Att genomföra och bedöma statistiska undersökningar

Pär Nyman och Marcus Österman
2016-01-20

Statistiska urvalsundersökningar är fascinerande. Genom att intervjua endast något tusental personer kan vi skaffa oss en relativt god uppfattning om hela den svenska befolkningens åsikter och levnadsförhållanden. Men vad ska man tänka på när man genomför en sådan undersökning? Och hur kan vi bedöma pålitligheten i de undersökningar som andra genomför?

I den här texten tar vi upp de aspekter av statistiska undersökningar som vi inte tycker att Teorell och Svensson behandlar tillräckligt ingående. Vi använder framför allt exempel från opinionsundersökningar, eftersom det är något ni alla är bekanta med, men resonemangen är lika giltiga för andra typer av statistiska undersökningar.

Innehållet i texten sträcker sig över flera av kursens föreläsningar, men vi rekommenderar i stort att ni läser texten inför seminarium 2. Det vi i texten skriver om standardfel och normalfördelningen är svårt och kommer inte att examineras på tentan. Syftet med dessa avsnitt är att ge en förståelse för logiken bakom konfidensintervall. Texten avslutas med avsnittet ”Tolka signifikans försiktigt” som framför allt är relevant när vi senare på kursen går in på regressionsanalys. Den delen av texten rekommenderar vi därför att ni läser inför seminarium 3 eller 4.

Först introduceras slumpmässiga och systematiska mätfel, innan vi utifrån dessa två typer av mätfel diskuterar urvalsmetoder, bortfall, intervjuarsituationen samt frågeformulering. Under alla dessa rubriker diskuterar vi möjliga strategier för att minska risken för systematiska mätfel. Såvida vi inte genomför en totalundersökning och intervjuar hela den population vi är intresserade av, kommer vår urvalsosäkerhet alltid att ge upphov till slumpmässiga mätfel. I den här texten beskriver vi hur vi kan hantera slumpmässiga mätfel genom att beräkna konfidensintervall och testa om skillnader och samband är statistiskt signifikanta. Eftersom statistisk signifikans är så centralt i kvantitativa studier avslutar vi med en diskussion om när en övertro på signifikanta resultat kan leda till missvisande slutsatser.

Slumpmässiga och systematiska mätfel

Alla urvalsundersökningar brottas med två grundläggande problem. För det första vill vi minimera risken för slumpmässiga mätfel. Slumpmässiga mätfel innebär att vi ibland underskattar och ibland överskattar det vi är intresserade av, men om vi skulle upprepa undersökningen ett oändligt antal gånger skulle vi i genomsnitt pricka rätt. Annorlunda uttryckt så snedvrider inte ett slumpmässigt mätfel resultatet åt något särskilt håll utan lägger till ett generellt "brus" i våra resultat.

Urvalsstorleken är den viktigaste faktorn för hur stora de slumpmässiga felen är. Ju fler personer som tillfrågas i en opinionsundersökning, desto mindre är risken att något parti "råkar" vara kraftigt överrepresenterat bland de som svarar. Om vi har stora slumpmässiga mätfel är sannolikheten hög för att två i övrigt identiska undersökningar ska visa mycket olika resultat, bara för att de intervjuade personerna var olika. När våra beräkningar dras med stora slumpmässiga mätfel säger vi att vi har dålig *precision*.

För det andra vill vi minimera förekomsten av systematiska mätfel. Sådana mätfel innebär att vi på grund av vår metod tenderar att antingen under- eller överskatta det vi vill mäta. Systematiska mätfel minskar inte genom att vi ökar storleken på urvalet. Denna typ av mätfel kan bland annat orsakas av systematiska bortfall, intervjuareffekter och ledande frågeformuleringar. Ett systematiskt mätfel uttrycks ofta som att vi har en *bias* eller att resultaten är *snedvridna*.[1]

I stora kvantitativa undersökningar utgör systematiska mätfel vanligtvis ett större problem för en forskare än slumpmässiga. Slumpmässiga mätfel kan uppskattas och hanteras med hjälp av statistiska hjälpmedel (mer om detta senare i texten). Systematiska mätfel är det däremot normalt svårt att uppskatta storleken på och kan inte heller avhjälpas genom att använda större stickprov.

Ett slumpmässigt mätfel kan liknas vid ett tärningsslag. Slår jag en sexsidig tärning en gång är alla utfallen ett till sex lika sannolika. Slår jag den tio gånger är osäkerheten om medelvärdet för slagen fortfarande stor. Om jag däremot slår tärningen 1 000 gånger kommer medelvärdet vara nära 3,5. Ju fler gånger jag slår, desto närmare kommer jag komma 3,5. Ett systematiskt mätfel kan istället liknas med att slå en fusktärning där jag t.ex. får oproportionerligt många sexor. Hur många gånger jag än slår kommer sexorna vara överrepresenterade. Jag får alltså en snedvridning eller en bias i mina "resultat" jämfört med om jag använt en riktig tärning. I t.ex. en opinionsundersökning är det dock sällan så här uppenbart att vårt mätverktyg ger upphov till systematiska fel.

Under kursens första vecka har vi bekantat oss med begreppen reliabilitet och validitet. En fördel med dessa begrepp är att de förekommer i nästan

[1] Vill man vara mer korrekt kan man säga att resultaten inte är *väntevärdesriktiga*.

alla typer av metodlitteratur och att de flesta studenter och forskare därför har åtminstone en vag idé om dess innebörd. En nackdel är att begreppens exakta betydelse och definition kan skilja sig mellan metodologiska kontexter, metodböcker och discipliner. Det är vanligt att prata om reliabilitet och validitet när man pratar om definitioner, indikatorer och mätinstrument, men när vi beskriver ett statistiskt estimat (en uppskattning av ett värde) brukar vi i regel använda begrepp där det finns mindre ovisshet om begreppens innebörd. Därför pratar vi om precision i stället för reliabilitet och bias i stället för validitet, även om innebörden är snarlik.

Urvalsmetoder och bortfall

Vi kommer nu titta på de mer praktiska aspekterna av en statistisk undersökning, som till exempel hur vi gör vårt urval, vad konsekvenserna blir av att alla inte vill delta i undersökningen samt vad man ska tänka på när man utformar frågorna och intervjusituationen. Samtliga ämnen kommer att diskuteras utifrån hur de påverkar våra slumpmässiga och systematiska mätfel.

Genomgången nedan avser alla typer av statistisk data som bygger på ett urval, oavsett om uppgifterna är insamlade genom t.ex. intervjuer, telefonintervjuer, postenkäter eller onlineundersökningar. Det kan också vara material som forskaren själv skapat genom att ha kodat in uppgifter från en annan typ av analysenhet än personer. Det kan t.ex. röra sig om rättsfall eller protestbrev till politiker.

Urvalsosäkerhet

Det finns många olika typer av urval, men i den här texten kommer vi att prata om den grundläggande skillnaden mellan å ena sidan sannolikhetsurval och å andra sidan urval där det varken är forskaren eller ren slump som avgör vilka som ingår i urvalet.

