Föreläsning 4: Beskrivande statistik

Pär Nyman

4 september 2015

Både föreläsning 4 och 5 innehåller en del matematik. På Studentportalen finns därför några sidor med räkneövningar, vilka riktar riktar sig till personer som inte tidigare har läst någon statistik och som även har svårt för eller upplever ett motstånd mot matematik. Vi vill som vanligt understryka att matematiken aldrig är det viktigaste, men att vi erbjuder räkneövningar för att det är där som era förkunskaper skiljer sig mest åt.

1 Att göra beskrivningar

Beskrivningar utgör fundamentet i samhällsvetenskapen. Det är genom beskrivningar vi har strukturerat världen för att göra den greppbar och förståelig. Öppnar man en samhällsvetenskaplig lärobok är den förmodligen full av beskrivningar, såsom välfärdsstatstypologier, demokratiindex och BNP-jämförelser. Goda beskrivningar är dessutom en förutsättning för att kunna genomföra förklarande studier. Utan god kännedom om de fenomen vi studerar kan vi inte gärna analysera hur de påverkar varandra.

Oavsett vilken typ av beskrivningar man gör handlar det till stor del om att reducera information. Exempelvis väljer vi kanske att kalla kristdemokraterna för ett socialkonservativt parti, snarare än att rada upp alla deras ställningstaganden i olika frågor. Det betyder inte att Kristdemokraterna är exakt samma sak som alla andra socialkonservativa partier, eller att de i varje avseende är just socialkonservativa.

Sådana kategorier behandlade vi på förra föreläsningen. Idag ska vi prata om hur vi reducerar information med hjälp av beskrivande statistik. Vad ska vi tänka på när vi reducerar den komplicerade politiska situationen i Ryssland till en åtta på en tiogradig demokratiskala eller summerar prisutvecklingen för alla varor i Sverige med att inflationen under 2012 var 0,9 procent?

2 Skalnivåer

Skalnivåer anger hur en variabels variabelvärden förhåller sig till varandra. Anledningen till att vi bryr oss om det är att skalnivåerna avgör vad vi kan utföra för typ av analyser med den data vi har. Jag kommer idag att prata

om fyra skalnivåer: nominalskala, ordinalskala, intervallskala och kvotskala. Det är vad jag upplever som vanligast och jag tror att en del av er har stött på denna uppdelning tidigare. Eftersom man kan använda intervallskalor till det mesta man kan göra med kvotskalor, gör Teorell och Svensson ingen åtskillnad mellan dessa skalnivåer utan använder ordet intervallskala om båda nivåerna. Det är helt upp till er om ni vill följa kursboken och prata om tre skalnivåer eller göra som mig och dela upp variablerna i fyra olika skalnivåer.

Nominalskalan är den första skalnivån. Variabler på denna skalnivå kallas ibland för kvalitativa eller kategoriska variabler. Definitionen av nominalskalan är att variablerna har värden vi inte kan rangordna, såsom yrke (snickare, lärare, polis), inriktning på en utbildning (samhällsvetenskaplig, humanistisk, naturvetenskaplig) eller arbetsmarknadsstatus (arbetslös, sysselsatt, ej i arbetskraften). Ett annat sätt att uttrycka samma sak är att det handlar om artskillnader och inte gradskillnader. Vilken skalnivå variabeln befinner sig på kan också variera med sammanhanget. För en fysiker är det kanske självklart att rangordna färger efter ljusets våglängd, medan färger för en samhällsvetare är tydliga nominalskalor. Eftersom vi på en nominalskala inte kan rangordna de observerade värdena kan vi inte heller säga vilket värde som befinner sig i mitten – vi kan alltså inte beräkna en median. Vi kan däremot säga vilket värde som är vanligast och därmed ange ett typvärde.

Nästa skalnivå kallas för ordinalskala och kräver att man kan rangordna variabelvärdena, men att avståndet mellan dem inte är konstant. Vanliga exempel på ordinalskalor är utbildningsnivå (förgymnasial, gymnasial, kandidat, master) och svaren på många enkätfrågor (t.ex. varje dag, varje vecka, varje månad).

Om vi inte bara kan rangordna variablerna, utan dessutom kan anta att avståndet mellan de möjliga variabelvärdena är konstant, då har vi antingen en intervallskala eller en kvotskala. Det som skiljer dem åt är att kvotskalor, till skillnad från intervallskalor, har en absolut nollpunkt.

