Tidsskriftet Tidsskrift for Den norske legeforening

Gjennomsnitt og standardavvik eller median og kvartiler?

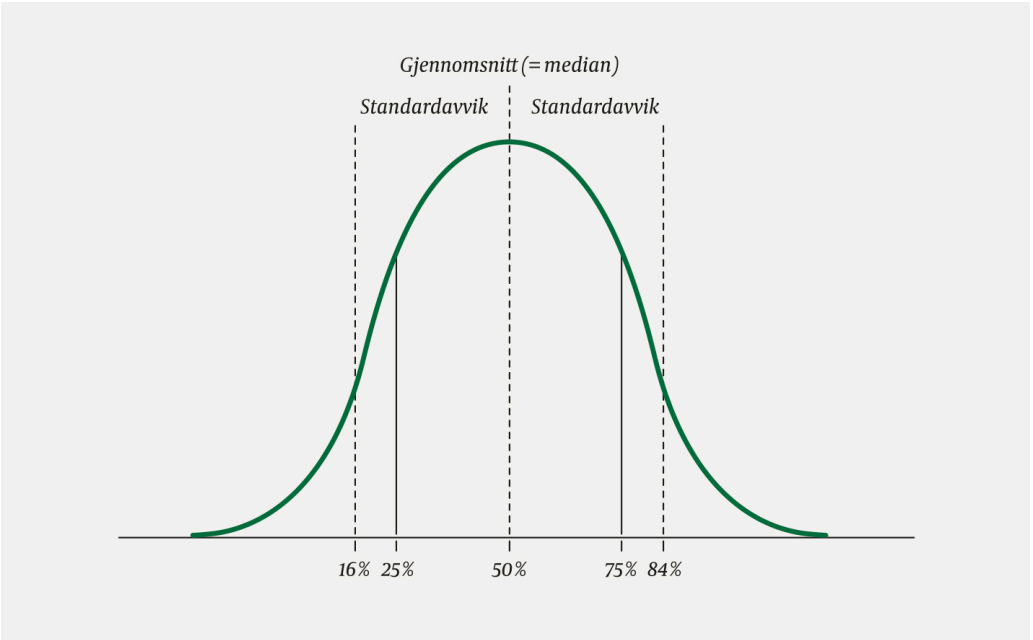
MEDISIN OG TALL

STIAN LYDERSEN

stian.lydersen@ntnu.no
Stian Lydersen er dr.ing. og professor i medisinsk statistikk ved Regionalt kunnskapssenter for barn og unge – psykisk helse og barnevern (RKBU Midt-Norge) ved Institutt for psykisk helse, NTNU.
Forfatteren har fylt ut ICMJE-skjemaet og oppgir ingen interessekonflikter.

Gjennomsnitt og standardavvik er mye brukte mål på sentraltendens og variasjon i data fra skalavariabler. Dersom dataene ikke er normalfordelt, vil enkelte foretrekke å oppgi median og kvartiler isteden. Men gjennomsnitt og standardavvik har nyttige egenskaper og kan være relevant også når dataene ikke er normalfordelt.

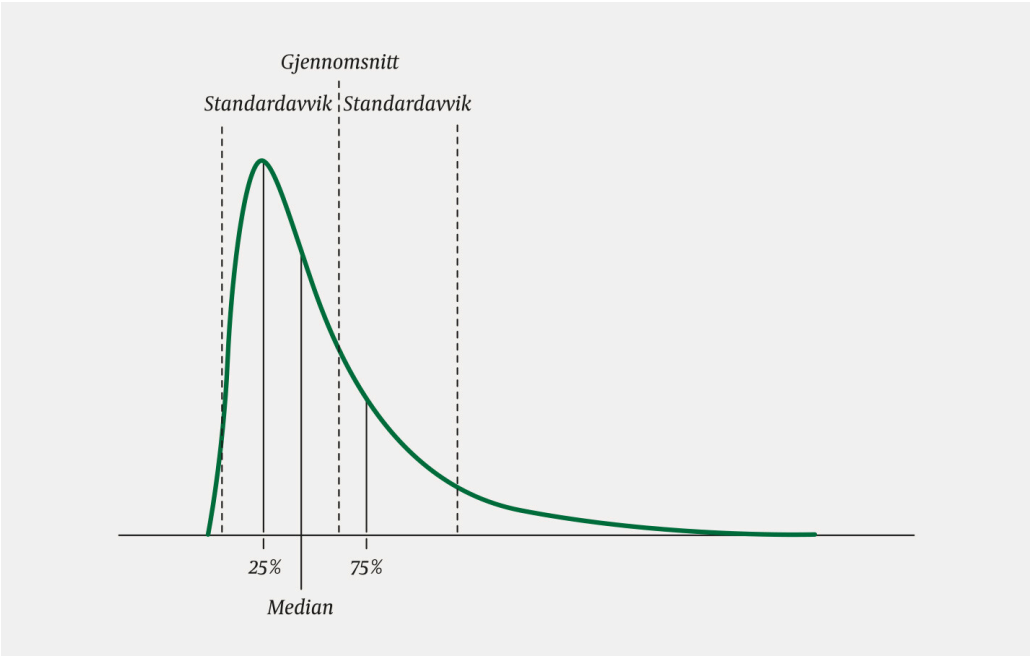
La oss starte med å se på normalfordelingen, som er vist i figur 1. Dersom dataene er normalfordelt, vil cirka 16 % av observasjonene være lavere enn ett standardavvik under gjennomsnittet. Og tilsvarende vil cirka 84 % av observasjonene være lavere enn gjennomsnittet pluss standardavviket. Dersom dataene er normalfordelt, vil standardavviket altså være direkte knyttet til 16-prosentilen og 84-prosentilen. Hva med medianen og kvartilene? Siden fordelingen er symmetrisk, vil medianen være lik gjennomsnittet. Kvartilene er per definisjon lik 25-prosentilen og 75-prosentilen, og disse angir derfor i normalfordelingen et litt smalere intervall enn ett standardavvik på hver side av gjennomsnittet.



