Om missbruket av statistik

Olle Häggström

Stephen T. Ziliak och Deirdre McCloskey (2008) *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives*, The University of Michigan Press, Ann Arbor, MI.

Att känna igen när ett rovdjur lurar i busken, och mer generellt att observera och dra riktiga slutsatser om omvärlden, hade för stenåldersmänniskan ett stort överlevnadsvärde. Det är därför knappast överraskande att evolutionen försett oss med långtgående kognitiva förmågor av det slaget. Men evolutionen är av flera skäl inte någon perfekt optimeringsalgoritm, och de system vi begåvats med har i själva verket en rad konstruktionsfel – buggar – vilka på senare tid rönt stort intresse bland psykologer och kognitionsforskare. En sådan bugg är vår tendens att tycka oss se mönster i vad som egentligen är meningslöst brus. En annan är hur vi systematiskt överskattar säkerheten i våra slutsatser; det finns studier som visar hur vi i vissa slags situationer har fel i cirka 40% av de slutsatser vi säger oss vara 98% säkra på.

Vetenskapen kan, om vi så vill, ses som ett organiserat försök att kringgå dessa buggar i sökandet efter tillförlitlig information om hur världen är beskaffad. Ett oundgängligt redskap i detta sammanhang är de matematisk-statistiska metoder som utvecklats för att hjälpa forskare se skillnad mellan brus och meningsfulla mönster, och att kvantifiera hur säkra våra slutsatser är. Dessa metoder har under 1900-talet kommit att alltmer genomsyra vetenskapen, så till de grad att vi idag sällan ser en forskare som pekar på sina data (det kan t.ex. handla om långtidstrender i det arktiska snötäcket, eller om hur effektiv den nya blodtryckssänkande medicinen är jämfört med placebo) och torgför sina slutsatser på basis av magkänsla – ”här ser vi ju en tydlig effekt!” – utan att backa upp denna med en strikt statistisk analys.

Det kan knappast råda någon tvivel om att denna utveckling bidragit starkt till att göra vetenskapen mer tillförlitlig. De båda amerikanska nationalekonomerna Stephen Ziliak och Deirdre McCloskey hävdar emellertid i sin nya bok *The Cult of Statistical Significance* att bruket av statistiska metoder på många håll har övergått i ett ritualiserande som kommit att bli mer av ett hinder än ett hjälpmedel för vetenskapens framåtskridande.

För att förstå deras budskap behöver vi lite statistisk terminologi. En forskare formulerar en *nollhypotes* (t.ex. att det arktiska snötäckets utbredning inte har någon trend vare sig nedåt eller uppåt, eller att den nya medicinen varken är bättre eller sämre än placebo). Med hjälp av ett *signifikanstest* försöker hon påvisa att nollhypotesen inte stämmer – ett förfarande i linje med Poppers falsifikationism. Den statistiska analys hon därvid genomför resulterar i ett *p-värde*, som är ett mått på hur osannolikt det vore att få minst så extrema data som hon faktiskt fick, om nollhypotesen var sann. Om p-värdet blir lägre än en på förhand föreskriven gräns (oftast 0,05 eller 0,01), sägs utfallet vara *statistiskt signifikant*. Ett statistiskt signifikant resultat är långt ifrån liktydigt med ett definitivt motbevis av nollhypotesen, men kan ändå ofta anses tala emot densamma.

