Statistiek voor tandartsen (V)

Betrouwbaarheidsintervallen

A.H.B. Schuurs, tandarts[1]
H.J. Duivenoorden, methodoloog[2]
M.A. van 't Hof, statisticus[3]

Uit [1]de vakgroep Cariologie en Endodontologie van het Academisch Centrum Tandheelkunde Amsterdam (ACTA), [2]de vakgroep Medische Psychologie en Psychotherapie, Faculteit der Geneeskunde, Erasmus Universiteit te Rotterdam en [3]de Medisch Statistische Afdeling, Faculteit der Geneeskunde en Tandheelkunde van de Katholieke Universiteit te Nijmegen.

Trefwoord: Statistiek

Datum van acceptatie: 9 november 1990.

Adres: Dr. A.H.B. Schuurs, ACTA, Louwesweg 1, 1066 EA Amsterdam.

Samenvatting. Het gemiddelde van een steekproef wijkt vrijwel zeker af van het gemiddelde van de populatie waaruit de steekproef is getrokken, maar in het gemiddelde van de populatie is men nu juist geïnteresseerd. Het is mogelijk een soort minimum- en maximumwaarde, een 'betrouwbaarheidsinterval', te berekenen, waarbinnen het populatiegemiddelde μ waarschijnlijk zal liggen - het is dus niet uitgesloten dat μ een waarde buiten het interval heeft. Het interval geldt als een schatting van het gemiddelde van de populatie. Met enkele rekenvoorbeelden is verduidelijkt hoe een en ander wordt uitgevoerd. De daarbij gebruikte formules lijken ingewikkelder dan zij zijn. De hier beschreven procedure betreft geen toets, maar behoort wel tot de verklarende (schattende) statistiek.

SCHUURS AHB, DUIVENVOORDEN HJ, VAN 'T HOF MA. Statistiek voor tandartsen (V). Betrouwbaarheidsintervallen. Ned Tijdschr Tandheelkd 1991; 98: 474-6.

1 Inleiding

Een tandarts wil weten hoelang de door hem gemaakte amalgaamrestauraties gemiddeld meegaan. Daartoe noteert hij in de maand september elke falende amalgaamvulling en gaat na hoelang die gefunctioneerd heeft. Het (fictieve) onderzoek liet zien dat er 35 amalgaamrestauraties moesten worden vervangen, na gemiddeld $\bar{X}$ = 11,77 jaar gefunctioneerd te hebben, met een standaarddeviatie (SD) van 4,56 jaar. In een vorige publikatie is opgemerkt dat $\bar{X}$ en SD de onderzoekgegevens goed beschrijven, op voorwaarde dat de gegevens normaal verdeeld zijn.[1]

We weten dat het gemiddelde van een steekproef naar alle waarschijnlijkheid *niet* exact overeenkomt met het gemiddelde van de populatie, dus het gemiddelde van alle patiënten in deze praktijk. Als een nieuwe steekproef zou worden getrokken, bijvoorbeeld in de maand november, zou, dankzij steekproeffluctuaties (= steekproeffouten), een ander gemiddelde worden gevonden. Het populatiegemiddelde μ wijkt daarom vrijwel zeker af van het steekproefgemiddelde $\bar{X}$. Met andere woorden, het steekproefgemiddelde behoeft helemaal niet zo'n goede schatting van het populatiegemiddelde te zijn. Maar het is nu juist het gemiddelde van de populatie dat men graag wil kennen.

2 Intervallen

Uit het gevonden gemiddelde (11,77) en SD (4,56) kan worden berekend dat het merendeel (circa 95%) van de amalgaamrestauraies een levensduur heeft die ligt tussen 2,6 en 20,9 jaar ($\bar{X} \pm 2$*SD). Dit interval wordt ook het concentratie-interval genoemd, omdat de data zich daar concentreren. Het is duidelijk dat het populatiegemiddelde in dit interval zal liggen. Een betrouwbaarheidsinterval voor het populatiegemiddelde kan gerust wat smaller zijn dan het concentratie-interval, maar om te beweren dat het populatiegemiddelde gelijk is aan 11,77 is te optimistisch.