Sannolikhetsurval innebär att varje person i undersökningspopulationen har en känd sannolikhet för att ingå i urvalet. Den enklaste metoden är att dra ett slumpmässigt urval där alla personer har samma sannolikhet att bli utvalda. Om det slumpmässiga urvalet dras från hela populationen man vill uttala sig om kommer urvalsosäkerheten då endast att ge upphov till slumpmässiga mätfel. En mer avancerad form av sannolikhetsurval innebär att vissa grupper ges en högre sannolikhet för att bli dragna än andra. Detta görs bl.a. för att ha möjlighet att närmare undersöka en grupp som utgör en liten del av befolkningen men som är särskilt intressant för undersökningen. T.ex. om vi vill få en god uppfattning om arbetslösheten även i små kommuner kan vi ge personer i dessa kommuner en större sannolikhet att ingå i urvalet än personer boende i stora kommuner. Men eftersom forskaren känner till de

olika urvalssannolikheterna kan denne kompensera för detta i sin statistiska analys och undvika systematiska mätfel.

Alternativet till sannolikhetsurval är ett urval där sannolikheten för att en viss person ska ingå är okänd för forskaren. Detta kallas för icke-sannolikhetsurval. Ofta innebär detta att sannolikheten varierar kraftigt mellan olika grupper och eftersom forskaren inte känner till hur denna variation ser ut är det omöjligt att kompensera för den.

Denna typ av "okontrollerade" urval som inte bygger på en slumpmässig dragning ur en större population har blivit allt vanligare. Exempelvis genomför många tidningar undersökningar bland alla som besöker tidningens hemsida, flera opinionsinstitut använder onlinepaneler av personer dit intresserade kan anmäla sig själva och en del studenter går runt på stan för att samla enkätsvar till sina C-uppsatser. Ett uppenbart problem med dessa typer av urval är att vi inte kan säkerställa att urvalet är representativt för den population vi vill studera. De som anmäler sig till en onlinepanel om politiska frågor är troligen mer politiskt intresserade än andra; vissa personer rör sig ute på stan i större utsträckning än andra, benägenheten att svara på enkäter som personer ger ut på gatan varierar o.s.v. Dessa urval är därmed förknippade med en urvalsosäkerhet som kan orsaka systematiska mätfel i undersökningen. Detta kallas ofta för att det finns en *selektionsbias* i urvalet. I praktiken riskerar forskaren att istället för den tänkta populationen – såsom hela Sveriges befolkning – uttala sig om en särskild population som består av politiskt intresserade eller personer som rör sig på stan och gillar att fylla i enkäter. Skillnaden i hur man svarar mellan dessa grupper och den tänkta befolkningen ger upphov till ett systematiskt mätfel.

Att dessa urvalsmetoder trots detta blir allt vanligare beror förmodligen på att det innebär betydligt billigare undersökningar (eller större urval) och att ökande bortfall har skadat pålitligheten för undersökningar som utförs med sannolikhetsurval. Med hjälp av viktning kan självrekryterade paneler göras mer representativa (Martinsson m.fl 2013). Det innebär, enkelt uttryckt, att om det t.ex. är färre äldre med i vårt urval än i populationen så låter vi svaren från de äldre som är med undersökningen räknas mer (viktas upp). Vi låter alltså de äldre som faktiskt svarat också representera dem som inte är med. Dock kan vi inte veta om de äldre som är med faktiskt svarat likadant som de äldre som inte är med skulle gjort. Ofta finns det skäl att tro att det finns en samvariation mellan de egenskaper som gör att en person är med i ett självselekterat urval, såsom t.ex. en onlinepanel, och hur denne svarar på frågorna. Att exempelvis uttala sig om internetanvändning bland äldre utifrån en onlinepanel torde innebära en kraftig överskattning av hur mycket de äldre använder internet.

Det ofrånkomliga problemet med viktning och uppräkning av vissa grupper är att det enda sättet att veta säkert om metoderna fungerar är att utvärdera dem i förhållande till studier som använder sannolikhetsurval eller andra typer av datakällor, som exempelvis registerdata och faktiska valre-

sultat. Men om vi redan har tillgång till mer pålitlig information, vad är då syftet med att genomföra mindre pålitliga undersökningar? Ett svar är kanske att vi kan använda billigare metoder för att studera snabba förändringar i exempelvis partisympatier, medan vi regelbundet måste utvärdera dem med andra metoder för att säkerställa att de fortfarande fungerar.

Bortfall

Med bortfall avses de personer som inte deltar i en undersökning trots att avsikten var att de skulle delta. Bortfall måste bedömas i förhållande till den s.k. *urvalsramen*. Urvalsramen är en specificering av populationen från vilken vi drar vårt urval. Det kan t.ex. vara röstberättigade till svenska riksdagsval mellan 18 och 80 år. I en klassisk frågeundersökning finns det fyra typer av bortfall (Holmberg & Petersson 1980):

- *Naturligt bortfall* (eller bortdefinierade). Denna grupp utgörs av personer som inte har haft fysisk möjlighet att besvara enkäten men som ingår i urvalsramen. Hit räknas personer som är långvarigt bortresta, har funktionshinder som omöjliggör deltagande, allvarligt sjuka eller som inte förstår undersökningens språk. Även personer som avlidit efter att urvalsramen fastställdes brukar räknas hit. Vanligtvis redovisas inte denna grupp som bortfall utan räknas istället bort från det ursprungliga urvalet. Det brukar uttryckas som att denna grupp bortdefinieras från urvalet.

- *Ej anträffade*. Personer som intervjuaren inte fått något svar från eller har haft någon kontakt med.

- *Vägrar deltagande*. Personer som intervjuaren får kontakt med men som inte överhuvudtaget vill vara med i undersökningen.

- *Svarsvägran*. Personer som är med i undersökningen men som inte vill svara på vissa frågor.

Många undersökningsinstitutet har rapporterat om att bortfallet har ökat kraftigt under 2000-talet och det utgör nu ett stort problem i så gott som samtliga undersökningar. Svenska SCB har jämfört med andra statistikproducenter ett lågt bortfall, särskilt i ett internationellt perspektiv, men även där ligger bortfallet i de flesta undersökningar på omkring 40 procent. I undersökningar som genomförs av privata institut är bortfallet ofta betydligt högre, men det är långt ifrån alltid som instituten uppger den typen av uppgifter.

Den viktigaste aspekten av bortfall är huruvida bortfallet är slumpmässigt eller systematiskt. I ett slumpmässigt bortfall skiljer sig inte svaren från respondenterna på ett systematiskt sätt från hur personerna som inte deltog i undersökningen skulle ha besvarat frågorna. Slumpmässigt bortfall är endast

ett problem därför att det minskar antalet observationer och därför leder till större slumpmässiga mätfel. Systematiskt bortfall innebär att benägenheten att delta i undersökningen samvarierar med hur man skulle besvara frågorna om man deltog. Systematiska bortfall medför därför att våra resultat blir snedvridna oavsett hur stora urval vi använder. Vi får ett systematiskt mätfel. Exempelvis har benägenheten att delta i opinionsundersökningar varit lägre bland Sverigedemokraternas väljare än bland övriga partiers sympatisörer, vilket medfört att de flesta opinionsinstitut har underskattat partiets stöd i väljarkåren.

