Statistiska termer som används i forskningsstudier: En lathund för journalister

Peter M. Dahlgren, JMG

2017-05-21

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

Statistikkurser finns på alla universitet, men dessvärre är det något som man snabbt går igenom och sedan lika snabbt glömmer bort. Men för både journalister och kommunikatörer är det 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 forskning. 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, men även för att förhålla sig skeptisk till resultaten i studierna.

Samband och orssakasamband

De flesta studier försöka upprätta ett **samband** mellan två saker (så kallade **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 saken *påverkar* den andra (A leder till B), och inte tvärtom (B leder till A). Steget från att bara se ett samband till att också säga att det är ett orsakssamband är väldigt långt.

Ibland säger man **korrelation är inte kausalitet**. Det betyder att samband inte nödvändigtvis är orsakssamband. Det betyder att två saker kan ha ett samband utan att den ena saken orsakar den andre.

Det finns flera kriterier för att bedöma om ett orsakssamband existerar. Här är några vanliga:

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

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

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 beskriver den data som har undersökts, utan också försöker säga något om människor i allmänhet, i framtiden eller kanske i andra länder.

Grundläggande statistiska begrepp

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

- En **population** är den grupp man vill säga någonting om. Det kan vara vad som helst: alla tidningsartiklar i Dagens Nyheter, svenska befolkningen eller studenter som har gått naturvetenskapliga programmet i Sävsjö. I opinionsundersökningar är det vanligt att populationen är "svenska folket", men ska man vara petig är det oftast "svenska medborgare som är 18 år eller äldre".
- 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 urvalet.
- Det finns två huvudtyper av urval: **slumpmässiga** och **icke-slumpmässiga**. Vid ett slumpmässigt urval väljs personer som ska delta slumpmässigt, medan ett icke-slumpmässigt urval ofta är konstruerade för att återspegla egenskaperna hos populationen. Man ser

till att hälften kvinnor och hälften män deltar, att lika många gamla som unga deltar, och så vidare. Det finns flera olika typer av slumpmässiga respektive icke-slumpmässiga urval, var och en med sina fördelar och nackdelar. Slumpmässiga 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** eller gör en **inferens** till populationen. Detta kan endast göras när samplet verkligen är representativt för hela populationen, alltså när det är ett slumpmässigt urval.
- När man generaliserar resultat från ett sampel till populationen måste man ta hänsyn till **variansen** i urvalet. Varians är ett annat ord för variation. Varje gång ett sampel tas slumpmässigt från en population, finns det en också en viss slumpmässig variation i samplet jämfört med populationen. Genom att beräkna en **felmarginal** eller **konfidensintervall** kan man undersöka hur stor denna slumpmässiga variation är, och på så vis få reda på var någonstans det sanna värdet finns. 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." Den faktiska andelen som röstar på Socialdemokraterna kan då vara så låg som 44 % eller så hög som 50 % (vilket är ±3 procentenheter från 47 %). Felmarginalen är då 3 procentenheter. Konfidensintervallet är då 44 till 50 procent.
- Ju större **sampelstorlek** (hur många som har tillfrågats i undersökningen, antalet nyhetsartiklar etc.), desto mindre blir felmarginalerna. Då kan vi också vara säkrare 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.
- De flesta studier undersöker förhållandet mellan två **variabler**, till exempel att exponering för bekämpningsmedel har ett samband med lägre födelsevikt.
- **Signifikanstest** är ett statistiskt test som används för att förkasta slumpmässiga skillnader. Man kan enklast jämföra det med en rättegång. Utgångspunkten är att den åtalade är oskyldig. Det är sedan åklagarens ansvar för att lägga fram tillräckligt mycket bevis mot denna utgångspunkt att personen är osykldig. Signifikanstestet fungerar precis likadant. Utgångspunkten är att det inte finns något samband, vilket kallas **nollhypotesen**. Det man vill testa är om det finns ett samband, vilket kallas den **alternativa hypotesen**.
- Från signifikanstestet får man fram ett **p-värde** som anger hur extremt resultatet är, under förutsättningen att nollhypotesen är sann (att det inte finns något samband). Om p-värdet är 0,05, är det 5 % sannolikhet att få så extrema resultat givet slumpen (att det inte finns något samband). Om p-värdet är 0,01, är det 1 % sannolikhet att få så extrema resultat givet slumpen, och så vidare. Om resultatet är tillräckligt extremt, förkastar man nollhypotesen och får stöd för den alternativa hypotesen.

- 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. Om forskaren får välja fritt finns det också möjligheten att forskaren väljer det som passar teorin.
- Ett annat vanligt problem är **problemet med flera jämförelser**. Varje gång man analyserar samma data med statistik ökar sannolikheten att hitta samband. Detta leder till att den vetenskapliga litteraturen överdriver antalet samband som faktiskt existerar.
- När två variabler rör sig tillsammans, sägs de vara **korrelerade**. En **positiv korrelation** innebär att om en variabel stiger eller faller, gör den andra variabeln också likadant. 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 endast i en riktning. Det är solljuset som *orsakar* temperaturförändringen. Detta kallas i dagligt tal för orsak och verkan.
- **Korrelationskoefficient** är ett mått på sambandets styrka och vanligtvis också riktning. Det varierar vanligen från -1 till +1. När korrelationskoefficienten är närmare 0 säger man att det inte finns något samband, så kallat **nollsamband**. Ju närmare korrelationskoefficienten är -1, desto mer negativ korrelation. Ju närmare +1, desto mer positiv korrelation.
- **Regressionsanalys** är ett sätt att kvantifiera sambandet mellan många variabler, se och hur pass starka dessa samband kan vara. Fördelen med regressionsanalys är att man kan se hur mycket varje enskild variabel bidrar. Om man vill veta hur mycket utbildning påverkar inkomst kan man även se hur mycket kön påverkar inkomst. Man säger då att man tar "tar hänsyn till kön" eller "kontrollerar för kön" i sin analys. Även om många forskare skriver att flera variabler *påverkar* en variabel finns det dock ingenting i regressionsanalys som sådant som kan visa orsakssamband. I sin mest absolut mest grundläggande form består regressionsanalys av två variabler.
- 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. Flera underhållande exempel finns

på webbplatsen spurious correlations.

- När man har etablerat att det finns ett orsakssamband, eller när man letar efter 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** visar variationen från ett medelvärde. **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 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 att de flesta löner ligger runt medelvärdet på 20 000 kronor. En hög standardavvikelse innebär att många löner är utspridda från medelvärdet på 20 000 kronor.
- **Procenttal** och **procentenheter** 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 fyra lika stora grupper. Man kan också dela upp datan i grupper om tio (**deciler**). När forskare säger att de undersökt "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 forskare säger att de undersökt "personer med en inkomst i den 99:e percentilen" så menar de personer som har en inkomst som är högre än 99 % av alla andra personernas inkomst.

Notera att en förståelse för statistiska termer inte innebär att du bör krydda 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
- How do you know a paper is legit?

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 Killes text (bland annat den vanliga missupfattningen att p-värdet visar att nollhypotesen är sann).