Statistiska termer som används i forskningsstudier: En lathund för journalister

Peter M. Dahlgren, JMG

2017-03-15

Vid bedömning av akademiska studier konfronteras journalister inte bara av siffror, utan också begrepp som "p-värde", "statistisk inferens" och "regression".

Statistikkurser finns på de flesta universitet, men ofta är det något som man tvingas genomlida, snabbt går igenom och sedan lika snabbt glömmer bort. Men för både journalister och kommunikatörer är det absolut nödvändigt att göra mer än att bara läsa forskningsstudiens sammanfattning. Man bör förstå de metoder och begrepp som ligger till grund för akademiska studier för att kunna bedöma forskningen. Även om man inte behärskar statistik fullt ut, är det ändå viktigt att ha grundläggande kunskaper för att åtminstone kunna formulera bättre och mer kritiska frågor till experter, och kanske mer specifikt att bli mer skeptisk till resultaten i studierna.

Samband och orssakasamband

De flesta studier försöka upprätta ett **samband** mellan två variabler, till exempel hur läraren kan ha ett *samband* med studenternas betyg. Eller hur vikten på en bil har ett *samband* med dödsolyckor.

Men att upptäcka ett sådant förhållande är bara ett första steg. Det yttersta målet är att fastställa **orsakssamband**, att den ena av de två variablerna *påverkar* den andra, och inte tvärtom.

Det är en vanlig fras som kan vara bra att tänka på, nämligen att **korrelation är inte kausalitet**. Eller ännu enklare: *samband är inte nödvändigtvis orsakssamband* eftersom orsakssamband tar lång tid att fastställa.

Det finns flera viktiga kritier för att bedöma om ett orsakssamband existerar:

- *Nära i tid.* Förändringen i den ena variabeln kan inte påverka den påverkade variabeln allt för långt senare.
- *Nära i rum.* Förändringen i den ena variabeln kan inte vara allt för geografiskt avlägset den påverkade variabeln.

- *Logisk koppling.* Man måste teoretiskt kunna argumentera för att den ena variabeln faktiskt orsakar en förändring i den påverkade variabeln.
- *Regelbundenhet.* Förändringen i den ena variabeln bör upprepade gånger leda till förändringar i den påverkade variabeln.
- *Utesluta andra förklaringar.* Det bör inte finnas några andra förklaringar som kan ge upphov till förändringen i den påverkade variabeln.
- *Orsaken måste föregå effekten.* Förändringen i den ena variabeln måste ske *före* den påverkade variabeln förändras.

Uttala sig om datan eller bortom datan

En annan viktig skillnad att komma ihåg är att studier kan antingen utforska observerade data (**deskriptiv statistik**) eller använda observerade data för att förutsäga vad som är sant *bortom* den observerade datan (**inferentiell statistik**).

Uttalandet "Från 2005 till 2015 har antalet anmälda stöldbrott ökat med 70 %" är deskriptiv statistik. Denna typ av uttalande är inte förknippad med särskilt mycket osäkerhet. Uttalande av detta slag är ofta antingen sant eller falskt.

Uttalandet "Om du tar en högskoleexamen ökar din livsinkomst med 50 %" är inferentiell statistik. Denna typ av uttalande är *alltid* förknippat med en osäkerhet eftersom uttalandet inte bara rör de som har undersökts, utan också människor i framtiden och kanske i andra länder.

Grundläggande statistiska begrepp

Här är några andra grundläggande statistiska begrepp som journalister och kommunikatörer bör känna till:

- Ett **sampel** (också kallat **urval**) är en del av en hel **population**. Inferentiell statistik försöka göra förutsägelser om en population baserad på resultaten som observerats i ett urval av denna befolkning.
- Det finns två huvudtyper av urval: **slumpmässiga** och **icke-slumpmässiga**. Vid ett slumpmässigt urval väljs personer helt av en slump, medan ett icke-slumpmässigt urval ofta är konstruerade för att återspegla egenskaperna hos befolkningen i stort (man väljer personer så att man får hälften kvinnor och hälften män, utifrån deras ålder, etnisk tillhörighet, etc.). Det finns flera olika typer av slumpmässiga respektive icke-slumpmässiga urval, var och en med sina fördelar och nackdelar. Ett slumpmässigt urval är **representativa** för befolkningen, vilket innebär att forskaren undersöker en minikopia av populationen.
- När man har analyserat ett sampel och vill uttala sig om populationen utifrån detta sampel så kallas det att man **generaliserar**. Detta kan

endast göras när samplet verkligen är representativt för hela befolkningen, det vill säga slumpmässigt.

