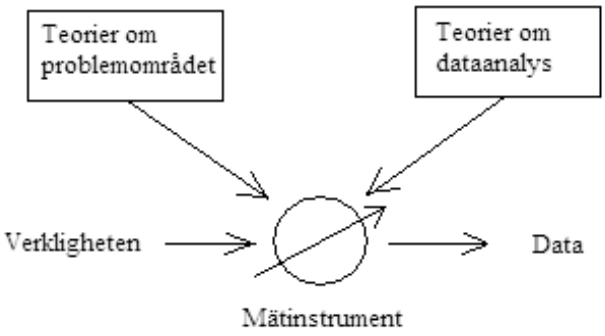
9. DATABILDNING

9.1 Allmänt om databildning

För att lära oss något om verkligheten måste vi förstås göra observationer på verkligheten. Att göra observationer är därför en viktig del av den empiriskt baserade kunskapsbildningen. När vi skall göra observationer ställs vi inför en rad frågor, t ex Vad skall vi observera, dvs vilka är våra undersökningsvariabler? Hur skall vi observera, dvs vilken teknik och vilket mätinstrument skall vi använda? Vilken del av omvärlden skall vi observera, dvs vilka individer skall ingå i undersökningen? När vi observerar värdet på undersökningsvariablerna för de individer som ingår i undersökningen bildas vad vi kallar data. Den process som har till uppgift att producera de data kunskapsbildningen behöver kallar vi databildning. I figur 9.1 visas en grafisk illustration av databildningen.



Figur 9.1 En modell över databildningen

Syftet med en empirisk studie är ofta att på något sätt avgöra om en modell är en rimlg beskrivnng av något vi studerar eller om en modell ger en rimlig

förklaring av något fenomen i verkligheten. Ofta innebär det att samband mellan variabler behöver uppskattas. Det är därför viktigt att datainsamlingen är kopplad till den teori man har om problemområdet. Den existerande teorin om problemområdet spelar alltså en viktig roll i databildningen. T ex besäms valet av undersöknngsvariabler av teoretiska överväganden. Hur bra mätningen än är, är den värdelös ifall resultaten inte kan placeras in i det sammanhang som är aktuellt. Teorin kan även antyda hur variablerna skall mätas, dvs hur mätinstrument skall konstrueras.

Det är emellertid inte enbart våra teorier om problemområdet som styr mätinstrumentets utformning och konstruktion. Även vår kunskap om tekniker för att tolka, bearbeta och analysera data är av stor betydelse. Det är ju tämligen meningslöst att samla observationer som vi sedan inte kan tolka. Hur än ett datamaterial skall analyseras ställer den analysteknik vi använder vissa krav på mätinstrumentet och dess egenskaper. Är inte dessa krav uppfyllda förlorar analysresultaten, och därmed kanske hela undersökningen, sitt värde. Det är därför viktigt att ha den tänkta dataanalysen klar för sig och känna till vilka krav dataanalysen kommer att ställa på mätinstrumentet redan då mätinstrumentet konstrueras.

Resultatet från våra mätnngar kallar vi data. Data består oftast av en följd av symboler, t ex tal eller bokstäver. I sig själv är data därför tämligen ointressanta. Ett exempel på data är

22 M 45 62 84

något som uppenbarligen inte säger oss särskilt mycket. Om vi däremot lägger till information om vad som har mätts och vår kunskap om de uppmätta storheterna och deras eventuella samband kan vi utläsa en hel del. Vi kan t ex få veta att ovanstående data är observationer på fem variabler hos en person och att variablerna är ålder, kön, skonummer samt puls före respektive efter en statistiklektion. Vi kan då sluta oss till att observationerna är gjorda på en 22-årig man med rätt stora fötter och att statistiklektionen var en upplevelse som inte lämnade personen i fråga oberörd. Från detta lilla exempel lär vi oss att data endast kan uttolkas och ge oss information om de kan kombineras med vår existerande kunskap om problemområdet och om mätinstrumentet. Vi säger att data är bärare av information, men att data i sig själv inte är information.

Ofta har vi observationer på flera individer, t ex

22 M 45 62 84

24 K 38 70 75

23 K 37 68 74

22 K 38 76 68

24 M 42 72 72

Här representerar varje rad observationer på en individ och varje kolumn observationer på en variabel. Den andra raden är alltså observationer på en 24 årig kvinna medan den andra kolumnen säger att vi har observationer på tre kvinnor och två män.

Empiriska undersökningar kan grovt delas in i två olika typer av undersökningar, experimentella studier och observationsstudier. I expertmentella studier utsätter vi medvetet de undersökta individerna för speciella behandlingar för att studera hur individern reagerar på behandlingarna. Samtidigt försöker man hålla andra betingelser som kan påverka utfallet så lika som möjligt för alla individer som ingår i försöket. Experimentella undersökningar kan därför ge svar på frågeställningar som "Kan acetylsalicylsyra minska risken för hjärtinfarkt?" och "Hur påverkas draghållfastheten i papper av inblandning av olika sorts fibrer i pappersmassa?". Syftet med en experimentell stydie är att undersöka om en viss behandling orsakar en förändring i responset. I en observationsstudie, å andra sidan, försöker man att låta bli att påverka de individer man observerar. Man strävar efter att passivt observara ett antal individer, deras tillstånd eller förändring över tiden med syfte att få en uppfattning om tillståndet hos en population eller hur individerna i den förändras. Observationsstudier kan t ex besvara frågor som hur stor andal av Sveriges väljarkår sympatiserar med ett visst politiskt pari och hur har priset på smör utvecklats under de senaste fem åren. Det är viktigt att de studerade individerna påverkas så lite som möjligt vid en observationsstudie, för om t ex syftet med en studie är att få en uppfattning om andelen individer som sympatiserar med ett visst politiskt parti, skulle man förmodligen få ett snedvridet resultat om man samtidigt förser urvalets individer med propagandamaterial om det studerade partiet.

En viktig typ av observationsstudier är urvalsundersökningar. Idén med en urvalsundersökning är att få information om en population av individer genom att observera endast en del, ett urval, av individerna i populationen. Det är mycket viktigt att inte blanda ihop population och urval. Populationen är alltså den grupp vi vill studera medan urvalet är den grupp vi faktiskt studerar. Urvalet är dessutom en del av populationen. Om vi t ex vill studera sysselsättningsgrad och begynnelselön hos nyexaminerade civilekonomer är populationen mängen av alla som tagit civilekonomexamen inom en viss tidsperiod före mättillfället. Det är viktigt att komma ihåg att individerna som ingår i ett urval inte är valda för att de är särskilt intressanta, utan för att de representerar sin population. Det är t ex inte de mest eller de minst avlönade civilekonomerna som är intresanta att studera om vi vill beskriva fördelningen av löner hos civilekonomer. I ett sådant fall är det viktigt att alla individer med alla storlekar av löner finns med i urvalet och att det finns fler individer i urvalet med de vanligaste lönerna än individer med extrema löner. För att informationen från ett urval skall vara meningsfullt måste vi veta vilken population urvalet representerar. Uppgifter om löner för t ex nyexaminerade civilekonomer ger knappast någon god beskrivning av fördelningen av löner för populationen av alla civilekonomer. I avsnitt 9.5 diskuterar vi mer om observationsstudier och i avsnitt 9.6 om urvalsundersökningar.

