09

Proefopzet en statistiek

Hein A. van Lith, Universiteit Utrecht
Franz Josef van der Staay, Universiteit Utrecht

1 Ontstaan en formulering van de onderzoeksvraag

Dierexperimenteel onderzoek is vrijwel altijd deel van een project dat wetenschappelijke vragen wil beantwoorden en wetenschappelijke hypothesen wil toetsen. We kunnen twee wegen bewandelen om wetenschappelijke gegevens te verzamelen:

- via beschrijvend onderzoek, dat steunt op systematische, nauwkeurige observaties en tellingen;
- via experimenteel onderzoek, dat steunt op de resultaten bij groepen mensen, dieren of cellen die aan verschillende behandelingen werden toegewezen, of die een verschillend genotype hebben, zoals bij vergelijken van inteeltstammen.

Vaak gaat beschrijvend onderzoek vooraf aan experimenteel onderzoek. Het beschrijvende onderzoek kan al aanwijzingen opleveren voor mogelijke oorzaak en gevolg verbanden. Zo heeft men gevonden dat mensen die planten of vee behandelden met organofosforverbindingen (zogenoemde organofosfaten en organothiofosfaten) tegen insectenvraat en bloedzuigende insecten, chronische ziektesymptomen ontwikkelden en een opmerkelijke daling van de leer- en geheugencapaciteit vertoonden. Bij het sproeien van organo(thio)fosfaten over planten en bij het onderdompelen van schapen in een waterbad met organo(thio)fosfaten, ademden zij deze stoffen in.

Deze observaties leidden tot de hypothese dat chronische blootstelling aan organo(thio)fosfaten schadelijk kan zijn voor mensen en hun geestelijke vermogens. Om deze hypothese te toetsen, kunnen dierexperimenten worden uitgevoerd. In zo'n studie stellen we bijvoorbeeld knaagdieren langdurig bloot aan verschillende concentraties organo(thio)fosfaten, en een andere groep niet; de controlegroep. Na enkele maanden vergelijken we de leer- en geheugenprestaties van de blootgestelde groepen met die van de controlegroep.

Dierexperimenteel onderzoek kent een aantal fases die elkaar logisch opvolgen:

- de stand van zaken inventariseren op het onderzoeksterrein via een grondige literatuurstudie of een systematische review;
- de precieze vraagstellingen formuleren;
- het onderzoek ontwerpen en het proefplan schrijven;
- het onderzoek uitvoeren;
- de resultaten analyseren;
- de resultaten en rapportage interpreteren;
- eventueel de resultaten in replicatiestudies bevestigen.

Voor de voorbereiding van dierproeven raden we de PREPARE (*Planning Research and Experimental Procedures on Animals*)-richtlijnen aan (zie: https://vimeo.com/358069203 en https://journals.sagepub.com/doi/full/10.1177/0023677217724823).

2 Literatuurstudie

Voorafgaand aan elk onderzoek op of met dieren moet je gedegen literatuuronderzoek doen. In de wetenschappelijke literatuur is een schat aan informatie over bijna elk aspect van het beoogde experiment te vinden. Publicaties hebben als doel wetenschappelijke inzichten met anderen te delen. Op de informatie uit relevante publicaties kun je een onderzoek volgens de huidige stand van de kennis plannen en uitvoeren. Bovendien moet literatuuronderzoek ervoor zorgen dat je valkuilen vermijdt waar anderen tijdens hun onderzoek in terecht zijn gekomen. Helaas wordt dit soort informatie lang niet altijd gepubliceerd. Het kan dan ook zinvol zijn een of meer laboratoria te bezoeken waar al onderzoek op het beoogde gebied is gedaan en waar de beoogde onderzoekstechnieken werden toegepast, en expliciet te vragen naar mogelijke technische problemen en naar de oplossingen die de onderzoekers vonden.

2.1 Oriënterend lezen

Het verdient aanbeveling een onderwerp eerst in een breder kader te bestuderen. Overzichtsartikelen (reviews) bieden vaak een goede mogelijkheid om met een onderzoeksvraag of onderzoeksrichting bekend te raken en snel een algemeen overzicht te krijgen van de stand van zaken op het onderzoeksgebied. Bovendien bevat de literatuurlijst van een overzichtsartikel meestal verwijzingen naar belangrijke originele publicaties. Deze kun je dan raadplegen om de diepte in te gaan.

2.2 Grondig en kritisch lezen

Een goed criterium om de waarde van een publicatie te beoordelen, is door te kijken naar de kwaliteit van de beschrijving van methoden en technieken. Deze moet alle informatie bevatten die nodig is om een experiment te kunnen herhalen. Ontbreken essentiële details, dan kun je een publicatie niet op waarde inschatten. Publicaties waarin de resultaten van dierproeven worden weergegeven en waarbij de wetenschappers de zogenaamde ARRIVE (*Animal Research: Reporting of In Vivo Experiments*)-richtlijnen hebben gehanteerd, zijn doorgaans van goede kwaliteit (zie: https://www.nc3rs.org.uk/arrive-guidelines en https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.1000412;). Soms wordt voor details van de toegepaste methoden en technieken verwezen naar een eerdere publicatie of naar aanvullende informatie die met een artikel (online) wordt gepubliceerd (vaak *supplemental information* genoemd). Het is dan noodzakelijk ook deze informatie te raadplegen.

Publicaties en vaktijdschriften zijn van zeer uiteenlopend wetenschappelijk niveau. Vanzelfsprekend verdient een hoogwaardige publicatie meer aandacht dan een minderwaardige publicatie. Het aantal citaties van een oudere publicatie in recenter onderzoek is vaak een goede indicatie voor de waarde van een publicatie. Uiteraard is de kans dat zeer recente publicaties al worden geciteerd erg klein. Ook zal over nieuwe onderzoeksrichtingen nog niet veel zijn gepubliceerd.

2.3 Systematische review

In een systematische review verzamel je een groot aantal wetenschappelijke publicaties over een specifieke onderzoeksvraagstelling. De publicaties haal je uit tenminste twee grote literatuurdatabases, bijvoorbeeld *PubMed* en *Scopus*. Andere literatuurdatabases zijn *EMBASE* en *Web of Science*. De selectie van de publicaties die je in het systematische review opneemt, moet aan een aantal strikte, vooraf vastgestelde inclusie- en exclusiecriteria voldoen. Hierdoor voorkom je dat je onvolledige studies of studies van onvoldoende kwaliteit meeneemt. Aan de hand van een of meer onderzoeksvragen evalueer je de geselecteerde publicaties systematisch, waarbij je gebruikmaakt van een gestructureerde methodologie. Deze methodologie is in een aantal publicaties over het uitvoeren van systematische reviews uitvoerig beschreven, bijvoorbeeld door SYRCLE, *Systematic Review Center for Laboratory Animal Experimentation* (https://journals.sagepub.com/doi/full/10.1258/la.2011.011087).

Globaal maak je een systematische review volgens deze stappen:

- formuleer de vraagstelling(en);
- stel een protocol op. Hierin leg je vast volgens welke zoekbegrippen in welke literatuurdatabases je zoekt. Het protocol bevat de volgende punten:
  - een beschrijving van de achtergrond van de vraagstelling(en), de achtergrondinformatie;
  - een beschrijving van de doelstelling(en) van de review;
  - een lijst van de inclusie- en exclusiecriteria voor de publicaties die je meeneemt;
  - een beschrijving van de te volgen zoekstrategie, dat wil zeggen welke literatuurdatabases en zoektermen je gebruikt;
  - een beschrijving van de analysemethode(n) en statistische modellen die je gebruikt.
- selecteer de publicaties die aan de gestelde criteria voldoen. Presenteer en beoordeel deze literatuur systematisch en analyseer de uitkomsten aan de hand van de vooraf geformuleerde vraagstelling(en) en hypothesen;
- schrijf de systematische review.

Systematische reviews kunnen nieuwe inzichten bieden en tekortkomingen in eerder onderzoek – en de oplossing ervan – aan het licht brengen. Door rekening te houden met de uitkomsten van systematisch onderzoek kun je de kwaliteit en de relevantie van nieuw te plannen onderzoek verhogen.

2.4 Preregistratie

Naar analogie van publieke preregistraties van klinische trials is een systeem opgezet voor publieke preregistratie van dierproeven met een volledig onderzoeksprotocol en data-analyseplan. Dit systeem vind je op https://www.preclinicaltrials.eu/. Op deze website kunnen onderzoekers hun studie met proefdieren gratis aanmelden en registreren. Iedereen met een account krijgt toegang tot de database met proefdierstudies. In de database kun je gericht zoeken naar specifieke dierproeven en hun resultaten. Onderzoekers kunnen eenvoudig zien welke dierproeven al gedaan zijn en met welke resultaten, ook als die resultaten niet gepubliceerd zijn. Hierdoor kun je onnodige herhaling van dierproeven tegengaan.

3 Wat zijn diermodellen?

Dierexperimenteel onderzoek wordt meestal verricht met diermodellen. Een diermodel definiëren we zo: een diermodel met biologische en/of klinische relevantie is een levend, dierlijk organisme dat wordt ingezet om fenomenen (bijvoorbeeld fysiologische processen, afwijkend gedrag) onder gecontroleerde omstandigheden te bestuderen. Het uiteindelijke doel is inzicht te verkrijgen en voorspellingen mogelijk te maken over deze fenomenen bij mensen of andere species dan die waarmee het onderzoek werd verricht, of bij dezelfde species onder andere omstandigheden dan die waaronder het onderzoek werd uitgevoerd.

Deze definitie bevat verschillende sleutelcomponenten:

- de biologische en/of klinische relevantie. Dit is een van de belangrijkste thema's voor dierexperimenteel werkende onderzoekers én een hoofdpunt van kritiek van mensen en actiegroepen die zich terughoudend of afwijzend opstellen ten opzichte van de waarde en toelaatbaarheid van dierexperimenteel onderzoek. Dierexperimenteel onderzoek moet erop gericht zijn wetenschappelijk gefundeerde antwoorden op vragen te vinden, bijvoorbeeld over de oorzaken, het verloop en de mogelijke behandeling van ziekten bij mens en dier;
- levend, dierlijk organisme benadrukt dat we in diermodellen omgaan met levende wezens. Anders dan bij levenloze modellen zoals computersimulaties en mechanische modellen, is werken met diermodellen onlosmakelijk verbonden met een aantal (ethische) waarden. Deze zijn voor een gedeelte verwoord in de drie V's (Russell & Burch, 1959 – https://caat.jhsph.edu/principles/the-principles-of-humane-experimental-technique) die we moeten respecteren. Niet alle experimenten met levende, dierlijke organismen zijn volgens de Wet op de dierproeven ook daadwerkelijk dierproeven;
- gecontroleerde omstandigheden hebben betrekking op de controle van storende variabelen (ook interveniërende of tussenkomende variabelen genoemd) en op standaardisatie van de huisvestings- en onderzoeksomstandigheden. Standaardisatie verhoogt de betrouwbaarheid, de reproduceerbaarheid en de vergelijkbaarheid van resultaten. Standaardisatie van een experiment en de controle van storende (interveniërende) variabelen verhogen de interne validiteit van het experiment;
- inzicht verkrijgen en voorspellingen mogelijk maken over deze fenomenen bij mensen of andere species dan die waarmee het onderzoek werd verricht, heeft betrekking op de generaliseerbaarheid van resultaten, ook wel externe validiteit genoemd. We generaliseren naar de mens of andere species op basis van het concept over evolutie van Darwin, het analogieprincipe. Dit ligt ten grondslag aan de vergelijking tussen species, bijvoorbeeld met betrekking tot neuro-anatomische structuren en hun functies;
- andere omstandigheden dan die waaronder het onderzoek werd uitgevoerd verwijst naar de generaliseerbaarheid van de resultaten (de externe validiteit). Van belang is dat de resultaten ook op situaties buiten het laboratorium van toepassing zijn.

3.1 Natuurlijke en gemaakte diermodellen

We kunnen drie klassen diermodellen onderscheiden:

- modellen die gebaseerd zijn op het 'normale' dier;
- modellen die gebruikmaken van natuurlijk voorkomende variaties (deviaties, afwijkingen);
- modellen waarin afwijkingen experimenteel zijn geïnduceerd (zie tabel 1).

Tabel 1. *Klassen van diermodellen.*

'Normale' dieren	Natuurlijk optredende afwijkingen	Experimenteel geïnduceerde afwijkingen
Dieren zonder enige observeerbare/meetbare afwijking	Genetische lijnen	Dieren met specifieke laesies
	Geselecteerde extremen uit een populatie, zoals dieren met een extreem hoog of laag lichaamsgewicht, met een hoge of lage agressie, of met een hoge of lage resistentie tegen ziekteverwekkers, *et cetera*	Dieren met specifieke, experimenteel geïnduceerde afwijkingen of ziekten, zoals veroorzaakt door toediening van farmaca, experimentele hypoxie, besmetting met ziekteverwekkers, *et cetera*
	Oude dieren (bijvoorbeeld voor onderzoek naar ziekten of afwijkingen die leeftijdsafhankelijk zijn)	Genetisch gemodificeerde dieren (*knock-outs** en transgene dieren)

* *onder knock-outs vallen ook de via CRISPR/Cas9 technologie gegenereerde knock-outs.*

Normale dieren gebruiken we bijvoorbeeld in onderzoek naar de teratologische en toxicologische risico's van een nieuw medicijn, of om de farmacokinetiek van een medicijn te bepalen. Normale dieren gebruiken we ook in onderzoek naar de effecten van huisvesting, voeding en verzorging: enerzijds voor dierenwelzijnsonderzoek en anderzijds om managementsystemen in de veehouderij te optimaliseren.

Maar meestal modelleert een diermodel de symptomatologie, of andere aspecten van een ziekte of afwijking bij de mens of een andere diersoort. Hiervoor staan diermodellen ter beschikking waarin de symptomen spontaan optreden, of waarin deze experimenteel worden geïnduceerd. Bijvoorbeeld spontane tumoren bij de hond als kankermodel voor de mens, of lepraonderzoek aan de hand van het negenbandig gordeldier dat is geïnfecteerd met *Mycobacterium leprae*.

4 Doel van de dierproef

Dierproeven worden voor veel uiteenlopende doelen uitgevoerd. Zo winnen we antistoffen of hormonen als geneesmiddel uit dieren die al dan niet werden behandeld. Hiervoor gebruiken we bij voorkeur species die een voldoende grote opbrengst van de stoffen in de grootst mogelijke zuiverheid garanderen. Grote aantallen dieren zetten we in om substanties te screenen die nieuwe geneesmiddelen kunnen opleveren. Voor de kwantitatieve bepaling van biologisch werkzame stoffen, door middel van toediening van deze stoffen aan levende organismen onder standaard omstandigheden, worden ook grote aantallen dieren gebruikt; we noemen dit *in vivo* bioassays.

Drug screening en bioassays zetten we in om potentieel effectieve therapeutica te vinden, zonder dat ze een model zijn voor symptomen van de ziekte die met deze therapeutica zal worden behandeld.

Meestal gebruiken we diermodellen om fundamentele wetenschappelijke vragen te beantwoorden, zoals: 'welke processen liggen ten grondslag aan een bepaalde ziekte?' Diermodellen gebruiken we ook om nieuwe geneesmiddelen en therapieën voor mens of dier te vinden, te karakteriseren en te optimaliseren.

Het doel van veel diermodellen is dan ook:
- (dieper) wetenschappelijk inzicht krijgen in fenomenen zoals (patho)fysiologische processen en afwijkend gedrag;
- de effecten bepalen van potentiële geneesmiddelen en andere therapeutische benaderingen. Het gaat hier om de gewenste therapeutische werking, maar ook om de mogelijke teratologische en toxicologische risico's en het verslavingspotentieel.

