

Statistiek en meten: wat moet je daarover weten?

Onder redactie van Diana Dolmans, Cees van der Vleuten, Albert Scherpbier en Ineke Wolfhagen

Bij het lezen van publikaties over onderzoek van onderwijs worden lezers regelmatig geconfronteerd met allerlei statistische begrippen, waarvan de betekenis veel lezers niet geheel duidelijk is. Daarom heeft de redactie besloten een reeks artikelen te publiceren waarin belangrijke onderwerpen op het gebied van statistiek en meten aan de orde worden gesteld. De redactie van deze reeks wordt gevormd door twee gastredacteuren (Diana Dolmans en Cees van der Vleuten) en twee leden van de BMO-redactie (Albert Scherpbier en Ineke Wolfhagen).

De keuze van onderwerpen is gebaseerd op veel gebruikte statistische technieken en begrippen. De nadruk ligt op de betekenis en interpretatie hiervan. Op deze manier wordt getracht een bijdrage te leveren aan de verdere professionalisering van docenten binnen het medisch onderwijs. In het achtste artikel van deze reeks staat power-analyse centraal. Dit is het laatste artikel van de reeks. We hopen dat deze reeks een bijdrage heeft geleverd aan een beter inzicht in een aantal statistische onderwerpen.

De bepaling van de steekproefgrootte: power-analyse

L.W.T. Schuwirth

Termen die aan bod komen:

hypothesetoetsing, type I-fout (α), type II-fout (β), power, standard error of the mean (Se), Z-transformatie.

In het artikel van Van Breukelen over statistische toetsen en betrouwbaarheidsintervallen kwam onder andere de bepaling van de statistische power van een steekproef aan bod.¹ Deze beschrijving was gebaseerd op de bepaling hiervan nadat het experiment was uitgevoerd. Dit artikel poogt hierop een kleine aanvulling te zijn, door een beschrijving te geven van een methode, waarmee een schatting vooraf gemaakt kan worden van het benodigde aantal waarnemingen om een gewenste power te verkrijgen. Om het geheel voor alle lezers in een duidelijk kader te plaatsen zullen een aantal basisideeën over hypothesetoetsing kort beschreven worden.

Steekproeven

In wetenschappelijke studies wordt vaak geprobeerd om op basis van een steekproef een uitspraak te doen die buiten het bereik van de steekproef ligt. Men wil de resultaten van de steekproef generaliseren. Dit stelt eisen aan de wijze waarop de steekproef wordt samengesteld en gebruikt. Bijvoorbeeld: een actualiteitenrubriek wil voorafgaande aan de kamerverkiezingen een inschatting maken van wat de zetelverdeling in de Tweede Kamer zal zijn. Een onderzoeker ondervraagt 25 Nederlanders hierover (allen mooi verdeeld naar leeftijd, geslacht, streek van herkomst, etcetera) en maakt waarschijnlijk de fout een te kleine steekproef genomen te hebben. Een tweede onderzoeker ondervraagt 2500 Amsterdammers en heeft een steekproef die misschien groot genoeg is maar niet representatief voor de Nederlandse samenleving. Een derde onder-

zoeker vraagt 2500 Italianen (allen mooi verdeeld naar leeftijd, etcetera) en heeft een steekproef die helemaal geen deel uitmaakt van de te onderzoeken populatie.

Voor de rest van dit artikel zullen we er vanuit gaan dat we spreken over steekproeven die een goede afspiegeling zijn van de te onderzoeken populatie met een meting die redelijk valide en betrouwbaar is. Dan blijft de vraag naar de grootte van de steekproef. Dit artikel poogt een uitleg te geven over de manier waarop de benodigde grootte van een steekproef vooraf bepaald kan worden.

