Ska vi sluta använda oss av statistisk signifikans?

MATTIAS NORDIN
är doktor i nationalekonomi och forskare i statistik vid Uppsala universitet. Han bedriver forskning inom experimentell design, men även tillämpad forskning inom, bl a, politisk ekonomi och arbetsmarknadsekonomi.
mattias.nordin@statistik.uu.se

På senare tid har en rad röster höjts för att statistisk signifikans inte längre ska användas som ett kriterium för att dra slutsatser om vetenskapliga hypoteser. I denna artikel ger jag en kort översikt av denna strömning och presenterar argument för att det finns anledning att avskaffa användandet av statistisk signifikans vid test av hypoteser. Jag visar att statistiskt signifikanta effekter är överestimat av sanna effekter och diskuterar även hur p-värdet kan misstolkas. Slutligen diskuterar jag vad det finns för alternativ till användandet av statistisk signifikans.

För ett år sedan publicerades ett upprop i tidskriften *Nature* med uppmaningen att man inte längre ska prata om ”statistisk signifikans” (Amrheim m fl 2019). Ungefär samtidigt publicerade tidskriften *The American Statistician* (som drivs av The American Statistical Association, ASA) ett temanummer med sammanlagt 43 artiklar som på ett eller annat sätt berörde statistisk signifikans, hypotestest och *p*-värden. I introduktionsartikeln till numret, ”Moving to a World Beyond’$p < 0.05$’”, argumenterade även Wasserstein m fl (2019) för ett avskaffande av statistisk signifikans. De skrev så här:

> We conclude, based on our review of the articles in this special issue and the broader literature, that it is time to stop using the term ’statistically significant’ entirely. Nor should variants such as ’significantly different’, ’$p < 0.05$ ’, and ’nonsignificant’ survive, whether expressed in words, by asterisks in a table, or in some other way. (s 2)

Givet att hypotestestet har varit en grundbult i den statistiska verktygslådan under de senaste hundra åren finns det all anledning att fråga sig var denna helomvändning kommer ifrån. Jag ämnar inte ge en fullständig översikt över alla argument för och emot användning av statistisk signifikans, utan hänvisar till temanumret, samt övrig litteratur refererad i denna artikel för den intresserade läsaren. I stället vill jag lyfta ett par poänger som jag anser relevanta, speciellt för den tillämpade nationalekonomiska forskaren.

Det har under 2000-talet argumenterats mer och mer för att det finns allvarliga problem inom den empiriska forskningen. Ett tidigt exempel på denna strömning är Ioannidis (2005) som argumenterade för att den mesta empiriska forskningen som baseras på statistisk signifikans är ”felaktig”. Replikationskrisen inom psykologi (Open Science Collaboration 2015) lyfte frågan, om huruvida empirisk forskning baserad på statistisk

Jag är mycket tacksam för värdefulla kommentarer från Adrian Adermon, Anna Dreber, Mikael Elinder, Georg Graetz och Per Johansson.

signifikans ger tillförlitliga resultat, högst upp på den vetenskapliga dagordningen.[1]

Anledningarna till varför det finns skäl att misstro empirisk forskning är flera. Det är allmänt accepterat att statistiskt signifikanta resultat är lättare att publicera, inte minst i topp-tidskrifter som till stor del sätter den vetenskapliga dagordningen. Exempelvis visar Brodeur m fl (2016) att det finns en lägre andel vetenskapliga resultat publicerade i tidskrifterna *American Economic Review*, *Journal of Political Economy* och *Quarterly Journal of Economics* med *p*-värden mellan 0,25 och 0,10 än vad man borde förvänta sig, medan det finns fler resultat med *p*-värden precis under 0,05.

Tyvärr finns det en självförstärkande effekt i denna *publikationsbias*. För en nationalekonom ligger det nära till hands att studera forskarens incitament. Om tidskrifter i huvudsak publicerar statistiskt signifikanta resultat så torde en rationell agent agera på ett sätt som gör det mer troligt att erhålla sådana resultat. Det finns flera mer eller mindre accepterade sätt att uppnå detta. Exempelvis kan en forskare välja att avbryta ett projekt om man efter en preliminär undersökning inte finner statistisk signifikans. Varför ska man som forskare lägga ner mycket tid och energi på ett projekt som i slutändan inte går att publicera (s k *file-drawer effect*)?