Alla fyra typerna av bortfall ovan kan potentiellt vara av slumpmässig eller systematisk karaktär. Vanligen finns det dock skäl att tro att orsaken till att en person inte kan eller vill vara med i en undersökning samvarierar med vissa egenskaper som också påverkar hur man skulle besvara frågorna. Beroende på frågornas art kan denna samvariation vara mer eller mindre stark för de olika typerna av bortfall. Den som bortdefinieras p.g.a. sjukdom som innebär en långvarig sjukhusvistelse har troligen en annan syn på frågor rörande sjukvårdens kvalitet än befolkningen i stort. Kanske är däremot frånvaro av svar som beror på att enkätutskicket kom bort i postgången av mer slumpmässig karaktär. I princip måste dock en forskare alltid räkna med risken att ett bortfall leder till systematiska mätfel.

För att hantera problemen med systematiska bortfall används ofta någon typ av bortfallsanalys. En typ av bortfallsanalys innebär att vi anstränger oss för att få tag på ett slumpmässigt urval av bortfallet, exempelvis genom upprepade kontaktförsök eller ekonomiska incitament. Genom att ställa ett begränsat antal frågor till dessa kan vi bedöma om – och i sådana fall hur – bortfallet skiljer sig från dem som svarade. Det är dock dyrt och förmodligen återstår ändå en stor grupp av personer som aldrig kommer att delta. Ett vanligare angreppssätt är att identifiera grupper där bortfallet är stort och därefter undersöka om resultatet i dessa grupper avviker från resultatet i andra grupper. Ett grundläggande problem med denna typ av bortfallsanalys är att bortfallet alltid kan ske på andra grunder än de vi observerar. Om vi vet att bortfallet är systematiskt *mellan* grupper, så att exempelvis lågutbildade svarar i lägre utsträckning än högutbildade, men är slumpmässigt *inom* gruppen av lågutbildade, är det relativt enkelt att hantera. Då kan vi både upptäcka systematiken och åtgärda den genom att vikta upp svar från grupper med stort bortfall. De lågutbildade som besvarat enkäten får helt enkelt representera även de lågutbildade som inte besvarade enkäten. Men om systematiken sker längs andra linjer än vad vi observerar, så är risken stor att vi inte upptäcker det systematiska bortfallet och därför får svårt att hantera det. Kanske skiljer sig de lågutbildade som besvarade enkäten från de lågutbildade som avstod från att svara?

Vad är frågan och hur ställs den?

Det är en lika sann som uttjatad klyscha att ”som man frågar får man svar”. Frågeformulering, såväl som hur frågan ställs och av vem kan ge upphov till stora skillnader i svar.

Frågeformuleringar

Hur en fråga formuleras kan vara helt avgörande för hur respondenterna svarar. Även om formuleringen är särskilt viktig när det rör ämnen där de flesta inte har tagit tydlig ställning, finns det också en risk att frågeformuleringar påverkar våra svar i frågor där vi tror oss fatta övervägda beslut. Exempelvis var partierna oense om hur valsedlarna i folkomröstningen om EMU skulle utformas, eftersom de misstänkte att ordvalen kunde påverka utfallet. Till exempel föreslog Moderaterna att svenskarna skulle rösta ”ja eller nej till euron”, samtidigt som Vänsterpartiet ville att valsedeln skulle fråga om ”Valutaunionen EMU” och Miljöpartiet motsatte sig ”ja” och ”nej” som svarsalternativ. De ville istället att alternativen skulle vara att ”bibehålla svenska kronor” eller ”införa euro”. Anledningen till detta var förmodligen att *euron* är ett mer positivt laddat ord än *valutaunion* samt att forskning har visat att många är mer benägna att svara ja än att svara nej.

När man bedömer andras undersökningar kan man inte räkna med att den som genomfört undersökningen är särskilt intresserad av att återge en sanningsenlig bild. Att det är ett välkänt opinionsinstitut som genomfört undersökningen är ingen garanti för att de inte har använt ledande frågeformuleringar. I regel är det viktigare vem som har beställt undersökningen än vem som utför den. Det är då viktigt att både granska frågeformuleringen och vilka svarsalternativ som gavs till respondenten.

Men frågeformuleringar kan ställa till det även för ärliga forskare och uppsatsstudenter. När det gäller känsliga frågor, eller frågor där det finns tydliga normer för vilket beteende som är mest önskvärt, kan man nästan vara helt säker på att respondenterna inte svarar sanningsenligt. Vi vet till exempel att detta gäller frågor om kost- och träningsvanor, huruvida man röstade i det senaste valet samt vilket parti man i så fall röstade på.

Det finns flera strategier man kan använda för att i högre grad få sanninsenliga svar från respondenterna. För det första kan vi i frågeformuleringen berätta att beteenden som avviker från normen är vanligt. Kanske svarar människor mer ärligt om sitt valdeltagande ifall vi inleder frågan med ”många personer valde att inte rösta i det senaste valet, eftersom de inte tyckte att något parti representerade deras åsikter...”? För det andra svarar vi förmodligen mer sanningsenligt på konkreta frågor om hur många gånger vi tränade förra veckan, än på generella frågor om hur ofta vi brukar träna. För det tredje kan vi kombinera flera olika typer av data för att skapa oss en helhetsbild. Exempelvis rapporterar handeln om högre konsumtion av

alkohol och tobak än vad hushållen gör. För det fjärde finns det en mängd snillrika metoder som gör det omöjligt att identifiera vem som har svarat vad. En sådan metod kommer ni läsa om i texten om experiment.

Även den ordning i vilken frågor ställs kan ha stor påverkan på hur respondenterna svarar. Ett klassikt exempel på det återfinns i Hyman och Sheatsley (1950). De ställde dels en fråga om huruvida Sovjetunionen borde släppa in amerikanska journalister i landet och låta dem rapportera fritt från vad de observerar, men också den motsatta frågan om ifall USA borde släppa in sovjetiska journalister. Bland de respondenter som först fick frågan om amerikanska journalister i Sovjet var det 73 procent som ansåg att USA borde släppa in ryska journalister, men när frågorna ställdes i motsatt ordning var motsvarande siffra endast 36 procent. För att tidigare frågor inte ska påverka utfallet kan det därför vara en bra idé att alternera ordningen på frågorna och se om det har någon påverkan på resultaten.