2.1 Betrouwbaarheidsintervallen

Een goede methode om met het effect van steekproeffluctuaties rekening te houden, is de bepaling van het zogenoemde betrouwbaarheidsinterval, dat wil zeggen dat we uitgaande van het gemiddelde $\bar{X}$ van de steekproef een boven- en onderwaarde bepalen waarbinnen het populatiegemiddelde μ waarschijnlijk zal liggen.

Het meest bekend en gebruikt is het 95% betrouwbaarheidsinterval (B.I., Confidence Interval, C.I.). Als dit interval is berekend, geeft het de volgende informatie: als een groot aantal malen, zeg 100 maal, aselect een steekproef uit de populatie zou worden getrokken, zullen 95 van de 100 daarbij behorende betrouwbaarheidsintervallen ook inderdaad het populatiegemiddelde bevatten. We mogen dan ook aannemen dat het ware gemiddelde (te vinden door *alle* restauraties van alle elementen (personen) van de populatie te onderzoeken) met een grote mate van waarschijnlijkheid, namelijk van 0,95, binnen het geschatte 95% B.I. ligt. Het ware gemiddelde kan dus eventueel buiten het berekende betrouwbaarheidsinterval vallen, hoewel die kans klein is, namelijk 1 - 0,95 = 0,05.

Hebben we van doen met een 99% betrouwbaarheidsinterval, dan zullen bij herhaalde steekproeftrekkingen 99% van de berekende betrouwbaarheidsintervallen μ bevatten. Het ware gemiddelde ligt dan nagenoeg zeker (waarschijnlijkheid 0,99) in dat interval. Het is duidelijk dat het 99% B.I. breder is dan het 95% B.I.

Let wel, het gemiddelde van een populatie varieert niet; het gaat slechts om het feit dat de gemiddelden van verschillende steekproeven uit een en dezelfde populatie zullen verschillen.

Concreet, van de in paragraaf 1 beschreven steekproef bleek het steekproefgemiddelde $\bar{X}$ = 11,77 te zijn. Als het daaraan gerelateerde 95% betrouwbaarheidsinterval loopt (zoals straks zal blijken) van 10,2 tot 13,3 dan is de kans 95% dat het traject 10,2 tot 13,3 het werkelijke gemiddelde bevat; μ kan buiten het interval liggen, maar die kans is slechts 5%.

3 Het berekenen van betrouwbaarheidsintervallen

Voor het berekenen van betrouwbaarheidsintervallen voor een gemiddelde van een steekproef bestaan twee mogelijkheden.

- De standaardafwijking van de populatie (σ) is onbekend, hetgeen gewoonlijk het geval is. In het nu volgende wordt de berekening voor deze situatie uitgewerkt, waarbij het van belang is een onderscheid te maken naar grote (3.1) en kleine steekproeven (3.2).
- De standaardafwijking σ van de populatie is bekend, maar dit is slechts zelden het geval. Mocht σ bekend zijn, dan verloopt de berekening hetzelfde als bij grote steekproeven met onbekende σ.

3.1 Grote steekproef

Als we een steekproef trekken is de standaardafwijking σ van de populatie gewoonlijk niet bekend. Maar het is wel mogelijk

om de standaarddeviatie SD van de steekproef te berekenen. Bij grote steekproeven is de SD een goede benadering van σ en kan SD dienen om het 95% betrouwbaarheidsinterval voor het gemiddelde te berekenen.

