

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 niet geheel duidelijk is. Daarom heeft de redactie besloten om 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 tweede artikel van deze reeks staan statistische toetsen en betrouwbaarheidsintervallen centraal.

Statistische Toetsen en Betrouwbaarheidsintervallen

G. van Breukelen

Termen die aan bod komen:

gemiddelde, proportie, standard error, normale verdeling, steekproef en populatie, gepaarde en ongepaarde steekproeven, nulhypothese, toetsing, voorspellings- en betrouwbaarheidsinterval, significant, p-waarde, type I en II fout, 'power'

In het vorige artikel in deze reeks over beschrijvende statistiek, maakten we kennis met steekproeven, de normale verdeling en maten als gemiddelde en standaarddeviatie. Statistische rapportage van onderzoek begint altijd met beschrijvende statistiek. Hierbij worden de gegevens die verzameld zijn bij het onderzoek, samengevat in figuren en tabellen aan de hand van eerder genoemde maten. Indien er een *steekproef* is genomen, wil men meestal ook uitspraken doen over de *populatie* waar de steekproef uit stamt. We willen bijvoorbeeld op basis van de bloeddruk in een 'aselecte' (willekeurige) steekproef van honderd bejaar-

de Nederlanders iets zeggen over de bloeddruk in de hele populatie Nederlandse bejaarden. Maar de ene bejaarde is de andere niet, en dus is de ene steekproef de andere niet, waardoor 'steekproeftoeval' optreedt. Daarom is toetsende statistiek nodig, en daarover gaat dit artikel. De basisaanname achter toetsende statistiek is dat de steekproef 'aselect' (willekeurig) is getrokken uit een zekere populatie waarover men uitspraken wil doen, bijvoorbeeld Nederlandse bejaarden, kinderen met leukemie, of CARA-patiënten. Ook wordt verondersteld dat de populatie veel groter is dan de steekproef.

We beginnen met een eenvoudige toets, namelijk die voor één *gemiddelde*. Toetsen voor het verschil tussen twee gemiddelden of percentages zijn uitbreidingen hiervan. We houden vast aan een goed gebruik: Statistische maten berekend op de steekproef worden aangeduid met gewone letters, maar maten met betrekking tot de populatie worden aangeduid

met griekse letters. Op deze manier is altijd duidelijk of we spreken over de populatie of de steekproef.

Steekproevenverdeling en standard error

Stel we observeren in een aselechte steekproef van honderd Nederlandse bejaarden een gemiddelde diastolische bloeddruk van 84 mm Hg. Dit gemiddelde wordt aangeduid met de letter M ('mean'). Wat zegt dit nu over de gemiddelde diastolische bloeddruk in de hele populatie van Nederlandse bejaarden, aangeduid met de griekse letter μ ? We vermoeden dat μ ongeveer 84 zal zijn, maar enige afwijking is mogelijk door steekproeftoeval. Een nieuwe steekproef kan immers een andere M opleveren. Deze variatie in M wordt weergegeven in de 'steekproevenverdeling' van M . Deze 'steekproevenverdeling' is de frequentieverdeling van alle steekproefgemiddelden, M -waarden, die gevonden wordt indien zeer vaak een aselechte steekproef getrokken wordt uit de populatie en indien voor elk van deze steekproeven het steekproefgemiddelde, M , wordt berekend. Het gemiddelde van al deze steekproefgemiddelden is gelijk aan het populatiegemiddelde μ . De standaarddeviatie (spreiding) van deze steekproevenverdeling heet de 'standard error of the mean' (SE). Deze *standard error* is gelijk aan de standaarddeviatie van de individuele bloeddrukwaarden in de populatie (σ) gedeeld door de wortel uit het aantal mensen in de steekproef (N). In formulevorm: $SE = \sigma / \sqrt{N}$. Dit betekent dat de standard error van het steekproefgemiddelde toeneemt, naarmate de standaarddeviatie van de individuele bloeddrukwaarden in de populatie (σ) groter wordt, en de steekproef minder mensen omvat (N kleiner).

Indien alle bejaarden dezelfde bloeddruk hebben ($\sigma = 0$), levert elke steekproef hetzelfde gemiddelde (M) op, dat bovendien gelijk is aan het populatiegemiddelde (μ). Er is dan geen steekproeftoeval en we kunnen volstaan met een omvang van de steekproef van $N = 1$. Is

daarentegen sprake van een grote spreiding in bloeddruk (heterogene populatie), dan maakt het voor het steekproefgemiddelde veel uit welke bejaarden toevallig in de steekproef meegenomen worden. Bij steekproeven van $N = 1$ is het steekproefgemiddelde (M) gelijk aan de bloeddruk van slechts één persoon, en is de SE gelijk aan de standaarddeviatie van de individuele bloeddrukwaarden in de populatie (σ). Naarmate N stijgt, daalt de SE tot bijna 0. De afwijkingen van de steekproefgemiddelden ten opzichte van het populatiegemiddelde worden dus steeds kleiner bij een toename van het aantal personen in de steekproef.

Meestal is er echter sprake van slechts één steekproef en één steekproefgemiddelde (M). Wat zegt dit dan over het populatiegemiddelde (μ)? Door een betrouwbaarheidsinterval te berekenen kan bepaald worden binnen welke marges μ ligt en met behulp van statistische toetsen kan bepaald worden of een veronderstelde μ -waarde met onze steekproefuitslag te rijmen is. Hiervoor moet echter aan minstens één van de volgende voorwaarden zijn voldaan. Ten eerste, de variabele die onderzocht wordt, in dit geval de bloeddruk, moet normaal verdeeld zijn in de populatie. Ten tweede, de steekproefomvang moet voldoende groot zijn (minimaal $N = 30$).