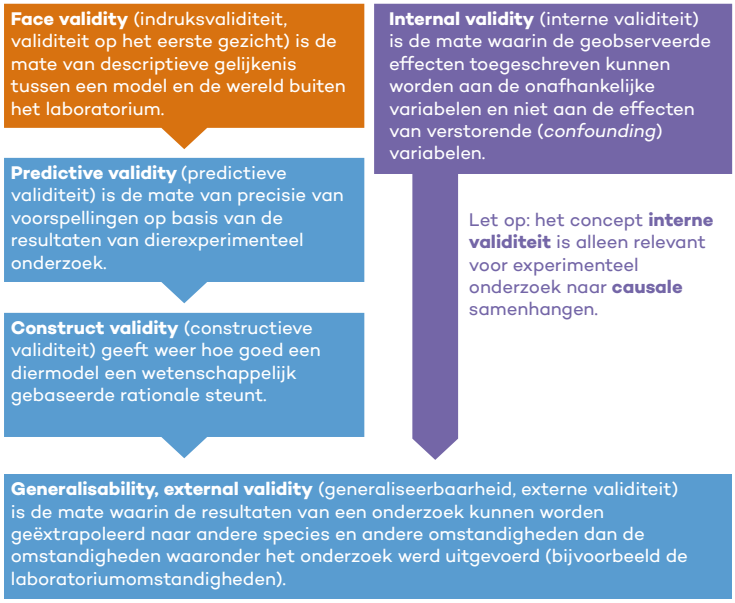
Het doel van de dierproef bepaalt de keuze van de modelspecies en de methoden en technieken die in een proef worden toegepast.

4.1 Diermodellen valideren

Door validering stellen we de waarde vast van het diermodel, ofwel de biologische en klinische relevantie. In een valideringsproces wordt beoordeeld of het model aan de criteria van een aantal validiteiten voldoet.

Validiteit wordt gedefinieerd als: de overeenkomst tussen een testscore of meting en het fenomeen dat je beoogt te meten.

Het meest gebruikelijke onderscheid is dat tussen *face validity* (indruksvaliditeit; validiteit op het eerste gezicht), *predictive validity* (predictieve of voorspellende validiteit), *construct validity* (constructieve of begripsvaliditeit) en *external validity* (externe validiteit; zie afbeelding 1). Als het gaat om dierexperimenteel onderzoek dat gericht is op het bestuderen van causale verbanden, passen we ook de tweedeling *internal validity – external validity* toe.



Afbeelding 1. *De gebruikelijkste hiërarchie is face-, predictive- en construct validity. Er wordt bij hypothesetoetsende experimenten ook onderscheid gemaakt tussen interne en externe validiteit. Externe validiteit of generaliseerbaarheid kan ook aan de meest gebruikte classificatie (de linkerkolom) worden toegevoegd.*

Face validity zegt iets over de mate van de ogenschijnlijke gelijkenis tussen het model en het fenomeen dat wordt gemodelleerd; het wordt ook wel fenomenologische validiteit genoemd. Niet iedereen beschouwt *face validity* als een belangrijk criterium voor de validatie van een model en is daarom in afbeelding 1 als oranje vak weergegeven.

Predictive validity zegt iets over de mate waarin de resultaten van het model een voorspellende waarde hebben voor het bestudeerde fenomeen dat het modelleert.

Construct validity geeft de mate aan waarin een model op een goede wetenschappelijke basis staat.

External validity geeft aan in hoeverre de inzichten verkregen uit onderzoek met diermodellen kunnen worden geëxtrapoleerd naar andere species (meestal de mens) en andere omstandigheden (buiten het dierexperimenteel laboratorium). Generaliseerbaarheid (extrapolatie) is het uiteindelijke doel van een dierproef.

4.2 Generaliseerbaarheid (extrapolatie)

De kennis over het verband tussen behandeling door de onderzoeker (de onafhankelijke variabele) en de gemeten waarden (de afhankelijke variabelen) bepalen in sterke mate de kwaliteit van de extrapolatie van dierexperimenteel onderzoek. *Laesie* (beschadiging) versus *shamlaesie* is zo'n onafhankelijke variabele. Afsluiting van een hersenarterie bij het varken is een voorbeeld van een *laesie* en de grootte van een herseninfarct is een voorbeeld van een afhankelijke variabele. Hoe groter de afstand tussen het modeldier en de species die het modelleert, des te geringer de mogelijkheid tot extrapolatie. Ook naarmate er meer ruis zit in de meetwaarden neemt de mogelijkheid tot extrapolatie af. Welke modeldierspecies je kiest, is dus afhankelijk van de vraagstelling en de beoogde 'vertaling'. Operatietechnieken die bij de mens toegepast gaan worden, ontwikkelen en onderzoeken we niet in knaagdieren, maar in een species die meer op de mens lijkt, bijvoorbeeld qua grootte, zoals het varken.

4.3 Grootte, verkrijgbaarheid, kosten

Kies een diermodel altijd met het oog op het doel van het onderzoek. Als de resultaten van een dierproef direct geëxtrapoleerd moeten worden naar de mens, dan kies je een diersoort met een zo klein mogelijke extrapolatieafstand (ook wel translationele afstand genoemd). Dit is de afstand tussen het modeldier en de species die het modelleert. Maar er zijn beperkende factoren bij de keuze van een diermodel. Als er geen geschikt diermodel of een acceptabel alternatief bestaat, dan moet een nieuw diermodel ontwikkeld en gevalideerd worden, wat een erg tijdrovend en kostbaar proces kan zijn. Pas daarna kan het model worden ingezet voor de beantwoording van de vraagstelling.

Diermodelspecies met de kleinste extrapolatieafstand tot de mens zijn primaten (halfapen en apen). Deze orde van zoogdieren staan ook fylogenetisch gezien dicht bij de mens. Er zijn erg weinig primaten beschikbaar voor dieronderzoek. Zo is het sinds 2003 in Nederland verboden om mensapen als proefdier te gebruiken. Bovendien zal bij de ethische toetsing van het onderzoek bijzonder kritisch worden gekeken naar de redenen om juist primaten te gebruiken. Dit geldt ook voor onderzoek met andere grotere zoogdieren zoals de hond.

Werken met grotere diersoorten – bijvoorbeeld resusaap, varken, hond en kat – stelt bijzondere eisen aan de voorzieningen (ruimte voor huisvesting) en de kennis en kunde van dierverzorgers en biotechnici. Lang niet elk dierverblijf en laboratorium is geschikt voor grotere diermodelspecies. De kosten van onderzoek met grotere species zijn aanzienlijk hoger dan die bij het werken met kleinere species, zoals knaagdieren.

5 Keuze van de experimentele eenheid

Belangrijk bij het plannen van een dierproef is de keuze van de experimentele eenheid. Elke experimentele eenheid moet aan verschillende behandelingscondities toegewezen kunnen worden. De experimentele eenheid kan het proefdier zijn, een groep dieren of delen van dieren. Hieronder een paar voorbeelden ter verduidelijking.

In een onderzoek bestuderen we de vraag welke effecten langdurige stress tijdens de zwangerschap heeft op de gezondheid, de ontwikkeling en het gedrag van de nakomelingen. We kiezen het varken als proefdier. We verdelen drachtige zeugen in drie groepen:

- een controlegroep;
- een groep die we behandelen met een lage dosis van het stresshormoon cortisol;
- een groep die we behandelen met een hoge dosis cortisol.

Omdat we de cortisol per injectie toedienen, krijgt de controlegroep een injectie met fysiologische zoutoplossing. We willen namelijk de stress van de injectie in alle groepen zeugen gelijk houden.

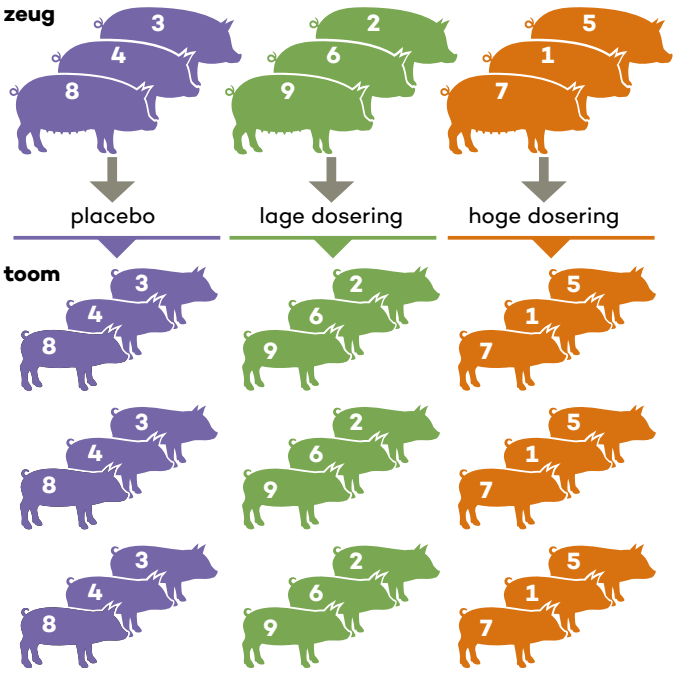
Het onderzoek richt zich op de effecten die de behandeling van de zeug op de nakomelingen heeft. Die effecten zijn voor alle nakomelingen van een bepaalde zeug gelijk. De hele worp biggen (de toom) is dus de experimentele eenheid, want je kunt geen individueel dier binnen de toom aan een van de andere experimentele condities toewijzen. In het voorbeeld in afbeelding 2a werden 9 zeugen over de 3 groepen (controle, lage dosering en hoge dosering cortisol) verdeeld; dus 3 zeugen per groep. Uit iedere worp werden 3 biggen aselect (willekeurig) aan een behandelingsgroep toegewezen. Nu is iedere groep van 3 nakomelingen uit dezelfde toom de experimentele eenheid.

Je zou ook het effect van cortisol op het gedrag van varkens kunnen onderzoeken. Voor dit onderzoek worden 9 varkens per toeval aan een van de 3 behandelingen (controle, lage dosering en hoge dosering cortisol) toegewezen (zie afbeelding 2b). Hier zijn de individuele dieren de experimentele eenheid.

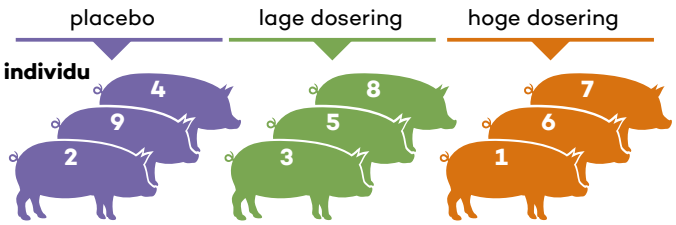
Cortisol heeft een genezend effect op een aantal huidaandoeningen. De huid van een varken waarbij experimenteel een huidaandoening is opgewekt, wordt onderverdeeld in velden die met placebo, een lage of een hoge dosis cortisol worden behandeld (zie afbeelding 2c). De verdeling van de velden wordt per toeval bepaald. In het voorbeeld zijn alle behandelingen in vier verschillende velden toegepast. Delen van het dier zijn de experimentele eenheid in dit onderzoek, in ons voorbeeld elk van de velden die een van de 3 behandelingen ondergaan.

Keuze van de experimentele eenheid

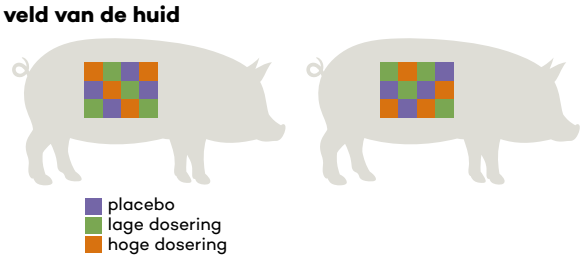
De experimentele eenheid is bij **a** de toom, bij **b** het individu en bij **c** een veld van de huid.



Afbeelding 2a. *Negen zeugen worden van 1 tot 9 genummerd, waarna een toevalsvolgorde wordt bepaald: 8-4-3-9-6-2-7-1-5. De eerste drie nummers (8-4-3) worden een placeboconditie toegeschreven, de tweede serie (9-6-2) aan de lage dosering en de derde serie (7-1-5) aan de hoge dosering.*



Afbeelding 2b. *De dieren worden van 1 tot 9 genummerd, waarna een toevalsvolgorde wordt bepaald: 2-9-4-3-5-8-1-6-7. De eerste drie nummers (2-9-4) worden een de placeboconditie toegeschreven, de tweede serie (3-5-8) aan de lage dosering en de derde serie (1-6-7) aan de hoge dosering.*



Afbeelding 2c. *Hier wordt een huidaandoening experimenteel geïntroduceerd. De drie velden in elk van de vier kolommen van een stuk huid worden per toeval aan de behandelingen toegewezen. De zwaarte van de experimenteel geïnduceerde huidaandoening kan per varken verschillen. Het is daarom ook nodig de factor 'varken' in de statistiche analyse op te nemen.*

5.1 Controle van de variatie

Bij dierexperimenteel onderzoek zal elke experimentele eenheid een eigen meetwaarde opleveren, die doorgaans verschilt van de waarden gemeten in de andere experimentele eenheden. Daardoor ontstaat een zekere variatie in de resultaten binnen elke behandelingsgroep. Deze binnengroepvariatie wordt gezien als een maat voor de meetfout. De variatie wordt veroorzaakt door een aantal factoren. Wanneer we *outbred*populaties inzetten – dus geen inteeltstammen of F1-hybriden – zal een deel van de variatie de genetische verschillen tussen de individuele dieren weerspiegelen. Ook heeft elk individu een net iets andere omgeving dan de andere dieren in een groep, zelfs in laboratoria waar alle omgevingscondities strikt zijn gestandaardiseerd (zie afbeeldingen 3 en 4). De positie in de baarmoeder (*intra-uteriene* omgeving) kan van invloed zijn op de ontwikkeling van een foetus. Ook de positie van de kooi in een schap, de sociale status in groepsgehuisveste dieren, *et cetera* kunnen een verschillende invloed uitoefenen op ieder individu. Verder kan de volgorde waarmee we dieren uit de kooi halen variabele effecten geven. Bovendien is er een samenspel (interactie) tussen het genotype en de omgeving. Dit samenspel bepaalt het uiteindelijke fenotype. Je kunt de binnengroepvariatie reduceren door dieren met hetzelfde genotype in een strikt gestandaardiseerde omgeving in een experiment in te zetten, door strikte standaardisatie van de dierverzorging en door gestandaardiseerde testcondities en testprocedures. Maar hierdoor kan de generaliseerbaarheid van onderzoeksresultaten worden gereduceerd, omdat de kans bestaat dat de gevonden effecten beperkt zijn tot de specifieke omstandigheden waaronder het onderzoek werd uitgevoerd. Ook is het lastig om alle factoren te standaardiseren en te benoemen. Veel zaken worden niet genoteerd. Zo is muziek in dierverblijven gangbaar in Nederland (om te voorkomen dat dieren schrikken van mensen die binnenkomen), maar dat is in andere Europese landen niet zo gebruikelijk.

Afbeelding 3. *Factoren die van invloed kunnen zijn op de resultaten van een dierexperimentele studie.*

In afbeelding 3 zijn de factoren weergegeven die de resultaten van een dierexperiment kunnen beïnvloeden. De grootste bron van variatie in de testresultaten is nagenoeg altijd de mens die het dier verzorgt en test. Goede scholing en werken volgens protocollen helpen de variatie zo laag mogelijk te houden. Bij kwaliteitssystemen zoals *Good Laboratory Practice* (GLP), *Good Manufacturing Practice* (GMP) en *Good Clinical Practice* (GCP) hoort een protocol of *Standard Operating Procedure* (SOP). In de SOP's worden de toegepaste methoden en technieken in detail beschreven.

In het proefplan beschrijf je alle methoden en technieken die worden toegepast tijdens een onderzoek: transport van de dieren, de huisvesting, de gezondheidscontrole, de toewijzing tot proefgroepen, criteria voor opname of uitsluiting, testapparatuur, testmethoden (hierbij kun je verwijzen naar SOP's), het lot van de dieren na afloop van het onderzoek, dataregistratie, controle van de ruwe data, data-analyse, rapportage, duur en plaats van de archivering van (ruwe) data en van het rapport, archivering van monsters. Die laatsten kunnen histologische preparaten zijn, weefsel- en bloedmonsters, (test)substanties, voermonsters. In een proefplan zijn vaak de ervaringen uit eerdere proeven verwerkt. Het proefplan is bindend voor het uitvoeren van het onderzoek. Door de variatie binnen de groep zoveel mogelijk te beperken, vergroot je de kans dat je tussen proefgroep(en) en controlegroep verschillen vindt. Deze tussengroepvariatie is een maat voor behandelingseffecten.

