Lögn och statistik

Per Flensburg

Professor i informationssystem

I en föregående artikel har jag berättat om John Ioannidis som rent matematiskt har visat att de flesta slutsatser vid hypotesprövning med kontrollgrupper är felaktiga. I denna artikel ska jag berätta om några vanliga statistiska misstag jag stött på som granskare av vetenskapliga artiklar och konferensbidrag. Jag avgränsar mig till hypotesprövningar baserade på antingen enkäter eller experiment med kontrollgrupp. Slutsatsen är att det finns många fallgropar som är lätta att falla i.

För litet urval

Ett vanligt fel är att för få objekt undersöks. Många tror att några tiotal är tillräckligt. Det kan det iofs vara, t.ex. om den totala populationen är under 100 objekt och man vill ha en uppskattning av förekomsten av en viss egenskap hos populationen. Men med alltför få objekt blir statistiken lätt missvisande eftersom slumpmässiga variationer ger större utslag. Hanne Kjöller ger ett exempel i sin bok där hon berättar om Amy Cuddys experiment där försökspersonerna fick omväxlande sträcka på sig och sjunka ihop (Kjöller, 2020). Cuddy påstod att två minuter i endera positionen ökade eller minskade testosteronhalten och kortisol och därmed också förändrade riskbenägenheten. Men hon hade bara 42 personer hon testade på. När experimentet gjordes om med 200 personer kunde man inte påvisa någon statistisk skillnad.

Icke representativt urval

Ett annat vanligt fel är att urvalet inte är representativt. Om man t.ex. vill undersöka ungdomars drogberoende är studenter inte en representativ grupp. Här är det viktigt man gör klart för sig hur man vill indela målgruppen t.ex. efter kön, inkomst, ålder, bostadsort, utbildning etc. Indelningen beror på vad man vill undersöka och vilka variabler man vill ta hänsyn till. Detta i sin tur beror på den teoretiska bakgrunden, dvs den teori man vill testa. Ett intressant fall är de svenska väljarundersökningarna. SIFO anses vara mest tillförlitlig trots att man tillfrågar ca 800 personer. De andra instituten som gör väljarundersökningar frågar betydligt men deras urval är inte lika representativt. Det fanns till och med en som gjorde undersökningen på internet och missade därmed de 10% som på den tiden inte hade tillgång till internet.

Låg svarsprocent

Det är väldigt vanligt att svarsprocenten är väldigt låg. Man tycks tro att det är antalet svar som är det viktiga. Jag granskade en gång en artikel där man hade sänt ut 3 000 enkäter och var väldigt stolt över att man fått in 300. Mitt avslag kom förmodligen som en chock! Man kan dock inte fastlägga att en viss svarsprocent är tillräcklig. Om man i mitt exempel hade gjort en bortfallsanalys och visat att bortfallet var representativt hade 10% varit tillräckligt. Men en sådan analys är inte lätt att göra. Om personen ifråga inte vill besvara enkäten är det inte troligt att de vill svara på frågan om varför de inte svarar. Hög svarsprocent är för övrigt den största förklaringen till att SIFO:s väljarundersökningar är mer tillförlitliga än andra. Man

lägger ner betydligt mer möda på att få varje person att besvara undersökningen. De övriga instituten frågar i princip en ny person om de saknar något svar. Men det är inte säkert att denna hör till samma grupp som den ursprunglige.

Plocka russin ur kakan

Detta kallas cherry-picking på engelska och innebär att man med ett statistikprogram undersöker sambanden mellan samtliga variabler i undersökningen och så presenterar man de intressantaste med vanligtvis en signifikansnivå på 95%. Men 5% av de framtagna sambanden är felaktiga. Det innebär att om du slumpmässigt väljer ett samband så är det korrekt med 95% sannolikhet. Men om du väljer ut det intressantaste sambandet så är det inget slumpmässigt urval och du kan inte säga ett dugg om dess korrekthet. Du måste göra en helt ny undersökning av just detta samband.

Ett nära besläktat fenomen är att man i efterhand gör olika indelningar av undersöknings-materialet för att få fram ett samband. Antag att vi undersöka om en viss medicin sänker blodtrycket och undersökningen visar detta inte är korrekt. Då kan man undersöka män och kvinnor för sig, ger inte heller detta något resultat kan man göra åldersindelningar och hjälper inte det kan man t.ex. avgränsa till medelålders män som har gikt. Där kan man kanske påvisa ett samband. Men återigen: Undersökningen var designad för att undersöka en hel population, den indelning man gjort efterhand i syfte att visa på ett samband kullkastar statistiska slutsatser.

Studien är inte reproducerbar

Ett kännetecken på den typ av undersökningar vi diskuterar här att en annan forskare ska kunna upprepa undersökningen och komma till samma resultat. Då är undersökningen reproducerbar. Man kan tänka sig att en annan forskare får tillgång till alla enkätsvar, hur urvalet gått till och vilka som svarat. Detta är dock väldigt sällan fallet för forskare tenderar att hålla hårt på sitt grundmaterial. Ett undantag är dock arbetet med att fram vaccin mot covid-19, där en sällan skådad delning av grundmaterial och preliminära slutsatser ägde rum. Visserligen var många undersökningar undermåliga men hellre riskera det än att vänta två år på att Lancet & C:o skulle publicera de goda undersökningarna.

För att en studie ska vara reproducerbar måste den metod man använt beskrivas rigoröst och utförligt. Vid många tillfällen finns dock kraftiga sidbegränsningar och det finns det inte plats till i de 10 sidor forskaren har till förfogande. En sådan redovisning gör att studien inte blir reproducerbar.

Fokus på frågor

Jag har tidigare varit inne på att frågorna i en enkät ska relatera till den teori som forskaren vill testa. Väldigt ofta är denna koppling osynlig, man får intrycket att forskaren ställt samman ett antal vanliga frågor och så lagt den eller de som berör det vederbörande vill testa. I praktiken innebär det att man mäter en massa variabler som inte har med problemet att göra och frestelsen att hitta några russin i denna kaka är överhängande.

Mäta rätt sak

Säg vi vill undersöka om ett visst preparat sänker blodtrycket. Undersökningen visar inte det, men däremot kan man se att det sänker kolesterolhalten. Eftersom högt blodtryck kan hänga

samman med hög kolesterolhalt, så presenteras effekten av preparatet som sänkt kolesterolhalt. Men detta var inte vad man undersökte och urval och mätmetoderna är inte designade för denna slutsats. Vad man måste göra är att göra en ny undersökningen med just denna hypotes.

Om vi fortsätter med samma exempel så kan det tänkas man har en kontrollgrupp som får sockerpiller. Låt oss vidare anta att studien visar att det testade preparatet har effekt. Det sänker blodtrycket med 95% sannolikhet. Men är det detta som är det viktiga? Är det inte viktigare att veta om preparat A är bättre än preparat B? Borde man inte testa mot andra preparat istället för mot sockerpiller?

En annan egenskap är undersökningens styrka. Den är en kombination av urval och skillnad mellan hypotes och nollhypotes. Ta återigen exemplet med blodtrycksmedicin. Om det är stor sänkning av blodtrycket ökar det undersökningens styrka. Om man har använt många testpersoner ökar det också styrkan.

Låt oss fortsätta med blodtrycksexemplet. Låt oss anta att preparatet har en del biverkningar, t.ex. att fötterna svullnar så man inte kan ha vanliga skor. Det gör att folk inte vill använda medlet även om det är bra och i praktiken har man då undersökt något som är oanvändbart.