Som tillämpade forskare vet, finns det dessutom inte bara ett enda sätt man kan genomföra en statistisk undersökning på. I observationsstudier finns det ofta ett stort antal olika regressionsmodeller som är rimliga, där ingen modell klart framstår som bättre än någon annan. I en sådan situation har forskaren ett stort antal frihetsgrader att välja de resultat som leder till statistisk signifikans. Detta kan vara genom att direkt välja ut statistiskt signifikanta resultat (s k *p-hacking*) eller genom att man övertygar sig själv att de resultat som ser mest intressanta ut också kommer från den mest rimliga modellen. Att forskare svarar på dessa typer av incitament tyder resultaten i Brodeur m fl (2016) på, då överdensiteten av statistiskt signifikanta resultat framför allt verkar finnas i artiklar skrivna av forskare som inte har fast tjänst, för vilka incitamenten att publicera sig är som allra högst.

Det finns i dag en ökad medvetenhet om riskerna med de snedvridande effekter som beskrivs ovan och flera åtgärder har genomförts för att hantera problemet (för en översikt kring detta inom det nationalekonomiska fältet, se Christensen och Miguel 2018). Exempelvis rekommenderar numera många tidskrifter att man inte indikerar statistiskt signifikanta effekter med stjärnor i tabeller (detta är exempelvis numera fallet i *American Economic Review*). Dessutom krävs allt oftare att forskare förregistrerar studier innan

[1] Replikationsgraden inom nationalekonomisk forskning verkar, enligt en studie (Camerer m fl 2016) vara högre. Det ska dock noteras att denna studie enbart studerade ett relativt litet antal lab-experiment och det finns stor anledning att tro att problemen är större i studier baserade på observationsdata där forskarens frihetsgrader är större. För en utmärkt översikt och diskussion av dessa studier hänvisas läsaren även till Dreber och Johannesson (2019). Något som är intressant att notera är att forskare själva verkar bra på att förutsäga vilka resultat som kommer att replikera (se t ex Forsell m fl 2019). Detta faktum tyder på att forskare själva har mer information än vad som erhållits genom hypotestestande.

data har erhållits i fall där detta är möjligt och ibland kan även *peer-reviews* genomföras innan resultat erhållits (detta har exempelvis testats i *Journal of Development Economics*).

1. Statistiskt signifikanta effekter är överdrivna

Om bara statistiskt signifikanta resultat publiceras och forskare väljer att avsluta projekt som inte leder till signifikans är det lätt att se att detta kan leda till snedvridna resultat där effektstorlekarna överdrivs. Problemen associerade med publikationsbias är dock inte en kritik mot hypotestest i sig utan handlar snarare om hur hypotestest används. En vanligt förekommande kommentar bland statistiker är att problemen med hypotestestande handlar om hur metoden används, snarare än att det finns ett fundamentalt problem med metoden. Med bra statistisk undervisning tillsammans med incitament för forskare att inte överdriva resultat skulle problemen med hypotestest, enligt detta argument, försvinna.

Jag skulle dock vilja hävda att riktigt så enkelt är det inte. Även om vi utgår från att det inte finns någon file-drawer-effekt, "*p*-hacking" eller publikationsbias så kan det fortfarande finnas snedvridande effekter om man använder sig av statistisk signifikans som kriterium för att utvärdera en vetenskaplig hypotes. Problemet, som jag ser det, är att användandet av statistisk signifikans dikotomiserar något som inte bör dikotomiseras. Det är därför värdefullt att studera hur det ofta lärs ut att hypotestest ska användas i ett enkelt exempel.