I en urvalsundersökning undersöker man bara en del av individerna i populationen. I en totalndersökning, å andra sidan, undersöker man alla individerna i populationen. Totalundersökningar av stora populationer blir ofta dyra och tar lång tid att genomföra. Det kan till och med inträffa att en totalundersökning tar så lång tid att genomföra så att resultaten är inaktuella när de är färdiga. Tid och pengar är alltså faktorer som talar till förmån för urvalsundersökningar. En annan faktor som är till fördel för urvalsundersökningar är att eftersomm antalet individer ofta är betydligt mindre än i totalundersökningar, är det möjligt att undersöka och mäta fler variabler på varje individ. På så vis kan en urvalsundersökning ge en mer nyanserad information om de undersökra individerna och därmed också en mer detaljerad information om hela populationen än vad en totalundersökning skulle ge. Ytterligare en anledning att föredra urvalsundersökningar före totalundersökningar är då man har så kallade förstörande prov. Om populationen består av tillverkade fyrverkeriraketer och undersökningen syftar till att

kontrollera kvaliteten på tillverkningen är det inte lämpligt att pröva alla raketerna: en totalundersökning skulle inte lämna kvar några raketer till försäljning.

Övning 9.1

Vilken typ av undersökning, urvalsundersökning, totalundersökning eller experimentell undersökning, passar bäst för att belysa följande problem:

a) Är det en bättre pedagogisk metod att studenterna själva kommer på hur statistikuppgifter skall lösas (learning by doing) än att lärarna demonstrerar uppgfternas lösning (learning by observing)?

b) Är studenterna nöjda med undervsningen?

c) Presterar sutdenter bättre på tentamina om de får lyssna på musik under tentamen?

d) Finns det några skillnader i hur Internet används (hur länge, hur ofta, vilken typ av sidor som besöks etc) mellan studenter på statistikkurser och övriga studenter?

e) Favoriseras hemmalag av domarna i ishockey?

f) Hur stor andel av deltagarna i Stockholm maraton använder skor av ett visst märke?

g) Blir livet längre av ett gott skratt?

9.2 Mätning

Statistik handlar om att beskriva och analysera en del av verkligheten genom att göra observationer. För att kunna få observationer måste vi göra mätningar på de variabler som vi är intresserade av. Vilka variabler vi skall mäta bestäms givetvis av den teori vi har om vårt problemområde. Allt för ofta tar vi lite för lätt på kopplingen till den bakomliggande teorin och gör mätninar på variabler som är lätta att mäta, snarare än varabler som är relevanta. Vi använder här ordet mätning i betydelsen att vi tillordnar ett tal till en viss egenskap hos en en individ. Mätning är altså den process som översätter en

teoretisk egenskap till ett tal. Varje möjlig egenskap hos en individ representeras alltså med ett visst tal och olika individer med olika egenskaper får olika tal. När vi mäter kroppslängd hos personer, t ex, representerar talen 168 och 179 två olika längder. På motsvarande sätt kan talet 88 representera en viss intelligens. För att utföra en mätning använder vi ofta ett instrument. För längdmätningar kanske vi använder ett måttband som intrument. Längd är en välkänd egenskap för oss och därför är det ofta lätt för oss att förstå vad vi menar med en viss längd. Om vi t ex ställer en person med längden 168 bredvid en person med längden 179 kan vi se med blotta ögat att den med längden 168 är något kortare än den med längden 179. Det instrument som används för att mäta intelligens består ofta av en uppsättning problem som en person skall försöka lösa. Beroende på hur många rätta lösningar personen lyckas lösa tillordnas personen ett visst tal som representerar personens intelligens. Att mäta intelligens är dock betydligt mer kontroversiellt än att mäta längd, eftersom det är inte helt klart vad vi menar med intelligens. De mätningar vi gör har också ofta en viss sort. Längder, t ex, kan mätas i sorten cm och intelligens i IQ.

Vi använder skolbetyg för att sortera bland de som söker en högskoleutbildning därför att vi tror att höga betyg ger bättre förutsättningar att klara studierna. De sökande med de högsta betygen antas till en viss utbildning. Däremot använder vi inte kroppsvikt för att bestämma vilka som skall antas till en utbildning. Anledningen till det är att vi tror inte att kroppsvikt har något med möjlig studieframgång att göra. Vi säger att skolbetygen är ett relevant mått på studieframgång eller att studentbetyg har hög validitet för begreppet studieframgång. Kroppsvikt, å andra sidan, är ett irrelevant mått på studieframgång eller en variabel som har låg validitet för begreppet studieframgång. En mätning är valid för en egenskap om den ger en relevant representation av den egenskapen.

En mätnings användbarhet avgörs av hur väl en variabel återspegla den egenskap vi vill studera, samt om variabeln är känslig för betydelsefulla variationer i egenskapen. Det är därför på sin plats att varna för ett alltför lättvindigt och slentrianmässigt användande av mätprocesser. Att t ex motivera användningen av en viss mätprocess med att den tidigare använts i liknande undersökningar är inte hållbart, om ändrade förutsättningar har medfört att mätprocessen har förlorat sin förmåga att representera den studerade egenskapen. I många situationer är det svårt att bestämma hur en

mätning skall utformas och hur väl de aktuella mätvärdena representer den egenskap vi önskar mäta.

En mätprocess kan vara mer eller mindre valid för en viss egenskap, det är inte så att en mätprocess antingen är vald eller så är den det inte. Däremot är det ofta svårt att avgöra hur valid en mätning är, det är sällan möjligt att mäta validitet. I ett viktigt fall är det dock möjligt att mäta validiteten hos en mätprocess och det är när den egenskap man mäter skall förutsäga något. Om vi använder gymnasiebetyg som mått på framtida studieframgångar i högskolan kan vi jämföra betygen med de poäng eller de examina som var och en klarar. Om det finns en stor överensstämmelse mellan studentbetygen och senare studieframgång i högskolan säger vi att skolbetygen har en stor prediktiv validitet, vi vet att gymnasiebetygen är ett bra mått på hur högskolestudierna komer att lyckas. Om det däremot inte är någon god överensstämmelse säger vi att gymnasiebetygen har en dålig prediktiv validitet för studieframgång. Någon liknande bedömning av ett inteligenstest kan inte göras eftersom vi inte har något att jämföra erhållna poäng på intelligens med. Det är därför svårt att avgöra om ett visst intelligenstest är en valid mätning av egenskapen intelligens.

En annan viktig aspekt på mätningar är hur precis en mätning är. Att mäta längd med ett måttband ger mätningar med stor validitet, men noggrannheten kan vara bristfällig. Om t ex måttbandet har en trasig början så att två centimeter saknas, kanske alla mätningar blir två centimeter för långa, dvs

$$\text{observarad längd} = \text{sann längd} + 2 \text{ cm}$$

Varje mätning kommer alltså att ge ett mätvärde som är exakt två cm mer än den sanna längden. När en mätprocess ger en sådan systematisk skilland mellan sant värde och uppmätt värde säger vi att mätningarna är biased. I detta exempel är storleken på biasen två cm. Förutom att observerade värden kan innehålla en bias, kan de också innehålla ett slumpmässigt fel. Det innebär att uprepade mätningar av samma egenskap, med samma mätinstrument och på samma indivd kan ge olika mätvärden. Då t ex kroppslängd mäts med ett måttband kan måttbandet sträckas ut olika mycket vid olika mätningar, förändringar i temperaturen kan ge olika längder mellan skalstrecken på måttbandet osv. Ett modell som beskriver observerade värden skulle därmed

kunna skrivas

$$\text{observerat värde} = \text{sannt värde} + \text{bias} + \text{slumpfel}$$

Slumpfelet kan vi betrakta som en stokastisk variabel. Vi betecknar den här med ε. Vilken sannolikhetsfördelning ε kan tänkas ha beror på en mängd olika faktorer, bl a det mätinstrument som används och den egenskap som skall mätas. I många tillämpningar tänker man sig att slumpfelet ε har en normalfördelning med väntevärdet 0 och variansen σ^2. Om vi betecknar det sanna värdet med μ och bias med b kan vi skriva

$$Y = \mu + b + \varepsilon$$

dvs resultatet från en mätning, ett observerat värde, kan betraktas som en stokastisk variabel, Y. Om $\varepsilon \sim N(0, \sigma^2)$ ser vi att $Y \sim N(\mu + b, \sigma^2)$.

