Hvorfor p-verdien er signifikant

P-verdien kan oppleves som et orakel som bedømmer våre resultater. Er p-verdien 0,05 eller lavere, utropes resultatet som signifikant, men er den over 0,05, er resultatet ikke-signifikant og forbigås ofte i stillhet. Hva er egentlig p-verdien, og hvorfor er 0,05 så viktig?

Are Hugo Pripp
apripp@ous-hf.no

> Se lederartikkel side 1424

Få statistiske estimater har gitt så mye glede og så mye skuffelse som p-verdien. Dens størrelse kan ha stor betydning for blant annet klinisk praksis, økonomiske vilkår for legemiddelprodusenter og den enkelte forskers karriere og publisering. Siden den ofte spiller en hovedrolle når resultater diskuteres, er den utstrakte bruken og fortolkningen omstridt.

Bør vi være takknemlige for at p-verdien enkelt kan gi oss svar på om våre resultater er signifikante – eller bør vi forlate hele konseptet om signifikanstesting og forby p-verdier?

Hva er p-verdien?

P-verdien tar utgangspunkt i to hypoteser. Den ene er nullhypotesen, der man vanligvis antar at det er ingen forskjell eller ingen effekt av en behandling eller eksponering. Det betyr at selv om resultatene «rent tallmessig» viser forskjell eller effekt, går vi ut fra at dette skyldes tilfeldig variasjon og dermed ikke en reell statistisk forskjell.

Den andre og dermed den alternative hypotesen er ofte en antakelse om at nullhypotesen ikke er sann. Dernest følger matematisk-statistiske metoder som kan beregne sannsynligheten for det man observerer hvis det faktisk er slik at nullhypotesen er korrekt. Sannsynligheten for det man observerer eller større avvik fra nullhypotesen – gitt at nullhypotesen er korrekt – er vår p-verdi. Hvis p-verdien er lavere enn en på forhånd spesifisert verdi, forkastes nullhypotesen og vi sier at resultatet er statistisk signifikant og påstår at den alternative hypotesen er sann. På den annen side – når resultatet ikke er statistisk signifikant, forkaster vi ikke nullhypotesen.

Hva forteller statistisk hypotesetesting om det vi egentlig ønsker å undersøke med en medisinsk eller klinisk studie? I de aller fleste forskningsstudier er hypotesen at en behandling eller eksponering har en effekt. Den «medisinske» nullhypotesen likner mer på vår statistiske alternative hypotese. Hensikten med en medisinsk eller klinisk forskningsstudie er vanligvis å si noe om sannsynligheten for at en behandling eller eksponering har en effekt og hvor stor denne effekten er. P-verdien er ikke noe mål for denne sannsynligheten og gir lite informasjon om den medisinske effekten. Altså – p -verdien gir informasjon om sannsynligheten for våre observasjoner, gitt nullhypotesen, mens målet med en forskningsstudie er å gi informasjon om sannsynligheten for vår medisinske hypotese, gitt våre observasjoner.

Disse to sannsynlighetene er forskjellige fra hverandre. For å gjøre det ytterligere komplisert, så er det i klinisk forskning nesten alltid en liten forskjell eller effekt selv om denne verken er målbar eller klinisk relevant, noe som betyr at nullhypotesen ytterst sjelden er sann. Hypotesetesting og p-verdier gir egentlig statistiske svar på spørsmål vi ikke stiller oss på bakgrunn av antakelser som ikke stemmer. Likevel bruker vi denne metodikken i de fleste studier – og ofte med stor anvendt nytte.

Hva er p-verdien ikke?

Selv om hypotesetesting og p-verdier er anvendt i medisinsk forskning, så er den statistiske definisjonen forholdsvis kompleks. Det kan derfor være oppklarende å diskutere hva p-verdien faktisk *ikke* er. Såfremt ikke helt spesielle antakelser om dataene gjelder, er følgende vanlige *misforståelser* av p-verdien (1):

- P-verdien er sannsynligheten for at nullhypotesen er sann
- (1 – p-verdien) er sannsynlighet for at den alternative hypotesen er sann
- En lav p-verdi viser at resultatene er gjentakbare
- En lav p-verdi viser at effekten er stor eller at resultatet har teoretisk, klinisk eller praktisk betydning
- Et ikke-signifikant resultat, der vi ikke forkaster nullhypotesen, er et bevis for at nullhypotesen er sann
- Ikke-signifikante resultater er tegn på en mislykket studie

Hvorfor akkurat 0,05?

Æren for et signifikansnivå på 5 % tilskrives statistikeren Ronald A. Fisher (1890–1962). Fisher var en av grunnleggerne av moderne forskningsmetode og statistisk analyse. Hans metoder ble utviklet for bruk i landbruksforskning og genetikk og er siden anvendt innen mange vitenskaper. Han er best kjent for utvikling av variansanalyse og randomiserte studier (2).

I 1925 utga han boken *Statistical methods for research workers*, der han skriver at et signifikansnivå på 5 % er et passende valg (3): «The value for which P = .05, or 1 in 20, …; it is convenient to take this point as a limit in judging whether a deviation is to be considered significant or not» (3, s. 45).

