

Bland tomtar, troll och forskare

I vilken utsträckning handlar forskningsgärningen om att berätta mer eller mindre sannolika sagor som ”passar” data? Och vad skulle hända om vi lade mindre vikt på berättandet och lät forskningsresultaten tala för sig själva?

Tidigare i år skrev Magnus Johannesson och jag om de projekt vi bedriver tillsammans med en grupp medförfattare. I dessa gör vi om – replikerar – tidigare studier på nya och större urval för att studera andelen falska positiva resultat (resultat som felaktigt förkastar nollhypotesen). Vi testar också om forskare kan förutse replikationsresultaten (Dreber och Johannesson 2018). Nu har vi nyligen publicerat resultaten från ett projekt där vi replikerade 21 studier publicerade i de allmänvetenskapliga tidskrifterna *Nature* och *Science* (Camerer m fl 2018, med bl a nuvarande och tidigare HHS-doktoranderna Adam Altmejd, Eskil Forsell, Emma Heikensten och Siri Isaksson). De 21 studierna är experimentella och testar samhällsvetenskapliga hypoteser.

Vi har i den nya studien betydligt högre statistisk styrka än i tidigare stora replikationsprojekt inom psykologi och experimentell ekonomi. Statistisk styrka är sannolikheten att hitta ett statistiskt signifikant resultat givet att hypotesen är sann och beror på antal observationer, variansen i data och den förväntade effektstorleken. De tidigare replikationerna hade ca 90 procent styrka att hitta 100 procent av den ursprungliga effektstorleken. Det kan ju låta bra, men om de ursprungliga effektstorlekarna är överskattade även för sanna positiva resultat, vilket mycket tyder på, finns det en risk att det ändå blir en betydande andel falska negativa resultat på grund av låg styrka.

I det senaste projektet genomfördes replikationer i två steg. I ett första steg av datainsamling siktade vi på att ha 90 procents styrka att hitta 75 procent av den ursprungliga effektstorleken, vilket var en ökning i styrka relativt tidigare projekt. Om replikationsresultatet inte var statistiskt signifikant ($p < 0,05$) i den ursprungliga studiens riktning fortsatte vi med datainsamlingen så att vi totalt hade 90 procents styrka att hitta 50 procent av den ursprungliga effektstorleken. I dessa fall var urvalsstorlekarna i genomsnitt ca sex gånger så stora som i originalstudierna.

Vi finner i den nya studien att 13 av 21 resultat går att replikera. Om vi endast hade genomfört den första replikationsomgången hade vi missat att hitta stöd för två studier som endast går att replikera när vi har betydligt högre styrka. Vi finner också att genomsnittlig relativ effektstorlek – hur stor effekten är i replikationsstudien jämfört med den ursprungliga studien

– för de effekter som vi kunde replikera är 74 procent, medan icke-replikerbara studier i genomsnitt har en relativ effektstorlek kring noll. Det här är ännu ett exempel på att även sanna positiva resultat har överdrivna effektstorlekar när de först publiceras.

Med hjälp av prognosmarknader och enkäter kan vi också konstatera att forskare i hög utsträckning kan förutse replikationsresultaten. Det tyder på att något systematiskt kännetecknar de studier som inte går att replikera och som troligen innehåller falska positiva resultat. Frågan blir då varför dessa och liknande resultat ändå publiceras så framgångsrikt. Vad händer i granskningsprocessen? Är gruppen av två till fem granskare, som de flesta tidskrifter använder sig av, för liten för att vi ska kunna få någon *wisdom of crowds* som på prognosmarknaden, eller är de flesta granskare statistiskt naiva jämfört med våra prognosmarknadsdeltagare? Det kan också vara så att forskare över tid får en känsla för vilka typer av studier eller forskningsgrupper som är mer pålitliga än andra – prognosmarknadsdeltagarna vet vilka ursprungsförfattarna är.

Tidskrifter kanske också premierar något annat än replikerbarhet. Under en av mina medförfattares seminarier där våra resultat presenterades sa en redaktör för en topp fem-tidskrift i nationalekonomi att hen inte var ute efter att publicera replikerbara resultat. Det redaktören kanske menade var att hen vill publicera nya och utmanande idéer. Vem vill inte det? Men jag undrar om redaktören även, precis som så många andra, var ute efter något annat. Kanske eftersökte hen en bra *story* – en tydlig och koherent berättelse som kan förklara just de resultat som rapporteras i den publicerade artikeln. Detta trots att resultaten ofta kunde ha blivit annorlunda i både riktning och storlek och trots att förklarande mekanismer likaså kunde ha varit annorlunda än de som presenterats. Hade resultaten blivit annorlunda skulle berättelsen också ha sett annorlunda ut. Ofta ger den här typen av lättillgänglig berättelse ett falskt intryck av säkerhet kring hur resultaten passar in i tidigare litteratur – inom många fält finns det så mycket tidigare forskning med motstridiga resultat att det är relativt enkelt att hitta studier som backar upp resultat i alla möjliga riktningar. Ofta ges dessutom ett falskt intryck av säkerhet i resultaten med flera decimaler i angivna effektstorlekar och p-värden.

Nu försöker många stressade doktorander, liksom tidigare år, komma in på den akademiska jobbmarknaden. Och med det kommer alla dessa diskussioner om hur resultat ska ”säljas” och paketeras. Självklart är det bra om man, när man har skrivit en uppsats, kan förstå varför frågan är viktig, vad bidraget är och vad studien finner. Men är och bör vi vara ”försäljare” av våra resultat? Och hur mycket premieras övertygande marknadsföring över transparens och ärlighet kring vad vi inte vet och hur osäkra skattningarna oftast är? För om forskningsresultaten kan tala för sig själva, och ärlighet och öppenhet kring resultaten är viktigast, varför lägger vi och våra doktorander så mycket tid på berättandet? Och vad skulle hända om vi slutade med det? Det är nog naivt att tro att det här är något som kommer att för-

ändras till det bättre inom en snar framtid. Men om vi vill att forskningsresultat ska fortsätta få tyngd i olika debatter bör vi nog tänka mer kring hur sanningsmaximerande våra normer faktiskt är.

Anna Dreber Almenberg

REFERENSER

Camerer, C F m fl (2018), "Evaluating the Replicability of Social Science Experiments in Nature and Science between 2010 and 2015", *Nature Human Behaviour*, 27 augusti 2018.

Dreber Almenberg, A och M Johannesson (2018), "Vilka forskningsresultat kan vi lita på?" , *Ekonomisk Debatt*, årg 46, nr 2, s 17-28.