5.2 Omgeving

Proefdieren moeten zich gedurende hun leven herhaaldelijk aanpassen (adapteren) aan een nieuwe omgeving. De eerste omgeving is het opfokbedrijf. Vervolgens worden de dieren overgebracht naar het onderzoekslaboratorium, waarvoor nagenoeg altijd een transport nodig is. Transport veroorzaakt stress en stress kan de resultaten van onderzoek beïnvloeden. Het is daarom nodig de dieren aan de nieuwe omgeving (het dierverblijf) te laten adapteren voordat ze worden ingezet in het onderzoek. Gebruikelijk is een adaptatieperiode van ten minste een week. De derde omgeving is de testomgeving. Het is niet altijd gebruikelijk om dieren aan hun testomgeving te laten wennen. Toch kan het dan zinvol zijn dieren aan bepaalde (be)handelingen te laten wennen voordat je ze toepast. Denk aan injecties en aan testapparaten. Je kunt sommige dieren zelfs trainen om mee te werken aan bepaalde (afname)technieken. Ook is gewenning aan verzorger, biotechnicus of onderzoeker vaak van belang, bijvoorbeeld door ze het dier vooraf te laten aanhalen of oppakken.

5.3 Mensen: de belangrijkste bron van variatie

Dierverzorgers en biotechnici die metingen en andere waarnemingen verrichten, hebben een belangrijke rol om de variatie binnen groepen te verminderen. Goede scholing, heldere instructies, een duidelijk proefplan en zorgvuldig en nauwkeurig werken, helpen daarbij. Vooral bij vragen die gebaseerd zijn op subjectieve beoordelingen (bijvoorbeeld 'hoe is de toestand van de vacht of van de veren van het proefdier?') is een goede scholing vereist. Door verschillende biotechnici onafhankelijk van elkaar iets te laten beoordelen, kun je de mate van overeenstemming en de betrouwbaarheid van zulke metingen bepalen. Als het verschil in beoordeling door biotechnici te groot is, zal verdere scholing vereist zijn. Een betere beschrijving van de beoordelingscriteria vergroot soms de overeenstemming tussen beoordelaars. Je kunt duidelijke foto's of korte video's gebruiken als aanvullend scholingsmateriaal. Beoordelaars moeten de waarnemingen doen zonder weet te hebben van de groepsindeling, ze moeten dus 'geblindeerd' beoordelen.

6 Aantal groepen in een dierexperiment

Het aantal behandelingsgroepen is afhankelijk van twee factoren:
- het aantal behandelingen in een onderzoek;
- of elke behandeling in een gescheiden groep wordt toegepast (een tussen-subjectenopzet/ tussen-subjectenontwerp), of dat de dieren van een groep aan meer behandelingen worden blootgesteld (een binnen-subjectenopzet/ binnen-subjectenontwerp).

Voor een binnen-subjectenontwerp heb je in het algemeen de minste proefdieren nodig, maar een nadeel is dat de dieren meer ongerief kunnen hebben omdat ze meer behandelingen moeten ondergaan. Bovendien kunnen eerdere tests de resultaten van volgende tests beïnvloeden. Zo zou de behandeling met een testsubstantie langdurige fysiologische veranderingen kunnen veroorzaken. Ook kan de ervaring van een proefdier met een testsituatie via leren of angst een invloed hebben op het resultaat van een volgende behandeling. Je kunt een proef ook ontwerpen als een mix van een tussen- en een binnen-subjectenontwerp.

In een onderzoek moet je controlegroepen meenemen. Die zijn nodig om te kunnen schatten of de behandelingen effect hebben gehad. Bovendien kun je met controlegroepen onderzoeken of bepaalde onderdelen van de behandeling op zich al effect hebben. In een laesie-experiment is het bijvoorbeeld gebruikelijk dat dieren van een controlegroep een *sham*-operatie krijgen: ze worden onder anesthesie gebracht en er wordt een incisie gemaakt die weer wordt gehecht en verzorgd. In de proefgroep wordt na de incisie ook een *laesie* aangebracht (bijvoorbeeld afsluiting van een kransslagader bij de hond), waarna de wond wordt gesloten en verzorgd. Het verschil tussen de *sham*- en de *laesie*dieren is dus de *laesie*. Je weet met deze opzet alleen niet of de anesthesie en de incisie een effect hebben gehad. Om dit te kunnen beoordelen, moet je een onbehandelde extra controlegroep opnemen: de dieren in deze groep blijven ongemoeid, maar worden wel getest met de *sham*- en *laesie*dieren.

6.1 Schatting van het aantal benodigde dieren

De Centrale Commissie Dierproeven (CCD) – die gebruik maakt van een gemotiveerd advies van een dierexperimentencommissie (DEC) – eist dat je een schatting maakt van de aantallen dieren die je per groep nodig hebt. Deze berekening, de *power*analyse, kun je baseren op de uitkomsten (gemiddelden en standaarddeviaties) die zijn gevonden in eerder onderzoek. Bovendien kun je de grootte van een interessant geacht behandelingseffect laten meewegen in deze berekeningen.

De effectgrootte is een statistische maat voor hoe sterk het effect van bijvoorbeeld een behandeling is op een – in dit geval – populatie proefdieren, waarbij de vergelijking wordt gemaakt met een andere vergelijkbare proefdierpopulatie waarop die behandeling niet wordt toegepast (de controlegroep). Er zijn verschillende indices voor effectgrootte, de bekendste voor continue variabelen is de Cohen's *d*. Deze dimensieloze index is gebaseerd op de gemiddelden en standaarddeviaties van de twee te vergelijken proefdiergroepen.

In de loop der jaren is er conventie voor interpretatie ontstaan waarbij men:
- een absolute waarde van Cohen's *d* ($|d|$) < 0,20 te verwaarlozen vindt;
- spreekt van een klein effect als 0,20 ≤ $|d|$< 0,50;
- spreekt van een middelgroot effect als 0,50 ≤ $|d|$ < 0,80;
- boven de 0,80 spreekt van een groot effect.

Deze interpretatie is van toepassing op de meeste studies met proefdieren.

Maar in dierstudies met genetisch en microbiologisch gedefinieerde laboratoriumratten of -muizen die onder gestandaardiseerde omstandigheden worden gehouden, interpreteert men doorgaans de absolute waarden van Cohen's *d* als volgt:
- klein effect als $|d| \leq 0{,}5$;
- middelgroot effect als $0{,}5 < |d| < 1{,}0$;
- groot effect als $1{,}0 \leq |d| < 1{,}5$;
- zeer groot effect als $|d| \geq 1{,}5$.

Niet elk statistisch significant behandelingseffect zal ook biologisch of therapeutisch van belang zijn. Soms zijn significante grote effecten dat niet, terwijl significante kleine effecten dat wel kunnen zijn. Het kan daarom zinvol zijn bij de berekening van het aantal in te zetten dieren per groep uit te gaan van de minimale grootte van een effect dat ook biologisch of therapeutisch relevant is. Voor de schatting van de groepsgrootte zijn procedures in geautomatiseerde statistische programma's zoals het vrij toegankelijke programma G*Power beschikbaar (https://www.psychologie.hhu.de/arbeitsgruppen/allgemeine-psychologie-und-arbeitspsychologie/gpower.html). De groepsgrootten schatten, is onderdeel van je proefopzet.

6.2 Dieren verdelen over de groepen

Proefdieren moet je aselect (volledig bepaald door toeval) verdelen over de verschillende groepen. Dit kan met lotingtabellen die bestaan uit series getallen die per toeval zijn gegenereerd (*random numbers*). Lotingtabellen vind je in statistische handboeken. Ook met sommige statistische programma's kun je lotingtabellen maken, vaak zelfs voor het specifieke experimenteel design dat je voor je proef hebt gekozen.

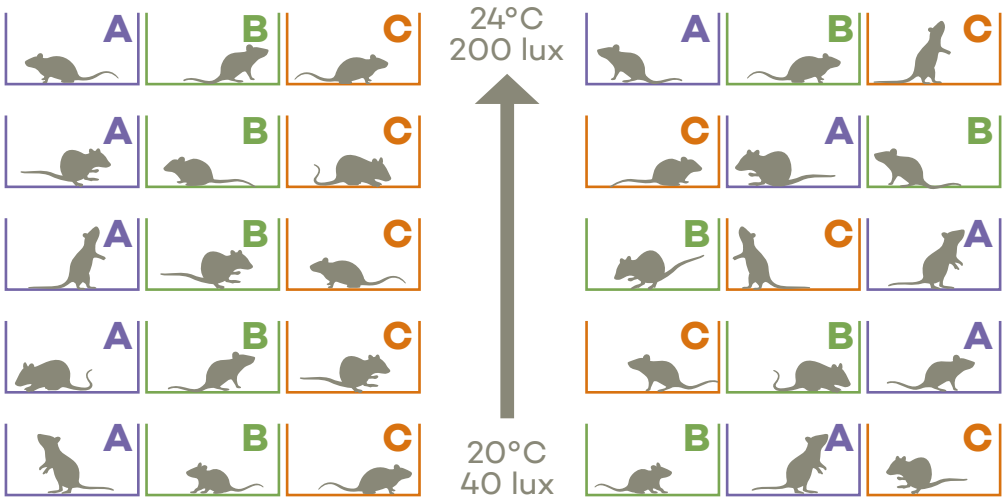
De nummering van de dieren kun je het best zo doen, dat er niet uit af te leiden valt welk dier aan welke behandelingsgroep is toegewezen. Hierdoor wordt gewaarborgd dat de personen die de proef uitvoeren en de metingen doen, niet weten met welke behandelingsgroep ze van doen hebben, en er dus 'blind' voor zijn. Uiteraard moet de uitvoerder van het onderzoek dan iemand anders zijn dan degene die de dieren heeft genummerd en aan behandelingscondities heeft toegewezen. Dit is van groot belang bij proeven waarin het meetresultaat (de afhankelijke variabele) gebaseerd is op subjectieve beoordeling, bijvoorbeeld het gedrag van de proefdieren of histologische afwijkingen in coupes.

6.3 Blinderen en randomiseren

Om te voorkomen dat dieren bewust of onbewust aan de controlegroep of aan een van de behandelingsgroepen worden toegewezen, worden de dieren per randomisatie (*ad random*, per toeval) verdeeld over de groepen in een experiment. Hierdoor kan worden voorkomen, dat er al voor de studie een systematische fout (*bias*) in de studie sluipt. Een tweede methode om *bias* te voorkomen is het geheimhouden van de toewijzing tot een groep tot het einde van de

studie. Pas als alle metingen zijn verricht en de statistische analyses worden uitgevoerd, wordt bekendgemaakt welk dier tot welke groep behoorde. Deze methode wordt *blinding* genoemd. De onderzoeker die de studie uitvoert, heeft geen idee welk dier welke behandeling heeft ondergaan en kan zo niet onbewust de resultaten beïnvloeden. Denk bijvoorbeeld aan een studie waarbij pijnuitingen een uitleesparameter zijn voor de effectiviteit van de behandeling met een pijnstiller. Het is goed mogelijk dat de onderzoeker onbewust de pijnuitingen verschillend interpreteert en scoort als hij of zij weet welke dieren tot de controlegroep behoren en welke dieren met welke dosering van de pijnstiller werden behandeld.

Blinderen betekent dat degene die de observaties en metingen aan het dier verricht niet weet welke behandeling het dier ondergaat. Een systematische nummering van de dieren is daarom af te raden. Geef dus niet bijvoorbeeld in een experiment met 3 behandelingsgroepen en 8 dieren per groep, de dieren van groep 1 de nummers 11 tot 18, die van groep 2 de nummers 21 tot 28, en die van groep 3 de nummers 31 tot 38. Het is veel beter de dieren zo te nummeren dat de nummering geen indicatie geeft over toewijzing tot behandelingscondities (zie ook afbeeldingen 2 en 4, de rechter voorbeelden).



Afbeelding 4. *Verdeling van dieren uit behandelingsgroepen A, B en C over de stellingen. In de linker stelling zijn de dieren in kolommen per behandelingsconditie gehuisvest. In de rechter stelling zijn er per schap de 3 behandelingscondities per loting toegewezen.*

6.4 De groepen in een dierverblijf

Ondanks dat dierverblijven voorzien zijn van airconditioning en een geregeld dag-nachtritme in de verlichting hebben, zijn de omgevingscondities niet op elke plek identiek. De temperatuur zal – afhankelijk van de kwaliteit van de airconditioning – in zekere mate variëren. In stellingen is de temperatuur in de kooien op de bovenste schappen iets hoger dan op de laagste schappen. Ook de lichtintensiteit verschilt en is het hoogst dicht bij de verlichtingsbron. De in- en uitlaten van de airconditioning produceren vaak een sterke ruis, soms met zeer hoge frequenties die de mens niet hoort, maar die een knaagdier wel kan waarnemen. Deze omgevingsinvloeden kunnen het welzijn en de gezondheid van het dier beïnvloeden en daardoor indirect invloed hebben op de resultaten van het onderzoek.

Door de dieren per behandelingsgroep verticaal over de schappen van een stelling te huisvesten, verdeel je de omgevingsinvloeden gelijkmatiger over de behandelingsgroepen (zie afbeelding 4). Door de kooien op de schappen door loting (toeval) een bepaalde plaats te geven, kan degene die de metingen uitvoert de behandelingsconditie niet meer achterhalen. Het onderzoek wordt hierdoor 'blind' uitgevoerd. Het per loting bepalen van de positie van de kooien op de schappen vereist wel grote zorgvuldigheid, omdat de kans op fouten hierbij reëel is.

Uiteraard moet degene die de behandelingen toepast wel weten welk dier aan welke behandelingsgroep is toegewezen. Maar ook dit kan 'blind'. Een onafhankelijke persoon kan de injectiespuiten met placebo en met de verschillende doseringen van een testsubstantie voorbereiden en coderen (bijvoorbeeld met A, B of C). De behandelingscondities die hierdoor zijn gecodeerd, worden 'achter slot en grendel' gehouden tot alle gegevens in het experiment zijn verzameld en de data kunnen worden geanalyseerd.

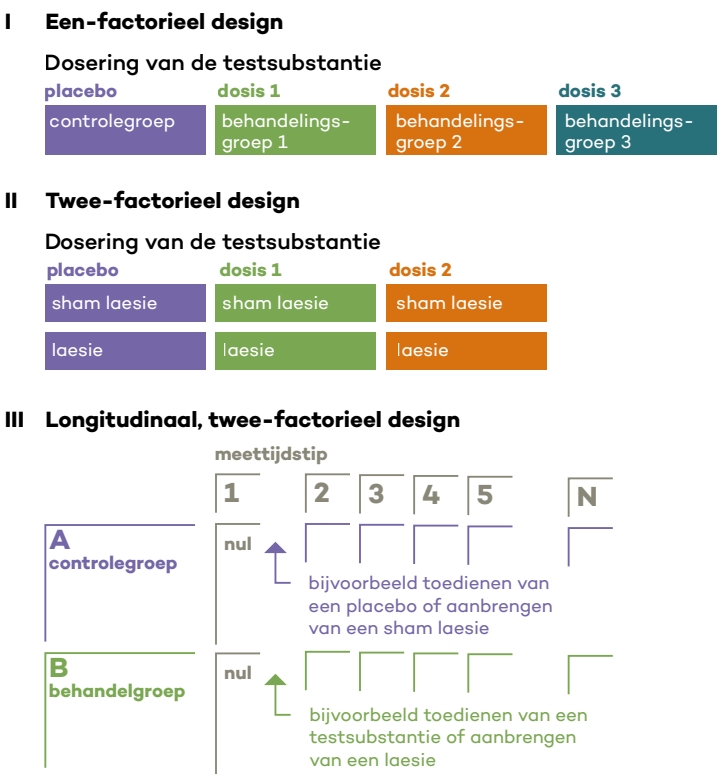
7 Proefopzet

De proefopzet is de manier waarop we de dieren (experimentele eenheden) aan de behandelingscondities toewijzen. De proefopzet heeft grote invloed op de nauwkeurigheid van de experimentele resultaten.

