Minder dieren door een gerichte volgorde in de vraagstellingen

Steven Teerenstra[1], Rienke Uijen[2], Cathalijn Leenaars[3]
[1]Department for Health Evidence, section Biostatistics, Radboud Institute for Health Sciences, Radboudumc, Nijmegen, Nederland, contact: steven.teerenstra@radboudumcn.nl,
[2]Radboudumc, Nederland
[3]Hannover Medical School, Duitsland

In veel dierstudies gebruiken we meerdere groepen. Deze groepen corresponderen bijvoorbeeld met verschillende behandelingen en/of verschillende soorten dieren (bijvoorbeeld met een verzwakt of een normaal immuunsysteem). Vaak willen we deze groepen met elkaar vergelijken, en daarvoor gebruiken we statistische toetsen. Misschien doe je als biotechnicus niet heel vaak statistiek, maar je kan wel meedenken bij het opzetten van nieuwe dierproeven. We geven hier een idee hoe je in sommige gevallen dieren kunt besparen door de vraagstellingen voor de verschillende groepen in een gerichte en slimme volgorde te beantwoorden.

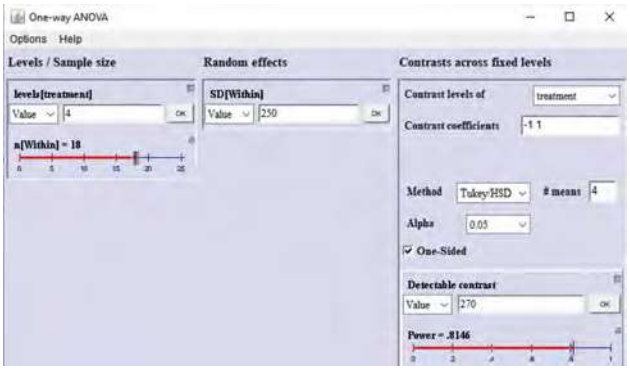
Laten we gelijk naar een voorbeeld gaan.

De onderzoeksvraag is: kan het effect van een actief middel verbeterd worden door het toevoegen van een hulpmiddel (versterker)? We denken aan een experiment met de volgende vier groepen muizen: 1. controle, 2. actief middel, 3. versterker en 4. combinatie van actief middel en versterker. Een verschil van 270 (van de betreffende meeteenheid, bijvoorbeeld micrometer) tussen groep 4 en groep 2 willen we in ieder geval kunnen detecteren. We weten dat de standaarddeviatie 250 is. De standaardberekening gaat als volgt. Met het programma 'PiFace' van Rus Lenth (https://homepage.stat.uiowa.edu/~rlenth/Power/#download)
kiezen we het menu 'Balanced Anova>Differences/Contrast' en vullen dan in (Afb. 1):

- levels=4: dit is het aantal groepen;
- SD[within]=250: dit is de standaarddeviatie;
- one-sided test: we willen alleen een uitspraak doen als er een verbetering is; als we ook de mogelijkheid willen hebben om te concluderen dat er een verslechtering is (als dat zo is), dan moeten we tweezijdig toetsen.
- alpha=0.05: type I-fout, dat is de maximale kans op een vals positieve uitkomst die we willen toelaten;
- detectable contrast=270: het minimale verschil.

We hebben bij de 'Method' verschillenden keuzes, maar de standaardkeuze 'Tukey/HSD' ligt het meest voor de hand ('Scheffé' en 'Tukey/HSD' verschillen niet wezelijk; 'Bonferroni' is al snel

strenger dan 'Tukey/HSD'; 'Dunnett' gaat op als we alleen groepen willen vergelijken met één en dezelfde controle, en de keuze 't' is voor exploratief onderzoek).
Door n[Within] (het aantal muizen per groep) te variëren vinden we dat de methode Tukey/HSD 18 muizen per groep nodig heeft, dus 4 x 18 = 72 in totaal, om een power van tenminste 80 % te krijgen.



Afbeelding 1. De powerberekening voor een verschil van 270 middels one-way ANOVA, Tukey/HSD in PiFace.

Iets preciezer gezegd, we hebben met Tukey's methode 80% kans ('power') om alle verschillen te detecteren die 270 of meer zijn. 'Alle verschillen' in de zin dat het niet uitmaakt of dat verschil tussen groep 1 vs 2, 1 vs 3, 1 vs 4, 2 vs 3, 2 vs 4, of 3 vs 4 zit.