Låt oss anta att vi vill studera den kausala effekten av en behandling, W, på ett utfall Y. Vi genomför ett randomiserat experiment där vissa individer slumpmässigt får behandlingen ($W = 1$) och andra får placebo ($W = 0$)[2]. Vårt estimat av den genomsnittliga behandlingseffekten, τ , är medelvärdesskillnaden mellan de två grupperna, $\hat{\tau} = \bar{Y}_1 - \bar{Y}_0$, där $\bar{Y}_1$ är medelvärdet av Y i behandlingsgruppen och $\bar{Y}_0$ är medelvärdet av Y i kontrollgruppen. Med hjälp av $\hat{\tau}$ genomför vi sedan hypotestestet

$$H_0: \tau = 0$$
$$H_a: \tau \neq 0.$$

Vi kan nu genomföra ett *t*-test för medelvärdesskillnaden mellan de två grupperna. Baserat på observerad data tar vi fram teststatistikan från vilket *p*-värdet erhålls. Vi har specificerat en beslutsregel som säger att om *p*-värdet är mindre än en viss gräns (t ex 0,05) så förkastar vi nollhypotesen, medan om vi erhåller ett *p*-värde över gränsen så förkastar vi inte nollhypotesen.[3] Det vill säga, det finns en dikotomi av slutsatser (förkastar/förkastar ej).

Idén att nollhypotesen bara kan förkastas eller inte förkastas stäm-

[2] Argumentet som presenteras här är på inget sätt begränsat till ett randomiserat experiment utan gäller i princip för vilken statistisk analys som helst.

[3] Ibland sägs även att nollhypotesen accepteras. Det ska dock inte tolkas som att nollhypotesen är sann, utan endast att det inte finns tillräcklig evidens för att förkasta den. För en tidig kritik av idén att det finns en nollhypotes som antingen förkastas eller accepteras, se Fisher (1955).

mer väl in med Poppers *falsifikationism*. Det vill säga, det är inte möjligt att bekräfta en vetenskaplig hypotes utan det går endast att förkasta den. Det är dock ytterst tveksamt om man kan se det statistiska hypotestestandet som att det följer idén om falsifikationism. Min erfarenhet är att det är vanligast att det är alternativhypotesen som är den vetenskapliga hypotesen som forskare vill studera. Det vill säga, den vetenskapliga hypotesen tar ofta formen ”behandlingen har en kausal effekt på Y” medan det är betydligt ovanligare med hypotesen ”behandlingen har *inte* en kausal effekt på Y”.[4]

Varifrån kommer då snedvridningen? Om vi inte förkastar nollhypotesen så förhåller vi oss skeptiska till att det finns en effekt och väljer att inte tolka storleken på effektskattningen, det är ju trots allt ett resultat som ofta skulle kunna observeras när det inte finns en effekt! Om vi i stället kan förkasta nollhypotesen så drar vi slutsatsen att det finns en kausal effekt av W på Y. Det blir då naturligt att ställa sig frågan hur stor effekten är. Om vi har en väntevärdesriktig estimator så kan det tyckas självklart att vi kommer att få en korrekt uppskattning av effekten. Tyvärr är dock detta inte fallet. Tvärtom så går det att visa att *statistiskt signifikanta effekter i genomsnitt är överestimat av den sanna effekten*. Hur kommer sig detta?

För att förstå detta så måste vi förstå vad väntevärdesriktighet betyder. Väntevärdesriktighet som koncept brukar definieras utifrån idén om återupprepad sampling.[5] Det vill säga, över fördelningen av alla tänkbara utfall som vi skulle kunna observera, samplingfördelningen, så gäller att det genomsnittliga värdet kommer att hamna vid det sanna värdet. Men om vi enbart väljer att tolka de estimat som är statistiskt signifikanta så kommer vi att bortse från de estimat i samplingfördelningen som ligger nära noll eftersom dessa inte kan vara signifikanta. Det faktum att vi inte tolkar alla möjliga estimat i samplingfördelningen är det som skapar snedvridningen.