Rena intervallskalor är ovanliga. Det vanligaste exemplet är temperatur mätt i Celsius, men även årtal befinner sig på en intervallskala. Notera att vi inte kan prata om relativa skillnader när variablerna befinner sig på en intervallskala. 24 grader är inte tre gånger så varmt som 8 grader. Och när Sverige spelade 2–2 mot England i 2006 års världsmästerskap i fotboll, skedde inte det dubbelt så sent som den danske kungen Sven Tveskägg invaderade samma land, vilket han gjorde år 1003. Det som gör intervallskalor intressanta, trots att de är så ovanliga, är att många variabler är så lika en intervallskala att vi kan hantera dem *som om* de vore intervallskalor. Vi antar då att avståndet mellan variabelvärdena är konstant. Så länge antagandet inte är helt orimligt, och avstånden därför bör vara ungefär lika stora, är detta i regel ganska oproblematiskt.

För att vi ska kunna prata om relativa skillnader, såsom ”hälften så mycket” eller ”50 procent högre”, krävs det att variabeln befinner sig på en

Tabell 1: De fyra skalnivåerna

Skalnivå	Egenskaper och exempel på variabler
Nominalskala	Kan ej rangordnas
	Kön, yrke, favoritfilm
Ordinalskala	Kan rangordnas men ej avståndsbedömas
	Utbildningsnivå, många enkätfrågor
Intervallskala	Kan avståndsbedömas men saknar absolut nollpunkt
	Temperatur i Celsius, årtal
Kvotskala	Kan avståndsbedömas och har absolut nollpunkt
	Alla antal och andelar

kvotskala. Ett annat sätt att beskriva kvotskalor är att de, utöver att de kan rangordnas och avståndsbedömas, har en absolut nollpunkt. Med absolut nollpunkt menas att värdet 0 är naturligt bestämt; att det betyder just total frånvaro av något i en absolut mening. Exempelvis innebär en förmögenhet på 0 en total frånvaro av pengar och temperaturen 0 på kelvinskalan innebär total frånvaro av termisk energi. Det är inte samma sak som att värdet aldrig kan bli negativt, även om de två ofta sammanfaller. Exempelvis kan resultatet i en årsredovisning vara negativt, trots att variabeln befinner sig på en kvotskala där vi kan prata om ”dubbelt så stor vinst som föregående år”. Detsamma gäller BNP-tillväxten, vilket är en kvotskala som antar negativa värden när ekonomin befinner sig i en recession. De flesta skalor vi kan avståndsbedöma (i strikt mening) är kvotskalor, såsom längd, tid, arbetslöshet, antal och andelar.

Många menar att distinktionen mellan intervallskala och kvotskala är oviktig. Jag vill ändå nämna att det krävs en kvotskala för att man ska kunna använda räknesätten multiplikation och division inom skalan, beräkna vissa centralitets- och spridningsmått som till exempel geometriskt medelvärde, variationskoefficient och percentilkvot samt studera relativa samband såsom elasticiteter. Det mesta av detta är visserligen sådant som inte lärs ut på kursen, men ändå tillräckligt viktigt för att motivera en distinktion mellan de två skalnivåerna.

Notera att en variabel inte behöver befinna sig på intervall- eller kvotskala bara för att den har variabelvärden som är siffror eller för att det är praktiskt möjligt att beräkna ett medelvärde. Många nominalskalor är kodade med siffror, men bara för att det är praktiskt möjligt att beräkna ett medelvärde betyder det inte att det är en bra idé.

2.1 Dikotoma variabler

En variabel som bara kan anta två olika värden brukar kallas för dummyvariabel, binär variabel eller dikotom variabel. Kärt barn har många namn. En

Tabell 2: Dela upp en kategorisk variabel i dummyvariabler

Facktillhörighet		LO	TCO	SACO	Annat
LO-medlem		1	0	0	0
TCO-medlem	$\Rightarrow$	0	1	0	0
SACO-medlem		0	0	1	0
Annat/Osäker		0	0	0	1
Ej medlem		0	0	0	0

anledning till dummyvariablernas popularitet är att de kringgår problemen med skalnivåer. Eftersom de bara har ett skalsteg – skillnaden mellan det ena och det andra värdet – är det ett oproblematiskt antagande att alla skalsteg är lika stora.

