Handbok 2008:1

Statistiska centralbyrån Statistics Sweden

Urval – från teori till praktik

Urval
– från teori till praktik

Statistiska centralbyrån
2008

Handbook 2008:1

Sampling – from Theory to Practice

Statistics Sweden
2008

Tidigare publicering
Previous publication

Registerstatistik – en introduktion
Register statistics – an introduction
Experimentell utvärdering – en handbok
Evaluation through Experiments – a handbook

Producent
Producer

SCB, utvecklingsavdelningen
Statistics Sweden, Research and Development
Box 24300
104 51 Stockholm
+46 8 506 940 00

Förfrågningar
Inquiries

Dan Hedlin, +46 8 506 943 34
dan.hedlin@scb.se

Annica Isaksson. +46 13 28 22 63
anisa@ida.liu.se

Tiina Orusild, +46 8 506 942 03
tiina.orusild@scb.se

Jörgen Svensson, +46 19 17 64 99
jorgen.svensson@scb.se

ISSN 1654-7268 (online)
ISSN 1654-7276 (print)
ISBN 978-91-618-1412-1 (print)
URN:NBN:SE:SCB-2008-X99BR0801_pdf (pdf)

Printed in Sweden
SCB-Tryck, Örebro 2008.03

Förord

Denna handbok, *Urval – från teori till praktik*, är resultatet av ett projekt som startade i januari 2005. Syftet var att ta fram en text om metoder för att göra statistiska urval från ändliga populationer. Boken skulle bidra till att överbrygga gapet mellan teori och praktik när det gäller urvalsmetodik. Den är avsedd som vägledning vid undersöknings- och urvalsplanering för sådana statistiska undersökningar som utförs vid SCB.

I projektet ingick dispositionsarbete, inventering av urvalsundersökningar på SCB, anordnande av fokusgrupper med statistiker på SCB, litteraturinsamling och en skrivfas.

Många personer bistod på olika sätt under arbetets gång. Carl-Erik Särndal tjänade som en mentor redan från start. Deltagarna i fokusgrupperna, ledda av Ing-Mari Boynton och Ulla Backman, gav viktiga insikter i förväntningar och önskemål om skriftens innehåll och utformning. Gunnar Arvidson gjorde stora insatser bland annat under inventeringen.

Referensgruppen bestående av Martin Axelson, Roland Friberg, Annika Lindblom, Sixten Lundström, Lennart Nordberg och Marianne Ängsved gav allmän vägledning och expertgranskade olika versioner av den framväxande texten. Clara Ferdman, Marina Jansson, Jenny Karlsson, Åsa Rosendahl, Frida Videll och Anna-Karin Westöö hjälpte till att öka läsvänligheten och användbarheten hos den framväxande texten. Martin Ribe har bidragit med kloka synpunkter under slutfasen av arbetet. Arne Orrgård har designat omslaget och flertalet figurer. Även många andra personer har hjälpt till med att kontrollera fakta, svara på frågor och bistå med formuleringar om verksamheter inom SCB som projektgruppen inte har varit förtrogen med.

Projektgruppen vill framföra ett varmt tack till alla nämnda och till alla andra som på olika sätt hjälpt till att ro projektet i hamn.

Avsikten är att texten ska hållas levande, som stöd för utformning och utförande av urvalsprocesser inom SCB.

Arbetet med denna handbok har utförts av följande projektgrupp:
Dan Hedlin, Annica Isaksson, Tiina Orusild och Jörgen Svensson.

Statistiska centralbyrån i mars 2008

Folke Carlsson

Innehåll

Sammanfattning

Denna handbok syftar till att överbrygga gapet mellan teori och praktik när det gäller urvalsmetoder för statistiska undersökningar.

De två första kapitlen ger en överblick. Kapitel 3 tar upp skillnaden mellan ett register och en urvalsram. Begrepp såsom intressepopulation, målpopulation och rampopulation tas upp. Andra frågor som diskuteras är bland annat när och hur en urvalsram ska tas fram.

Kapitel 4 innehåller praktiska råd om klassiska urvalsmetoder såsom obundet slumpmässigt urval, stratifierat urval och urval i flera steg eller faser. Avsnittet om πps-urval går igenom systematiskt πps-urval och de moderna metoderna Pareto πps-urval och PoMix-urval. Kapitlet tar också upp urval ur kontinuerliga populationer och icke-sannolikhetsurval.

Kapitel 5 handlar om samordning av urval, ramar och undersökningar. Metoder som diskuteras är panelurval och samordning genom permanenta slumptal. Ett avsnitt beskriver SAMordnad Urvalsdragning, SAMU, som är det system som SCB använder för samordning av urvalsramar och urval mellan olika företagsundersökningar eller samordning över tid inom en företagsundersökning. En annan fråga som tas upp i kapitlet är en förändrings beståndsdelar i en upprepad undersökning.

I kapitel 6 ligger tyngdpunkten på valet mellan urvalsmetoder och bestämning av urvalsstorlek. Olika sätt att skatta populationsvariansen redan på planeringsstadiet och andra praktiska frågor diskuteras. Kapitlet går också igenom starka och svaga sidor hos urvals- och totalundersökningar, registerbaserade undersökningar och kvalitativa studier.

Handboken innehåller många exempel hämtade från SCB. De första sex kapitlen har åtskilliga mindre exempel i nära anslutning till det stoff som gås igenom. Kapitel 7 innehåller fem större exempel.

De tre sista kapitlen innehåller råd om statistisk konsultationsteknik, som är användbar i planering av undersökningar, en checklista med ingångar till handboken och rekommendationer av böcker om teorin för urvalsundersökningar.

Ett stort antal referenser ger tips om ytterligare fördjupning.

1 Inledning

Statistiska undersökningar kan ha många skepnader. Denna handbok fokuserar på urvalsundersökningar, dvs. undersökningar där man gör urval av objekt ur en större grupp och använder insamlade data för att dra slutsatser till hela gruppen. I första hand behandlas undersökningar som baseras på slumpmässiga urval från en urvalsram. Vidare fokuserar framställningen på urvalsmetoder, och avstår därmed från att ägna alltför stor uppmärksamhet åt urvalsundersökningens många andra moment, såsom datainsamling och estimation. Denna avgränsning är artificiell och svår att genomföra helt konsekvent, eftersom alla moment i själva verket hänger nära samman. När man gör urvalet måste man exempelvis ta hänsyn till hur data ska samlas in, och valet av urvalsmetod har stora konsekvenser för estimationen. Där så krävs tar handboken ett vidare grepp och behandlar frågor som inte bara rör urvalsmetodik. I kapitel 10 finns ett antal referenser till böcker om urval och estimation.

Denna text gör inga anspråk på att ersätta eller konkurrera med de många utmärkta böcker som finns i urvalsteori. Mycket har hänt sedan Neymans klassiska artikel om stratifierat urval på 1930-talet, och urvalsteorin erbjuder i dag en stor verktygslåda av metoder och resultat. När teori ska omsättas i praktik måste dock bedömningar och avvägningar göras. Det kan handla om hur man i praktiken konstruerar en bra urvalsram eller väljer en lämplig urvalsmetod för sin situation. Bland erfarna statistiker på SCB finns lång erfarenhet av att göra denna typ av bedömningar. I den här skriften är en avsikt att sätta något av denna samlade erfarenhet på pränt.

Som en del av förarbetet till denna skrift inventerades vilka urvalsundersökningar som genomförs på SCB. Här återfinns en rik flora av undersökningar och exempel på många olika urvalsmetoder. Olika typer av undersökningar visade sig ofta ha specifika urvalsproblem. Valet av metod beror bland annat av

- vilket slags *objekt* man undersöker (t.ex. företag eller organisationer, individer eller hushåll, eller fenomen i naturen)
- vilket *ämnesområde* man behandlar (t.ex. arbetsmarknad, välfärd eller energianvändning)
- med vilken *frekvens* undersökningen genomförs (t.ex. löpande undersökningar eller engångsundersökningar)
- om undersökningen är en stor löpande undersökning eller en liten uppdragsundersökning.

Framställningen begränsas inte avsiktligt till någon viss typ av undersökning. Däremot finns inte möjlighet att gå till botten med alla urvalsproblem i alla olika sorters urvalsundersökningar som genomförs på SCB.

Det här är som sagt ingen lärobok i urvalsteori, utan närmast en guide till hur befintlig teori kan omsättas i praktiken under de förutsättningar som gäller på SCB. Därför saknar framställningen också härledningar och bevis, men erbjuder en hel del referenser för den vetgirige.

I samråd med en referensgrupp för projektet har målgruppen för handboken definierats som i första hand metodstatistiker som arbetar med urvalsundersökningar på SCB och som är nya i denna roll. Läsaren förutsätts ha förkunskaper i urvalsteori motsvarande en grundläggande universitetskurs. Läsare som saknar denna grund rekommenderas att parallellt läsa någon god bok i urvalsteori, exempelvis Cochran (1977) eller Lohr (1999) eller någon annan av de böcker som nämns i kapitel 10. Den bok som oftast åberopas i denna handbok är Särndal et al. (1992). Det förutsätts inte att läsaren är insatt i den boken, men det kan vara praktiskt att ha den till hands som referensbok.

2 Översiktligt om urvalsplanering

I det här kapitlet är avsikten att ge en översikt över olika begrepp och frågeställningar i anslutning till urvalsundersökningar. Mer detaljerade genomgångar kommer i kapitel 3–6.

2.1 Hur planeras en undersökning?

En statistisk undersökning avser att ge sammanfattande mått på de relevanta egenskaperna hos objekten i målpopulationen. Om undersökningen avser dem som var folkbokförda i Sverige ett visst datum är *objekten* de folkbokförda individerna, och populationen är befolkningen. *Målpopulationen* är den population man vill uttala sig om.

Statistiska undersökningar är av olika karaktär, bland annat beroende på vilka objekten och egenskaperna är. Hur en undersökning planeras beror även på en mängd andra faktorer. Det här avsnittet går kortfattat igenom de viktigaste faktorerna. I följande kapitel ges mer djupgående diskussioner.

Objekten som undersöks kan vara människor, företag, institutioner, konsumentvaror, lastbilar osv. Objekten kan också vara t.ex. sjöar, djur eller växter av någon viss art. En egenskap eller *variabel* hos en individ kan vara individens inkomst. En viktig variabel hos en konsumtionsvara är dess pris. Hos lantbruk är det bland mycket annat skördens storlek som är intressant.

Det som ska skattas i en statistisk undersökning kallas statistisk storhet eller *parameter*. Parametern kan involvera mer än en variabel. Exempelvis utgör kroppsvikt delat med längd i kvadrat *Body Mass Index* (BMI). Parametern kan även bero av något som utvecklas över tid, t.ex. prisförändring. Populationen kan då vara varor som funnits vid två tillfällen, variablerna är pris vid första och andra tillfällena, och parametern är kvoten av genomsnittspriserna vid de två tillfällena.

Den beräkningsregel som används i parametern kallas *statistiskt mått*. Det kan vara t.ex. medelvärde eller median, eller kvoten av två medelvärden. Ett annat vanligt statistiskt mått är en *populationstotal* av någon variabel, dvs. summan av variabelvärdena för alla objekt i populationen.

Begreppen diskuteras mer utförligt i Statistiska centralbyrån (2001a, bilaga 1).

Hur en undersökning planeras, styrs även av praktiska faktorer. En av dessa är kontaktvägen till uppgiftskällan. Anta att syftet är att dra slutsatser om Sveriges befolkning, som då alltså är målpopulation. Då kan man dra ett urval ur befolkningsregistret, som inkluderar adresser till alla registrerade individer. Hur urvalet ska göras i praktiken beror på undersökningens syfte och förutsättningar. Säg att undersökningen avser fisken i Östersjön. Om uppgiftslämnarna är fiskare kan datainsamlingen ske via enkät. Ett problem är då att fiskare inte kan identifieras i befolkningsregistret. Om data samlas in via direkt observation, t.ex. genom provfiske, görs både urval och insamling av uppgifter på ett helt annat sätt.

Det är enklast att dra ett urval om det finns en lista över alla objekt i populationen. En sådan lista kallas *urvalsram* eller kortare *ram*. Om listan dessutom innehåller en praktisk kontaktväg till objekten, t.ex. telefonnummer eller adress, kan även en storskalig undersökning genomföras relativt smidigt. Under vissa förutsättningar går det att dra ett urval även om man saknar en ram i vanlig bemärkelse. Det gäller exempelvis vid urval ur kontinuerliga populationer, som behandlas i avsnitt 4.8.

Rampopulationen är den mängd objekt som kan nås med hjälp av den information som finns i ramen. Rampopulationen skiljer sig ofta något från målpopulationen. Låt oss säga att målpopulationen är alla aktiva företag i Sverige. Det finns ett register över alla företag i Sverige, men det registret innehåller uppgifter som ännu inte har uppdaterats. Företag startas, läggs ned och ombildas. Vissa objekt i målpopulationen finns därmed inte i ramen; de är inte nåbara, åtminstone inte utan extra insatser. Omvänt ingår vissa objekt i ramen inte i målpopulationen. Det kan gälla t.ex. nedlagda företag som kan kvarstå i registret med aktiv status. Med ramens *täckning* avses hur väl ramen täcker målpopulationen. I kapitel 3 beskrivs några av SCB:s register.

Sannolikhetsurval kan dras på många sätt. Det enda grundläggande kravet är att varje nåbart objekt i målpopulationen kan ingå i urvalet och detta med en känd, positiv sannolikhet. Det finns många etablerade metoder att dra sannolikhetsurval; ett exempel är stratifierat obundet slumpmässigt urval. Kapitel 4 ger en detaljerad genomgång av de vanligaste och viktigaste urvalsmetoderna vid SCB.

När man planerar hur urvalet ska göras i en undersökning behöver man beakta vilken sorts objekt som ingår i populationen, vilka kontaktvägar som finns till objekten och vad för sorts information man söker om populationen. Det är stor skillnad mellan det slags urval som är optimalt för att uppskatta – eller kortare *skatta* – hur stor andel av populationen av företag som har en viss egenskap, och den sorts urval man behöver för att skatta alla företags samlade produktion. Ett urval ur en ram med objekt som är heterogena med avseende på det man vill undersöka dras på ett helt annat sätt än ett där objektens variabelvärden inte skiljer sig mycket åt.

Upprepas undersökningen vid många tillfällen behöver man en rutin för att successivt byta ut objekten i urvalet, inte minst därför att man annars löper risken att urvalet inkluderar allt flera uppgiftslämnare som tröttnat på att medverka i undersökningen. Ett par tekniker för att behålla respektive byta ut objekt – *panelurval* och urval genom användande av *permanenta slumptal* – behandlas i avsnitt 5.4.

När uppgifterna har samlats in förestår en mängd beräkningar. Dessa utgör länken mellan de enskilda uppgifterna, dvs. objektens egenskaper, och den sammanfattande informationen, som är statistiken. Den algoritm som man använder kallas *estimator* eller *skattningsförfarande*. Det tal som beräkningen ger kallas *skattning* eller *estimat*.

Det finns sällan ett enda vedertaget sätt att beräkna en skattning av en och samma parameter. Olika tänkbara estimatorer har olika egenskaper och ett val måste göras. Valet av estimator påverkar ofta valet av urvalsdesign och vice versa. I ramen finns ofta uppgifter om objekten, såsom adress, kön,

ålder och taxerad inkomst för individer. Dessa uppgifter kan användas när man drar urvalet eller beräknar skattningarna.

I början av detta avsnitt togs för givet att någon har bestämt vilka de intressanta parametrarna är. Det kan ligga ett omfattande planeringsarbete bakom detta beslut. Inför undersökningsplaneringen har man ett sakproblem som ska omformuleras till ett statistiskt problem. Sakproblemet kan t.ex. vara att belysa hur skörden av havre per hektar förändras från år till år. I den statistiska formuleringen av problemet definieras begreppen skörd och areal och den population av lantbruksföretag som avses. Begreppen kan behöva operationaliseras på olika sätt beroende på datainsamlingsmetod. Exempelvis kan man samla in information på en mer detaljerad nivå vid direkt observation än i en telefonintervju. I det statistiska problemet ingår även hur man vid små förändringar av skattad havreskörd ska avgöra om det är en slumpens nyck eller en verklig förändring.

Mängden av skattningar som undersökningen avses att resultera i kallas *tabellplan*. I en tabellplan finns parametrarna, de populationsstorheter som ska skattas. En parameter kan i detta sammanhang sägas vara definierad av en specifikation med fyra komponenter: grupp (population och redovisningsgrupper), objekt, variabler och statistiskt mått. Tabellplanen används för att konkretisera och kommunicera målet för undersökningen. För vissa undersökningar kan tabellplanen behöva vidgas till en mer långtgående *analysplan*, även om det i SCB-sammanhang väsentliga oftast stannar vid tabellplanen.

Det övergripande målet när man planerar en undersökning är att få så bra kvalitet som möjligt med tillgängliga resurser. Det är viktigt att redan i planeringen tänka på den önskade statistiken och de resurser undersökningen kan behöva. Vilka krav som ställs på statistiken bestäms ytterst av dess avsedda användning. Det finns dock inbyggda konflikter mellan kraven, bland annat mellan snabbhet och tillförlitlighet. För officiell statistik gäller dessutom åtskilliga formella krav på t.ex. arkivering, dokumentation, röjandekontroll av enskilda uppgifter och öppenhet vad gäller metoder och resultat.

2.2 Hur utformas urvalet?

2.2.1 Begreppen sannolikhetsurval och inklusionssannolikhet

I denna skrift utgår vi vanligen från att sannolikhetsurval används för att välja ut objekt för observation. Som nämndes i avsnitt 2.1 är de grundläggande kraven för sannolikhetsurval att varje nåbart objekt i målpopulationen kan ingå i urvalet och det med en känd, positiv sannolikhet. En formell definition av sannolikhetsurval finns i Särndal et al. (1992, s. 8–9). Om man använder sig av någon av de standardmetoder för urval som behandlas i avsnitt 4.1–4.8, kan man vara säker på att kriterierna för sannolikhetsurval är uppfyllda. Om man däremot utvecklar en urvalsmetod på egen hand är det nödvändigt att stämma av den mot den formella definitionen för att utröna om den verkligen åstadkommer ett sannolikhetsurval.