Hypothesetoetsing

Een onderzoeker wil in een deelstudie vaststellen of er een verschil bestaat in moeilijkheid tussen open vragen en gesloten vraagvormen (in een examen). Hij neemt hiervoor de proef op de som met een gesplitste toets van 80 vragen (40 open vragen en dan 40 multiple-choice vragen). Hij vindt in zijn studie, dat de gemiddelde score op de open vragen 33.0% is en op de multiple-choice vragen 43.2%. In beide gevallen is de standaarddeviatie (SD) 10.0%. Hij staat nu voor de vraag of hij hieruit mag concluderen dat dit verschil in moeilijkheid alleen binnen deze meting (door toeval) bestaat. Dit gaat de onderzoeker na aan de hand van *hypothesetoetsing*.

De hypothese van de onderzoeker is dat er geen werkelijk verschil bestaat. Dit wordt de nulhypothese (H_0) genoemd. In formule ziet het er zo uit: $\mu_1 = \mu_2$ (waarbij bijvoorbeeld μ_1 het gemiddelde van alle mogelijke multiple-choice vragen aangeeft en μ_2 dat van alle mogelijke open vragen, binnen het desbetreffende kennisdomein). Er bestaat dan in feite ook een alternatieve hypothese (H_a) die waar wordt wanneer de nulhypothese verworpen wordt: $\mu_1 \neq \mu_2$ (er is wel sprake van een verschil in moeilijkheid). De onderzoeker kan dus twee foute beslissingen nemen. Hij kan een zogenaamde *type I fout* maken: hij kan de nulhypothese verwerpen (dus zeggen dat er een verschil

Tabel 1. Hypotheses op basis van een steekproef en de bijbehorende foutensoorten

| Steekproef | Werkelijkheid | |
|--|--|--|
| | Geen verschil | Wel verschil |
| Geen verschil (H_0 niet verwerpen) | Terechte beslissing ($1 - \alpha$) | Type II fout β |
| Wel verschil (H_0 verwerpen) | Type I fout α | Terechte beslissing ($1 - \beta = \text{power}$) |

bestaat), terwijl er in werkelijkheid toch geen verschil bestaat. Maar hij kan ook een zogenaamde *type II fout* maken: hij kan de nulhypothese aannemen (zeggen dat er geen verschil bestaat) terwijl er in werkelijkheid wel een verschil is. Zulke fouten kunnen altijd optreden, maar de onderzoeker kan een grens stellen waaronder de kans op de fouten moet vallen. Als de kans hoger is dan die grens dan noemt de onderzoeker zijn bevindingen niet significant. Deze grens wordt aangeduid met respectievelijk α en β . De α geeft dus de kans aan dat ten onrechte gezegd wordt dat er wel een verschil in gemiddelden bestaat. De β geeft de kans aan dat ten onrechte gezegd wordt dat er geen verschil in gemiddelden bestaat. De kans dus om een verschil te vinden, als dit daadwerkelijk ook bestaat is $1 - \beta$. Deze waarde wordt de *power* genoemd. In tabel 1 zijn deze mogelijkheden weergegeven.

Steekproefgemiddelden

De steekproef die de onderzoeker met de examens heeft getrokken is een van de vele mogelijke steekproeven die hij had kunnen trekken. De gemiddelde score die hij vond op de open vragen is een van de vele gemiddelden die hij had kunnen vinden. In feite heeft hij maar één steekproef en is het moeilijk voor te stellen dat hij hieraan conclusies kan verbinden.

Wel is het zo dat als alle mogelijke steekproefgemiddelden een normale verdeling vertonen, de onderzoeker de kans kan inschatten dat hij een bepaalde waarde zou vinden gegeven een bepaald werkelijk populatiegemiddelde. Bijvoorbeeld indien een onderzoeker zou willen vaststellen of medisch onderwijskundigen gemiddeld een hogere diastolische tensie hebben dan normaal geacht wordt. Ook kan de kans worden ingeschat dat twee gemiddelden echt van elkaar verschillen, zoals in ons voorbeeld van de twee vraagsoorten. Steekproefgemiddelden benaderen een normale verdeling wanneer de variabele normaal verdeeld is. Maar dit geldt ook wanneer de verdeling van de variabele niet normaal is mits de steekproef groter is dan 30.