Om det slumpmässiga felet är litet kommer upprepade mätningar på samma individ med samma mätinstrument att ge ungefär samma mätvärden. Vi säger då att mätningarna är reliabla. Observera här att variansen, eller standardavvikelsen σ, då är ett mått på mätningarnas reliabilitet. En mätprocess som alltid ger samma mätvärden då mätningar sker på samma individ är perfekt reliabel, även om mätprocessen är biased. En hög reliabilitet innebär alltså bara att mätvärdena är upprepbara. Bias och låg reliabilitet är två olia sorts brister.

Antag att vi önskar göra observationer på en viss variabel och att detta motiveras av den teori vi stöder undersökningen på. Det visar sig då ofta att den definition teorin ger på variabeln inte ger tillräckligt entydiga anvisningar om vilket värde en viss individ skall tillordnas på variabeln då vi gör en mätning. Vad som krävs är en beskrivning av hur man skall gå till väga, vilka operationer man måste utföra, för att kunna göra mätningen. Vi säger då att vi ger en operationell definition av variablerna.

Variabler som längd och vikt är relativt lätta att operationellt definiera. För dessa finns ofta lämpliga standardskalor och standardprocedurer att tillgå. Många andra variabler är däremot betydligt besvärligare att operationalisera. Om vi t ex önskar göra observationer på den teoretiska variabeln "inkomst" finns det flera olika operationella definitioner som är möjliga. T ex kan

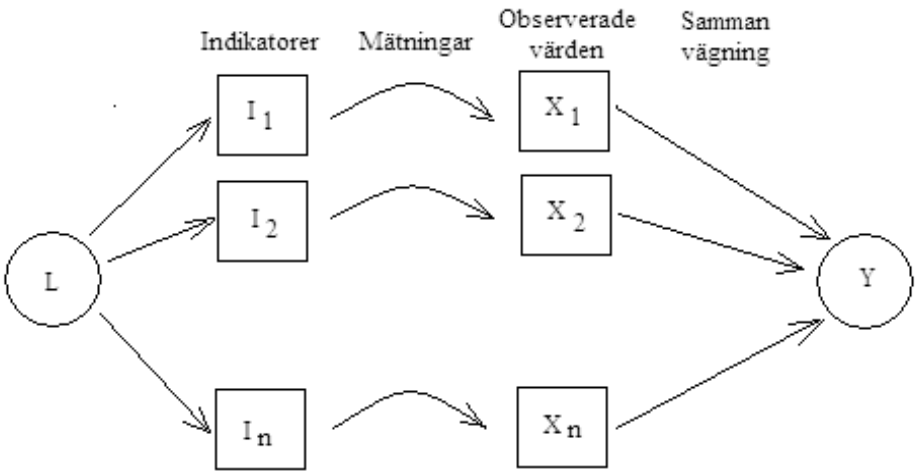
"inkomst" mätas som "bruttoinkomst" (summan av vad en individ får från sin/sina arbetsgivare) eller som "inkomst efter skatt och transfereringar".

Operationalisering innebär således att mätregler definieras. Vi har tidigare noterat att de mätvärden som erhålls vid mätningarna, data, skall kunna förmedla ett budskap, eller med andra ord vara informationsbärare. För att detta skall vara möjligt måste de mätregler vi sätter upp vara standardiserade, dvs vara rutinmässigt kända och/eller noga angivna. Utan standardiserade mätregler är det svårt att förstå och inordna mätvärden på variabler i det egna referenssystemet för tolkning. Exempelvis förmår temperaturangivelsen 60^o en svensk (med Celsiusskalan som referenssystem) att svettas medan en amerikan (med Farenheitskalan som referenssystem) inte kommer att tycka att temperaturen är särskilt märkvärdig. Sålunda innebär standardisering bl a att ett mätvärde ges en unik mening.

Operationalisering innebär inte enbart gradering av ett mätinstrument som t ex ett måttband. Även betingelserna i övrigt som påverkar mätningarna måste specificeras. T ex kan en högre omgivningstemperaturer göra ett måttband längre. Likaså kan en högre använd dragkraft vid mätning töja ut ett måttband och på så vis påverka mätresultaten.

I en del situationer inträffar det att den variabel vi önskar mäta inte kan observeras direkt. Sådana variabler brukar kallas latenta variabler. För att få en uppfattning om värdet på en latent variabel brukar man göra mätningar på så kallade indikatorer. Variabeln "betalningsförmåga", t ex, kan mer eller mindre väl representeras med någon av indikatorerna "bruttoinkomst", "inkomst efter skatt och transfereringar" och "skulder". För variabler som "intelligens", "attityd" och "livskvalitet" anses ofta en indikator vara otillräcklig varför flera indikatorer måste användas. En indikator är således en konkret och mätbar variabel som kan representera någon latent variabel.

Figur 9.2 visar schematiskt en situation då flera indikatorer används. Goda indikatorer är sådana som har ett starkt samband mellan den studerade latenta variabeln L och observerat värde, Y. Ju starkare sambandet är, desto bättre avspeglas variationer i L av variationer i Y. I situationer då flera indikatorer används beror styrkan inte enbart på vilka indikatorer som används, utan också på hur observerade värden på indikatorer kombineras till Y.



Figur 9.2 Mätning med flera indikatorer på en latent variabel.

I samband med statistiska undersökningar är det såluna viktigt att göra klart för sig

1. Hur är de ingående variablerna (operationellt) definierade?

2. Är variablerna valida beskrivningar av de egenskaper de avser att representera?

3. Är mätprocesen reliabel?

Övning 9.2

Hur kan man mäta trafiksäkerhet?

Övning 9.3

Diskutera hur variabeln ”antal invånadre i Sverige” kan definieras. Ta reda på hur den definieras i den officiella statistiken?

Övning 9.4

Diskutera hur variabeln ”arbetslöshet” kan definieras. Ta reda på hur den definieras i den officiella statistiken?

Övning 9.5

Ge ett exempel på en mätprocess som är har hög validitet men har en stor bias. Ge också ett exempel på en mätprocess som har liten validitet men har en hög reliabilitet.

9.3 Datainsamling

När undersökningens problemformulering är klar och variabler är definierade, blir uppgiften att samla in data. Man brukar skilja mellan två typer av data, nämligen primärdata och sekundärdata. Primärdata är nya data som undersökaren själv samlar in genom att använda sig av någon datainsamlingsmetod. Sekundärdata är data som insamlats av andra och som kan vara mer eller mindre tillgängliga. I detta avsnitt diskuterar vi något kring insamling av primärdata och något om för och nackdelar med att använda sekundärdata.

