

Prosentiler til besvær

Sammendrag

Bakgrunn. I mange medisinske sammenhenger har referanseverdier en sentral rolle når beslutninger om diagnose og behandling tas. Ofte baseres referanseverdiene på prosentiler. Vi har undersøkt hvordan bruk av ulike statistikkprogrammer kan påvirke estimatet for prosentiler.

Materiale og metode. Sædparametere fra fertile menn har vært benyttet til å etablere referanseverdier for sædvariabler. Vi har reanalysert våre tidligere publiserte data ved hjelp av ulike statistiske metoder for estimering av prosentiler. Disse metodene finnes i programmet R i funksjonen *quantile* og omfatter blant annet de som finnes i vanlig brukte statistikkpakker, som SAS, SPSS og Minitab.

Resultater og fortolkning. Analyse av to sædvariabler med ulike ikke-parametriske metoder avdekket betydelige forskjeller i estimerte 5- og 10-prosentiler for spermiekonsentrasjon. Forskjellene var langt mindre for spermimotoilitet. Våre resultater viser at når referanseverdier oppgis, bør det samtidig gis informasjon om hvilke statistiske metoder som benyttes til estimering av prosentiler. I klinisk praksis kan valg av metode avgjøre om et prøvesvar nær referanseverdien faller innenfor eller utenfor referanseområdet. Sammenlikning av referanseintervaller etablert ved ulike laboratorier kan dessuten gi et feilaktig bilde hvis forskjellige statistiske metoder er benyttet.

Engelsk sammendrag finnes i artikkelen på www.tidsskriftet.no

Oppgitte interessekonflikter: Ingen

Referanseverdier for biologiske variabler spiller en sentral rolle i mange medisinske sammenhenger og danner ofte grunnlaget for diagnostikk og behandling. Det typiske innen medisinsk biokjemi er å sammenlikne en persons analyseresultater med verdier basert på friske individer, tilpasset alder og kjønn. Et omfattende arbeid ligger bak etablering av referanseintervaller og inkluderer standardisering av analysemetoder, utvelgelse av referansepersoner, innsamling og analyse av materiale samt statistisk bearbeiding av data. Ulike statistiske metoder kan benyttes i estimering av referanseverdier (1), men prosentiler er ofte brukt ved fastsettelse av et referanseintervall. Forskjellige statistikkprogrammer kan benyttes til beregning av en prosentil, og man burde kunne forvente tilnærmet like verdier. Imidlertid er dette ikke alltid tilfellet.

I denne artikkelen gis det først en gjennomgang av det teoretiske grunnlaget for beregning av prosentiler og referanseverdier. Videre illustreres forskjellene ved beregning av prosentiler ved et eksempel hentet fra etablering av referanseverdier for sædvariabler (2).

Prosentiler

Medianen er den midterste observasjonen i et datasett og er et velkjent mål for den sentrale verdien i materialet. Den er ofte et alternativ eller supplement til gjennomsnittsverdien. Det sentrale statistiske begrepet i denne artikkelen er prosentil (percentile), og medianen svarer til 50-prosentilen. I tillegg er en rekke andre begreper som fraktil, kvantil og kvartil også spesialtilfeller av prosentiler. Prosentil kan defineres uten matematisk formalisme, slik vi har gjort i figur 1. Et velkjent eksempel er følgende: I en standard normalfordeling er $-1,64$ lik 5-prosentilen fordi 5 % av arealet ligger til venstre for $-1,64$.

En mulig fremgangsmåte for å estimere prosentiler er å anta en bestemt fordeling for et datasett, slik som normalfordeling, og deretter anvende vanlige statistiske metoder, som for eksempel sannsynlighetsmaksimering. Dette er en *parametrisk* fremgangsmåte, som ikke forfølges videre her (1). Det er imidlertid viktig å påpeke at ved denne fremgangsmåten benyttes hele datasettet og ikke bare en eller to observasjoner slik tilfellet er for beregning av medianen.

Vi diskuterer i dette arbeidet *ikke-parametrisk* estimering av prosentiler. Det betyr at observasjonene ordnes fra den minste til den største, og at disse ordnede observasjonene danner grunnlag for estimeringen. For

dataene 1, 2, 5, 8, 10 estimeres følgelig medianen som 5, den midterste ordningsobservatoren. Dersom dataene er 1, 2, 8, 10, er det vanlig å estimere medianen som $(2+8)/2 = 5$.

Hyndman & Fan (3) er en autorativ referanse når det gjelder estimering av prosentiler generelt. Hyndman har sammen med Ivan Frohne implementert estimeringen av prosentiler i programmet R i funksjonen *quantile*. Programmet R, som er fritt tilgjengelig på Internett (4), har en meget sentral betydning for statistikere og andre som utvikler nye metoder. Andre programmer, slik som Minitab, SAS og SPSS, er bedre kjent for brukere av statistiske metoder. Funksjonen *quantile* inneholder ni forskjellige metoder og omfatter blant annet de som brukes i de nevnte programmene. Disse ni metodene har det til felles at prosentilestimatene kan skrives som et *veiet* gjennomsnitt av to ordningsobservatorer. Medianen er gjennomsnittet av de to midterste observasjonene dersom det er et likt antall observasjoner. Hvis andre prosentiler skal estimeres, benyttes andre relevante ordningsobservatorer enn de midterste. Det er generelt viktig å oppgi usikkerheten i estimer, og prosentilestimer er ikke noe unntak.