Wanneer we een groot aantal, zeg 500, steekproeven uit de populatie zouden trekken en we zouden voor elke steekproef het gemiddelde berekenen, dan tonen die gemiddelden een normale verdeling. Uit deze normale verdeling kunnen we zicht krijgen op de mate waarin de gemiddelden fluctueren, en wel door de standaarddeviatie SD van de steekproefgemiddelden te berekenen. Voor een normale verdeling geldt dat 95% van alle waarden ligt tussen het ware gemiddelde (μ) plus of min tweemaal de spreiding, dus tussen μ - 2*σ en μ + 2*σ.

Het zou wel zeer omslachtig zijn om 500 steekproeven te moeten trekken ter bepaling van een betrouwbaarheidsinterval. Dat is ook niet nodig; er bestaan formules waarmee op grond van de data van één aselect getrokken steekproef de berekening kan plaatsvinden.

3.1.1 95% Betrouwbaarheidsinterval van een grote steekproef

Voor de bepaling van het 95% B.I. wordt de *standaardfout* SE (Standard Error) gebruikt. De standaardfout is een kwantificering van de grootte van de fout in het gemiddelde. Hij wordt gevonden door de standdaarddeviatie SD te delen door de wortel uit het aantal onderzochten. Omdat het hier om de standaardfout in het gemiddelde gaat, wordt deze vaak SEM (Standard Error of the Mean) genoemd. Er geldt dus: SEM = SD/√N. De minimumwaarde van het 95% betrouwbaarheidsinterval is:

$$\bar{X} - 2 * \frac{SD}{\sqrt{N}} \quad \text{(formule 1)}$$

De maximumwaarde van het interval is gelijk aan:

$$\bar{X} + 2 * \frac{SD}{\sqrt{N}} \quad \text{(formule 2)}$$

Stel dat in een steekproef van N = 900, het gemiddelde $\bar{X}$ = 21,1 en de standaarddeviatie SD = 15,0. Dan wordt als uitkomst van formule 1 gevonden:

$$21{,}1 - 2 * \frac{15}{\sqrt{900}} = 20{,}1$$

De uitkomst van formule 2 is:

$$21{,}1 + 2 * \frac{15}{\sqrt{900}} = 22{,}1$$

Dus met een waarschijnlijkheid van 95% geldt: $20{,}1 < \mu < 22{,}1$.

3.1.2 99% Betrouwbaarheidsinterval van een grote steekproef

Voor een normale verdeling geldt dat circa 99% van alle waarden ligt tussen μ - 2,6 * σ en μ + 2,6 * σ. Het 99% B.I. voor het populatiegemiddelde wordt berekend met de formules 1 en 2, dus:

$$\bar{X} - 2{,}6 * \frac{SD}{\sqrt{N}} \text{ èn}$$

$$\bar{X} + 2{,}6 * \frac{SD}{\sqrt{N}} \text{ en geldt } 19{,}8 < \mu < 22{,}4.$$

Het 99% betrouwbaarheidsinterval is, zoals hier is aangetoond (en per definitie geldt), breder dan het 95% B.I.

3.2 Kleine steekproeven

De gebruikte formule $\bar{X} \pm 2 * SD/\sqrt{N}$ is eigenlijk alleen geldig als de SD gelijk is aan werkelijke spreiding σ, en dat is bij benadering juist voor grote steekproeven. Maar als de steekproefomvang klein is, zal de gevonden SD kunnen afwijken van σ. Dit brengt een extra onzekerheid met zich mee, die ondervangen kan worden door wat meer slagen om de arm te houden. Dat gebeurt door de vermenigvuldigingsfactor, die in de formules (1 en 2) gelijk is aan 2, een grotere waarde te geven. Daarbij geldt: hoe kleiner de steekproef, des te groter de vermenigvuldigingsfactor, die wordt aangeduid met het symbool t_N.

Tabel I. Enkele waarden van de vermenigvuldigingsfactor t_n voor verschillende steekproefomvangen (n), nodig bij de berekening van een 95% en 99% betrouwbaarheidsinterval.