De problem vi diskuterat ovan torde främst ge upphov till systematiska mätfel, men dåliga frågeformuleringar kan dessutom leda till mer slumpmässiga mätfel. Oklara svarsalternativ eller frågor som innehåller flera olika påståenden där respondenten endast instämmer i vissa kan tänkas innebära att slumpen får en större betydelse för hur folk svarar. Samtidigt kan vi sällan på förhand anta att en olycklig frågeformulering bara leder till slumpmässiga mätfel.

Intervjuer och formulär

Enkätdata insamlas i huvudsak genom intervjuer, post- eller webbenkäter. Intervjuerna kan ske öga mot öga eller över telefon. De olika insamlingsteknikerna kan påverka resultaten och innebära olika stora risker för systematiska mätfel (se t.ex. Newman m.fl. 2002).

Till att börja med ställer olika sätt att samla in data olika krav på respondenterna för att det överhuvudtaget ska vara möjligt för dem att svara på undersökningen. Post- eller webbenkäter kräver en distributionskanal i form av postgång och en fast adress, alternativt tillgång till internet och en e-postadress eller motsvarande. Tillgång till telefon för en telefonintervju är kanske mer spridd i vissa sammanhang. Besöksintervjun ställer i sig inga sådana tekniska krav, även om det förstås krävs ett sätt för att inledningsvis komma i kontakt med respondenterna ifråga.

En respondent erbjuds nästan alltid anonymitet i en enkät. I kraft av anonymiteten antas en respondent ge ärligare svar på känsliga frågor. Uppenbart upplevs dock anonymiteten annorlunda i en enkät som samlas in genom intervjuer – i synnerhet ifråga om besöksintervjuer – jämfört med självifyllda enkäter. Här har post- och webbenkäter en fördel.

Nackdelen med enkäter som respondenten fyller i själv är att forskaren har begränsade möjligheter att kontrollera om respondenten förstår frågorna och besvarar dem som tänkt. Det finns inte heller några möjligheter att förtydliga

en fråga vilket kan göra språkförståelse till en större faktor. Likaså kan det faktum att respondenten överhuvudtaget själv behöver läsa frågan också vara av betydelse. T.ex. kan läsförmågan variera mellan olika respondenter. Att respondenten själv läser frågan öppnar också upp för ”fuskläsning” och snabba och slarviga svar. Alla dessa faktorer kan ge upphov till systematiska mätfel då det är troligt att språkförståelse m.m. samvarierar med hur man besvarar undersökningens frågor. En del av dessa problem går att komma åt genom att ställa kontrollfrågor i enkäten vilka kan avslöja respondenter som inte läser (eller förstår) frågorna. Det ger åtminstone en fingervisning om problemets omfattning.

Att samla in data genom intervjuer kan ge en ökad kontroll över datainsamlingen för forskaren. Forskaren får feedback på hur själva insamlingen fungerar och frågor kan upprepas och förtydligas. Å andra sidan är en intervjusituation mer komplex än att fylla i en enkät och gör att många fler faktorer kan påverka. Genast kan alla möjliga personliga egenskaper vara av betydelse och hur olika respondenter reagerar på dessa. T.ex. intervjuarens språk och dialekt, sätt att betona frågor, kulturella skillnader och likheter mellan intervjuare och respondent m.m. Här finns klara risker för systematiska mätfel. För ett större datamaterial är det omöjligt för forskaren själv att genomföra alla interjvuer, om några överhuvudtaget. I dessa fall är det viktigt att intervjuarna är professionella och har klara instruktioner som borgar för att intervjuerna blir så lika som möjligt. För att det t.ex. ska vara en fördel att en intervjuare kan förtydliga en fråga så gäller det alla gör detta på ett likartat sätt och det oavsett respondentens egenskaper.

Checklista för systematiska mätfel

De möjliga källorna till systematiska mätfel är snarast oändliga och beror helt på en undersöknings karaktär. Diskussionen ovan har tagit fasta på några av de centrala riskerna som man oftast stöter på. Till skillnad från slumpmässiga mätfel finns det färre klara metoder för hur systematiska mätfel ska hanteras. Istället blir detta vanligen en fråga för forskaren att diskutera och bedöma utifrån den specifika undersökningen. Checklistan nedan kan utgöra en utgångspunkt för en sådan bedömning.

- Hur har urvalet gjorts?
  - Sannolikhetsurval eller inte?
  - I det senare fallet, hur pass stora är riskerna för ett snedvridet urval med den använda urvalsmetoden? Presenterar datainsamlaren någon information om urvalets representativitet med avseende på relevanta kända faktorer?
- Hur stort är bortfallet?

  - Finns det skäl att tro att de som inte svarat skulle svara annorlunda jämfört med dem som har svarat?
  - Genomför datainsamlaren någon form av bortfallsanalys? Resultat?
  - Kompenserar datainsamlaren för bortfallet genom viktning? Är det rimligt att tro att de vars svar viktas upp kan representera dem som inte svarat?

- Vad för fråga är det exakt som respondenterna har svarat på och vilka var svarsalternativen?
  - Kan formuleringen ha påverkat hur personer svarat? Kan i princip samma fråga ställas på ett annat, mer neutralt sätt, som torde ge andra svar?
  - Vilka har beställt undersökningen?
  - Rör frågan ett känsligt ämne där respondenterna kan ha skäl att inte svara sanningsenligt? Har datainsamlaren i sådana fall vidtagit några åtgärder för att underlätta sanningsenliga svar?

- Hur har undersökningen genomförts? Självifyllda enkäter eller intervjuer?
  - Har respondenterna fått vara anonyma?
  - Hur stor är risken att respondenterna missförstått frågan? Kan frågan vara särskild svårbegriplig för vissa respondenter?
  - Om intervjuer, vilka är det som genomfört intervjuerna och hur är deras relation till respondenterna? Risk för intervjuarefffekter?

Hantering av slumpmässiga mätfel

Alla urvalsundersökningar dras med slumpmässiga fel. Så länge vi inte intervjuar samtliga personer i den population vi är intresserade av, så kommer vårt urval att skilja sig något från populationen. Ju fler vi intervjuar, desto mindre blir denna skillnad, men vi kan aldrig undkomma problemet helt och hållet. Inte heller vet vi för ett enskilt urval om de slumpmässiga felen orsakat en underskattning eller överskattning. Som tur är kan vi använda statistisk teori för att åtminstone beräkna hur stora slumpmässiga fel vi kan räkna med. Men hur gör vi detta? Först måste vi bekanta oss med de tre nya begreppen standardfel, normalfördelning och säkerhetsnivå.

Standardfel

Vi kommer att använda flera olika metoder för att hantera statistisk osäkerhet, men alla metoder bygger på något som kallas för standardfel. Standardfelet

uttrycker hur god precision vår undersökning har och är ungefär samma sak som hur stora de slumpmässiga felen skulle vara i genomsnitt om vi drog ett oändligt antal urval. Ju större slumpmässiga mätfel, desto större standardfel. Det bästa sättet att minska standardfelet i en urvalsundersökning är i regel att öka storleken på urvalet.

