Några kommentarer till 'Signifikansmissbruket i pseudovetenskapen' (Folkvett 2010:4; Qvintensen 2010:3)

Lars Ängquist[1]

28 januari 2011

[1](Filosofie doktor; Matematisk statistik, 2007) Nuvarande anställning: Institut for sygdomsforebyggelse, Köpenhamn, Danmark. Hemadress: Mariedalsvägen 33, S-21745, Malmö, Sverige. Telefon: (+46) 070-3586673. E-mail: Lars.Angquist@telia.com.

Innehåll

1 Inledning och bakgrund

Inom ramen för denna text följer nedan några kommentar till Sven Ove Hanssons (SOH) artikel som publicerats, i något sinsemellan avvikande versioner, i både *Folkvett 2010:4* [1] och *Qvintensen 2010:3* [2]. I det förstnämnda fallet rör det sig om, och kring, vad som däri hänvisas till som Missförstånd 4 (s. 30-31) och i det senare fallet istället som Missförstånd 3 (s. 10).

> *Missförstånd: Signifikant avvikelse från nollhypotesen är ett adekvat mått på graden av vetenskaplig bevisning.*

SOH rekommenderar i avsnittet att man vid en testningssituation visavi en levande försöksperson (till exempel en individ med påstådda övernaturliga förmågor) går till väga såsom att man initialt frågar personen om hur stor effekt (i någon mening) som den anspelar på att kunna visa och argumenterar sedan, om jag har förstått det hela korrekt, för att man senare konkret använder sig av denna effekt som guidning till att finna en vattendelare (gräns) mellan lyckade/misslyckade testutfall. I samband med detta används då även antalet testobservationer (försök) till att kalibrera försökuppställningen i akt och mening att få denna delningströskel till att motsvara signifikans (med avseende på nollhypotesen).

Genom hela texten, om ej annat anges, så antar vi nu här att ett test består av att testpersonen a priori skall välja ett av två möjliga utfall, där en förmåga innebär att man ovanligt ofta gissar rätt, och en motsvarande oförmåga består av att man lika gärna gissar fel som rätt (båda alternativen varandes lika sannolika). Om detta utförs som en serie oberoende likafördelade försök så leder detta då till att antalet rätta svar, som vi kallar n, motsvarar en observation från en binomialfördelning, $Bin(N, p)$, där N är seriens längd (antalet försök) och p den sanna sannolikheten att personen gissar rätt.[1]

Låt mig här slutligen kort påpeka att jag till fullo i det stora hela håller med SOH om de viktiga förtydliganden, om vådan av att tolka statistisk signifikans, som fint och välformulerat läggs fram i ovan nämnda artikel.

2 Motivering av kommentarer

Som jag ser det hela så blir det ovan beskrivna tillvägagångssättet lätt förvirrande då det, som jag tolkar det, leder till en något otydlig variant av argumentsamling baserad på de väldefinerade hypotestestsbegreppen signifikansnivå och statistisk styrka. En viktig del av problemet är att det kan argumenteras för att det ter sig lite oklart vad man egentligen testar (vilka antaganden om detta som man gör).

Enligt de argument som SOH för fram så skulle man kunna tolka det huvudsakliga intresset för att använda sig av informationen om uppskattad effekt (från försökspersonen; för exemplet antas här, som av SOH, 75% korrekta svar) på två olika sätt: antingen att man vill testa om någon effekt överhuvudtaget existerar

$$\text{nollhypotes } H_0 : p = 0.5 \text{ ; } \quad \text{mothypotes } H_A : p > 0.5$$

och att man då även använder denna information till att kalibrera den statistiska styrkan[2] i syfte att ge den testade rimliga förutsättningar att lyckas påvisa en uttalad förmåga (vilket ökar trovärdigheten/kvalitéten på testet) eller så sätter man istället resolut denna uppskattning till att vara nollhypotes – som då leder till en fullständig testformulering genom

$$\text{nollhypotes } H_0 : p = 0.75 \text{ ; } \quad \text{mothypotes } H_A : p > 0.75$$

[1]En icke-existerande förmåga motsvarar då $p = 0.5$ och en reell förmåga, av någon storlek, $p > 0.5$.

[2]Scensättande: Givet att man definerat vilka utfall som anses som signifikanta/icke-signifikanta så är *signifikansnivån* och den *statistiska styrkan* sannolikheterna för att man får ett signifikant utfall om nollhypotesen H_0 respektive en speciell instans som tillhör mothypotsen H_A är sann. Notera att styrkan beror på vilken instans (tillhörandes mothypotesen) som avses och är, vagt uttryckt, en växande funktion av avståndet mellan den valda instansen och H_0.

vilket ändrar innebörd på testet, dvs man testar inte längre om någon *som helst effekt* föreligger utan istället om effekten är *större/lika med* den uppskattade förmågan.

3 Räkneexempel

Antag, som i SOHs exempel, att vi observerar binära utfall (korrekt/fel gissning) och att man, som beskrivits ovan, är intresserad av effekter som innebär att man oftare än slumpen svarar korrekt. Antag vidare initialt att även vi, som SOH, bestämmer oss för att köra en serie om 40 oberoende försök - ledandes (teoretiskt sett) till en observation från en $Bin(40, p)$ fördelning. Låt oss nu konsultera tre olika fall:

Fall 1 Antag $H_0 : p = 0.5$. Det krävs nu $n \geq 26$ för att få ett signifikant utfall med ett ensidigt test på signifikansnivån 0.05.

