PEDAGOGISK FORSKNING I SVERIGE 2010 ÅRG 15 NR 2/3 S 242–247 ISSN 1401-6788

Om användningen av inferensstatistik i vetenskapliga undersökningar

HORST LÖFGREN
Malmö högskola

WIDAR HENRIKSSON
Institutionen för tillämpad utbildningsvetenskap, Umeå universitet

I många undersökningar används inferensstatistik, det vill säga signifikansprövningar och angivande av probabilitetsvärden på ett högst tveksamt eller till och med felaktigt sätt. Även många konsumenter av forskning saknar förståelse för vad signifikans eller signifikanta skillnader innebär. Ofta uppfattas signifikansbegreppet som ett tillförlitlighetsmått, som visar att de skillnader som framkommer i den kvantitativa analysen av data från den studerade gruppen i någon mening är sanna, är vetenskapligt belagda. Så är det inte! Signifikansanalyser ska i flesta fall användas för att pröva möjligheten att utifrån stickprov kunna generalisera resultat till en större och bakomliggande population. Även om man har representativa stickprov i sin undersökning ger erhållna p-värden vid signifikansanalyser inte tillräcklig information. Man vill självfallet, utöver att skillnader finns, också veta något om hur stora skillnaderna är.

BEGREPPET STATISTISK INFERENS

Begreppet statistisk inferens refererar till en situation, där man av olika anledningar väljer att (eller inte kan) studera hela populationen. Man bygger på sannolikhetsteori för att kunna dra slutsatser om förhållanden i populationen. Dock finns alltid en risk, om än liten, för att man, utifrån observation av förhållanden i stickprovet, drar en felaktig slutsats om motsvarande förhållanden i populationen. Risken för felaktiga slutsatser av så kallade »typ I-fel» reglerar man exempelvis via val av signifikansnivå.

Ett krav är att man har tillgång till stickprov som baseras på ett slumpmässigt urval från en population. Inom samhällsvetenskap är det synnerligen ovanligt, att man kan utgå från en undersökningsuppläggning som bygger på ett obundet slumpmässigt urval, och man brukar därför oftast utgå från begreppet representativitet i stället. Den frågeställning som då aktualiseras är om stickprovet är representativt för den definierade populationen. För att

kunna genomföra denna representativitetsprövning krävs att man (både för population och stickprov) har tillgång till så kallade stratifieringsvariabler, det vill säga variabler som kan antas påverka studiens utfall.

Om man exempelvis, utifrån tillgänglig kunskap, vet att kön är en variabel som är av betydelse för utfallet så utnyttjas denna (plus ytterligare variabler) vid prövningen av representativitet. Denna representativitetsprövning är en statistisk inferens, det vill säga man prövar om skillnaden är statistiskt signifikant eller icke signifikant för respektive stratifieringsvariabel. Givet signifikans är stickprovet inte representativt för populationen i den valda stratifieringsvariabeln och vice versa. Slutsatsen är således att statistisk inferens alltid förutsätter en definierad population och tillgång till ett representativt stickprov från denna population. Om så inte är fallet, ska inte inferensstatistik användas.

Deskriptiva undersökningar, där man jämför utfallet i olika undersökningsgrupper och beskriver skillnaderna med hjälp av lämpliga storleksmått, behöver ingalunda vara av mindre värde för kunskapsutvecklingen än stickprovsundersökningar från vilka man kan dra generella slutsatser om förhållanden i en definierad population. Däremot kan undersökningar, i vilka man använder och anger resultat i form av värden i testfunktioner (t ex F-kvot, t-värde, etc), signifikansangivelser i form av p-värden eller stjärnsystem, vara av tveksamt värde, om man inte också anger storleksmått på erhållna skillnader. Detta gäller även om stickproven är korrekt dragna ur en definierad population. Vid tillräckligt stora stickprov kan nämligen även marginella skillnader bli högst signifikanta. Storleksmått påverkas dock inte av stickprovsstorlek.

Om man gör en populationsundersökning, det vill säga undersöker alla i populationen, ska man självfallet inte göra några signifikansanalyser. Man har ju all information och ingen bakomliggande population att generalisera sina resultat till.

Även om *statistisk signifikans* fortfarande kan väcka viss vördnad bland okunniga, är det hög tid att reagera mot felaktig användning av inferensbegreppet. Att felaktig användning av inferensstatistik fortfarande är så frekvent beror sannolikt på att inferensmetoder alltför ofta betonats i grundkurser i statistisk dataanalys och att man inte riktigt har förstått signifikansbegreppet. Det är därför angeläget att mått på effektstorlek förs in i grundläggande kurser i kvantitativ dataanalys.

