Att undersöka kontroversiella fenomen

Jesper Jerkert

Kontroversiella fenomen avfärdas ibland på teoretiska grunder. Telepati, homeopati och astrologi sägs vara omöjliga eller i varje fall ytterst osannolika, givet vad vi sedan tidigare vet om världen och människan. Denna typ av argumentation är ofta nyttig och intressant, men den kan inte ge slutgiltiga besked. Det är ju fullt möjligt att ett till synes osannolikt fenomen ändå är äkta. Om så är fallet kan det dessutom öppna nya forskningsfält och tvinga fram nya teorier, och därmed öka vår förståelse för hur världen fungerar. Vetenskapen har tidigare tagit flera sådana vändningar. Även en djupt skeptisk person bör därför, i princip, gå med på att det säkraste sättet att ta reda på om kontroversiella påståenden är sanna eller ej, är att empiriskt undersöka dem.[1]

Det finns förstås en del påståenden som är omöjliga att undersöka empiriskt. Det finns t.ex. inget sätt att empiriskt avgöra om en person som påstår sig stå i direktkontakt med jultomten talar sanning. Men hur ska man bete sig när det är görligt? Tyvärr kan ingen allmän anvisning följas i alla sammanhang. Olika påståenden kan kräva vitt skilda slags undersökningar. Här ska jag ändå försöka peka på några metodfrågor som man ofta behöver tänka på, särskilt när det gäller kontroversiella fenomen. Råden är ofta tillämpbara även på undersökningar av okontroversiella fenomen. Texten behandlar frågor som i de flesta fall är gemensamma för många vetenskapsgrenar. Extra vikt har dock lagts vid undersökningar inom medicinen.

Före undersökningen

Först några varningsord. Berättelser om fantastiska händelser cirkulerar ständigt. Innan man sätter igång med mer avancerade undersökningar bör man tillämpa vanlig källkritik på påståendena. Varifrån kommer påståendet, och hur trovärdig är källan? Existerar personerna som omtalas? Har händelsen inträffat? Existerar platsen där den sägs ha inträffat? Om man inte gör sådana elementära källkontroller riskerar man att hamna i kniviga undersökningssituationer helt i onödan. Antag att en av dina vänner påstår sig ha upptäckt varför så många fartyg och flygplan försvinner i Bermuda-triangeln. Orsaken är att lätta gaser strömmar upp från havsbottnen. När gasen når havsytan minskar bärkraften, och fartyg som råkar passera där sjunker omedelbart. Om gasen fortsätter upp i luften drabbas passerande flygplan av samma öde. Det låter som en spännande hypotes, och din vän uppmanar dig att undersöka den närmare. Hur ska du gå tillväga? Skaffa en bassäng och experimentera med någon lätt gas och modellbåtar?

Nej, först bör du kontrollera påståendena om Bermuda-triangeln. Stämmer det verkligen att båtar och flygplan försvunnit under mystiska omständigheter i triangeln? Svaret är nej. Myten om Bermuda-triangeln, som går ut på att ett oförklarligt stort antal fartyg och flygplan försvunnit i ett havsområde i Atlanten, fick stor spridning genom Charles Berlitz bok *The Bermuda Triangle* 1974 (på svenska *Dödens triangel*). Redan 1975 utkom dock en bok av Larry Kusche, *The Bermuda Triangle Mystery – Solved*, som systematiskt visar att många av Berlitz påståenden är felaktiga. Olycksfrekvensen i Bermudatriangeln är inte större än i andra sjöområden med lika stor trafik. Det finns inget mysterium att förklara. Därmed blir en undersökning av gashypotesen överflödig.

Exemplet Bermuda-triangeln illustrerar också en allmän princip som är viktig att ha i bakhuvudet: Bevisbördan ligger alltid på den som kommer med påståenden som strider mot tidigare erfarenheter och kunskaper. En undersökning bör därför gå ut på att ta reda på om det fantastiska påståendet är sant eller ej. Detta måste vara huvudfrågan. Huru-

vida en teori som sägs förklara de fantastiska påståendena är sann eller ej är en bifråga. Inte sällan vänder företrädare för udda teorier på bevisbördan, och hävdar att det åligger skeptikerna att visa att teorin är felaktig.

Ytterligare ett exempel erbjuder s.k. psykisk kirurgi (engelska *psychical surgery*), vilket kanske hellre borde kallas ”mirakelkirurgi” på svenska. Vissa personer påstår sig kunna utföra avancerade operationer inuti kroppen på t.ex. cancerpatienter utan att använda några kirurgiska instrument. Patienten ligger ned på en brits. Mirakelkirurgen tränger in i buken med sina bara händer, fiskar upp diverse köttslamsor och avlägsnar dem. Det hela sker under ganska stor blodspillan, men till slut torkar mirakelkirurgen bort blodet och – simsalabim! – buken är hel igen, utan något tecken på att någonsin ha varit öppnad. Vissa desperata patienter har betalat stora summor för att få sådana operationer utförda, främst i Brasilien och på Filippinerna.

Är sådana kirurgiska ingrepp verkligen möjliga? Flera dylika operationer finns fångade på bild och videoband. Finns någon naturlig förklaring? Ja, det kan röra sig om skickligt utförda illusionsnummer. Mirakelkirurgen böjer sina fingrar så att det ser ut som att han tränger in i buken. Blodet förs dit med hjälp av rödfärgsampuller dolda i händerna. När kirurgen hämtar handdukar för att torka upp, smugglas ännu mer rödfärg dit. Köttslamsorna smugglas dit på samma sätt. (Vid något tillfälle har en köttbit omhändertagits för analys efter en mirakeloperation. Den visade sig komma från kyckling.) Trollkonstnären James Randi har offentligt demonstrerat att mirakelkirurgi kan utföras som illusionsnummer.[2] Försvarare av mirakelkirurgi har erkänt att Randis föreställning liknar mirakelkirurgernas, men de tvivlar ändå på att jämförelsen är rättvis: Är det verkligen möjligt att mirakelkirurgerna med illusionistmetoder skulle kunna föra dit *så* mycket blod och kött som syns under deras operationer?

Försvararnas invändning missar att frågan om trollerimetoder inte är den viktigaste. Kritikernas starkaste argument är att mirakelkirurgerna aldrig har låtit sig närmare undersökas och övervakas. Bevisbördan ligger ju hos mirakelkirurgerna, inte hos kritikerna.