Låt oss anta att vi genomför en opinionsmätning där vi tillfrågar 500 personer om vilket parti de skulle rösta på om det var riksdagsval idag. Låt oss vidare anta att 30 procent av svenskarna är socialdemokrater. Standardfelet för andelen socialdemokrater skulle i detta fall vara ungefär 2 procentenheter (den beräkningen går vi igenom på föreläsningen om generaliseringar). I frånvaro av systematiska mätfel innebär det att om vi drog ett oändligt antal urval, skulle dessa i genomsnitt avvika med 2 procentenheter från 30 procent. I de flesta urval skulle mellan 28 och 32 procent av respondenterna uppge att de röstar på Socialdemokraterna, men ibland skulle det vara färre än 28 procent eller fler än 32 procent. Problemet är förstås att vi i regel endast gör ett urval och inte vet hur stor avvikelsen var i just detta urval. För att veta exakt hur vanligt det är med olika avvikelser från populationens medelvärde måste vi anta att urvalen följer en normalfördelning.

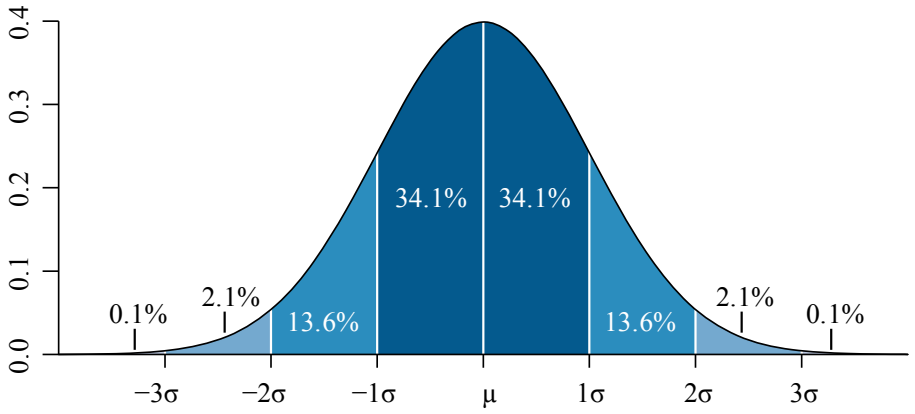
Normalfördelningen

Om vi drar ett oändligt antal urval kan vi vanligtvis anta att medelvärdena för urvalen fördelar sig enligt en normalfördelning (eller enligt den närbesläktade t-fördelningen).[2] När man ritar utseendet på en sannolikhetsfördelning brukar man visa de möjliga värdena på den horisontella axeln och hur vanliga de olika värdena är på den vertikala axeln. Det kännetecknande för normalfördelningen är att den kan beskrivas med hjälp av fördelningens medelvärde, μ, och standardavvikelse, σ. Detta illustreras i figur 1. Eftersom det i just detta fall handlar om en fördelning av urval, kallas fördelningens standardavvikelse även för standardfel.[3]

Som ni ser påminner normalfördelningen om en kulle. Att den är högst vid μ innebär att de mest sannolika utfallen är att urvalets medelvärde ligger nära populationens medelvärde. Som de mörkblå fälten i figur 1 visar, ligger medelvärdet i 68,2 (34,1+34,1) procent av urvalen inom ett standardfel från populationens medelvärde. Om vi summerar alla fält mellan -2σ och 2σ ser vi att urvalets medelvärde i 95,4 ($2\times$(13,6+34,1)) procent av urvalen befinner sig inom 2 standardfel från populationens medelvärde.

[2] Att vi kan göra detta antagande beror på den s.k. *centrala gränsvärdessatsen*. Ett matematiskt teorem som ni kan läsa en kortare beskrivning av i Teorell & Svensson s. 130-132.

[3] Mer precist uttryckt är standardfelet samma sak som standardavvikelsen för de olika urvalens medelvärde, vilket är ungefär samma sak som hur mycket de olika urvalen i genomsnitt avviker från medelvärdet (hur man beräknar en standardavvikelse går vi igenom på föreläsningen om beskrivande statistik).



Figur 1: Normalfördelningen

Om vi känner till urvalens standardfel kan vi beräkna hur sannolika olika urval är. Normalfördelningen har nämligen den egenskapen att medelvärdet i 90 procent av alla urval befinner sig mindre än 1,65 standardfel från populationens medelvärde, i 95 procent av urvalen befinner sig mindre än 1,96 standardfel från populationens medelvärde och i 99 procent av alla urval befinner sig mindre än 2,58 standardfel från medelvärdet. 1,65, 1,96 och 2,58 kallas för kritiska värden och ni kommer att stöta på dem många gånger under kursen.

Med andra ord, för varje möjligt urval är sannolikheten 90 procent att medelvärdet i urvalet avviker mindre än 1,65 standardfel från populationens medelvärde. Eller omvänt uttryckt, i 90 procent av fallen kommer populationens medelvärde ligga mindre än 1,65 standardfel från urvalets medelvärde. Det innebär att om vi beräknar ett intervall runt urvalets medelvärde, som spänner från 1,65 standardfel över medelvärdet till 1,65 standardfel under medelvärdet, kommer populationens medelvärde återfinnas inom detta intervall för 90 procent av urvalen. Sådana intervall kallas för *konfidensintervall*.

Konfidensintervall

Konfidensintervall är en metod för att kvantifiera den osäkerhet som uppstår när vi generaliserar punktestimatet i ett urval till att gälla en hel population. Låt oss anta att vi vill ta reda på hur stor andel av svenskarna som skulle rösta på Socialdemokraterna om det vore val idag. Eftersom vi inte kan tillfråga samtliga svenskar måste vi göra ett urval och utifrån det urvalet uttala oss om populationen. Genom att beräkna ett konfidensintervall kan vi dra slutsatser som att ”medelvärdet i populationen befinner sig mellan A (konfidensintervallets lägre ändpunkt) och B (konfidensintervallets övre ändpunkt).”

Vi kan emellertid inte vara helt säkra på att populationens medelvärde återfinns i konfidensintervallet, utan vi måste alltid acceptera en viss risk för att vi drar ett missvisande urval. När vi avgör hur stor risk vi är beredda att acceptera brukar vi säga att vi väljer säkerhetsnivå. Exempelvis innebär en säkerhetsnivå på 95 procent att, om vi drog ett oändligt antal urval och beräknade ett konfidensintervall för varje urval, skulle 95 procent av konfidensintervallen innehålla populationens medelvärde. Ju högre säkerhetsnivå, desto större kritiskt värde. Det ger i sin tur ett bredare konfidensintervall och en lägre risk att populationens medelvärde befinner sig utanför intervallet.