Låt oss anta att vi har en variabel kvinna som antar värdet 1 för kvinnor och värdet 0 för män. Trots att detta är en nominalskalevariabel kan vi beräkna ett medelvärde, vilket i detta fall motsvarar andelen kvinnor. Vi kan också undersöka hur en ökning av variabeln med ett skalsteg – alltså att vara kvinna i stället för man – påverkar värdet på en annan variabel. Hur det går till kommer vi att prata mer om när vi kommer in på regressionsanalysen.

Även variabler som har flera naturliga kategorier kan omvandlas till dummyvariabler, exempelvis med syfte att inkludera dem i en regressionsanalys (detta gäller förstås även kön, eftersom det inte är en självklar dikotomi). Antalet dikotoma variabler som behövs är alltid en mindre än antalet kategorier i den ursprungliga variabeln (könsvariabeln hade två kategorier och då räckte det med en variabel). Låt oss anta att vi har en variabel som mäter facktillhörighet, vilken kan anta värdena *LO-medlem*, *TCO-medlem*, *SACO-medlem*, *Annat/Osäker* och *Ej medlem*. Detta är en typisk nominalskalevariabel. Vill vi använda variabeln i exempelvis en regressionsanalys måste vi därför omvandla den till fyra dummyvariabler (den kategoriska variabeln kan anta fem olika värden). Tabell 2 visar hur det skulle kunna gå till. Den översta raden visar variablerna och de övriga raderna visar variabelvärden. Vi går alltså från en variabel med fem möjliga värden till fyra variabler som alla har två möjliga värden.

2.2 Några ord om antaganden

Som forskare uttrycker vi ofta att vi gör antaganden, i synnerhet när vi gör kvantitativa undersökningar. Vi har diskuterat några antaganden ovan, som att alla skalsteg på en skala är lika stora, och vi kommer göra ännu fler antaganden framöver. Men vad menar vi egentligen med ett antagande? Betyder det att vi tror att det är exakt så verkligheten ser ut?

De flesta statistiska metoder vi använder förutsätter att vissa antaganden är sanna, för att metoden ska ge helt korrekta resultat och erbjuda de

statistiska egenskaper som gjort metoden populär. I regel är det emellertid inget stort problem om dessa antaganden inte är helt korrekta, så länge avvikelsen är så liten att den endast marginellt påverkar resultaten. Det är i allmänhet också så att när vi utför formella test av våra antaganden, räcker det inte med marginella avvikelser för att vi ska avfärda antagandet som felaktigt. Vi vet därför sällan om våra antaganden är helt korrekta. När vi gör ett antagande menar vi således, att medan modellens egenskaper i strikt mening förutsätter att antagandet stämmer, tror vi som forskare endast att antagandet är tillräckligt nära verkligheten för att inte snedvrida resultaten alldeles för mycket.

Ytterligare en aspekt av antaganden handlar om hur man tror att resultaten skulle påverkas om antagandet inte håller. Låt oss anta att vi gör ett tveksamt antagande, som om det är fel kommer få ett samband att framstå som svagare och mindre signifikant än vad det egentligen är. Om vi trots detta hittar en signifikant effekt, kan vi i viss mån försvara oss med att sambandet i själva verket kanske är ännu starkare. Om vi i stället inte hittar någon effekt, skulle vi ha svårt att övertyga andra om resultatet eftersom det skulle kunna bero på det tveksamma antagande vi gjort. Men även om resonemangen påminner lite om logiken bakom kritiska fall – vi får ett starkare argument om vi ger vår hypotes svåra förutsättningar – bör vi i regel undvika att göra orimliga antaganden.

Avslutningsvis bör det poängteras att god forskningstradition föreskriver att antaganden motiveras och testas samt att forskaren även redovisar hur känsliga resultaten är för de antaganden som gjorts.