Detta fenomen går att illustrera i en enkel Monte Carlo-simulering. Låt oss anta att Y genereras på följande sätt:

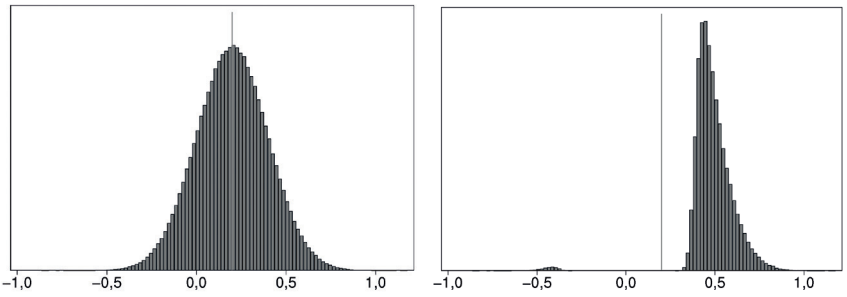
$$Y = \tau W + \varepsilon,$$

där ε följer en standardnormalfördelning. Det finns 100 individer i vårt experiment och vi låter 50 slumpmässigt få behandling, $W = 1$, medan 50 får tillhöra kontrollgruppen som inte får behandling, $W = 0$. $\tau = 0{,}2$ är den behandlingseffekt vi vill estimera och $\hat{\tau}$ är medelvärdesestimatorn definierad ovan. Vi använder ett vanligt t-test för att undersöka om det finns en statistiskt signifikant effekt av behandlingen.

Den vänstra grafen i figur 1 visar samplingfördelningen utifrån 100 000 oberoende stickprov. Vi vet att denna fördelning ska vara centrerad runt 0,2

[4] Det är också värt att notera att Popper själv i strikt mening inte ansåg det möjligt att avfärda en vetenskaplig hypotes på sannolikhetsmässig grund. Han menade i stället att enbart om en hypotes förbjuder vissa observationer kan den ses som en vetenskaplig hypotes. Han insåg dock att detta i praktiken inte var möjligt, varför han medgav att forskare kan behöva sätta upp sannolikhetsmässiga kriterier för när en hypotes ska förkastas. Se diskussionen på s 66 i Godfrey-Smith (2003).

[5] En alternativ tolkning är att betrakta stickprovet som fixt, men att stokastiken kommer ifrån randomiseringen till behandling. Båda dessa tolkningar leder dock till samma slutsats.



Figur 1
Samplingfördelning för alla estimat (vänster) och statistiskt signifikanta (höger)

Källa: Egna beräkningar.

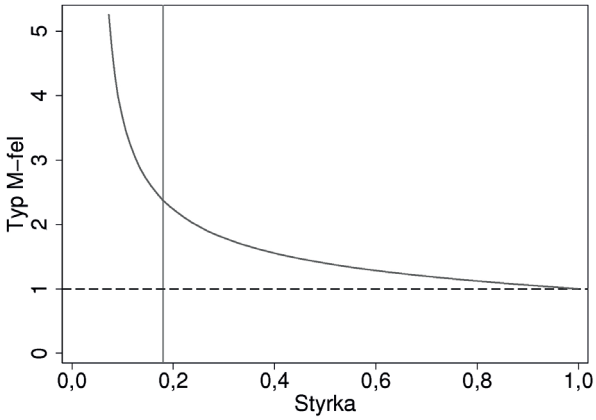
och följa en *t*-fördelning med 100 – 2 = 98 frihetsgrader (approximativt normalfördelad), något vi ser är fallet. Det intressanta är dock vad som händer i den högra grafen. Här ser vi fördelningen av enbart statistiskt signifikanta estimat (fem procent signifikansnivå). Dessa estimat är inte alls centrerade runt 0,2. I stället har fördelningen två delar. Den första är de få estimat som är statistiskt signifikant negativa. Detta sker endast i ett fåtal fall eftersom den sanna effekten är positiv. Den huvudsakliga massan ligger i stället för de positivt statistiskt signifikanta estimat som i huvudsak erhålls när $\hat{\tau}$ är större än 0,4.