Blindhet

Låt oss nu anta att en undersökning verkligen ska genomföras. En person påstår sig ha en paranormal förmåga som yttrar sig i att han kan berätta vilket kort som ligger överst i en uppochnedvänd blandad kortlek. Vi blandar en kortlek ordentligt och lägger den på bordet, och ber personen att tänka på det kort han tror ligger överst. Vi vänder sedan upp det översta kortet och frågar honom om det är rätt kort. Vore detta en bra undersökning? Nej, givetvis inte. Vi måste förstås fråga försökspersonen vilket kort han tror ligger överst *innan* vi vänder på kortet. Mer allmänt uttryckt så får inte personen som testas känna till de rätta svaren i förväg, och han ska inte heller på något annat sätt kunna ta reda på svaren annat än med den förmåga som testas (i detta fall en paranormal förmåga). Om detta krav är uppfyllt sägs undersökningen vara *blind, blindad* eller *maskerad*.

Att undersökningar ofta bör blindas verkar självklart, men det är inte alltid så lätt att försäkra sig om fullständig blindhet. Låt oss fortsätta med kortexemplet. Antag att försöksledaren tittar på det översta kortet men inte visar det för försökspersonen. Denne får sedan tala om vilket kort han tror att det är, varefter försöksledaren visar upp kortet så att båda kan se om svaret var rätt eller fel. Detta förfarande verkar vara blint – det är ju bara försöksledaren som känner till svaret, inte försökspersonen. Men kan man vara alldeles säker? Finns det månne en spegel eller ett fönster bakom försöksledaren? Eller kan kortet speglas i försöksledarens ögon eller glasögon? Eller är någon annan person

närvarande och kan se kortet samtidigt som försöksledaren och sedan omärkligt meddela det rätta svaret till försökspersonen?

Att låta försöksledaren titta på kortet innan försökspersonen avgivit sitt svar öppnar också dörren för subtila signaler som inte behöver ha något med fusk att göra. Exempelvis är det välkänt att en person som koncentrerar sig på ett ord eller begrepp ofta tenderar att nästan omärkligt och helt omedvetet röra läpparna och struphuvudet som om han skulle uttala ordet. Sådana här rörelser, som i sig själva är små och omedvetna men som ändå är styrda av en medveten idé, kallas ideomotoriska.[3] I vårt exempel kan man tänka sig att försöksledarens omedvetna läpprörelser kan hjälpa försökspersonen, och detta även utan att någon av dem är medveten om det. Vissa ledtrådar kan alltså både skapas och varseblivas omedvetet, vilket gör dem mycket svåra att upptäcka. I vårt exempel kan vi eliminera alla ovannämnda bekymmer genom att inte låta någon titta på kortet förrän försökspersonen har gjort sin bedömning. Vidare krävs förstås att alla kortbaksidor är exakt likadana.

Behovet av blind undersökningsmetod brukar särskilt framhållas inom medicinen. Om man utvärderar ett nytt läkemedels eventuella effekt så innebär blindhet att inga av de deltagande försökspersonerna vet huruvida de får läkemedlet eller placebo (eller kanske ett konkurrerande läkemedel).

Antag att vi istället vill testa en helare som säger sig kunna avgöra om människor har levercancer enbart genom att hålla handen ovanför buken och känna ”vibrationerna”. Blindhet i detta fall skulle innebära att helaren inte får reda på vilka testpersoner som har levercancer i förväg. Men det räcker inte. Helaren ska inte kunna avgöra det på något annat sätt heller, annat än med hjälp av ”vibrationerna”. Det innebär sannolikt att personerna måste täckas över, eftersom en långt gången cancersvulst kan vara så stor att den syns genom buken. Vidare kan de sjuka patienterna ha genomgått strål- eller cellgiftsbehandling och t.ex. ha tappat håret. Mer subtila ledtrådar är att patienter med leverstörningar kan få en gulaktig hy och att de kan lukta på ett speciellt sätt. I en väl upplagd undersökning måste alla dessa ledtrådar elimineras, annars är undersökningen inte blindad. Dessutom får helaren naturligtvis inte prata med personerna. I detta liksom många andra fall är det inte alldeles trivialt att uppnå fullständig blindhet.

Undersökningar kan vara mer eller mindre blinda. Om helaren i förväg får veta vilka personer som har levercancer, så blir studien givetvis helt oblind. I fallet med läpprörelserna är det däremot inte säkert, kanske inte ens troligt, att den testade personen kan se att försöksledaren tänker på ”spader dam”. Men om han åtminstone ser att läpprörelserna *inte* stämmer för t.ex. klöver, så innebär det ändå en viss hjälp. Den testade personen kan då utesluta alla klöver ur sina möjliga svar. På så sätt ökar hans chans att säga rätt, även om han inte kunnat läsa av exakt rätt kort av läpprörelserna. Så länge möjligheten att läsa på läpparna kvarstår kan därför undersökningen inte sägas vara tillfredsställande blindad.

Dubbelblindhet

I exemplet med korten ovan framhölls att några möjliga ledtrådar var försöksledarens läpprörelser eller att en annan person såg kortet och meddelade det rätta svaret till den som skulle testas. Det gemensamma för dessa två felkällor är att andra personer än den som ska testas är närvarande och vet rätt svar. Eftersom det är mycket vanligt att åtminstone någon annan än försökspersonen är närvarande, så finns ett speciellt begrepp som täcker in kravet att även sådana personer ska vara okunniga om det rätta svaret: *dubbelblindhet*.

I exemplet med korten blir undersökningen alltså dubbelblind om vi ser till att ingen vet vilket det översta kortet är förrän den testade personen avgivit sitt svar. Vi förutsätter då att den person som blandade kortleken (om det överhuvudtaget gjorts manuellt) inte heller tagit reda på vilket kort som hamnat överst; för säkerhets skull bör blandaren förbjudas att närvara. I läkemedelsstudier betyder blindhet, såsom nämnts, att försökspersonerna är okunniga om huruvida de får läkemedlet, medan dubbelblindhet betyder att även de som behandlar försökspersonerna och de som bedömer deras hälsotillstånd är okunniga om detta.