I praktiken tvingas man till avsteg från sannolikhetsurval. Ofta undantas en del objekt i målpopulationen från undersökningen trots att de är nåbara. Ett vanligt exempel är småföretag i företagsundersökningar. De kommer

inte att ha någon positiv sannolikhet att väljas. Det förekommer också andra planerade avsteg från sannolikhetsurval: det kan handla om provundersökningar eller andra undersökningar där det är särskilt bråttom eller där urvalsram saknas. Detta gör det motiverat att behandla även icke-sannolikhetsurval, se avsnitt 4.9. Även bortfall leder till avvikelser från sannolikhetsurval; bortfall behandlas enbart kortfattat i denna bok.

Producenter av officiell statistik, såväl i Sverige som utomlands, brukar så långt det är möjligt sträva efter att använda sannolikhetsurval i sina urvalsundersökningar. Det finns två huvudskäl till detta. Dels är sannolikhetsurval ett sätt att undvika att introducera systematisk avvikelse mellan skattningen och populationsparameterns värde: en avvikelse som beror på att man systematiskt väljer ut – eller inte väljer ut – objekt med vissa egenskaper. Dels uppfattas sannolikhetsurval i allmänhet som objektiva, vilket ger ökad tilltro till den statistik man producerar.

Sannolikheten för ett objekt att ingå i urvalet brukar kallas *inklusionssannolikhet*. Det finns inget krav på att inklusionssannolikheterna måste vara lika för att urvalet ska få kallas ett sannolikhetsurval. Det är tvärtom vanligt att de är olika. Inklusionssannolikheterna används i skattningsförfarandet. Mer formellt talar man om första ordningens inklusionssannolikheter när det gäller sannolikheter för enskilda objekt, och andra ordningens inklusionssannolikheter när det gäller sannolikheter för par av objekt. Inklusionssannolikheter av högre ordningar definieras på motsvarande sätt. En *urvalsdesign* definieras av funktionen $p(\cdot)$, där $p(s)$ ger sannolikheten att välja s, och s är ett av alla möjliga urval under given design. Olika urvalsdesigner kan ha samma värden på första ordningens, men olika värden på andra ordningens, inklusionssannolikheter. Exempelvis kan första ordningens inklusionssannolikheter sättas till 0,1 för alla objekt om man drar ett obundet slumpmässigt urval (OSU), ett stratifierat OSU med proportionell allokering eller ett systematiskt urval. Eftersom andra ordningens inklusionssannolikheter är olika skiljer sig variansskattningarna åt.

Ett begrepp med nära koppling till inklusionssannolikhet är *designvikt*. Designvikten för ett objekt beräknas som inversen av första ordningens inklusionssannolikhet. Ett objekt som har stor sannolikhet att väljas har alltså liten designvikt och vice versa. En tolkning av designvikten är att ett objekt som ingår i urvalet representerar så många objekt som designvikten anger. Om designvikten t.ex. är fem representerar objektet sig självt och fyra andra, inte utvalda, objekt. Ett vanligt sätt att skatta en populationstotal är att summera de observerade värdena för utvalda objekt dividerade med deras inklusionssannolikheter. Detta ger samma resultat som om man summerar objektens värden multiplicerade med deras designvikter. Vid förekomst av bortfall, ramproblem och liknande justeras dock vikterna.

Ett dragningsschema eller *urvalsschema* är en algoritm för urvalsdragning. Olika urvalsscheman kan användas för att åstadkomma samma urvalsdesign. I kapitel 4 finns exempel på urvalsscheman för OSU och några varianter av πps-urval.

2.2.2 Urval i ett eller flera steg

Urvalsmetoder kan klassificeras på olika sätt. En indelning grundar sig på hur många urvalssteg som används, och detta styrs i sin tur i allmänhet av vilka urvalsramar man har tillgång till. Om man har en bra ram över

objekten i populationen kan man göra urvalet direkt från denna. Detta kallas att man gör ett *direkturval*. I Sverige är vi väl försedda med register och kan ganska ofta göra direkturval. Ibland är situationen dock sådan att man önskar eller måste göra urval av *kluster*, dvs. grupper av objekt. Utvalda kluster kan man antingen totalundersöka eller dra ett urval från. Det senare brukar kallas för ett *tvåstegsurval*. Det finns inget i teorin som hindrar att man fortsätter att göra urval i tre eller flera steg. Med ett samlingsnamn kallas sådana urval för *flerstegsurval*. Läs mer om klusterurval i ett och flera steg i avsnitt 4.6.

2.2.3 Användning av hjälpinformation i urvalsdesignen

Ju mer kunskap man har om objekten i populationen innan urvalet dras, desto större är handlingsutrymmet vad gäller utformningen av urvalsdesignen. Om urvalet ska göras från en ram som endast består av en lista av objekt är man hänvisad till enkla metoder som obundet slumpmässigt urval (OSU) eller systematiskt urval. Sådana urvalsmetoder leder ofta till skattningar med stor varians.

Den kännedom man har om egenskaper och förhållanden hos objekten i populationen utöver det man undersöker kallas sammanfattningsvis för *hjälpinformation*. Vidare är *hjälpvariabler* sådana variabler som kan användas till att förbättra urvalsdesign och skattningar.

Om ramen innehåller någon eller några relevanta hjälpvariabler öppnar sig möjligheter till stratifiering och till olika typer av urval proportionella mot något storleksmått. Om hjälpvariabeln har något samband med det man vill undersöka kan man på detta sätt, med samma urvalsstorlek, åstadkomma skattningar med mindre varians än vid OSU. En undersökning har i allmänhet flera undersökningsvariabler och flera redovisningsgrupper. En viss hjälpvariabel kan ha ett starkt samband med en viss undersökningsvariabel men knappast något samband alls med en annan. I avsnitt 6.4 diskuteras användning av hjälpinformation mer utförligt.

2.2.4 Bestämning av urvalsstorlek

En fråga som en statistiker ofta ställs inför är ”hur stort urval ska jag dra?”. Frågan är enkel att ställa men inte fullt lika enkel att besvara, eftersom en rad faktorer behöver beaktas. Här behandlas några av de viktigaste. En mer utförlig beskrivning finns i avsnitt 6.1.

Precision: Hur stor varians i mina skattningar är jag beredd att acceptera?

Rent allmänt gäller att ju större varians man kan godta, desto mindre urval behöver man. Frågan om varians och urvalsstorlek blir särskilt viktig om skattningar krävs för små redovisningsgrupper.

En vanlig missuppfattning är att populationens storlek alltid har betydelse för hur stort urvalet behöver vara. Faktum är att populationens storlek i många urvalsundersökningar bara har marginell betydelse för urvalsstorleken.

Kostnad: Vad får datainsamlingen kosta? Kostar alla objekt lika mycket att undersöka?

I de flesta läroböcker i urvalsteori kan man hitta färdiga formler för storleken på urvalet. Formlerna är lösningar till ett matematiskt problem: för

vilken urvalsstorlek minimeras variansen givet den budgetrestriktion man har? Det förekommer också att man i stället formulerar problemet ”för vilken urvalsstorlek blir undersökningen så billig som möjligt utan att variansen överskrider ett förutbestämt värde?”. Anledningen till att det finns flera formler är att variansen ser olika ut beroende på vilken urvalsdesign man tänker använda och vilken parameter man vill skatta. Budgetrestriktionen uttrycks vanligen som en funktion av en fast kostnad och en rörlig styckkostnad för att undersöka ett utvalt objekt. I praktiken ser kostnadsbilden ofta betydligt mer komplicerad ut. På SCB samlas exempelvis stora mängder data in genom telefonintervjuer. En realistisk kostnadsfunktion för en sådan undersökning bör förutom kostnaden för genomförd intervju innehålla bland annat kostnad per intervju fram till första kontaktförsöket (inkluderar t.ex. kostnad för spårning och introduktionsbrev) och kostnad för varje ytterligare kontaktförsök.

Övertäckning i ramen och bortfall: Hur stort kommer mitt slutliga urval av svarande objekt tillhörande målpopulationen att bli?

Läroböckernas vanliga formler för bestämning av urvalsstorlek förutsätter en ideal värld där det går att observera alla objekt som väljs ut. I praktiken lider de allra flesta urvalsundersökningar av bortfall i större eller mindre utsträckning.

Uppgiftslämnarbörda: Hur många uppgiftslämnare är det försvarbart att belasta?

En annan faktor, utöver bortfall, som läroböckerna inte brukar ta hänsyn till vid bestämning av urvalsstorlek är uppgiftslämnarbördan (uppgiftslämnarkostnaden). Denna faktor är extra problematisk i och med att den drar i en annan riktning än precisionen: från precisionssynpunkt önskas så stort urval som möjligt, men med tanke på kravet att minska uppgiftslämnarbördan önskas så litet urval som möjligt. Här måste det alltså till en kompromiss. Ett sätt att hålla uppgiftslämnarbördan nere för enskilda uppgiftslämnare är att samordna sin undersökning med andra undersökningar. Läs mer om detta i kapitel 5.

2.2.5 Läs mer

Psykologer har studerat hur människor uppfattar urvalsstorlekens betydelse för hur säkra slutsatser man kan dra om det fenomen man studerar. Det har då bland annat visat sig att människor har svårt att intuitivt inse betydelsen av urvalsstorleken på en skattnings varians. Ett försök till förklaring av vad detta kan bero på finns i Sedlmeier och Gigerenzer (1997).

2.3 Hur görs uppskattningar för populationen?

2.3.1 Parametrar och estimatorer

En parameter kan som nämndes i avsnitt 2.1 normalt sägas vara definierad av en specifikation med fyra komponenter: grupp, objekt, variabler och statistiskt mått.

Olika typer av parametrar ställer olika krav på estimationen. Till de parametrar som är enklast att skatta hör de som inbegriper totaler (summor), medelvärden och andelar. Dessa är också enkla funktioner eller specialfall av varandra: ett medelvärde är en total dividerad med populationsstorle-

ken; en andel är ett medelvärde för en undersökningsvariabel som bara kan anta värdena ett och noll. Andra parametrar är mer komplicerade, då punktestimatorernas varianser är besvärliga att härleda. Hit hör exempelvis medianer (och andra kvantiler och funktioner av kvantiler), Gini-koefficienter, som brukar användas som mått på ojämlikhet, och vissa index.

Ett sätt att skatta en populationstotal är med hjälp av den s.k. *Horvitz–Thompson-estimatorn* (*HT-estimatorn*). Med den estimatorn skattas totalen med summan av produkten av de observerade värdena för utvalda objekt och objektens designvikter. Se Särndal et al. (1992, Result 2.8.1) för en formell beskrivning. För parametrar som är funktioner av totaler kan de ingående totalerna skattas med HT-estimatorer. Enkla exempel på sådana parametrar är populationsmedelvärden och populationsandelar.

Om man använder HT-estimatorn för att skatta en total så garanterar teorin att skattningarnas värden i genomsnitt kommer att pricka populationstotalens värde. Här menas ett genomsnitt över skattningarna baserade på alla möjliga urval valda med samma design. Denna träffsäkerhet, som kallas *väntevärdesriktighet*, gör att HT-estimatorn brukar användas som en referenspunkt som man jämför andra tänkbara estimatorer med. Observera dock att HT-estimatorn är väntevärdesriktig bara under förutsättning att det inte finns ramfel, bortfall eller andra icke-urvalsrelaterade problem som på ett systematiskt sätt påverkar resultaten.

HT-estimatorn är inte alltid den *effektivaste* möjliga: den har med andra ord inte alltid minsta möjliga varians. Ett sätt att minska variansen utan att öka urvalsstorleken är att utnyttja eventuell hjälpinformation i estimationen. Detta går i sin tur att göra på olika sätt: man kan använda sig av en poststratifierad estimator, en regressionsestimator eller någon annan variant av den generaliserade regressionsestimatorn (GREG-estimatorn), se Särndal och Lundström (2005, formel (4.7)) eller Särndal et al. (1992, formel (6.4.1)). GREG-estimatorn utnyttjar hjälpinformation och är normalt mer effektiv än HT-estimatorn om hjälpinformationen samvarierar med aktuell undersökningsvariabel.

2.3.2 Statistikens osäkerhet

I den teoretiska bedömningen av hur bra en estimator är brukar man titta på dess *tillförlitlighet*. Tillförlitlighet handlar om skillnaden mellan skattning och ”sant” värde. Detta sanna värde är ett okänt tal som enkelt kan uppfattas som resultatet av en tänkt "ideal" undersökning, som inte störs av några osäkerheter eller felkällor. Skillnaderna kan uppstå till följd av urval, ramfel, bortfall, mätfel m.m. – se vidare Statistiska centralbyrån (2001a). Ett sammanfattande mått på en estimators kvalitet är dess totalfel eller *medelkvadratavvikelse* (eller *medelkvadratfel)*. Medelkvadratavvikelsen definieras som estimatorns kvadrerade bias plus dess varians. Andra förekommande uttryck för bias är *skevhet* och *systematiskt fel*. I valet mellan olika estimatorer föredrar man i allmänhet den som har minst medelkvadratavvikelse, i den mån detta går att bedöma.

Metoden att dra ett sannolikhetsurval leder till ett *urvalsfel*, som mäts med estimatorns varians. Man tänker sig (hypotetiskt) att man drar alla möjliga urval med vald urvalsdesign och storlek, och beräknar estimat för vart och ett av dem. Estimatorns varians är ett mått på spridningen av estimaten.

När statistiken publiceras kan man använda något av följande precisionsmått:

- **Konfidensintervall.** Den vanligaste formen av osäkerhetsmått är konfidensintervallet. Anta att man drar alla möjliga urval med vald urvalsdesign och storlek, och beräknar ett konfidensintervall för vart och ett av dem. Om konfidensgraden är 95 procent kommer de erhållna intervallen att innesluta det sanna parametervärdet i ungefär 95 fall av 100. Halva längden på ett 95-procentigt konfidensintervall brukar kallas *osäkerhetsmarginal* eller *felmarginal*.
- **Relativ osäkerhetsmarginal.** Kvoten mellan osäkerhetsmarginal och punktskattning.
- **Skattat relativt medelfel.** Kvoten mellan estimatorns skattade medelfel och punktskattning. Medelfelet är detsamma som skattningens standardavvikelse, dvs. kvadratroten ur variansen.

Mer om mått på precision finns i avsnitt 6.1.1. Ovanstående mått behandlar endast osäkerheten till följd av att man gör ett urval i stället för att totalundersöka. Övriga osäkerhetskällor, såsom bortfall, ramfel och mätfel, måste förstås också kommenteras och i möjligaste mån kvantifieras. Ibland tvingas man nöja sig med enkla kvalitetsindikatorer såsom svarsandel.

2.3.3 Hantering av bortfall och mätfel

På SCB (liksom hos de flesta andra nationella statistikbyråer) är bortfall i undersökningar ett växande problem. Sedan 1980-talet har bortfallsnivåerna nära nog fördubblats i många SCB-undersökningar. Bortfallet kan påverka skattningarnas precision, men framför allt kan det leda till bias i skattningarna. Biasen riskerar att uppstå om de objekt för vilka data saknas skiljer sig på ett systematiskt sätt från de svarande med avseende på det man undersöker. Precisionen försämras i och med att bortfallet reducerar antalet svarande till en lägre nivå än urvalsstorleken. Utöver detta fördyrar bortfallet undersökningen genom att bortfallsuppföljning kostar pengar.

Det finns flera sätt att hantera bortfall i skattningarna. Två huvudansatser är imputering och vägning. Kombinationer av de två förekommer också, såsom att imputering används för att hantera partiellt bortfall och vägning för att hantera objektbortfall. På SCB används ofta en speciell vägningsmetod som kallas *kalibrering*.

Bortfallsreduktion och hantering av bortfall i urval och estimation ligger utanför ramen för denna skrift, men behandlas utförligt i Japec et al. (1997) och Lundström och Särndal (2001). Vad gäller kalibrering hänvisas även till Särndal och Lundström (2005).

En av de viktigaste processerna i statistiska undersökningar är mätprocessen. Med mätprocess avses då alla former av datainsamling. Fel i data leder till slumpmässiga eller systematiska mätfel. En stor del av Biemer och Lyberg (2003) handlar om mätfel och hur risken för sådana minskas så mycket som möjligt.

2.3.4 Läs mer

Tillförlitlighet förutsätter förekomsten av ett sant värde – ett begrepp som inte är helt oproblematiskt. Är det exempelvis meningsfullt att tala om ett sant värde när det gäller attityder till olika fenomen? Frågan är gammal och diskuterades redan på 1950-talet. För en operationell definition av sant värde, se Hansen et al. (1951).

3 Urvalsramar

Teorin för urvalsundersökningar förutsätter en väl definierad population. Om man dessutom använder sig av en bra urvalsram kan man göra effektiva urval och skattningar. Här diskuteras mer ingående vad detta innebär.

3.1 Vad är en population?

3.1.1 Populationen är en mängd av objekt

I statistiska sammanhang är en *population* vanligen en mängd av objekt som en undersökning ska uttala sig om. Exempelvis kan en opinionsundersökare vara intresserad av populationen röstberättigade. Objekt kallas ibland även enheter eller element. Denna skrift uppehåller sig främst vid ändliga populationer, i vilka antalet objekt är begränsat. I avsnitt 4.8 behandlas dock urval ur kontinuerliga (”oändliga”) populationer.

I samband med att man fastställer populationen för en undersökning måste man även välja hur objekten ska definieras. Detta moment är inte alltid självklart, utan måste tänkas igenom i det specifika fallet. I Företagsdatabasen, exempelvis, finns ett antal olika typer av enheter som kan användas som objekt. Juridiska enheter är lämpligast att använda i vissa undersökningar, medan de s.k. företagsenheterna är lämpligast i en del andra sammanhang. Se vidare avsnitt 3.3.2 för olika typer av objekt i företagsstatistik. I vissa undersökningar är man intresserad av att redovisa resultat för flera typer av objekt. I Arbetskraftsundersökningarna tar man exempelvis fram skattningar för både individer och hushåll.

Ett objekt som man hämtar in uppgifter om kallas ett *observationsobjekt*, och variabler som man samlar in värden på kallas *observationsvariabler*. En instans från vilken uppgifter inhämtas kallas en *uppgiftskälla*. Om uppgiftskällan är en person brukar man kalla henne eller honom för *uppgiftslämnare* eller *respondent*. Begreppen observationsobjekt, observationsvariabel, uppgiftskälla och uppgiftslämnare diskuteras mer utförligt i Statistiska centralbyrån (2001a, bilaga 1). Populationsobjekten och observationsobjekten sammanfaller oftast, men det finns undantag. Anta exempelvis att man är intresserad av vissa undersökningsvariabler kopplade till företag sådana att informationen inte finns tillgänglig centralt utan endast hos företagens regionala enheter. Observationsobjekten blir då de regionala enheterna. Ett annat exempel kan vara en undersökning av hushåll. Man är intresserad av variabeln hushållsinkomst, som för ett hushåll är summan av hushållsmedlemmarnas inkomster. Observationsobjekten utgörs av hushållsmedlemmar medan populationsobjekten är hushåll.