Det genomsnittliga statistiskt signifikanta estimatet ligger på 0,488. Det vill säga, givet att ett estimat är statistiskt signifikant så kan vi förvänta oss att det är ett överestimat med en faktor 0,488/0,2=2,44. Gelman och Carlin (2014) beskriver detta fenomen som ett typ M-fel (*magnitude error*).[6]

Problemet kommer alltså ifrån dikotomiseringen som användandet av statistisk signifikans innebär. Genom att hantera statistiskt signifikanta resultat på ett annat sätt än icke-signifikanta resultat så snedvrids tolkningen. Varför blir snedvridningen så stor i detta fall? Svaret är att i enbart 16,7 procent av fallen var estimatet statistiskt signifikant. Styrkan (dvs sannolikheten att förkasta en falsk nollhypotes) i det statistiska testet var alltså långt under den som vanligtvis brukar rekommenderas på 80 procent. Man skulle därför kunna invända att jag har presenterat något av en halmgubbe: med så extremt låg styrka så är det inte konstigt att resultaten blir missvisande. Tyvärr finns det mycket som tyder på att detta är vanligt förekommande. Ioannidis m fl (2017) fann att medianstyrkan i nationalekonomiska studier låg på 18 procent och typiskt sätt överdrivs estimaten med en faktor två. I en tredjedel av studierna överdrevs estimaten med en faktor fyra eller mer!

Figur 2 illustrerar sambandet mellan typ M-felet och den statistiska styrkan. Vi ser att när styrkan är nära ett så överdrivs inte resultaten alls. Anledningen är helt enkelt att ingen selektion introduceras när vi enbart tolkar de signifikanta resultaten; alla resultat är ju signifikanta! När styrkan

6 Tekniskt sett är typ M-felet absolutvärdet av statistiskt signifikanta estimat delat på det sanna värdet. Eftersom så få statistiskt signifikanta estimat är negativa är det i detta fall i stort sett samma sak. Typ M-felet är här 0,497/0,2 = 2,48 där 0,497 är det genomsnittliga absolutvärdet av statistiskt signifikanta estimat.



Figur 2
Samband mellan typ M-fel och statistisk styrka

Källa: Egna beräkningar.

minskar så ökar typ M-felet först långsamt för att sedan växa snabbt när styrkan är låg. Den vertikala linjen indikerar en styrka på 18 procent i enlighet med Ioannidis m fl (2017). Med sådan styrka är typ M-felet knappt 2,4 vilket innebär att magnituden på en statistiskt signifikant effekt är i genomsnitt 2,4 gånger större än den sanna effekten.

Ett förslag som framkommit för att hantera problemet med ”*p*-hacking” är att sänka konventionen för vad som räknas som statistiskt signifikanta effekter från 0,05 till 0,005 (Benjamin m fl 2018). För en given stickprovsstorlek innebär en sådan förändring att styrkan på det statistiska testet går ner vilket gör att typ M-felet ökar ytterligare. Även om en mer konservativ signifikansnivå innebär färre falska positiva test, så kommer överdriften av de resultat som faktiskt är statistiskt signifikanta att bli ännu större. Det är därför inte uppenbart att en mer konservativ signifikansnivå leder till mindre snedvridning av resultat. Benjamin m fl (2018) föreslår därför att man i experimentsituationer bör öka stickprovsstorleken så att styrkan hålls konstant. För observationsstudier, å andra sidan, är stickprovsstorleken ofta fix vilket gör att det inte är möjligt att behålla styrkan på det statistiska testet.

Läsaren kanske vid detta ögonblick invänder att det är missvisande att säga att statistiskt insignifikanta effekter inte tolkas alls och att vi faktiskt spenderar tid med att tolka storleken även på insignifikanta effekter. Min erfarenhet är dock att det trots allt är relativt vanligt att forskare säger att det ”inte finns en effekt” när ett statistiskt insignifikant resultat erhålls. Å andra sidan: om man faktiskt tolkar koefficienten oavsett om resultatet är statistiskt signifikant eller inte, varför använder vi då ett hypotestest överhuvudtaget? Antingen gör vi en uppdelning baserad på statistisk signifikans, i viket fall vi, enligt argumentet ovan, drar inkorrekta slutsatser, eller så gör vi inte en sådan uppdelning. Men om ingen uppdelning görs har hypotestestet spelat ut sin roll och fyller inte längre någon funktion.