- När man generaliserar resultat från ett sampel till populationen måste man ta hänsyn till **variansen** i urvalet. Varians är ett annat ord för variation. Även om samplet är helt slumpmässigt, finns det fortfarande en viss variation inom populationen som kommer att kräva att dina sampel också innehåller en **felmarginal**. Till exempel kan resultaten av en opinionsundersökning av väljare ge felmarginalen i procentenheter: "47 % av de tillfrågade röstar på Socialdemokraterna, med en felmarginal på 3 procentenheter." Så, den faktiska andelen som röstar på Socialdemokraterna kan vara så låg som 44 % eller så hög som 50 % (vilket är ±3 procentenheter från 47 %).
- Ju större **sampelstorlek** (hur många som har tillfrågats i undersökningen), desto säkrare kan vi vara på resultatet. En bra tumregel att ha i huvudet: Om det är runt 1 000 personer i en undersökning blir felmarginalen högst ±3,1 procentenheter (givet att det är ett slumpmässigt urval).
- De flesta studier undersöker förhållandet mellan två **variabler**, till exempel att exponering för bekämpningsmedel har ett samband med lägre födelsevikt. Detta kallas den **alternativa hypotesen**. Studier försöka ofta motbevisa **nollhypotesen**, i detta fall att exponering för bekämpningsmedel inte har något samband med lägre födelsevikt.
- **Signifikanstest** visar sannolikheten att resultatet är så extremt att det sannolikt inte beror på slumpen. Då får man fram ett **p-värde** som anger hur extremt resultatet är, under förutsättningen att nollhypotesen är sann (att det inte finns något samband). En vanlig tolkning är följande: om p-värdet är under 0,05, är det 5 % sannolikhet att få så extrema resultat om nollhypotesen är sann. Om p-värdet är 0,01, är det 1 % sannolikhet att få så extrema resultat om nollhypotesen är sann, och så vidare.
- Ett vanligt problem inom forskningsstudier är så kallad **bias**. Bias kan översättas med *fel*, *skevhet* eller *partiskhet* och kommer i många former, men den vanligaste är valet av svarspersoner. Om svarspersoner inte väljs slumpmässigt, utan får själva välja om de vill svara på en opinionsundersökning, är samplet inte slumpmässigt och därmed nödvändigtvis inte generaliserbart till populationen.
- När två variabler rör sig tillsammans, sägs de vara **korrelerade**. **Positiv korrelation** innebär att om en variabel stiger eller faller, gör den andra variabeln också det. Till exempel, om du äter mycket mat ökar din vikt. Matintag och vikt är då positivt korrlerade. **Negativ korrelation** innebär att två variabler rör sig i motsatta riktningar. Till exempel fordonshastighet och restid. Ju högre fordonshastighet, desto mindre restid. Så om en forskare skriver "inkomst är negativt korrelerad med fattigdom," säger forskaren att när inkomster stiger, så minskar fattigdom.
- **Kausalitet** är när förändringen i en variabel *orsakar* förändring i en annan variabel. Till exempel är lufttemperaturen och solljus korrelerade (när solen är uppe, stiger temperaturen), men orsakssamband går ndast i en riktning. Det är solljuset som *orsakar* temperaturförändringen. Detta

kallas också orsak och verkan, eller ibland också prediktor och effekt.