I undersökningar med flerstegsurval behöver man använda en hierarki av flera nivåer av populationer, som består av olika typer av objekt. Se vidare avsnitt 4.6.

Med ett s.k. nätverksurval kan man exempelvis nå en population av hushåll ”bakvägen” via en population av individer. *Urvalsobjekten*, i detta fall individer, är alltså inte lika med observationsobjekten som är hushåll. Se vidare avsnitt 4.5.

Ibland kan populationen bestå av alla objekt som har funnits någon gång under en referensperiod, varvid man kan tala om en flödes- eller periodvariant av en population. En population som består av objekt som finns vid en specifik referenstidpunkt kallas en stock- eller tidpunktsvariant av en population. I ekonomisk statistik är man exempelvis intresserad av flödesvariabler, såsom omsättningen under kalenderåret 2006. Även företag som bara existerat under en del av året har haft en omsättning. För att skatta omsättning används dock en tidpunktsvariant, varvid en viss underskattning uppstår.

Det finns sedan ett antal år flödesvarianter av företagsregistret, kallade ”årgångsramar”. Dessa tas fram årsvis och används bland annat för företagsdemografi som publiceras av Eurostat. Årgångsramarna innehåller alla företag som är eller har varit aktiva under året och har funnits i Företagsdatabasen (FDB). I december år t tas en definitiv version fram för kalenderåret t –1.

3.1.2 Olika populationsbegrepp

Man talar om olika slags populationer. Nedanstående begrepp hämtas främst från Statistiska centralbyrån (2001a), vars terminologi följs i denna handbok.

Med *intressepopulation* menas den mängd av objekt som man idealt skulle vilja uttala sig om, dvs. den objektgrupp som perfekt ansluter till användarens sakproblem. I praktiken kan man sällan rikta sig till hela intressepopulationen, eftersom det skulle bli för dyrt eller ta för lång tid.

Eftersom man inte rimligen kan nå hela intressepopulationen, riktar man sig i stället till en s.k. målpopulation. Med *målpopulation* avses den mängd av objekt som man väljer att använda i sin undersökning. Den överensstämmer vanligen inte helt med intressepopulationen, men bedöms ligga tillräckligt nära denna. Jämfört med denna ideala population, är målpopulationen lättare och billigare att undersöka. Den avgränsas från intressepopulationen genom ett aktivt val av undersökningsplaneraren.

När man avgränsar en målpopulation från intressepopulationen, kan man bara uttala sig om de objekt som ligger i målpopulationen, dvs. man kan enbart ta fram skattningar för målpopulationen. Intressepopulationen kan vara de i Sveriges befolkning som är över 16 år vid en viss tidpunkt, medan man i målpopulationen undantar människor som är permanent bosatta på institution. I diskussioner med beställare av undersökningar är det ofta fruktbart att diskutera i termer av intresse- och målpopulation.

Exempelvis exkluderas små objekt ibland ur målpopulationen. Om gränsen sätts relativt högt med avseende på objektens storlek, slipper man att ”jaga” små objekt i undersökningen. Då minskas uppgiftslämnararbetet och förbilligas undersökningen. Men nackdelen är att man då inte kan uttala sig om de små objekten och alltså får en brist i statistikens innehåll. Är gränsen alltför hög blir statistiken således inte helt relevant för användaren.

En annan populationsaspekt är tiden: ibland avser intressepopulationen senaste datum, medan målpopulationen avser tillgängligt datum.

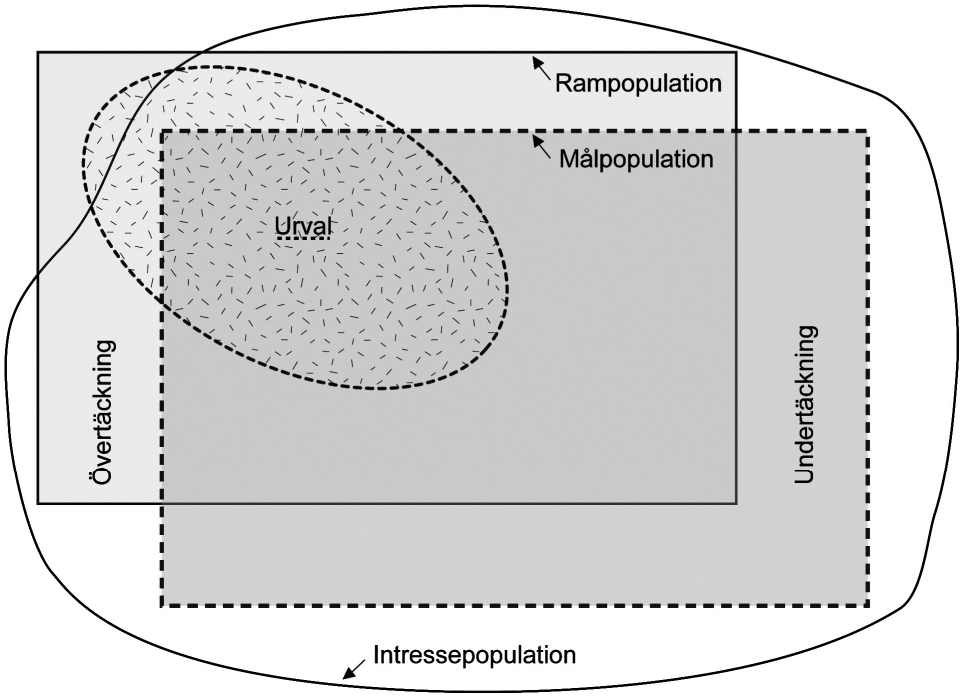
Ett principiellt annorlunda, vanligt förekommande, fall på SCB är att man gör ett s.k. *cut-off-urval*. Man avgränsar en del av målpopulationen som man inte samlar in några uppgifter från. Exempelvis avstår man i företagsundersökningar ibland från att samla in data från små företag. Trots denna avgränsning uttalar man sig om hela målpopulationen. Bidraget till aktuell parameter från den del av målpopulationen som inte alls undersökts skattas med hjälp av någon form av modell. För kvantitativa variabler förekommer att bidraget sätts till noll, vilket kan ses som en enkel form av modellantagande. Alternativt kan en sådan imputering med noll ses som ett sätt att "definiera bort" problemet, eller rättare sagt att övervältra problemet på den som ska tolka statistiken. Man kan även betrakta gruppen av objekt som inte undersöks som en form av avsiktlig undertäckning. Läs mer om cut-off-urval i avsnitt 6.2.4.

Ramen är den förteckning eller det utdrag ur register man utgår från i de flesta statistiska undersökningar. En karta kan också fungera som en ram. På SCB används ofta ramar i form av register som är utdrag ur större register. Med *rampopulation* menas den mängd objekt som kan nås med hjälp av den information som finns i ramen. Ramen är alltså den konkreta listan över objekt, medan rampopulationen är samma objekt sedda som möjliga undersökningsobjekt. Den nåbara delen av målpopulationen är den delmängd av rampopulationen som faller inom målpopulationen. Denna kallas *undersökningspopulation*.

Rampopulationen skiljer sig vanligen från målpopulationen, vilket leder till täckningsfel eller multiplicitet, dvs. förekomst av dubbletter i ramen. Ett exempel på det mindre vanliga fallet att rampopulationen är exakt lika med målpopulationen är när man vill utvärdera kvaliteten i ett register genom en urvalsundersökning; objekten i registret utgör då både mål- och rampopulation.

I figur 3.1.1 illustreras begreppen intresse-, mål- och rampopulation samt urval (som dras ur rampopulationen). Rektangeln med streckad kant är målpopulationen, och rektangeln med heldragna sidor är rampopulationen. Intressepopulationen representeras av en oregelbunden form därför att den är en ideal population som inte kan användas i den verkliga undersökningen. Övertäckning och undertäckning behandlas i avsnitt 3.2.

Figur 3.1.1 Populationer och urval



EXEMPEL 3.1.1 *Strukturundersökning för lantbruket*
I undersökningen av lantbrukets struktur, som numera utförs av Statens jordbruksverk, har man avgränsat en målpopulation från den ideala intressepopulationen genom att sätta några olika storleksgränser (”trösklar”). Intressepopulationen omfattar samtliga åkerfält och husdjur inom svenskt lantbruk. Det har beslutats att målpopulationen utgörs av de lantbruksföretag som har minst 2,1 hektar åkermark, stor djurbesättning eller stor trädgårdsodling (enligt vissa detaljerade kriterier).

Lantbruksföretagen är alltså objekt i undersökningen. De definieras som företag med lantbruksverksamhet driven under en och samma ledning. Notera att lantbruksföretag med t.ex. 1,5 hektar åkermark ingår i intressepopulationen men inte i målpopulationen, enligt de avgränsningar som här gjorts. I ärlighetens namn kan dock noteras att det ofta kan vara något svävande exakt vilken avgränsning av intressepopulationen som är den mest relevanta. Den frågan bör ses i ett användarperspektiv. Rampopulationen utgörs av de företag som ingår i den senaste versionen av Lantbruksregistret (LBR) och innefattar en del nedlagda lantbruk. Skattningarna som tas fram avser målpopulationen, dvs. i det här fallet alla bestående företag som ligger över storleksgränserna och alltså har chans att komma med i undersökningen samt nya företag. Som uppgiftskällor används dels brukaren direkt, dels administrativa register vars uppgifter härrör från brukaren. □

3.2 Vad menas med en urvalsram, och hur ska den vara?

Detta avsnitt går igenom olika begrepp som rör egenskaper hos ramar. Begreppen är väsentligen giltiga även för register ur vilka ramar framställs. Hur det går till att framställa ramar ur register beskrivs i avsnitt 3.4.

3.2.1 Urvalsramar

Som beskrivits i avsnitt 3.1.2 är *ramar* de förteckningar, register eller kartor man oftast utgår från i statistiska undersökningar. Ofta talas om *urvalsramar* i samband med urvalsundersökningar. Syftet med en urvalsram är att ge en väg till de objekt som ska observeras.

Ibland utgörs målpopulationen av enbart teoretiskt listbara objekt, varför ramen inte kan upprättas fysiskt. Ett exempel på detta är när man i undersökningen ”Inkommande besökare i Sverige” före passkontrollen vid bland annat flygplatser drar ett urval av utländska besökare som lämnar Sverige. Det finns även andra situationer då man vill dra urval även om man saknar en ram i vanlig bemärkelse. Se vidare avsnitt 3.2.3.

I en undersökning med flerstegsurval har man flera ”nivåer” av populationer. Därför upprättas en ram på varje nivå. För steg 1 upprättas en ram över hela populationen av steg 1-objekt. För steg 2 upprättas en ram som innehåller alla steg 2-objekt för utvalda steg 1-objekt. På motsvarande sätt upprättas ramar för objekten i eventuella efterföljande steg. Att man inte behöver upprätta ramar över alla objekt i de steg som följer på steg 1 är en praktisk fördel och ger kostnadsbesparingar. För exempelvis ett trestegsurval av skolor – klasser – elever skapas alltså en ram av skolor, sedan en ram av klasser för utvalda skolor och till sist en ram av elever för utvalda klasser. Man slipper upprätta ramar över alla klasser eller alla elever. Se vidare avsnitt 4.6 om klusterurval.

I en undersökning med nätverksurval utgörs ramen av ramobjekt som inte sammanfaller med, men har kopplingar till, observationsobjekten. I en hushållsundersökning kan ramen innehålla individer (ramobjekten), medan det som observeras är hushåll (observationsobjekten, kopplade till individerna).

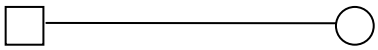
Vid planeringen av en undersökning behöver man ofta fundera på vad man ska välja som populations-, ram- och observationsobjekt. De tre olika typerna av objekt samt uppgiftskällan behöver inte sammanfalla. Se Statistiska centralbyrån (2001a). Anta att man vill undersöka en målpopulation av förskoleelever och använder ett tvåstegsförfarande. I första urvalssteget används en ram av förskolor. Ramobjekten i första steget är då förskolor. För varje utvald förskola listas alla dess elever, och ett urval av elever görs från denna lista. Ramobjekten i andra steget är då elever. Utvalda elever undersöks och utgör således populations- och observationsobjekt. Uppgifterna om eleverna hämtas dock från aktuellt kommunkontor, som därmed utgör uppgiftskälla.

3.2.2 Relationer mellan ramobjekt och populationsobjekt

Ramstrukturen kan logiskt anta fyra olika utseenden om man betraktar relationen mellan ramobjekt och populationsobjekt. Se figurerna nedan, fritt återgivna från Dalenius (1985, kapitel IV).

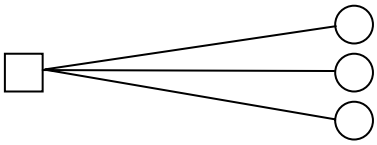
1. *Ett-till-ett-relation*
   Varje ramobjekt är kopplat till ett och endast ett populationsobjekt. Olika typer av direkturval genomförs. Exempelvis drar man lantbruksföretag ur Lantbruksregistret i undersökningar där lantbruksföretagen är populationsobjekt.

   Ramobjekt *Populationsobjekt*

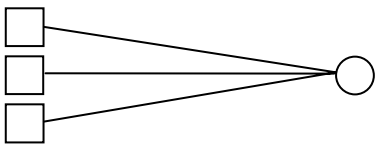
2. *Ett-till-många-relation*
   Ett ramobjekt är kopplat till flera populationsobjekt; inget populationsobjekt relaterar till mer än ett ramobjekt. Ramobjekten kallas vanligen för kluster, och olika typer av klusterurval genomförs. Se vidare avsnitt 4.6.

   Ramobjekt *Populationsobjekt*

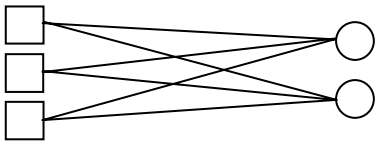
3. *Många-till-ett-relation*
   Varje ramobjekt är kopplat till ett och endast ett populationsobjekt; däremot kan ett populationsobjekt kopplas till flera ramobjekt. I detta fall kan s.k. nätverksurval tillämpas, se vidare avsnitt 4.5. Exempelvis dras individer ur befolkningsregistret, varefter hushåll associeras till dessa individer.

   Ramobjekt *Populationsobjekt*

4. *Många-till-många-relation*
   Ett ramobjekt kan kopplas till flera populationsobjekt, och vice versa. Avancerade typer av nätverksurval kan användas. Detta fall är dock mer sällsynt.

   Ramobjekt *Populationsobjekt*

Det är inte ovanligt att man i en och samma undersökning arbetar med mer än en av relationstyperna ovan. Exempelvis kan man vilja redovisa statistik för både hushåll och individer baserat på en och samma undersökning. Man kan göra ett direkturval av individer ur en ram som framställs ur registret över totalbefolkningen (RTB) och har då en ett-till-ett-relation mellan ram- och populationsobjekt (fall 1). Man gör vidare ett nätverksurval av de utvalda individernas hushåll (fall 3).

3.2.3 Ramar för urval i rum och tid

Vid framtagande av areell statistik, trafikstatistik, miljöstatistik och liknande blir det ibland aktuellt att dra urval från kartor eller flygbilder eller listor över geografiska områden: *areella urval*.

Ramar för *urval i rummet* kan avse

- punkter (direkt för observation på punkten eller indirekt som koordinat för t.ex. en provyta)
- linjer (t.ex. linjer som följs vid skogsinventering eller vägar representerade som linjer eller kurvor)
- ytor eller områden avseende geografiska regioner, åkerfält eller rutor
- volymer.

Vid urval i tiden förekommer ofta ingen listbar ram. Man knyter olika objekt och händelser till tidpunkter eller tidsintervall. Ibland drar man ett urval av händelser, exempelvis kundbesök.

Ramar för *urval i tiden* kan avse

- tidpunkter (att användas för att mäta något med ett urval med kontrollerad spridning över tiden, exempelvis mätning vid på förhand bestämda tidpunkter)
- tidsperioder (t.ex. mätveckor i undersökningen ”Inrikes och utrikes trafik med svenska lastbilar” – se vidare avsnitt 7.5)
- objekten i ett flöde (t.ex. att man väljer ut var k:te person som passerar en viss punkt).

EXEMPEL 3.2.1 *Nedskräpning*
Nedskräpningen på trottoarer o.d. i svenska städer anses ha ökat på senare år. Stiftelsen Håll Sverige Rent har därför gett SCB i uppdrag att försöka hitta en metod för att skatta förekomsten av skräp i stadsmiljön. Provundersökningar har genomförts i tre kommuner under sommaren 2007. I skrivande stund planeras kommunerna erbjudas ett system med urval av vägsträckor, representerade som linjeobjekt i geografiska informationssystem. Räkning av skräpföremål kan sedan göras på endera sidan eller båda sidorna om vägen eller gatan. Stratifiering kan behövas för homogenisering eller för att förbättra precisionen i skattningar för eventuella redovisningsgrupper. Inom respektive stratum verkar det lämpligt att dra systematiskt urval, för att få en areell spridning. □

3.2.4 Egenskaper hos en urvalsram

Urvalsramens uppgift är att tillhandahålla en koppling till de objekt man vill observera. En bra urvalsram speglar den aktuella målpopulationen, identifierar och ger en sökväg till objekten samt innehåller relevant hjälpinformation.

Täckningsproblem och multiplicitet
När rampopulationen inte helt speglar målpopulationen uppstår täckningsproblem. *Övertäckning* innebär att rampopulationen innehåller objekt som inte hör till målpopulationen. Det kan handla om nedlagda företag, eller om utflyttade eller avlidna personer. I en undersökning uppstår i allmänhet bortfall. För objekt för vilka data saknas råder ofta osäkerhet om huruvida objekten tillhör bortfallet eller utgör övertäckning. För ett resonemang om denna distinktion, se Japec et al. (1997, avsnitt 2.1).

Undertäckning innebär att det finns objekt i målpopulationen som inte kan nås från ramen. Ramen kan exempelvis sakna uppgift om nybildade företag eller inflyttade personer. Undertäckning kan vara ett stort problem i många undersökningar, i synnerhet som den är svår att åtgärda i efterhand. Ett begrepp besläktat med undertäckning är *mörkertal*, som förekommer inom den händelsebaserade statistiken. Mörkertal kan exempelvis utgöras av trafikolyckor som inte rapporterats till polisen.

