Hva vil vi med alle p-verdiene?
Man må vite hvilket spørsmål man stiller før man utfører en signifikanstest

Spør først, regn siden

«Spør først, forsk siden,» oppfordret Are Brean i en lederartikkel i Tidsskriftet nr. 12–13/2012 (1). Han baserte seg på Televerkets gamle råd. Dette rådet kan med hell videreføres til statistisk analyse av medisinske data og beregning av p-verdier.

På et internasjonalt møte erfarte jeg en gang en «interessant» bordplassering. Det var 30 plasser ved et hesteskoformet bord; 12 komitémedlemmer på høyre side, seks ved hovedbordet (formann og sekretariat) og 12 på venstre side. Flere reagerte med latter på at de ni kvinnene var fordelt med sju til venstre, én ved hovedbordet og én til høyre. Noen stilte spørsmålet «hva er p-verdien»? Det lar seg selvsagt beregne. Avhengig av om vi deler i to eller tre sider, eller om vi tar hensyn til at en mannlig deltaker ikke møtte, blir p-verdien mellom 0,005 og 0,03. Aha! Statistisk signifikans på 5 %-nivå! Men hva så? Hva var egentlig spørsmålet eller hypotesen vi testet? Og hva blir konklusjonen?

En p-verdi gir svar på følgende spørsmål: Hva er sannsynligheten for at det resultatet vi observerer (eller et enda «skjevere» resultat) har oppstått på grunn av tilfeldighet? I eksemplet med bordplasseringen sier p-verdien at det er lite sannsynlig at en så skjev fordeling skyldes tilfeldighet. Men vi vet at kjønnsfordelingen er tilfeldig fordi representantene var plassert etter nasjonal tilhørighet med landene i alfabetisk rekkefølge.

Å angi p-verdier i tabeller over demografiske variabler i randomiserte forsøk hører hjemme i samme kategori. Antakelig ønsker man å vise at fordelingen av viktige variabler er lik i behandlingsgruppene, men en p-verdi har ingen plass her. Hensikten er neppe å teste en hypotese om at randomiseringen ikke har «virket». Vi vet jo at pasientene er tilfeldig fordelt mellom gruppene når forsøket er randomisert.

Det er ikke p-verdien som er hovedpoenget med bordplasseringshistorien, men det at vi regner ut en p-verdi uten egentlig å ha klart for oss hva vi spør om Dette har en parallell i forskningsrapporter som skal beskrive «tingenes tilstand». Det er ikke opplagt at signifikanstester har en plass her. Hvis vi ikke stiller et eksplisitt spørsmål, kan vi heller ikke formulere en nullhypotese og en alternativ hypotese. Det er det man implisitt gjør når man beregner en p-verdi og trekker en konklusjon på grunnlag av denne. Dette er selvsagt noe «alle» har lært på kurs, men som dessverre ofte blir glemt. Resepten er: 1) Still spørsmålet. 2) Bestem signifikansnivå (vanligvis 5 %). 3) Formuler en nullhypotese (det du ønsker å motbevise) og en alternativ hypotese (det du tror eller håper er tilfellet). 4) Utfør testen. 5) Sammenlikn p-verdien med signifikansnivået, og forkast nullhypotesen hvis $p < 0,05$.

Ved første øyekast virker det merkelig å akseptere 5 % sannsynlighet for å trekke en gal konklusjon. Hvorfor settes ikke grensen lavere? Det skyldes at det er to typer feil man kan gjøre når man utfører en test; man kan komme til å «oppdage» en effekt som ikke er reell (type I-feil) eller man kan komme til å gå glipp av en forskjell som faktisk finnes (type II-feil). Jo lavere risiko man aksepterer for den ene feilen, desto større blir sannsynligheten for den andre dersom antall observasjoner ikke endres. Sannsynligheten for en type I-feil reguleres med grensen for p-verdien (signifikansnivået). Med andre ord er det 5 % sannsynlighet for å «oppdage» en effekt som ikke er reell (falskt positivt funn).

Gitt et fast signifikansnivå reguleres sannsynligheten for type II-feil av antall pasienter i studien; med flere pasienter er det lettere å avdekke sanne effekter, og teststyrken øker. En p-verdi blir høy når vi har få observasjoner fordi en reell effekt «drukner» i tilfeldig variasjon. I store forsøk med mange pasienter eller i metaanalyser av mange studier kan p-verdier derimot bli veldig små også om de faktiske effektene er minimale og uten klinisk interesse. Skal man vurdere klinisk relevans, gir et effektestimat med tilhørende konfidensintervall mye mer informasjon enn en p-verdi.

Er det viktig å stille spørsmålet før man begynner å analysere data? I bordplasseringseksempelet var det ikke noen hypotese om kjønnsfordeling før vi satte oss, og spørsmålet kom opp «post hoc», altså basert på noe vi allerede hadde observert. Vi analyserte data bare fordi vi syntes fordelingen var urimelig skjev, ikke egentlig for å besvare et spørsmål. Selvsagt må man ha mulighet til å forfølge uventede funn, men en p-verdi danner et mye svakere grunnlag for en konklusjon når analysen er drevet frem av observasjoner i et datasett enn når man først stiller hypotesen og deretter samler data for å besvare spørsmålet (2).

Hvorfor rapporteres det ofte så mange p-verdier? Behandlingseffekt kan måles og rapporteres på forskjellige måter og på forskjellige tidspunkter, og det kan være hensiktsmessig å analysere pasientgrupper med forskjellige karakteristika separat, men verdien av ett positivt funn blant en rekke tester i forskjellige undergrupper er begrenset. Jo flere tester vi utfører, desto større er sannsynligheten for ett eller flere falskt positive funn. I en studie av to behandlinger der det i virkeligheten ikke er noen effektforskjell, vil vi forvente én signifikant p-verdi hvis vi utfører 20 tester. Gjør vi «bare» ti forskjellige (uavhengige) tester, er sannsynligheten for minst ett falskt positivt funn så stor som 40 %. Hvis man tester alt, blir p-verdiene verdiløse. Konklusjoner må baseres på mer enn en p-verdi fra en enkelt studie, særlig hvis en forhåndsspesifisert plausibel hypotese mangler.

Nye ideer som er basert på uventede funn, kan av og til vise seg å være veien til økt kunnskap, og selvsagt er det lov å undersøke og beskrive tingenes tilstand uten å stille konkrete spørsmål på forhånd. Men straks man gir seg i kast med beregning av p-verdier, er det fort gjort å miste gangsynet og ende med å trekke gale konklusjoner. Så spar på signifikanstestene, og still spørsmålet før du regner.

Eva Skovlund
eva.skovlund@farmasi.uio.no

Eva Skovlund (f. 1959) er avdelingsdirektør ved Nasjonalt folkehelseinstitutt og professor II ved Universitetet i Oslo.
Forfatter har fylt ut ICMJE-skjemaet og oppgir ingen interessekonflikter.

Litteratur
1. Brean A. Spør først, forsk siden. Tidsskr Nor Legeforen 2012; 132: 1425.
2. Mills JL. Data torturing. N Engl J Med 1993; 329: 1196–9.