När undersökaren själv skall samla in sina data, skapa primärdata, finns det ett antal olika metoder för datainsamling att välja mellan. De två vanligaste metoderna är experimentella studier och observationsstudier. Dessa två metoder för datainsamling diskuteras närmare i avsnitten 9.4 och 9.5.

Vid sidan av planerade försök och observationsstudier förekommer ytterligare ett antal datainsamlingsmetoder. I medicinska undersökningar, t ex, förekommer det ofta att patienter som uppsöker en viss klinik för någon åkomma får ingå i en undersökning. Liksom i en observationsstudie noteras ett antal egenskaper hos patienterna. Vad som dock skiljer detta förfarande från en observationsstudie är att det är här oklart vilken population individerna tillhör och hur urvalet har gått till. Detta gör att förutsättningarna för analys av observationerna är helt annorlunda än i observationsstudier. Vi diskuterar inte denna datainsamlingsmetod ytterligare här.

Slutligen skall nämnas att simulering ibland anges som en datainsamlingsmetod. Simulering innebär att en modell används för att generera data, dvs det genererade datat speglar inte något verkligt förlopp eller förhållande. Syftet med att generera simulerade data är att studera den modell som har genererat datat och är således inte någon metod att studera empiriska förhållanden.

I många fall finns redan tillgängliga data, sekundärdata, som kan belysa den

valda problemställningen. Det är då bortkastad tid att inhämta egna data. Innan man tar ställning till att använda ett sekundärdatamaterial måste man emellertid ta reda på om materialet lämpar sig för att belysa den valda problemställningen. Sekundärdata är ofta insamlade för helt andra syften och kan vara utmärkta för att belysa andra frågeställningar men mindre lämpliga för den aktuella undersökningen. Det handlar här om ett datakvalitetsproblem. Vilken relevans har ett givet sekundärdatamaterial för vår undersökning och vilken noggrannhet har det?

När det gäller vissa typer av problemställningar har man inget annat val än att använda sig av sekundärdata. Skall man studera tidigare förhållanden, historiska trender eller sociala förändringar har man vanligtvis inget alternativ till sekundärdata. Samma sak gäller om man är intresserad av att studera makroekonomiska förhållanden som t ex ett lands ekonomiska utveckling.

En stor fördel med att använda sekundärdata är att man spar både pengar och tid. Att själv samla in data är i allmänhet dyrt och tar ofta lång tid.

De operationella definitioner som används insamling av sekundärdata kanske inte helt täcker in de teoretiska definitioner som är aktuella vid den egna undersökningen. Sådana förhållanden försämrar sekundärdatats relevans. Det är också viktigt att kunna avgöra noggrannheten hos sekundärdata. Detta förutsätter då att insamlingen av sekundärdata har dokumenterats och redovisas, något som tyvärr inte alltid är fallet.

När man skall använda sekundärdata från flera olika källor uppstår problem med jämförbarheten. Samma problem uppstår när man skall jämföra data från flera länder. Förutom att begreppen kan ha operationaliserats på olika sätt, så kan praxis vid registrering variera från land till land.

Ett annat problem som kan uppstå vid användningen av sekundärdata är att variabler som är tänkta att ingå i undersökningen helt kan saknas eller att den kategoriindelning som gjorts inte är relevant för den aktuella undersökningen.

Ett speciellt problem som kan uppstå är att datakällor ofta saknar uppgifter om de så kallade mörkertalen. Kriminalstatistiken t ex omfattar inte all kriminalitet, utan endast den som anmäls. På motsvarande sätt omfattar den ekonomiska statistiken inte data från den “svarta” ekonomin.

En stor fördelen med att använda offentlig statistik är att den redan finns

insamlad och att man kan få data från perioder långt tillbaka i tiden.

Den viktigaste producenten av sekundärdata är Statistiska Centralbyrån (SCB). Statistiska Centralbyrån redovisar sin produktion i en rad publikationer. En viktig sådan publikation är "Vägledning i Sveriges officiella statistik". Den ger en systematisk översikt över de ämnesområden som täcks in av SCB's datainsamling. I "Publikationer utgivna av Statistiska Centralbyrån" och "Månadens tryck" redovisas de publikationer som getts ut det senaste året respektive den senaste månaden. Publikationen "Allmän månadsstatistik" innehåller tabeller med löpande månads och kvartalsstatistik.

Vid sidan av SCB är många kommuner och andra offentliga förvaltningar viktiga dataproducenter.

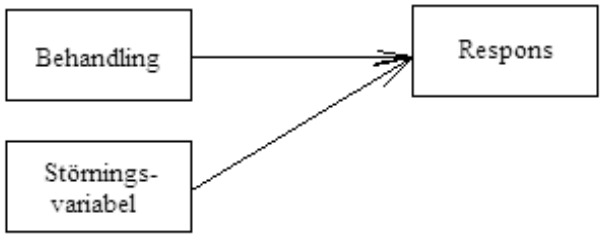
9.4 Experimentella studier

I ett experimentella studier utsätts ett antal individer, försöksenheter, medvetet för ett antal olika behandlingar med syftet att studera behandlingarnas effekter på försöksenheterna. Ett exempel på försöksenheter är områden av odlad mark som utsätts för olika gödselmedel för att studera gödslingens inverkan på avkastningen. Ett annat exempel på försöksenheter är människor med en viss sjukdom som utsätts för olika medicinska behandlingar för att studera vilken behandling som ger den i någon mening bästa läkningen.

Den variabel vi använder för att avläsa resultatet av en behandling brukar kallas för respons. De individer som utsätts för behandlingarna, försöksenheterna, fördelas slumpmässigt på de olika behandlingarna med hjälp av någon slumpmekanism. Detta kallas randomisering och används för att undvika att systematiska fel i undersökningsresultaten skall uppstå. Tilltron till en undersökning av en ny hostmedicin skulle inte vara särskilt stor om t ex alla som får den nya medicinen har en lindrig hosta medan alla som får en konventionell medicin har en svår hosta. Man skulle då kunna misstänka att de med lindrig hosta skulle ha blivit friska även utan behandling medan vi inte vet hur behandlingen skulle ha verkat för de med svår hosta. I det fallet finns det alltså en annan variabel än behandlingen, en störningsvariabel som vi kan kalla sjukdomsgrad, som påverkar utfallet av behandlingen. Om vi inte kan skilja på effekterna av en behandling och en annan variabel säger vi att vari-

ablerna är sammanblande, eller confoundade, se figur 9.3 för en illustration. Randomisering, dvs att fördela försöksenheterna slumpmässigt på de olika behandlingarna, är en metod för att undvika att effekter av behandlingar sammanblandas med effekter av störningsvariabler.

Ett viktigt karaktäristikum för experimentella studier är att de skall kunna upprepas. Upprepbarhetskravet medger att försökssituationen kan beskrivas med hjälp av en sannolikhetsmodell och att resultaten kan analyseras med hjälp av statistiska metoder. Om försöken inte vore upprepbara skulle resultaten vara giltiga endast för det studerade försökstillfället och några generella slutsatser skulle alltså inte vara möjliga att dra.