I vissa situationer blir det meningslöst att tala om dubbelblindhet, t.ex. när det inte finns två olika grupper av personer som ska hållas i okunnighet, utan bara en. Om man testar egenskaperna hos en kemisk substans, så kan man inte ”blinda” substansen. Däremot kan man blinda de personer som kan påverka resultatet. Ett instruktivt exempel på bristande blindning erbjuder Jacques Benvenistes homeopatiundersökning.

Jacques Benveniste var en mycket aktad forskare när han tillsammans med medarbetare 1988 publicerade en artikel i *Nature*.[4] Där rapporterades att när mänskliga basofiler, ett slags blodcell med antikroppar av typ immunoglobulin E (IgE), utsätts för antikroppar mot IgE, så ändrar de färg. Det märkliga var att detta skedde även om antikropparna tillsattes i extremt utspädd form, så utspädd att inte en enda antikroppsmolekyl bör ha funnits i lösningen. Resultatet stred mot etablerad kunskap, men skulle kunna ge vetenskapligt stöd åt homeopatin, en alternativmedicinsk skola som använder extrema utspädningar. *Nature* publicerade artikeln på villkor att man fick besöka Benvenistes laboratorium och bevittna upprepningar av experimentet.

En tremannakommitté besökte laboratoriet och publicerade snabbt en rapport.[5] Vid upprepningen blev resultatet fullständigt negativt. Kommittén fann flera tänkbara felkällor i det tidigare experimentet. Mest intressant för oss är att räkningen av antalet färgförändrade basofiler gjordes manuellt (i mikroskop). Personen som räknade var i många fall samma person som gjort spädningen. Personen var fullt medveten om vad olika räkningsresultat skulle innebära, vilket strider mot blindningskravet. Benvenistes studie är således inte trovärdig. För säkerhets skull gjorde en helt annan forskargrupp flera år senare en upprepningsstudie som utföll negativt.[6]

Tyvärr går det inte alltid att ordna strikt dubbelblindhet i den meningen att alla närvarande är okunniga om relevanta fakta. Låt oss återknyta till helaren med påstådd förmåga att känna ”vibrationer” av levercancer. Ett test av förmågan måste gå ut på att helaren ska peka ut sjuka personer i en grupp med både sjuka och friska. Men det kan vara svårt att finna levercancersjuka personer som inte själva vet att de har åkomman. (Möjligen kunde man låta helaren undersöka personer som misstänks ha levercancer, men där detta ännu inte är fastslaget med säkra metoder.) I de fall där det är ofrånkomligt att någon närvarande har kunskap om det rätta svaret, får man helt enkelt nöja sig med att söka eliminera möjligheten att kunskapen medvetet eller omedvetet kommuniceras.

Ytterligare ett exempel på bristande dubbelblindhet är ett autentiskt och ganska subtilt fall. Parapsykologerna Russell Targ och Harold Puthoff utförde experiment med fjärrsyn (”remote viewing”) under 1970-talet. En av de mest omtalade experimentserierna gick till så här: Pat Price, personen vars paranormala förmåga skulle testas, befann sig tillsammans med en försöksledare på Stanford Research Institute (SRI). Vid en förutbestämd tidpunkt skulle Price beskriva den plats där två till fyra andra personer befann sig. Platsen hade valts slumpvis ur en lista över mer än 100 platser (som låg någorlunda nära SRI). Varken Price eller försöksledaren visste vilken plats det var. Prices beskrivning spelades in på band. Därefter fick Price åka ut till platsen för att omedelbart

kunna bilda sig en uppfattning om hur väl han lyckats med beskrivningen. Proceduren upprepades med nio olika platser. För att utvärdera Prices förmåga skrevs hans beskrivningar ut, och dessa utskrifter jämte en lista över de nio platserna gavs till oberoende bedömare. Bedömarna besökte platserna och skulle för var och en av dem välja vilken beskrivning de tyckte passade bäst. En statistisk utvärdering visade att bedömarnas ihopparningar blev så bra att sannolikheten att resultatet uppkommit av en slump var mindre än en på miljarden. Targ och Puthoff menade att de därmed fått bevis för ett paranormalt fenomen.[7]

Det fanns två avgörande fel i försöksuppläggningen. Det första felet var att Price fick åka ut till målplatserna efter varje försök. Detta hjälpte visserligen knappast Price själv att genomskåda vilken nästa plats skulle bli, men det kunde i högsta grad hjälpa dem som skulle para ihop beskrivningar och platser, ty Prices beskrivningar kan då innehålla referenser till de besökta platserna, eller till platsernas ordning. Flera sådana referenser fanns i utskrifterna. Price sade att han tyckte att uppgiften var svår och att han tvivlade på att han skulle lyckas. Det uttalandet gällde mål nr 1. Om mål nr 2 sade han att det var den ”andra platsen för dagen”. Vid beskrivningen av mål nr 3 hänvisade han till ”gårdagens två målplatser”. Vid mål nr 7 nämnde han marinan, som hade varit mål nr 4. Och det fanns ännu fler exempel. Bedömarna fick därmed många ledtrådar om platsernas ordning. Targ och Puthoff redigerade inte bort sådana uttalanden, men även om de hade gjort det, är det inte säkert att undersökningen blivit invändningsfri. Price kan nämligen (medvetet eller omedvetet) ha påverkats av de tidigare besökta platserna att låta senare beskrivningar skilja sig så mycket som möjligt från dem. Om målplats nr 1 var en skogsdunge, så kanske Price undvek att nämna träd i de senare beskrivningarna. Om det störtregnade vid Prices besök på målplats nr 2 så kanske han särskilt talade om den uppklarnande himlen vid beskrivningen av mål nr 3, och så vidare. Det enda säkra sättet att undvika sådana här ledtrådar rörande platsernas ordning hade varit att inte låta Price besöka någon av platserna förrän efter det att alla försök genomförts.[8]

Det andra felet var att den förteckning över platser som delgavs bedömarna upptog platserna i korrekt besöksordning, inte i slumpmässig ordning. I kombination med Prices referenser till tidigare besökta platser blev bedömarnas uppgift därmed ovanligt enkel. Detta fel kunde man inte upptäcka genom att läsa Targs och Puthoffs rapport, utan det uppdagades först sedan en besökande forskare försökt replikera experimentet.[9] Ingen av bedömarna i Price-experimentet protesterade mot den bristande blindningen. Tydligen tänkte ingen på att information om ordningen kunde utvinnas ur Prices beskrivningar.