Eftersom ordet *signifikant* kan användas i andra betydelser än den man använder i statistiska sammanhang, borde man vara noga med att använda *statistisk signifikans* eller att något är *statistiskt signifikant*, om det är detta man menar.

Om man har tillgång till ett slumpmässigt urval eller åtminstone ett representativt urval måste man göra allt man kan för att undvika bortfall. Det finns visserligen metoder för att hantera bortfall, men risken är alltid att bortfall leder till att urvalet inte längre kan betraktas som slumpmässigt.

Här bör kanske nämnas att det finns en grupp av test (goodness-of-fit), där visserligen resultatet av en prövning anges i sannolikhetstermer, men där man inte är ute efter att generalisera från ett individstickprov till en bakomliggande individpopulation. Det kan exempelvis gälla att pröva ett teoretiskt antagande

om en viss fördelning eller att pröva om en teoretisk modell kan vara giltig för insamlade empiriska data.

Även om man har en definierad population och ett representativt stickprov, så kan det tyvärr ändå bli fel!

Om man har små stickprov (under c:a 30 observationer) kan signifikanstestningar ändå bli tvivelaktiga, eftersom då *statistical power*, som är en funktion av stickprovsstorlek, effektstorlek och vald signifikansnivå (p), är för låg för att upptäcka skillnader. Det så kallade typ II-felet (att felaktigt behålla nollhypotesen, trots att den är fel) riskerar att bli för stort.

Här bör också nämnas att det finns en grupp av statistiska metoder, så kallade icke-parametriska metoder, som med fördel kan användas vid små stickprov och i de situationer där antaganden, som den statistiska prövningen med en parametrisk metod baseras på, inte är uppfyllda. En nackdel med flera av de icke-parametriska metoderna är dock att det är svårare att finna index på effekternas storlek.

Även om styrkan i analysen är hög, kan signifikansen ändå bli av tveksamt värde!

Även om man har stora stickprov, och typ II-felet är lågt, kan signifikansanalyser ändå bli missledande. Vid stora stickprov blir nämligen även små och triviala skillnader signifikanta. Detta är kanske den största faran i många undersökningar. Stickproven är så stora, att även den minsta skillnad blir signifikant.

Stickprovets storlek

Stickprovets storlek avgör hur stort utrymme slumpfaktorer får att operera. Ju större stickprov desto mindre risk för slumpfaktorer att påverka. Dock bör man inte misstro ett ganska litet men korrekt draget stickprov. Man kan bestämma lämpligt stickprovsstorlek utifrån typ II-fel och önskad styrka i det statistiska testet. Dock tar man i de allra flesta fall till i överkant, när man väljer stickprovsstorlek.

Vanligt fel vid vissa typer av urval

Om man inte har ett stickprov av individer utan baserar urvalet på ett antal valda grupper, till exempel klasser, skolor eller någon annan gruppering, måste man ta hänsyn till detta. Ett vanligt fel, när man har gruppurval, är att man analyserar data som om de kommer från ett obundet slumpmässigt urval. Om man exempelvis i en pedagogisk undersökning valt ut ett antal klasser och samlat in data från eleverna, kan man inte utgå ifrån att dessa observationer är oberoende. Gruppen, klasskamraterna och lärarna kan ha betydelse, och om så är fallet kan man inte använda statistiska analyser, som förutsätter oberoende observationer. Dock finns metoder att hantera data som baseras på gruppurval.

Sammanfattande råd

- I de flesta fall bör man mer uppmärksamma effektstorlek än signifikansanalyser.
- Har man ett representativt stickprov från någon definierad population, kan man använda statistiska signifikansanalyser. Dock räcker det inte med att endast ange p-värden. Man ska även redovisa effektmått samt precisionsmått för effektskattningen, helst också ange konfidensintervall för effekten.
- Man bör ta ställning till om den valda statistiska metodens villkor och antaganden är uppfyllda. Om villkor för en parametrisk metod inte är uppfyllda, bör man använda en motsvarande icke-parametrisk metod.

Krav på storleksmått

I forskningsmetodiska kurser inom det kvantitativa området betonas, som ovan nämnts, alltför mycket signifikansprövningar, och alltför lite ägnas åt mått på effektstorleksmått och konfidensintervallskattning. I flera decennier har man försökt få forskare att förstå missbruket av signifikansanalyser, när inte förutsättningarna är uppfyllda, för att i stället använda mått på effektstorlek. Om förutsättningarna är uppfyllda för signifikansprövningar bör ändå dessa alltid följas av mått på effektstorlek. Det är självfallet viktigt att veta hur mycket, än att bara veta att något med en viss grad av säkerhet finns i populationen. Från och med den femte upplagan av *Publication Manual*, utgiven av American Psychological Association (APA), finns ett tydligt krav vid publicering av forskning att förutom eventuella signifikansanalyser också ange lämpliga storleksmått.