Figur 9.3 I ett försök kan effekten av en behandling sammanblandas med effekten av en störningsvariabel

Om vi vill studera effekten av en behandling delar vi in försöksenheterna i två grupper, en grupp som får behandlingen och en kontrollgrupp som inte får behandlingen. En enkel situation med två värden på störningsvariabeln kan då formellt beskrivas på föjlade sätt. Beteckna effekten av behandingen med τ och antag att effekterna av störningsvariabeln är $a - s$ och $a + s$. Antag vidare att vi randomiserar så att störningsvariabeln har effekten $a - s$ på en försöksenhet är 0,5 och effekten $a + s$ på en försöksenhet är 0,5. Vi kan då betrakta effekten av störningsvariabeln som en stokastisk variabel, S, med sannolikhetsfördelningen $P(S = a - s) = P(S = a + s) = 0,5$. Uppenbarligen är $E(S) = 0.5(a - s) + 0,5(a + s) = a$. Individer i gruppen som får behandlingen får då responset

$$Y = \tau + S$$

med väntevärdet $E(Y) = E(\tau + S) = \tau + E(S) = \tau + a$. I kontrollgruppen får individerna responset

$$Y = S$$

med väntevärdet $E(Y) = E(S) = a$. Skillnaden mellan grupperna i förväntat respons är således $(\tau + a) - a = \tau$, den behandlingseffekt vi vill studera.

De flesta försök syftar till att jämföra effekter av flera olika behandlingar. Vi använder då fler behandlingsgrupper och vanligen en kontrollgrupp och fördelar försöksenheterna slumpmässigt över alla grupper.

I vissa försök, speciellt i samband med utprovning av medicinska behandlingar, uppträder psykologiska effekter hos försöksenheterna. Att över huvet taget få en behandling, vilken som helst, kan bidra till att patienter blir friska. Den effekten kallas placeboeffekten och är en störningsvariabel. För att kunna skilja på behandlingseffekten och placeboeffekten brukar man därför låta en av de studerade behandlingarna vara en s k placebo behandling, dvs en medicinskt verkningslös behandling. Vidare får patienterna inte veta vilken behandling de har fått, placebo eller en med förmodad medicinsk verkan. Man säger då att försöket är blint. För att vara helt säkra på att inte resultaten påverkas av försöksledarna brukar det vara obekant för både patienter och försöksledare hur behandlingarna har fördelats då försöket genomförs. Behandlingarna förses då med en kod och det är först efter att försöket har genomförts och alla responser har avlästs som koderna bryts och behandlingseffekterna analyseras. Ett sådant försök kallas ett dubbelt blint försök.

Ett annat sätt att kontrollera för störningsvariabler är att försöka konstanthålla dem. Om man vet att t ex temperatur är en variabel som påverkar responset kan man låta alla försöksenheter vara utsatta för samma temperatur under försöket. Då kommer också effekten av temperatur att bli densamma för alla försöksenheter och, på samma sätt som ovan, skillnader i respons mellan olika behandlingsgrupper kommer endast att innehålla effekter av behandlingar. Ibland är det svårt att låta alla försöksenheter ha samma värde på störningsvariabler. Då man t ex provar en metod för viktminskning kan den ha olika stor effekt beroende på försökspersonernas initialvkt, ålder och kön. Om vi skulle försöka randomisera över dessa störningsvarabler skulle vi behöva tilldela försökspersonerna en viss initialvikt, en viss ålder

och ett visst kön. Detta kan givetvis omöjligen göras, en försöksperson har redan en vikt, en ålder och ett kön och de kan inte ändras. Vi försöker i stället konstanthålla störningsvariablerna. Ett sätt att göra det vore att genomföra försöket med försökspersoner med samma initialvikt, samma ålder och samma kön. De resultat man skulle komma fram till skulle då vara fri från inverkan av störningsvariablerna initialvikt, ålder och kön, men också endast vara giltig för personer med just den använda initialvikten, åldern och könet. Vill man att resultaten skall vara giltiga för ett större grupp personer måste man genomföra motsvarande försök med personer med andra initialvikter, åldrar och kön. Ett sådant delförsök brukar kallas block och data från hela försöket kan analyseras med statistiska metoder.

En speciellt enkel form av blockförsök inträffar då man har två behandlingar och man har två försöksenheter i varje block. Försöksenheterna i ett par skall då vara så lika varandra som möjligt med avseende på störningsvariabler man vill kontrollera för. Man säger då att man har matchade par. De två behandlingarna randomiseras sedan över de två försöksenheterna.

Exempel 9.1

Ett företag erbjuder kurser som avser att förbereda studenter inför en tentamen i statistik, För att utvärdera kursen gjordes ett försök med 12 studenter. Studenterna grupperades i sex par på såant sätt att båda medlemmarna i varje par hade samma kön, samma ålder och så lika studentbetyg som möjligt. I tabell 9.1 visas data för deltagarna i försöket. Vi ser att deltagare 1 och 5 matchar varandra och kan bilda ett par. En av dessa väljs slumpmässigt ut och får genomgå den förberedande kursen, medan den andra studenten i paret får förbereda sig på vanligt sätt inför tentamen. På samma sätt matchar deltagare 2 och 12 varandra och en av dem väls slumpmässigt ut för att delta i kursen. De par som bildades visas i tabell 9.2. Där visas också de poäng respektive student fick på tentamen.

Tabell 9.1 Data för deltagare i försök för att utvärdera en kurs

Deltagare	Kön	Ålder	Betyg
1	K	20	18,2
2	K	28	17,8
3	M	19	18,2
4	M	27	17,4
5	K	20	18,2
6	K	24	17,8
7	M	27	17,5
8	M	22	17,4
9	K	24	17,8
10	M	19	18,3
11	M	22	17,3
12	K	28	17,9

Tabell 9.2 Resultat från försöket för att utvärdera en kurs

Par	Deltagare	Kurs	Ej kurs	Differens
1	1 och 5	82	75	7
2	2 och 12	73	71	2
3	3 och 10	93	83	10
4	4 och 7	59	52	7
5	6 och 9	69	70	-1
6	8 och 11	48	46	2

Observera att detta försök har precis de karakteristika ett planerat försök skall ha: försöket är upprepbart, bahandlingarna (delta i förberedande kurs eller traditionell förberedelse) fördelas slumpmässigt över försöksenheterna (studenterna som deltar i försöket) och det finns en kontrollgrupp (traditionell förberedelse). Vidare har det skett en matchning av deltagarna så att de med lika eller ungefär lika värden på ett antal matchningsvariabler har förts samman till ett par.

Antag nu att den poäng en student får på tentamen är en observation på en normalfördelad stokastisk variabel med väntevärdet μ och variansen σ^2. Att poängen är en stokastisk variabel motiveras av att den beror av en lång rad störningsvariabler som vi inte känner men som slumpmässigt påverkar

studenterna och som genom randomiseringen fördelas slumpmässigt över behandlingarna. Om försöket upprepas får vi nya effekter av de okända störningsvariablerna. Poängens väntevärde, i sin tur, beror dels på störningsvariablerna kön, ålder och studentbetyg och dels på om studenten har gått den förberedade kursen eller inte. För den student i det $j-$te paret, $j = 1, 2, \ldots, 6$, som genomgått den förberedande kursen kan vi skriva väntevärdet av poängen som

$$\mu_j = s_j + \tau,$$

där s_j är effekten av de konstanthållna störningsvariablerna i det $j-$te paret och τ är effekten av kursen. För den andra studenten i det $j-$te paret blir då väntevärdet av poängen

$$\mu_j = s_j.$$

Eftersom båda studenterna i ett par har samma konstanthållna störningsvariabler får de också samma effekt, s_j, från dem. Däremot får endast den student som har genomgått den förberedande kursen en effekt av den, τ. Om vi nu bildar differensen av tentamenspoängen inom ett par, såsom visas i det högra kolumnen i tabell 9.2, får vi tydligen observationer på en normalfördelad stokastisk variabel med väntevärdet

$$(s_j + \tau) - s_j = \tau,$$

dvs väntevärdet är lika med effekten av den förberedande kursen, och variansen

$$\sigma^2 + \sigma^2 = 2\sigma^2.$$

Detta kommer sig av att variansen för skillnaden mellan två oberoende stokastiska variabler är lika med summan av varianserna, såsom det redogjordes för i kapitel 6. Hur data från försöket analyseras diskuteras i kapitel 17. ■