Argumentet ovan bygger på att vi enbart har en enda nollhypotes som

vi antingen förkastar eller inte förkastar. Det finns dock inget som säger att vi enbart kan förhålla oss till en hypotes. Ett bättre alternativ är att vi tänker oss att det finns en kontinuerlig uppsättning hypoteser och vi kan, exempelvis, studera vilka hypoteser som inte går att förkasta. För klassiska Neyman-Pearson-test så är detta definitionen av ett *konfidensintervall*. Med konfidensintervall skapas inte någon dikotomisering utan intervallet ger oss i stället en uppsättning rimliga värden för den effekt vi studerar. På så sätt undviker vi den snedvridning som uppkommer med hypotestestet.[7] En dikotomisering ska dock inte ersättas med en annan: Parametervärden utanför konfidensintervallet är inte fullständigt inkompatibla med data, bara mindre kompatibla än parametervärden inom konfidensintervallet. Dessutom är inte alla parametervärden inom intervallet lika kompatibla med data; de värden som ligger närmare punktskattningen är mer kompatibla än de som ligger längre ifrån.

2. Är *p*-värden jämförbara mellan studier?

Ovanstående diskussion behandlar hypotestestet och uppdelningen mellan signifikanta och icke-signifikanta resultat. Det är dock viktigt att påpeka att även om hypotestestet får mindre inflytande, eller rent av avskaffas, så kan det ändå vara värdefullt att använda sig av *p*-värden.[8]

Det är dock viktigt hur man tolkar ett *p*-värde. Mycket har skrivits om hur detta värde misstolkas. Det kanske vanligaste misstaget är att tolka *p*-värdet som sannolikheten att nollhypotesen är sann, något som definitivt inte stämmer. Men även när *p*-värdet tolkas på ett tekniskt korrekt sätt så kan det leda till en övertolkning av i vilken utsträckning *p*-värdet ger information om nollhypotesen.

Följande exempel får användas för att förtydliga hur denna övertolkning uppstår: Låt oss säga att vi är intresserade av effekten av en variabel X_1 på Y där vi betingar på X_2, ..., X_k. Vi sätter upp följande regressionsmodell:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_K X_K + u.$$

Om vi nu vill ta fram *p*-värdet för ett dubbelsidigt test för hypotesen $\beta_1 = 0$ så brukar vi lära att ut att:

> Under antagandena att ovanstående modell är korrekt specificerad, att $E(u|X_1, X_2, \ldots, X_K) = 0$, det inte finns någon perfekt kollinearitet och att *u* är oberoende fördelad med konstant varians så kommer *p*-värdet ge oss sannolikheten att, över återupprepad sampling, observera ett absolutvärde på β_1 större än det värde vi observerade givet att $\beta_1 = 0$.

[7] Det är dock viktigt att hålla tungan rätt i mun när vi tolkar ett konfidensintervall. Strängt taget kan ett konfidensintervall inte tolkas som sannolikheten att parametern ligger i intervallet, utan det är ett intervall som över återupprepad sampling kommer att täcka parametern med en viss sannolikhet. För en diskussion om hur felaktig tolkning av konfidensintervall leder till problem – och hur Bayesiansk analys kan hantera dessa problem – se Morey m fl (2016).

[8] Det finns dock alternativ till *p*-värden som kan vara enklare att tolka, såsom Shannons *S*-värde, se t ex diskussionen i Greenland (2019).

Inte minst brukar vi som är lärare i ekonometri och statistik vara bra på att lära studenter att kunna lista upp alla antaganden som ”måste” göras. Det är strikt taget inget fel på denna beskrivning av *p*-värdet. Men i sättet att skriva har vi gjort en uppdelning mellan antaganden och hypotes, där vi säger att vi testar hypotesen givet antagandena. Rent matematiskt finns det dock ingen sådan uppdelning. Vi skulle lika gärna kunna säga att:

> Under antagandena att ovanstående modell är korrekt, att $E(u|X_1, X_2, \ldots, X_K) = 0$, det inte finns någon perfekt kollinearitet och att $\beta_1 = 0$ så kommer *p*-värdet ge oss sannolikheten att, över återupprepad sampling, observera ett absolutvärde på $\hat{\beta}_1$ större än det värde vi observerade givet att *u* är oberoende fördelad med konstant varians.