Låt oss återgå till vårt tidigare exempel, där vi frågade ett slumpmässigt urval svenskar om vilket parti de skulle rösta på om det var val idag. I exemplet var andelen socialdemokrater i Sverige 30 procent och standardfelet för urvalen var 2 procentenheter. Om vi genomförde ett oändligt antal undersökningar, skulle medelvärdet i 90 procent av urvalen ligga mellan 26,7 (30-2×1,65) och 33,3 (30+2×1,65) procent. För dessa 90 procent av urvalen skulle också ett konfidensintervall innehålla populationens medelvärde. Fem procent av urvalen skulle dock ha ett högre medelvärde än 33,3 procent och för dessa urval skulle konfidensintervallet börja ovanför 30 procent. Fem procent av urvalen skulle ha ett lägre medelvärde än 26,7 procent och för dessa urval skulle konfidensintervallet sluta under 30 procent. Med andra ord skulle vi i 10 procent av fallen beräkna konfidensintervall som inte omfattar populationens medelvärde. Om vi i stället sätter säkerhetsnivån till 95 procent, och därför beräknar ett konfidensintervall som sträcker sig 1,96 standardfel åt varje håll, skulle endast 5 procent av konfidensintervallen inte innehålla populationens medelvärde. För att minska denna andel till 1 procent måste konfidensintervallet sträcka sig 2,58 standardfel från urvalets medelvärde. Det är samma sak som att vi sätter säkerhetsnivån till 99 procent.

Statistisk signifikans

Med hjälp av ett vanligt konfidensintervall kan vi uttala oss om exempelvis populationsmedelvärdet för en variabel. Men minst lika vanligt är att vi vill uttala oss om hur värdena på en variabel beror på värdena på en annan variabel. I de enklaste fallen vill vi jämföra medelvärdet i två olika grupper eller vid två olika tidpunkter. Då kanske vi drar ett urval från varje grupp, eller från varje tidpunkt, och jämför sedan de två urvalen. Men kan vi vara säkra på att de två grupperna är olika, eller kan skillnaden mellan urvalen bero på slumpen?

När media rapporterar om opinionsundersökningar kommenterar de ofta huruvida en förändring är *statistiskt säkerställd*. I vetenskapliga sammanhang använder vi oftare begreppet *statistiskt signifikant*, men begreppen betyder samma sak. Vanligtvis innebär det att vi med en viss bestämd säkerhet kan utesluta möjligheten att någonting är noll. En skillnad mellan två grupper är statistiskt signifikant, eller statistiskt säkerställd, om skillnaden mellan

urvalen är större än vad vi tror kan bero på slumpen. Det innebär i så fall att skillnaden mellan dessa grupper i populationen inte kan vara noll. På samma sätt är en förändring mellan två tidpunkter statistiskt säkerställd om skillnaden mellan två urval dragna vid två olika tidpunkter är större än vad slumpen rimligtvis skulle ge upphov till.

På kursen kommer vi undersöka om skillnaden mellan två grupper är signifikant genom att beräkna ett konfidensintervall för hur stor skillnaden mellan grupperna är i populationen. Om värdet 0 återfinns i intervallet, innebär det att vi inte kan utesluta möjligheten att de två grupperna inte skiljer sig från varandra. Detta gäller på samma sätt för om vi beräknar skillnaden mellan två medelvärden som skillnaden för två proportioner. Vi kan helt enkelt inte utesluta att skillnaden i medelvärde eller proportion mellan grupperna är noll. Men om värdet 0 inte återfinns i intervallet, då är vi tämligen säkra på att de två grupperna har olika medelvärde även i populationen. Det är samma sak som att skillnaden är statistiskt signifikant eller statistiskt säkerställd. Skillnaden mellan urvalen var helt enkelt större än vad som rimligtvis kan bero på slumpmässiga mätfel.

När vi senare på kursen lär oss regressionsanalys kommer vi prata om statistiskt signifikanta effekter. Då menar vi att värdena på den beroende variabeln samvarierar så kraftigt med värdena på den oberoende variabeln att det inte kan vara en slump. Vanligtvis argumenterar vi då för att samvariationen beror på att den oberoende variabeln påverkar den beroende variabeln, men mer om det senare.

Tolka signifikans förnuftigt

Många, även etablerade forskare, lägger så stor vikt vid att enstaka resultat är statistiskt signifikanta att deras tolkningar av resultaten blir missvisande. I det här avsnittet diskuterar vi några av de vanligaste fallgroparna. Eftersom problematiken är vanligast i samband med regressionsanalyser rekommenderar vi att ni läser denna del av texten först inför seminarium 3 eller 4.

Signifikanstest löser inte allt

Konfidensintervall och signifikanstester är bra verktyg för att hantera slumpmässiga mätfel. De hjälper oss emellertid inte med att hantera systematiska mätfel, omvänd orsaksriktning eller andra typer av problem vi stöter på i kvantitativa studier. Om en enkätstudie har stora systematiska mätfel, kan därför konfidensintervallet vara direkt missvisande. Och när vi genomför regressionsanalyser säger statistisk signifikans ingenting om orsaksriktningen (sambandet skulle förmodligen vara signifikant även om vi bytte plats på den oberoende och den beroende variabeln) eller huruvida vi har kontrollerat för

alla relevanta variabler. Statistisk signifikans innebär endast att sambandet förmodligen inte beror på slumpen.

Glöm inte det substantiella

Att ett resultat är statistiskt signifikant betyder inte automatiskt att det är intressant. Oavsett om det handlar om skillnaden mellan två grupper eller den estimerade effekten i en regressionsanalys, så innebär statistisk signifikans endast att skillnaden eller effekten vi intresserar oss för inte är noll. Även om effekten är större än noll, kan den ju fortfarande vara så liten att den är försumbar. Eller ännu värre, det vi undersökte kanske inte ens är särskilt intressant.

"Men det är ju signifikant" är helt enkelt inget universalsvar på frågor om ens resultat är betydelsefulla. Signifikans är en sak, men man bör också kommentera storleken på de estimerade skillnaderna eller effekterna, så att läsaren kan bedöma hur pass intressanta resultaten är.

Det är viktigt att vara medveten om att möjligheterna till att få signifikanta resultat i hög grad beror på hur mycket data man har. Med större datamaterial så kommer, allt annat lika, standardfelen att bli mindre och det gör det lättare att finna signifikanta skillnader eftersom även mindre skillnader kan skiljas från slumpen. Det innebär, grovt sagt, att minsta lilla skillnad blir signifikant med tillräckligt stora stickprov.

Frågan är alltså inte bara om ett resultat är signifkant eller inte utan om det är av *substantiellt* intresse. Vad som egentligen är "substantiellt intressant" är nu ingen enkel fråga och vi kommer komma tillbaka till den under kursen. Svaret på frågan handlar dock oftast om att jämföra ens resultat med existerande forskning för att se om de innebär ett betydande bidrag. Tyvärr brukar både etablerade forskare och studenter glömma detta.[4]

Tolka icke-signifikanta resultat försiktigt

Om en skillnad mellan två grupper är statistiskt signifikant, innebär det att skillnaden inte är noll. Men om motsatsen inträffar, och en skillnad inte är signifikant, innebär det inte att skillnaden är noll. Det innebär endast att vi inte kan utesluta möjligheten att skillnaden är noll. När detta glöms bort orsakar det ofta tolkningar som saknar stöd i empirin. Här tar vi upp två exempel på sådana missförstånd.