Övning 9.6

En tillverkare av livsmedel förpackar sina produkter i en sorts plastförpackning. Förpackningarna försluts genom att en upphettad tång nyper ihop förpackningens öppning. Den kraft som behövs för att öppna förpackningen beror på den

temperatur som tången har vid förslutningen. För att studera effekten av tångens teperatur skall ett försök med totalt 20 förpackningar användas och fem olika temperaturer, 75, 90, 105 och 120 grader, skall prövas. Identifiera vad som är försöksenheter, behandlingar och responsvariabel. Förklara hur randomiseringen kan gå till och vilken betydelse den har för försöket.

9.5 Observationsstudier

Vid observationsstudier väljs ett antal individer ut ur en population av individer och några av individernas egenskaper noteras med syftet att studera egenskapernas fördelning i populationen. Exempelvis kan individerna vara människor som tillhör populationen av alla röstberättigade i landet vid ett visst tillfälle. Den studerade egenskapen kan vara individernas partisympatier och syftet med datainsamlingen att skapa underlag för slutsatser om storleken på de olika politiska partiernas sympatier i populationen. Ett annat exempel på individer i en observatuionsstudie är tillverkade enheter valda från populationen av alla tillverkade enheter under en viss tidsperiod. Den studerade egenskapen kan i detta exempel vara individens kvalitetsegenskaper med syftet att dra slutsatser om kvaliteten hos alla enheter i populationen eller, alternativt, dra slutsatser om tillverkningsprocessen. Det är viktigt här att notera att individerna som väljs ut att ingå i en observationsstudie inte påverkas av undersökaren på något sätt. Behandlingarna i den experimentella studien ersätts här av ett passivt observerande.

Det finns i huvudsak två typer av observationsstudier, nämligen totalundersökningar och urvalsundersökningar. En totalundersökning ger förstås den största informationen om populationen, under förutsättning att mätningarna sker utan fel, medan en urvalsundersökning ger information från endast en del av populationen. Trots det är det nästan alltid bättre att använda sig av en urvalsundersökning. I det här avsnittet diskuterar vi fyra skäl som gör att urvalsundersökningar är att föredra, nämligen kostnadsaspekten, tidsfaktorn, populationsstorleken och förstörande prov.

Eftersom varje observation för med sig en viss kostnad blir en undersökning billigare om antalet observationer är färre. En totalundersökning är därför den dyraste formen av undersökningar och faller ofta utanför undersökningsbutgetens ram. Om det är möjligt att erhålla tillräckligt mycket information

från ett urval, så är alltså en urvalsundersökning att föredra av kostnadsskäl.

Om populationen är stor tar en totalundersökning mycket längre tid att genomföra än en urvalsundersökning. Detta beror givetvis på att en totalundersökning har avsevärt större antal observationer att samla in och att bearbeta. Resultaten kan därför erhållas snabbare från en urvalsundersökning. I vissa fall tar en totalundersökning så lång tid att genomföra, att resultaten är inaktuella då undersökningen är klar. Om det är viktigt att begränsa undersökningstiden, är sålunda en urvalsundersökning att föredra.

Om populationen är mycket stor kan det vara praktiskt omöjligt att göra en totalundersökning. Detta gäller speciellt oändliga populationer. Då är en urvalsundersökning det enda möjliga sättet att skaffa information om populationen. Vi har tidigare nämnt den oändliga populationen "mängden av alla kast som kan göras med en tärning". Information om den populationen kan därför endast fås genom att studera ett urval, dvs ett visst ändligt antal kast med tärningen.

I vissa fall förstörs individen då observationen görs. Detta kallas förstörande prov. Ett exempel på detta är kvalitetskontroll av fyrverkeriraketer. Då ett parti av fyrverkeriraketer skall kvalitetsprovas före försäljning måste en urvalsundersökning användas. Vid en totalundersökning skulle alla fyrverkeriraketerna avfyras och inga blir kvar till försäljning. Vid förstörande prov är därför en urvalsundersökning att föredra.

Det finns flera olika sätt att dra ett urval från en population, dvs att välja ut de individer som skall ingå i urvalet. Då populationen består av människor och mätinstrumentet är ett frågeformulär förekommer det ibland att man låter frågeformulär vara tillgängliga för individerna i populationen och låta de som känner sig manade svara på frågorna. Det är uppenbart att detta förfaringssätt för med sig en lång rad problem. Dels är det svårt att veta om det verkligen är individer ur populationens som svarar. Dels kommer antagligen inte urvalets individer att vara representativa för populationen.

Ett annat sätt att dra sitt urval på är att ur en uppräkning av populationens individer välja ut de individer som man anser är representativa för hela populationen. Detta är givetvis mycket svårt, särskilt som syftet med en undersökning ofta är just att undersöka vad som är typiskt för populationen. Trots detta finns det situationer då tekniken är den enda möjliga. T ex vore

det svårt att basera konsumentprisindex på någon annan urvalsteknik. Vid beräkning av konsumentprisindex väljer man ut särskilda så kallade representantvaror som fungerar som ett urval av de varor och tjänster som konsumeras i riket. Det som sedan studeras hos representantvarorna är varornas prisutveckling. Det är möjligt att välja representantvaror på detta sätt eftersom man har god kännedom om hur stor konsumtionen av olika varor faktskt är. I allmänhet leder dock analyser av data från urval bestående av individer man anser vara typiska för poplationen till problem som försvårar möjligheterna att dra slutsatser.

Den vanligaste urvalstekniken grundar sig på att slumpmässigt välja ut de individer som skall ingå i urvalet.

9.6 Slumpmässiga urval

De flesta statistiska tekniker som finns för att dra slutsatser från urval är utformade för slumpmässiga urval. En grundläggande form av slumpmässigt urval är s k *obundet slumpmässigt urval*, OSU. Vi säger att vi har ett OSU om varje urval om n individer ur en poplation om N individer har samma sannolikhet att bli dragen. Exempelvis kan ett urval bestå av $n = 400$ röstberättgade från en kommun bestående av $N = 100000$ röstberättigade och syftet med undersökningen är att uppskatta hur stor proportion av kommunens röstberättigade som är för ett visst förslag. Ett annat exempel är att välja ut $n = 50$ verifikat om ekonomiska transaktioner ur en population om $N = 5000$ verfikat i samband med en revision. Vid en revision kan man dels studera hur stor andel av verifikaten som överensstämmer med bokförda värden och dels kan man studera storleken på skillnaden är mellan de värden som anges på verifikaten och de bokförda värdena.

Utgångspunkten för att dra ett OSU är en förteckning av alla individer i populationen, en så kallad *urvalsram*. Varje individ förses med en unik identifikation som kan anges i urvalsramen. Om populationen består av människor kan identifikationen t ex bestå av namn eller personnummer. Ett slumpmässigt urval kan i sin enklaste form utgöras av ett vanligt lotteri. I lotteriet finns lika många lotter som individer i populationen och på varje lott har identifikationen hos en individ markerats. Efter att ha blandat lotterna väl dras ett erfoderligt antal lotter. Urvalet får bestå av de individer som svarar mot

identifikationerna på de dragna lotterna.