Ibland, särskilt i medicinska sammanhang, kan man tala om ”trippelblindhet”, eller t.o.m. om ”kvadrupelblindhet”.[10] Med det menas att det utöver vanlig dubbelblindhet finns ytterligare en eller två relevanta grupper av människor som hållits blindade. Den tredje gruppen kan vara den eller de personer som utfört statistisk utvärdering av resultaten. Man har alltså särskilt sett till att statistikerna inte vet vad siffrorna egentligen betyder eller vilket resultat som förväntas. En fjärde grupp skulle kunna vara personerna som författar rapporten. (Men ofta är det förstås samma personer som både utför experimentet, analyserar det och skriver rapporten.) Trippelblindhet och kvadrupelblindhet kan inte betraktas som allmänt vedertagna begrepp, åtminstone inte utanför medicinen, och man bör använda dem med försiktighet. Framför allt ska man komma ihåg att det finns situationer där inte ens dubbelblindhet kan uppnås. Låt oss återvända till levercancerundersökningen. Även om t.ex. helaren och de statistiska utvärderarna vore blindade, så bör man inte på denna grundval kalla testet ett dubbelblindtest. Dubbelblindhet

syftar specifikt på att alla som direkt kan påverka utfallet är okunniga om de rätta svaren, inte på att vilka två grupper som helst varit blindade.

Statistik

Ofta kan man få höra att ett resultat är ”signifikant” eller ”skiljer sig signifikant från slumpen”. Vad betyder det? Det betyder att avvikelsen från det statistiskt förväntade värdet är så stor att sannolikheten är liten att en sådan avvikelse (eller en ännu större) uppkommit av en slump. Som exempel kan vi ta en experimentuppläggning från parapsykologin, använd främst under första halvan av 1900-talet. Den går ut på kortgissning, men istället för en vanlig kortlek används den s.k. zenerkortleken (uppkallad efter en person vid namn Zener), innehållande 25 kort, 5 vardera med symbolerna kors, stjärna, cirkel, våg, fyrkant.

Antag att vi gör 100 försök. Mellan varje försök blandas kortleken enligt konstens alla regler, och försökspersonens uppgift är att tala om vilken symbol som ligger överst i varje försök (givetvis utan att kunna se kortet eller ta reda på det på något annat normalt sätt). Ett slumpmässigt resultat vore att få ungefär 20 rätt. Frågan är hur mycket resultatet måste avvika från 20 för att det ska kunna kallas osannolikt. Det beror förstås på hur osannolika resultat vi kräver för att ta dem på allvar. Ett mått på detta är den s.k. *signifikansnivån*. En vanlig nivå är 5%. Resultat som är signifikanta på 5%-nivån avviker så mycket från slumpresultaten att *om* slumpmodellen är korrekt så uppkommer såpass avvikande resultat i högst 5% av fallen. Man kan tycka att 5% är alltför generöst, och i så fall kan man använda en lägre nivå, 1%, 0,1% eller ännu lägre.

I vårt zenerkortexempel förväntar vi oss ungefär 20 rätt utifrån hypotesen att den testade personen inte har någon paranormal förmåga. En sådan hypotes, att en viss förmåga *inte* föreligger, att det *inte* är något annat än slumpen som råder, eller att ett nytt läkemedel *inte* är bättre än ett etablerat, brukar kallas *nollhypotes*. Den som vill hävda något annat måste visa att nollhypotesen är fel. Om den testade personen har en paranormal förmåga borde det yttra sig som ett resultat bättre än 20 rätt. Dock bör man notera att det utifrån slumpmodellen (nollhypotesen) vore lika överraskande med ett ovanligt lågt antal korrekta gissningar som med ett ovanligt högt. Därför vill vi kanske fördela signifikansen så att vi betraktar resultatet som statistiskt signifikant om det *antingen* är så dåligt att sannolikheten för ett så dåligt resultat eller sämre är högst 2,5% *eller* är så bra att sannolikheten för ett så bra resultat eller bättre är högst 2,5%. Signifikansnivån för detta tvåsidiga test blir då 5%. (Testet kallas tvåsidigt eftersom det finns två distinkta områden inom vilka vi förkastar nollhypotesen, dels om resultatet blir klart sämre än ett slumpresultat, dels om det blir klart bättre.) Exakt hur man räknar ut gränserna är inte viktigt i denna framställning – det viktiga är att de *går* att räkna ut.

Ibland uttrycker man området inom vilket man tror att det korrekta värdet ligger som ett *konfidensintervall*. Med ett 95-procentigt konfidensintervall menas ett intervall som med sannolikheten 95% innehåller det sanna värdet. Konfidensintervallet kan också uppfattas som de värden för vilka nollhypotesen inte skulle ha förkastats på signifikansnivån 5%. Konfidens betyder här alltså motsatsen till signifikans. Detta kan vara förvirrande, men begreppen används om olika saker: signifikans om ett *test*, konfidens om ett *intervall*.

Tilläggas bör att man ibland har anledning att beräkna konfidensintervall utan att man har någon hypotes som man vill testa. Exempelvis kan man göra ett flertal mätningar med ett instrument för att bestämma dess noggrannhet, vilken kan uttryckas med ett konfidensintervall.

I normalfallet bör signifikansnivån bestämmas i förväg. Likaså bör man i förväg bestämma exakt vilka statistiska tester man tänker underkasta datamaterialet. Det är annars lätt hänt att man prövar många statistiska metoder och att man överskattar betydelsen av funna resultat. Att pröva många statistiska metoder på måfå brukar benämnas *datafiske*, på engelska vanligen *data mining* eller *data dredging* (mining = 'gruvbrytning', dredging = 'muddring'). Om man utför tillräckligt många olika slags tester kan man nämligen förvänta sig att finna signifikanta effekter även i ett fullständigt slumpmässigt datamaterial. Fenomenet kallas ibland *massignifikans*.

Det finns flera sätt att erhålla massignifikans. Dels kan man använda olika statistiska metoder på ett och samma material. Dels kan man använda samma statistiska test (eller flera olika) på olika delar av datamaterialet. Antag t.ex. att materialet består av uppgifter om ett stort antal människor. Man kan då utföra ett statistiskt test som omfattar alla personer, men också enbart kvinnor, män, personer som är 20-30 år gamla, pensionärer, utrikes födda, rödhåriga, osv. Med så många tester, och en inte alltför snål signifikansnivå, vore det inte underligt om man någonstans finner signifikanta resultat. Numera krävs regelmässigt inom läkemedelsindustrin att forskare redan under planeringsfasen av undersökningen anger vilken statistisk metod de tänker använda och hur den ska tillämpas.