3 Beskrivande statistik

Som jag argumenterade i inledningen, handlar beskrivningar i hög utsträckning om att reducera information. Om någon ber oss att beskriva svenskarnas inkomster, svarar vi förhoppningsvis inte med en lista över alla svenskar och deras taxerade inkomster. I stället skulle vi ta fram statistik på medel- eller medianinkomsten samt något mått på hur jämnt eller ojämnt inkomsterna är fördelade. Detta skulle vara enklare att greppa, praktiskt hanterbart och underlätta jämförelser över tid eller med andra länder. Utmaningen ligger i att uppnå detta utan att så mycket information går förlorad att beskrivningen blir missvisande.

När vi beskriver en fördelning på det här viset använder vi oss av centralitets- och spridningsmått. Centralitetsmått anger det typiska eller mest representativa värdet i en fördelning, vilket kan handla om exempelvis det vanligaste värdet eller ett genomsnitt av samtliga värden. Spridningsmått anger hur långt ifrån varandra observationerna ligger.

3.1 Centralitetsmått

Typvärdet är det enklaste centralitetsmåttet och anger det vanligaste värdet. Fördelen med typvärdet är att det kan beräknas oavsett vilken skalnivå en variabel befinner sig på. Typvärdet är användbart när antalet observationer är stort i förhållande till antalet möjliga variabelvärden eller när det finns ett värde som av någon anledning är särskilt vanligt. Då kan typvärdet tolkas som det mest sannolika värdet. Om antalet observationer är litet i förhållande till antalet möjliga variabelvärden, vilket fallet nästan alltid är när vi använder kontinuerlig data, är typvärdet ett godtyckligt mått som man med fördel undviker. Typvärde kallas även för modalvärde.

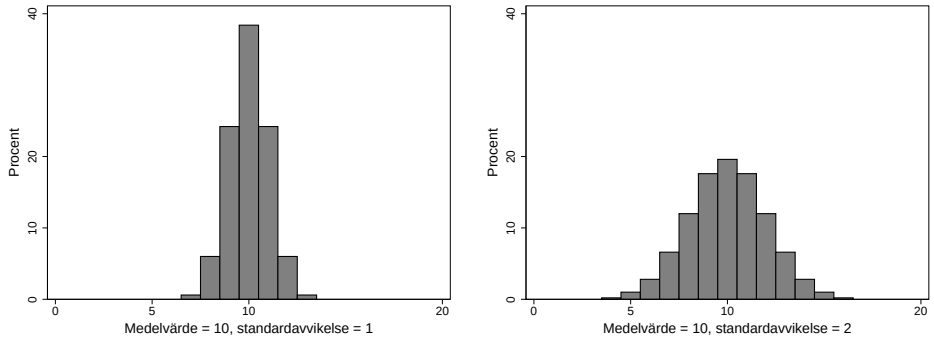
Medianen anger det mittersta värdet i en fördelning. Om man sorterar alla värden efter storlek hittar vi medianen i mitten. Ett annat sätt att göra samma sak är att om vartannat stryka det största och det minsta värdet tills bara ett värde återstår. Då har man funnit medianen. Med andra ord befinner sig alltid halva fördelningen över medianen och den andra halvan under medianen, om vi bortser från de observationer med samma värde som medianen. Om variabeln har ett jämnt antal värden beräknas medianen som medelvärdet av de två mittersta värdena.

Medelvärdet är samma sak som det genomsnittliga värdet i en fördelning. Det får vi genom att summera samtliga analysenheters värden och därefter dela på antalet analysenheter. Medelvärdet tar alltså hänsyn till samtliga värden och kan till skillnad från medianen därför påverkas av extrema värden.

Så vilket centralitetsmått är att föredra? I de flesta fall är typvärdet ett sämre alternativ än de två övriga måtten och används därför huvudsakligen när vi arbetar med nominalskalor eller som komplement till medianen på en ordinalskala. Ett viktigt undantag är när det finns vissa specifika värden som är mer sannolika än andra värden. Då kan typvärdet vara ett intressant mått. Medianen och medelvärdet är identiska om fördelningen är symmetrisk, men så fort den är skev åt något håll – alltså har en svans med höga eller låga värden som inte motsvaras av en liknande svans i den andra änden av fördelningen – kommer de två måtten att ge olika svar. Vilket mått som är mest lämpligt beror då helt på sammanhanget, men ofta är det bra att ange båda två. Det enda som är självklart är att ju mer skev fördelningen är, desto viktigare blir valet av centralitetsmått.