Om populationen är stor blir detta förfaringssätt praktiskt ohanterligt. Ett bättre sätt är då att numrera individerna i urvalsramen från ett och i stigande ordning. Därefter skaffar man sig erfoderligt antal slumptal från en så kallad slumptalsgenerator eller en slumptalstabell. De individer som har ett nummer svarande mot ett erhållet slumptal får då utgöra urvalet.

Ett annat sätt att dra ett urval är att använda dra ett s k systematiskt urval. Den första individen dras då slumpmässigt bland de första individerna i urvalsramen. De övriga individerna blir precis de individer som står på ett visst givet och fixt avstånd mellan varandra och från den först dragna i urvalsramen. På så sätt kan t ex individ nummer 3 dras först och sedan dras var femte med början vid individ nummer 3, dvs urvalet består av individerna med nummer 3, 8, 13,... i urvalsramen. Avståndet mellan individerna bestäms dels av populationsstorleken och dels av urvalsstorleken.

Systematiska urval kan ibland ge upphov till ett systematiskt fel. Detta uppkommer då urvalsramen har någon form av periodicitet som överensstämmer med urvalets period. Om t ex populationen består av gifta par, mannens namn först följd av hustruns namn på nästa rad i urvalsramen, kommer urvalet antingen att bestå av enbart män eller enbart kvinnor om ett udda antal individer hoppas över då ett systematiskt urval dras. Om individerna är slumpmässigt ordnade i urvalsramen blir de statistiska egenskaperna desamma som för OSU.

Antag vi vill undersöka sympatierna för ett visst politiskt parti i en population av röstberättigade och att proportionen sympatisörer i populationen är p. Om vi slumpmässigt väljer en individ ur populationen är den individen antingen en sympatisör eller inte en sympatisör med det studerade pertiet. Vi låter X vara lika med 1 om personen sympatiserar med pariet och 0 om hon inte gör det. Innan individen är vald är alltså X en stokastisk variabel som antar värdet 1 med sannolikheten p och värdet 0 med sannolikheten $(1-p)$, eller i formler

$$f(x) = p^x (1-p)^{1-x}, \; x = 0, 1$$

X är sålunda en Bernoullifördelad stokastisk variabel med väntevärdet

$$E(X) = p$$

och variansen

$$V(X) = p(1-p).$$

Om vi slumpmässigt väljer flera individer låter vi X_1 vara den stokastiska variabel som svarar mot den första individen, X_2 den som svarar mot den andra osv. Om vi sammanlagt observerar n individer får vi på det sättet n stycken stokastiska variabler, $X_1, X_2, \ldots, X_n$. Invididerna kan antingen väljas *med återläggning* eller *utan återläggning*. Om de väljs med återläggning innebär det, som namnet antyder, att en individ som har blivit vald läggs tillbaka till urvalsramen och kan således bli vald en gång till. Dragning med återläggning innebär att de stokastiska variablerna $X_1, X_2, \ldots, X_n$ blir stokastiskt oberoende av varandra. Vid dragning utan återläggning, å andra sidan, läggs en vald individ inte tillbaka till urvalsramen och kan då inte bli vald mer än en gång. En konsekvens av detta är att urvalsramen ändras efterhand som individer dras och läggs till urvalet. De stokastiska variablerna $X_1, X_2, \ldots, X_n$ är stokastiskt beroende vid dragning utan återläggning.

Medelvärdet av variablerna,

$$\overline{X} = \frac{1}{n}(X_1 + X_2 + \ldots + X_n) = \frac{1}{n}\sum_{i=1}^{n} X_i$$

är en stokastisk variabel med väntevärdet

$$E\left(\overline{X}\right) = p.$$

Detta följer av räknereglerna för väntevärden enligt kapitel 6. Variansen för $\overline{X}$ blir olika beroende på om vi har ett urval med eller utan återläggning. Om urvalet är med återläggning kan vi använda räknereglerna för varianser i kapitel 6 för att få

$$V\left(\overline{X}\right) = \frac{p(1-p)}{n}.$$

Om urvalet är utan återläggning gör det stokastiska beroendet att det tillkommer en faktor till variansuttrycket och vi får

$$V\left(\overline{X}\right) = \frac{p(1-p)}{n}\left(\frac{N-n}{N-1}\right),$$

där N är antalet individer i hela populationen.

Exempel 9.2

Antag vi har en kommun med $N = 100000$ röstberättigade som skall folkomrösta om en speciell fråga och $0,70\%$ av de röstberättigade är positiva till det förslag man skall rösta om. I ett slumpmässigt urval utan återläggning om $n = 400$ individer kommer då proportionen individer som är positiva till förslaget att vara en stokastisk variabel $\overline{X}$ med väntevärdet

$$E\left(\overline{X}\right) = 0,70$$

och variansen

$$V\left(\overline{X}\right) = \frac{0,70\,(1-0,70)}{400}\left(\frac{100000-400}{100000-1}\right) = 0,0005229. \blacksquare$$

Antag nu att populationen kan delas in i grupper, så kallade strata, där individerna inom ett stratum är ungefär lika varandra med avseende på den studerade egenskapen, medan individerna i olika strata är tämligen olika. Ett så kallat stratifierat urval erhålles om oberoende OSU dras från varje stratum. Fördelen med stratifierade urval är att de kan ge en större precision i undersökningsresultaten jämfört med ett OSU från hela populationen. Förutsättningen för detta är att man har tillräckligt bra information om individerna före undersökningen, så att indelningen i strata kan ske på ett lämpligt sätt.

Exempel 9.2 (fortsättning)

Antag nu att en älv delar kommunen i två lika stora delar så att $N_1 = 50000$ röstberättigade bor norr om älver och $N_2 = 50000$ röstberättigade bor söder om älven. Antag också att proportionen positiva till förslaget bland de röstberättigade som bor norr om älven är $p_1 = 0,90$ medan de röstberättigade som bor söder om älven är mer tveksamma, proportionen positiva där är $p_2 = 0,50$. I hela kommunen är då antalet positiva till förslaget $0,90 \cdot 50000 + 0,50 \cdot 50000 = 45000 + 25000 = 70000$ så att proportionen positiva till förslaget är fortfarande $p = 70000/100000 = 0,70$.