Täckningsfel hanteras på olika sätt i estimationen. Övertäckning kan hanteras med hjälp av teorin för redovisningsgrupper (Särndal et al., 1992, kapitel 10); undertäckning eventuellt genom en kalibreringsansats (Lundström och Särndal, 2001, kapitel 11; Särndal och Lundström, 2005, kapitel 14). Begreppen över- och undertäckning illustreras i figur 3.1.1.

Om populationen är snabbt föränderlig blir ramen snabbt inaktuell. Det gäller då att hålla ett register med täta uppdateringar för att få en användbar ram. I många av de korttidsundersökningar på företagsområdet som SCB genomför tas en urvalsram fram och urval dras en gång per år. I snabbt föränderliga populationer innebär detta att urvalet vid periodens slut är gammalt, dvs. populationen har förändrats så mycket att urvalet inte längre representerar den population som det avser att representera. Täckningsfelen är helt enkelt stora. Ett sätt att komma runt detta är att dra urval oftare från en uppdaterad ram. Ett annat sätt är att i skattningen försöka korrigera för täckningsproblemen. Att dra urval oftare ger dock merarbete i produktionen. Med nya uppgiftslämnare blir bortfallet ofta större, och mer tid går åt till att kontakta, lära upp och hjälpa de nya.

Ett annat tänkbart ramproblem är *multiplicitet* (dubblettförekomst, duplettförekomst), som innebär att vissa populationsobjekt förekommer mer än en gång i ramen. Dessa populationsobjekt har alltså omotiverat råkat få mer än ett ramobjekt vardera kopplat till sig. Man brukar kalla detta för att man har *replikat* eller dubbletter (dupletter) i ramen. Ibland kan man eliminera replikat redan innan den slutliga urvalsramen fastställs. Annars kan man eventuellt justera för multiplicitet i skattningarna. Detta fordrar i allmänhet information om hur många gånger utvalda objekt förekommer i ramen. Se vidare Lessler och Kalsbeek (1992, avsnitt 5.2). Replikat som inte identifieras vare sig i ramen eller i urvalet riskerar att leda till systematiska fel.

Hjälpinformationens kvalitet
Hjälpinformation i ramen kan bland annat utgöras av

- kategorivariabler (t.ex. region och kön, eller bransch i företagsstatistik) som kan användas för att dela in populationen i strata eller redovisningsgrupper, för bestämning av avgränsningar av populationen eller i samband med cut-off-urval

- storleksvariabler som kan användas för att dela in populationen i strata eller redovisningsgrupper (efter storleksgrupper), för avgränsningar av populationen, i samband med cut-off-urval eller för urval proportionella mot storlek
- kategori- eller storleksvariabler som kan användas i estimationen.

I avsnitt 6.4 diskuteras mer utförligt hur hjälpinformation kan användas. En aspekt på kvaliteten i hjälpinformationen är dess aktualitet. Aktualiteten beror på hur ofta registret som ligger till grund för ramen uppdateras. Ett vanligt fall är att man använder fjolårets registeruppgifter som hjälpinformation. Man får avgöra från fall till fall om ramen innehåller alltför gamla eller felaktiga variabelvärden för att hjälpinformationen ska vara användbar.

Allmänt kan man bedöma ramkvaliteten utifrån s.k. evalveringsstudier. En översikt av metoder som kan användas för att bedöma ramkvalitet finns i Andersson et al (1999). Det är önskvärt att evalveringsstudier görs med jämna mellanrum för återkommande, större undersökningar.

EXEMPEL 3.2.2 *Varuflöden*
I Varuflödesundersökningen skattas summor av transporterade varors vikt och värde. Undersökningen avser en period av fyra kvartal. Ca 3000 arbetsställen dras inför varje kvartal. Ramen uppdateras alltså inför varje urvalsdragning, vilket ger aktuellare information för urvalet och säkrare skattningar både för hela perioden och för de flesta kvartal, jämfört med om man endast använt *en* ram per undersökningsomgång. □

EXEMPEL 3.2.3 *Djurräkningen*
I Djurräkningen avseende juni 2001 undersöktes förekomsten av nötkreatur, svin, får och höns på svenska lantbruksföretag. Målpopulationen utgjordes av företag med mer än 2 hektar åkermark under eget bruk eller med större djurbesättningar. Som ram användes en del av Lantbruksregistret (LBR) avseende 2000. Urvalsstorleken var 15 100 företag.

Ungefär 36 000 av de 77 000 företagen i ramen hade inga djur år 2000. Företagen hade alltså värdena noll på hjälpvariablerna avseende antal djur av olika slag. Denna grupp avsattes som ett eget stratum, kallat ”nollstratum”. Från stratumet drogs ett urval omfattande 400 företag. Urvalsfraktionen blev därmed mycket lägre än i övriga strata. Problem uppstod med att några företag i urvalet hade införskaffat ett stort antal djur mellan åren 2000 och 2001, och att dessa observationer multiplicerades med en så hög designvikt som 90,6. Detta riskerade att leda till vissa extrema bidrag till skattningarna av djurantalen i vissa redovisningsgrupper. Fem sådana företag betraktades därför som outliers (avvikare) och behandlades som övertäckning i sina ursprungliga strata. De placerades därefter i ett totalundersökt stratum för nya företag. Nackdelen med hanteringen är att skattningsproceduren, med reducerade designvikter, orsakar bias. Dessutom finns ett inslag av godtycke, om inte tydliga regler för ändring av designvikter upprättas.

Ett alternativ till ovanstående ansats hade varit att placera företagen utan djur, enligt ramen, under en cut-off-gräns och sedan skatta djurantalen år 2001 hos dessa med en modellbaserad skattning. Se avsnitt 6.2.4 om cut-off-urval. □

Situationer med särskilda ramproblem
I vissa urvalssituationer är ramframställningen särskilt problematisk. Några sådana situationer diskuteras i denna skrift:

- Undersökningar av ovanliga företeelser; se avsnitt 6.2.6.
- Urval ur rörliga populationer; se avsnitten 3.2.3 och 4.3.
- Urval ur kontinuerliga populationer; se avsnitten 3.2.3 och 4.8.

3.2.5 Läs mer
Lessler och Kalsbeek (1992, kapitel 3) diskuterar ramproblem och definierar vanliga termer. Ramproblem i företagsundersökningar behandlas i Cox et al. (1995, del A).

3.3 Vilka är de viktigaste registren på SCB?
I detta avsnitt beskrivs kortfattat några viktiga register på SCB. I avsnitt 3.4 behandlas sedan processen från generellt register till ram för en undersökning.

Ett *register* är en fullständig förteckning över objekten i en viss objektmängd eller population, medan en *ram*, enligt vad som tidigare sagts, är en förteckning av objekt som man utgår från i en undersökning.

SCB:s registersystem behandlas i Wallgren och Wallgren (2004, 2007). I det följande ges kortfattade beskrivningar av några av de viktigaste registren som används för att ta fram urvalsramar.

Fastighetstaxeringsregistret (FTR) används för ramframställning bland annat i undersökningar som riktar sig till fastighetsägare, i bostadsstatistik och i energistatistik. I företagsundersökningar eller undersökningar om kommuner och landsting etc. används oftast Företagsdatabasen (FDB) som utgångspunkt vid ramframställning. Men även andra register kan utnyttjas. Lantbruksregistret används exempelvis i undersökningar till lantbruksföretag. I individ- och hushållsundersökningar används Registret över totalbefolkningen (RTB) ofta som utgångspunkt.

3.3.1 Fastighetstaxeringsregistret
Skatteverket framställer varje år ett fastighetstaxeringsregister som baseras på särskilda, allmänna och förenklade fastighetstaxeringar. SCB får en fullständig kopia av Skatteverkets fastighetstaxeringsregister. Uppgifter om juridisk form, bransch (SNI-kod) och ägarkategori hämtas från Företagsdatabasen, se avsnitt 3.3.2. Från Geografidatabasen hämtas uppgift om fastighetskoordinater, koder för SAMS, tätort, postnummer och NYKO. Uppgifter om befolkning hämtas från RTB. (SAMS, Small Area Market Statistics, är en rikstäckande indelning av homogena bostadsområden som är framtagen av SCB i samarbete med kommunerna. NYKO, nyckelkodsystemet, är en finare geografisk indelning än län, kommun och församling. Det finns ca 90 000 delområden (NYKO) i Sverige.) Uppgifterna bearbetas av SCB och lagras i en databas, Fastighetstaxeringsregistret (FTR), som innehåller uppgifter om landets samtliga taxerings- och värderingsenheter.

Allmän fastighetstaxering (AFT) genomförs enligt ett rullande system med sex års intervall för varje fastighetstyp. Från och med 2003 infördes ett system med förenklade fastighetstaxeringar (FFT) för småhus, hyreshus och lantbruk vart sjätte år mitt emellan AFT.

FTR avser läget den 1 januari respektive år. Registret innehåller uppgifter om ett antal olika typer av objekt:

- taxeringsenhet
- värderingsenhet
- fastighet
- ägare.

Kopplingsvariabler finns som möjliggör koppling mot ägare som är företag eller fysiska personer och mot fastighet i RTB och Geografidatabasen.

I några av SCB:s undersökningar görs urval från FTR. Registret används också för att komplettera andra undersökningar med uppgifter om fastigheter och byggnader.

Definitioner och avgränsningar av objekten i FTR
Taxeringsenhet är vad som taxeras för sig och motsvarar som regel en fastighet. Om delar av en fastighet har olika ägare redovisas fastigheten uppdelad i skilda taxeringsenheter. Fastighet kan också delas upp på flera taxeringsenheter om det föreligger olika skattepliktsförhållanden eller användningssätt för skilda delar av egendomen. Vidare kan flera fastigheter inom en kommun slås samman till en taxeringsenhet om de har samma ägare och har samma skattepliktsförhållande och användningssätt. Taxeringsenhet är vad som normalt utgör objekt för överlåtelse och upplåtelse. För lantbruksenhet gäller att taxeringsenhet ska utgöras av egendom som ingår i en och samma brukningsenhet.

Värderingsenhet är den egendom inom en taxeringsenhet som ska värderas för sig. Varje tomt utgör som regel en värderingsenhet. Varje byggnad med ett värde om minst 50 000 kronor utgör en värderingsenhet. Vart och ett av ägoslagen åkermark, betesmark, skogsmark och skogsimpediment utgör också en värderingsenhet. För hyreshusenheter gäller att bostadsdelen och lokaldelen i ett hyreshus bildar separata värderingsenheter.

Alla fastigheter är skattepliktiga med undantag av s.k. specialbyggnader. Principen för undantagen från skatteplikt kan mycket grovt sägas vara att fastigheternas användning i någon mening anses vara av särskild samhällsnytta, vilket exempelvis gäller för skolor och vårdbyggnader. För skattefri fastighet ska inte fastställas något taxeringsvärde. Det finns heller inte uppgifter om byggnaderna för de icke-skattepliktiga enheterna.

Variabelinnehåll i FTR
Taxeringsenhet
- Belägenhet (län, kommun, församling).
- Typkod (lantbruksenhet, småhusenhet, hyreshusenhet, industrienhet, täktenhet, elproduktionsenhet samt specialenhet).
- Samband med fastighet (delning, sammanföring).
- Ägare (deklarationsansvarig och samtliga övriga ägare).
- Areal.
- Totalt taxeringsvärde.

Värderingsenhet (beroende på typ av värderingsenhet)

- Byggnader
  - Typ av byggnad.
  - Nybyggnadsår och värdeår.
  - Area.
  - Uppgifter om standard/beskaffenhet.
  - Taxeringsvärde.
- Tomter
  - Tomtareal.
  - Strandbelägenhet (småhus).
  - Vatten- och avloppsförhållanden (småhus).
  - Area byggrätt (hyreshus).
  - Taxeringsvärde.
- Ägoslag
  - Typ av ägoslag (skog, skogsimpediment, åker, bete).
  - Areal.
  - Beskaffenhet (skog, åker, bete).
  - Virkesförråd (skog).
  - Taxeringsvärde.
- Övriga typer av värderingsenheter
  - Olika variabler beroende på använda värdefaktorer.
  - Taxeringsvärde.

Ägare

- ID-nummer (person- eller organisationsnummer).
- Juridisk form.
- Ägarkategori (inklusive kodning av allmännyttiga bostadsföretag med hjälp av SCB-register).
- Ägarandel.
- Folkbokföringskommun (för fysisk person, deklarationsansvarig ägare).

Fastighet

- Belägenhet (län, kommun, församling).
- Fastighetsbeteckning.
- Areal.
- Samband med taxeringsenhet (delning, sammanföring).

Registret kan kopplas till fastighetsbeteckning och befolkningen på fastigheten och ge information om fastighetens huvudsakliga typ av taxeringsenhet och ägarkategori. Ägarstrukturen för bostäder, lokaler, skog, åker m.m. kan analyseras. Sambandet mellan taxeringsenhetens belägenhetskommun och ägarens boendekommun kan studeras.

För mer detaljerad information, se www.scb.se/BO0601. Där finns bland annat en länk till en fullständig förteckning över typkoder.

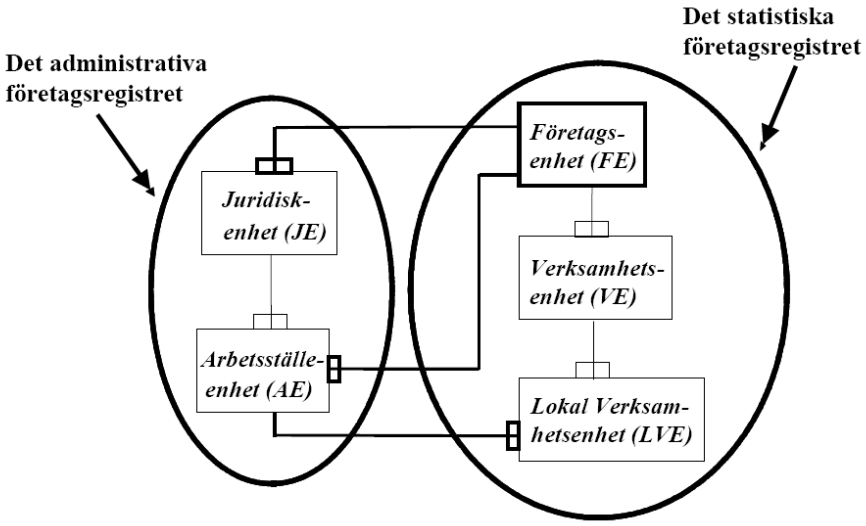
3.3.2 Företagsdatabasen

Företagsdatabasen, FDB, innehåller alla företag i Sverige som bedriver någon form av ekonomisk verksamhet samt statliga myndigheter, kommuner och landsting. I FDB finns kontaktuppgifter samt ett antal variabler som kan användas för att ta fram urvalsramar och vid urvalsdragning, bland annat näringsgren, antal anställda, sektor, juridisk form, ägarkategori, omsättning och företagets status (verksam, ej verksam, etc.). Variabeln näringsgren kallas i vardagslag även för bransch.

FDB uppdateras kontinuerligt. Dels ändras uppgifter för enskilda företag, dels används administrativa material, t.ex. från Skatteverket, för att uppdatera större delar av registret. Enkäter skickas också varje år till företag, främst till de större företagen.

I FDB finns ett antal olika objekttyper som hänger ihop – se figur 3.3.1. Vilken typ av objekt som används i en undersökning beror på vilka uppgifter som ska samlas in, vilka redovisningsgrupperna är och till viss del också på om det är en årsundersökning eller en korttidsundersökning. De objekttyper som används i undersökningar är juridisk enhet, JE, företagsenhet, FE, verksamhetsenhet, VE, lokal verksamhetsenhet, LVE, och arbetsställeenhet, AE. De olika enheterna är kopplade till varandra, t.ex. består en FE av en eller flera JE, men en JE ingår endast i en FE. Objekttyperna JE och AE utgör tillsammans det *administrativa* företagsregistret; FE, VE och LVE det *statistiska* företagsregistret.

Figur 3.3.1 Relationer mellan objekttyper i företagsundersökningar. Källa: Statistiska centralbyrån (2002)



JE används ibland i korttidsstatistik. FE är en statistisk enhet som konstrueras på SCB. Varken JE eller FE behöver vara ”branschrena”; de kan alltså ha verksamhet i flera olika branscher. En VE är konstruerad så att den ska vara branschren och bör användas då branschren statistik ska tas fram. En LVE används för både geografisk indelning och branschstatistik. En AE är en geografisk enhet och kan användas för geografisk indelning.

En beskrivning av innehåll, variabler, enheter och uppdateringar i FDB finns på SCB:s webbplats:
http://www.scb.se/templates/Standard____180452.asp.

3.3.3 Registret över totalbefolkningen

Registret över totalbefolkningen, RTB, innehåller uppgifter om den folkbokförda befolkningen i Sverige. De variabler som finns i RTB är

- personnummer, namn och adress
- folkbokföringsförhållanden (län, kommun, församling och fastighet)
- civilstånd
- relationer (maka/make, registrerad partner, biologisk far, biologisk mor, adoptivföräldrar, vårdnadshavare)
- samhörighet (gift/sammanboende, ensamstående/annan boende, barn/fosterbarn till någon i hushållet)
- medborgarskap och födelseland
- sammanräknad inkomst
- föräldrarnas födelseland
- grund för bosättning (utomnordiska medborgare vilka invandrat sedan 1980-talets senare hälft).

En uppgift som däremot saknas i RTB är telefonnummer, vilket gör att urval från RTB behöver kompletteras med denna information om telefonintervjuer ska göras.

RTB uppdateras kontinuerligt. Månads-, kvartals- och årsversioner av RTB sparas och kan användas för t.ex. framställning av urvalsramar.

Läs mer om RTB på SCB:s webbplats:
http://www.scb.se/templates/Standard____22840.asp.

3.4 När och hur ska en urvalsram tas fram?

Här fokuseras på en undersöknings *ramförfarande*, dvs. proceduren för att fastställa observationsobjekt och komma i kontakt med uppgiftskällor, se Statistiska centralbyrån (2001a). En ram kan bestå av alla eller en delmängd av objekten i ett eller flera register. I många av SCB:s undersökningar kan en urvalsram tas fram från något av de register som finns tillgängliga. I andra fall finns inte något färdigt register att använda.

För att ta fram en urvalsram krävs information om en rad olika saker: vad undersökningen avser, vilka objekt som ingår i populationen, vilken typ av objekt som ska observeras, referenstid och vilken hjälpinformation som behövs för avgränsning, urvalsdesign och skattning. Detta avsnitt koncentreras på undersökningar där det är möjligt att utgå från ett eller flera register för att framställa ramen.