Bij het plannen en opzetten van een proef spelen maatregelen die meetonnauwkeurigheden en andere storende invloeden verkleinen, een centrale rol. Vergelijk de behandelingseffecten in je experimenteel onderzoek altijd met een geschikte controle. Daarvoor staat een groot aantal proefschema's ter beschikking. Al deze proefschema's bevatten op zijn minst een controleconditie die nodig is om een goede schatting van de behandelingseffecten te kunnen maken.

7.1 Schema's als elk dier één behandeling ondergaat

Enkele proefschema's waarin elk dier slechts een behandeling (of combinatie van behandelingen) ondergaat, zie je in afbeelding 5. In deze figuur staat 'factorieel' voor een behandelingsconditie.



Afbeelding 5. *Enkele voorbeelden van een proefopzet.*

Schema I: *De dieren worden per toeval aan een van vier groepen toegewezen (een placebo controlegroep en drie groepen met oplopende doseringen van de testsubstantie).*

Schema II: *Hier worden zes groepen dieren ingezet.*

Schema III: *Voorbeeld van een longitudinale studie met herhaalde metingen. De nulmeting kan worden gebruikt om de dieren met behulp van 'random matched assignment' over de behandelingsgroepen te verdelen (voor verdere uitleg, zie tekst). Na de nulmeting worden er nog (N –1)-metingen verricht.*

Schema I is een één-factorieel design waarin het toeval de dieren toewijst aan de behandelingscondities. Zet je 10 dieren per behandelingsgroep in en heb je 4 groepen, dan gebruik je dus 40 dieren. Een van de groepen dient als controlegroep. De andere groepen krijgen elk een van de 3 doseringen van de testsubstantie toegediend. Dit onderzoek zou bijvoorbeeld de vraag kunnen beantwoorden welke dosering de beste gewenste (therapeutische) effecten oplevert of vanaf

welke dosering ongewenste bijwerkingen optreden. In het laatste onderzoek kies je een dosering met therapeutische werking en logaritmisch (grondtal 10) oplopend, minimaal twee hogere doseringen.

Schema II geeft een twee-factorieel design weer. Hier zou bijvoorbeeld de vraag kunnen worden onderzocht of een potentieel therapeuticum na een experimenteel geïnduceerd herseninfarct de schade kan reduceren. Er is gekozen voor twee doseringen van het therapeuticum. De proef bestaat dus uit twee factoren. De eerste factor is de *laesie* (experimenteel geïnduceerd herseninfarct) met als controle de *shamlaesie* waarbij het dier de hele operatieprocedure ondergaat, maar geen herseninfarct krijgt. De tweede factor is het toegediende therapeuticum. De controle bestaat uit een placebobehandeling waarin de dieren exact dezelfde behandeling ondergaan als de dieren die het therapeuticum krijgen toegediend, maar nu dienen we geen farmacologisch werkzame substantie toe. De controlecondities en controlegroepen zijn voor elk experiment van cruciaal belang. Afhankelijk van de complexiteit van de vraagstelling kan het nodig zijn meer dan één controlegroep op te nemen. Bij *laesie*-experimenten ondergaat een controlegroep een invasieve *sham*-operatie. Om de effecten van deze operatie te kunnen meten, is soms een extra controlegroep nodig met dieren die géén operatie hebben ondergaan.

Schema III is een voorbeeld van een twee-factorieel longitudinaal (herhaalde metingen) design. Hiermee onderzoek je of de toediening van een substantie of het opwekken van een specifieke hersenlaesie langdurige effecten heeft. Het toeval wijst de dieren toe aan een van beide behandelingsgroepen: de controlegroep A of de experimentele of behandelingsgroep B. Na een nulmeting behandel je de controlegroep met een placebo (meestal het middel waarin de testsubstantie wordt opgelost, dus zonder actieve componenten). Groep B krijgt de testsubstantie. Op verschillende tijdstippen na behandelingen worden metingen verricht. De tijdstippen waarop metingen worden verricht, representeren een herhaalde metingenfactor.

Bij experimenten met een nulmeting kun je dieren op basis van de meting aan de behandelingsgroepen toewijzen (*matching*). Je rangschikt hierbij de dieren op de waarden van de nulmeting. Daarna wijst in deze ranglijst het toeval elk van de twee dieren van een paar, van de hoogste naar de laagste waarde bij de nulmeting toe aan de twee behandelingsgroepen.

Als de behandelingen vóór de eerste meting worden gegeven, moeten de dieren vooraf per toeval aan de behandelingsgroepen worden toegewezen. Omdat in de meeste gevallen maar op één variabele wordt gematcht, is het niet zeker dat de groepen ook voor andere (wellicht relevante) variabelen gelijke gemiddelden en spreidingen hebben. Matchen op meer dan één variabele tegelijk is mogelijk, maar is een zeer complex proces. Bij matching op slechts één variabele moet je met statistische methoden onderzoeken of dit voor alle relevante variabelen gelijkwaardige groepen oplevert. Een bijkomend voordeel van nulmetingen is dat je de behandelingseffecten binnen een behandelingsgroep kunt meten als veranderingen ten opzichte van de nulmeting.

Er bestaat nog een groot aantal andere experimentele ontwerpen. In ontwerpen met meer dan één factor kunnen we interacties tussen factoren onderzoeken. Zo zou bijvoorbeeld het effect van een geneesmiddel geslachtafhankelijk kunnen zijn, dus in het ene geslacht groter dan in het andere. Je moet in een onderzoek routinematig de twee geslachten opnemen om deze interacties te kunnen detecteren. Door beide geslachten mee te nemen worden tevens fokoverschotten beperkt en wordt het aantal in voorraad gedode proefdieren verminderd. Dit betekent dat we aan de voorbeelden van afbeeldingen 5 en 6 (zie hierna) een extra factor moet toevoegen: geslacht.

Met het oog op de kosten (in materiaal, proefdieren, tijd) van een studie kan het erg verleidelijk zijn de controlegroepen weg te laten, zeker wanneer er resultaten van controlegroepen en behandelingsgroepen uit eerder onderzoek zijn (historische data). Maar er zijn altijd verschillen tussen experimenten. Het kan dus zijn dat alle waarden uit een experiment hoger of lager uitkomen dan eerder werd gemeten. Vergelijken we deze nieuwe data dan met historische controledata, dan zouden we ten onrechte kunnen concluderen dat er een behandelingseffect was, of juist niet.

7.2 Schema's als de dieren meer behandelingen ondergaan

Het is ook mogelijk de effecten van meer behandelingen – zoals verschillende doseringen van een proefsubstantie – in hetzelfde dier te onderzoeken. Het Latijns vierkant is een proefopzet waarin dit mogelijk is (zie afbeelding 6).

I Latijns vierkant met twee groepen

groep	testperiode 1	testperiode 2
A	placebo	dosis 1
B	dosis 1	placebo

II Latijns vierkant met drie groepen (voorbeeld 1)

groep	testperiode 1	testperiode 2	testperiode 3
A	placebo	dosis 1	dosis 2
B	dosis 1	dosis 2	placebo
C	dosis 2	placebo	dosis 1

III Latijns vierkant met drie groepen (voorbeeld 2)

groep	testperiode 1	testperiode 2	testperiode 3
A	placebo	dosis 2	dosis 1
B	dosis 1	placebo	dosis 2
C	dosis 2	dosis 1	placebo

IV Uitbreiding van Latijns vierkant tot een schema met zes groepen

groep	testperiode 1	testperiode 2	testperiode 3
A	placebo	dosis 1	dosis 2
B	placebo	dosis 2	dosis 1
C	dosis 1	dosis 2	placebo
D	dosis 1	placebo	dosis 2
E	dosis 2	placebo	dosis 1
F	dosis 2	dosis 1	placebo

Afbeelding 6. *Twee of meer behandelingen kunnen in hetzelfde dier worden onderzocht met behulp van (uitbreiding van) het Latijns vierkant.*

In schema I is het meest eenvoudige Latijns vierkant weergegeven. Twee behandelingen worden in twee groepen van dieren onderzocht. In schema II worden drie behandelingen in drie groepen van dieren onderzocht. De groepen van dieren worden per loting aan een van deze volgorden toegewezen. Bij drie of meer behandelingen zijn er alternatieve Latijnse vierkanten mogelijk. Schema's II en III zijn varianten van een Latijns vierkant met drie behandelingen en drie proefgroepen. De proefopzet in schema IV is een uitbreiding van het basisschema die het mogelijk maakt overdrachts- of 'cross-over'-effecten te detecteren.

Kenmerkend voor het Latijns vierkant is dat het aantal groepen (N) gelijk is aan het aantal behandelingen of testperioden. N mogelijke volgorden van behandelingen worden onderzocht. In een Latijns vierkant bevat elke rij en elke kolom maar één keer een bepaalde behandeling.

Bij twee behandelingen A en B (bijvoorbeeld '*placebo*' en 'dosis 1') zijn als volgorden A-B en B-A mogelijk. Je ziet dat in afbeelding 6, schema I. Bij drie behandelingen A, B en C (bijvoorbeeld '*placebo*', 'dosis 1' en 'dosis 2') zijn de volgorden A-B-C, B-C-A, C-A-B, A-C-B, B-A-C en C-B-A mogelijk, zoals je ziet in afbeelding 6, schema's II en III. Elke groep dieren wordt door loting aan een van deze volgorden toegewezen.

Behandelingseffecten in een Latijns vierkant worden binnen ieder individu berekend als verschillen in scores tussen de behandelingen. Deze verschillen in scores worden statistisch geanalyseerd. Opeenvolgende behandelingen kunnen elkaar beïnvloeden. Dit effect wordt overdrachtseffect of *cross over*-effect genoemd. Overdrachtseffecten kun je in de Latijnse vierkanten in het voorbeeld van afbeelding 6, schema II niet onderzoeken, omdat er alleen de volgorden A-B, B-C en C-A voorkomen, maar niet de volgorden B-A, C-B en A-C. Deze volgorden komen in het voorbeeld van afbeelding 6, schema III wel voor. Hier ontbreken dan weer de volgorden A-B, B-C en C-A. Door vergelijking van volgorden A-B en B-A zou je overdrachtseffecten kunnen detecteren. Zonder overdrachtseffecten moeten de effecten van behandeling A gelijk zijn, ongeacht de volgorde. Dit geldt uiteraard ook voor behandeling B.

Om overdrachtseffecten te kunnen detecteren, moeten paren van behandelingen elkaar in de twee mogelijke volgorden opvolgen. Dit bereik je door uitbreiding van het Latijns vierkant, zoals je ziet in afbeelding 6, schema IV: de schema's II en III zijn hier in elkaar geschoven, waardoor beide mogelijke volgorden per paar van behandelingen voorkomen.

Het aantal dieren in een (uitbreiding van het) Latijns vierkant is altijd een veelvoud van het aantal behandelingen. Bijvoorbeeld: bij 3 behandelingen met 12 metingen per behandeling, zet je 4 dieren per behandelingsvolgorde (van 2 behandelingen) in. Elke groep draagt 4 metingen bij aan het totaal van 12 metingen. In totaal zet je in dit voorbeeld dus 12 dieren in. Bij een proefschema waarin elk dier een enkele behandeling ondergaat, zouden er 3 groepen met elk 12 dieren nodig zijn, in totaal 36 dieren. Met behulp van een (uitbreiding van het) Latijns vierkant kun je het aantal proefdieren sterk reduceren. Maar hier staan beperkingen en nadelen tegenover. Je kunt dit proefschema niet gebruiken als een dier na of vanwege de meting moet worden geëuthanaseerd, bijvoorbeeld voor histopathologisch onderzoek. Dit schema is ook onbruikbaar als een behandeling onomkeerbare (irreversibele) veranderingen in het dier veroorzaakt (bijvoorbeeld door toxische effecten), of wanneer door een meting leerprocessen optreden die het gedrag in een volgende meting beïnvloeden. Ook als er geen irreversibele veranderingen optreden, moet je tussen opeenvolgende testperioden een *wash out*-periode inlassen. Deze periode moet lang genoeg zijn om eventueel in het dier aanwezige testsubstanties volledig te laten verdwijnen (weg te wassen), voordat het effect van de volgende dosering kan worden onderzocht.

Een nadeel van het Latijns vierkant en uitbreidingen daarvan kan zijn dat de dieren meer ongerief ondervinden, omdat ze vaker worden behandeld en getest.

8 Protocollen ontwikkelen

Een protocol beschrijft alle handelingen en technieken en hun volgorde in een dierexperimenteel onderzoek. Een protocol ontwikkelen is een taak van de proefleider. Een goed protocol is een handleiding voor het onderzoek waarin alle relevante informatie is vastgelegd. Ook kan erin worden verwezen naar bijvoorbeeld SOP's, waarin deze handelingen gedetailleerd zijn beschreven. De auteur van een protocol doet er goed aan dierverzorgers en biotechnici in een vroeg stadium bij de ontwikkeling van een protocol te betrekken. Ervaringen, waaronder fouten, die in een eerder onderzoek werden gemaakt, zijn een goede bron om een protocol te verbeteren. Een zorgvuldige analyse van bestaande publicaties die relevant zijn voor de te onderzoeken vraagstelling kan bijvoorbeeld aanwijzingen opleveren voor valkuilen en problemen en aanbevelingen hoe deze te vermijden zijn. Een protocol moet uitvoerbaar zijn en moet daarom rekening houden met de omstandigheden in de dierverblijven en de laboratoria en met de dagelijkse routines.

8.1 Het nut van protocolleren

Kwaliteitssystemen zoals GLP, GMP en GCP steunen op protocollen waarin alle (proef)dierenmanagement handelingen en alle (dier)experimentele handelingen zeer gedetailleerd zijn vastgelegd. Eventueel wordt verwezen naar SOP's, waarin het materiaal en de methoden bindend en nauwkeurig zijn gedocumenteerd. De term SOP gebruiken we alleen bij onderzoek onder de genoemde kwaliteitssystemen. In andere gevallen gebruiken we termen als voorschriften, handleidingen, protocollen.

Ook voor onderzoek dat niet volgens de regels van deze kwaliteitssystemen wordt uitgevoerd, is een protocol onmisbaar. Het protocol kan uitgangspunt zijn voor de aanvraag van het onderzoek bij de dierexperimentencommissie.

Wanneer een onderzoek met of voor een buitenlandse partner of opdrachtgever wordt uitgevoerd, is het protocol waarschijnlijk in het Engels geschreven. In dit geval worden weleens Nederlandstalige uittreksels voor de werkvloer gemaakt. Hoewel deze de communicatie op de werkvloer kunnen vergemakkelijken, is alleen het originele – Engelse – protocol bindend!

Het verdient aanbeveling om voor de start van een onderzoek een *kick off*-meeting te organiseren. De projectleider informeert hierin alle medewerkers aan het onderzoek over het doel van het onderzoek, over kritieke punten in het protocol en ieders specifieke taken. Op deze bijeenkomst kunnen vragen over het protocol worden gesteld. Ook kunnen afspraken worden gemaakt over de communicatielijnen, dat wil zeggen, wie wanneer wordt geïnformeerd over problemen tijdens het onderzoek. Communicatielijnen kunnen ook bindend op schrift worden gesteld en deel uitmaken van het protocol.

Voor de kwaliteitsborging van een onderzoek en ter voorkoming van fouten in toekomstig onderzoek, verdient het aanbeveling om na afsluiting van een onderzoek een evaluatiebijeenkomst met alle betrokken medewerkers te houden. Hierin informeert de projectleider de medewerkers over de resultaten, over problemen en hun oplossing, en over fouten.