Vi ordnar nu urvalet så att vi drar ett OSU om $n_1 = 150$ individer från den norra sidan och ett OSU om $n_2 = 250$ individer från den södra sidan. Beteckna proportionen positiva i respektive urval med $\overline{X}_1$ respektive $\overline{X}_2$ och en ihopvägd proportion med $\overline{X}$, där $\overline{X}$ definieras genom

$$\begin{aligned}\overline{X} &= \frac{N_1}{N}\overline{X}_1 + \frac{N_2}{N}\overline{X}_2 = \frac{50000}{100000}\overline{X}_1 + \frac{50000}{10000}\overline{X}_2 \\ &= 0,5\overline{X}_1 + 0,5\overline{X}_2\end{aligned}$$

Det är nu lätt att se att väntevärdet för $\overline{X}$ är

$$\begin{aligned}E\left(\overline{X}\right) &= 0,50E\left(\overline{X}_1\right) + 0,50E\left(\overline{X}_2\right) = 0,50 \cdot 0.90 + 0,50 \cdot 0,50 \\ &= 0,45 + 0,25 = 0,70\end{aligned}$$

dvs $\overline{X}$ har samma väntevärde som när vi tar ett OSU från hela populationen och är lika med proportionen positva till förslaget i hela populationen. Variansen för var och en av proportionerna är

$$V\left(\overline{X}_1\right) = \frac{0,90\left(1-0,90\right)}{150}\left(\frac{50000-150}{50000-1}\right) = 0,0005982$$

$$V\left(\overline{X}_2\right) = \frac{0,50\left(1-0,50\right)}{250}\left(\frac{50000-250}{50000-1}\right) = 0,0009950$$

Det gör att variansen för $\overline{X}$ blir

$$\begin{aligned}V\left(\overline{X}\right) &= V\left(0,50\overline{X}_1 + 0,50\overline{X}_2\right) = 0,50^2V\left(\overline{X}_1\right) + 0,50^2V\left(\overline{X}_2\right) \\ &= 0,25 \cdot 0,0005982 + 0,25 \cdot 0,0009950 = 0,0003983\end{aligned}$$

Variansen är alltså mindre för den sammanvägda proportionen än för proportionen då vi tar ett OSU från hela populationen. Hur stora urval man bör ta från respektive stratum beror på variansen i varje stratum och storleken av varje stratum. ■

Övning 9.7

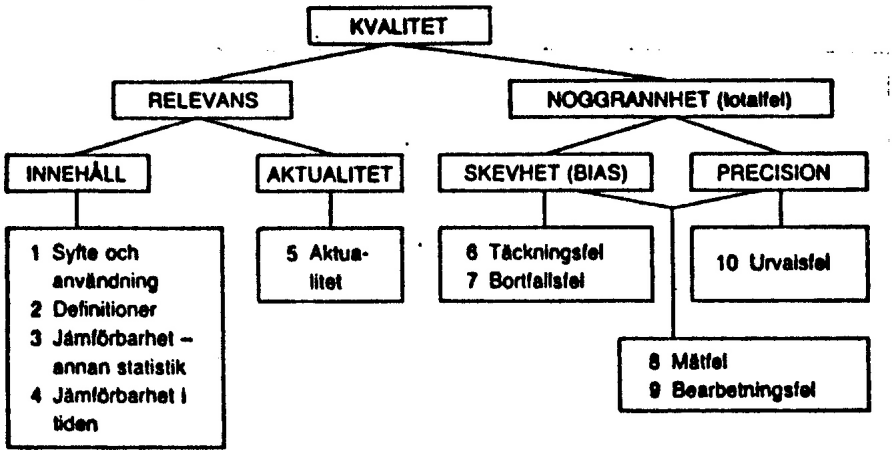
Ett företag har under en tidsperiod handlagt 5000 fakturor. De ekonomiska transaktonerna under perioden skall revideras. Beskriv hur ett urval om 50 fakturor kan dras. Förklara också vilken betydelse det har att urvalet är slumpmässigt.

9.7 Datakvalitet

9.7.1 *Datakvalitetsbegreppet*

De flesta varor som produceras och säljs har en kvalitetsdeklaration. Sådana kvalitetsdeklarationer har kommit till för att vi som konsumenter lättare skall kunna avgöra produktens användbarhet i olika situationer och hur produkten skall hanteras för att bättre komma till sin rätt. Produktion av data utgör i detta sammanhang inte något undantag. För att data skall kunna användas i kunskapsbildningsprocessen krävs att datas kvalitet är känd och uppfyller de krav som undersökningens syfte ställer. Vad som avses med datakvalitet är inte självklart, men bedömningen av datakvaliteten brukar beröra kopplingen mellan de egenskaper data har och de behov som användaren har.

Ett schema för kvalitetsbegreppets innehåll vid dataproduktion visas i figur 9.3. Schemat har utvecklats av Statistiska Centralbyrån. Här har kvalitetsbegreppet delats upp i en relevansaspekt och en noggrannhetsaspekt. För ett givet problem innebär en hög datakvalitet att data både är relevanta för problemet och har en hög noggrannhet. Notera att detta innebär att ett datamaterial som har en hög kvalitet för ett problem kan visa sig ha en låg kvalitet för ett annat problem. Vi skall nu diskutera dessa två kvalitetsaspekter något närmare.



Figur 9.3 Kvalitetsbegreppet. Källa:

9.7.2 *Relevansfaktorer*

Begreppet relevans sammanhänger med behov och användning och bedöms således relativt den aktuella undersökningen. Vi delar här upp relevansaspekten på datakvalitet i två delar, innehåll respektive aktualitet.

För att undvika relavansfel med avseende på innehåll är det viktigt att dataproducenten anger dataproduktionens syften och förutsättningar. Det är således nödvändigt att lämna en redogörelse om använda definitioner och om använda mätinstrument (frågeformulär etc) vid dataproduktionen.

Allt eftersom samhället förändras över tiden förändras även människors värderingar, attityder och beteenden. Exempelvis är listan över brottsliga aktiviteter och de straffsatser som utmäts för brotten annorlunda idag än de som fanns för 20 år sedan. Likaså finns skillnader mellan olika länder vid samma tidpunkt. Något som är brottsligt i ett land behöver inte nödvändigtvis vara det i ett annat. Vi inser därmed att okritiska jämförelser av förändringar över tiden och olikheter mellan länder kan vara mycket vanskliga. Vi behöver omfattande kunskaper om problemområdet för att minska risken för allvarliga fel i slutsatserna. Jämförbarhet med annan statistik och jämförbarhet i tiden är således två viktiga komponenter i innehållsaspekten.

Vid sidan av innehållsaspekten är datas aktualitet en viktig relevansfaktor. Självfallet har data som speglar förhållanden från en annan tid än den som studeras i undersökningen ett begränsat värde. Detta är ett stort problem för de samhällsinstitutioner där snabba reaktioner på förändringar är nödvändiga om de data som skall ligga till grund för besluten har en mycket lång produktionstid.

9.7.3 *Noggrannhetsfaktorer*

Noggrannhetsbegreppets huvudkomponenter är skevhet och precision. När datamaterialets skevhet diskuteras i detta sammanhang avses felkällor som kan ge upphov till systematiska fel. Vid bedömning av precision, å andra sidan, beaktas risken tlll eventuella slumpmässiga fel.

Täckningsfel och bortfallsfel är två felkällor som bidrar till skevheten. Täckningsfel

uppstår då populationens individer och den förteckning över dem som avvänds i undersökningen (t ex en urvalsram) inte överensstämmer. Dels kan det finnas individer som tillhör populationen men inte finns med i förteckningen (undertäckning), dels kan det finnas individer i förteckningen som inte tillhör populationen (övertäckning). Bortfallsfel uppstår då uppgifter från de individer man vill mäta inte kan erhållas (t ex på grund av sjukdom och svarsvägran).

Urvalsfel uppkommer då urvalsundersökningar används i stället för totalundersökningar. Om sannolikhetsurval används blir urvalsfelet slumpmässigt och urvalsfelets storlek kan kontrolleras. Därför räknar vi urvalsfelet som en precisionsfaktor vid bedömningen av datakvaliteten.

Mätfel och bearbetningsfel kan båda bidra till skevhet och precisionsförluster. Mätfel beror på fel och oklarheter i mätinstrument. Bearbetningsfel (kodningsfel, granskningsfel, utskriftsfel etc) uppstår under bearbetningen av data.