3.4.1 Val av objekt

I vissa undersökningar måste ett val av objekt göras, vilket kan fordra en del överväganden:

- Vad ska redovisas?
- Vilka objekt finns tillgängliga i register och för vilken typ av objekt går önskade uppgifter att samla in?
- Efterfrågas redovisningar för särskilda geografiska områden? Då bör objekten väljas så att detta blir möjligt.

- Hos vilken typ av företagsenhet finns uppgifterna tillgängliga? Detta kan i sin tur bero på om undersökningen är en korttids- eller årsundersökning. (Med en korttidsundersökning avses inom företagsstatistiken vanligen en månads- eller kvartalsundersökning.)
- Vilka objekt finns tillgängliga i register? Exempelvis finns i Sverige i skrivande stund inte tillgång till ett register över hushåll. När en undersökning behöver samla in uppgifter om hushåll görs därför ofta detta via ett urval av individer. Ramobjekten är alltså individer i stället för hushåll.

Självklart finns det också undersökningar där valet av objekt är oproblematiskt, t.ex. individundersökningar där individer utgör objekt och tillräcklig information om individerna finns i RTB.

3.4.2 Val av tidpunkt

Valet av tidpunkt för framställning av urvalsramar från register beror på en rad faktorer, bland annat

- referenstiden för undersökningen
- när registret uppdateras avseende objekt och annan information
- samordning med andra undersökningar.

Det är viktigt att samordna urvalsramar då resultaten från olika undersökningar ska jämföras eller då undersökningar hänger ihop men de olika undersökningarna riktar sig till olika delmängder av samma population. Om olika versioner av ett register används kan objekten ha bytt grupptillhörighet så att en del objekt inte finns med i någon av undersökningarna och en del finns med i flera.

I många företagsundersökningar används SAMU-systemet, SAMordnad Urvalsdragning, för urvalsdragning. I SAMU används fyra olika versioner av FDB varje år. Nya versioner av FDB hämtas in i mars, maj, augusti och november. I regel brukar årsundersökningar utgå från den version som tas fram i slutet av november det år som undersökningen avser, medan åtminstone en del korttidsundersökningar utgår från den version som tas fram i mars (återigen det år som undersökningen avser).

Alla undersökningar som utnyttjar SAMU utgår från någon av de fyra versionerna av FDB för ramframställning. Om två undersökningar drar urval i samma version av FDB så blir ramarna samordnade. Läs mer om SAMU i avsnitt 5.5.

3.4.3 Avgränsning av en ram

Avgränsningar av en ram som används för en företagsundersökning kan handla om att endast inkludera en del av alla branscher eller endast företag av en viss storlek, i termer av antal anställda exempelvis. I en individundersökning kan ramen avgränsas så att endast individer i ett visst åldersintervall tillåts ingå.

Avgränsning kan också göras genom att man med hjälp av en cut-off-gräns medvetet undantar delar av målpopulationen. Läs mer om användning av cut-off-gränser i avsnitt 6.2.4.

3.4.4 Sammanfogning av register och administrativa data

Vid framställning av en urvalsram används ofta ett register som utgångspunkt. För att få mer aktuell information om vissa variabler eller för att komplettera med variabler som inte ingår i registret matchas registerobjekten med administrativa data från andra källor. I exempel 3.4.2 nedan kompletteras företagsregistret med information från en rad olika källor för att framställa en ram. Den tillkommande informationen kan exempelvis användas för avgränsning av ramen, stratifiering eller i skattningarna.

Det krävs lagstöd för att behandla personuppgifter. Att sammanfoga register är ett exempel på ”behandling” av uppgifter. Det kan finnas anledning att kontakta en jurist innan sammanfogning görs.

3.4.5 Multipla ramar

När två eller flera urvalsramar används till samma undersökning, kallas detta för användning av *multipla ramar*. Idén är att använda ett antal ramar som var för sig inte täcker målpopulationen men som tillsammans gör det. I praktiken förekommer nästan alltid viss undertäckning. Användning av multipla ramar görs i förhoppningen att dessa tillsammans ger mindre undertäckning än ramarna var för sig. Om ramarna tillsammans ger full täckning av målpopulationen så ingår varje objekt i målpopulationen i åtminstone en av urvalsramarna. I en urvalsundersökning dras ett urval från varje ram och i det resulterande urvalet kan objekt som ingår i flera ramar komma med flera gånger. Detta justeras för i estimationen. På SCB är användningen av multipla ramar inte vanligt förekommande, men exempel på användning finns, se exempel 3.4.4 nedan. Se avsnitt 3.4.7 för några referenser som ger mer information om multipla ramar.

3.4.6 Exempel

Nu följer exempel på hur ramframställningen görs i några SCB-undersökningar:

EXEMPEL 3.4.1 *Forskning och utveckling i företagssektorn*
Undersökningen ”Forskning och utveckling i företagssektorn” (”FoU i företag”) genomförs vartannat år. Referensperioden är kalenderår, och denna beskrivning avser undersökningen med referensperiod 2005. Bland annat skattas totala FoU-utgifter för hela populationen, samt för redovisningsgrupper definierade av bransch och storleksklasser, där storleksklasserna bestäms av det antal anställda företagen har.

Målpopulationen utgörs av alla företag av en viss storlek inom vissa branscher och sektorer. Storleken mäts i antal anställda. Företagens sektorindelning beskrivs i Statistiska centralbyrån (2001b). De sektorer som ingår i ”FoU i företag” är icke-finansiella företag, delar av den finansiella företagssektorn och hushåll med företagarinkomst från rörelse med anställda. Övriga sektorers FoU-utgifter samlas in i andra undersökningar; t.ex. samlas FoU-utgifter för kommunal sektor in i undersökningen ”FoU i kommuner och landsting”. I vissa branscher ingår inte vissa sektorer som ingår i andra branscher. Företag med färre än tio anställda ingår inte i undersökningens målpopulation.

Objekt
FE (företagsenheter) används som objekt. I en del fall observeras alla ingående JE (juridiska enheter) i en FE. Redovisning görs på FE-bransch och på storleksgrupp efter antal anställda på FE.

Ramframställning
Som utgångspunkt används alla verksamma FE i FDB. Avgränsningar görs med avseende på bransch kombinerat med sektor och storlek, dvs. de avgränsningar som definierar målpopulationen används. Bransch och antal anställda används även för stratifiering.

Referensperiod och tidpunkt för ramframställning
Referensperioden är ett år, och ramen tas fram i slutet av referensperioden. SAMU-systemet används för ramframställning och urvalsdragning. Den version av FDB som används i novemberomgången av SAMU används som utgångspunkt vid ramframställningen.

Samordning med andra undersökningar
Positiv samordning (se avsnitt 5.1) görs med undersökningen ”Företagens ekonomi”, FEK. Denna undersökning är också årlig. □

EXEMPEL 3.4.2 *Utrikeshandel med tjänster*
I undersökningen ”Utrikeshandel med tjänster” samlas uppgifter in om företagens tjänsteimport, tjänsteexport och transfereringar till och från utlandet. Målet är att skatta total tjänsteimport, tjänsteexport och transfereringar för olika tjänsteslag och olika typer av transfereringar till och från utlandet. I dagsläget skattas endast totaler för hela populationen, inte för delmängder av densamma.

I målpopulationen ingår alla företag, men det är långt ifrån alla som har import eller export av tjänster eller transfereringar med utlandet. I princip skulle en urvalsram kunna konstrueras som hela FDB vid en given tidpunkt, men då skulle urvalsramen till största delen bestå av företag som är ointressanta för undersökningen, nämligen företag som inte har denna typ av utlandsverksamhet.

Objekt
Den typ av objekt som används är juridisk enhet, JE. Redovisning på bransch har inte varit aktuell, dvs. branschrena enheter har inte behövts, och undersökningen är en korttidsundersökning.

Ramframställning
För att ta fram en urvalsram görs diverse avgränsningar med målet att de företag (här JE) som ingår i ramen ska vara de företag som har tjänsteimport, tjänsteexport eller transfereringar. Som utgångspunkt används alla företag i FDB vid en given tidpunkt. För avgränsning används ett antal administrativa material och register; bland annat företagens skattedeklarationer, där utrikeshandel med tjänster specificeras, och register över företag som bedriver utrikeshandel med varor. Om ett företag ingår i någon av de externa källor som används får det ingå i urvalsramen. Företag som endast finns med i registret från utrikes varuhandel och som har liten omsättning exkluderas ur ramen. Kommuner och landsting ingår endast via Sveriges Kommuner och Landsting.

Variabler baserade på omsättning med olika referensperioder används för avgränsningar av urvalsramen, men även i urvalsdesignen för stratifiering och allokering (se vidare exempel 4.2.2 och 6.1.6). Omsättningsvariablerna i

urvalsramen hämtas från olika källor. Bransch och sektor används också i stratifieringen och dessa variabler hämtas från FDB.

Den del av företagspopulationen som ligger utanför ramen men ingår i målpopulationen antas ha värde noll för de variabler som samlas in. Avgränsningarna i urvalsramen definieras alltså som en cut-off-gräns (se vidare avsnitt 6.2.4). Då undersökningen startade fanns tillgång till data om betalningar till och från utlandet. Det ramförfarande som används utvecklades bland annat genom att man såg till att de källor som användes täckte in de företag som hade utlandsbetalningar.

Referensperiod och tidpunkt för ramframställning
Referensperioden är kvartal. Ramen tas fram i mars, och ett nytt urval dras i samband med att ramen tas fram. *Ett* urval används för alla kvartal det år som urvalet dras.

Samordning med andra undersökningar
SAMU-systemet används för urvalsdragning. Samordning av urvalsramar och urval görs med andra undersökningar, men det ställs inga krav på det. □

EXEMPEL 3.4.3 *Skörd av spannmål, trindsäd och oljeväxter*
I undersökningen ”Skörd av spannmål, trindsäd och oljeväxter” samlar man in uppgifter om lantbruksföretagens totala bärgade kvantiteter av varje gröda och kvantiteternas vattenhalt.

I målpopulationen ingår alla lantbruksföretag med odling av spannmål, trindsäd eller oljeväxter, vilka brukar mer än 2 hektar åkermark och har minst 0,3 hektar av undersökningsgrödorna.

Objekt
Objekten är lantbruksföretag. Med lantbruksföretag avses en inom jordbruk, husdjursskötsel eller trädgårdsodling bedriven verksamhet under en och samma driftsledning.

Ramframställning
För att ta fram en urvalsram används Statens jordbruksverks arealregister, som baseras på stödansökningar. Urvalsramen är en delmängd av registret och utgörs av jordbruksföretag med mer än 5 hektar åkermark och minst 0,3 hektar av undersökningsgrödorna. Det aktuella årets uppgifter om lantbruksföretag och grödarealer används som underlag. Referenstiden är kalenderår.

Samordning med andra undersökningar
Positiv samordning görs med Gödselmedelsundersökningen; se exempel 5.5.1. Ramarna för de två undersökningarna är dock inte identiska, eftersom man även behöver efterfråga uppgifter för djurgårdar i Gödselmedelsundersökningen. □

EXEMPEL 3.4.4 *Identifiering av anläggningar med häst och uppskattning av antal hästar i Sverige*
Undersökningens syfte var att skatta antalet hästar och anläggningar med häst fördelat på län och tätort respektive glesbygd.

Objekten i populationen var anläggningar med häst. När ramen konstruerades var målet att identifiera anläggningar med häst den 14 oktober 2004. För att göra detta definierades fyra delpopulationer, och urvalsramar togs

fram för varje delpopulation. Målet var att skapa ramar med god täckning och liten överlappning.

Delpopulation 1 bestod av anläggningar med häst med vissa företagstyper, nämligen ridskolor samt professionella trav- och galopptränare. Urvalsramen togs för ridskolor fram med hjälp av offentliga telefonregister och via kontroll mot uppgifter om ridklubbsanknutna ridskolor från Ridsportförbundets webbplats. För professionella tränare användes uppgifter från Svenska Travsportens Centralförbund och Svensk Galopp. Företag som fanns med i delpopulation 4 (se nedan) uteslöts ur delpopulation 1.

Delpopulation 2 bestod av lantbruksföretag. Urvalsramen skapades utifrån Lantbruksregistret (LBR). Lantbruksföretag som fanns med i delpopulation 1 och 4 uteslöts ur delpopulation 2. Företag med häst enligt LBR 2003 ingick med uppgiften från 2003. Bland övriga företag gjordes ett urval.

Delpopulation 3 bestod av fastigheter ur FTR. Endast delar av registret användes som urvalsram. Avgränsningar gjordes utifrån storlek, typ av fastighet, om fastigheten var bebyggd, m.m. Uppgifter från LBR användes för att, så långt det gick, ta bort jordbruksföretag som ingick i delpopulation 2. Objekt från delpopulation 1 uteslöts ur delpopulation 3, genom uppmaning att inte svara på blanketten. Fastigheter belägna i kommuner som ingick i delpopulation 4 uteslöts också.

Delpopulation 4 bestod av anläggningar med häst i kommuner där kommunförvaltningen hade god kännedom om hästbeståndet. Totalt ingick 17 kommuner i denna del.

Från respektive ram drogs ett urval eller gjordes en totalundersökning. □

3.4.7 Läs mer

Mecatti (2007) ger många litteraturhänvisningar till artiklar som behandlar tekniken att utföra en undersökning med urval som dras från flera ramar. Cox et al. (1995) och Lessler och Kalsbeek (1992) innehåller översikter av området.

4 Olika sätt att dra urval

I avsnitt 4.1–4.4 i detta kapitel beskrivs några av de viktigaste urvalsmetoderna som används på SCB. Här förutsätts för enkelhetens skull att ram- och målpopulation sammanfaller, och kan benämnas ”population”.

I avsnitt 4.5–4.8 behandlas sätt att hantera situationer då man antingen helt saknar en ram över de objekt man vill undersöka, eller (i fallet tvåfasurval) t.ex. har behov av att komplettera befintlig ram med hjälpinformation för att kunna göra ett effektivt urval.

Kapitel 4 avslutas med ett avsnitt om icke-sannolikhetsurval.

Olika urvalsmetoder jämförs i avsnitt 6.2.5. Avsnitt 7.5 beskriver ett sätt att dra urval som inte sammanfaller med någon av metoderna i kapitel 4, men som kan approximeras med en av dem. Det finns naturligtvis fler urvalsmetoder som inte tas upp här. Se litteraturtipsen i kapitel 10.

4.1 Obundet slumpmässigt urval

I detta avsnitt ges en introduktion till *obundet slumpmässigt urval* (OSU) – dess egenskaper och många användningsområden. OSU används oftast i kombination med något annat, såsom stratifiering. I detta avsnitt betraktas endast dragning utan återläggning, dvs. fallet att ett objekt kan komma med i urvalet högst en gång.

4.1.1 Beskrivning av metoden

Ett OSU definieras enligt följande:

Varje delmängd, av fix storlek n av distinkta objekt, av en ändlig population har samma sannolikhet att utgöra urvalet.

Definitionen innebär bland annat att första ordningens inklusionssannolikheter är lika för alla objekt. Inklusionssannolikheterna är helt enkelt n/N, där n är önskad urvalsstorlek och N är antalet objekt i populationen. Att första ordningens inklusionssannolikheter är lika för alla objekt i populationen innebär att OSU är ett *självvägande* urval. Det finns flera urvalsmetoder som kan ge upphov till självvägande urval. Se vidare avsnitt 4.6.4.

4.1.2 Användningsområden

OSU är en möjlig urvalsmetod att använda om variabelvärdena inte varierar särskilt mycket, dvs. om populationen är homogen med avseende på undersökningsvariabelns värden för de ingående objekten. Detta gäller i första hand då skattningar av parametrar önskas för hela populationen.

Även då skattningar för redovisningsgrupper ska beräknas kan OSU fungera, förutsatt att redovisningsgrupperna är tillräckligt stora, att urvalsstorlekarna i redovisningsgrupperna är tillräckligt stora och att populationen är relativt homogen.

Om hjälpinformation saknas kan OSU vara den enda möjliga eller rimliga urvalsmetoden att använda. I undersökningar där något av SCB:s register används för att framställa en urvalsram finns oftast bra hjälpinformation

som kan användas i urvalsdesignen. I undersökningar där ramen framställs på annat sätt kan sådan information saknas eller vara av så dålig kvalitet att det enda rimliga är att använda OSU. Även då något av SCB:s register används för att konstruera en urvalsram kan ”bra” hjälpinformation saknas för en specifik undersökning.

I praktiken är det vanligast att OSU tillämpas i kombination med något annat, dvs. att OSU används i någon del av en mer komplicerad urvalsdesign. OSU kan exempelvis användas i delmängder av populationen, se avsnitt 4.2 om stratifierat OSU, eller som urvalsmetod i något av stegen i ett flerstegsurval, se avsnitt 4.6. OSU kan också användas i första eller andra fasen i ett tvåfasurval, se avsnitt 4.7.

I jämförelser mellan olika urvalsmetoder används ofta OSU som referensmetod, se avsnitt 6.1.3 om *designeffekt*.

4.1.3 Fördelar

Enkelhet
OSU kan ur flera olika aspekter betraktas som den enklaste urvalsmetoden: teorin är väl utvecklad vad gäller bestämning av urvalsstorlek, skattning m.m. Det är också enkelt att utföra en urvalsdragning.

Behöver inte hjälpinformation
Ingen hjälpinformation behövs för att dra urvalet – det räcker med en förteckning över de objekt som ingår i populationen. För detaljer om hur man kan dra ett OSU, se avsnitt 4.1.5.

Underlättar analys
Vid analys av surveydata, t.ex. regressionsanalys eller multivariat analys, uppstår frågan om urvalsvikter ska ingå i beräkningarna eller inte (se avsnitt 4.9.4). Om urvalet dragits med OSU behövs inga sådana vikter. På så sätt är OSU en urvalsdesign som är särskilt enkel vid analys. Det är inte helt ovanligt att användare vill ta fram egna skattningar och göra analyser; de föredrar då ofta att objekten har samma vikter, vilket de har vid bland annat OSU.

4.1.4 Nackdelar

Svårt att kontrollera precisionen i redovisningsgrupper
I de flesta undersökningar skattas parametrar både för hela populationen och för redovisningsgrupper. Om OSU används finns inga garantier för att alla redovisningsgrupper är tillräckligt representerade i urvalet. Detta kan leda till att skattningar för vissa redovisningsgrupper inte kan göras alls eller inte kan göras med tillräckligt bra precision.