De standaarddeviatie van de verdeling van alle mogelijke steekproefgemiddelden wordt *standard error of the mean (Se)* genoemd, en wordt berekend door de standaarddeviatie te delen door de wortel uit het aantal waarnemingen. Deze Se speelt een belangrijke rol bij het schatten van kansen bij beslissingen ten aanzien van gemiddelden, maar hierover later meer.

Z-transformatie

De kans dat in een normaalverdeling een bepaalde waarde wordt gevonden, wordt weergegeven door de hoogte van de curve bij die waarde. Eigenlijk is het zo, dat de kans dat waarden binnen een bepaald interval worden gevonden, weergegeven wordt door de oppervlakte onder de curve. Nu is de formule van een normaalverdeling nogal ingewikkeld en zou dit voor iedere statistische toets zeer complexe berekeningen inhouden. Indien men met tabellen zou willen werken zouden deze gemaakt moeten worden voor iedere grootte van gemiddelde en standaarddeviatie. Daarom wordt de curve gestandaardiseerd via een *Z-transformatie*. Hierbij wordt de curve in feite eerst heen en weer geschoven, zodanig dat het gemiddelde op 0 valt, en vervolgens wordt de

curve zodanig breder of smaller gemaakt dat de standaarddeviatie 1 wordt. Zo krijgt iedere waarde in de oorspronkelijke curve een bijbehorende Z-waarde. Voor deze Z-curve of standaard normaalverdeling is wel een tabel gemaakt waarin de oppervlakten onder de curve van verschillende intervallen beschreven zijn. Het is dus mogelijk om bijvoorbeeld het gebied te vinden waarbinnen 5% van de waarnemingen vallen die het verst van het gemiddelde verwijderd liggen (hoogste en laagste 2.5% in dit geval). Dit wil dus zeggen dat als een waarde in dat gebied gevonden wordt de kans maar 5% is dat bij dit gemiddelde een dergelijke waarde door toeval gevonden wordt.

In feite kan deze hele procedure ook gedaan worden voor de verdeling van steekproefgemiddelden (die ook vaak normaal is zoals we zagen). Hierbij wordt de zaak dan vaak wel een beetje omgedraaid, er wordt een Z-verdeling rond het steekproefgemiddelde bepaald en berekend wordt hoe groot het interval rond dit gemiddelde is waarbinnen (met een 95% kans) het werkelijke gemiddelde ligt.

Power-analyse

Nu zou het jammer zijn als de onderzoeker zijn hele experiment heeft afgerond, en tot de conclusie komt dat er noch een voldoende grond is om de hypothese aan te nemen of te verwerpen (dat noch een acceptabele α als β gehaald worden). Of andersom als de steekproef te groot blijkt te zijn geweest en met de helft van de observaties ook al een hoog significantieniveau zou zijn bereikt. Met behulp van power-analyse is een goede schatting te maken van het aantal benodigde waarnemingen. Hiervoor moeten echter wel een viertal grootheden gekozen of geschat worden.

De eerste twee grootheden die gekozen moeten worden zijn natuurlijk de gewenste α en de power ($1 - \beta$). Percentages die hiervoor gebruikt worden, worden meestal arbitrair gekozen (voor de α zijn 0.05 of 0.01 en voor de β 0.10 of 0.20 veel gebruikte waardes). De

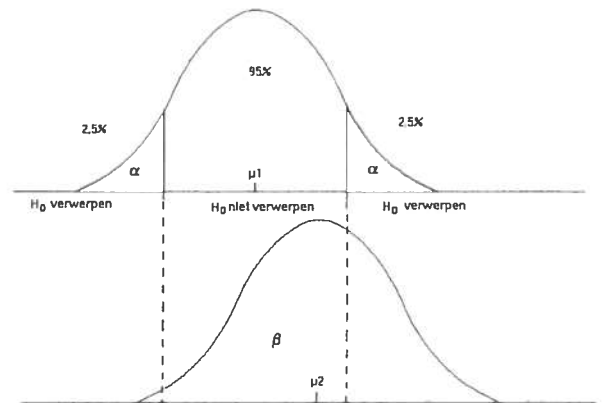
percentages geven hierbij de kans aan dat een beslissing ten onrechte gemaakt wordt; een α van 0.05 betekent dus dat de kans 5% is dat ten onrechte geconcludeerd wordt dat er een verschil in gemiddelden bestaat. Vaak wordt aangeraden deze beide grootheden zo klein mogelijk te kiezen.