8.2 Afwijkingen in het protocol (amendement, proefafwijking)

Nadat het protocol is goedgekeurd, kunnen in de loop van het onderzoek twee soorten afwijkingen voorkomen:

- afwijkingen die op grond van voortschrijdend inzicht worden gepland (amendementen). Deze moeten als afwijkingen van het oorspronkelijke protocol vooraf worden beschreven en goedgekeurd door de proefleider en eventueel de opdrachtgever én door de CCD of DEC, voordat je ze mag toepassen;
- afwijkingen die tijdens het onderzoek of achteraf worden geconstateerd (proefafwijkingen, deviaties). Zulke afwijkingen zijn nagenoeg altijd het resultaat van fouten, het niet goed lezen en omzetten van het protocol of van incidenten. Ze kunnen een grote invloed hebben op de resultaten en daarmee ook op de interpretatie van het onderzoek.

Daarom is een eerlijke, nauwkeurige en volledige documentatie nodig. Hierbij wordt tevens de vraag gesteld – en zo mogelijk beantwoord – welke effecten de afwijking op de validiteit van de onderzoeksresultaten kan hebben gehad.

Bij onderzoek volgens kwaliteitssystemen zijn de procedures over constateren en rapporteren van afwijkingen van het onderzoeksprotocol strikt geregeld. Ze zijn een integraal en uiterst belangrijk aspect van deze kwaliteitssystemen. Een onderzoek verliest de GLP- of GCP-status wanneer niet strikt volgens de regels wordt gehandeld. De resultaten van dit onderzoek kunnen dan niet worden meegenomen in het dossier dat nodig is voor de eventuele toelating van een nieuw geneesmiddel op de markt. Het onderzoek moet soms zelfs worden overgedaan en kan dus leiden tot onnodig proefdiergebruik, tot extra kosten en tot een vertraging van het toelatingsproces van een geneesmiddel. Als dierverzorger of biotechnicus draag je dus een grote verantwoordelijkheid!

9 Statistiek

Statistiek levert methoden om onderzoek op te zetten, en om conclusies te trekken uit gegevens die uit onderzoek zijn verkregen. Het is een van de weinige deelgebieden van de wiskunde dat in nagenoeg elke vorm van wetenschap gebruikt wordt. Statistiek is in elk geval onontbeerlijk voor het biowetenschappelijk (dier)onderzoek:

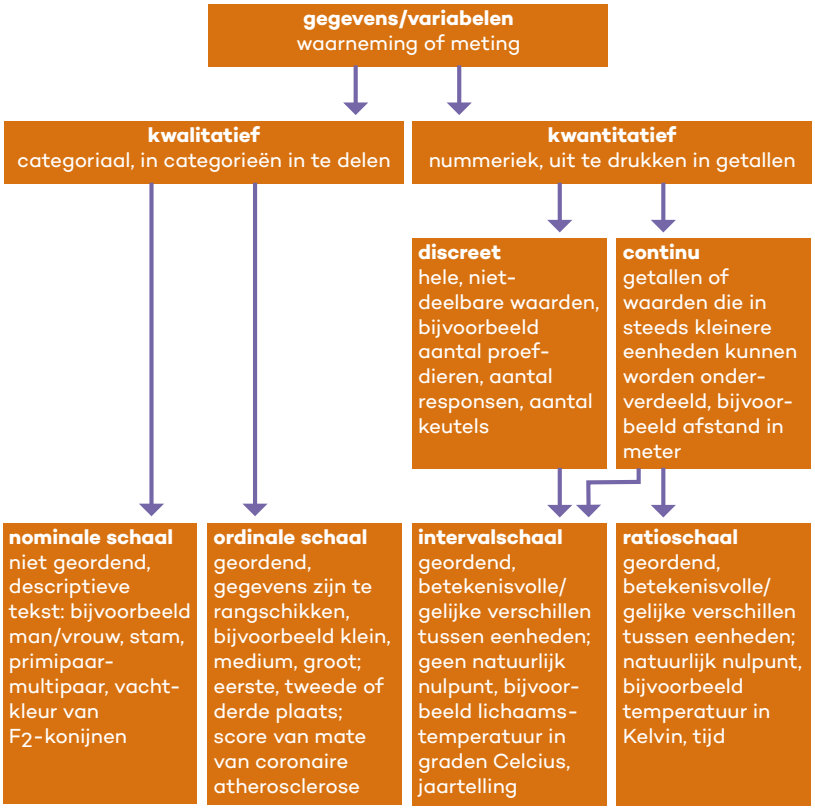
- *in vivo* (in een levend organisme in zijn geheel);
- *ex vivo* (buiten het lichaam van een organisme);
- *in vitro* (in buisjes of schaaltjes van het laboratorium).

Statistiek is een middel om de waarde van resultaten van experimenten vast te stellen. Voordat je resultaten statistisch kunt analyseren, moet duidelijk zijn wat voor soort resultaten het zijn, hoe ze zijn verzameld, *et cetera*. De bruikbaarheid van de resultaten is natuurlijk afhankelijk van een goede opzet van het experiment; in paragraaf 7 *Proefopzet* lees je daar meer over. Daarom is het belangrijk om advies in te winnen bij een (bio)statisticus voordat je aan een experiment begint. Statistiek is geen tovermiddel waarmee de data van een slecht experiment tot iets moois kunnen worden gemaakt. Het is slechts een hulpmiddel om gegevens te analyseren die je op een verantwoorde, op de vraagstelling toegesneden manier hebt verkregen, zodat de onderzoeksvragen beantwoord kunnen worden. Hierna komen verschillende zaken aan de orde die betrekking hebben op de statistische analyse van resultaten. Bijvoorbeeld de verschillende typen experimentele resultaten en de manieren waarop je ze kunt indelen en weergeven.

Er wordt onderscheid gemaakt tussen beschrijvende statistiek en verklarende of toetsende statistiek. In de beschrijvende statistiek worden de gegevens geordend en samengevat, met als doel een overzicht van de gegevens te krijgen. Bij de verklarende statistiek gaat het om de statistische analyse van de resultaten en het trekken van conclusies; deze kan een voorspellend karakter hebben.

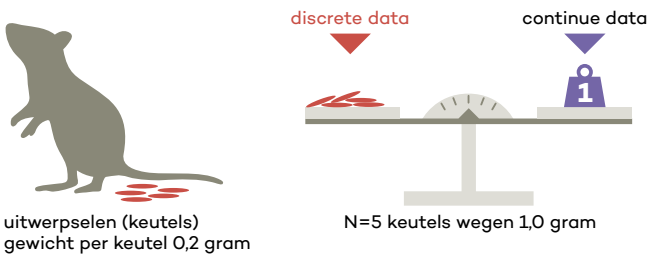
10 Soorten gegevens

Het is voor elk onderzoeksproject essentieel de gegevens te beschrijven die je hebt verkregen door waarneming of meting. Hier zijn veel redenen voor, zoals onlogisch hoge of lage waarden (uitbijters) opsporen, of de aannamen controleren die vereist zijn om statistisch te kunnen toetsen. De aard van de gegevens is kwalitatief of kwantitatief (zie afbeelding 7). Kwantitatieve ofwel numerieke gegevens kun je op grond van hun aantal (kwantiteit) op volgorde rangschikken. Bij kwalitatieve gegevens kan dat niet altijd. Als het wel kan, noemen we de gegevens ordinaal en als het niet kan, nominaal. Kwalitatieve gegevens horen tot de zogenoemde discrete grootheden of variabelen.



Afbeelding 7. *Soorten van gegevens.*

Kwantitatieve uitkomsten worden onderscheiden in discrete (discontinue) en continue variabelen (zie afbeelding 8). Numerieke discrete variabelen zijn meestal tellingen of aantallen. Een continue variabele kan elke waarde (uitkomst) zijn, meestal wel in een bepaald bereik. Bij continue variabelen kunnen we nog een verder onderscheid maken tussen die op een ratioschaal en die op een intervalschaal. Bij een ratioschaal is er een vast nulpunt en is het zinvol ratio's (verhoudingen) te bestuderen. Het nulpunt is meestal een natuurlijk gegeven waar je niet onder kunt zakken: negatieve waarden hebben hier geen betekenis. Bij een intervalschaal is het nulpunt willekeurig en in feite verschuifbaar.



Afbeelding 8. *Een discrete en een continue variabele.*

11 Betekenisvolle cijfers en afronden

Alleen cijfers die met zekerheid bekend zijn, zijn van betekenis (significant). Het is een goed gebruik om alleen die decimalen van een uitkomst te melden die je op grond van de nauwkeurigheid van de meting kunt verantwoorden. Zo zul je bij het meten van de lengte (in centimeters) van een muizenstaart niet de neiging hebben om meer dan één decimaal te vermelden. Meer decimalen hebben geen betekenis en spelen dus ook geen rol bij conclusies trekken. Met het aantal decimalen geef je de nauwkeurigheid van de waarneming of meting weer.

Afronden is het aan het eind van een getal weglaten van cijfers die geen betekenis hebben. Daar zijn enkele vuistregels voor:

- eindigt het af te ronden getal op een 5 of hoger, dan rond je naar boven af naar het dichtstbijzijnde getal: 4,56 rond je af naar 4,6;
- eindigt het getal op een cijfer lager dan 5, dan rond je naar beneden af: 4,54 rond je af naar 4,5;
- het resultaat van een berekening heeft niet meer decimalen dan de waarneming met het kleinste aantal decimalen. Dus als je een meetwaarde op 2 decimalen nauwkeurig weergeeft, heeft het geen zin een gemiddelde te vermelden in 3 decimalen.

Afronden is geen kwestie van gemakzucht; door correct af te ronden, vermijd je dat er een grotere precisie gesuggereerd wordt dan verantwoord is. Rond af als laatste stap in de rekenprocedure.

12 Tabellen en grafieken

In onderzoek gaat het vaak om grote aantallen gegevens. Als je snel iets wilt kunnen aflezen is het handig om de gegevens te ordenen, bijvoorbeeld van klein naar groot, alfabetisch, per categorie, van vroeg naar laat, *et cetera*. Gegevens worden meestal als frequentieverdelingen weergegeven in grafieken en tabellen. Het belang van een tabel is het geven van exacte getallen (met een precisie zoals uitgelegd in paragraaf 11 *Betekenisvolle cijfers en afronden*). Bij grafieken moet je de waarden van meetpunten schatten door bijvoorbeeld naar de assen van een grafiek te kijken.

Er is geen duidelijk criterium voor de keuze tussen een tabel of een grafiek voor het laten zien van gegevens. Eigenlijk moet je jezelf altijd afvragen welke 'boodschap' je met die gegevens wilt overbrengen, en of dit het beste kan via een grafiek of een tabel. Over het algemeen is een grafiek wat duidelijker. Wanneer het mogelijk is de resultaten overzichtelijk in een grafiek weer te geven, moet je dat doen. In sommige gevallen heb je erg veel grafieken nodig en kun je beter voor de tabel kiezen om de informatie overzichtelijk te presenteren.

Gegevens verwerken in tabellen of grafieken is uitgebreid behandeld in twee artikelen (#2 – *Ordenen van gegevens door middel van tabellen* en #3 – *Ordenen van gegevens door middel van figuren*), die in 2003 verschenen zijn als onderdeel van de serie – *[basale] statistiek en dierexperimenten* – in het tijdschrift *Biotechniek*. We raden je aan deze twee artikelen te raadplegen.

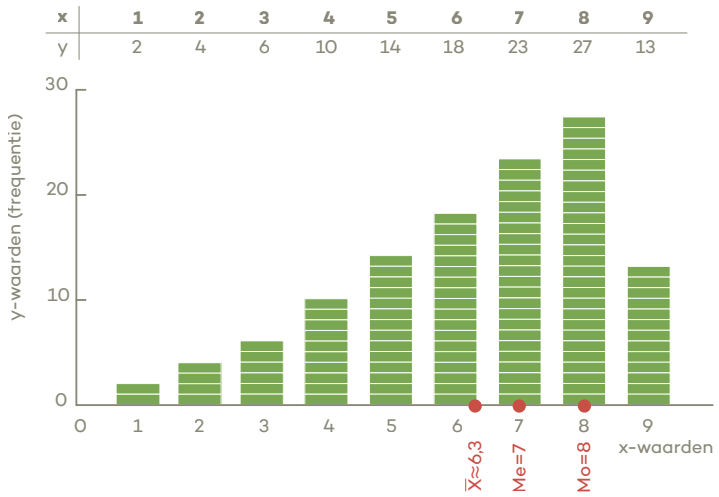
13 Kengetallen

In paragraaf 12 *Tabellen en grafieken* schreven we dat je gegevens kunt reduceren tot frequentieverdelingen. In deze paragraaf voeren we deze datareductie nog verder door. Je zoekt dan naar kenmerkende getallen voor de frequentieverdeling. Zulke kengetallen zijn er voor de ligging (ook wel centrummaten of locatiematen genoemd) en voor de spreiding. Om zinvol gebruikt te kunnen worden, moeten de kengetallen de volgende eigenschappen bezitten:

- de kengetallen moeten eenduidig gedefinieerd zijn;
- alle waarnemingen spelen een rol bij de bepaling van het kengetal;
- de interpretatie van het kengetal moet eenvoudig en inzichtelijk zijn;
- de kengetallen moeten niet al te gevoelig zijn voor steekproeftoevalligheden;
- met de kengetallen moeten algebraïsche bewerkingen mogelijk zijn.

13.1 Centrummaten

Bij het beschouwen van een reeks gegevens vraagt men zich vaak af: 'welke waarde komt het meest voor?', 'welke waarde komt gemiddeld genomen het meest voor?' of 'wat was gemiddeld genomen de waarde?' Dit soort vragen gaat over een centrale waarde die een goede indruk geeft van de verzameling gegevens. Goede parameters voor een centrale waarde zijn de modus, de mediaan en het (rekenkundig) gemiddelde (zie afbeelding 9). Deze stellen eisen aan het meetniveau van de variabele. De modus kun je op alle niveaus bepalen, de mediaan vanaf ordinaal niveau en het gemiddelde vanaf intervalniveau.



Afbeelding 9. *Een scheve frequentieverdeling met maatstaven voor het centrum.*

13.1.1 Modus

De modus wordt vaak afgekort als Mo. Het is de waarde die het meest voorkomt, ofwel de categorie met de hoogste frequentie. Indien we te maken hebben met een interval- of ratioschaal, dan kun je van de klasse met de grootste frequentie het midden van de klasse als modus opgeven. In een frequentieverdeling met een ongelijke klassenbreedte is de modale klasse de klasse met

de hoogste frequentie. Soms komt de hoogste frequentie bij twee waarden of klassen voor. Dan zeggen we dat de verdeling bimodaal is. Als dat bij meer dan twee waarden of klassen zo is, wijzen we meestal geen modus aan. De modus is alleen een zinvolle maatstaf als je veel waarnemingen hebt gedaan. Bij weinig waarnemingen kan het zijn dat alle uitkomsten maar één keer voorkomen; je kunt dan geen modus bepalen. De modus is een vrij onstabiele centrummaat, omdat bij verandering van het aantal waarnemingen de modus gemakkelijk kan verschuiven naar een andere waarde of klasse.

13.1.2 Mediaan

Dit is een ander veel gebruikt kengetal om het centrale niveau aan te geven. We noteren de mediaan vaak als Me. De mediaan is de waarde waarbij de helft van de scores lager en de andere helft hoger is. In een geordende reeks is de mediaan de middelste waarde. Bij een even aantal waarnemingen in een geordende reeks is de mediaan het gemiddelde van de middelste twee waarnemingen. De mediaan voor het aantal keren dat een muis zichzelf poetst gedurende een gedragstest van 5 minuten {7 muizen: 1, 2, 3, 3, 5, 8, 10} is dus gelijk aan 3; die van 6 muizen: {2, 3, 3, 5, 8, 10} is gelijk aan 4.

13.1.3 Rekenkundig gemiddelde

Het rekenkundig gemiddelde, ook wel aritmetisch gemiddelde genoemd, is de meest gebruikte centrummaat en wordt weergegeven door een x met een horizontaal streepje erboven ($\bar{x}$). Je vindt het rekenkundig gemiddelde door alle waarnemingen op te tellen en de som te delen door het aantal waarnemingen. We laten het voorvoegsel 'rekenkundig' meestal weg. Behalve het rekenkundig gemiddelde zijn er nog enkele andere gemiddelden die soms nuttig zijn om te gebruiken, zoals het meetkundig of geometrisch gemiddelde en het harmonisch gemiddelde.