Ineffektivitet
Hjälpinformation används inte i designen av ett OSU. Detta kan leda till att urvalsstorleken blir onödigt stor för en given önskad precision, eller att variansen blir onödigt stor för en given urvalsstorlek. Ibland kan man kompensera för denna bristande effektivitet genom att använda hjälpinformation i estimatorn i stället. Om man har god hjälpinformation drar man i allmänhet bäst nytta av den genom att använda den i såväl urval som skattningar.

Bäddar ibland för otursamma urval
Med OSU finns en risk att få dåliga urval i den meningen att en skattning baserad på urvalet skiljer sig starkt från parametervärdet. Om t.ex. FDB används som urvalsram är sannolikheten stor att endast små företag ingår i urvalet, eftersom stora delar av ramen består av små företag. Med ett sådant urval går det inte att ta fram bra skattningar av t.ex. totaler för variabler som beror på storleken på företagen. Variansskattningarna blir också missvisande.

Dyrare intervjuer
Om besöksintervjuer ska göras kan användningen av OSU medföra stora kostnader, då de utvalda objekten kan vara utspridda över stora geografiska områden.

4.1.5 Dragningsscheman

En enkel beskrivning av att dra ett OSU är att dra objekten ”som lotter ur en tombola”. Med ramobjekten som ”lotter” i en väl omskakad tunna kunde man på måfå plocka ut så många av dem som ska ingå i urvalet. Detta enkla förfarande skulle ordentligt utfört vara fullt korrekt, men något arbetsamt. I praktiken dras OSU med datorbaserade metoder som har motsvarande mål som tombola-idén men som går något andra vägar.

De datorbaserade metoderna använder någon funktion i datorn för generering av *slumptal.* Slumptalsfunktionen är i grunden deterministisk och alltså inte ”äkta” slumpmässig i en striktare mening, och slumptalen kallas därför ibland för *pseudoslumptal*. Slumptalsfunktionen är dock så konstruerad att en mängd av tal som den genererat kan användas som om den vore genererad av slumpen.

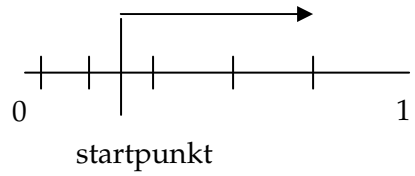
I litteraturen beskrivs ett antal olika *dragningsscheman* (urvalsscheman) för OSU. Ett dragningsschema för OSU är helt enkelt en algoritm för att dra ett OSU. Här beskrivs endast två olika scheman.

Urvalsdragning via sortering:

1. Till varje objekt i populationen genereras ett slumptal. Slumptalen generas oberoende av varandra från en likformig fördelning på intervallet [0,1].
2. Sortera med avseende på slumptalen.
3. Objekten med de n minsta slumptalen utgör ett OSU av storlek n.

I stället för att välja ut de objekt som har de n minsta slumptalen kan man välja de som har de n största. Dessa objekt utgör också ett OSU. Mer generellt, genom att välja en startpunkt på intervallet [0,1] och sedan ta de n objekt med slumptal som ligger till höger eller vänster om startpunkten fås ett OSU, se figur 4.1.1. Detta dragningsschema kan även användas för att åstadkomma urvalsrotation, vilket beskrivs i avsnitt 5.4.3.

Figur 4.1.1 Ett urval dras från en linje där fem ramobjekt är inlagda med slumptal. Objektens position markeras med lodräta sträck. Urvalet består av *n* = 3 objekt som ligger till höger om startpunkten



Ett annat schema för att dra ett OSU är följande s.k. listsekventiella schema, se t.ex. Särndal et al. (1992, s. 66). Anta att objekten i populationen finns i en lista med etiketter 1, 2, …, *N*.

1. Generera oberoende slumptal, $\varepsilon_1, \varepsilon_2, \ldots$ från en likformig fördelning på intervallet [0,1].
2. För objekt 1:

   om $\varepsilon_1 < \frac{n}{N}$ väljs objekt 1.

3. För objekt k = 2, 3, 4, …, N:

   om $\varepsilon_k < \frac{n - n_k}{N - k + 1}$ väljs objekt k,

   där n_k är antalet objekt som valts ut bland de k–1 första objekten i listan.

Båda metoderna generar som sagt OSU. Den förstnämnda kräver sortering, vilket skulle kunna vara ett problem om populationen är mycket stor. I den andra metoden behöver man oftast inte gå igenom hela populationen, eftersom man slutar då man har dragit n objekt.

4.1.6 Exempel

EXEMPEL 4.1.1 *Hushållens utgifter*
Som urvalsram i undersökningen ”Hushållens utgifter” används alla individer i åldrarna 0–79 år i RTB vid urvalstidpunkten. Urvalet dras som ett OSU av 4000 individer.

Urvalsobjekten är individer men de viktigaste parametrarna avser hushåll. Ett hushållsurval hade varit mer lämpligt, men eftersom ett hushållsregister inte finns tillgängligt används individer som urvalsobjekt. Se även avsnitt 4.5 om nätverksurval.

Vissa parametrar avser individer, t.ex. genomsnittliga utgifter per individ och kostnadsslag. De redovisningsgrupper som förekommer täcks in av urvalet, trots att ingen hänsyn tas till dessa i urvalsdragningen. Resultaten av undersökningen används i analyser, vilket är ett skäl till att OSU har valts som urvalsmetod. □

EXEMPEL 4.1.2 *Hundar och katter*
I ett utredningsuppdrag om märkning och registrering av hundar och katter genomfördes bland annat en postenkät. Till denna drogs ett OSU av 10 100 individer i åldrarna 16–89 år ur RTB. Anledningarna till att använda

OSU var att ingen hjälpinformation finns i RTB om förekomst av sällskapsdjur och att ingen regional redovisning krävdes. □

4.2 Stratifierat urval

I de flesta stora, nationella undersökningar är populationen naturligen indelad i delmängder. De kan utgöras av branscher i företagsundersökningar eller kön i individundersökningar. Om avsikten är att samtliga delmängder ska undersökas med ett urval eller med totalundersökning, säger man att rampopulationen är stratifierad.

4.2.1 Beskrivning av metoden

Anta att populationen delas in i icke-överlappande (disjunkta) delmängder som tillsammans täcker hela populationen. Urval görs sedan oberoende av varandra inom delmängderna. Detta förfarande kallas stratifierat urval. En delmängd som man drar urval ur kallas *stratum*; bestämd form i singular är *stratumet;* pluralformen är *strata* eller stratum; bestämd form i plural är strata eller *stratumen* (källa: Svenska Akademiens ordlista).

Strata kan undersökas med olika urvalsmetoder. Det är inte alltid nödvändigt att undersöka samtliga delmängder i en population. Om endast ett urval av delmängderna avses att undersökas, säger man att man drar ett *klusterurval* (se avsnitt 4.6).

Stratifieringens hantverk beskrivs i avsnitt 4.2.5–4.2.9.

4.2.2 Användningsområden

Stratifierat urval är mycket vanligt. Urvalsmetoden används ofta för att så långt det är praktiskt möjligt avgränsa redovisningsgrupper, där ett eller flera strata motsvarar en redovisningsgrupp. Stratifiering används ofta då populationen är heterogen. Indelningen görs då så att objekten inom strata är mer lika varandra än vad objekt i olika strata är, med avseende exempelvis på undersökningsvariabler, svarsbenägenhet eller liknande. Det finns ofta även praktiska och administrativa skäl att dela upp en stor undersökning i mindre delar, och ett enkelt sätt att hantera detta är via stratifiering. Ibland vill man använda olika frågeformulär eller olika datainsamlingsmetoder för olika grupper av populationen. Det kan då vara praktiskt att stratifiera med avseende på detta.

Omvänt kan man fråga sig om det finns skäl att *inte* stratifiera. Om hjälpinformationen är otillräcklig finns det strängt taget bara administrativa skäl att stratifiera. För undersökningar som har stora eller inga redovisningsgrupper, stort urval och en population som är homogen i alla viktiga avseenden är behovet av stratifiering svagt. Det kan i så fall räcka med att dra ett OSU eller ett systematiskt urval.

Även om populationen är starkt heterogen kan man tänka sig att dra ett OSU och skatta med en poststratifierad estimator. För beskrivning av poststratifiering, se någon av de läroböcker som omtalas i kapitel 10. Det är ett alternativt förfarande som i praktiken får nästan samma (eller samma) variansreducerande effekt som stratifierat OSU. En fördel med att helt förlita sig på poststratifiering är den flexibilitet som det medför. Man kan poststratifiera på olika sätt för olika skattningar. Nackdelen är att man inte har garanterat tillräckliga urvalsstorlekar i viktiga redovisningsgrupper. I

värsta fall kan bristen på kontroll leda till att man inte får något utvalt objekt alls i någon viktig redovisningsgrupp. Skattningar för variabler med sned fördelning kräver att de största objekten kommer med i urvalet, vilket inte nödvändigtvis sker med OSU. Om populationen är sned ska man om möjligt ta hänsyn till snedheten i urvalsdesignen. Som nämndes i avsnitt 4.1.4 finns även risken, som ibland kan vara stor, att få otursamma urval med OSU.

Det kan nämnas att stratifiering förekommer även i situationer där man måste tillgripa urvalsmetoder som inte är sannolikhetsurval, vilket är mindre vanligt på SCB. Stratifieringen avses då i någon mån stävja den urvalsbias som ett sådant urvalsförfarande ger upphov till (se avsnitt 4.9.2).

4.2.3 Fördelar

Det finns väsentliga fördelar med stratifiering. Här diskuteras några av dem.

Möjligt att kontrollera precisionen i redovisningsgrupper
Det är mycket vanligt att skattningar ska tas fram för olika redovisningsgrupper. För att kunna garantera en tillräcklig urvalsstorlek inom varje redovisningsgrupp, brukar man låta – så långt det är möjligt – varje sådan utgöra ett stratum eller en grupp av strata.

Homogenisering
Man försöker i allmänhet forma strata som är inbördes mer *homogena* än populationen som helhet. Det är lättare att hantera en homogen än en heterogen population när man drar urval och gör skattningar. Bland annat krävs en mindre urvalsstorlek för att få en given precision om man drar ett stratifierat urval från inbördes homogena strata än om man drar ett urval direkt ur hela populationen. Dessutom minskar risken för att få ett otursamt urval (jämför med OSU avsnitt 4.1.4).

Populationen kan vara heterogen med avseende på en mängd förhållanden. Undergrupper kan vara mer eller mindre ”svåra” eller kostsamma att få kontakt med. Svarsbenägenheten kan också variera inom populationen. Se Särndal och Lundström (2005, avsnitt 10.3) för jämförelser av robustheten mot bortfall hos några vanliga estimatorer. Det kan även finnas stora skillnader i förekomst av – och storlek på – mätfel i olika delar av populationen.

När ger stratifierat urval bättre precision? Anta att undersökningsvariabeln y är kvantitativ och att man formar strata utifrån objektens storlek. Om det finns ett samband mellan objektens storlek och y, är det nästan alltid så att *populationsvariansen* av variabeln y i stratum h, $S^2_{yU_h}$, är mindre än inom populationen som helhet, dvs. $S^2_{yU_h} < S^2_{yU}$. Se formel (4.2.6) nedan för hur S^2_{yU} beräknas. Populationsvariansen $S^2_{yU_h}$ är en av de faktorer som påverkar precisionen i skattningarna. Vid given urvalsstorlek leder stratifiering till ökad precision i skattningen om stratifieringsvariabeln är väl korrelerad med undersökningsvariabeln.

Ämnesskäl
En undersökning täcker ett ämnesområde, men detta kan i sin tur bestå av mer eller mindre starkt sammanhängande ämnen som har stort egenintresse var för sig. Varje ämne kan kräva speciell behandling, t.ex. i form av olika blanketter. I undersökningen ”Utrikeshandel med tjänster” och i en del andra företagsundersökningar används branschanpassade blanketter.

Administrativa skäl
Fältarbetet i intervjuundersökningar underlättas av geografisk stratifiering där datainsamlingen organiseras inom strata. SCB:s fältintervjuare är fördelade på sju regioner. I en specifik undersökning kan en region omfatta ett eller flera strata. I Konsumentprisindex används olika datainsamlingssätt i olika branscher, som utgör strata på övergripande nivå.

Allmänt gäller att stora undersökningar är som stora projektarbeten: de tjänar oftast på att delas upp i hanterliga bitar. De administrativa skälen och ämnesskälen går ofta in i varandra.

4.2.4 Nackdelar

Tidskrävande
Att konstruera strata kan vara tidsödande. I återkommande undersökningar krävs också att stratifieringen underhålls. Objekt kan ändra sin stratumtillhörighet över tiden, och stratumindelning och allokering (fördelning av urvalsstorlek över strata) behöver ses över.

Kräver god hjälpinformation
För att uppnå stratifieringens alla fördelar måste ramen innehålla en eller flera hjälpvariabler som kan användas för att avgränsa relevanta strata.

Försvårar analys
Det är vanligt att man drar urval med olika inklusionssannolikheter i olika strata. Analys av data kan då bli mer komplicerad än vid OSU. Se vidare avsnitt 4.9.4.

4.2.5 Stratumkonstruktion

För att kunna dra ett stratifierat urval måste tre huvudfrågor besvaras: hur många strata man ska ha, hur stratumgränserna ska väljas och vilken urvalsdesign och urvalsstorlek som ska användas inom stratumen när dessa väl konstruerats. Dessa huvudfrågor kan inte alltid behandlas sekventiellt.

Det finns i princip två sorters stratumindelningar: dels de som bestäms av någon eller några *kategorivariabler* (t.ex. region och kön, eller bransch i företagsstatistik), dels de som baseras på en eller flera *kvantitativa variabler* (t.ex. antal anställda eller inkomst).

Kategori- och storleksstrata
De strata som blir resultatet efter stratifiering med en kategorivariabel benämns *kategoristrata*. Här följer några allmänna råd om utformning av kategoristrata:

- De ska kunna gå att kombinera så att varje viktig redovisningsgrupp motsvarar en union av kategoristrata så väl som möjligt. Exempelvis kan strata utgöras av län, medan redovisningsgrupperna är län eller länsgrupper. Om redovisningsgrupperna är många och överlappande kan det dock bli svårt att utforma kategoristrata. Då kan man kanske fokusera på de viktigaste redovisningsgrupperna.
- De kan förbättra precisionen och minska biasen i skattningarna genom homogenisering, se avsnitt 4.2.3.
- De ska vara anpassade till administrativa krav.
- De ska inte vara så många att urvalsstorlekarna inom kategoristrata blir för små. Se vidare nedan under rubriken ”Antal strata”.

Efter att kategoristrata har upprättats kan man gå vidare med finare stratifiering inom varje sådant stratum. Den finare stratifieringen kan göras med ytterligare kategorivariabler eller utifrån kvantitativa variabler.

Om det finns en kvantitativ variabel i ramen är det ofta bra att stratifiera efter den inom kategoristrata. De strata som fås vid stratifiering efter en kvantitativ variabel kan kallas *storleksstrata*, eftersom dessa strata är ordnade efter objektens storlek. På engelska är termen *size strata* etablerad. Skälet att införa storleksstrata är framför allt homogenisering. Variabler som används för indelning i kategori- eller storleksstrata kallas för *stratifieringsvariabler*.

Den första frågan är vilken eller vilka stratifieringsvariabler som ska användas. Oftast använder man en enda variabel inom ett kategoristratum för storleksstratifiering. Det ska vara en kvantitativ variabel som uppfyller två krav: den ska ha hög korrelation med viktiga undersökningsvariabler och den ska ha ett registrerat värde för alla eller nästan alla objekt i ramen (se exempel 4.2.1 och 4.2.3 samt även avsnitt 7.5 för hur resterande objekt kan behandlas). Även om korrelationen är låg kan storleksstratifiering ge en viss precisionsvinst (Cochran 1977, formel 5.37, s. 100).

Finns det mer än en kandidat till stratifieringsvariabel kan man överväga att stratifiera med hjälp av en korsklassificering efter två eller flera variabler, eller bilda en funktion av variablerna. Exempelvis kan man tänka sig att använda det aritmetiska medelvärdet av de tillgängliga registervariablerna som stratifieringsvariabel. I litteraturen finns ett stort antal andra förslag på hur man kan använda mer än en kvantitativ variabel, men de metoderna är ganska opraktiska. Det behövs mer utredning för att belysa om nyttan motsvarar de extra insatserna. Se avsnitt 4.2.10 för referenser.

EXEMPEL 4.2.1 *Djurräkningen*
I Djurräkningen (se exempel 3.2.3) avseende juni 2001 stratifierades huvudsakligen efter fjolårsantalet nötkreatur, svin och får, enligt ramen som avgränsats från Lantbruksregistret 2000. Varje djurslag utgjorde en egen stratifieringsvariabel för storleksstratifiering. Därutöver fanns några kategoristrata för särskilda kategorier av djurhållare, t.ex. slaktkycklingsuppfödare. Vissa av de planerade stratumen slogs samman för att uppnå en minimistorlek på ungefär 20 företag för alla strata. Sedan drogs ett urval på 15 100 företag, fördelat över 98 strata. Minst 40 företag drogs i respektive urvalsstratum.

Att använda flera stratifieringsvariabler med relativt få storleksklasser för varje befanns vara mycket effektivare än att finindela enligt *en* stratifieringsvariabel. Anledningen till detta var att flera parametrar – främst parametrar för nötkreatur, svin och får – skulle skattas och att dessa byggde på relativt svagt korrelerade variabler. □

Givet att man har bestämt stratifieringsvariabel och antal storleksstrata – hur ska storleksstrata konstrueras? Det är skillnad på om strata ska undersökas med urval eller om de ska totalundersökas, dvs. om alla objekt i stratumet ska undersökas.

Bestämning av gränser för strata som undersöks via urval
Här behandlas bestämning av stratumgränser för storleksstrata som undersöks genom ett urval. Allmänt kan man säga att det inte spelar så stor roll för precisionen exakt hur gränserna för sådana strata dras. Även om gränserna flyttas en hel del, påverkas inte precisionen i skattningarna nämnvärt – se Dalenius (1957) och Hedlin (1998). Det viktigaste undantaget är om något storleksstratum används även för att avskilja en redovisningsgrupp. Även när gränsen gäller ett stratum som ska totalundersökas kan det ibland bli känsligt hur den dras.