Vervolgens moet een keuze gemaakt worden voor het minimale verschil dat opgemerkt moet kunnen worden. De grootte van het minimale verschil moet in relatie tot de vraagstelling bekeken worden. Stel dat men in het kader van de preventie van hart- en vaatziekten van een groep managers wil bepalen of zij een intensief stress-reductieprogramma moeten volgen, waarbij als maat voor het risico de diastolische tensie wordt gebruikt. Het heeft dan geen zin om een groep mensen te identificeren bij wie de diastolische tensie 2 mmHg hoger ligt dan normaal en deze dan vervolgens deze cursus te laten volgen. De verschillen die interessant zijn voor de bepaling of iemand een stress-reductieprogramma zou moeten volgen liggen eerder in de orde van grootte van 10 - 20 mmHg.

Tenslotte moet een inschatting gemaakt worden van de standaarddeviatie van de populatie. Dit kan gedaan worden op basis van de literatuur of op basis van eerdere (pilot) experimenten die door de onderzoeker zijn uitgevoerd.

Bij de berekening van de benodigde waarnemingen om een bepaalde power te bereiken wil men dus een inschatting verkrijgen van de kansen dat wanneer er een verschil bestaat dit ook gevonden wordt. Dit is dus de combinatie van de gebieden die buiten de Z-waarden voor α en buiten de Z-waarde voor β liggen. Figuur 1 is hiervan een voorstelling. Alle gebieden die niet gearceerd zijn in het onderste gedeelte van de figuur geven de power aan van de steekproef.

Er zijn nu twee mogelijkheden: de onderzoeker wil weten of een bepaalde gevonden waarde binnen de normaalwaarden valt (zoals in het voorbeeld met de diastolische tensie van



Figuur 1: Het gebied waarin de type I fout (α) en de type II fout (β) valt (power is alles wat niet binnen het gebied van β valt: $1 - \beta$)

medisch onderwijskundigen) of een onderzoeker wil weten of twee verschillende steekproefgemiddelden ook twee gemiddelde verschillen in de populatie representeren (zoals in het voorbeeld met de verschillende soorten toetsvragen). Voor beide vraagstellingen wordt bij power-analyse een iets andere formule gebruikt. Het voorbeeld van de diastolische meting van de tensie staat uitgewerkt in voorbeeld 1.

Tenslotte

Hoewel het doen van power-analyse weer een extra berekening in de toch al vrij ingewikkelde materie van hypothesetoetsing inhoudt, blijkt het in de praktijk waardevol. Niet alleen kan het aangeven dat meer proefpersonen nodig zijn dan oorspronkelijk geschat (en zo dus een studie redden die anders geen tot weinig waarde zou hebben), ook het omgekeerde kan gebeuren, wat geld en tijd spaart. In een extreem geval kan zelfs duidelijk worden dat zoveel observaties nodig zijn om de vraagstelling beantwoord te krijgen dat het logistiek niet haalbaar is en de studie beter niet gestart kan worden.

Een voorbeeld

Een onderzoeker zou willen vaststellen of medisch onderwijskundigen gemiddeld een hogere diastolische tensie hebben dan normaal geacht wordt. In dit geval betreft het dus een vergelijking van een speciale groep met een normaalwaarde. De onderzoeker kiest zijn α en β respectievelijk op 0.05 en 0.10, dat wil zeggen hij accepteert een kans van 5 procent dat hij onterecht zou kunnen concluderen dat er wel een verschil was, en een kans van 10 procent dat hij onterecht zou concluderen dat er geen verschil was. De power is hierbij dan $100\% - 10\% = 90\%$. Het minimale verschil dat de onderzoeker wil kunnen vinden bedraagt 10% en op grond van de literatuur stelt hij vast dat de standaarddeviatie circa 15 mmHg is.