	95% interval	99% interval
n= 2	12,71	63,66
3	4,30	9,93
4	3,18	5,84
5	2,78	4,60
6	2,57	4,03
7	2,45	3,71
8	2,37	3,50
9	2,31	3,36
10	2,26	3,25
15	2,15	2,98
20	2,09	2,86
25	2,06	2,80
30	2,05	2,76
35	2,03	2,73
40	2,02	2,71
45	2,02	2,69
50	2,01	2,68
60	2,00	2,66
75	1,99	2,64
100	1,98	2,63
150	1,98	2,61
200	1,97	2,60
300	1,97	2,59
∞	1,96	2,58

* Gemodificeerd naar Gardner & Altman (1989)[2]

3.2.1 95% betrouwbaarheidsinterval voor kleine steekproef

Het 95% B.I. voor het populatiegemiddelde μ, uitgaande van het steekproefgemiddelde wordt aan de onderkant begrensd door:

$$\bar{X} - t_N * \frac{SD}{\sqrt{N}} \quad \text{(formule 3)}$$

en aan de bovenzijde door:

$$\bar{X} + t_N * \frac{SD}{\sqrt{N}} \quad \text{(formule 4)}$$

In deze formules staat $\bar{X}$ weer voor het steekproefgemiddelde, N voor de steekproefomvang, SD voor de standdaarddeviatie en de term t_N voor de vermenigvuldigingsfactor. Om de waarde van t_N te bepalen zijn tabellen gemaakt, die berusten op de zogenaamde t-verdelingen van Student (Student was de schuilnaam van de betreffende statisticus). De waarde van t_N hangt af van de steekproefomvang N en vanzelfsprekend ook van de gewenste betrouwbaarheid van het interval. In tabel I zijn een aantal waarden van t_N gepresenteerd, voor het 95% en 99% B.I. Uit tabel I blijkt dat bij N > 50 gerust met een vermenigvuldigingsfactor = 2 gerekend kan worden.

In het onderzoek naar de duurzaamheid van amalgaamrestauraties was N = 35, $\bar{X}$ = 11,77 en SD = 4,56. Raadpleging van tabel I voor N = 35, laat zien dat t_N = 2,03. Invulling van deze waarden in formule 3, voor berekening van het 95% betrouwbaarheidsinterval, geeft:

$$\bar{X} - t_N * \frac{SD}{\sqrt{N}} = 11{,}77 - 2{,}03 * \frac{4{,}56}{\sqrt{35}} = 10{,}2$$

en invulling van formule 4 resulteert in een waarde van 13,3. De schatting voor de gemiddelde levensduur van de amalgaamrestauraties in de onderzochte praktijk is: $10{,}2 < \mu < 13{,}3$.

3.2.2 99% Betrouwbaarheidsinterval voor kleine steekproef

Voor het 99% B.I. is $\alpha = 0{,}01$. We zoeken t_N op in de tweede kolom van tabel I voor N = 35 en vinden dat t_N = 2,73. Toepassing van formules 3 en 4 levert 9,7 als ondergrens en 13,9 als bovengrens van het 99% betrouwbaarheidsinterval op.

(Bij de berekening van het 99% betrouwbaarheidsinterval voor een grote steekproef werd als vermenigvuldigingsfactor 2,6 gebruikt, zijnde de afgeronde waarde van t_N in tabel I voor N = ∞.)