13.1.4 Wanneer welke centrummaat?

De koppeling van centrummaten aan meetniveau is niet het enige argument om te kiezen. Van belang is ook de verdeling van de scores: symmetrisch of niet. Bij een symmetrische verdeling vallen modus, mediaan en gemiddelde samen. Als er sterk eruit springende waarden ofwel uitbijters zijn, of als je hierover twijfelt, dan kun je beter de mediaan kiezen; dit geldt ook bij interval- en ratiovariabelen.

13.2 Spreidingsmaten

Spreiding (ook wel strooiing genoemd) doet zich voor als de waarden van resultaten van elkaar verschillen. Hoe groter de verschillen, des te groter de spreiding. Twee verzamelingen waarnemingsresultaten kunnen hetzelfde rekenkundig gemiddelde hebben, maar een zeer verschillende spreiding. Voor de spreiding van uitkomsten is een aantal maatstaven ontwikkeld. Er zijn spreidingsmaten voor alle meetniveaus, zelfs voor nominale variabelen (die we in dit handboek niet bespreken). De keuze van een spreidingsmaat is gekoppeld aan die van een centrummaat.

13.2.1 Spreidingsbreedte of variatiebreedte

Het ligt voor de hand alleen de uiterste waarden te vermelden. De grootste (hoogste) minus de kleinste (laagste) score geeft de spreidingsbreedte. Deze past bij de modus als centrummaat. De spreidingsbreedte zegt niets over hoe de scores gespreid liggen. Eén uitschieter kan volledige vertekening van de spreiding geven. De spreidingsbreedte is alleen goed hanteerbaar wanneer de waarnemingen regelmatig gespreid zijn. De spreidingsbreedte is systematisch afhankelijk

van het aantal waarnemingen: in de regel neemt de spreidingsbreedte toe met het aantal waarnemingen.

13.2.2 Gemiddelde absolute afwijking

Deze maatstaf maakt gebruik van alle waarnemingsuitkomsten. Elke waarneming wijkt meer of minder af van het gemiddelde. Eerst bereken je het rekenkundig gemiddelde van de waarnemingen. Daarna bereken je voor elke uitkomst het absolute verschil met dit gemiddelde; laat bij een negatief verschil het minteken weg. Tel de absolute verschillen op en deel ze door het aantal waarnemingen. Deze spreidingsmaat, die past bij het gemiddelde als centrummaat, gebruik je als de variantie of de standaardafwijking die we hierna bespreken, niet voldoet.

13.2.3 Variantie en standaardafwijking

De meest gebruikte spreidingsmaat is de standaardafwijking of -deviatie (SD). Ook deze spreidingsmaat is gebaseerd is op het rekenkundig gemiddelde. Eerst bepaal je voor elke waarneming weer de afwijking tot het gemiddelde. Daarna kwadrateer je deze afwijkingen. Tel de kwadraten op en deel ze door het aantal waarnemingen. Het zo verkregen gemiddelde van de gekwadrateerde afwijkingen heet de variantie. De SD is nu gelijk aan de wortel van de variantie.

De op deze manier berekende variantie en SD gelden voor een set waarnemingen die tegelijk de gehele populatie vormen, bijvoorbeeld alle ratten van een bepaalde stam in de wereld. Gaat het om een steekproef, deel dan de som door het aantal waarnemingen minus 1.

In plaats van de SD gebruiken we ook wel de standaardfout van het gemiddelde (*standard error of the mean*, vaak afgekort tot SEM) . Deze bereken je uit de SD door deling door de kwadraatwortel uit het aantal waarnemingen. Hierdoor is de SEM kleiner dan de SD. De SEM gebruik je als je een indicatie wilt geven van de nauwkeurigheid, waarmee het steekproefgemiddelde het populatiegemiddelde schat.

13.2.4 Variatiecoëfficiënt

Deze geeft aan hoe groot de relatieve spreiding is: het is de SD in procenten van het gemiddelde. Formule: (SD/gemiddelde) x 100%. De coëfficiënt wordt veel gebruikt om uitspraken te doen in de trant van 'de verschillen in deze groep zijn relatief groter dan in een andere groep'.

13.2.5 (Inter)kwartielafstand

In een geordende reeks bij continue variabelen is de mediaan de middelste waarde; de helft van de scores is lager en de andere helft hoger. De mediaan van de onderste helft heet het eerste kwartiel (Q1). Die van de bovenste helft heet het derde kwartiel (Q3). Het verschil tussen Q1 en Q3 is de (inter)kwartielafstand. Deze maat is weinig gevoelig voor uitbijters.

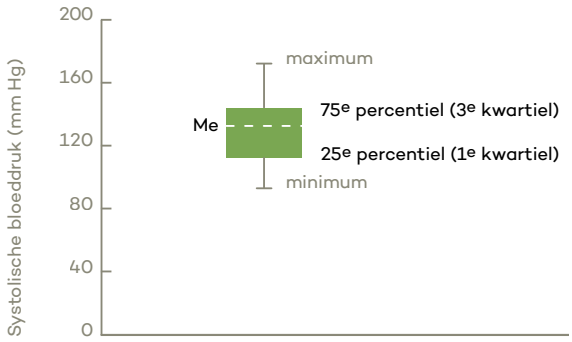
14 Kengetallen weergeven

Je kunt niet alleen frequentieverdelingen grafisch voorstellen. Het is ook mogelijk de centrum- en spreidingsmaten van een frequentieverdeling weer te geven in een grafiek.

14.1 Boxplot

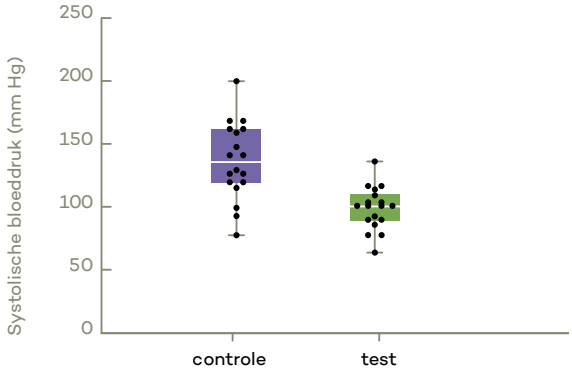
Een boxplot is een presentatie van een (scheve) verdeling waarvoor je gebruikt (zie afbeelding 10):

- de kleinste gemeten waarde (minimum);
- het eerste kwartiel (Q1);
- de mediaan;
- het derde kwartiel (Q3);
- de grootste gemeten waarde (maximum).



Afbeelding 10. *Boxplot: systolische bloeddruk van 50 HsdCpb:WU ratten.*

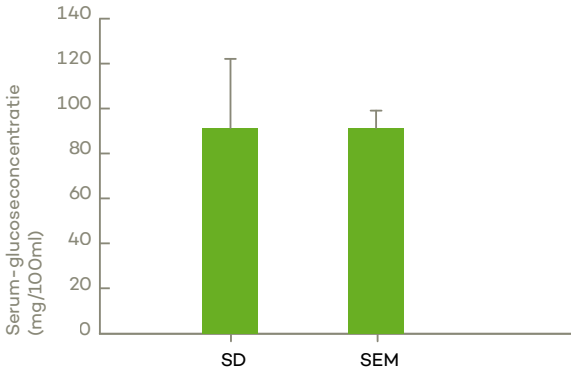
In de box bevindt zich ongeveer 50% van de waarnemingen. Een boxplot wordt ook wel snorredoos genoemd. Een nog fraaiere weergave bereik je door een puntendiagram te combineren met een boxplot (zie afbeelding 11).



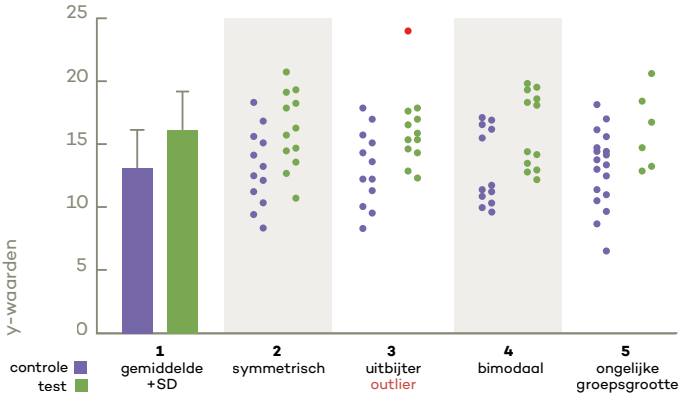
Afbeelding 11. *Puntendiagram gecombineerd met boxplot (ieder symbool in het puntendiagram representeert een individueel meetpunt).*

14.2 Kolomdiagram met standaardafwijking

Dit is een presentatie van een symmetrische frequentieverdeling. Het rekenkundig gemiddelde wordt uitgebeeld als kolom en de standaardafwijking (SD) in de vorm van een T boven op de kolom. In plaats van de SD kun je ook de SEM weergeven (zie afbeelding 12). Let op als je een kolomdiagram gebruikt: datareductie kan misleidend zijn. De kolommen 1 tot en met 5 van afbeelding 13 – die voor wat betreft de verdeling verschillend zijn – resulteren allen in dezelfde gemiddelde waarde en SEM, zoals je kunt zien in kolom 1.



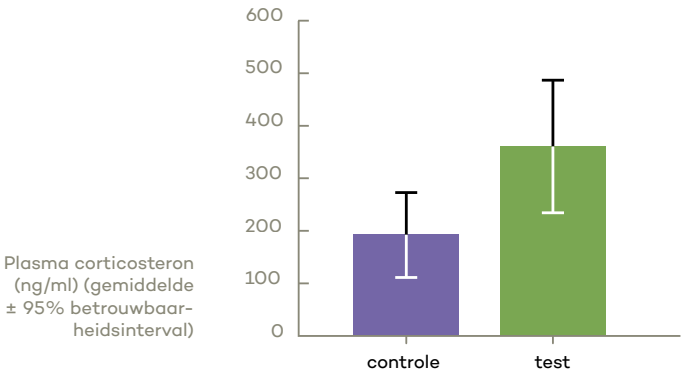
Afbeelding 12. *Gemiddelde serum glucoseconcentratie van 16 gerbils; SD versus SEM.*



Afbeelding 13. *Verschillende verdelingen resulteren in dezelfde gemiddelde waarde en SD.*

14.3 Kolomdiagram met betrouwbaarheidsinterval

Steeds vaker zie je dat bij de uitkomstmaat (meestal het gemiddelde) het betrouwbaarheidsinterval (BI, *of confidence interval*, CI) wordt weergegeven. De BI gebruik je om aan te geven hoe zeker je bent van een geschatte waarde (bijvoorbeeld de gemiddelde waarde) van een groep dieren, waarbij een dier de experimentele eenheid is. Het is een interval waarbinnen je verwacht dat de werkelijke waarde ligt. Vaak wordt het 95%-BI gebruikt, hetgeen populair gezegd betekent dat er 95% kans is dat de gemiddelde waarde van de groep dieren binnen het bereik ligt dat door bovenste en onderste grenswaarde wordt afgebakend (zie afbeelding 14).



Afbeelding 14. *Kolomdiagram: gemiddelde bloedplasma-corticosteron concentratie van C57BL/6J muizen (controle, n = 17; test, n = 23) met bijbehorend 95% BI.*

15 Samenhang tussen twee kenmerken

Als je in het onderzoek 2 of meer variabelen gelijktijdig waarneemt, komt de vraag aan de orde hoe deze variabelen met elkaar samenhangen De samenhang van variabelen bestuderen, is een zeer belangrijk onderdeel van de statistiek. Statistisch onderzoek naar deze samenhang doe je met zogenoemde bivariate technieken: '*bi*' verwijst naar 'twee' en '*variaat*' naar 'veranderlijk' of 'aan het toeval onderhevig'. Onderzoek naar de samenhang van meer dan 2 variabelen voer je uit met multivariate technieken. Er bestaan veel maten voor samenhang; welke je gebruikt, hangt af van het meetniveau van de variabelen. Bij variabelen met minimaal ordinaal niveau en (bijna) volledige ordening onderzoeken we rangcorrelatie, ook Spearman-correlatie genoemd. Bij variabelen met interval- of ratiomeetniveau onderzoeken we productmomentcorrelaties, ook Pearson-correlatie genoemd.

15.1 Kruistabellen

Deze gebruiken we om de samenhang van 2 of meer nominale kenmerken te analyseren. Ze worden ook wel contingentietabellen genoemd: contingent betekent toevallig, of mogelijk. Het aantal mogelijke waarden van beide variabelen moet niet al te groot zijn, omdat de kruistabel anders onoverzichtelijk wordt. Kruistabellen zijn primair bedoeld voor nominale gegevens, maar je kunt ze ook toepassen als je een nominale en een ordinale variabele hebt, ook gemengde verzameling variabelen genoemd. Je plaatst de categorieën van de ene variabele in de rijen, en die van de andere variabele in de kolommen. Als de waarden van de afhankelijke variabele 2 worden bepaald door die van onafhankelijke variabele 1, bestaat er een lichte voorkeur de afhankelijke variabele 2 in de rijen te plaatsen. Een tabel met 2 variabelen, waarbij variabele 1 vier categorieën heeft en variabele 2 drie categorieën, heeft 12 cellen met getallen. Zie tabel 2.

Tabel 2. *Waargenomen frequenties. Vier biotechnici voeren een eenmalige orbitapunctie bij ratten onder ethernarcose uit.*

	Effecten op de stand van de oogbol* (gepuncteerde zijde)			
Biotechnicus	**Exophthalmus**	**Enophthalmus**	**Normaal**	**Totaal**
A	1	2	99	**102**
B	6	7	137	**150**
C	6	10	134	**150**
D	7	16	49	**72**
Totaal	**20**	**35**	**419**	**474**

* *Exophthalmus: het naar voren geplaatst zijn van de oogbol;*
Enophthalmus: het terugzakken van de oogbal in de oogholte.

Het is verstandig om niet naar de absolute aantallen te kijken, maar naar percentages. Uit de absolute frequenties kun je diverse relatieve frequenties berekenen:

- rijpercentages: de frequenties van elke cel als percentage van het bijbehorende rijtotaal;
- kolompercentages: de frequenties van elke cel als percentage van het bijbehorende kolomtotaal;
- totaalpercentages: de frequenties van elke cel als percentage van het totaal generaal.

Als de beide variabelen onafhankelijk van elkaar zijn en er dus totaal geen samenhang is, zal voor elke cel de verwachte (voorspellende) frequentie gelijk zijn aan het product van de randtotalen gedeeld door het totale aantal.

Als in elke cel de waargenomen frequentie gelijk is aan de verwachte frequenties, berust de verdeling blijkbaar op toeval en zijn de indelingscriteria van de variabelen onafhankelijk van elkaar. In de praktijk is er nagenoeg altijd wel een verschil te zien tussen waargenomen en verwachte celfrequenties.

Als het verschil groot is, berusten de waargenomen frequenties waarschijnlijk niet op toeval en is er sprake van een samenhang. Met een statistische toets kun je nagaan of het verband tussen twee of meer nominale kenmerken significant is. We spreken van een significant verband als die in sterke mate de veronderstelling ondersteunt dat het verband niet door toeval is ontstaan, maar door iets anders. Voor de goede orde, niet elk statistisch significant verband zal ook biologisch of therapeutisch van belang zijn. De grootheid die je op basis van de verwachte en waargenomen frequenties kan uitrekenen heet het Chi-kwadraat.

Meer informatie hierover vind je in #7 *Het verband tussen twee nominale kenmerken: kruistabellen*; 2004 van de serie *[basale] statistiek en dierexperimenten* in het tijdschrift '*Biotechniek*'.

15.2 Rangcorrelatie

Bij ordinale (dus kwalitatieve) gegevens gaat het bij de vergelijking met een tweede variabele niet zozeer om de vergelijking van de waarden, maar om vergelijking van de volgorde. We spreken dan van rangcorrelatie. Zijn beide variabelen van interval- of ratiomeetniveau, dan ligt overigens de zogenoemde productmomentcorrelatie van Pearson meer voor de hand.