Referanseverdier

Verdens helseorganisasjon oppfordrer hvert enkelt laboratorium som utfører sædanalyser, til å etablere egne referanseverdier for sædvariabler fordi metodevariasjoner og geografiske eller etniske forskjeller kan gjøre sammenlikninger mellom laboratorier vanskelige (5). På bakgrunn av dette ble det ved Andrologisk laboratorium ved Rikshospitalet etablert referanseverdier for sædvariabler basert på et utvalg av menn som nylig hadde gitt opphav til graviditet (2). Ulike prosentiler ble estimert ved hjelp av SPSS og brukt som utgangspunkt for referanseverdier. Ofte defineres referanseområdet som det sentrale 95 %-intervallet, men i mange kliniske sammenhenger er det mer hensiktsmessig å forskyve intervallet hvis det er patologiske funn



Hovedbudskap

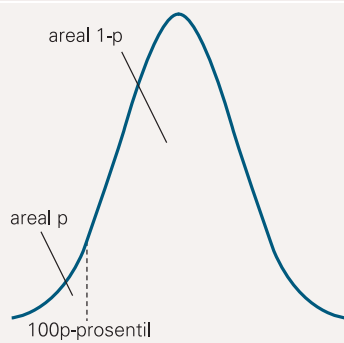
- Forskjellige statistikkprogrammer kan gi ulike estimer for prosentiler
- Ved etablering av referanseområder er det viktig å informere om statistiske metoder som er brukt

Thore Egeland

Avdeling for medisinsk genetikk
Ullevål universitetssykehus

Trine B. Haugen

trine.b.haugen@hf.hio.no
Avdeling for helsefag
Høgskolen i Oslo
Postboks 4 St. Olavs plass
0130 Oslo

Figur 1

Prosentiler er definert med utgangspunkt i figuren. Dersom $p = 0,05$, fremkommer 5-prosentilen, som er definert ved at 5 % av arealet skal være til venstre for den stipulede linjen og de resterende 95 % til høyre. I dette tilfellet blir 100p-prosentilen $-1,64$ på x-aksen

ved kun lave eller høye verdier. På denne bakgrunn anses 5-prosentilen å være mest relevant ved estimering av referanseverdien for spermiekonsentrasjonen. Det kan imidlertid være informativt å oppgi flere prosentiler basert på referansematerialet.

Metode og resultater

Vi analyserte dataene ($n = 82$) for de to sædvariablene spermiekonsentrasjon og raskt progressiv spermimotoilitet (2, 5) ved hjelp av funksjonen quantile i programmet R (4). Ni ulike ikke-parametriske metoder ble benyttet for å estimere prosentiler.

Tabellene 1a og 1b gjengir et utvalg prosentiler for spermiekonsentrasjon og spermimotoilitet basert på de ni metodene. Vi observerte forskjeller for alle prosentiler estimert for spermiekonsentrasjonen i dette eksemplet, bortsett fra medianverdien. Det var et betydelig avvik for 5-prosentilen mellom metode 2, som er en av SAS-oppsjonene, og metode 6, som brukes i SPSS. Tallene var henholdsvis 14,0 og 10,6. De fem minste observasjonene i datasettet var 0,9, 9,0, 10,0, 10,0 og 14,0. Metode 2 estimerer 14 ved den femte minste observasjonen. Opsjonene 4–9 kan oppfattes som ulike lineære interpolasjoner mellom fjerde og femte minste observasjon, altså mellom 10 og 14.

Analyse av spermimotoilitet ved bruk av de ulike metodene resulterte kun i små forskjeller (tab 1b). Her var resultatene sammenfallende for prosentiler opptil 1 %, mens for eksempel medianverdiene (50 %) ikke var helt identiske.

Diskusjon

Vi har vist at forskjellige statistiske metoder kan gi ulike estimater for prosentiler. Det skyldes i stor grad at ordningsobservatorene vektet ulikt (3). I vårt eksempel kan det innvendes at det kan være for få observasjoner til å gi stabile estimater av prosentiler, spesielt for prosentiler som er langt fra senteret i fordelingen. Imidlertid viser analyse av dataene for spermiekonsentrasjonen og spermimotoiliteten, som er basert på samme antall observasjoner, at forskjellene varierer avhengig av variabel. Fordelingen til data spiller følgelig en vesentlig rolle også ved ikke-parametriske

metoder. Det er ikke uten videre enkelt å forutsi hvor mange observasjoner som bør inkluderes for å minke forskjellene, og det er viktig å kjenne til at ulike programmer kan gi forskjellige resultater.