Figur 1 Normalfordeling med gjennomsnitt (= median), standardavvik og kvartiler (25 % og 75 %).

Skjevfordelte data

Figur 2 viser en fordelingskurve som er høyreskjev. Slike fordelinger kan skyldes målinger som ikke kan være negative, som for eksempel plasmakonsentrasjon. I en høyreskjev fordeling vil gjennomsnittet være høyere enn medianen. Og standardavviket er ikke knyttet til bestemte prosentiler, slik det var i normalfordelingen.



Figur 2 Høyreskjev fordeling med gjennomsnitt, standardavvik, median og kvartiler (25 % og 75 %).

Hva er relevante mål på sentraltendens og variasjon hvis dataene ikke er normalfordelt? De matematiske uttrykkene for å beregne gjennomsnitt og standardavvik forutsetter ingenting om fordelingen, og er veldefinert også for data som ikke er normalfordelt. La oss se på et tenkt talleksempel, hentet fra (1): Anta at vi har registrert antall dager på sykehus for 13 pasienter med en gitt diagnose (hhv. 3, 9, 10, 10, 10, 12, 13, 14, 18, 21, 27, 38 og 62 dager). Her blir gjennomsnittet 19 dager, mens medianen blir 13 dager. Standardavviket blir 15,8 dager, og nedre og øvre kvartil blir hhv. 10 og 24 dager. Hvis vi ønsker å estimere kostnad eller behov for personell, er gjennomsnittet en mer relevant størrelse enn medianen. Hvis man ønsker å si noe om «typisk» liggetid for en enkelt pasient, vil medianen kunne være mere relevant.

Man ser at enkelte forfattere bare oppgir interkvartilbredden, som her vil være 24–10 = 14 dager, istedenfor å oppgi kvartilene. Dette er mindre informativt enn å oppgi kvartilene, som sammen med medianen også gir innsikt i hvor skjev fordelingen er. I vårt eksempel ser vi at medianen på 13 dager er nærmere nedre kvartil på 10 dager enn øvre kvartil på 24 dager, og dette indikerer en høyreskjev fordeling, liknende den som er vist i figur 2. I noen sammenhenger kan det være fornuftig å oppgi minimums- og maksimumsverdien istedenfor, eller i tillegg til, kvartilene. Men man bør være bevisst på det faktum at i motsetning til kvartilene så vil avstanden mellom minimum og maksimum forventes å øke med utvalgsstørrelsen.

Hva bør rapporteres?

Hvilke mål bør man oppgi hvis man ikke har normalfordelte data? Et kriterium kan være å se på hva som er relevant i den aktuelle anvendelsen, som i eksempelet med liggetid. Men hva med beskrivende statistikk for bakgrunnsdata i en studie? Enkelte forskere hevder at det generelt er feil å oppgi gjennomsnitt og standardavvik når dataene ikke er normalfordelt. Dette er et synspunkt som er vanskelig å forsvare. Ikke bare er disse størrelsene generelt definert for alle typer fordelinger, det er også disse størrelsene som trengs for å oppsummere data i for eksempel fremtidige metaanalyser. Dette er en god grunn til å rapportere gjennomsnitt og standardavvik for skalavariabler, også når dataene ikke er normalfordelt. Og så kan man oppgi median og kvartiler i tillegg der det er relevant.

Når dataene er kategoriske med få kategorier, for eksempel med de mulige verdiene 1, 2, 3 og 4, vil median og kvartiler være uegnet for å beskrive fordelingen. Det vil vi komme tilbake til i en senere artikkel i Medisin og tall.

LITTERATUR

1. Skovlund E. Bootstrapping – å løfte seg selv etter håret? Tidsskr Nor Legeforen 2019; 139. doi: 10.4045/tidsskr.19.0413. [PubMed][CrossRef]

Publisert: 11. juni 2020. Tidsskr Nor Legeforen. DOI: 10.4045/tidsskr.20.0032