Omdat het een vergelijking van een gemiddelde van een speciale groep met een normaalwaarde betreft wordt de volgende formule gebruikt:

$$n = \left[\frac{(Z_{\alpha} - Z_{\beta}) \sigma}{\mu_1 - \mu_0} \right]^2$$

In deze formule is:

- Z_{α} de Z-waarde die een gebied uitsluit van 5%
- Z_{β} de Z-waarde die een gebied uitsluit van 10%
- μ_1 de gemiddelde waarde van de te onderzoeken groep
- μ_0 de gemiddelde waarde van de norm groep
- σ de standaarddeviatie
- n het aantal benodigde waarnemingen.

De Z_{α} moet hierbij tweezijdig een gebied van 5% uitsluiten, dus 2.5% aan iedere zijde van de verdeling. De Z-waarden van de grenzen van deze 2.5%-gebieden zijn -1.96 en +1.96. De Z_{β} daarentegen hoeft alleen eenzijdig een gebied uit te sluiten. De grens hiervan ligt bij een Z-waarde van 1.28.

Wanneer de formule nu ingevuld wordt dan ziet ze er als volgt uit:

$$n = \left[\frac{(-1.96 - 1.28) 15}{10} \right]^2 = \left[\frac{(3.24) 15}{10} \right]^2 = \left[\frac{48.6}{10} \right]^2 = 23.62$$

Het aantal waarnemingen dat hieruit komt hoeft natuurlijk geen geheel getal te zijn maar wordt naar boven afgerond. De onderzoeker heeft voor zijn vraagstelling de diastolische tensie van 24 proefpersonen nodig.

Literatuur

1. Van Breukelen G. Statistische Toetsen en Betrouwbaarheidsintervallen. Bulletin Medisch Onderwijs 1993; 12: 73-78.

Aanbevolen literatuur

- Dawson-Saunders B, Trapp RG. Basic and Clinical Biostatistics. 2^e editie. Norwalk: Appleton & Lange, 1994.
- Hays WL. Statistics for the social sciences. Plymouth en Londen: Clarke, Doble & Brendon Ltd., 1978.
- Slotboom A. Statistiek in woorden. De meest voorkomende termen en technieken. Groningen: Wolters-Noordhoff, 1987.

DE AUTEUR

L. Schuwirth is als arts verbonden aan het Evaluatie Project Geneeskunde van de Faculteit der Geneeskunde en houdt zich bezig met de evaluatie van studieresultaten.

Correspondentie-adres:

L.W.T. Schuwirth. Vakgroep Onderwijsontwikkeling en Onderwijsresearch, Rijksuniversiteit Limburg, Postbus 616, 6200 MD Maastricht.

Opdrachten

1. Leg uit wat de "eenheid" is van de Z-waarden in de standaard normaal verdeling.
2. Een onderzoeker is geïnteresseerd om een verschil tussen twee populaties te vinden van 10% met behulp van twee steekproeven. Hij doet power-analyse en komt uit op een benodigd aantal waarnemingen van 72. Hij besluit nu dat hij liever toch een verschil van maar 5% zou willen vinden (met dezelfde α en dezelfde power). Leg uit wat dit voor effect heeft op het benodigde aantal waarnemingen.
3. Een onderzoeker heeft bij het doen van een power-analyse geschat dat de standaarddeviatie in beide te vergelijken populaties 10 bedraagt, op basis hiervan heeft hij zijn steekproefgrootte bepaald. Later bemerkt hij dat deze standaarddeviatie eigenlijk maar 5 is. Is zijn schatting van de steekproefgrootte nu te hoog of te laag geweest. Leg uit waarom.