc-toetsen maar bijzaak zijn.
4. Geen overall ANOVA: de overall ANOVA wordt niet uitgevoerd. Je richt je meteen volledig op de post-hoc-toetsen, en gebruikt dan zelfs de gewone ongepaarde t-toets, met wel of geen correctie van de onbetrouwbaarheid.

De praktische manier ligt erg voor de hand, zeker als veel condities met elkaar worden vergeleken. Let wel: iedere benadering heeft zijn eigen gevolgen voor de resultaten van de toetsen en voor de power.

Poweranalyse bij 1-weg ANOVA in de praktijk

De meeste artikelen die gaan over steekproefgrootteberekening, gaan over de primaire onderzoeksvraag ‘Is er een effect tussen de k condities’, m.a.w. de overall ANOVA. Maar wat als de post-hoc-toetsen de primaire onderzoeksvragen zijn?

Het is bekend dat de meeste post-hoc-toetsen de overall onbetrouwbaarheid op 5% houden, maar dat er wel verlies van power optreedt. Om de gewenste power voor de post-hoc-toetsen te behouden moet bij de bepaling van minimale steekproefgrootte de onbetrouwbaarheid (α) ook aangepast worden, zodat de steekproefgrootte iets groter wordt.

> ***Voorbeeld***: Onderzoekers willen een *effect size* aantonen van ES = 1.5 (dit was de kleinste verwachte *effect size* van de 4 specifieke post hoc toetsen) met een power van 90%. Voor hun onderzoek zijn ze alleen geïnteresseerd in 4 specifieke post-hoc-toetsen. Als post-hoc-toets kiezen ze dan voor de Bonferroni methode met α=5/4%=1.25%.
> Als ze bij het bepalen van de minimale steekproefgrootte de onbetrouwbaarheid niet hadden aangepast dan was n=11. Bij de analyse echter zou dit bij dezelfde *effect size* hebben geleid tot een power van 77% en dus een verlies van 13%. Hadden ze bij de bepaling van de minimale steekproefgroottebepaling ook gekozen voor de aangepaste onbetrouwbaarheid α=1.25% dan was n=15 en hadden ze geen verlies van power.

Samenvattend heeft de onderzoeker drie opties:

1. Indien de primaire onderzoeksvraag 'is er een verschil tussen de k condities' is, dan zou de steekproefgroottebepaling gebaseerd moeten worden op de overall ANOVA-analyse. De post-hoc-toetsen zijn dan slechts secundaire onderzoeksvragen of worden beschouwd als een leuk extraatje dat interessant is voor de toets, maar niet voor de power.
2. Echter, zijn je primaire onderzoeksvragen gericht op het 'verschil tussen bepaalde condities' dan kan de minimale steekproefgroottebepaling gebaseerd worden op een ongepaarde t-toets met als spreiding de gepoolde spreiding van alle condities en
    a. zonder een aangepaste onbetrouwbaarheid
    b. met een aangepaste onbetrouwbaarheid

Let wel: we gaan dan uit van de Bonferroni post-hoc-toets.

Als we dit toepassen op bovenstaand voorbeeld dan is de minimale steekproefgrootte in geval van optie 1 n=3, voor optie 2a n=11 en voor optie 2b n=15. De keuze tussen optie 2a en optie 2b is een afweging tussen grootte van de steekproef en verlies aan power. Een lastige afweging, waarbij het verlies aan power ook nog eens bepaald wordt door de keuze van het aantal post-hoc-toetsen, maar ook de werkelijke *effect size*.

VI. Ten slotte

Deze notitie is geschreven voor de gemiddelde onderzoeker met een basiskennis van de statistiek. De achterliggende theorieën zijn terug te vinden in basisstatistiekboeken, Wikipedia of onderstaande referenties.

Meer Informatie:

Websites

Over de bovenstaande onderwerpen kun je veel lezen op de volgende website: www.3Rs-reduction.co.uk. Hier kun je je kennis ook toetsen met een zelftest.

Literatuur

Amor, S., Baker, D., (2012) Checklist for reporting and reviewing studies of experimental animal models of multiple sclerosis and related disorders. Mult Scler Relat Disord. 1(3)
Een artikel met daarin een volledige lijst van informatie die je zou moeten geven wanneer je je resultaten presenteert in een wetenschappelijk tijdschrift. De focus is op MS, maar de informatie is ook geschikt voor andere vakgebieden.

Lara-Pezzi, E et al., (2015) Guidelines for Translational Research in Heart Failure J. of Cardiovasc. Trans. Res. 8(1)
Een artikel met daarin vooral de focus op de translatie van modellen voor hartfalen in dieren en effectieve suggesties voor het ontwerpen van dergelijke studies De focus is op hartfalen, maar de informatie is ook geschikt voor andere vakgebieden.

Steward, O., Balice-Gordon, R. (2014) Rigor or mortis: best practices for preclinical research in neuroscience. Neuron. 84(3)
In dit artikel worden best practices op het gebied van experimental design en statistiek in preklinische studies op het gebied van de neurologische en psychiatrische aandoeningen besproken. Ook is er aandacht voor datamanagement. De focus van het artikel is op neurologische en psychiatrische aandoeningen, maar de informatie is ook geschikt voor andere vakgebieden.

Festing, M. F. W., Altman, D.G., (2002) Guidelines for the Design and Statistical Analysis of Experiments Using Laboratory Animals. *ILAR 43(4)*
Dit artikel helpt je stap voor stap om je vraagstelling te kunnen beantwoorden met verschillende typen experimenten. Het biedt manieren om fouten te voorkomen en zinvolle data te verkrijgen. Het is speciaal geschreven voor het gebruik van dieren in onderzoek en legt daarbij de nadruk op de 3V's en goede statistische analyse.

Aban, I.B., George, B., (2015) Statistical considerations for preclinical studies, Exp. Neurol. 270
Dit artikel bespreekt statistische begrippen met als doel een betere kwaliteit van dierstudies. Dit artikel is speciaal geschreven voor het gebruik van dieren in preklinische studies zodat de data geschikt is als voorbereiding op de klinische fase van onderzoek.

Hirst, J.A., et al. (2014) The Need for Randomization in Animal Trials: An Overview of Systematic Reviews. PLoS ONE 9(6)
In dit artikel wordt aan de hand van een meta-analyse aangetoond hoe belangrijk het is om dierstudies gerandomiseerd, met het blind toewijzing van interventies en geblindeerd uit te voeren.

Tweel, I. van der (2006) Sample size determination. Intern Report nr 4
(http://portal.juliuscentrum.nl/Portals/2/Disciplines/Biostatistics/SAMPLE%20SIZE%20DETERMINATION_electronic%20version.pdf)
In dit report wordt uitleg gegeven over de meest simpele steekproefgrootte berekening.

Bate, S.T. & R.A. Clark, (2014) The design and statistical analysis of animal experiments, Cambridga University Press
Dit boek bespreekt allerlei aspecten met betrekking tot het opzetten en analyseren van dierproeven.