P-värdet handlar om slump

P-värden tillskrivs en väldigt stor betydelse vid bedömning av resultat i kliniska studier och gränsen mellan rätt och fel, sant och falskt, går knivskarpt vid 0,05. Eller gör den det? Vad är egentligen ett p-värde och vilken betydelse ska man tillmäta ett statistiskt signifikant resultat? **Anna Törner,** statistiker och verkställande direktör i Scandinavian Development Services, tar ett grepp om de magiska p-värdena.

I en klinisk studie har vi undersökt om ett nytt läkemedel kan lindra symptomen vid behandling av en sjukdom, till exempel depression. Resultaten visar att patienterna som fick det nya läkemedlet har en lägre depressionspoäng på en skala än de som fick placebo. Men detta är inte tillräckligt för att dra slutsatsen att det nya läkemedlet har effekt. Varför?

När det gäller patienterna i den kliniska studien är det tydligt att de som fick aktiv behandling har lägre depressionspoäng. Men, avsikten med studien är ju att göra en bedömning av den sanna skillnaden i effekt, den som vi skulle ha observerat om vi hade behandlat ett oändligt stort antal patienter. Vår kliniska studie är egentligen bara ett litet stickprov av verkligheten och ett verktyg för att kunna dra slutsatser om den sanna effekten av ett nytt läkemedel.

Och för att kunna dra dessa slutsatser behöver man göra statistiska tester. Statistiska tester och p-värden är verktyg för att kvantifiera slumpen och hjälpa oss att avgöra om en skillnad i resultaten kan vara slumpartad eller om den ena behandlingen faktiskt är bättre än den andra.

Något formellt är p-värde sannolikheten för att observera en så stor, eller större, skillnad i effekt mellan två behandlingar som vi faktiskt observerar, om utgångspunkten är att behandlingarna är likvärdiga.

Om sedan sannolikheten för en sådan skillnad blir orimligt liten, till exempel mindre än 5 procent ($p<0,05$), handlar det troligen om en verklig skillnad i effekt. Ett p-värde säger alltså inget om hur stor eller kliniskt viktig den observerade effekten är – bara om det är sannolikt att den kan ha uppkommit genom en slump.

Många faktorer avgör statistiskt signifikans

Flera olika faktorer påverkar om en effektskillnad i en klinisk studie blir statistiskt signifikant eller inte. En viktig faktor är studiens storlek. Stora studier med många patienter ger generellt sett mer trovärdiga resultat än studier med få patienter. En effektskillnad i en stor studie innebär en mindre risk för att den observerade skillnaden beror på slumpen. Det betyder att många patienter i en studie generellt sett ger mindre p-värden, med andra ord mindre sannolikhet för att skillnaden vi observerar kan ha orsakats av slumpen.

En annan viktig faktor är hur stor effektskillnaden är. Om studien visar på en mycket stor effekt av en ny behandling jämfört med kontrollgruppen så är det naturligtvis mer sannolikt att det är en verklig skillnad i effekt.

Ytterligare en viktig faktor är variabilitet i respons, det vill säga hur deltagarna svarar på behandlingen. Om de flesta patienterna reagerar positivt på en ny behandling medan de flesta patienterna i jämförelsegruppen reagerar sämre blir det mer sannolikt att det är en verklig skillnad, även om den genomsnittliga skillnaden inte är så stor. Medan den motsatta situationen, att responsen på behandlingen varierar mycket både mellan patienterna som får ny behandling och inom jämförelsegruppen, gör att det är svårare att uttala sig om slumpen kan ha gett upphov till skillnaden.

Slutsatsen är att statistisk signifikans beror på många faktorer och statistisk signifikans betyder inte nödvändigtvis att skillnaden i effekt mellan två behandlingar är stor och kliniskt viktig.

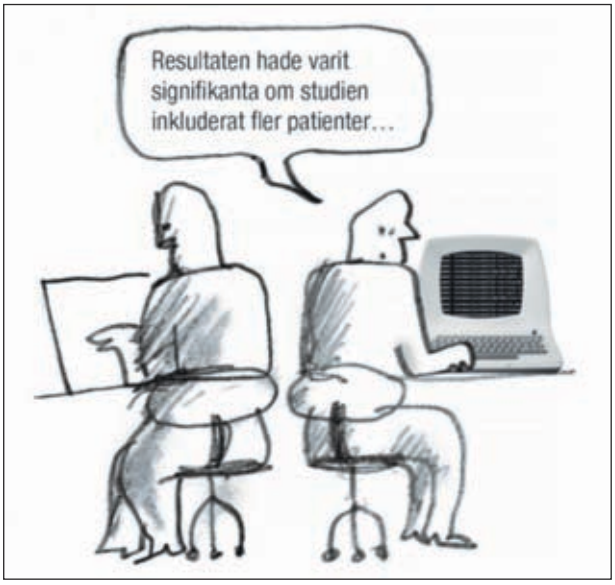
– inte om klinisk effekt

> **Statistiska tester och p-värden är verktyg för att kvantifiera slumpen och hjälpa oss att avgöra om en skillnad i resultaten kan vara slumpartad eller om den ena behandlingen faktiskt är bättre än den andra.**