For spermiekonsentrasjonen i vårt eksempel var det betydelige forskjeller i estimert 5-prosentil avhengig av metode for estimering. Konsekvensen av å benytte 14 millioner spermier per ml ejakulat som referanseverdi vil være at flere pasienter anses som subfertile i forhold til om 10 millioner per ml blir brukt som verdi. Sammenlikning av referanseintervaller etablert ved ulike laboratorier kan dessuten gi et feilaktig bilde hvis forskjellige statistiske metoder er benyttet.

Ved ikke-parametriske metoder er det vanskelig å argumentere for at én estimator er best i en stringent matematisk forstand. Imidlertid kan man argumentere for at noe er mer rimelig enn noe annet ut ifra visse kriterier. En slik diskusjon vil innebære en avveining mellom estimatorens presisjon og skjevhet. Det er også mulig å sammenlikne metodene i tilfeller der fasiten er kjent, og det leder til å se på simulerte data. Fan & Hyndman (3) foretrekker metode 8 i tabell 1, men det er langt fra konsensus om dette. Denne metoden gir tall som ligger mellom det vi har kalt SPSS- og SAS-estimer, men nærmere førstnevnte. Det hadde vært en fordel, slik mange forfattere har påpekt, om man kunne ha blitt enig om en felles fremgangsmåte, en felles estimator. Det er muligens ikke realistisk, da de forskjellige vitenskapelige fagretningene ofte lever i temmelig atskilte vitenskapelige verdener. Det man bør tilstrebe, er at forskere som arbeider med liknende anvendelser, bruker like metoder når verdier skal sammenliknes. Dette gjelder ikke minst sammenfatning av data, som i metaanalyser.

Konklusjon

Ved etablering av referanseområder for biologiske variabler er det viktig å gi informasjon både om referansegruppe og de statistiske metodene og programmene som er brukt til estimering. Det er derfor viktig å kjenne til at ulike programmer kan gi forskjellige referanseverdier.

Litteratur

1. Solberg HE. Establishment and use of reference values. I: Burtis CA, Ashwood ER, Bruns DE, red. Tietz textbook of clinical chemistry and molecular diagnostics, 4. utg. Philadelphia, PA: Elsevier, 2006: 425–48.
2. Haugen TB, Egeland T, Magnus Ø. Semen parameters in Norwegian fertile men. J Androl 2006; 27: 66–71.
3. Hyndman RJ, Fan Y. Sample quantiles in statistical packages. Am Stat 1996; 50: 361–5.
4. The R project for statistical computing: www.r-project.org/ (6.11.2006).
5. World Health Organization. WHO laboratory manual for the examination of human semen and sperm-cervical mucus interaction. 4. utg. Cambridge: Cambridge University Press, 1999.

Manuskriptet ble mottatt 6. 11. 2006 og godkjent 2. 1. 2007. Medisinsk redaktør Michael Bretthauer.

Tabell 1 Ulike ikke-parametriske metoder benyttet i estimering av prosentiler, basert på 82 observasjoner. Hver av de ni linjene svarer til en estimator i funksjonen quantile i programmet R (4), mens hver kolonne svarer til en estimert prosentil for spermiekonsentrasjon (millioner/ml) (a) eller raskt progressiv spermimotoilitet (%) (b). Estimator 6 er implementert i programmene SPSS og Minitab, mens 2 og 3 finnes i SAS. Presise definisjoner av alle metoder finnes i (4)

a)		0,1 %	0,5 %	1 %	2 %	5 %	10 %	50 %
Estimator								
1		0,9	0,9	0,9	9,0	14,0	19,0	70,0
2		0,9	0,9	0,9	9,0	14,0	19,0	70,0
3		0,9	0,9	0,9	9,0	10,0	16,0	70,0
4		0,9	0,9	0,9	6,1	10,4	16,6	70,0
5		0,9	0,9	3,5	9,1	12,4	18,1	70,0
6		0,9	0,9	0,9	6,2	10,6	16,9	70,0
7		1,6	4,2	7,5	9,6	14,1	19,4	70,0
8		0,9	0,9	2,2	8,8	11,8	17,7	70,0
9		0,9	0,9	2,5	9,0	12,0	17,8	70,0
b)								
Estimator		0,1 %	0,5 %	1 %	2 %	5 %	10 %	50 %
1		10,0	10,0	10,0	10,0	16,0	18,0	35,0
2		10,0	10,0	10,0	10,0	16,0	18,0	35,5
3		10,0	10,0	10,0	10,0	15,0	18,0	35,0
4		10,0	10,0	10,0	10,0	15,1	18,0	35,0
5		10,0	10,0	10,0	10,4	15,6	18,0	35,5
6		10,0	10,0	10,0	10,0	15,2	18,0	35,5
7		10,0	10,0	10,0	11,9	16,0	18,1	35,5
8		10,0	10,0	10,0	10,0	15,5	18,0	35,5
9		10,0	10,0	10,0	10,1	15,5	18,0	35,5