3.2 Spridningsmått

Figur 1 visar två fördelningar som trots att de ser väldigt olika ut har samma värden på de olika centralitetsmåtten. Anledningen är att de två fördelningarna har olika spridning. I det vänstra diagrammet är spridningen liten och de flesta observationer ligger nära varandra och medelvärdet. I det högra diagrammet är spridningen större, vilket betyder att observationerna ligger längre ifrån varandra.

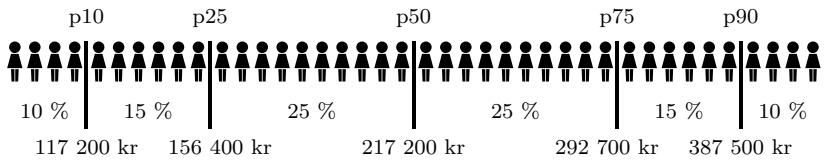


Figur 1: Två fördelningar med olika spridning

Ibland är det rent av spridningen vi är intresserade av, som när vi analyserar inkomstskillnader i olika länder. Men även om vi saknar ett direkt intresse av spridningen är den viktig för oss. Vet vi att spridningen är låg, kanske vi kan göra trovärdiga generaliseringar trots ett relativt litet urval. Men om spridningen är större måste vi intervjua fler personer eller undersöka fler kommuner för att kunna uttala oss om hela populationen.

Ett sätt att beskriva spridningen i ett datamaterial är att ordna alla värden från lägst till högst och sedan ange värden från olika delar av fördelningen. Observationen som har ett högre värde än exakt x procent av alla observationer kallas för percentil x. Exempelvis är den 50e percentilen samma sak som medianen. Den 25e percentilen kallas ibland för den första kvartilen, eftersom exakt en fjärdedel (en kvart) av observationerna har ett lägre värde. På samma sätt kan vi kalla den 50e och den 75e percentilen för den andra och den tredje kvartilen.

Nedan visas ett exempel på hur den svenska befolkningens disponibla inkomster är fördelade (data är från 2013, inkluderar kapitalvinster och avser årsinkomst per konsumtionsenhet). Längst till vänster är personen med lägst inkomst och längst till höger är personen med högst inkomst. Den tionde percentilen är uppmätt till 117 200 kr, vilket innebär att tio procent av befolkningen har mindre än 117 200 kr per år att röra sig med. Medianinkomsten är ungefär dubbelt så stor: 217 200 kr.



Vi har nu reducerat alla svenskars inkomster till ett fåtal tal, men även detta kan vara för mycket information om vi vill studera utvecklingen över tid eller

jämföra ett stort antal länder. Vi sammanfattar då gärna spridningen i ett mått.

Valet av spridningsmått beror på flera saker. För det första måste vi fundera på om vi är mest intresserade av absoluta eller relativa skillnader. ”Lisa tjänar 1000 kr mer än Kalle” är ett exempel på en absolut skillnad, medan ”Lisa tjänar 10 procent mer än Kalle” är ett exempel på en relativ skillnad. Om vi skulle fördubbla alla värden i en datamängd skulle de absoluta spridningsmåtten indikera en ökad spridning (Lisa tjänar då 2000 kr mer än Kalle), medan relativa spridningsmått skulle förbli oförändrade (Lisa tjänar fortfarande 10 procent mer än Kalle). Vilken egenskap i spridningsmåttet som är mest önskvärd beror helt på vad vi studerar.

För det andra måste vi fundera på vilka delar av fördelningen som är viktigast för spridningen. Vill vi att alla observationer ska vägas in, eller finns det skäl att exempelvis studera spridningen i mitten av fördelningen? För det tredje har vissa spridningsmått värdefulla statistiska egenskaper, vilket i synnerhet gället standardavvikelsen, men det kommer vi inte prata så mycket om på den här kursen.

I min föreläsning pratar jag om fyra olika spridningsmått: percentilavstånd, percentilkvot, standardavvikelse och variationskoefficient. Percentilavstånd och standardavvikelse är exempel på absoluta spridningsmått, medan percentilkvot och variationskoefficient är relativa mått.