Toetsing, voorspellingsinterval en betrouwbaarheidsinterval

Bij een *normale verdeling* ligt 95% van de waarnemingen binnen een afstand van 2 (exact: 1.96) standaarddeviaties ten opzichte van het gemiddelde. Dit impliceert dat 95% van de aselechte steekproeven uit de populatie een steekproefgemiddelde (M) oplevert dat minder dan 2 SE van het populatiegemiddelde (μ) ligt, en slechts 5% van de steekproeven een gemiddelde oplevert dat verder dan 2 SE van μ af ligt (onder de voorwaarde dat de variabele die onderzocht wordt normaal verdeeld is in de populatie of de steekproefomvang groter is

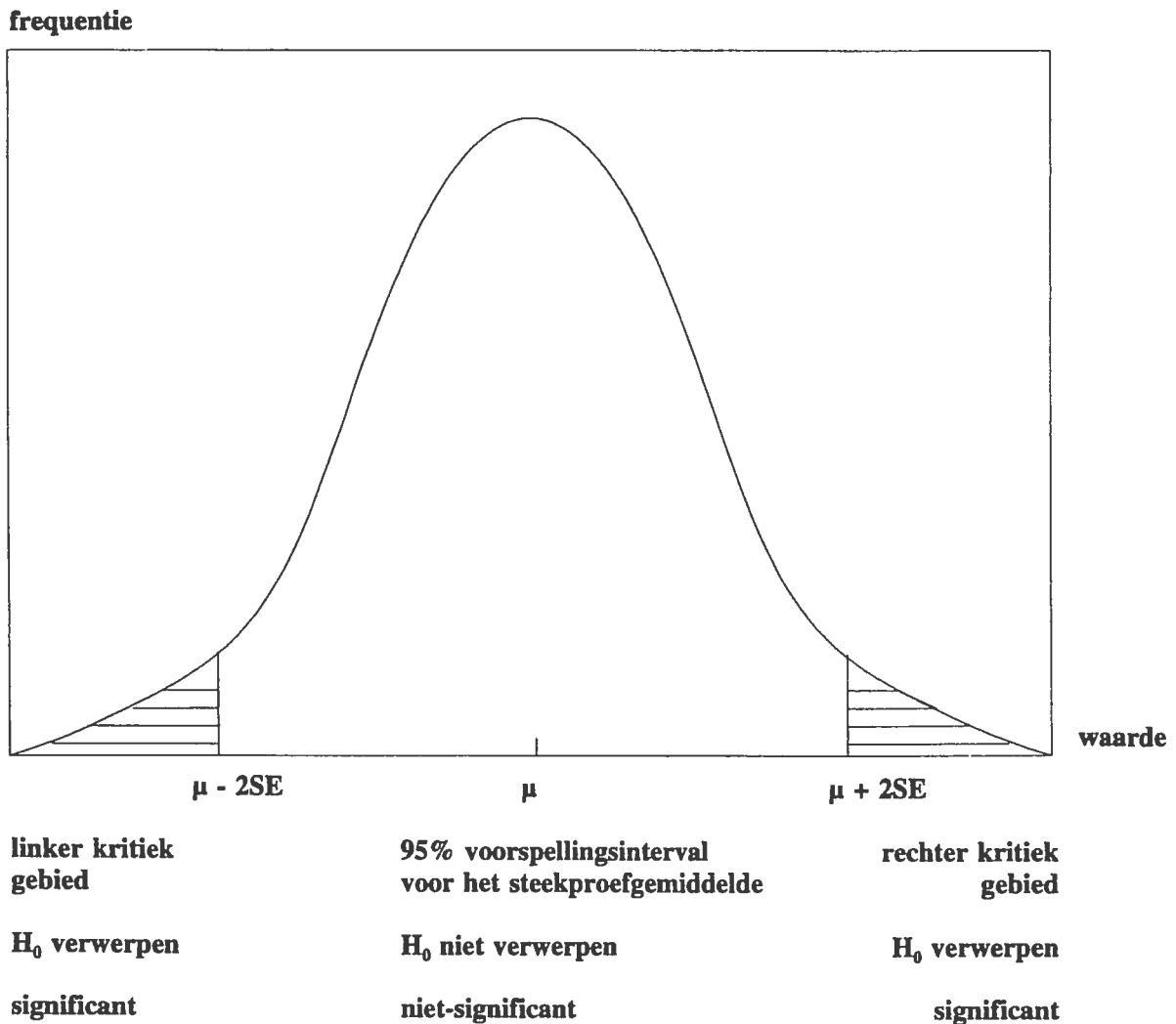
dan 30). Hierop zijn toetsing en betrouwbaarheidsinterval gebaseerd.