Det kan förstås finnas situationer när det är intressant och högst rimligt att undersöka mindre delar av det totala datamaterialet, och det är givetvis inte förbjudet att utföra många tester, men resultaten måste tolkas försiktigt om man inte använt en pålitlig matematisk metod för att kompensera signifikansnivåerna för upprepad testning. (Flera sådana metoder existerar, men vi går inte in på dem här.) Det finns också situationer när det är tillåtet att inte ha bestämt exakt vilka tester som ska utföras, men då sysslar man inte längre med strikt hypotesprövning. En undersökning där man letar efter signifikanser utan att i förväg ha bestämt vad man ska söka efter, kan kallas *explorativ* (utforskande). Explorativa undersökningar är helt legitima när det är svårt att i förväg veta vilka hypoteser som är rimliga att testa. Om man finner intressanta resultat i en explorativ undersökning bör man dock sedan göra en ny undersökning, med nya data, för att strikt testa om effekterna kvarstår när man bestämt sig för att undersöka just dem.

Matematiken är förstås viktig vid statistisk hypotesprövning. Lika viktigt är dock att man tänker igenom förhållandet mellan modell och verklighet. Man måste kunna ställa upp en slumpmodell som det experimentella utfallet ska jämföras med *på ett meningsfullt sätt*. I fallet med kortgissning är jämförelsen med slumpmodellen meningsfull (förutsatt att inga metodfel föreligger). Men om man krånglar till situationen en smula kan det meningsfulla lösas upp i intet. Ett autentiskt exempel är en slagruteundersökning från 1971 av Duane G. Chadwick och Larry Jensen vid Utah State University.[11]

Chadwicks och Jensens experiment utfördes först i en fruktträdgård med långa rader av äppelträd. Teststräckan förlades mellan två sådana rader. På ett ställe längs sträckan grävdes en järnstav ned på ca 15 centimeters djup. Platsen täcktes över så att det skulle vara omöjligt att upptäcka att den rörts. Försökspersonerna var ovetande om järnstaven. De flesta hade aldrig tidigare använt slagruta. Alla fick ett flertal utslag längs sträckan. Om man prickar in dessa i en figur ser man att det knappast finns någon ansamling av utslag just vid järnstaven. Samma slags experiment utfördes även längs ett järnvägsspår på en bangård samt på en stor gräsmatta i en park. I båda fallen grävdes järnstavar ned, men inte i något fall påverkade stavarna utslagens placering. Utslag förekom längs hela sträckorna. Sedd som ett test av orutinerade rutgängares förmåga att finna underjordiskt järn gav studien alltså ett tydligt negativt besked. Men Chadwick och Jensen undersökte utslagens fördelning längs teststräckorna närmare. Utslag förekom visserligen längs

hela teststräckorna, men de var ändå inte slumpmässigt utspridda. Utslagen ser ut att vara grupperade, visserligen i ganska många grupper, men dock i grupper. Statistiska beräkningar bekräftade detta, och följaktligen – menade författarna – fanns något oförklarligt med slagrutan. Men dessa statistiska beräkningar är inte relevanta för den undersökta frågan. De förutsätter att slumpmässig spridning av utslagen är vad man kan förvänta sig av helt naturliga orsaker, men så är inte fallet. Om man sätter slagrutor i händerna på försökspersoner och ber dem gå längs en sträcka och markera var de får utslag, så kommer de troligen att få utslag på ställen som av någon anledning är anmärkningsvärda – kanske vid ett ovanligt stort äppelträd, vid en extra bred järnvägssyll, vid ett hål i marken, vid en sten, vid ett ogräs, vid ett bortkastat kolapapper, och så vidare. Att utslagen bildade många små grupper i denna undersökning betyder antagligen bara att det fanns många småsaker att lägga märke till för försökspersonerna.

Kontrollgrupper

Vi har redan berört kontrollgrupper en smula. Angående helaren och vibrationerna så konstaterades det att ett test måste gå ut på att helaren ska peka ut sjuka personer i en grupp med både sjuka och friska. Man kan då se de friska som en kontrollgrupp i förhållande till de sjuka. Det är ju möjligt att helaren pekar ut 75% av alla sjuka som sjuka, men det är mindre imponerande om han samtidigt pekar ut 75% av de friska som sjuka. Undersökningen måste göras med både sjuka och friska. Om man bara använder personer från den ena gruppen så kan man inte tolka resultaten.

Begreppet kontrollgrupp är välkänt från medicinska undersökningar. Vid läkemedelsprövningar får vissa personer läkemedlet och andra placebo (eller ett konkurrerande läkemedel med kända effekter). De senare personerna utgör kontrollgrupp. Om man inte hade någon kontrollgrupp skulle man visserligen kunna konstatera att ”si och så många procent mådde bättre efter att ha tagit läkemedlet”, men man skulle inte veta om det var ett bra eller dåligt resultat. Människors hälsa kan bero på så många, svårkontrollerade faktorer. Idén med en kontrollgrupp är att man genom jämförelse ska kunna filtrera bort inverkan från dessa faktorer och enbart få kvar läkemedlets specifika effekt. En förutsättning för att detta ska kunna stämma är att kontrollgruppen är så lik läkemedelsgruppen som möjligt, i alla relevanta avseenden.

Antag att vi vill ta reda på vilka sjukdomar som rökare drabbas av. Vi kan t.ex. följa ett stort antal rökare under lång tid och låta göra hälsokontroller med jämna mellanrum. Resultat kan bli av typen ”X% av rökarna drabbas av hjärtinfarkt inom 10 år”. Men ett sådant resultat är inte speciellt intressant om vi inte kan jämföra med icke-rökare. Vi bör därför parallellt följa en kontrollgrupp bestående av personer som är lika rökarna i alla avseenden utom just när det gäller rökningen. Ett förekommande missförstånd är att kontrollgrupper bör bestå av ”ett representativt urval av befolkningen” eller något liknande. Så är det alltså i regel inte.