Percentilavstånd anger den absoluta skillnaden mellan två percentiler och uttrycks i samma enheter som variabeln är mätt. Två vanliga exempel på percentilavstånd är kvartilavstånd och variationsbredd. Kvartilavståndet anger skillnaden mellan den 75e percentilen (den tredje kvartilen) och den 25e percentilen (den första kvartilen). I exemplet ovan kan vi beräkna kvartilavståndet till 136 300 kr ($292700 - 156400$). Eftersom detta avstånd inte påverkas av hur små de lägsta värdena är, eller hur stora de högsta värdena är, är kvartilavståndet okänsligt för extrema värden. Huruvida det är en önskad egenskap beror på vilken typ av spridning som är intressant. Variationsbredden anger skillnaden mellan det största och det minsta värdet. Det är därför mycket känsligt för extremvärden. Ordinalskala är tillräckligt för att sortera variabelvärden och därmed för att ange percentiler, men för att bedöma avståndet mellan två percentiler krävs det att vi vet hur stora skalstegen är.

Percentilkvoter påminner om percentilavstånd men beräknas i stället som den ena percentilen genom den andra. Ifall både värdena skulle fördubblas, skulle percentilkvoten förbli oförändrad. Det är därför ett relativt spridningsmått. En percentilkvot kan uttryckas som att den ena percentilgränsen är x gånger större än den andra eller räknas om till en procentuell skillnad. $p90/p10$ och $p90/p50$ är två vanliga percentilkvoter. Den första kan tolkas som den relativa inkomstskillnaden mellan en höginkomsttagare och en låginkomsttagare, medan den andra anger hur mycket mer än höginkomsttagare tjänar jämfört med medianinkomsten. I exemplet ovan kan de beräknas till

Tabell 3: Skalnivåer, centralitets- och spridningsmått

	Nominal	Ordinal	Intervall	Kvot
Centralitetsmått				
Typvärde	x	x	x	x
Median		x	x	x
Medelvärde			x	x
Spridningsmått				
Percentilavstånd			x	x
Percentilkvot				x
Standardavvikelse			x	x
Variationskoefficient				x

3.31 (387500/117200) och 1.78 (387500/21720). Med andra ord tjänar en person vid den 90e percentilen 78 procent mer än medianinkomsten och mer än tre gånger så mycket som en person vid den 10e percentilen.

Standardavvikelsen är det överlägset mest populära spridningsmåttet. Det har en mängd attraktiva statistiska egenskaper, men nu nöjer vi oss med att det anger den typiska avvikelsen från medelvärdet. Det sägs ibland att det är den genomsnittliga, absoluta avvikelsen, men det är bara nästan rätt. Nedan visas ekvationen för standardavvikelsen.

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} \tag{1}$$

Om du är ovan vid att läsa ekvationer av den här typen kan det vara enklare att gå igenom beräkningen steg för steg. x_i står för värdet på en variabel x för observation nummer i. Medelvärdet för samma variabel betecknas $\bar{x}$. Precis som med alla ekvationer säger räkneordningen att vi först beräknar parenteser, därefter potenser, multiplikation och division samt allra sist addition och subtraktion. Ett liggande bråkstreck tolkas som att det är en parentes runt uttrycken på vardera sida av strecket, så vi beräknar en sida i taget. Summatecknet ($\sum_{i=1}^{n}$) betyder att vi adderar allt som står till höger om summatecknet från den första observationen ($i = 1$) till den sista ($i = n$). I den här ekvationen innebär summatecknet att vi summerar alla kvadrerade avvikelser från medelvärdet ($(x_i - \bar{x})^2$).

1. Beräkna avvikelsen mellan varje observation (x_i) och medelvärdet ($\bar{x}$).
2. Kvadrera alla dessa avvikelser.
3. Summera de kvadrerade avvikelserna.
4. Dividera med antalet observationer (n) minus ett.

Tabell 4: Amerikanska presidenters tid som president

President	År som president
John F. Kennedy	3
Lyndon B. Johnson	5
Richard Nixon	5
Gerald Ford	3
Jimmy Carter	4
Ronald Reagan	8
George H.W. Bush	4
Bill Clinton	8
George W. Bush	8
Barrack Obama	7

5. Dra kvadratroten ur kvoten du just beräknade.

Den relativa motsvarigheten till standardavvikelsen är variationskoefficienten. Den beräknas som standardavvikelsen genom medelvärdet.