Eerst zal ingegaan worden op *toetsing*. We veronderstellen dat het populatiegemiddelde (μ) 90 is. We noemen die waarde μ_0 en de veronderstelling dat μ_0 gelijk is aan 90, de nulhypothese (H_0). De *alternatieve hypothese* (H_a) luidt: μ is niet gelijk aan 90. Uitgaande van H_0 verwachten we een steekproefgemiddelde (M) dat niet verder dan 2 SE van 90 afligt. Bijvoorbeeld, indien de standaarddeviatie van de populatie (σ) gelijk is aan 10 en de steekproefomvang (N) gelijk is aan 100, is de SE gelijk aan 1, immers $SE = 10 / \sqrt{100}$. We verwachten hierbij een steekproefgemiddelde tussen 88 en 92 (respectievelijk: $\mu_0 - 2 SE$ en $\mu_0 + 2 SE$). Indien het gevonden steekproefgemiddelde binnen dit *voorspellingsinterval* ligt, is er geen aanleiding om de veronderstelde nulhypothese, ($H_0: \mu = 90$), te verwerpen. Vinden we echter een steekproefgemiddelde buiten dit interval van 88 tot 92, dan verwerpen we H_0 ten gunste van de alternatieve hypothese H_a . Volgens H_0 zal slechts 5% van de aselechte steekproeven (met een steekproefomvang van 100) een gemiddelde buiten dit interval opleveren. Indien de eerste de beste steekproef een gemiddelde buiten dit interval oplevert, zal H_0 wel fout. In het voorbeeld werd een steekproefgemiddelde van 84 gevonden. Dit gemiddelde valt buiten het voorspellingsinterval, waardoor de nulhypothese ($H_0: \mu = 90$) verworpen wordt. De gemiddelde diastolische bloeddruk van bejaarden is kennelijk niet gelijk aan 90 mm Hg, maar lager.

Het gebied buiten het voorspellingsinterval heet het 'kritieke gebied' en de grenzen van het interval heten 'kritieke waarden'. Dit is weergegeven in figuur 1. Indien het steekproefgemiddelde in het kritieke gebied ligt, wordt H_0 dus verworpen. We spreken dan van een 'statistisch significant resultaat'. Indien het steekproefgemiddelde niet in het kritieke gebied ligt, wordt H_0 niet verworpen en is er sprake van een 'statistisch niet-significant resultaat'. Dit is eveneens weergegeven in figuur 1.

Bij toetsing redeneren we vanuit een aanname omtrent de populatie, de nulhypothese (H_0), naar de steekproef toe. We voorspellen namelijk binnen welke marges het steekproefgemiddelde zal liggen, uitgaande van een verondersteld gemiddelde van de populatie (μ_0). Bij een *betrouwbaarheidsinterval* redeneren we omgekeerd. We nemen niets aan over het populatiegemiddelde, maar berekenen vanuit het steekproefgemiddelde een interval dat met een grote kans (95%) het ware, onbekende populatiegemiddelde omvat. Indien het steekproefgemiddelde (M) minder dan 2 SE van het populatiegemiddelde (μ) afligt, ligt μ ook minder dan 2 SE van M af. Bijvoorbeeld als de SE gelijk is aan 1, dan levert een steekproefgemiddelde (M) van 84 mm Hg een 95% betrouwbaarheidsinterval op van 82 tot 86 mm Hg (respectievelijk $M - 2 SE$ en $M + 2 SE$). Dit interval omvat dus waarschijnlijk het ware, onbekende populatiegemiddelde (μ). Het veronderstelde populatiegemiddelde van 90 ligt buiten het betrouwbaarheidsinterval van 82 tot 86 mm Hg. Dit komt overeen met de eerder uitgevoerde toets waarin de nulhypothese, die veronderstelde dat het populatiegemiddelde gelijk is aan 90 mm Hg, werd verworpen. Zowel de statistische toetsing als het betrouwbaarheidsinterval tonen aan dat het populatiegemiddelde niet gelijk is aan 90. Het betrouwbaarheidsinterval geeft bovendien aan dat het ware populatiegemiddelde vermoedelijk ergens tussen 82 en 86 mm Hg ligt.

Soms wordt niet uitgegaan van 95%, maar van 90% of 99% betrouwbaarheidsintervallen. De grenzen van het betrouwbaarheidsinterval worden dan niet berekend op basis van 2 SE (exact: 1.96 SE), maar op basis van 1.65 SE indien het een 90% betrouwbaarheidsinterval betreft, of op basis van 2.58 SE, indien het een 99% betrouwbaarheidsinterval betreft. Een grotere zekerheid gaat dus gepaard met een breder en minder nauwkeurig interval.



Figuur 1. Het 95% interval en het kritieke gebied voor het gemiddelde van een steekproef

Z-toets, T-toets, een- en tweezijdig toetsen, p-waarde

In de praktijk wijkt statistische toetsing op een aantal punten af van het bovenstaande. We bespreken hieronder kort de belangrijkste afwijkingen. Om te beginnen wordt de toets uitgevoerd via de zogenaamde Z-grootheid. Deze Z-grootheid wordt berekend door het veronderstelde populatiegemiddelde (μ_0) af te trekken van het gevonden steekproefgemiddelde (M) en dit te delen door de 'standard error of the mean' (SE). In formule: $Z = (M - \mu_0) / SE$. De Z-grootheid geeft dus de afwijking weer tussen het gevonden steekproefgemiddelde (M) en het veronderstelde populatie-