Gränsen 0,05

Det vanligaste är att vi säger att resultat är statistiskt signifikanta om $p<0,05$, det vill säga om det är max 5 procents chans att observera en så stor (eller större) skillnad av en slump. Kvantitativt är det ingen viktig skillnad mellan $p=0,051$ och $p=0,049$, och att hamna på rätt sida av 0,05 tillmäts ofta en överdriven betydelse. Nivån 0,05 är en konvention och en allmänt accepterad gräns för statistik signifikans, inget ”magiskt” händer vid just $p=0,05$. Ibland nämns p-värden som ligger strax över 0,05 som att ”det är en trend mot signifikant skillnad”. Det är en lite olycklig formulering eftersom trend antyder att något ändras, till exempel över tid, eller att det följer ett mönster för olika dosnivåer. Då är det kanske bättre att skriva att ett p-värde strax över 0,05 är ”nära statistisk signifikans” och ange det aktuella p-värdet.

Ibland tolkas avsaknad av statistisk signifikans, det vill säga att $p>0,05$, som att man då i stället kan dra slutsatsen att det inte är någon skillnad i effekt mellan två behandlingar. Detta är fel, anledningarna kan vara att det inte är någon skillnad i effekt, men kan också beror på att studien är underdimensionerad för att kunna visa om det finns intressanta skillnader i effekt.



Det är inte så enkelt som att fler patienter i en studie ger en bättre signifikans. Det går inte att säga att fler patienter hade visat samma skillnad i klinisk effekt. Å andra sidan kan en studie med ett oändligt stort antal patienter leda till att vilken skillnad som helst bli statistiskt signifikant. Men då blir ju också resultatet ett helt annat och måste bedömas på nytt. Illustration: Anders Gunér

Ibland används Carl Sagans berömda citat (där han säkert tänkte på något annat än p-värden) för att elegant förklara varför icke-signifikanta p-värden inte kan tolkas som bevis för likhet: ”Absence of evidence is not evidence of absence.”

> ”Ett p-värde säger alltså inget om hur stor eller kliniskt viktig den observerade effekten är – bara om det är sannolikt att den kan ha uppkommit genom en slump.

Klinisk relevans och statistisk signifikans

Om en klinisk studie har inkluderat ett mycket stort antal patienter och man bara observerar en mindre skillnad i klinisk effekt så kan man ofta ändå vara ganska säker på att det faktiskt är skillnad i effekt – det stora antalet patienter garanterar att skillnaden inte har uppkommit av en slump. I den situationen kommer ett statistiskt test ge ett lågt p-värde även om skillnaden i effekt mellan två behandlingar inte är så stor. Ett litet p-värde betyder alltså inte nödvändigtvis att skillnaden i effekt är stor – det kan lika gärna betyda att resultatet är baserat på ett mycket stort antal patienter.

Ta till exempel en studie på 20 000 patienter där man studerar risk för venös trombos. Förekomsten av venös trombos visar sig vara 5,3 procent respektive 4,6 procent i de två behandlingsgrupperna och p-värdet beräknas till 0,022, med andra ord statistiskt signifikant. Men är skillnaden kliniskt relevant? Det är förstås en medicinsk och klinisk bedömning om en reduktion i risken för trombos med 0,7 procentenheter är kliniskt relevant. I den bedömningen får man också väga in över hur lång behandlingstid man uppnår den effekten och om det finns andra skillnader, till exempel biverkningar. Statistisk signifikans betyder inte alltid att resultaten är medicinskt relevanta.

Fler patienter kan ge statistisk signifikans

Ibland presenteras resultat från ”nästan-signifikanta” studier som om allt hade ordnat sig om man hade inkluderat ytterligare ett antal patienter. Men så enkelt är det inte eftersom det inte går att säga att en studie med fler patienter hade visat samma skillnad i klinisk effekt. Det är omöjligt att förutspå vad resultatet hade blivit. Å andra sidan kan en studie med ett oändligt stort antal patienter leda till att vilken skillnad som helst bli statistiskt signifikant. Men då blir ju också resultatet ett helt annat och måste bedömas på nytt.

P-värden ger inte hela bilden

Slutsatsen är att p-värden är ett bra verktyg för att kunna bedöma om en skillnad mellan olika behandlingar i en klinisk studie möjligen kan förklaras som ett slumpmässigt fynd eller om resultaten tyder på att det faktiskt är skillnad i effekt mellan två behandlingar. P-värdet säger emellertid ingenting om hur stor skillnaden i effekt är och heller inget om klinisk relevans av observerade resultat. Att blint abdikera från att konstruktivt värdera effektestimat och konfidensintervall och bara trycka på att resultaten är statistiskt signifikanta blir helt fel.

ANNA TÖRNER
statistiker, verkställande direktör
Scandinavian Development Services

Statistiska tester används för att avgöra hur sannolikt det är att skillnaden mellan två interventioner i en studie har uppkommit av ren slump. Om sannolikheten är liten, ofta max 5 procent, väljer man att tro att skillnaden inte förorsakats av slumpen. Många statistiska tester är uppbyggda som en kvot på formatet:. Här är den observerade skillnaden i effekt, s ett uttryck för variabilitet i respons och n patientantal. Om kvoten blir tillräckligt stor så tyder det på att den observerade skillnaden inte uppkommit av en slump och resultaten betecknas som statistiskt signifikanta, p<0,05.