En annan fråga som är viktig i undersökningar av företag och organisationer är uppgiftslämnarbördan. Objekten i totalundersökta strata kommer med i varje omgång av undersökningen och bördan blir stor med tiden. Omfattningen kan möjligen undvikas i någon mån med mindre totalundersökningsstrata. Ett större problem med tung uppgiftsbörda uppstår dock för små och medelstora företag i små branscher. Där kan urvalsfraktionerna bli orimligt stora.

Den vanligaste metoden för att konstruera urvalsundersökta storleksstrata kallas $cum\sqrt{f}$ -metoden. Särndal et al. (1992, s. 463–464) ger en bra beskrivning som här återges kort. Två numeriska exempel ges i Cochran (1977, s. 129–130).

$cum\sqrt{f}$ -metoden

1. Sortera objekten i storleksordning efter stratifieringsvariabeln.
2. Dela in objekten i ett antal klasser med samma klassbredd. Det finns ingen regel som säger hur många klasser man behöver. Det beror på hur många objekten är. Det brukar gå bra med fyra till sex gånger flera än antal storleksstrata inom varje kategoristratum.
3. Räkna ut $\sqrt{f_j}$ för varje klass där f_j är antalet objekt i klass *j*.
4. Bilda större grupper genom att slå samman närliggande klasser på så sätt att summan av $\sqrt{f_j}$ blir ungefär lika stor för alla grupper.
5. Varje slutlig grupp blir ett stratum.

Ekmans regel är en annan stratifieringsmetod som ibland används inom SCB. Hedlin (2000) beskriver Ekmans regel, vilken först publicerades i Ekman (1959). Ekmans regel och $cum\sqrt{f}$ -metoden ger liknande resultat (Hedlin 1998, avsnitt 5.5.3).

Den situation som Ekmans regel och $cum\sqrt{f}$ -metoden är designade för är HT-estimatorn (se avsnitt 2.3.1), Neymanallokering (se avsnitt 4.2.7), små urvalsfraktioner och mycket hög korrelation mellan stratifierings- och undersökningsvariabel. Men eftersom storleksstratifiering är ganska okänslig för avvikelser, är metoden mer generellt användbar än så.

En annan metod för storleksstratifiering är Wrights modellbaserade stratifiering. Den beskrivs på ett bra sätt i avsnitt 12.4 i Särndal et al. (1992). En enkel variant är *the equal aggregate σ rule*. Den kan utformas på olika vis, men i många fall är det lämpligt enligt den metoden att stratifiera på sådant sätt att för varje färdigt storleksstratum gäller

$$\sum_{U_h}\sqrt{x_k} \approx \sum_{U}\sqrt{x_k}\,/H \qquad (4.2.1)$$

där x_k är stratifieringsvariabelns värde för objekt *k*, summan $\sum_{U_h}\sqrt{x_k}$ tas över de objekt som hör till stratum *h*, och *H* är antalet strata. Se Särndal et al. (1992, punkt 2 på s. 462). För att använda denna stratifieringsregel flyttar man stratumgränser tills formel (4.2.1) är ungefärligen uppfylld. Metoden innebär att summan av $\sqrt{x}$ -variabeln ska vara ungefär lika i alla strata.

Om man har tänkt totalundersöka en del av populationen, undantas denna del av populationen från urvalsdragningen. Sedan används stratifieringsregeln som uttrycks i formel (4.2.1) på den resterande delen av populationen. Uttryckt i en formel blir den modifierade stratifieringsregeln

$$\sum_{U_h}\sqrt{x_k} \approx \sum_{U-U_H}\sqrt{x_k}\,/(H-1), \qquad (4.2.2)$$

där summan $\sum_{U-U_H}\sqrt{x_k}\,/(H-1)$ tas över objekt i $U-U_H$, dvs. den del av populationen som urvalsundersöks. De objekt som har inklusionssannolikhet 1 läggs i det totalundersökta stratumet *H.* Eftersom ett stratum har ”gått åt” till ett totalundersökningsstratum, är det resterande antalet strata *H*–1.

I båda fallen, (4.2.1) och (4.2.2), gäller att urvalet allokeras på så sätt att varje urvalsundersökt stratum får lika urvalsstorlek. Om denna stratifieringsmetod väljs, används stratifieringsvariabeln med fördel även i estimationen i någon form av kalibrering, t.ex. en kvotestimator (för beskrivning av kvotestimator, se någon av de böcker som nämns i kapitel 10; för kalibrering, se Särndal och Lundström 2005).

Ytterligare ett alternativ, som är vanligt inte minst internationellt, är samma stratifieringsregel och estimator som ovan men med $\sqrt{x_k}$ utbytt mot x_k, se Särndal et al. (1992), punkt 3 på s. 462. Metoden dikterar alltså att summan av *x*-variabeln ska vara ungefär lika i alla strata. Det är dock mindre lämpligt i företagsstatistik, eftersom stratumen då blir av väldigt olika storlek. Metoden med lika *x*-summa fungerar bäst om variationskoefficienten $S_{xU_h}/\bar{X}_h$, där $\bar{X}_h$ är medelvärdet av *x*-variabeln i stratum *h*, inte ändrar sig mycket ens om man ruckar på stratumgränserna ganska kraftigt. Den generella varianten av Wrights modellbaserade stratifiering är beskriven i Särndal et al. (1992, s. 459–462).

Det bör också påpekas att *πps*-urval ofta är ett alternativ till storleksstratifiering följt av obundet slumpmässigt urval. Med *πps*-urval slipper man att avgöra hur många storleksstrata som behövs och hur dessa ska konstrueras. I undersökningen ”Företagens ekonomi” används *πps*-urval inom kategoristrata (bransch) i stället för stratifierat OSU just av dessa praktiska skäl. Inom varje bransch dras ett *πps*-urval oberoende av urvalen inom andra branscher. Se vidare avsnitt 4.4.

EXEMPEL 4.2.2 *Utrikeshandel med tjänster*
I undersökningen ”Utrikeshandel med tjänster” är det viktigt att sätta inklusionssannolikheten relativt högt för företag som troligen handlar med tjänster internationellt och lågt för andra företag. Det görs genom att en indikator skapas i ramen, där de olika förhållanden som är kända och som tyder på utrikes tjänstehandel tas med i indikatorn. Denna kan anta värdet 0 eller 1 och bildar på så sätt två kategoristrata. Dessa i sin tur stratifieras finare med avseende på bransch och storlek. Urvalsmetoden är stratifierat OSU. Urvalsfraktionen är högre i strata där indikatorns värde är ett. Som stratifieringsvariabel för storleksstrata används numera summan av omsättningen och värdet av tjänsteexporten enligt momsregistret. Tidigare användes endast omsättning. □

EXEMPEL 4.2.3 *Omsättning inom detaljhandeln*
I undersökningen ”Omsättning inom detaljhandeln” används omsättning som stratifieringsvariabel för storleksstrata. De företag som saknar värde i ramen på omsättning placeras i ett särskilt kategoristratum, vari storleksstratifieringen görs efter lönesumma. I båda fallen är urvalsmetoden stratifierat OSU. □

Bestämning av gränser för strata som totalundersöks
Det finns flera tänkbara skäl till att totalundersöka ett stratum:

- Objekten i stratumet är extremt stora. Nästan alla företagsundersökningar har något totalundersökt stratum bestående av riktigt stora företag.
- Stratumet omfattar ett fåtal objekt.
- Stratumet består av ett fåtal specialbevakade objekt som befaras generera stora eller avvikande värden på undersökningsvariablerna.
- Stratumet utgör urvalsobjekten från en pilotundersökning. Då är det viktigt att pilotundersökningens tidpunkt och metoder ligger nära huvudundersökningens.

En vanlig metod för att avgränsa totalundersökta strata för stora objekt är att utföra stratifiering med $cum\sqrt{f}$ -metoden, allokera urvalet med Neymanallokering och konstatera att ett eller flera strata blev totalundersökta. Metoden är inte helt teoretiskt korrekt, eftersom $cum\sqrt{f}$ -metoden, liksom andra metoder av samma slag, förutsätter att inget stratum är totalundersökt. Emellertid finns det erfarenhetsmässiga skäl att tro att den stratifiering man får på så sätt ändå ofta är relativt bra, framför allt därför att en urvalsundersökning är ganska okänslig för exakt hur stratumgränserna bestäms – se t.ex. Hedlin (1998).

I avsnitt 4.2.10 finns referenser till flera metoder för att bestämma gränser för strata som ska totalundersökas. Metoderna har – till skillnad från $cum\sqrt{f}$ -metoden – teoretiskt stöd. Ingen av metoderna är dock särskilt praktisk. Metoderna kan vanligen bara användas i vissa specifika situationer.

Antal strata
Så återstår frågan om hur många strata som behövs. Här är det återigen skillnad på kategori- och storleksstrata. Främst i det senare fallet är det nödvändigt att även göra skillnad på om strata ska undersökas med urval eller om de ska totalundersökas. Det finns argument för att man sällan behöver undersöka flera än ungefär sex storleksstrata inom ett kategoristratum via urval, se Cochran (1977, s. 132–134). Ofta räcker det med fyra eller fem eller ännu färre. Att ha för många strata kan leda till problem: främst kan urvalen inom vissa strata bli för små (se avsnitt 4.2.10 för referenser till vidare läsning). Rådet är alltså att ha få storleksstrata inom varje kategoristratum, åtminstone inte flera än omkring sex urvalsundersökta storleksstrata. Det finns undantag, främst då vissa storleksstrata ska användas även som redovisningsgrupper.

Det är svårare att ge precisa råd för hur många kategoristrata som behövs. Det är vanligt att man efter en preliminär stratifiering i kategoristrata får för många strata. Då kan man slå samman vissa av dessa preliminära strata eller utesluta vissa av stratifieringsvariablerna. Se exempel 4.2.5.

EXEMPEL 4.2.4 *Lantbrukets struktur*
I strukturundersökningen för lantbruket 1997 stratifierades av homogenitetsskäl efter driftsinriktning och storlek på lantbruksföretagen. Län utgjorde en viktig redovisningsgrupp, men att stratifiera även på län bedömdes ge alltför många, små strata. Därför tillämpades en kompromiss, med stratifiering på tre länsgrupper. Således användes tre stratifieringsvariabler, vilka resulterade i totalt 74 strata. Urvalsstorleken var 32 000 företag. □

EXEMPEL 4.2.5 *Gymnasieungdomars studieintresse*
I undersökningen ”Gymnasieungdomars studieintresse läsåret 2002/03” var två av stratifieringsvariablerna region och kön. Ålder används ofta i individundersökningar som ytterligare stratifieringsvariabel. I den här populationen har nästan alla samma ålder, varför den variabeln inte kan användas för stratumindelning. Studieprogram på gymnasiet är en viktig variabel i en undersökning av intresse för högre studier. För att minska antalet strata från 273 till 32 slogs studieprogram samman till grupper av liknande program. □

4.2.6 Urvalsdesign inom stratum

För varje stratum måste en urvalsdesign bestämmas. Den viktigaste principen är att urvalet i ett stratum dras oberoende av urvalen i alla andra strata. Det finns dock urvalsmetoder som utgör undantag från denna princip: se avsnitt 4.2.10 om *controlled selection*.

Urvalet kan dras med helt olika metoder i olika strata. Det är dock ovanligt inom SCB. Det är tänkbart att använda *πps*-urval i strata där hjälpinformationen är stark och OSU i strata med ingen eller mindre stark hjälpinformation. I länder som saknar befolkningsregister är det vanligt att stratifie-

ringen på den grövsta nivån avser stad respektive landsbygd. Urvalsdesignen inom dessa två övergripande strata kan vara helt olika.

Det i praktiken vanligaste inom SCB är att dra ett OSU inom varje stratum, så kallat *stratifierat OSU*. Andra vanliga designer är att dra systematiskt urval eller *πps*-urval inom varje stratum.

4.2.7 Allokering

Problemet att bestämma urvalsstorleken tas upp i avsnitt 6.1. Vid ett stratifierat urval måste man dessutom fördela urvalet över strata. Den processen kallas *allokering*. Om man vill allokera så att man får bästa möjliga precision för skattningar som avser hela populationen, givet en fast, förutbestämd kostnad, utför man *optimal allokering*. Nedan finns formler för sådan allokering samt även för situationen att man i stället vill allokera så att man får lägsta kostnad givet en fix precision. Allokering beror bland annat på urvalsmetod och estimator inom stratum. Generella formler finns i resultat 3.7.3–3.7.4 i Särndal et al. (1992, s. 105).

Varianter på optimal allokering

Grundformeln för optimal allokering under stratifierat OSU lyder

$$n_h = n\frac{N_h S_{yU_h}\Big/\sqrt{c_h}}{\sum_{h=1}^{H} N_h S_{yU_h}\Big/\sqrt{c_h}} \qquad (4.2.3)$$

där n är den totala urvalsstorleken, N_h är antal objekt i stratum h i populationen, $S^2_{yU_h}$ är populationsvariansen av undersökningsvariabeln y i stratum h, och c_h är kostnaden för att undersöka ett objekt i stratum h (alla objekt i samma stratum antas ha samma undersökningskostnad). Med andra ord väljs urvalsstorleken n_h i stratum h proportionellt mot stratumstorleken multiplicerad med stratumets standardavvikelse och dividerad med kvadratroten ur kostnaden. Kostnaden kan antingen ligga på producentens sida eller på uppgiftslämnarens. I det senare fallet talar man om *uppgiftslämnarbörda* eller *uppgiftslämnarkostnad*. För bestämning av den totala storleken på urvalet, n, se avsnitt 6.1.

Optimal allokering är optimal just för en HT-estimator (se avsnitt 2.3.1) vid stratifierat OSU om man känner spridningen av undersökningsvariabeln, populationsvariansen $S^2_{yU_h}$. Eftersom populationsvariansen $S^2_{yU_h}$ är okänd måste den skattas eller gissas. Se avsnitt 6.1.10 för metoder och tips. I allokeringen används någon skattning av eller substitut för S_{yU_h} som här allmänt betecknas med S_{xU_h}.

Om man ersätter S_{yU_h} med S_{xU_h} i formel (4.2.3), där S_{xU_h} beräknas för hjälpvariabeln x, får man s.k. *x-optimal allokering.* Variabeln x kallar vi *allokeringsvariabel*.

Om alla objekt i ramen har samma undersökningskostnad, oavsett stratum, förenklas formeln för x-optimal allokering till

$$n_h = n\frac{N_h S_{xU_h}}{\sum_{h=1}^{H} N_h S_{xU_h}}. \qquad (4.2.4)$$

På SCB används termen Neymanallokering för såväl den teoretiska formeln

$$n_h = n\frac{N_h S_{yU_h}}{\sum_{h=1}^{H} N_h S_{yU_h}}$$

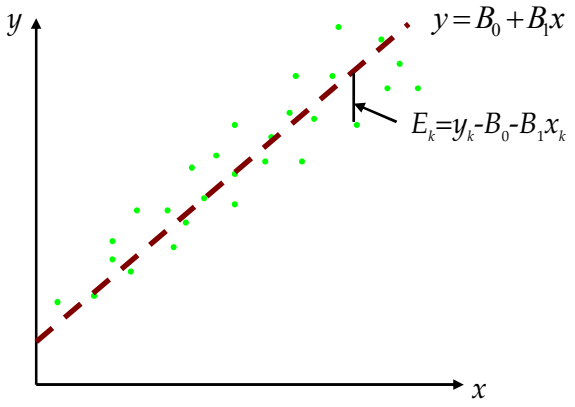
som den i praktiken använda motsvarigheten där S_{yU_h} skattas med S_{xU_h}. Termen *x*-optimal allokering används däremot sällan på SCB. I litteraturen brukar termen Neymanallokering reserveras för den teoretiska formeln.

Neymanallokering kan i de flesta fall användas även om undersökningskostnaden inte är lika för alla objekt. Det är nämligen bara vid mycket stora skillnader i kostnader som undersökningens effektivitet påverkas av att allokeringen tar hänsyn till kostnaderna. Valliant et al. (2000, s. 171–172) ger ett räkneexempel där det i vissa strata är upp till tre gånger så dyrt att undersöka varje objekt som i andra strata. Ändå påverkas inte undersökningens effektivitet nämnvärt av att man under samma budget allokerar antingen med eller utan hänsyn till skillnaderna i undersökningskostnad per objekt. Större skillnader i kostnader uppstår framför allt om somliga objekt undersöks på ett annat sätt än övriga objekt. Ibland blandar man datainsamlingssätt, t.ex. telefon- och besöksintervju. Ibland har man ”lång blankett” och ”kort blankett” som kan ge stora kostnadsskillnader.

Om c_h avser uppgiftslämnarbörda och inte statistikproducentens kostnad, kan inte skillnader i c_h försummas lättvindigt, eftersom det då inte enbart är en budget- och effektivitetsfråga.

En allokering med någon av formlerna (4.2.3) och (4.2.4) är inte optimal om någon annan estimator än HT-estimatorn ska användas. För en GREG-estimator (se avsnitt 2.3.1) ges den optimala allokeringen av samma formler, men med S_{xU_h} utbytt mot S_{EU_h}, dvs. populationsstandardavvikelsen av residualerna E_k vid en tänkt anpassning av en regressionslinje till hela populationens värden. Residualerna E_k åskådliggörs i figur 4.2.1. De definieras i Särndal et al. (1992, formel (6.5.14)) och diskuteras i avsnitt 6.7 i samma bok. Liksom är fallet med S_{yU_h} är S_{EU_h} okänd. Samma metoder som används för att skatta eller gissa S_{yU_h} kan användas för S_{EU_h}, se avsnitt 6.1.10. Exempelvis kan man i en löpande undersökning som inte ändrar stratumgränserna få en uppfattning om S_{EU_h} genom att räkna på det senaste urvalet.

Figur 4.2.1 Residualer vid en tänkt anpassning av en regressionslinje till alla objekt i hela målpopulationen



Vet man inte något om storleken på S_{yU_h} respektive S_{EU_h} kan det vara klokt att använda *proportionell allokering* i stället, dvs. att varje stratum får en urvalsstorlek i proportion mot det antal objekt som finns i ramen och som hör till stratumet. I en formel uttrycks den allokeringen så här:

$$n_h = n \frac{N_h}{\sum_{h=1}^{H} N_h} \qquad (4.2.5)$$

Om alla S_{xU_h} (eller S_{yU_h}) är lika reduceras alltså Neymanallokering till proportionell allokering, dvs. alla urvalsstorlekar ska sättas proportionella mot stratumstorlekarna. Om korrelationen mellan stratifieringsvariabeln och viktiga undersökningsvariabler inte är alltför stark är det ofta bra att använda proportionell allokering i stället för Neymanallokering. Särndal et al. (1992), s. 464, anger 0,8 som ungefärlig tumregel för hur stark korrelationen minst bör vara för att Neymanallokering ska löna sig framför proportionell allokering. Det gäller också om de olika S_{xU_h} är svåra att beräkna på grund av partiellt bortfall, avvikande värden eller förmodade mätfel. Det gäller även, som redan nämnts, om S_{xU_h} skattas på basis av ett urval och den skattningen kan misstänkas ha låg precision. Proportionell allokering ger relativt stora urval i strata med många objekt, dvs. i strata där N_h är stort. I företagsundersökningar kan dock proportionell allokering leda till orimligt stora urval i strata med många små företag.