gemiddelde (μ_0), gedeeld door de SE, ofwel een 'gestandaardiseerde afwijking'. Als het echte populatiegemiddelde inderdaad μ_0 is, dan is Z 'standaardnormaal' verdeeld, dat wil zeggen normaal verdeeld met een gemiddelde van 0 en een standaarddeviatie van 1. We verwachten dan een Z-waarde tussen -2 en +2. Vinden we een Z-waarde buiten dit interval, dan ligt het steekproefgemiddelde meer dan 2 SE van het veronderstelde populatiegemiddelde (μ_0) af. Dat wijst er op dat μ_0 niet het echte populatiegemiddelde is, en we H_0 dus moeten verwerpen. De waarden +2 en -2 worden dan ook de kritieke Z-waarden genoemd. De kans op een Z-waarde links van -2 is 2.5%, en de

kans op een Z-waarde rechts van +2 is 2.5%. De kans op een Z-waarde buiten het interval van -2 tot +2 is dus 5%. In ons voorbeeld levert een steekproefgemiddelde van 84 een Z-waarde op van -6 ($[84 - 90] / 1$). Deze waarde ligt links van -2 en dus duidelijk in het kritieke gebied, waardoor H_0 wordt verworpen.

Een tweede afwijking van het bovenstaande is dat statistische software niet zegt of H_0 wordt verworpen, maar een 'significantie-niveau' oftewel '*p-waarde*' (overschrijdskans) geeft. De *p-waarde* is de kans op de gevonden of een extremere Z, onder de aanname dat H_0 klopt. Indien de *p-waarde* kleiner is dan 0.05, ligt Z in het kritieke gebied, buiten het interval van -2 tot +2, waardoor H_0 verworpen wordt en er sprake is van een statistisch significant resultaat. Is de *p-waarde* groter dan 0.05, dan ligt Z in het interval, waardoor H_0 niet verworpen wordt en er sprake is van een statistisch niet-significant resultaat. Het weergegeven van de *p-waarde* is genuanceerder dan slechts weergegeven of H_0 wel of niet verworpen wordt op grond van de wat arbitraire 5% grens. Een *p-waarde* van 6% is immers vrijwel even klein als een *p-waarde* van 4%.

Een derde afwijking heeft betrekking op de aanname dat de standaarddeviatie in de populatie (σ) bekend is. In de praktijk is namelijk de standaarddeviatie in de populatie (σ) onbekend, evenals het populatiegemiddelde (μ), en vervangen we de standaarddeviatie in de populatie (σ) door een schatting. Deze schatting is gelijk aan de standaarddeviatie in de steekproef, aangeduid met S. De SE wordt dus geschat door de standaarddeviatie in de steekproef te delen door de wortel uit het aantal personen in de steekproef (S / \sqrt{N} in plaats van σ / \sqrt{N}). Hierbij is de ratio $(M - \mu_0) / SE$ niet meer standaardnormaal verdeeld en is er een iets andere kansverdeling, de T-verdeling, nodig voor statistische toetsing en berekening van betrouwbaarheids- en voorspellingsintervallen. De kritieke waarden van de T-verdeling zijn iets ruimer dan bij de Z-verdeling, vooral indien de steekproefomvang (N) klein is. In-

dien de steekproefomvang (N) groter is dan 30, komt de T-verdeling zo goed overeen met de Z-verdeling, dat we de Z-verdeling mogen gebruiken.

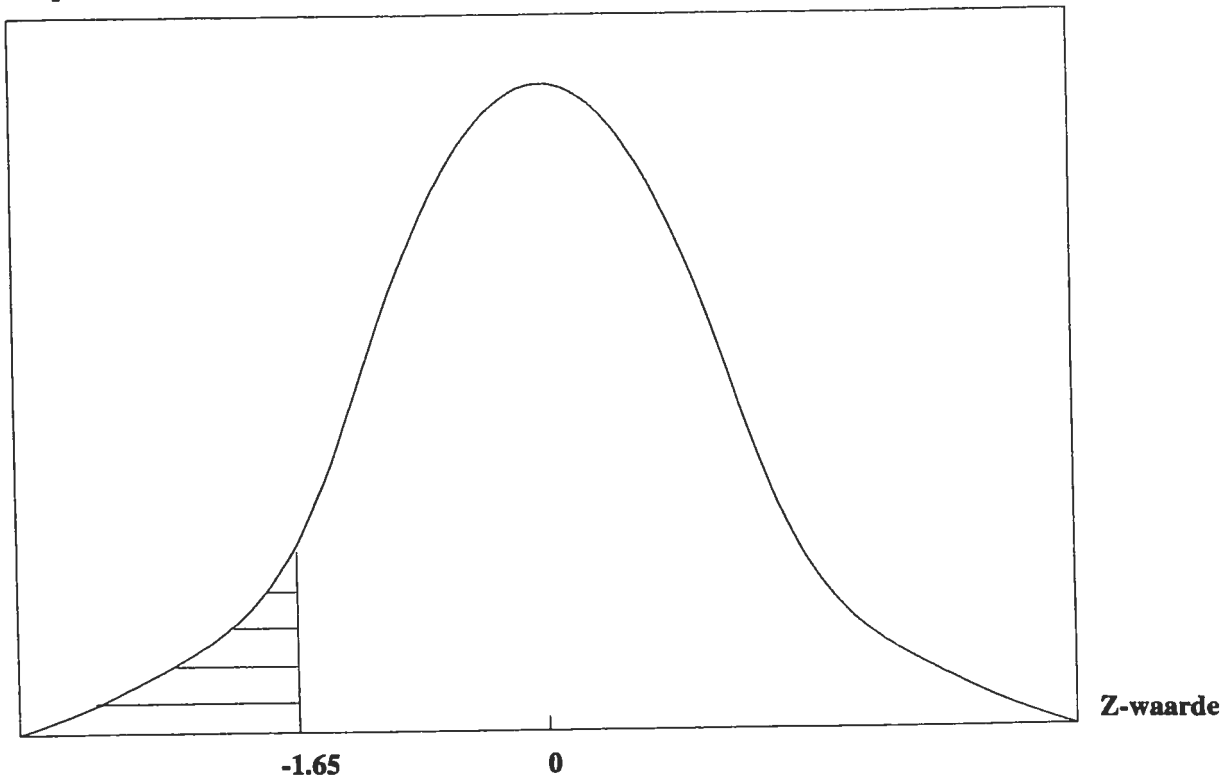
Een laatste afwijking heeft betrekking op eenzijdig toetsen. Vaak hebben we wel een vermoeden dat H_0 fout is en ook in welke richting. We vermoeden bijvoorbeeld dat de gemiddelde bloeddruk van bejaarden lager is dan 90. Bij toetsing wordt er dan van uitgegaan dat het populatiegemiddelde (μ) onder H_0 groter of gelijk is aan 90. De alternatieve hypothese (H_a) luidt dan dat het populatiegemiddelde (μ) kleiner is dan 90. Het kritieke gebied van 5% ligt nu aan de linkerkant van de normale verdeling, immers een steekproefgemiddelde dat groter is dan 90, wat overeen komt met een Z-grootte die groter is dan 0, kan nooit tot verwerping van H_0 leiden. De linker kritieke Z-waarde is dan -1.65. Dit is weergegeven in figuur 2. Het is uiteraard ook mogelijk rechts-eenzijdig te toetsen, bijvoorbeeld indien H_0 luidt dat μ kleiner of gelijk is aan 90 en H_a luidt dat μ groter is dan 90. De hierbij behorende kritieke Z-waarde is +1.65.