Antag nu att en ny blodtryckssänkande medicin skall utprovas, och att det faktiska förhållandet är att den har en positiv effekt jämfört med placebo, men att skillnaden är så liten att den inte har någon praktisk betydelse för patientens hälsa eller välbefinnande. Om försöket omfattar tillräckligt många patienter, så kommer likväl denna lilla skillnad med stor säkerhet att detekteras, och statistisk signifikans erhållas. Lärdomen av detta är att det i en medicinsk studie inte räcker att uppnå statistisk signifikans för att ge stöd åt det nya läkemedlet – vad som också skall till är att den observerade skillnaden är *medicinskt signifikant*. På samma sätt gäller för statistiska studier inom nationalekonomi

(eller psykologi, geologi, etc) att det inte räcker att se om statistisk signifikans erhålles – också den eventuella nationalekonomiska (respektive psykologiska, geologiska, etc) signifikansen måste beaktas. En av Ziliaks och McCloskeys huvudpoänger är att påtala det utbredda fenomenet att forskare är så fixerade vid att erhålla statistiska signifikans att de glömmer att alls befatta sig med frågan om de skillnader de observerat är stora nog att ha någon substantiell ämnesmässig signifikans. Detta fenomen kallas i boken *sizeless science*, och det kan också ses som en variant av vad jag i andra sammanhang kallat statistiska signifikanssjukan.

Ziliak och McCloskey diskuterar och exemplifierar bruket av sizeless science inom bl.a. medicin och psykologi, men främsta udden riktar de naturligt nog mot sina kollegor inom nationalekonomin. I en imponerande litteraturstudie har de gått igenom samtliga de 369 artiklar som under 1980- och 1990-talet publicerats i den prestigeladdade tidskriften American Economic Review och som inbegriper statistisk analys. Hela 70% av artiklarna under 80-talet visade prov på sizeless science, och vad värre är, motsvarande siffra för 90-talet är 79%. En rad andra, men oftast besläktade, slags missbruk av statistisk metodik studeras också. Resultaten är nästan genomgående nedslående.

Statistisk metodik synes kanske för de flesta vara ett alltför torrt ämne att frivilligt läsa en hel bok om, men Ziliaks och McCloskeys utpräglat polemiska avsikt ger liv och driv åt framställningen. Ibland går emellertid polemiken överstyr, som när de (s 240) frågar sig: ”Om nu signifikanstest är så idiotiskt som vi och dess övriga kritiker länge hävdat, hur kan det då ha överlevt?”. Låt gå för att det ensidiga användandet av signifikanstest är ett oskick – på denna punkt ger jag dem gärna rätt. Men den statistiska signifikansen är ett avgörande komplement till medicinsk/nationalekonomisk/etc dito, ty utan den har vi inte grund för att säga att de mönster vi tycker oss se är verkliga mönster och inte bara brus.

Till bokens läsvärdhet bidrar också ett längre historiskt avsnitt om signifikanstestets uppkomst och mer allmänt om den matematiska statistikens utveckling under 1900-talets tidigare hälft. Även här är framställningen polemisk, och något utrymme för annat än svart och vitt ges inte i deras bild av William Gossett som hjälte och Ronald Fisher (allmänt ansedd som 1900-talets viktigaste statistikteoretiker) som skurk. Det är med illa dold vällust författarna återger (s 222) vad Robert Oppenheimer lär ha sagt på tal om Fishers ankomst som gästforskare i Berkeley 1936: ”Jag tog en enda titt på Fisher och beslöt mig för att honom ville jag inte träffa”.

Vad tycker då Ziliak och McCloskey bör göras åt den rådande situationen? De är tydliga med att förespråka en högre grad av pluralism bland statistiska metoder. Viktigast enligt min egen uppfattning är att höja kunskapsnivån i statistik bland forskare och blivande forskare i empiriska ämnen. De missförhållanden som Ziliak och McCloskey påtalar beror i hög grad på att forskarnas kännedom om statistik oftast inskränker sig till ren procedurkunskap. Denna behöver kompletteras med kunskaper om den matematik och den statistikfilosofi som motiverar metoderna. Med en sådan höjning av kunskapsnivån skulle sizeless science-bruket automatiskt upphöra, och den blir ännu viktigare om den av Ziliak och McCloskey förordade pluralismen slår igenom, ty ju fler vapen en blind jägare har att tillgå desto större är risken att han i sin förvirring skjuter sig i foten.