För det första bör vi aldrig tolka icke-signifikanta effekter eller skillnader som att de är noll, utan vi måste alltid utgå ifrån konfidensintervallet för att se hur stora effekter som faktiskt är möjliga. Det är bara om endast försumbara effekter ryms inom intervallet som vi kan tolka effekten som försumbar. Låt oss till exempel anta ni har låtit 100 slumpmässigt valda svenskar svara på en enkät om inkomster. Därefter beräknar ni ett konfidensintervall för skillnaden

[4]Vi hänvisar den nyfikne till artikeln *Size matters* av Ziliak och McCloskey.

i inkomst mellan män och kvinnor. Om skillnaden inte är statistiskt signifikant betyder inte det att svenska män och kvinnor tjänar lika mycket. Det innebär endast att er undersökning inte kan *utesluta möjligheten* att kön är orelaterat till inkomst. Med andra ord, er undersökning visar inte att skillnaden i inkomst mellan män och kvinnor är noll utan att skillnaden inte kan vara större än konfidensintervallets gränser (givet den aktuella säkerhetsnivån). Intervallets ändpunkter visar alltså hur stora inkomstskillnader som ni kan utesluta baserat på resultaten i er studie.

För det andra är skillnader i statistisk signifikans sällan särskilt intressanta. Om variabel A har en signifikant och positiv effekt, medan variabel B inte är statistiskt signifikant, betyder inte det nödvändigtvis att variabel A har en större effekt än variabel B. För att återanvända exemplet i det förra stycket kan vi tänka oss att vi vid en första tidpunkt hittar en signifikant inkomstskillnad mellan könen, medan vi inte gör det vid en senare tidpunkt. Detta innebär att vi bara vid den första tidpunkten kan vara säkra på att män och kvinnor i genomsnitt tjänar olika mycket. Det innebär inte att vi vet att inkomstskillnaden mellan män och kvinnor har förändrats mellan de två tidpunkterna. Det kan nämligen vara så att en förändring av storleken på skillnaden mellan två grupper som *i sig* inte är signifikant, mycket väl kan påverka huruvida skillnaden som sådan mellan grupperna är signifikant. [5]

Många jämförelsepunkter

Eftersom vi inte kan använda 100 procents säkerhetsnivå löper vi alltid viss risk att felaktigt identifiera statistiskt säkerställda skillnader och samband. Vid 95 procents säkerhetsnivå är sannolikheten att vi hittar ett statistiskt signifikant samband, eller en statistiskt signifikant skillnad, 5 procent, givet att det inte finns någon skillnad eller samband i populationen.

En risk på 5 procent kan tyckas försumbar, men de verkliga problemen uppstår så fort signifikanta resultat ges större uppmärksamhet än icke-signifikanta resultat. Det kan hända om en forskare testar ett stort antal hypoteser och bara redovisar de resultat som var statistiskt signifikanta, men vanligare är kanske att såväl traditionella medier som vetenskapliga tidskrifter hellre rapporterar om statistiskt signifikanta resultat. Låt oss använda ett enkelt räkneexempel för att visa på hur missvisande våra statistiska test då kan bli. I Sverige har vi åtta riksdagspartier och sju stora opinionsinstitut, vilket innebär att det produceras 56 mätningar per månad. Om det inte har hänt någonting i opinionen sedan föregående månad, är risken för att en enskild mätning felaktigt indikerar en statistiskt säkerställd förändring endast 5 procent, givet en säkerhetsnivå på 95 procent. Men sannolikheten för att minst en av dessa 56 mätningar ska hitta en statistiskt säkerställd förändring,

[5] Gelman och Stern skriver bra om detta i artikeln *The Difference Between "Significant" and "Not Significant" is not Itself Statistically Significant.*

trots att ingen förändring har skett, är ca 94 procent.[6] Risken är förstås överhängande att mediarapporteringen kommer fästa större uppmärksamhet vid denna statistiskt säkerställda förändring än vid övriga 55 mätningar och därigenom ge en falsk bild av vad som hänt i opinionen. Det mest välkända exemplet på en sådan bias kommer från serien xkcd och återges här på sidan intill.