- **Regressionsanalys** är ett sätt att avgöra sambandet mellan många variabler och hur starka dessa samband kan vara. Även om många forskare skriver att flera variabler *påverkar* en variabel finns det dock ingenting i regressionsanalys som sådant som kan visa på orsakssamband. I sin mest grundläggande form består regressionsanalys av två variabler som kan visas i ett digram på varsin axel.
- **Korrelationskoefficient** är ett mått på sambandets styrka (och ibland också riktning). Det varierar från 0 till 1 (och ibland -1 till +1 om det finns riktning). När korrelationskoefficienten är närmare 0 säger man att det inte finns något samband, så kallat **nollsamband**.
- Medan orsakssamband är väldigt lätta att visa, är det betydligt svårare att visa vad som är orsaken till sambandet. Det kan i själva verket vara ett skensamband, ett så kallat **spuriöst samband**. Till exempel, de som är gifta tenderar att ha högre lön. Man skulle då kunna dra slutsatsen att giftemål leder till högre lön. Men det är snarare så att äldre personer har både högre lön och mer oftare är gifta.
- När man har etablerat att det finns ett orsakssamband, så är den faktor som driver förändringen den **oberoende variabeln**. Den variabel som påverkas är den **beroende variabeln**. I exemplet tidigare är alltså solljus den oberoende variabeln, medan lufttemperatur är den beroende variabeln.
- **Standardavvikelse** ger en inblick i hur mycket variation det finns inom en grupp av värden. Standardavvikelsen visar avvikelsen (skillnaden) från gruppens medelvärde. Standardavvikelsen är ett komplement till **medelvärdet** och **medianen**. Medelvärdet och medianen visar mittpunkten i en grupp av värden, medan standardavvikelsen visar hur stor variationen är från detta medelvärdet. Medelvärdet får du fram genom att plussa ihop alla värden och dividera med det antalet värden. Medianen får du fram genom att rangordna alla värden från det lägsta till det högsta, och sedan välja det värdet som är i mitten. Låt säga att medellönen är 20 000 kronor. En låg standardavvikelse innebär då att de flesta löner ligger runt medelvärdet på 20 000 kronor. En hög standardavvikelse innebär då att många löner är utspridda från medelvärdet på 20 000 kronor.
- Var uppmärksam på skillnaden mellan **procenttal** och **procentenheter**. De är inte samma sak. Till exempel, om 40 av 100 hem i en stad har låg inkomst, är andelen 40 % som har låg inkomst. Om 10 husägare får högre inkomst, återstår nu bara 30 hem med låg inkomst. Den nya andelen är då 30 %, en minskning med 10 *procentenheter* (40 - 30 = 10). Detta är dock *inte* 10 % mindre. I själva verket är minskningen 25 % (10 / 40 = 0,25 = 25 %).
- **Kvartiler** kan användas för att dela in data i grupper. Till exempel, att dela en lista över personer sorterade efter inkomst i två grupper (låginkomsttagare och höginkomsttagare) resulterar i två kvartiler. Medianen blir därmed skiljelinjen mellan de två grupperna. Vanligtvis delar man in datan i fyra grupper (*kvartiler*) så att det blir fyra lika stora grupper. Man kan också dela upp datan i grupper om tio (deciler). När en forskare säger att

> "De personer med en inkomst i den lägre decilen" så menar de personer som har en inkomst som är bland de lägsta 10 % personernas inkomst. Om man delar in datan i grupper om hundra så säger man *percentiler*. När en forskare säger att "De personer med en inkomst i den övre 83 percentilen" så menar de personer har en inkomst som är bland de övre 83 % personernas inkomst.

Notera att en förståelse för statistiska termer inte innebär att du bör salta din text med dem. Skriv på ett enkelt sätt som kan förstås av så många som möjligt, utan statistiskt fackspråk. "Samband" fungerar ofta lika bra som "korrelation". Det underlättar även för forskare!

Läs mer

- Statistics and Probability Dictionary
- Introduction to Statistics: Inference

Referenser

Denna text är inspirerad från Statistical terms used in research studies: A primer for media. Alla kredd till Leighton Walter Kille som skrev den artikeln, vilken också är licensierad med Creative Commons Attribution 3.0 Unported. I min text har jag dock ändrat de faktafel som fanns i den texten (bland annat den vanliga missupfattningen att p-värdet visar att nollhypotesen är sann).