Een belangrijke maat voor de sterkte van rangcorrelatie is de Spearman-rangcorrelatie-coëfficiënt, ook wel *Spearman's rho* genoemd, afgekort tot r_S; de index $_S$ refereert naar Spearman. De waarden van r_S lopen van -1 tot +1.

Als $r_S = 0$ dan is er geen verband tussen beide variabelen. Als r_S dicht bij +1 ligt, is er een sterk positief verband, en als r_S dicht bij -1 ligt is er een sterk negatief verband. Overigens maakt het voor de waarde van r_S niet uit of je van laag naar hoog dan wel van hoog naar laag rangschikt. Maar je moet het voor beide variabelen wel op dezelfde manier doen. Een redelijk positieve correlatie, bijvoorbeeld 0,65, kan een indicatie voor rangcorrelatie zijn. Het is geen bewijs voor een zeker lineair verband tussen beide variabelen. Of het resultaat significant is, dat wil zeggen afwijkt van 0, oftewel dat de hypothese dat er geen samenhang bestaat weerlegd wordt, moet je statistisch toetsen.

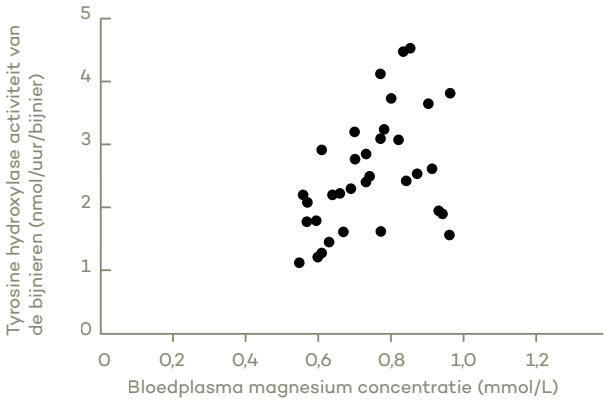
15.3 Productmomentcorrelatie van Pearson

Twee continue variabelen (interval- of ratiomeetniveau) kunnen een rechtlijnig verband met elkaar hebben. Het deelgebied van de statistiek dat de sterkte en de richting van deze lineaire relatie bestudeert, noemen we de (parametrische) correlatie.

15.3.1 Spreidingsdiagram

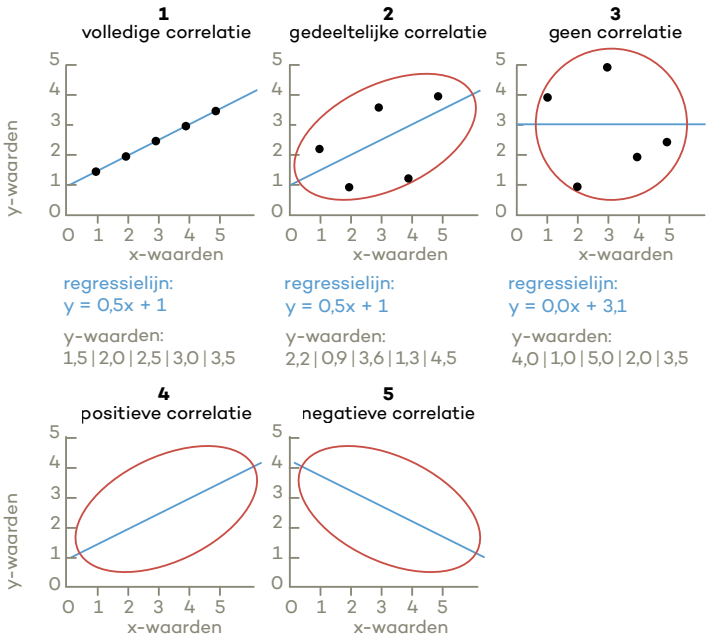
Wanneer het aantal categorieën per variabele klein is, geeft een kruistabel de resultaten op overzichtelijke wijze weer (zie paragraaf 15.1 *Kruistabellen*). In beginsel kun je zo'n tabel ook voor continue variabelen maken, maar dan moet je ze eerst in klassen indelen. Daardoor gaat informatie verloren.

Een handiger manier om de relatie tussen twee continue variabelen X en Y te laten zien is het spreidingsdiagram, ook wel strooidiagram genoemd. Dit is een tweedimensionale grafische weergave van de gegevens (zie afbeelding 15). Als er sprake is van een oorzakelijke variabele (niet-beïnvloedbare of onafhankelijke variabele) en een afhankelijke variabele, dan wordt doorgaans de horizontale x-as voor de onafhankelijke variabele X gebruikt; op de y-as wordt de afhankelijke variabele Y afgezet. Er ontstaan dan waarnemingsparen die elk worden weergegeven door 1 punt (elk punt heeft een x- en een y-waarde). Het spreidingsdiagram geeft een globaal beeld van de relatie tussen de beide variabelen.



Afbeelding 15. *Spreidingsdiagram: de variabele tyrosine hydroxylase activiteit in de bijnieren is uitgezet tegen de variabele bloedplasma-magnesiumconcentratie.*

De punten vormen samen een puntenwolk. Hoe meer de punten van zo'n wolk een rechte strook vormen, des te sterker is het verband (correlatie) tussen de beide variabelen. Als alle punten van de puntenwolk op een centrale, rechte lijn liggen, is er een volledige correlatie. Als er geen enkel verband bestaat tussen de twee grootheden is er geen correlatie. Vaak is sprake van een gedeeltelijke correlatie; hierbij vormen de punten in het spreidingsdiagram een puntenwolk waarvan de tendens kan worden aangegeven door een centrale lijn. Deze correlatie kan positief of negatief zijn (zie afbeelding 16). Een positieve correlatie betekent dat als de waarde van de variabele op de x-as groter wordt, dat bij de variabele op de y-as ook zal gebeuren. Bij een negatieve correlatie worden de waarden van de andere variabele dan juist kleiner.



Afbeelding 16. *De mate van correlatie en positieve en negatieve correlatie.*

15.3.2 Productmomentcorrelatiecoëfficiënt

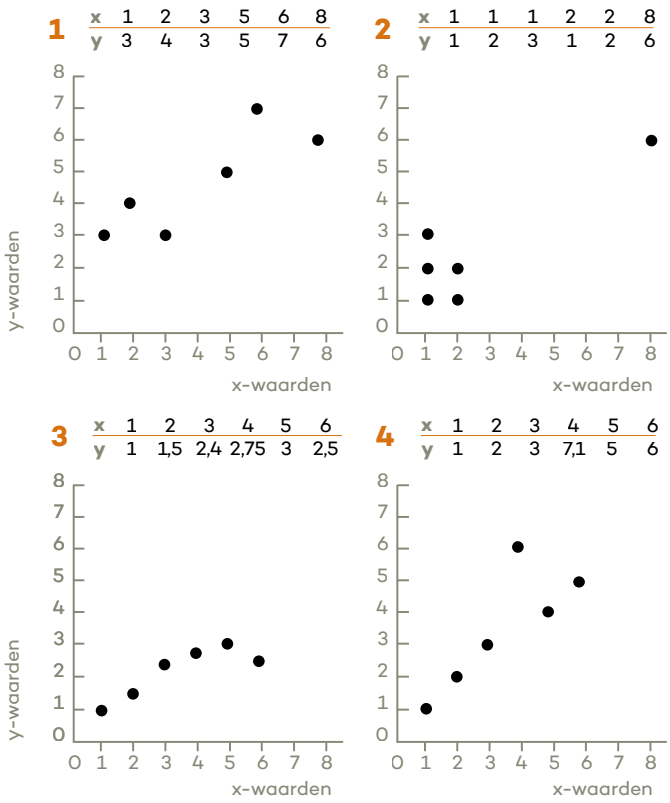
Het spreidingsdiagram geeft dus een indruk van de samenhang tussen beide variabelen. De mate van samenhang kan worden uitgedrukt in een getal. Het meest toegepast wordt hiervoor de productmomentcorrelatiecoëfficiënt volgens Pearson; kortweg de correlatiecoëfficiënt, afgekort tot *r*. De correlatiecoëfficiënt ligt altijd tussen -1 en +1 (zie tabel 3). In het algemeen zeggen we dat twee variabelen gecorreleerd zijn als $r < -0{,}5$ of $> 0{,}5$ is. Hoe dichter de correlatiecoëfficiënt ligt bij -1 of + 1, des te sterker het verband.

Tabel 3. *Interpretatie van de Pearson correlatiecoëfficient.*

Correlatiecoëfficient *r*	**Interpretatie**
Vanaf -1 tot -0,8	sterk negatief
Vanaf -0,8 tot -0,5	matig tot sterk negatief
Vanaf -0,5 tot -0,3	zwak tot matig negatief
Vanaf -0,3 via 0 tot 0,3	zwak negatief via geen tot zwak positief
Vanaf 0,3 tot 0,5	zwak tot matig positief
Vanaf 0,5 tot 0,8	matig tot sterk positief
Vanaf 0,8 tot en met 1	sterk positief

Een sterke correlatie is niet noodzakelijk een significant verband. Bij een zeer klein aantal waarnemingen is de kans zeer groot dat de waarnemingsparen ongeveer op een lijn liggen en dus een sterke correlatie suggereren. Voor een gegeven correlatiecoëfficiënt hangt het significantieniveau af van het aantal waarnemingen. Je moet daarom niet alleen *r* vermelden, maar ook het aantal waarnemingsparen en eventueel het significantieniveau.

Een significante correlatie tussen twee variabelen toont nooit een oorzakelijk verband aan. Bij een positieve correlatie tussen bloeddruk en cholesterolgehalte van het bloed weet je niet of een hoge bloeddruk tot een hoog cholesterolgehalte leidt of omgekeerd. Dit kun je mogelijk wel experimenteel aannemelijk maken. Wanneer een geïnduceerde stijging van bijvoorbeeld het cholesterolgehalte gepaard gaat met een significante stijging van de bloeddruk is er wellicht een direct oorzakelijk verband. Maar een hoog cholesterolgehalte kan zelf gecorreleerd zijn met een derde variabele zoals stress, en mogelijk is deze stress de echte oorzakelijke factor voor de waargenomen correlatie.



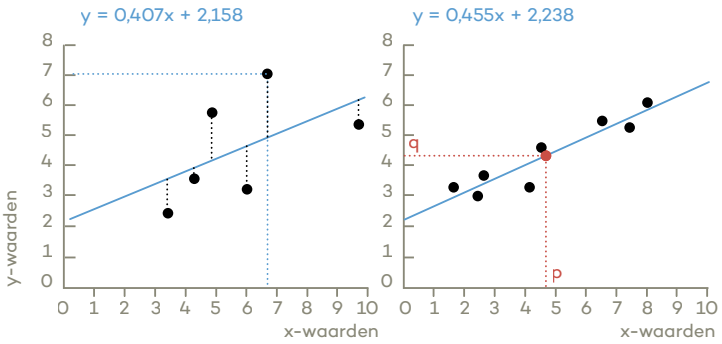
Afbeelding 17. *Vier scorepatronen met gelijke correlatiecoëfficiënt: r = 0,85.*

De interpretatie van de correlatiecoëfficiënt is niet altijd eenvoudig. Er zijn veel factoren met soms een aanzienlijke invloed op de grootte van de correlatiecoëfficiënt. Verschillende datasets (zie afbeelding 17) kunnen een zelfde *r* hebben, maar toch zeer verschillend van aard zijn. Het kwadraat van de correlatiecoëfficiënt geeft de gemeenschappelijke variantie van beide variabelen weer: zo is er bij een correlatie van 0,7 ongeveer 50% variantie.

15.4 Lineaire regressie

Wanneer in het (x,y)-spreidingsdiagram de puntenwolk de vorm van een strook heeft, kun je 'uit de hand' door het lengte-midden van de strook een rechte lijn trekken. Deze zogenoemde regressielijn benadrukt de algemene tendens van het verband tussen de beide variabelen. Het is natuurlijk beter om een best passende rechte lijn te berekenen (zie afbeelding 18). De algemene vergelijking van de regressielijn is y = a + bx. Hierin is b de richtingscoëfficiënt van de rechte. Regressiecoëfficiënt a is het intercept: daar waar de rechte de y-as doorsnijdt als x = 0. De regressielijn geeft bij elke x een passende y-waarde, met andere woorden: de x-waarde voorspelt

de y-waarde. Deze statistische methode – die als doel heeft de relatie tussen variabelen weer te geven via een rechte lijn –noemen we lineaire regressie.

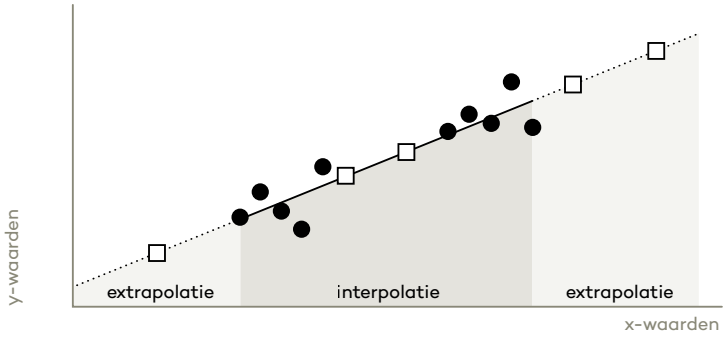


Afbeelding 18. *Afstand van punten tot de best passende lijn en het zwaartepunt van de regressielijn.*

De vergelijking van een regressielijn bereken je via de zogenoemde kleinste-kwadraten-methode. Het principe hiervan is dat de som van de kwadraten van de verticaal gemeten afstanden van de punten tot de lijn zo klein mogelijk moet zijn. Eén punt van de regressielijn is uit de waarnemingsgegevens te halen. Dit punt wordt bepaald door de gemiddelde waarde (p) van de gegevens op de x-as en door de gemiddelde waarde (q) van de gegevens op de y-as. Het punt waarin de ordinaten in p en q elkaar snijden, is het zwaartepunt (ook wel evenwichtspunt) van de regressielijn en de puntenwolk(zie afbeelding 18). Alle statistische analyseprogramma's en ook Excel® kunnen de lineaire regressiecoëfficiënten berekenen.

15.4.1 Lineair interpoleren en extrapoleren

Vaak ken je van een rechtlijnig verband slechts een beperkt aantal waarden. Tussenliggende waarden kun je bepalen via lineaire interpolatie (zie afbeelding 19). Wil je voorspellingen doen voor verder gelegen waarden, dan pas je extrapolatie toe. Hierbij is voorzichtigheid geboden, want het is helemaal niet zeker dat het lineaire verband zich bij vergroting of verlaging van de onafhankelijke variabele (x-as) op dezelfde wijze ontwikkelt. Daarom is het goed de geëxtrapoleerde regressielijn te stippelen.



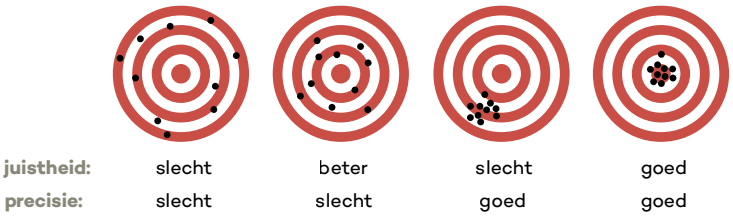
Afbeelding 19. *Extra- en interpolatie.*
● zijn gemeten waarden,
□ werden door inter- of extrapolatie berekend, onder de veronderstelling dat er een lineaire samenhang bestaat tussen de x- en y-waarden.

15.4.2 Kromlijnige verbanden

Vaak blijkt al uit het spreidingsdiagram dat er geen rechtlijnig verband is tussen twee variabelen. Als het verband kromlijnig is, kun je beter geen lineaire regressie toepassen. Soms kun je door transformatie van één of beide variabelen het verband toch rechtlijnig maken. Zo kun je bij een exponentieel verband via logaritmische transformatie een lineaire betrekking definiëren, waarna je lineaire regressie kunt toepassen.

