Tidsskriftet
DEN NORSKE LEGEFORENING

Fishers eksakte test – hvordan smaker teen?

MEDISIN OG TALL

STIAN LYDERSEN

E-post: 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.

MORTEN WANG FAGERLAND

Morten Wang Fagerland er ph.d. og leder for Seksjon for biostatistikk og epidemiologi ved Oslo universitetssykehus.
Forfatteren har fylt ut ICMJE-skjemaet og oppgir ingen interessekonflikter.

PETTER LAAKE

Petter Laake er professor emeritus ved Avdeling for biostatistikk ved Institutt for medisinske basalfag, Universitetet i Oslo, og professor II ved Avdeling for helse- og sosialfag ved Høgskolen i Molde.
Forfatteren har fylt ut ICMJE-skjemaet og oppgir ingen interessekonflikter.

Et enkelt eksperiment med te og melk kan ha inspirert den britiske statistikeren Ronald Aylmer Fisher (1890–1962) til å lage en av verdens mest kjente statistiske tester.

En gruppe menn og kvinner er samlet til ettermiddagste på Rothamsted forsøksstasjon for landbruk i slutten av 1920-årene. Psykologen Muriel Bristol hevder hun kan smake om det er teen eller melken som er blitt skjenket i koppen først. For å sjekke om hun har rett, skal Fisher ha servert henne åtte kopper te i randomisert rekkefølge – fire med melk først og fire med te først (1, s. 2 og s. 8). Historien står sentralt i Fishers egen bok *Design of Experiments* (2). Vi tenker oss at Bristol fikk vite at nøyaktig fire av åtte kopper var skjenket med melk først. Et mulig resultat av dette eksperimentet er vist i tabell 1. Radsummene er bestemt på forhånd (fire av hver type). Bristol vil annonsere hvilke fire kopper hun mener fikk melk først, altså er kolonnesummene også bestemt på forhånd. Greide hun å få flere riktige enn man kunne forvente av tilfeldig gjetting? La oss først anta at hun markerte fire av koppene med «melk først» helt tilfeldig. Hvor sannsynlig er det at hun ville få rett på minst tre av dem, slik som vist i tabell 1? I dette eksemplet blir p-verdien lik 0,243 med en ensidig Fishers eksakte test (3, kap. 4). Dette resultatet er ikke overbevisende evidens for at hun har rett i påstanden.

Tabell 1

Et mulig resultat av Muriel Bristols blindtesting av te.

Skjenket	Gjettet melk først	Gjettet te først	Sum
Melk først	3	1	4
Te først	1	3	4
Sum	4	4	8

Studier med få pasienter

Tabell 2 viser resultatet av et randomisert kontrollert forsøk (4). Hvis tallene hadde vært større, kunne vi brukt Pearsons khikvadrattest. Men her er det laveste forventede antallet $34 \cdot 8 / 68 = 4$, altså mindre enn 5, som er Cochrans kriterium, og testen kan ikke brukes. Nullhypotesen er at sannsynligheten for suksess (her: 24 timers overlevelse) er den samme i de to behandlingsgruppene. Men problemet er at denne felles sannsynligheten er ukjent selv om nullhypotesen er sann, og vi trenger den for å beregne p-verdien. Dette problemet løses i Fishers eksakte test ved å late som om det totale antallet suksesser er bestemt på forhånd. Denne teknikken kalles å betinge på kolonnesummene. Resonnementet er at hvis de to behandlingene var nøyaktig like gode, så ville det totale antallet suksesser (her: 8) bli det samme uansett hvilken behandling de enkelte pasientene ble allokert til. Dermed kan forsøket analyseres på tilsvarende måte som blindtestingen av te: Dersom det var like sannsynlig at suksessene kom i den ene eller andre behandlingsgruppen, ville den ensidige p-verdien bli sannsynligheten for at 7 eller flere (dvs. 8) suksesser var i den første behandlingsgruppen, hvilket blir $p = 0{,}027$. Men i medisinsk forskning beregner man vanligvis en tosidig p-verdi. Det er flere måter å beregne denne på, men den mest vanlige er å sette den lik to ganger den ensidige p-verdien, her $p = 2 \cdot 0{,}027 = 0{,}054$. Hvis man setter grensen for statistisk signifikans ved 0,05, viser dette ikke en statistisk signifikant forskjell mellom behandlingsgruppene.

Tabell 2

Behandling av barn med hjertestans. Høy dose versus standard dose adrenalin (4).

	24 timers overlevelse		
Behandling	**Ja**	**Nei**	**Sum**
Standard dose	7	27	34
Høy dose	1	33	34
Sum	8	60	68

Denne fremgangsmåten kan også brukes på 2×2-tabeller med andre design, som en kasus–kontroll-studie eller en tverrsnittstudie.

Finnes det bedre tester i små utvalg?

Fishers ide om å betinge på kolonnesummene er en snedig løsning på problemet med den ukjente felles sannsynligheten. Men det finnes alternative løsninger som gir andre tester som er bedre for små utvalg, fordi de har høyere statistisk styrke enn Fishers eksakte test. Dette vil vi beskrive i den siste av tre artikler om testing i 2×2-tabeller i Medisin og tall.

LITTERATUR:

1. Salsburg D. The lady tasting tea: How statistics revolutionized science in the twentieth century. W.H. Freeman / Owl Book, 2001.

2. Fisher RA. The design of experiments. 8. utg. Edinburgh/London: Oliver & Boyd, 1966.

3. Fagerland M, Lydersen S, Laake P. Statistical analysis of contingency tables. Boca Raton, FL: Chapman

and Hall/CRC, 2017.

4. Perondi MBM, Reis AG, Paiva EF et al. A comparison of high-dose and standard-dose epinephrine in children with cardiac arrest. N Engl J Med 2004; 350: 1722–30. [PubMed][CrossRef]

Publisert: 23. september 2019. Tidsskr Nor Legeforen. DOI: 10.4045/tidsskr.19.0237