Man kan få inntrykk av at en p-verdi på ≤ 0,05 og denne verdiens betydning i senere forskning skyldes at Ronald A. Fisher mer eller mindre vilkårlig syntes at et signifikansnivå på 5 % var passende. Hvis han heller hadde tenkt på 2 %, 7 % eller 10 %, ville da medisinsk forskning og klinisk praksis ha vært annerledes i dag? Er det slik at resultater og konklusjoner fra store deler av medisinsk forskning avhenger av hvilket tall en statistiker tenkte på for nesten hundre år siden?

Selv om Ronald A. Fisher utvilsomt har hatt en stor betydning for utvikling av forsøksmetodikk og statistikk, blir det en forenkling å gi ham all æren (eller skylden) for at det ble akkurat 5 %. Det er heller ikke riktig at han valgte dette nivået helt vilkårlig, andre statistikere arbeidet med tilsvarende verdier (4).

Cowles & Davis (5) undersøkte hvorfor Fisher valgte 5 % som signifikansnivå. De mener at han kun baserte seg på det som var et innarbeidet konsept. Karl Pearson (1857–1936), en annen grunnlegger av moderne statistikk, utviklet metoder for å vurdere hvor godt data er tilpasset en matematisk sannsynlighetsfordeling, noe som blant annet er grunnlaget for den mye brukte khikvadrattesten av krysstabeller. Han uttrykte at ved en sannsynlighet på 10 % (altså p = 0,1) er det ikke usannsynlig at de observerte data er tilfeldige og videre at ved en sannsynlighet på 1 % (p = 0,01) er det meget usannsynlig at de observerte data kan skyldes tilfeldigheter. Et passende punkt midt imellom er 5 %. William Gosset (1876–1937), som utviklet t-testen, antydet også 5 % som et naturlig signifikansnivå,

Illustrasjon © Espen Friberg/Yokoland

men uttrykte dette statistisk-matematisk noe annerledes (4, 5).

Er det noe spesielt med en sannsynlighet på 5 %? Inspirert av sine historiske undersøkelser rundt anbefalte signifikansnivåer utforsket Cowles & Davis om det finnes et intuitivt og naturgitt signifikansnivå (6). Hvor sjelden må en hendelse forekomme i forhold til det man forventer før man tenker at den opprinnelse antakelsen, altså nullhypotesen, er usann? De nevner et enkelt eksempel. Du og din kollega kaster mynt og kron om hvem som skal kjøpe kaffe til lunsjen, men du taper dag etter dag. Hvor mange dager vil du fortsette med å kjøpe kaffe til din kollega før du mistenker at tapet ikke skyldes tilfeldigheter? Jeg vil anta at mange vil godta dette i fire ($p = 0{,}0625$) eller fem ($p = 0{,}03125$) dager, men vil tro at få antar at det er kun tilfeldigheter hvis man ti dager etter hverandre taper og må kjøpe kaffe ($p < 0{,}001$).

For å undersøke dette systematisk utviklet de et psykologisk eksperiment (6). Frivillige deltakere var med på et pengespill. Foran dem var det tre kopper, og de ble fortalt at det var en liten rød knapp under en av dem. Hvis de gjettet på riktig kopp, vant de penger. Pengespillet ble gjentatt inntil deltakerne selv ønsket å avslutte det.

Den intuitive nullhypotesen for deltakerne er en sannsynlighet på en tredel for å gjette riktig kopp i hver spilleomgang. Det deltakerne ikke visste, var at ingen av koppene skjulte noen rød knapp og at de dermed ville tape hver gang, altså var den intuitive nullhypotesen usann. Eksperimentet gikk ut på å undersøke hvor mange ganger de gjentok spillet før de mistenkte at noe var feil, altså tvilte på nullhypotesen. Over halvparten av deltakerne var mistenksomme etter seks spilleomganger med gjentatt tap ($p = 0{,}088$) og nesten 90 % etter åtte omganger ($p = 0{,}039$). Eksperimentet tydet på at mange naturlig og intuitivt vil velge et signifikansnivå på rundt 5 %.

P-verdiens fremtid i medisinsk forskning

Enkelte kliniske og epidemiologiske forskere har vært meget kritiske til bruk av hypotesetesting og p-verdier i biologisk og medisinsk forskning. Noen mener at hele konseptet er en direkte plage og hindrer den vitenskapelige fremdriften (7, 8). Andre fremtredende forskere mener at p-verdien fremdeles har en viktig funksjon ved statistisk analyse av slike data, selv om de fremhever betydningen av konfidensintervaller og er skeptiske til en absolutt 0,05-regel (9). Det er ofte understreket at rapportering av deskriptiv statistikk og effektmål med konfidensintervaller er viktig og nødvendig i tillegg til p-verdien, deriblant også i anbefaling for bruk av statistikk i Tidsskriftet (10).