4 Methodologische kanttekening

Het onderzoek zoals beschreven in paragraaf 1 is eenvoudig en snel uit te voeren en wordt ook in de literatuur beschreven. Helaas geldt hiervoor vanuit methodologisch standpunt de kwalificatie 'Quick & Dirty'. Desalniettemin overdonderen 'experts' die dit soort 'faalonderzoek' willen stimuleren, de argeloze lezer, die echter toch aanvoelt dat er iets niet klopt. Binnen het kader van deze serie artikelen werd bewust gekozen voor dit 'invalide' voorbeeld: statistisch gezien heeft de werking van dat voorbeeld daar niet onder te lijden, terwijl het toont dat naast de statistiek ook de methode van onderzoek van groot belang is. Wij willen toelichten waarom het voorbeeld niet-valide is. Daartoe vergelijken we een pas drie jaar werkzame tandarts A met tandarts B, die al 25 jaar praktijk doet. De levensduur van de vullingen van tandarts A kan niet groter zijn dan drie jaar en het gemiddelde (op grond van een relatief gering aantal restauraties met weinig mislukkingen) zou bijvoorbeeld twee jaar kunnen zijn. Tandarts B zal een veel langere levensduur voor zijn restauraties vinden, tot 25 jaar, bijvoorbeeld gemiddeld acht jaar. Op grond van een dergelijk 'faalonderzoek' komt men ongetwijfeld tot de conclusie dat de oudere tandarts beter werkt, ongeacht werkelijke kwaliteitsverschillen.

Maar ook bij onderzoek binnen de praktijk van één tandarts stuit men op soortgelijke methodologische problemen. Stel dat een tandarts 25 jaar lang restauraties van amalgaam maakt en pas 10 jaar van composiet. De levensduur van de amalgaamrestauraties is maximaal 25 jaar en de gemiddelde levensduur bijvoorbeeld elf jaar. De maximale levensduur van de composietrestauraties is in dit geval tien jaar en de gemiddelde levensduur bijvoorbeeld zes jaar. Het zou dan fout zijn te concluderen dat amalgaam beter is dan composiet. Bij levensduuronderzoek is het namelijk cruciaal ook de goede restauraties mee te tellen.

5 Slot

Betrouwbaarheidsintervallen zijn bij allerlei schattingen te berekenen, bijvoorbeeld bij het verschil tussen de gemiddelden van twee populaties, correlaties en de mediaan. De wijzen van berekening zijn veelvuldig beschreven.[2-5] Met name het boekje van Gardner en Altman (1989) verdient aanbeveling.[2]

Mede om de relevantie van een onderzoeksresultaat aan te geven, wordt tegenwoordig door de meeste internationale medische tijdschriften gevraagd naast het significantieniveau (P-waarde) van de toets ook het 95% betrouwbaarheidsinterval voor het effect op te geven. Daardoor krijgt de lezer een beter beeld van het belang dat aan de bevindingen moet worden gehecht: immers, een effect, dat bij toetsen sterk significant is bevonden, valt door de mand als men ziet dat het betrouwbaarheidsinterval maar weinig boven de nulwaarde uitstijgt. Vermelding van alleen het steekproefgemiddelde, een zogenaamde puntschatting, geeft altijd een onduidelijk beeld, het betrouwbaarheidsinterval een meer volledig.

Summary

ESTIMATION OF THE POPULATION MEAN SCORE

Key word: Statistics

The mean score of a sample deviates most probably from the mean score of the population, in which one is most interested. It is possible to calculate a kind of minimum and maximum value, the so called confidence interval, which indicates the position of the mean score of the population. Some worked examples elucidate the procedure of constructing such confidence intervals for the population mean, using the standard error of the (sample) mean (SEM).

Literatuur

[1] SCHUURS AHB, DUIVENVOORDEN HJ, VAN 'T HOF MA. Appels plus peren. Meetniveau. Ned Tijdschr Tandheelkd 1990; 97: 505-8.
[2] GARDNER MJ, ALTMAN DG. Statistics with confidence. London: British Medical Journal, 1989.
[3] POCOCK SJ. Clinical trials. Chichester, John Wiley & Sons 1983: 206-10.
[4] INGELFINGER JA, MOSSTELLER F, THIBODEAU LA, WARE JH. Biostatistics in clinical medicine. New York, Macmillan Publishing Co. Inc. 1987: ch. 5.
[5] WEINBERG R, CHEUK SL. Introduction to dental statistics. New Jersey, Noyes Medical Publications 1980: ch. 4.