Ett annat tankefel om kontrollgrupper är att de måste vara lika stora som försöksgrupperna. Det är alls inte nödvändigt. Det viktiga är ju hur stora *andelar* av försöksgrupp och kontrollgrupp som vid utvärderingen hamnar i de olika kategorierna. Det går utmärkt att jämföra reaktionerna hos 231 personer som får ett nytt läkemedel med reaktionerna hos en kontrollgrupp på, säg, 157 personer som får placebo. Men lika sant som att försöks- och kontrollgrupp *inte måste* vara lika stora, lika sant är det att de oftast *bör* vara *ungefär* lika stora. Skälet är helt enkelt att man då får de säkraste statistiska resultaten. Om jag vid en läkemedelsprövning har 288 personer att fördela mellan försöks- och kontrollgrupp, så är det dumt att ge läkemedlet till 283 personer och placebo till

bara 5. Kontrollgruppen blir då så liten att de förväntade slumpvariationerna i resultatet för denna grupp blir stora, sedda i relation till gruppstorleken.

Ett undantag från den allmänna regeln att försöks- och kontrollgrupp bör vara ungefär lika stora är när försöksgruppen av något ofrånkomligt skäl måste vara väldigt liten. Säg att vi undersöker en mycket ovanlig sjukdom och att försöksgruppen består av personer med sjukdomen. Vi kanske bara kan skrapa ihop 15 personer som kan ingå i gruppen. Om det däremot är lätt att få ihop personer till kontrollgruppen så är det förmodligen dumt att låta den vara lika liten.

Ett tredje missförstånd är att man alltid behöver kontrollgrupper. Poängen är att en kontrollgrupp ska utgöra den referensnivå mot vilken den andra gruppen jämförs. Men ibland känner vi till den nivån ändå. I kortgissningsförsök vill vi veta om träffrekvensen skiljer sig från det slumpmässigt förväntade värdet. Det är slumputfallet som utgör referensnivå, och den kan vi beräkna exakt. Därför behöver vi ingen kontrollgrupp eller ”kontrollkortlek” eller dylikt.

Dock ska det sägas att det finns en fördel med att använda kontrollgrupper i situationer när de egentligen inte behövs: Man kan upptäcka metodfel. Om både försöks- och kontrollgrupp visar resultat som starkt avviker från det slumpmässigt förväntade, är det en fingervisning om att det kan finnas förbisedda fel i upplägget.

Slumpmässighet

Slumpen kommer ofta in i empiriska undersökningar på ett eller annat sätt. Den enklaste förklaringen av ”slumpmässig” är ”oförutsägbar”. Om jag kastar ett mynt tillräckligt högt och med tillräckligt stor rotation så blir det helt omöjligt att förutsäga vilken sida som ska komma upp. Resultatet blir antingen krona eller klave varje gång, men det går inte att säga vilketdera i förväg. Därför kan vi säga att resultatet är slumpmässigt. Men observera att det är en *praktisk* oförutsägbarhet: Det går inte i praktiken att förutsäga vilken sida som kommer upp, särskilt inte under den korta tid som myntet är i luften. Ur klassisk-fysikalisk synvinkel är myntets rörelse däremot deterministisk, och teoretiskt borde det vara möjligt att beräkna utfallet givet tillräcklig information om initialvärden. Ett fenomen som såvitt bekant även rent teoretiskt är oförutsägbart är radioaktivt sönderfall. Det är dock ganska krångligt att generera slumptal ur radioaktivt sönderfall, och därför nöjer man sig ofta med metoder som bara är praktiskt oförutsägbara. Exempel på sådana – om de utförs kompetent – är slantsingling, tärningskast, lottdragning, avläsning i slumptalstabell (finns publicerade i bokform) och användning av matematiska dataprogram.

Det finns förstås massor av metoder som *inte* ger slumpmässiga resultat. Att t.ex. låta varannan person komma till försöksgruppen är inte slumpmässigt, inte heller att ta den första halvan (såvida inte gruppen redan tidigare är slumpmässigt ordnad). Särskilt bör det betonas att man inte får slumpmässighet genom att göra ”som man känner”, eller ”efter eget huvud”, alltså att man utan hjälpmedel försöker simulera en fördelning som verkar slumpmässig. Människor är nämligen ganska dåliga på att simulera slump, fastän vi ofta tror motsatsen. Om man ber ett stort antal människor att tänka på ett ensiffrigt heltal, vilket som helst (alltså 0, 1, 2, …, 9), så kunde man tro att fördelningen skulle bli ganska jämn mellan de tio alternativen. Men så är det inte. I undersökningar har det visat sig att vissa tal är mycket vanligare än andra – 7 förefaller vara populärast.[12] Eller ett annat exempel: Ge en person i hemläxa att singla slant 60 gånger och skriva upp resultatet i form av ettor och nollor, där ”1” betyder krona och ”0” klave. Antag att vi får denna serie till svar:

100110101110110011010100101101101011000111010100100111010010

Skulle denna serie kunna uppkomma ur verklig slantsingling? Ja, självklart. Men om vi inte vet huruvida den är resultatet av verklig slantsingling eller personens fantasi, vad ska vi då tro? Serien har 32 ettor och 28 nollor. Fördelningen är alltså ganska jämn och trovärdig som äkta slumpserie.[13] Hur många ettor eller nollor i rad hittar vi som mest? Svaret är tre. Det är mindre trovärdigt. Om en 60 siffror lång följd skapas slumpmässigt kan man förvänta sig att någonstans finna minst fyra ettor eller nollor i rad.[14] Ur detta lär vi oss att slumpmässighet inte enbart har att göra med frekvensen av enstaka utfall, utan även med frekvensen av utfallskombinationer.

Man kan också uttrycka detta så att det är skillnad mellan slumpmässig *sammansättning* och slumpmässig *ordning*. Ett datamaterial kan mycket väl vara slumpmässigt till sin sammansättning men icke slumpmässigt ordnat. Ett datamaterial kan omvänt vara slumpmässigt ordnat utan att vara slumpmässigt till sin sammansättning. Tag t.ex. helaren och vibrationerna. Försöksgruppen behöver inte utgöra något slumpmässigt urval ens bland levercancersjuka (men kan göra det). Kontrollgruppen bör matcha försöksgruppen utom när det gäller sjukdomen, och bör därför inte utgöra ett slumpmässigt urval ur den friska befolkningen. Ingen av gruppernas sammansättningar behöver alltså ha något slumpmässigt över sig. Däremot är det helt avgörande att sjuka och friska blandas i slumpmässig ordning när de presenteras för helaren.