Vilka spridningsmått man använder beror alltid på vad man studerar. I regel är det en bra idé att följa de konventioner som finns inom det egna området när man vänder centralitets- och spridningsmått. Om man är osäker på vilka konventionerna är rekommenderas standardavvikelsen eftersom det är det absolut vanligaste måttet, men fundera också på hur spridningen bäst beskrivs i ditt fall. Är absoluta eller relativa skillnader mest relevant? Och hur känsligt vill du att ditt mått ska vara för extremvärden?

3.3 Räkneexempel

Vi ska nu beräkna de olika måtten för en variabel som mäter hur många år de tio senaste amerikanska presidenterna har spenderat vid makten. En del av måtten hade kanske varit mer meningsfulla med fler presidenter, men syftet med ett litet antal är att uträkningarna inte ska bli för tidskrävande.

Vi börjar med typvärdet, vilket är 8. Detta värde förekommer tre gånger. 4 och 5 är näst vanligast med två förekomster vardera. Notera att typvärdet är ett meningsfullt mått i det här exemplet, eftersom det inte är en slump vad värdet blev. Åtta år motsvarar nämligen två presidentperioder och är både den längsta och den mest sannolika tid en person kan vara president. På så vis kompletterar det de övriga centralitetsmåtten. För att beräkna medianen kan vi ställa upp värdena från minst till störst. I det här fallet har vi två mittersta värden, 5 och 5. Medianen är därmed 5.

3 3 4 4 5 5 7 8 8 8

Tabell 5: Amerikanska presidenters tid som president

President	År som president	$x - \bar{x}$	$(x - \bar{x})^2$
John F. Kennedy	3	-2.5	6.25
Lyndon B. Johnson	5	-0.5	0.25
Richard Nixon	5	-0.5	0.25
Gerald Ford	3	-2.5	6.25
Jimmy Carter	4	-1.5	2.25
Ronald Reagan	8	2.5	6.25
George H.W. Bush	4	-1.5	2.25
Bill Clinton	8	2.5	6.25
George W. Bush	8	2.5	6.25
Barrack Obama	7	1.5	2.25
Summa	55	0	38.5

För att beräkna medelvärdet summerar vi samtliga värden (55) och dividerar med antalet värden (10). Medelvärdet blir alltså 5.5. Låt oss nu övergå till spridningsmåtten. Genom att subtrahera det minsta värdet (2) från det största (8) erhåller vi variationsbredden, vilken i detta fall är 6. Några andra percentilavstånd beräknas sällan vid så få observationer. För att räkna ut standardavvikelsen manuellt är det en bra idé att ställa upp en tabell likt Tabell 5. Där visar den tredje kolumnen avvikelsen mellan varje observerat värde (x_i) och medelvärdet ($\bar{x}$). Den fjärde kolumnen visar det kvadrerade värdet av den tredje kolumnen ($(x_i - \bar{x})^2$). Genom att summera dessa får vi det som står ovanför bråkstrecket i formeln för standardavvikelsen ($\sum_{i=1}^{n}(x_i - \bar{x})^2$). Vi kan därmed stoppa in värdena och beräkna standardavvikelsen till 2,07. För att beräkna variationskoefficienten behöver vi vara dela standardavvikelsen med medelvärdet. Resultatet blir 0.38, vilket kan uttryckas som att standardavvikelsen var 38 procent av medelvärdet.

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{38.5}{9}} = \sqrt{4,68} = 2,07. \tag{2}$$

3.4 Nivåskattningar

Har Sverige en hög arbetslöshet? Är Kambodja en demokrati? För att svara på dessa frågor räcker det inte med att räkna antalet arbetslösa svenskar eller att besitta fullständig kunskap om den politiska processen i Kambodja. Vi behöver också relevanta referenspunkter med vilka vi kan jämföra det vi observerar.

Esaiasson m.fl. diskuterar tre typer av empiriska referenspunkter som

kan användas för att göra den typen av nivåskattningar. Dessa är *förändringsstrategin*, *populationsstrategin* och *referenspunktsstrategin*.

Förändringsstrategin handlar om att jämföra med andra tidspunkter. För att svara på om den svenska arbetslösheten är hög, kan vi bland annat konstatera att Sverige har högre arbetslöshet idag än vi haft under nästan hela efterkrigsperioden, men också att den är ungefär lika hög som för ett år sedan och lägre än under 1990-talskrisen. För att nämna ytterligare ett exempel skulle en korporatismforskare finna att Sverige inte är lika korporativistiskt som det var på 1970-talet.