Menigeen weet bij eenzijdige toetsen niet wat H_0 is en wat H_a . De regel die hiervoor gehanteerd moet worden is dat het '='-teken in H_0 genoemd wordt, ongeacht of men zelf gelooft in H_0 of H_a . Uitgaande van een μ -waarde die groter of gelijk is aan 90 versus een μ -waarde die kleiner is dan 90, geeft het eerste alternatief de veronderstelde H_0 weer en ligt het kritieke gebied links. Uitgaande van een μ -waarde die groter is dan 90 versus een μ -waarde die kleiner of gelijk is aan 90, geeft het tweede alternatief de veronderstelde H_0 weer en ligt het kritieke gebied rechts.

Type I en II fout, α , β

Een statistische toets dient om te beslissen of H_0 danwel H_a juist is. Bij deze beslissing kunnen we twee typen fouten maken. Van een *type I fout* spreken we indien H_0 wordt verworpen ten gunste van H_a , terwijl H_0 juist is. Van een

frequentie



linker kritiek
gebied

H_0 verwerpen

significant

95% voorspellingsinterval
voor het steekproefgemiddelde

H_0 niet verwerpen

niet-significant

Figuur 2. Het 95% interval en de linker kritieke Z-waarde bij links-eenzijdig toetsen

type II fout spreken we indien H_0 niet wordt verworpen, terwijl H_0 onjuist is. Dit is weergegeven in figuur 3. Hoe groot is de kans op deze fouten? De kans op een type I fout, aangeduid met α , is gelijk aan de grootte van het kritieke gebied (meestal 5%), H_0 wordt immers slechts verworpen indien de gevonden Z-waarde in het kritieke gebied ligt. De kans op een type II fout, met β aangeduid, is echter lastig te bepalen en is afhankelijk van een aantal factoren. Menigeen negeert de kans op een type II fout dan ook maar. Dat is jammer, want zolang we β niet kennen, zegt niet-verwerping van H_0 weinig. We weten immers niet of H_0 juist is, of dat er sprake is van een type II fout.

Hierop komen we nog terug. Vermeld zij nu reeds dat de kans op een type II fout stijgt naarmate de steekproefomvang kleiner is. Een steekproefomvang kleiner dan 100 is vaak te weinig.

Vanwege het risico van type II fouten verdienen betrouwbaarheidsintervallen vaak de voorkeur boven toetsen. Een niet-significant resultaat verleidt namelijk tot de conclusie dat H_0 klopt. Het betrouwbaarheidsinterval geeft aan dat de steekproefuitslag een hele 'range' van populatiewaarden toestaat, niet alleen de door H_0 veronderstelde waarde. Als we in het voorbeeld over bloeddruk vinden dat het steekproefgemiddelde gelijk is aan 89, levert dit een

Uitslag toetsing	Werkelijke populatietoestand	
	H_0 is waar	H_0 is niet waar
H_0 wordt verworpen	Foute beslissing	Juiste beslissing
	H_0 wordt ten onrechte verworpen	H_0 wordt terecht verworpen
	Type I fout (α)	(1- β)
H_0 wordt gehandhaafd	Juiste beslissing	Foute beslissing
	H_0 wordt terecht gehandhaafd	H_0 wordt ten onrechte gehandhaafd
	(1- α)	Type II fout (β)

Figuur 3. Een overzicht van de typen fouten die bij statistische toetsing kunnen optreden

95% betrouwbaarheidsinterval op van 87 tot 91 mm Hg. Dit betekent dat het populatiegemiddelde (μ) 90 kan zijn, maar ook 88. Indien de steekproefomvang niet 100 was, maar slechts 25, dan zou de SE (σ / \sqrt{N}) gelijk zijn geweest aan 2 ($10 / \sqrt{25}$). Bij een steekproefgemiddelde van 89, zou dat een betrouwbaarheidsinterval van 85 tot 93 ($89 - 2 \text{ SE}$ en $89 + 2 \text{ SE}$) opleveren. Een betrouwbaarheidsinterval waarschuwt dus veel beter dan een toets voor de gevolgen van een kleine steekproefomvang, omdat een kleine steekproefomvang leidt tot een breed interval en dus tot onnauwkeurige resultaten.