Fall 2 Antag $H_0 : p = 0.75$. I detta fall måste vi sträcka oss ända upp till $n \geq 35$, vid samma signifikansnivå, för att påvisa signifikanta utfall.

Fall 3 Antag SOHs procedur. Här sätts gränsen mittemellan de båda ovanstående alternativen, till $n \geq 30$, vilket inte riktigt kan motiveras annat än utifrån att gränsen (det diskreta tröskelvärdet) här var den initiala uppskattningen på effekten. Notera att detta är ett väldigt signifikant utfall ($p \approx 0.001$) under Fall 1 och ett högst alldagligt utfall ($p \approx 0.584$) under Fall 2.[3]

4 Potentiella problem med Fall 3

Om man generellt antar angreppssättet att deklarera signifikans utifrån den initialt uppskattade effekten så hamnar man lätt, om antalet observationer N sätts för lågt, till att man även under nollhypotesen väldigt enkelt kan finna signifikans (dvs signifikansnivån blir i dessa fall obrukbart hög). Typexempel: Antag situationen ovan fast med enbart 4 observationer – att få minst 75% korrekta svar motsvarar här $n \geq 3$ vilket har hela 0.3125 i sannolikhet att uppträda under $H_0 : p = 0.5$.

[3] Signifikansen under H_0 noteras även av SOH.

Det kan också ge ett lite godtyckligt intryck då det, enligt mig och antytt ovan, ter sig något oklart vilken nollhypotes man egentligen är intresserad av att testa. I SOHs exempel där 29 anses som ett oacceptabelt utfall för att ge evidens för en förmåga så motsvarar rent faktiskt detta ett väldigt starkt stöd (i standardmässig signifikansmening) vid test avseende Fall 1 ($p \approx 0.003$), men inte alls gällande Fall 2 ($p \approx 0.715$).[4] Naturligtvis är det generellt sett en skämspraxis, som SOH indikerar, att ändra hypoteser efter att man fått tillfälle att granska utfallet men problemet här torde snarare vara att man från början inte varit tydlig nog med vad man egentligen velat testa.

En tolkning där man, återigen enligt mig, med större tillförsikt kan använda Fall 3-proceduren är i vad man skulle kunna kalla för *vadslagningssituationen* (eller strukturellt identiska situationer) där man, givet utfall, antingen vinner eller förlorar. Om man då utger sig för att kunna påvisa en effekt av en viss storlek så skulle ett lämpligt alternativ för att skilja mellan dessa utfall vara att just kräva att man (allra minst) kan uppnå det som man säger sig kunna utföra.

5 Något om statistisk styrka och signifikansnivåer

Som tidigare antytts så hänger begreppen signifikansninå och styrka intimt samman. Vidare så är oftast, givet en viss signifikansnivå, styrkan en ökande funktion av antalet observationer. Detta leder till att man (något schematiskt) kan finjustera en testuppsättning genom att initialt ansätta en rimligt låg signifikansnivå och sedan (i mån det är praktiskt möjligt) välja ett så pass stort antal observationer att styrkan (helst även för minsta möjliga effektstorlek varandes av intresse) blir hög nog. Är det sedan de facto möjligt att samla in ytterligare observationer så kan man som en följd antingen ytterligare sänka signifikansnivån, öka styrkan, eller en aning utav båda delarna.[5]

Optimalt sett så vill man egentligen minimera båda de möjliga testfelriskerna – signifikansnivån och styrkekomplementet (1-styrkan) – reglerandes frekvensen av falskt positiva respektive falskt negativa utfall. I det förstnämnda fallet så ligger fokus *på testaren* i den meningen att det skall säkerställas att testet blir rigoröst, i det andra fallet är fokus istället på *den testade* då

[4]Inte heller utfallet $n = 30$ ger starkt stöd vid nyttjande av Fall 2, se siffra ovan samt vidare resonemang i nästa avsnitt.

[5]Genom att sänka signifikansnivån så sänks generellt sett samtidigt även styrkan på grund av att rejektionsregionen snävas av.

denna skall försäkras en rimlig chans att påvisa förmåga. Båda faktorerna ökar trovärdigheten/tillförlitligheten med avseende på testet.[6]

Ett relaterat implicit problem med den diskuterade proceduren, som skisserats som Fall 3 ovan, är att det blir lite oklart hur man förhåller sig till dessa faktorer. Om man vill testa nollhypotesen $H_0 = 0.5$ enligt Fall 1 så blir styrkan på testet vid gissat (tröskel)värde ($p = 0.75$) högst måttlig. Detta kan ses genom en generell observation att styrkan att exakt få, eller överskrida, antaget väntevärde i de allra flesta testfall per definition blir runt 0.50.[7]

Referenser

[1] S. O. Hansson: *Signifikansmissbruket inom pseudovetenskapen.* Folkvett 2010:4, s. 25-33 (2010).

[2] S. O. Hansson: *Signifikansmissbruket inom pseudovetenskap.* Qvintensen 2010:3, s. 9-10 (2010).

[6] Ett sätt att utvärdera ett tests effektivitet med avseende på att väga samman sådana möjiga val (avvägningar) ges genom att använda sig av så kallade ROC-kurvor; dessa kan t.ex antingen granskas grafiskt eller utnyttjas för tillhörande beräkning av area-underkurvan (av intresse i sig självt eller genom jämförelse mellan olika teststatistikor).

[7] För en kontinuerlig symmetrisk fördelning blir sannolikheten exakt.