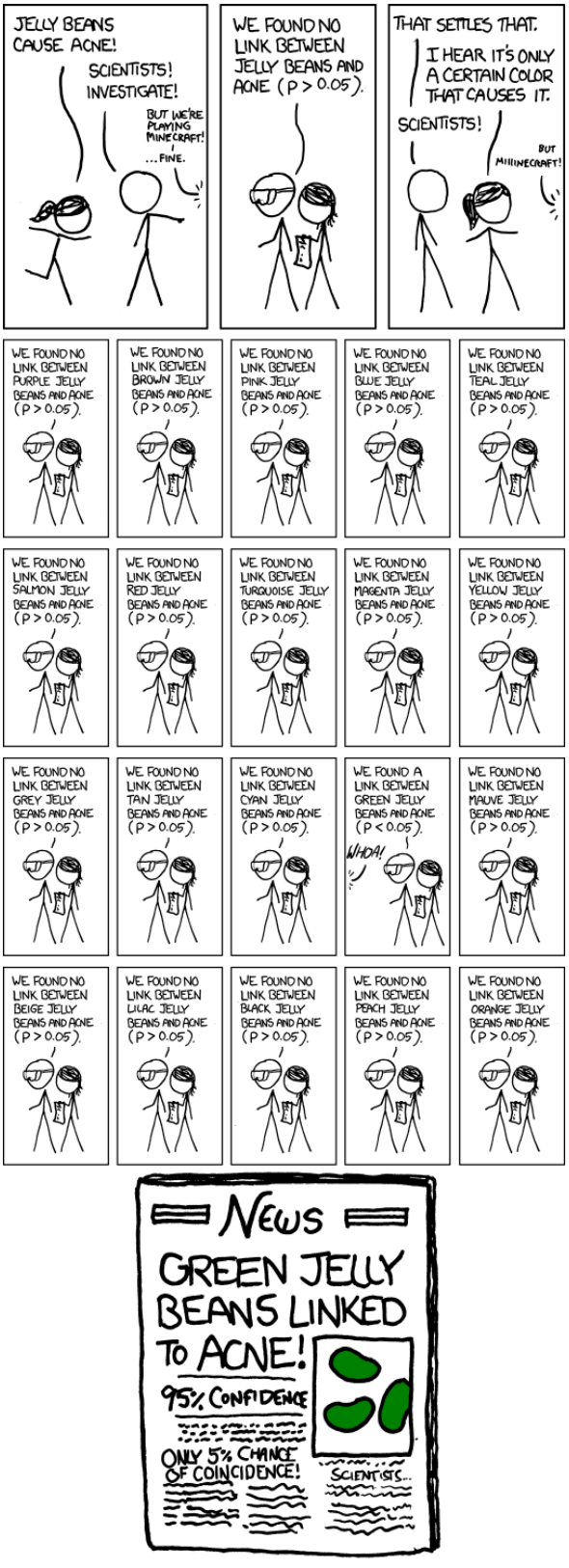
Publiceringsbias

Problematiken kring många jämförelsepunkter är nära förknippad med vad som brukar kallas *publiceringsbias* i vetenskapliga sammanhang. I vetenskapssamhället är det lättare att få genomslag för signifikanta resultat än motsatsen, ofta i form av en ny förklaringsfaktor till ett fenomen. Det slår helt enkelt högre att säga ”titta här, vi har missat den här faktorn som kan förklara i vilken utsträckning personer röstar” än att visa att en viss faktor inte har någon betydelse för personers röstningsbeteende. Vanligen får då huruvida resultaten är signifikanta eller inte utgöra måttstock för om detta är en intressant förklaringsfaktor och frågan om substantiellt intresse glöms mer eller mindre bort. Givet att signifikanta resultat uppmärksammas och publiceras i högre grad så finns det en omfattande risk för att den aktuella forskningen ger en snedvriden bild av kunskapsläget. Den existerande publicerade forskningen kan alltså bestå av en överdrivet stor andel signifikanta resultat som kan riskera att överskatta olika förklaringars betydelse för ett fenomen. Detta brukar kallas publiceringsbias.[7]

Problemet är omvittnat och har många orsaker som går i flera led, från den enskilda forskaren till forskningsfinansiärerna. För en forskare handlar det vanligtvis i slutändan om att kunna publicera sina resultat i vetenskapliga tidskrifter. Eftersom det är lättare att få gehör för nya signifikanta förklaringar är de också vanligtvis lättare att publicera i högt rankade vetenskapliga tidskrifter. Tidskrifterna vill också ha artiklar som uppmärkammas och citeras och har därför även själva intresse av “följa strömmen” och premiera signifikanta resultat. I många vetenskapliga discipliner kan också externa kommersiella motiv ha stor betydelse, om än kanske inte så ofta inom samhällsvetenskapen. Det mest klassiska exemplet är läkemedelsföretag som har stora intressen i att visa att nya preparat faktiskt är verkningsfulla. Det betyder både ett intresse av att finansiera studier som ger signifikanta

[6] För varje mätning är sannolikheten för att förändringen inte är statistiskt säkerställd 95 procent. Att detta ska upprepas för samtliga 56 mätningar är ca 6 procent ($0.95^{56} = 0.06$).

[7] Det finns en ”motrörelse” som vill lyfta fram icke signifikanta resultat, ofta kallade nollresultat (”null result” på engelska). Det har t.ex. tagit sig utryck genom särskilda tidsskrifter som bara publicerar nollresultat, t.ex. psykologitidsskriften *Journal of Articles in Support of the Null Hypothesis*. Andra har uttryckt en mer generell skepsis mot användandet av signifikanstester. Se t.ex Ziliak & McCloskey 2008. En annan psykologitidsskrift har t.o.m. helt avskaffat signifikanstester, *Basic and Applied Social Psychology* (Trafimow & Marks, 2015).


JELLY BEANS CAUSE ACNE!
SCIENTISTS! INVESTIGATE!
BUT WE'RE PLAYING MINECRAFT!
...FINE.
WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE (P > 0.05).
THAT SETTLES THAT.
I HEAR IT'S ONLY A CERTAIN COLOR THAT CAUSES IT.
SCIENTISTS!
BUT MIIIINECRAFT!
WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLUE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN SALMON JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN RED JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TURQUOISE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN MAGENTA JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE (P < 0.05).
WHOA!
WE FOUND NO LINK BETWEEN MAUVE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE (P > 0.05).
NEWS
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE
ONLY 5% CHANCE OF COINCIDENCE!
SCIENTISTS...

resultat och att sedan få dessa publicerade. Omvänt är det färre kommersiella medicinska forskningsfinansiärer som vill finansiera studier i syfte att visa att en substans *inte* har någon verkan eller som har intresse i att sådana studier publiceras. Sammantaget innebär detta att många aktörer inom vetenskapssamhället har starka intressen för att premiera signifikanta resultat framför icke signifikanta.

Detta innebär, något tillspetsat, att om en forskare gör 1 000 analyser och finner 10 signifikanta resultat och 990 icke signifikanta så är det de där 10 som blir publicerade och uppmärksammade. Det är dem som vi får ta del av i vetenskapliga tidskrifter. De 990 icke signifikanta resultaten hamnar istället på forskarens sophög. Det är något att ha i åtanke när ett signifikant resultat lyfts fram, även om det så är signifikant på 99 procents säkerhetsnivå. Har det föregåtts av 99 andra analyser kan det vara slumpen som spelar oss ett spratt.

Det finns inga enkla lösningar på den här typen av problem, men vi vill ge er två rekommendationer. För det första måste risken för att p.g.a. slumpen få signifikanta resultat tas på allvar, även om 95 eller 99 procents säkerhetsnivå används. Det innebär att vi alltid måste tolka enskilda mätningar och studier med stor försiktighet. Om upprepade replikationer visar samma sak som originalstudien, eller om flera olika opinionsmätningar visar liknande tendenser, ligger det förmodligen något i resultaten. För det andra pekar problemet på vikten av att basera hypoteser i teoretiska resonemang och att stärka de statistiska beläggen med andra former av argument och evidens. Med goda belägg för ett samband, som vilar på mer än statisk signifikans, kan vi minska risken för att vårt forskningsresultat bara beror på slump.

Referenser

Gelman, A., & Stern, H. (2006). "The difference between 'significant' and 'not significant' is not itself statistically significant". *The American Statistician*, 60(4), 328-331.

Holmberg, S. & O. Petersson. (1980). *Inom felmarginalen: En bok om politiska opinionsundersökningar*. Stockholm: Liber förlag.

Hyman, H. H., och Sheatsley, P. B. (1950). *The current status of American public opinion*.

Martinsson, Dahlberg & Lundmark (2013). *Is Accuracy Only For Probability Samples? Comparing Probability and Non-probability Samples in a Country with Almost Full Internet Coverage*. Conference paper, AAPOR conference in Boston.

Newman, J. C., Des Jarlais, D. C., Turner, C. F., Gribble, J., Cooley, P., & Paone, D. (2002). "The differential effects of face-to-face and computer interview modes". *American Journal of Public Health* 92 (2):294-7.

Trafimow, D. & M. Marks (2015). Editorial, *Basic and Applied Social*

Psychology, 37:1, 1-2,

Ziliak, S. T., & McCloskey, D. N. (2004). Size matters: the standard error of regressions in the American Economic Review. *The Journal of Socio-Economics*, 33(5), 527-546.

Ziliak, S. T., & McCloskey, D. N. (2008). *The cult of statistical significance: How the standard error costs us jobs, justice, and lives*. University of Michigan Press.