Tenslotte, toets en intervallen veronderstellen dat de variabele in de populatie, en dus ook in de steekproef ongeveer normaal verdeeld is of dat de steekproefomvang groter of gelijk is aan 30. Soms vindt men in de steekproef enige 'outliers' (ofwel uitbijters). Dit zijn extreme waarden die verder dan 3 standaarddeviaties van het steekproefgemiddelde, of geïsoleerd ten opzichte van de andere waarnemingen liggen, zoals een bloeddruk van 130 mm Hg als alle andere waarden tussen 70 en 110 liggen. Vooral bij een kleine steekproefomvang kunnen 'outliers' zowel het steekproefgemiddelde als de SE, en daarmee de toetsing en het be-

trouwbaarheidsinterval sterk vertekenen. Daarom moet men altijd controleren of er 'outliers' zijn. Indien er sprake is van 'outliers', en deze zijn niet veroorzaakt door meet- of typefouten, zijn er twee oplossingen mogelijk. Ten eerste, de resultaten van de toetsing kunnen tweemaal gerapporteerd worden, eenmaal op alle data, en eenmaal na eliminatie van de 'outliers'. Ten tweede kan er een 'non-parametrische' toets toegepast worden. Dergelijke toetsen zijn veel minder gevoelig voor 'outliers'.

Het verschil tussen twee gemiddelden

Vaak is men geïnteresseerd in het verschil tussen twee gemiddelden. Bijvoorbeeld: Is de bloeddruk van rokers hoger dan die van niet-rokers; leidt onderwijsmethode A tot hogere resultaten dan B? Om deze vragen te beantwoorden trekken we uit elk van beide populaties een steekproef en vergelijken we de gemiddelden met elkaar. De statistiek hiervoor is een eenvoudige uitbreiding van de besproken theorie. Hierbij dient wel een onderscheid gemaakt te worden tussen gepaarde en ongepaarde steekproeven.

Bij *gepaarde steekproeven* is er een een-op-een relatie tussen de waarnemingen in de ene

steekproef en de waarnemingen in de andere. Beide steekproeven zijn dan ook even groot. Bijvoorbeeld: het verschil tussen voormeting (1e steekproef) en nameting (2e steekproef) van de bloeddruk van een aantal behandelde patiënten; het verschil in lengte tussen de man (1e steekproef) en de vrouw (2e steekproef) van Nederlandse echtparen. Eigenlijk is in deze beide voorbeelden sprake van slechts één steekproef, die van patiënten respectievelijk echtparen, en meet men twee keer per eenheid (patiënt respectievelijk echtpaar).

Bij *ongepaarde steekproeven* is er niet sprake van een een-op-een relatie tussen de waarnemingen in beide steekproeven. We trekken een steekproef uit de ene populatie en los daarvan een steekproef uit de andere populatie. De omvang van beide steekproeven hoeft niet even groot te zijn. De vragen over rokers versus niet-rokers en onderwijsmethode A versus B leiden meestal tot ongepaarde steekproeven.

Over gepaarde steekproeven kunnen we kort zijn. Men berekent per paar de verschilscore, bijvoorbeeld de nameting minus de voormeting. Op deze wijze worden de gepaarde steekproeven gereduceerd tot één steekproef van verschilscores. De nulhypothese veronderstelt dan dat de gemiddelde verschilscore gelijk is aan nul ($H_0: \mu = 0$).

De statistiek voor het verschil tussen twee ongepaarde gemiddelden is in wezen hetzelfde als die voor één gemiddelde. Er is een aantal verschillen. Ten eerste, bij tweezijdig toetsen luidt H_0 niet dat $\mu = 100$ (of een andere waarde), maar dat het gemiddelde van de populatie waaruit de eerste steekproef stamt (μ_1) gelijk is aan het gemiddelde van de populatie waaruit de tweede steekproef stamt (μ_2), oftewel dat $\mu_1 - \mu_2 = 0$. Evenzo luidt H_a niet dat μ niet gelijk is aan 100, maar dat $\mu_1 \neq \mu_2$. Het tweede verschil is dat we niet kijken naar de afwijking tussen het steekproefgemiddelde (M) en het veronderstelde populatiegemiddelde (μ_0), maar naar de afwijking tussen de beide steekproefgemiddelden (M_1 en M_2). We vergelijken het gevonden verschil ($M_1 - M_2$) met het ver-

Tabel 1. De diastolische bloeddruk in mm Hg in een aselechte steekproef van bejaarde mannen en een aselechte steekproef van bejaarde vrouwen in Nederland

Statistische grootheid	Mannen	Vrouwen
Steekproefomvang (N)	36	49
Gemiddelde (M)	102	97
Standaarddeviatie (S)	18	14
Standard Error (SE)	3	2

onderstelde populatieverschil ($\mu_1 - \mu_2$), dat onder H_0 verondersteld wordt gelijk te zijn aan 0. Het derde verschil is dat de SE van ($M_1 - M_2$) gelijk is aan de wortel uit SE van M_1 in het kwadraat plus de SE van M_2 in het kwadraat, omdat beide steekproefgemiddelden aan toeval onderhevig zijn. In formule: $\sqrt{[(SE \text{ van } M_1)^2 + (SE \text{ van } M_2)^2]}$.