Här bör nämnas att ett konfidensintervall för effekten är intressant att studera även om effekten inte är statistiskt signifikant. Om man har ett »smalt konfidensintervall som innehåller nollan» blir slutsatsen att någon effekt inte kunnat påvisas, och om en effekt ändå skulle finnas torde den vara liten. Har man däremot ett »brett intervall som innehåller nollan» är slutsatsen återigen att någon effekt inte kunnat påvisas, men nu kan det finnas en effekt, som till och med kan vara ganska stor, som kanske inte kunnat påvisas på grund av att stickprovet varit för litet.

Vad visar det om ...?

- Storleksmåttet visar en obetydlig eller liten skillnad men signifikansprövningen att skillnaden är signifikant? Då har man haft ett onödigt stort stickprov!
- Storleksmåttet visar en påtaglig skillnad men signifikansprövningen att skillnaden inte är signifikant? Då har man haft ett för litet stickprov!
- Storleksmåttet visar en stor skillnad och signifikansprövningen en högst signifikant skillnad? Då har man haft ett lagom stort stickprov.

EFFEKTSTORLEK

Effektstorlek (ES) är en familj av mått som används för att beskriva resultatskillnader mellan behandlingsgrupper. Till skillnad från signifikansanalyser

påverkas inte dessa storleksmått av stickprovsstorlek. Man använder vanligen endera av följande:
(1) Den standardiserade skillnaden mellan två medelvärden.
(2) Sambandet mellan den oberoende behandlingsvariabeln och den beroende utfallsvariabeln.

I statistiska programvaror (t ex SPSS) finns en del storleksmått tillgängliga. På senare år finns effektstorleksmått med i handböcker om statistisk dataanalys, och många artiklar finns att hämta på Internet.

EXEMPEL PÅ STORLEKSMÅTT

Cohens d
Cohens $d = (M_1 - M_2)/s_{pooled}$, där $s_p = \sqrt{\frac{s_1^2 + s_2^2}{2}}$

Vid lika stora jämförelsegrupper och vid olika stora grupper blir

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

På Internet kan man finna hur man enkelt beräknar Cohens *d*.

Förslagsvis kan ES 0,20 betraktas som små skillnader, ES ca 0,40 som måttliga skillnader och ES 0,60 som stora skillnader. Cohen's *d* kan även beräknas utifrån resultat av t-test och F-test.

Eta-kvadrat
Eta-kvadrat är ett mått på relationen mellan oberoende och beroende variabel. Eta-kvadrat är den andel av den totala variansen som kan förklaras av variationen i den oberoende variabeln (betecknas ofta g^2). I en variansanalys erhålls eta-kvadrat via relationen $SS_{mellan\ grupper}/SS_{total}$, det vill säga kvadratsumman mellan grupper dividerat med den totala kvadratsumman. Eta och Eta-kvadrat kan erhållas direkt vid variansanalytisk bearbetning i till exempel SPSS. Bedömningen av vad som är en stor eller liten skillnad är relaterat till vad man har anledning att vänta sig. I påverkansundersökningar inom beteendevetenskaplig forskning, där man sällan finner stora effekter av insatta åtgärder, kan förslagsvis följande gränser användas:

< 0,05 liten skillnad
0,05–0,09 medelstor skillnad
> 0,09 stor skillnad

Eftersom eta^2 är mått på relationen mellan oberoende och beroende variabel kan man i stället tala om svag, påtaglig och stark relation.

Cramérs index
Detta index beräknas genom att ta kvadratroten ur Chi-kvadratvärdet dividerat med antalet observationer (N) multiplicerat med (s-1); s = minsta antalet av rader och/eller kolumner. För en fyrfältstabell blir s=1 och Cramérs index lika med den så kallade phi-koefficienten. Fördelen med Cramérs V är att detta sambandsmått inte har olika max-värden för tabeller av olika storlek utan alltid har variationsområdet 0–1.

Vad som kan betraktas som svagt resp. påtagligt samband beror självfallet på vad som studeras och vad man har anledning att vänta sig (jfr eta-värden ovan). Ett enkelt sätt att få förståelse för sambandsstorlek är att kvadrera korrelationen. En korrelation på exempelvis 0,30 innebär således att den ena variabeln endast kan förklara 9% av den andra, det vill säga hela 91% är oförklarad varians. I en situation där man förväntar sig att sambandet är närmast noll, till exempel samband mellan längd och begåvning, kan dock en korrelation på 0,30 uppfattas som mycket hög.

Det är vår förhoppning att denna artikel ska bidra till att minska på felanvändningen av statistiska signifikansanalyser och att storleksmått och konfidensintervall används vid resultatredovisning.