Det vill säga, vi kan lika gärna säga att vi testar om *u* är oberoende fördelad med konstant varians under antagandet att $\beta_1 = 0$ som att säga att vi testar huruvida $\beta_1 = 0$ under antagandet att *u* är oberoende fördelad med konstant varians. Uppdelningen mellan antaganden och hypotes är alltså något som vi lägger in som tolkning, snarare än något som gäller matematiskt (se t ex s 13 i Greenland 2017). Notera vad som skrivs i ASA:s officiella uttalande från 2016, ”ASA Statement on Statistical Significance and P-Values”, där de skriver att

> A *p*-value provides one approach to summarizing the incompatibility between a particular set of data and a proposed model for the data. (Wasserstein och Lazar 2016, s 131)

Ett lågt *p*-värde innebär alltså att modellen inte passar data. Men *p*-värdet kan inte svara på *vad* i modellen som inte passar data.

Att man inte kan bortse från antagandena i en regressionsmodell är dock knappast något som den empiriska nationalekonomiska forskaren behöver höra. Ett seminarium i nationalekonomi där ett empiriskt papper presenteras är ofta till stor del en enda diskussion om huruvida exogenitetsantagandet är uppfyllt. Men, i alla fall för min egen del, tycker jag att synsättet att det egentligen inte finns någon uppdelning mellan hypotes och antagande är hjälpsamt och tydliggör att det finns en begränsning av vilken mängd information som finns i *p*-värdet.

Med detta vill jag också ha sagt att *p*-värden inte nödvändigtvis är jämförbara mellan studier. I ett randomiserat experiment så är det möjligt att med, exempelvis, Fishers exakta test helt antagandefritt få fram ett *p*-värde som ger sannolikheten att få ett minst så extremt utfall som det observerade utfallet givet att det inte finns några behandlingseffekter. I ett sådant fall blir tolkningen av *p*-värdet relativt rättfram. I en studie där ett naturligt experiment analyseras å andra sidan så bygger analysen, även när exogenitetsantagandet är uppfyllt, ofta på asymptotiska och funktionsformsmässiga argument. Ett *p*-värde på 0,05 kan vara relativt stark evidens mot nollhypotesen i det förstnämnda fallet medan ett *p*-värde på 0,05 i det senare fallet troligen innebär svagare evidens mot nollhypotesen.

3. Vad ska man då göra?

Som Wasserstein m fl (2019) skriver så räcker det dock inte att ta fram pekpinnen och säga vad man inte ska göra, utan det är också viktigt att säga vad det är man faktiskt *ska* göra. Det viktigaste är enligt min mening att *acceptera osäkerhet*. Osäkerhet kommer dels av det faktum att vi har den statistiska osäkerheten, något som vi matematiskt kan beskriva med, exempelvis, *p*-värdet. Men det kommer också från *modellosäkerhet*. Det vill säga vi vet inte nödvändigtvis vad den sanna modellen är (eller vi kan inte estimera den) och vi måste ofta använda en förenklad modell där ytterligare en felkälla introduceras, något som *p*-värdet inte fångar upp.

Den andra slutsatsen jag drar är att vi inte bör dikotomisera slutsatser på det sätt som vi gör när vi förkastar eller inte förkastar en hypotes. Detta leder inte bara till incitament för forskare att på olika sätt påverka resultaten så att de framstår som mer intressanta, utan som vi såg ovan skapar själva dikotomiseringen i sig *bias*. Bättre är att i stället diskutera intervall av rimliga värden på parametrar av intresse.