16 Juistheid, precisie, nauwkeurigheid en uitbijters

De juistheid is de mate van overeenstemming tussen het gemiddelde van een reeks meetwaarden en de werkelijke waarde (of de als werkelijk aangenomen waarde) van de te bepalen grootheid. Als maat voor de juistheid wordt gewoonlijk het verschil in overeenstemming genomen: de juistheid geeft informatie over de systematische afwijking (bias) van een methode (zie afbeelding 20).



Afbeelding 20. *Juistheid en precisie.*

Precisie – ook wel toevallige afwijking genoemd – is de mate van spreiding (SD) in de meetresultaten die je krijgt door de meting een aantal malen, onder vastgelegde condities, op hetzelfde monster uit te voeren. Hierbij onderscheiden we herhaalbaarheid en reproduceerbaarheid. Zie afbeelding 20.

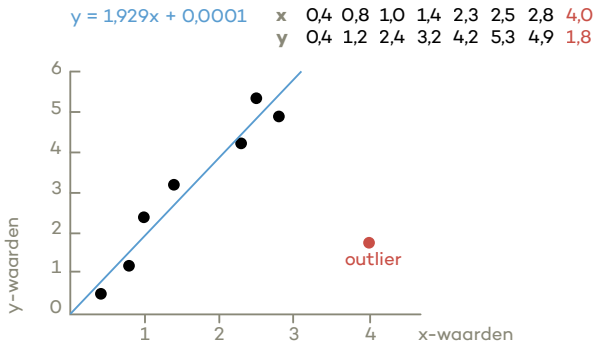
Herhaalbaarheid is de precisie die je krijgt met dezelfde methode bij een identiek materiaal onder dezelfde omstandigheden: dezelfde persoon voert die uit met dezelfde apparatuur, in hetzelfde laboratorium en op hetzelfde tijdstip of met een korte tussentijd.

Reproduceerbaarheid is de precisie die je krijgt met dezelfde methode bij identiek materiaal onder verschillende omstandigheden: verschillende personen voeren die uit met verschillende apparatuur, in verschillende laboratoria en op verschillende tijdstippen.

Nauwkeurigheid is de toevallige afwijking (precisie) en de systematische afwijking (juistheid) samen.

Een uitbijter of uitschieter (ook wel *outlier* genoemd) is een onverwacht hoge of lage uitkomst (zie afbeeldingen 13 en 21). Deze verbreekt de tendens van de overige waarnemingen uit een verzameling (dataset). Detectie van uitbijters is een van de belangrijkste taken na het verzamelen van gegevens. Uitbijters kunnen de resultaten van statistische analyses zeer sterk beïnvloeden. Soms zijn uitbijters gemakkelijk herkenbaar als simpele waarnemings- of registratiefouten, zoals overschrijffouten, rekenfouten, afleesfouten of het gebruik van verkeerde hoeveelheden. De documentatie raadplegen over afwijkingen van het protocol tijdens het uitvoeren van een onderzoek kan helpen uitbijters te identificeren. Zijn dieren bijvoorbeeld op een meetdag opvallend inactief geweest en blijkt dat het op deze dag in het testlaboratorium enorm lawaaierig

was door bouwwerkzaamheden, dan kan dit een oorzaak voor de lage activiteitswaarden zijn. Wellicht werd al in de documentatie genoteerd dat de dieren opvallend inactief waren.



Afbeelding 21. *Een uitbijter (outlier).*

Meestal is het lastiger te bepalen of een extreme uitkomst een uitbijter is. Gelukkig bestaat hiervoor een aantal statistische methoden zoals Dixon's Q-toets, de toets van Grubbs, berekenen van Cook's D waarden (bij lineaire regressie) en uitsluiting op basis van een betrouwbaarheidsinterval. Je kunt aan de correctheid van een meetgegeven twijfelen als die meetwaarde meer dan drie standaarddeviaties van het gemiddelde van de andere meetwaarden ligt. Houd altijd in gedachten dat veel ontdekkingen gebaseerd zijn op uitzonderlijke uitkomsten die alerte onderzoekers niet hebben weggegooid.

17 Schatten en toetsen

Onderzoekers proberen vaak informatie over een proefdierpopulatie te krijgen uit waarnemingen bij een beperkt aantal experimentele eenheden, dus uit een steekproef. Deze werkwijze wordt de verklarende statistiek genoemd. Hierin wordt een onderscheid gemaakt tussen schattingsmethoden en toetsingsmethoden.

17.1 Steekproef

Je kunt op verschillende manieren een steekproef verkrijgen. Als alle dieren uit de populatie dezelfde kans hebben om in de steekproef te worden opgenomen, spreken we van een aselecte steekproef. Worden de dieren niet op basis van toeval uit een populatie genomen, dan is sprake van een selecte steekproef. Zorg je ervoor dat in een steekproef uit een bepaalde proefdierpopulatie evenveel mannelijke als vrouwelijke dieren voorkomen, dan is de steekproef representatief voor het kenmerk geslacht. Met behulp van de schattingstheorie kun je op grond van waarnemingen bij de steekproef een statistische uitspraak doen over een populatieparameter, bijvoorbeeld de centrum- of spreidingsmaat.

17.2 Hypothesen

Een iets andere werkwijze volg je bij wetenschappelijke bewijsvoering: veronderstellingen (hypothesen) toetsen. Onderzoek begint met het formuleren van de onderzoeksvraag. Daarin geef je als onderzoeker zo concreet mogelijk aan welke vraag het onderzoek moet beantwoorden. Vanuit de onderzoeksvraag formuleer je vervolgens hypotheses. Bij een statistische toetsing stel je eerst een nulhypothese (H_0) en een alternatieve hypothese (H_A) op. Stel, je wilt onderzoeken of een bepaald therapeuticum effect heeft op virusinfecties bij laboratoriummuizen. Je formuleert dan een nulhypothese waarin wordt uitgedrukt dat het therapeuticum geen effect heeft op virusinfecties. Daarnaast stel je een alternatieve hypothese op die uitdrukt dat het therapeuticum wel effect heeft op virusinfecties. De nulhypothese is de hypothese die je tracht onderuit te halen met de statistische toets, terwijl de alternatieve hypothese geldt als de nulhypothese niet geldt. Welke statistische toets je gebruikt, is afhankelijk van de opzet van het onderzoek en het type uitkomstmaat.

Over het algemeen berust de keuze van de nulhypothese op de ervaring die je als onderzoeker hebt met soortgelijke onderzoekingen, en wil je onderzoeken of een nieuwe situatie opvallend afwijkt van bevindingen in het verleden. Kun je de nulhypothese op basis van de statistische toets niet verwerpen, dan zeggen we wel dat we deze accepteren, zij het bij gebrek aan bewijs. De nulhypothese verwerp je als je de waarnemingsuitkomsten ten opzichte van eerdere bevindingen als afwijkend aanmerkt. Dit komt er populair gezegd op neer dat de waarnemingsuitkomsten verschillen met wat verwacht was, en dat ze niet meer op toeval lijken te berusten. We spreken van een significante uitkomst als die in sterke mate de veronderstelling ondersteunt dat het verschil niet door toeval is ontstaan, maar door iets anders: je verwerpt dus de nulhypothese.

De bedoelde veronderstelling betreft meestal het verschil tussen groepen, vaak de controlegroep en de experimentele groep. Maar de toetsing kan tot goede en foute besluiten leiden. Zie tabel 4.

Tabel 4. *Mogelijke besluiten bij de toetsing van hypothesen.*

Besluit	Nulhypothese (H_0) juist	Nulhypothese onjuist
Nulhypothese niet verwerpen	goed	Type II-fout (ß)
Nulhypothese verwerpen	Type I-fout (α)	goed

17.3 Type I-fout

Een type I-fout (fout van de eerste soort of alfafout) is de fout die je maakt door met een toets een juiste nulhypothese toch te verwerpen (zie tabel 4). De kans op een type I-fout heet de onbetrouwbaarheid van de toets en is in feite de *P*-waarde. De *P*-waarde is een getal tussen 0 en 1, dat wordt berekend op basis van de toetsstatistiek, de verdeling van de toetsstatistiek en het aantal vrijheidsgraden van de toets. (De berekening van het aantal vrijheidsgraden – een maat voor het 'aantal stukken informatie in de toets' – hangt af van de soort toets.) Het wel of niet verwerpen van de nulhypothese vindt meestal plaats op basis van een waarde waarboven de onbetrouwbaarheid niet mag uitstijgen: de onbetrouwbaarheidsdrempel, ook wel alfa-niveau of significantieniveau genoemd. Vandaar dat we type I-fout ook wel de alfafout noemen. Bij het overschrijden van de alfa (dat wil zeggen $P < \alpha$) spreken we van een statistisch significant resultaat, met andere woorden zeggen we dat het resultaat waarschijnlijk wordt veroorzaakt door iets anders dan louter toeval (zie afbeelding 22). Het alfa-niveau kan ook in procenten worden uitgedrukt, bijvoorbeeld $\alpha = 5\%$ in plaats van $\alpha = 0{,}05$. Soms gebruiken we de volgende aanduidingen om aan te geven hoe significant (bij een $\alpha = 0{,}05$) een resultaat is (notatie: omschrijving, *P*-waarde):

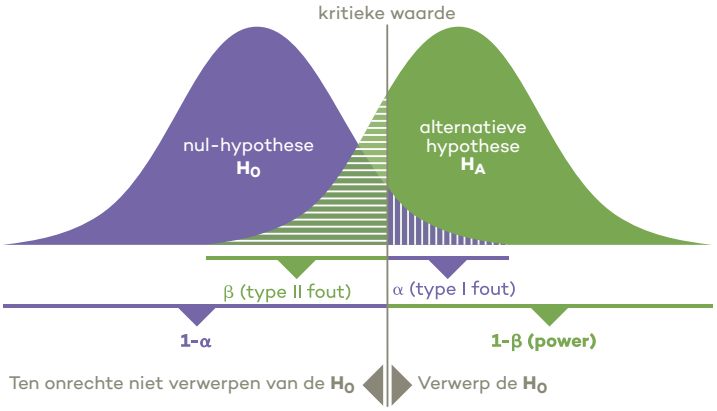
~: zwakke aanwijzing voor een verschil tussen de vergeleken condities of groepen – neiging tot significantie, $0{,}05 \leq P < 0{,}10$;

*: aanwijzing voor een verschil tussen de vergeleken condities of groepen – significantie, $0{,}01 \leq P < 0{,}05$;

**: sterke aanwijzing voor een verschil tussen de vergeleken condities of groepen – sterke significantie, $0{,}001 \leq P < 0{,}01$;

***: zeer sterke aanwijzing voor een verschil tussen de vergeleken condities of groepen – zeer sterke significantie, $P < 0{,}001$.

Maar let op bij het hanteren van het woord significant(ie): wat significant is voor de onderzoeker is misschien niet interessant voor het publiek!



Afbeelding 22. *Type I- en type II-fouten.*

17.4 Type II-fout

Een type II-fout (fout van de tweede soort of bètafout) maak je door een nulhypothese te accepteren en niet te verwerpen, terwijl deze niet klopt (zie tabel 4). De kans op een type II-fout duiden we meestal aan met β (bèta). Je zou ook kunnen zeggen: β is de kans die een onderzoeker accepteert om een bestaand effect niet te detecteren. Net als het alfa-niveau van een toets wordt het bèta-niveau ook uitgedrukt als een waarde tussen 0 en 1 (of tussen 0 en 100%). De kans om deze fout niet te maken (1-β, of 100-β%) noemen we het onderscheidingsvermogen (ook wel bewijskracht of *power*) van een toets (zie afbeelding 22). Een goed opgezette studie hoort de kans op een type II-fout te beperken. Het instellen van de *power* van een test is complex omdat de *power* afhangt van verschillende factoren:

- het alfa-niveau;
- de componenten van de berekening van de toetsingsgrootheid, bijvoorbeeld het verschil tussen de gemiddelden van de controle- en testgroep (= behandelingseffect), de grootte van de standaardafwijking;
- het aantal experimentele eenheden, meestal dieren.

Belangrijk is dat een toename van de steekproefgrootte (aantal dieren per groep) een toename van de *power* van de statistische toets tot gevolg heeft. De keuze van de steekproefgrootte is in dierexperimenteel onderzoek een belangrijke aangelegenheid.

17.5 Een- of tweezijdig toetsen

Je kunt tweezijdig of eenzijdig toetsen. Tweezijdig toetsen is aangewezen wanneer het doel van het experiment is een behandelingseffect te onderzoeken zonder te specificeren in welke richting dit effect gaat. Eenzijdig toetsen is aangewezen wanneer op grond van kennis vooraf er slechts van een eenzijdig alternatief sprake kan zijn. Oftewel: als je vooraf weet dat een afwijking van de nulhypothese slechts in één richting kan zijn, hoef je met de onbetrouwbaarheid van afwijkingen in de andere richting geen rekening te houden. Je kunt dan een eenzijdige toets uitvoeren. Bij terecht eenzijdig toetsen is de *power* groter. Eenzijdigheid of tweezijdigheid is een theoretische kwestie die zich afspeelt in het hoofd van de onderzoeker. Statistische computer-programma's kunnen niet weten of een vraagstelling eenzijdig is, en dus ook niet wat de richting dan is. Daarom geven deze programma's meestal een tweezijdige *P*-waarde. Je kunt een tweezijdige *P*-waarde omrekenen in een eenzijdige door de helft te nemen (*P*/2).

18 Verslaglegging van dierproeven

Een dierproef is pas voltooid als je het verkregen resultaat aan anderen kunt rapporteren. Dat betekent doorgaans dat je de bevindingen van je dierexperiment moet vastleggen in een wetenschappelijk tijdschrift of in een toegankelijk rapport. Biotechnici en dierverzorgers zijn meestal niet de verantwoordelijke auteurs van zo'n publicatie of rapport, maar dragen wel de gegevens aan. Op basis daarvan stelt de verantwoordelijke auteur zo'n artikel of rapport vervolgens op. Die gegevens kunnen het best aangeleverd worden via een duidelijk verslag. Je moet hierbij in het achterhoofd houden dat veel wetenschappelijke tijdschriften die resultaten van dierproeven publiceren, de zogenaamde ARRIVE-richtlijn volgen. Deze richtlijn, opgesteld door het NC3Rs (https://www.nc3rs.org.uk/), is daarom leidend bij de verslaglegging van dierproeven. Daarom moet je in elk geval in je verslag verwerken:

- een gedetailleerde beschrijving van het ontwerp van de dierproef met daarbij het aantal en de soort groepen (controle- en testgroepen), en ook wat de experimentele eenheid is;
- het aantal experimentele eenheden voor elke groep, én het totaal aantal experimentele eenheden voor de dierproef. Geef ook aan hoeveel dieren de studie in totaal omvat. Vermeld bovendien hoe de grootte van elke groep tot stand is gekomen;
- de criteria die je gehanteerd hebt om dieren in of uit te sluiten van het experiment, en of er al dan niet sprake was van (onverwachte) uitval van proefdieren;
- op welke manier de randomisatie en blindering heeft plaatsgevonden;
- welke registratie- en meetapparatuur er gebruikt zijn;
- welke uitleesvariabelen er vastgesteld zijn en wat de belangrijkste variabele is;
- details over de gebruikte proefdieren, zoals diersoort, populatie, geslacht, leeftijd, lichaamsgewicht, algemene gezondheidstoestand, microbiologische status, merktekens, voer- en wateropname;
- verdere aspecten van zowel de huisvesting als de omgeving van de proefdieren, bijvoorbeeld huisvestingsvorm, temperatuur, relatieve luchtvochtigheid, luchtbehandeling, verlichting (lichtintensiteit en licht-donkercyclus), geluid, soort voer en drinkwater, manier van voeder- en drinkwaterverstrekking, bedding-, nest- en verrijkingsmateriaal;
- gevolgde procedures. Beschrijf ze voor elke groep. Denk hierbij bijvoorbeeld aan de toedieningswijze en dosis van stoffen;
- samenvatting van de resultaten via kengetallen per groep in grafieken of tabellen;
- de manier waarop je de resultaten statistisch hebt geanalyseerd (als je dat hebt gedaan).