Verder verlopen toetsing en berekening van het betrouwbaarheidsinterval net als bij één gemiddelde. Hierbij worden M en μ vervangen door respectievelijk ($M_1 - M_2$) en ($\mu_1 - \mu_2$), en de SE is die van het verschil ($M_1 - M_2$), niet van een enkele M . Als beide steekproeven minstens 30 waarnemingen omvatten, kan men de ratio $(M_1 - M_2)/SE$ behandelen als Z -grootheid. Tabel 1 geeft een voorbeeld met betrekking tot de bloeddruk van bejaarde mannen en vrouwen.

De nulhypothese luidt dat de gemiddelde bloeddruk van bejaarde mannen en vrouwen gelijk is ($H_0: \mu_1 = \mu_2$). We vinden dat M_1 minus M_2 gelijk is aan 5, de SE is gelijk aan $(9 + 4) = 13$. De Z -grootheid is dan gelijk aan $1.39 (5/\sqrt{13})$. Dit betekent dat H_0 niet verworpen wordt. Het 95% betrouwbaarheidsinterval loopt van $\sqrt{5} - 2\sqrt{13}$ tot $5 + 2\sqrt{13}$, oftewel van -2.22 tot $+12.22$. De door H_0 veronderstelde waarde, namelijk een gemiddelde verschilscore gelijk aan 0, valt binnen dit interval. Dit betekent dat H_0 niet verworpen wordt en er sprake is van een statistisch niet-significant resultaat. Er is geen aanwijzing voor een verschil in bloeddruk tussen bejaarde mannen en vrouwen.

Het 'outlier' probleem speelt ook bij twee-steekproeven toetsen. De remedie is hetzelfde als bij de één-steekproef toets, dat wil zeggen dat toetsing met en zonder de 'outliers' dient plaats te vinden, of non-parametrische toetsen gebruikt moeten worden.

Het verschil tussen twee proporties of percentages

In medisch onderzoek wil men vaak twee groepen vergelijken op een binaire maat, zoals wel of geen kanker of wel of geen genezing. Men wil bijvoorbeeld weten of longkanker vaker voorkomt bij rokers dan bij niet-rokers, of het succespercentage van bestraling hoger is dan dat van chemotherapie. Men vergelijkt hierbij niet twee gemiddelden, maar proporties. Een *proportie* is echter het gemiddelde van een 0/1 variabele, waarbij '0' betekent geen kanker en '1' wel kanker betekent. Het gemiddelde van deze 0/1 variabele is gelijk aan de proportie mensen met kanker. We kunnen dus de beschreven theorie toepassen. Opgemerkt dient echter te worden dat het steekproefgemiddelde (M) wordt vervangen door een steekproefproportie (P). De standaarddeviatie S is daarbij gelijk aan $\sqrt{P(1-P)}$. Bovendien mag de normale benadering alleen gebruikt worden als NP en $N(1-P)$ beide minstens 5 zijn in elke steekproef. Verder zijn er enige subtiele verschillen tussen de statistiek voor proporties en gemiddelden waarop we niet ingaan.

Steekproefomvang, type II fout en 'power'

We zagen eerder dat bij toetsing twee fouten kunnen optreden. De type I fout waarin H_0 ten onrechte verworpen wordt en de type II fout waarin H_0 ten onrechte niet verworpen wordt. De kans op een type I fout kiezen we zelf, door de grootte van het kritieke gebied te kiezen (α). De kans op een type II fout, β , is lastig te bepalen. Vaak wordt echter niet over β gesproken, maar $(1-\beta)$. Deze waarde geeft de 'power' of het onderscheidend vermogen van de toets

weer; de kans dat H_0 wél wordt verworpen, gegeven dat H_0 onjuist is. Dit is weergegeven in figuur 3.

Voor het interpreteren van de 'power' van de toets geldt een aantal richtlijnen. Ten eerste, de 'power' is kleiner naarmate α kleiner is, omdat α gelijk is aan de grootte van het kritieke gebied. De 'power' is ook kleiner bij een tweezijdige toets dan bij een eenzijdige toets. Ten tweede, de 'power' is kleiner naarmate de foute H_0 dichter bij de waarheid zit. Ten derde, de 'power' is kleiner naarmate de SE groter is, omdat een grote SE gepaard gaat met een groot niet-kritiek gebied. Aangezien de grootte van de SE afhankelijk is van de standaarddeviatie in de populatie en de grootte van de steekproefomvang, kunnen nog een vierde en vijfde richtlijn worden geformuleerd voor de grootte van de 'power'. De vierde richtlijn is dat naarmate de standaarddeviatie in de populatie groter is, de 'power' kleiner wordt. De vijfde richtlijn is dat naarmate de steekproefomvang groter is, de 'power' eveneens groter is. Een steekproefomvang kleiner dan 100 is vooral bij toetsen voor proporties meestal te klein.