Populationsstrategin går ut på att jämföra med en population, vilken vi som forskare bedömer att det aktuella fallet är en del av. Exempelvis har Sverige lägre arbetslöshet än EU-genomsnittet, vilket visar på vikten av vilken referenspunkt vi väljer. Vi skulle också kunna jämföra en revolution som skett nyligen med alla tidigare revolutioner för att avgöra om den gick snabbare, var blodigare eller fick större konsekvenser än vad som brukar vara fallet.

Den tredje strategin, referenspunktsstrategin, går ut på att jämföra med ett relevant fall där det är allmänt vedertaget att det har vissa egenskaper. Vi kanske intresserar oss för om sommarens händelser i Egypten var en statskupp eller en del av en revolution. Ett tillvägagångssätt skulle kunna vara att jämföra med andra svårkategoriserade händelser, men där det är allmänt vedertaget att kategorisera skeendet som antingen en statskupp eller en revolution. För att hitta sådana referenspunkter skulle man kunna läsa den akademiska litteraturen om oktoberrevolutionen i Ryssland 1917 eller den orangea revolutionen i Ukraina 2005.

Ett annat exempel på referenspunktsstrategin är hur vi bedömer storleken på statsskuldsräntor med hjälp av så kallade räntedifferenser. Tyskland är känt för att ha låga räntor. Därför beräknar man ofta skillnaden mellan räntan i ett land jämfört med räntan i Tyskland. Om skillnaden är liten vet vi att räntan kan anses som låg. Namnet på strategin är emellertid olyckligt, med tanke på att alla jämförelser kräver en referenspunkt.

Utöver dessa tre empiriska strategier kan vi även tänka oss andra typer av referenspunkter, även om de kanske är lite skakigare. Har skalan ändpunkter eller finns det beskrivningar av vad skalans värden motsvarar? I så fall kan vi med vissa förbehåll jämföra observationerna med dessa. Men även när observationerna befinner sig i närheten av en skalas ändpunkt är det viktigt att vi är försiktiga. Exempelvis har USA allt sedan 1871 det högsta möjliga värdet (10 av 10) på Polity IVs demokratiskala, trots att kvinnor inte fick rösträtt förrän 1920. Kan vi konstruera en eller två idealtyper att jämföra med? I frånvaro av relevanta empiriska jämförelsepunkter kan en jämförelse med en sådan teoretisk referenspunkt vara fruktbar. Och kanske finns det tydliga förväntningar på vad vi borde observera eller utsagor om vilka värden vi borde observera? Sådana kan man hitta i media eller den politiska debatten såväl som i bedömningar av andra forskare. Att positionera sig i förhållande

till dem är också ett sätt att göra en nivåskattning.

Det är lätt hänt att man fastnar i ett uppradande av begrepp. Jag vill därför avsluta med att de viktigaste insikterna om nivåskattningar kan summeras i följande slutsatser.

- Vi måste jämföra.
- Jämförelsen måste vara relevant.
- Vi måste vara tydliga med vad jämförelsevärdet representerar. Ett erkänt högt eller lågt värde? Ett typiskt eller representativt värde? Ett gränsfall mellan två kategorier?

3.5 Grafer

För avsnittet om hur och varför vi visualiserar data så hänvisar jag till mina slides.

4 Liten repetition och ordlista

- Skalnivå: Ett sätt att kategorisera variabler efter hur deras variabelvärden förhåller sig till varandra. Skalnivån säger vad vi kan använda för metoder.
- Centralitetsmått: En typ av mått som anges för att visa på vilket det typiska värdet för en variabel är.
- Spridningsmått: En typ av mått som mäter hur olika våra analysenheter är med avseende på en viss variabel, dvs hur stor spridning denna variabel har.
- Nivåskattningar: Att avgöra huruvida något är att betrakta som högt eller lågt, stort eller litet, dvs. att sätta det i relation till något annat.
- Förändringsstrategin: Jämför med andra tidpunkter.
- Referenspunktsstrategin: Jämför med en allmänt vedertagen empirisk referenspunkt.
- Populationstrategin: Jämför med hela populationen.