Dat is veel meer informatie dan we nodig hebben. Wat we willen weten is:

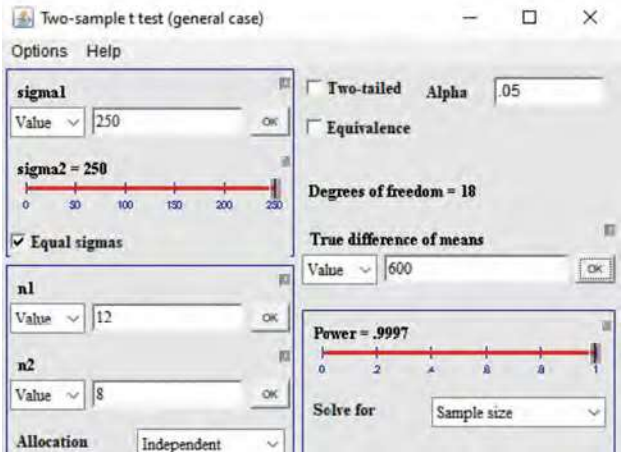
A. Is het actieve middel en versterker beter dan het actieve middel alleen (4 vs 2)?
B. Doet de versterker van zichzelf iets (3 vs 1), of alleen in combinatie met het actieve middel?
C. Is het experiment een technisch succes? Dat wil zeggen is het actieve middel wel actief in deze populatie muizen (2 vs 1)? Zo niet, dan is het experiment niet (goed) interpreteerbaar.

Een goede en gerichte volgorde lijkt: eerst c, dan a, dan b toetsen. (Natuurlijk doen we het experiment zelf wel met alle groepen tegelijkertijd om tijdseffecten/batch-effecten te voorkomen).

Bij zo'n vooraf vastgestelde testvolgorde kunnen we een 'fixed sequence multiple testing procedure' gebruiken: eerst toetsen we c bij alpha = 0.05; alleen als dat statistische significant is, toetsen we a bij alpha=0.05, en alleen als dat ook statistisch significant is, toetsen we b bij alpha=0.05.
Als je vooraf zo'n slimme testvolgorde plant, blijft de totale kans op één of meer vals positieve bevindingen (type I-fout) onder de 0.05. Dit principe wordt ook wel zo beschreven: als de toets van c statistisch significant is, dan blijft er nog 'ongebruikte alpha' van die toets over (alpha=0.05) en die kun je opnieuw gebruiken voor de toets van b, etc.
Het voordeel van de 'fixed sequence testing'-procedure is dat je de 0.05 niet hoeft te corrigeren voor meervoudig testen, waarvoor je meer dieren nodig zou hebben. De prijs die je betaalt is dat je de volgorde van toetsen moet vastleggen voordat je de data hebt (anders zou je opportunistisch kunnen kiezen). Een ander nadeel is: als bijv. c niet statistisch significant is, dan kun je a en b niet verder toetsen. Maar dat is in dit geval niet zo'n probleem: als het experiment geen technisch succes is (c), dan hoeven we niet verder te kijken. En als de versterker in combinatie met het actieve middel geen effect heeft (a), dan hoeven we ook niet te weten of een deel van het combinatie effect door de versterker alleen komt (b).

Wat betekent dit voor de power?

We weten uit vorig onderzoek dat het verschil tussen het actieve middel en controle groot is (gemiddeld 600). Verder zijn we pas geïnteresseerd in de versterker als een middel op zichzelf als het verschil tussen controle en versterker boven de 400 uitkomt. Met een t-toets bij alpha=0.05 (eenzijdig) en een SD=250 geven twaalf muizen met het actieve middel versus acht muizen met controlemiddel een power van 99.97 % (Afb. 2).



Afbeelding 2. De powerberekening in PiFace voor een t-test tussen placebo en actief middel.

Op dezelfde manier geven 12 vs 12 muizen een power van 82.07 % voor een verschil van 270 (combinatie vs actief) en 8 vs 8 muizen geven een power van 91.85 % voor een verschil van 400 (versterker vs controle). Deze powerniveaus gelden als de t-testen in drie aparte experimenten worden uitgevoerd. Echter de toetsen worden achter elkaar uitgevoerd en de toetsen gebruiken deels dezelfde groepen muizen. Een berekening waarbij we ons niet te rijk rekenen geeft een power van 99.98 % voor de eerste toets, 99.97 % x 82.07 %=82.05 % voor de tweede toets (als de eerste statistisch significant is) en 82.05 % x 91.85 %=75.36 % power voor de derde toets (als de eerste en tweede toets statistisch significant zijn). De werkelijke power is hoger doordat groepen gedeeld worden (correlatie tussen de toetsen). Bovendien wordt de power nog hoger als we de analyse met een ANOVA doen, omdat dan de SD en dus de effecten preciezer geschat worden.
Omdat de laatste vraag minder belangrijk is, vinden we wat minder power voor deze vraag acceptabel. (Als we ook voor deze vraag 80 % power willen hebben dan blijken vier extra muizen in de versterkergroep voldoende volgens een exactere berekening met een ANOVA-analyse).
In totaal dus 8 (controle) +12 (actief middel) + 8 (versterker) + 12 (combinatie)= 40 muizen in totaal in plaats van de oorspronkelijke 72.

Samenvattend: als we op een zinvolle manier de vraagstellingen na elkaar kunnen beantwoorden (bijvoorbeeld een logische volgorde of een prioritering), dan kunnen we kijken of een 'fixed sequence testing' ons een besparing van proefdieren oplevert.

Bron

1. Fixed Sequence Procedure: p. 56 in Multiple Testing Problems in Pharmaceutical Statistics. Edited By Alex Dmitrienko, Ajit C. Tamhane, Frank Bretz. Chapman & Hall. 2010.