Vaak wordt aan de 'power' geen aandacht gegeven. Soms wordt netjes berekend hoe groot de steekproefomvang moet zijn om een 'power' van 80% of 90% te halen. Dit is beter dan de type II fout negeren, maar impliceert nog altijd een kans op een type II fout van 20% respectievelijk 10%. Het verdient aanbeveling een β te hanteren die gelijk is aan α (5%), oftewel een 'power' van 95%. Waarom zou immers een type II fout minder erg zijn dan een type I fout? Een voorbeeld zal dit illustreren. In veel clinical trials wordt een nieuwe therapie vergeleken met een oude therapie. H_0 luidt dan dat beide therapieën even goed zijn. De vraag is dan: Wat is erger? Concluderen dat de therapieën niet even goed zijn, terwijl ze het wel zijn (type I fout), of concluderen dat ze even goed zijn als een van de twee superieur is (type II fout)? Als ze in werkelijkheid even goed zijn, maakt het niet uit welke therapie wordt geko-

zen. Maar als ze niet even goed zijn, wil iedereen toch de beste therapie krijgen?

Non-parametrische toetsen

De toetsen voor gemiddelden heten 'parametrisch', omdat ze uitgaan van een normale populatieverdeling. Hiertegenover staan 'non-parametrische' of 'verdelingsvrije' toetsen. Bekend zijn vooral de rangtoetsen. Non-parametrische toetsen worden gebruikt als alternatief voor parametrische toetsen indien de populatieverdelingen niet normaal zijn. Bij grote steekproeven is een niet-normale verdeling overigens niet zo erg. Een voordeel van non-parametrische toetsen is vooral dat ze minder gevoelig zijn voor 'outliers' en, met uitzondering van de rangtekentoets, ook geschikt zijn voor ordinale gegevens.

Vergelijking tussen meer dan twee groepen, en correlatie

Tot nu toe zijn alleen vergelijkingen aan bod geweest tussen twee groepen met betrekking tot een gemiddelde of proportie. Ook voor vergelijking tussen meer dan twee groepen bestaan er toetsen. Voor gemiddelden is dat variantie-analyse en de Kruskal-Wallis-toets (voor ongepaarde steekproeven), en de Friedman-toets (voor gepaarde steekproeven). Voor ongepaarde proporties kan de χ^2 toets voor kruistabellen gebruikt worden.

Indien we het verband tussen twee kwantitatieve variabelen, zoals bloeddruk en Quetelet-index willen weten, kunnen we mensen op basis van de ene variabele (Quetelet-index) indelen in groepen en nagaan of er verschil tussen die groepen is met betrekking tot de andere variabele (bloeddruk). Men gooit dan echter informatie weg. Het is vaak beter om het verband tussen beide variabelen uit te drukken in een maat. Pearson's correlatie en zijn

non-parametrische tegenhanger, Spearman's rangcorrelatie, zijn de bekendste maten.

Vaak moet men bij de bestudering van een verschil tussen twee groepen of een correlatie tussen twee variabelen rekening houden met andere factoren, zoals leeftijd, geslacht en leefwijze. Hiervoor bestaan uitbreidingen van de besproken technieken. Bekend zijn met name variantie- en regressieanalyse. Gebruik hiervan maakt duidelijk dat statistiek meer is dan toetsen, namelijk ook het via wiskundige modellen corrigeren van vertekeningen die inherent zijn aan vooral observationeel (niet-experimenteel) onderzoek.

Verantwoorde toepassing van deze complexe technieken vergt echter een goed begrip van methodologie en statistiek. Het aansturen van regressieanalyse op de computer kan men in een dag leren. Het kiezen van een correct model, controleren op schendingen van aannamen, en interpreteren van resultaten vereisen echter de hulp van een deskundige.

Opdrachten

1. Een kennistoets voor medische studenten bestaat uit 100 goed/fout vragen, en de te behalen score ligt tussen 0 en 100. Stel dat de score in de populatie eerstejaars studenten geneeskunde in Nederland anno 1985 normaal verdeeld is. We trekken een aselechte steekproef hieruit van $N = 50$, en vinden: $M = 71.0$ en $S = 10.0$.
 - a. Hoe groot is de 'standard error' van het steekproefgemiddelde?
 - b. Bereken het 95% betrouwbaarheidsinterval voor het populatiegemiddelde.
 - c. Wordt $H_0: \mu = 68.0$, verworpen? ($\alpha = 0.05$ tweezijdig). Kon deze conclusie al uit het resultaat van 1b afgeleid worden?
 - d. Wat voor aannamen en beperkingen zijn inherent aan de gebruikte toets en de formule voor het betrouwbaarheidsinterval?

Aanbevolen literatuur

- Johnson R. Elementary statistics. Boston: PWS-KENT, 1988.
- Matthews DE, Farewell VT. Using and understanding medical statistics. Basel: Karger, 1988.
- Pocock SJ. Clinical trials: a practical approach. Chichester: Wiley, 1983.
- Slotboom A. Statistiek in woorden. De meest voorkomende termen en technieken. Groningen: Wolters-Noordhoff, 1987.
- Wijvekate ML. Verklarende statistiek. Utrecht: het Spectrum, 1983.

DE AUTEUR

G. van Breukelen is als universitair docent verbonden aan de vakgroep Methodologie & Statistiek van de Rijksuniversiteit Limburg, Faculteit der Gezondheidswetenschappen, Maastricht.

Correspondentie-adres:

G. van Breukelen, Vakgroep Methodologie en Statistiek, Rijksuniversiteit Limburg, Postbus 616, 6200 MD Maastricht