I exemplet med helaren vill vi utsätta två olika grupper för samma behandling, nämligen helarens undersökning. Ofta är man intresserad av det motsatta, att utsätta likartade grupper för olika behandling. Det är en typisk situation när man prövar ett nytt läkemedel. Man har då ett stort antal personer med likartade egenskaper (t.ex. förkylda kvinnor i åldrarna 20-40 år). Med hjälp av någon slumpmekanism låter man vissa personer hamna i försöksgruppen och andra i kontrollgruppen. En sådan slumpvis placering i grupper kallas *randomisering*.

Några metodfel i kliniska prövningar

Med *klinisk prövning* (engelska *clinical trial*) brukar man mena en undersökning på friska eller sjuka personer för att studera effekterna av en behandlingsform, ofta ett läkemedel. Adjektivet ”klinisk” indikerar att undersökningen sker i praktisk medicinsk verksamhet, till skillnad från en studie i laboratorium. Kliniska prövningar genomförs vanligen i flera faser. Först ges läkemedlet till ett litet antal friska frivilliga, sedan till ett litet antal sjuka. Slutligen görs en större, kontrollerad undersökning. Att undersökningen är kontrollerad betyder bl.a. att man använder en kontrollgrupp. Det är sådana större studier man vanligen syftar på om man säger att ett läkemedel undersökts i en klinisk prövning. Man brukar ofta underförstå att studien varit dubbelblind och randomiserad.

En undersökning från 1948 har framhållits som den första kontrollerade kliniska prövningen.[15] Dock har vissa läkare varit medvetna om behoven av kontroll, blindhet och randomisering långt tidigare, och några forskare har lyft fram en dansk undersökning från 1898 som den första acceptabelt genomförda kliniska prövningen.[16] Den gjordes av läkaren Johannes Fibiger för att ta reda på om serumbehandling var effektiv mot difteri. De difteridrabbade patienter som under ett helt år 1896-97 kom in till Blegdamshospitalet i Köpenhamn fick antingen standardbehandling eller standardbehandling plus seruminjektioner. Vilken grupp patienten hamnade i bestämdes av ankomstdagen. (Det är inte en i alla lägen optimal randomiseringsmetod, men förmodligen fungerade den väl här.) Vissa patienter fick uteslutas ur studien sedan man inte kunnat påvisa difteribakterier hos dem. Av 239 patienter i serumgruppen dog 8, medan 30 av 245 patienter dog i kontrollgruppen. Någon statistisk hypotesprövning gjordes inte, eftersom sådana meto-

der var outvecklade vid denna tid. Fibiger menade ändå att resultatet var tydligt: Serum hjälpte mot difteri.

Kliniska prövningar är ganska komplicerade företag. De innefattar många olika moment, och det räcker att ett enda av dem sköts slarvigt för att hela undersökningen kan bli värdelös. De mest uppenbara felkällorna är förstås brott mot blindnings- och randomiseringskraven, eller användning av massignifikans, vilka alla har diskuterats ovan. Här följer ytterligare åtta fel som inte är helt ovanliga i kliniska prövningar. Flera av felen kan förstås återfinnas även i undersökningar utanför medicinens område.[17]

Först några fallgropar som har med den statistiska behandlingen att göra: (1) Trots randomisering kan det hända att försöks- och kontrollgrupperna skiljer sig åt i viktiga avseenden, t.ex. när det gäller initial hälsostatus. Om försöksgruppen från början har sämst hälsa så är det troligt att deras förbättring blir störst, alldeles oavsett om behandlingen är verksam eller ej. Om möjligt bör man därför kontrollera relevanta variabler i de båda grupperna vid studiens början, så att de inte skiljer sig mycket från varandra. (2) I alla studier med många personer inblandade förekommer bortfall, t.ex. patienter som inte längre vill delta eller som avlider. Underlåtelse att analysera bortfallet kan vara ett allvarligt fel. Om bortfallet t.ex. är mycket större i försöksgruppen än i kontrollgruppen så kan det säga något viktigt om den undersökta behandlingsmetoden. (3) I ett datamaterial måste alltid några värden utgöra extremer. Om dessa värden avviker kraftigt från de övriga brukar man tala om ”outliers” (även på svenska). Det är viktigt att kontrollera sådana värden, så att de inte uppkommit pga. mät- eller skrivfel. Man bör kontrollera *alla* outliers, och inte falla för frestelsen att bara undersöka dem som genom att ändras skulle kunna stödja en önskad slutsats.

Några punkter som har mer med själva upplägget att göra: (4) Man bör i förväg ha bestämt exakt hur omfattande undersökningen ska vara. Det är annars lätt hänt att man avbryter försöken när resultaten ser ovanligt bra ut, alternativt förlänger studien så att resultaten ska bli lite bättre. (5) Vid statistisk hypotesprövning försöker man oftast förkasta en nollhypotes som säger att två behandlingar är likvärdiga. För att statistiska test ska kunna användas fordras en tillräcklig datamängd. Alltför få data leder till att nollhypotesen svårligen kan förkastas, trots att de två behandlingarnas effektivitet i verkligheten kan skilja sig mycket. Ett lömskt sätt att lyfta fram en dålig behandling är därför att som nollhypotes anta att den är lika bra som en välkänt bra behandling, men sedan utföra en klinisk prövning som är så underdimensionerad att nollhypotesen inte kan förkastas. (6) Ett besläktat metodfel är att låta personerna i kontrollgruppen få en behandling som försämrar deras tillstånd och därmed gynnar försöksgruppen vid jämförelsen.

Några fel som mest har med presentationen av resultaten att göra: (7) Se upp för påståenden om gruppskillnader som inte åtföljs av upplysningar om eventuella statistiska signifikanser. Det är inte alls ovanligt att forskare som förväntat sig en tydlig skillnad mellan grupper framhåller att man verkligen fann en skillnad, men undviker att tala om att skillnaden inte var statistiskt signifikant. Särskilt misstänksam ska man vara när dylika resultat utan upplysningar om signifikanser framhävs redan i artikelns sammanfattning (abstract). (8) Kontrollgruppens behandling måste vara väl undersökt sedan tidigare, annars vet man inte vad jämförelsen mellan grupperna går ut på. Så här kan man få en dålig behandling att framstå i god dager: Jämför behandlingen med en fastslaget dålig behandling. Tala inte om det för läsaren. Om resultatet blir att behandlingarna är jämbördiga, övertyga läsaren om att båda är bra.