Ibland måste man dock fatta beslut baserat på statistisk analys. Exempelvis ska beslut fattas om ett läkemedel ska introduceras på marknaden utifrån kliniska försök och Riksbanken beslutar om ränteförändringar till viss del baserat på ekonometrisk modellering. Det finns dock metoder inom *statistisk beslutsteori* som är utformade för att hantera just detta. Något som är tilltalande för en nationalekonom är att man där explicit modellerar konsekvenserna av ett beslut med hjälp av en välfärdsfunktion. För en introduktion till denna litteratur hänvisas den intresserade läsaren till Manski (2019). För den mesta forskningen behöver dock inget beslut fattas och det finns därför starka skäl att avstå från användandet av statistisk signifikans.

REFERENSER

Amrhein, V, S Greenland och B McShane (2019), ”Retire Statistical Significance”, *Nature*, vol 567, 7748, s 305–307.

Benjamin, D J m fl (2018), ”Redefine Statistical Significance”, *Nature Human Behavior*, vol 2, s 6–10.

Brodeur, A, M Lé, M Sangnier och Y Zylberberg (2016), ”Star Wars: The Empirics Strike Back”, *American Economic Journal: Applied Economics*, vol 8, s 1–32.

Camerer, C F m fl (2016), ”Evaluating Replicability of Laboratory Experiments in Economics”, *Science*, vol 351, s 1433–1436.

Christensen, G och E Miguel (2018), ”Transparency, Reproducibility, and the Credibility of Economics Research”, *Journal of Economic Literature*, vol 56, s 920–980.

Dreber, A och M Johannesson (2019), ”Statistical Significance and the Replication Crisis in the Social Sciences”, *Oxford Research Encyclopedia of Economics and Finance*, Oxford University Press, Oxford.

Fisher, R A (1955), ”Statistical Methods and Scientific Induction”, *Journal of the Royal Statistical Society*, *Series B (Methodological)*, vol 17, s 69–78.

Forsell, E, m fl (2019), ”Predicting Replication Outcomes in the Many Labs 2 Study”, *Journal of Economic Psychology*, vol 75(A), 102117.

Gelman, A och J Carlin (2014), ”Beyond Power Calculations Assessing Type S (Sign) and Type M (Magnitude) Errors”, *Perspectives on Psychological Science*, vol 9, s 641–651.

Godfrey-Smith, P (2003), *Theory and Reality: An Introduction to the Philosophy of Science*, University of Chicago Press, Chicago.

Greenland, S (2017), ”For and against Methodologies: Some Perspectives on Recent Causal and Statistical Inference Debates”, *European Journal of Epidemiology*, vol 32, s 3–20.

Greenland, S (2019), ”Valid P-Values Behave Exactly as They Should: Some Misleading Criticisms of P-Values and Their Resolution with S-Values”, *The American Statistician*, vol 73, sup 1, s 106–114.

Ioannidis, J P A (2005), ”Why Most Published Research Findings Are False”, *PLOS Med*, vol 2, nr 8, e124.

Ioannidis, J P A, T D Stanley och H Doucouliagos (2017), ”The Power of Bias in Economics Research”, *Economic Journal*, vol 127, s F236–F265.

Manski, C F (2019), ”Treatment Choice with Trial Data: Statistical Decision Theory Should Supplant Hypothesis Testing”, *The American Statistician*, vol 73, sup 1, s 296–304.

Morey, R D, R Hoekstra, J N Rouder, M D Lee och E-J Wagenmakers (2016), ”The Fallacy of Placing Confidence in Confidence Intervals”, *Psychonomic Bulletin & Review*, vol 23, s 103–123.

Open Science Collaboration (2015), ”Estimating the Reproducibility of Psychological Science”, *Science*, vol 349, nr 6251.

Wasserstein, R L och N A Lazar (2016), ”The ASA Statement on *p*-values: Context, Process, and Purpose”, *The American Statistician*, vol 70, s 129–133.

Wasserstein, R L, A L Schirm och N A Lazar (2019), ”Moving to a World Beyond ’$p < 0.05$’”, *The American Statistician*, vol 73, sup 1, s 1–19.