Det forskes på metodisk utvikling og alternativer til p-verdien. Enkelte arbeider med metodikk basert på bayesiansk statistikk, der på forhånd antatte sannsynligheter om effekter og resultater tas hensyn til i de statistiske analysene. P-verdien kunne for eksempel justeres på bakgrunn av en medisinsk vurdering av sannsynligheten til nullhypotesen i studien (11). Disse vurderingene uttrykkes statistisk som såkalte apriorisannsynligheter. Apriorisannsynlighetene og de faktiske data fra studien brukes i de statistiske analysene. Et dilemma er selv-

følgelig å bli enig på forhånd om hvilken effekt som er sannsynlig og hvor mye hver enkelt p-verdi bør justeres.

Innen genetisk statistikk har man en utfordring med svært mange tester (multiple sammenlikninger). Selv om p-verdien brukes i slike analyser, gir det lite mening å bruke 5 % signifikansnivå når man tester for eksempel 10 000 gener. Hvis genene er uavhengige av hverandre og det ikke er noen forskjell mellom to grupper, så vil man likevel forvente 500 signifikante tester. Det er derfor utviklet egne metoder for å korrigere for multiple tester innen genetisk statistikk (12), men en systematisk korrigering av alle multiple tester i kliniske studier er omdiskutert (13).

P-verdien er bedre i praksis enn i teorien

Hypotesetesting og p-verdier gir som sagt ikke noe direkte svar på de spørsmål vi har i de fleste medisinske og kliniske forskningsstudier, men likevel brukes denne statistiske metodikken mye. Etter min vurdering er det fordi p-verdien har vist seg nyttig i praksis. Hvis sannsynligheten for våre observasjoner er lav når vi antar at nullhypotesen er sann, er dette et indirekte mål på at våre observerte effekter ikke skyldes tilfeldig variasjon. Det synes fornuftig å gjøre en innledende vurdering av hvorvidt så er tilfellet. Jeg mener at signifikanstesting på 5 %-nivå er som en første statistisk screening før videre vurdering av effektstørrelse og klinisk drøfting. Å gjøre denne vurderingen har vist seg nyttig i praksis og er allment akseptert, selv om den rent metodisk har svakheter med henblikk på dens anvendelse på kliniske data.

P-verdien er basert på statistisk-matematiske metoder og har som sådan liten likhet med et mystisk orakel. Likevel, et orakelsvar er i *Den Danske Ordbog* beskrevet som at det «er ofte tvetydig og kræver tolkning af særlig kyndige» (14). Den beskrivelsen passer også meget godt på bruk av p-verdier i medisinsk og klinisk forskning.

Are Hugo Pripp (f. 1971)
er forsker og biostatistiker ved Oslo Centre of Biostatistics and Epidemiology, Forskningsstøtteavdelingen, Oslo universitetssykehus. Forfatter har fylt ut ICMJE-skjemaet og oppgir ingen interessekonflikter.

Litteratur

1. Nickerson RS. Null hypothesis significance testing: a review of an old and continuing controversy. Psychol Methods 2000; 5: 241–301.
2. Box JFRA. Fisher, the life of scientist. New York: Wiley, 1978.
3. Fisher RA. Statistical methods for research workers. Edinburgh: Oliver and Boyd, 1925.
4. Curran-Everett D. Explorations in statistics: hypothesis tests and P values. Adv Physiol Educ 2009; 33: 81–6.
5. Cowles M, Davis C. On the origins of the. 05 level of statistical significance. Am Psychol 1982; 37: 553–8.
6. Cowles M, Davis C. Is the. 05 level subjectively reasonable? Canadian Journal of Behavioural Science/Revue Canadienne Des Sciences Du Comportement 1982; 14: 248–52.
7. Stang A, Poole C, Kuss O. The ongoing tyranny of statistical significance testing in biomedical research. Eur J Epidemiol 2010; 25: 225–30.
8. Mitchell MS, Yu MC, Whiteside TL. The tyranny of statistics in medicine: a critique of unthinking adherence to an arbitrary p value. Cancer Immunol Immunother 2010; 59: 1137–40.
9. Vanderweele TJ. Re: The ongoing tyranny of statistical significance testing in biomedical research. Eur J Epidemiol 2010; 25: 843–5, author reply 844–5.
10. Aamodt G, Gulbrandsen P, Laake P et al. Presentasjon av statistiske analyser i Tidsskriftet. Tidsskr Nor Lægeforen 2005; 125: 2183–7.
11. Cohen HW. P values: use and misuse in medical literature. Am J Hypertens 2011; 24: 18–23.
12. Montana G. Statistical methods in genetics. Brief Bioinform 2006; 7: 297–308.
13. Perneger TV. What's wrong with Bonferroni adjustments. BMJ 1998; 316: 1236–8.
14. Den Danske Ordbog. http://ordnet.dk/ddo/ordbog?query=orakelsvar (22.4.2015).

Mottatt 23.4. 2015, første revisjon innsendt 28.5. 2015, godkjent 10.6. 2015. Redaktør: Siri Lunde Strømme.

Engelsk oversettelse på www.tidsskriftet.no