Att tänka efter före

Jag hoppas ha visat att empiriska undersökningar av kontroversiella fenomen har många potentiella fallgropar. Samtidigt vill jag understryka att metoden för att undvika felen är att använda vanligt, förnuftigt tänkande. Det krävs ingen speciell utbildning för att lägga upp empiriska undersökningar, men det kräver en del planering. Därför är det viktigt att tänka igenom upplägget i förväg.

När man ska testa en speciell persons påstådda kontroversiella förmåga (såsom telepati eller healing) är det särskilt viktigt att försöksupplägget har godkänts i förväg av denna person. Försöksledare och försöksperson bör då vara överens om vad testet är tänkt att undersöka, exakt hur det ska gå till, hur omfattande det ska vara, tolkningsprinciper, statistisk metod, signifikansnivå, att alla parter ska acceptera resultatet och inte skylla på något som inte varit uppe till diskussion, och så vidare. Allt för att slippa hamna i oklarheter om hur resultatet ska tolkas.

För att komma överens om detta kan ingående diskussioner och förklaringar av vetenskaplig metod krävas. Försökspersonen bör självfallet inte tillåtas att styra undersökningen så att dess vetenskapliga värde försvagas. Om försökspersonen inte är beredd att medverka i en undersökning som är vetenskapligt invändningsfri, ska man *inte* ge efter för kraven om ett ovetenskapligt tillvägagångssätt.

Att tänka igenom undersökningsmetoden i förväg är bra i all empirisk forskning, inte bara när det gäller kontroversiella undersökningar. Det råder stor enighet inom vetenskapssamfundet att undersökningar med metodbrister är att betrakta som mer eller mindre värdelösa. Tyvärr är denna insikt ofta svår att förmedla till personer som har stark emotionell bindning till ovetenskapliga föreställningar.

Noter

[1] En principiellt positiv inställning till empiriska undersökningar betyder dock *inte* nödvändigtvis att sådana ska utföras till varje pris eller i obegränsad mängd. Det behöver ingalunda finnas någon motsättning mellan att vara positiv till empiriska undersökningar och att anse att det inte är lönt att satsa mer pengar på att undersöka om t.ex. astrologin fungerar, eftersom detta redan har undersökts i hundratals studier. Jag går inte närmare in på denna typ av avvägningsfrågor, men vill uppmärksamma läsaren på att de existerar.

[2] Se vidare Sven Ove Hanssons text om trolleri och bedrägeri i denna volym.

[3] Se vidare texten om slagrutor i denna volym.

[4] E. Davenas et al., ”Human basophil degranulation triggered by very dilute antiserum against IgE”, *Nature* 333, 1988, s. 816–818.

[5] J. Maddox, J. Randi & W. W. Stewart, ”’High-dilution’ experiments a delusion”, *Nature* 334, 1988, s. 287–290.

[6] S. J. Hirst et al., ”Human basophil degranulation is not triggered by very dilute antiserum against human IgE”, *Nature* 366, 1993, s. 525–527.

[7] R. Targ & H. Puthoff, ”Information transmission under conditions of sensory shielding”, *Nature* 251, 1974, s. 602–607.

[8] C. Scott, ”Remote viewing”, *Experientia* 44, 1988, s. 322–326.

[9] D. Marks & R. Kammann, ”Information transmission in remote viewing”, *Nature* 274, 1978, s. 680–681.

[10] Så t.ex. i A. Jadad, *Randomiserade kontrollerade kliniska prövningar*, Lund: Studentlitteratur 2000, s. 43.

[11] D. G. Chadwick & L. Jensen, *The Detection of Magnetic Fields Caused by Groundwater and the Correlation of Such Fields with Water Dowsing*. Progress Report 78:1. Utah Water Research Laboratory, 1971.

[12] Se t.ex. D. Marks, *The Psychology of the Psychic*, Amherst NY: Prometheus 2000, s. 311–317.

[13] Jag menar inte att fördelningen *måste* vara så här jämn för att vara trovärdig, bara att det hade varit osannolikt med en *väldigt* ojämn fördelning, t.ex. 50 ettor och 10 nollor. Faktum är att människor snarast överskattar sannolikheten att fördelningen blir mycket jämn i sådana här serier; se A. Tversky & D.

Kahneman, ”Belief in the law of small numbers”, i D. Kahneman, P. Slovic & A. Tversky (red.), *Judgment Under Uncertainty: Heuristics and Biases*, Cambridge: Cambridge University Press 1982, s. 23–31.

[14] Sannolikheten för detta är 99,2%. Sannolikheten att vid N myntkast någonstans få minst n likadana utfall i följd kan rekursivt tecknas $P(N, n) = P(N - 1, n) + 2^{-n}[1 - P(N - n, n)]$. När $N < n$ så gäller förstås $P(N, n) = 0$, medan för $N = n$ gäller $P(n, n) = 2^{1-n}$. Formeln har jag fått från matematikern Jonas Sjöstrand, KTH.

[15] Medical Research Council, ”Streptomycin treatment of pulmonary tuberculosis”, *British Medical Journal* ii, 1948, s. 769–782.

[16] A. Hróbjartsson, P. C. Gøtzsche & C. Gluud, ”The controlled clinical trial turns 100 years: Fibiger’s trial of serum treatment of diphtheria”, *British Medical Journal* 317, 1998, s. 1243–1245.

[17] Jag har sammanställt dessa fel huvudsakligen från följande två källor. (1) T. Greenhalgh, ”Statistics for the non-statistician. II: ’Significant’ relations and their pitfalls”, *British Medical Journal* 315, 1997, s. 422–425. (2) E. Ernst, ”How to show that an ineffective therapy works”, *Drug Discovery Today* 9, 2004, s. 99–100.

Denna text har publicerats i Jesper Jerkert & Sven Ove Hansson (red.), *Vetenskap eller villfarelse*, Stockholm: Leopard 2005, s. 289–305, 339–340. Sidnumreringen i detta pdf-dokument överensstämmer alltså inte med den tryckta versionen.