Allokering när undersökningsvariabeln är en kategorivariabel

Om undersökningsvariabeln är en kategorivariabel med två kategorier, gäller $S_{yU_h}^2 \approx P_h(1-P_h)$ där P_h är andelen i populationen som tillhör ena kategorin (se formlerna 4.2.6–4.2.8).

I planeringsstadiet av en undersökning vet man ofta inte vad P_h kommer att bli. Då kan det vara lämpligt att vara försiktig och sätta $P_h(1-P_h)$ till dess maximivärde, som är 1/4. Med avseende på en kategorivariabel med två kategorier är en population nämligen mest heterogen om hälften tillhör den ena klassen och hälften tillhör den andra. Följden blir proportionell allokering, eftersom $S_{yU_h}^2 \approx P_h(1-P_h)$ har satts till samma värde, $(1/2)^2 = 1/4$, i alla strata.

Populationsvariansen för en variabel y definieras som

$$S_{yU}^2 = \sum_{k=1}^{N} \left(y_k - \bar{y}_U\right)^2 / \left(N-1\right) \qquad (4.2.6)$$

där $\bar{y}_U$ är medelvärdet i populationen. Om y är en kategorivariabel som antar två värden med proportionerna P respektive 1–P gäller att

$$S_{yU}^2 = \left[N/\left(N-1\right)\right]P\left(1-P\right) \qquad (4.2.7)$$

Om $\frac{N}{N-1} \approx 1$ gäller approximationen

$$S_{yU}^2 \approx P\left(1-P\right), \qquad (4.2.8)$$

som i regel är användbar.

Allokering som ger orimligt stora eller små urvalsstorlekar
Om fördelningen av den variabel som används för allokeringen är sned, kommer troligen något n_h, som framräknas med formlerna för olika varianter av optimal allokering ovan, att bli större än antalet objekt i stratumet. Då totalundersöker man vanligen det stratumet (alltså $n_h = N_h$).

Vidare kan vissa urvalsstorlekar bli mycket små. Då ökas dessa till någon lämplig minsta urvalsstorlek. Sedan beräknas resterande urvalsstorlekar på nytt med

$$n_h = n' \frac{N_h S_{xU_h}}{\sum_{h=1}^{H'} N_h S_{xU_h}} \qquad (4.2.9)$$

där n' och H' är den urvalsstorlek och det antal strata för vilka det fortfarande återstår att bestämma urvalsstorlekar.

Det finns ingen bra tumregel för minsta urvalsstorlek i ett storleksstratum. Australian Bureau of Statistics har sex utvalda objekt som minimum (Brewer 2002, s. 60). SCB har ingen allmän regel, utan det avgörs från fall till fall. Faktorer som beaktas är hur stort bortfallet förväntas bli och hur stor uppgiftslämnarbördan är. En annan faktor att ta hänsyn till är vilken estimator som man planerar att använda och hur känslig den är för avvikande värden i hjälpinformationen. HT-estimatorn är helt okänslig, eftersom den inte använder hjälpinformation, medan en regressionsestimator baserad på en kvantitativ hjälpvariabel kan vara mycket känslig vid små urvalsstorlekar inom strata.

Notera att om stratifieringen avser primära urvalsenheter (kluster) i ett flerstegsurval gäller andra regler för minsta urval per stratum, se avsnitt 4.6.

Allokering för att få lägsta kostnad för given precision
Den formel som ger lägsta kostnad och som ger en förutbestämd varians av en skattning av en total lyder

$$n_h = \left(N_h S_{xU_h} \Big/ \sqrt{c_h}\right) \frac{\sum_{h=1}^{H} N_h S_{xU_h} \sqrt{c_h}}{V + \sum_{h=1}^{H} N_h S_{xU_h}^2} \qquad (4.2.10)$$

där V är den förutbestämda estimatorvariansen. Denna formel motsvarar alltså (4.2.3) men löser ett annat, snarlikt problem. Formel (4.2.10) förutsätter optimal allokering. Cochran (1977, avsnitt 5.9) ger även mer allmänna allokeringsformler för stratifierat OSU.

För att få fram ett lämpligt värde på V är ett sätt att beställaren eller intressenterna av statistiken talar om vilken osäkerhetsmarginal som behövs relativt skattningen. Till stöd för bedömningen kan man resonera utifrån vilka och hur säkra slutsatser som ska kunna dras av undersökningens resultat. Det kan gälla t.ex. hur stora skillnader mellan olika parametervärden som ska kunna beläggas signifikant. För en skattning av totalen kan osäkerhetsmarginalen relativt skattningen uttryckas allmänt som $1{,}96\sqrt{V}/\hat{t}$, där $\hat{t}$ är totalskattningen. Se vidare avsnitt 6.1.10 för råd om hur totalen t kan skattas eller gissas om det inte finns någon god förhandsuppfattning.

Allokering för andra stratifierade urvalsmetoder
Följande variansformel för en HT-estimator gäller exakt eller approximativt för de flesta stratifierade enstegsurval:

$$V\left(\hat{t}_y\right) = B + \sum_{h=1}^{H} A_h / n_h \qquad (4.2.11)$$

För stratifierat OSU gäller den exakt med

$$A_h = N_h^2 S_{yU_h}^2$$

$$B = -\sum_{h=1}^{H} N_h S_{yU_h}^2$$

För variansen i (4.2.11) är den optimala allokeringen:

$$n_h = n \frac{\sqrt{A_h}}{\sum_{h=1}^{H} \sqrt{A_h}} \qquad (4.2.12)$$

Allokeringen (4.2.12) kan användas även för Pareto *πps*-urval (som beskrivs nedan i avsnitt 4.4.1). Pareto *πps*-urval, dragna separat inom kategoristrata $h = 1, 2, \ldots H$, har variansen för en HT-estimator (Rosén 2000a):

$$V\left(\hat{t}_y\right) \approx \sum_{h=1}^{H} \frac{N_h}{N_h - 1}\left(D_h - F_h^2 / C_h\right) \qquad (4.2.13)$$

där

$$D_h = \sum_{U_h} \frac{y_k^2 \cdot \left(1 - \lambda_k\right)}{\lambda_k}$$

$$F_h = \sum_{U_h} y_k \cdot (1 - \lambda_k)$$

$$C_h = \sum_{U_h} \lambda_k \cdot (1 - \lambda_k)$$

där λ_k definieras i formel (4.4.2). Om $\lambda_k \approx n_h / N_h$ inom strata, kan (4.2.13) approximeras med (4.2.11) för stratifierat OSU. Då kan allokeringen (4.2.12) användas med $A_h = N_h^2 S_{yU_h}^2$ och $B = -\sum_{h=1}^{H} N_h S_{yU_h}^2$.

Nackdelen med denna approximation är att urvalsstorleken kan bli för stor i kategoristrata där λ_k varierar särskilt mycket. Ofta har man dock andra mål än minsta varians för skattningar för *hela* populationen. I stället är det vanligt att man vill att skattningen för varje kategoristratum ska nå en viss förutbestämd precision. Se ett exempel i avsnitt 7.4.

Allokering med flera mål

I de flesta undersökningar har man flera mål, och då får man söka sig fram till någon lämplig kompromiss. En vanlig situation är att man vill ta fram skattningar både på övergripande nivå och för redovisningsgrupper, och har precisionskrav på alla nivåer. En metod som i det sammanhanget många gånger fungerar bra kallas *power allocation*, se Särndal et al. (1992, s. 470–471) och Bankier (1988). Foreman (1991, s. 122) ger en tumregel, som är ett specialfall av power allocation,

$$n_g = n \frac{Z_g}{\sum_{g=1}^{G} Z_g} \qquad (4.2.14)$$

där Z_g är något storleksmått för redovisningsgrupp $g = 1, 2, \ldots, G$. T.ex. kan Z_g vara summan av en hjälpvariabel som finns tillgänglig i ramen. Denna allokering används även ibland för allokering till kategoristrata då urvalsmetoden inom kategoristrata är *πps*-urval. Denna allokering kallas *allokering proportionell mot z-totalen*.

Cochran (1977, avsnitt 5A.3 och 5A.4) diskuterar olika metoder för allokering när man har mer än en undersökningsvariabel att ta hänsyn till.

Ibland vill man ta fram skattningar för enskilda strata för att jämföra dessa sinsemellan. Målet kan t.ex. vara att visa att något stratum har högre medelvärde än andra strata. Då ska man allokera proportionellt mot S_{xU_h}, där $S_{xU_h}^2$ är populationsvariansen av en hjälpvariabel i stratum *h* som antas vara korrelerad med undersökningsvariabeln. Om det inte finns någon större skillnad på S_{xU_h} i olika strata, ska man alltså för det andra ändamålet ha lika stora urval i varje stratum.

Allokering som beaktar bortfall och övertäckning

Ibland kan man på förhand bedöma ungefär hur stort bortfallet kommer att bli i olika grupper. Samma sak kan gälla övertäckningen. Om man har förhandskännedom om någon grupp som riskerar att ha dålig svarsbenägenhet, exempelvis ”unga män i storstäder”, kan man göra ”överurval” i denna grupp, alltså välja extra många. På så sätt försöker man åstadkomma att gruppen svarande får en sammansättning som blir lik den som ett

"idealt", icke bortfallsdrabbat, urval skulle ha haft. Överurval kan även användas i syfte att försöka kompensera för övertäckning i urvalsramen.

Allokeringen ovan kan höja precisionen i skattningarna. Men risken för ett systematiskt bortfallsfel (bortfallsbias) kvarstår. Denna risk kan minskas genom att man delar in populationen i strata som speglar svarsbenägenheten för olika grupper av objekt, och sedan beräknar skattningen genom rak uppräkning inom dessa strata.

Allokeringen är viktig
Allokeringen är ofta viktig för hela undersökningen och det är i allmänhet värt att lägga ned tid på det momentet. Det brukar krävas många överväganden för att hitta en kompromiss mellan olika krav. Den allokering man får med räknereglerna ovan ska inte uppfattas som en naturlag. Tvärtom ska man gå igenom hela listan av urvalsstorlekar och bedöma dem. Exempelvis kan någon urvalsstorlek bli stor på grund av avvikande värden i hjälpvariabeln. Som också nämnts, kan urvalen bli för små i andra strata. Det är svårt att ge exakta instruktioner för hur bedömningen av erhållna urvalsstorlekar ska gå till.

4.2.8 Hur kontrolleras att stratumindelningen och allokeringen blev bra?

I efterhand går det att kontrollera om en genomförd stratifiering och allokering resulterat i ett urval som ger bättre precision i skattningarna än ett OSU. Populationsvariansen kan skattas utifrån insamlade data för urvalet, se Särndal et al. (1992, formel (5.9.11)). Den skattade populationsvariansen används sedan i formeln för variansen för en punktskattning. Exempelvis ges variansen för en HT-skattning av en total under OSU i Särndal et al. (1992, formel (2.8.13)). Slutligen jämförs variansen under det stratifierade urvalet (formel (4.2.11)) med motsvarande varians under OSU. Det kan visa sig att OSU skulle ha varit bättre, även om det är mindre vanligt för den typ av urval som avses här. Stratifiering som inte ger någon homogenisering försämrar alltså precisionen. Det kan då vara skäl att ompröva stratifieringens utformning.

Samma typ av jämförelse kan göras separat för varje redovisningsgrupp – se exempel 4.2.6.

4.2.9 Exempel

I det följande ges exempel på olika aspekter av stratifiering. Även kapitel 7 ger exempel på flera undersökningar som använder stratifierat urval.

EXEMPEL 4.2.6 *Hushållens inkomster*
I ”Hushållens inkomster” användes på 1980-talet ett stratifierat urval. Det vanliga är att ett stratifierat urval ger bättre precision än ett OSU, varför Rosén (1986) gjorde en kontrollberäkning som visade vilken precision ett tänkt OSU skulle ge i ”Hushållens inkomster”. Det visade sig att det stratifierade urvalet med befintlig allokering var dåligt: precisionen var nämligen sämre i flera redovisningsgrupper än vad ett OSU hade gett. Stratifieringen och allokeringen förbättrades sedan. □

EXEMPEL 4.2.7 *Djurräkningen*
I Djurräkningen avseende juni 2001 (se exempel 3.2.3 och 4.2.1) utfördes *x*-optimal allokering (Neymanallokering) för antalsvariablerna nötkreatur, mjölkkor, svin respektive får. Fjolårsvärden, enligt ramen, användes för att skatta stratumvarianserna $S^2_{xU_h}$ i formel (4.2.4) för de fyra olika variablerna. En kompromissallokering gjordes utifrån dels användarnas krav på skattningsprecision för olika djurslag (nötkreatur värderades högre än svin och svin högre än får), dels förmodad svarsbenägenhet från företag med de olika djurslagen. □

EXEMPEL 4.2.8 *Omsättning inom detaljhandeln*
I undersökningen ”Omsättning inom detaljhandeln” delas populationen in i branscher, eftersom det krävs separata skattningar inom respektive bransch, som alltså utgör redovisningsgrupper. Varje bransch delas sedan in i storleksstrata. Neymanallokering beräknas på alla storleksstrata över alla branscher samtidigt. Huvudanvändaren Nationalräkenskaperna prioriterar god precision för en skattning av omsättning för alla branscher tillsammans framför god precision för samtliga branscher var för sig. Neymanallokeringen applicerad på alla storleksstrata samtidigt leder till att branscher med låg omsättning får relativt små urval medan branscher med hög omsättning får jämförelsevis stora urval. Det är önskvärt i den här undersökningen. Allokeringsberäkningarna ses bara som vägledande. Manuell justering av enstaka urvalsstorlekar förekommer. □

EXEMPEL 4.2.9 *Företagens ekonomi*
I undersökningen ”Företagens ekonomi” allokeras urvalet över branscher med ett iterativt förfarande på så sätt att de viktigaste skattningarna i alla branscher får ungefär samma precision i absoluta tal. Förfarandet skiljer sig alltså från det i undersökningen ”Omsättning inom detaljhandeln” i exempel 4.2.8, eftersom användarkraven är olika. Avsnitt 7.4 innehåller en längre beskrivning av ”Företagens ekonomi”. □

4.2.10 Läs mer

Är det från varianssynpunkt någon skillnad mellan att använda sig av stratifierat OSU och OSU kombinerat med en poststratifierad estimator? Det beror på om man tar hänsyn till att urvalets fördelning över strata är slumpmässig vid poststratifiering. Holt och Smith (1979) argumenterar för att man inte ska ta hänsyn till denna källa till slumpmässighet. I så fall ger båda ansatserna samma varians.

Det finns enormt många metoder för stratifiering föreslagna i litteraturen, men det lönar sig knappast att sätta sig in i alla dessa. Gunning och Horgans (2004) metod är mycket enkel, men behöver testas mer utförligt innan den kan rekommenderas.

Det finns många förslag i litteraturen på stratifiering som utnyttjar mer än en stratifieringsvariabel utan att det blir för många strata totalt, bland annat Drew et al. (1985), Lavallée (1988), Roshwalb och Wright (1991), Rivest (2002), Lu och Sitter (2002) samt Fabrizi och Trivisano (2006). Skinner et al. (1994) ger en översikt.

En översikt över stratifiering av sneda populationer ges i Sigman och Monsour (1995), Hedlin (1998) och Horgan (2006). Ingen av dem ger dock en heltäckande bild av litteraturen på området.

Metoder för att avgränsa totalundersökta strata föreslås av Lavallée och Hidiroglou (1988), Hedlin (1998), Hidiroglou och Laniel (2001) samt Rivest (2002). Metoderna i de två första arbetena utgår från att korrelationen mellan stratifierings- och undersökningsvariabel är mycket hög. Om korrelationen är måttlig, kommer dessa metoder att ge ett alltför stort totalundersökt stratum. Rivest (2002) ger en länk till ett egenproducerat SAS-program som man kan använda för att avgränsa totalundersökta strata.

Optimal allokering för flera än en undersökningsvariabel diskuteras i Särndal et al. (1992, s. 467 och framåt). En ganska teoretisk framställning finns i Danielsson (1975).

I Cochran (1977, avsnitt 5A.7) diskuteras en metod att dra stratifierat urval där det finns beroende mellan urvalen i olika strata (*controlled selection*). Metoden kan ge lägre kostnader vid framför allt besöksintervjuer. Nackdelen är besvärliga variansskattningar. Se Valliant et al. (2000, s. 72) för referenser till praktisk användning av denna metod i USA.

4.3 Systematiskt urval

4.3.1 Beskrivning av metoden

Enkelt uttryckt innebär ett *systematiskt urval* att man väljer en startpunkt och därefter var *d*:te objekt (t.ex. var tionde objekt) i följden av objekt. Om urvalet är ett stratifierat systematiskt urval genomförs proceduren inom varje stratum. Då kan *d* ha olika värden i olika strata.

Hur objekten i ramen är ordnade (sorterade) är avgörande för hur effektivt det systematiska urvalet blir. Några extrema men belysande exempel kan ges. En undersökning går ut på att mäta människors arbetstid, och för det ändamålet dras ett urval av dagar och personer. För varje vald dag ställs frågan för ett urval av personer: Hur länge har du förvärvsarbetat i dag? Man kan tänka sig två olika metoder för att välja dagar:

Urvalsmetod 1: Valet av dagar görs systematiskt över årets dagar, med var sjunde dag som utvald.

Urvalsmetod 2: Årets dagar sorteras först slumpmässigt, och det systematiska urvalet dras från den slumpmässiga följden.

I urvalsmetod 1 kommer insamlade data att se mycket olika ut beroende på om urvalet består av årets alla söndagar eller måndagar. Steglängden för det systematiska urvalet sammanfaller på ett olyckligt sätt med en *periodicitet* i ramen. Systematiskt urval är här mindre effektivt än OSU. Urvalsmetod 2 däremot ger ett urval som är ekvivalent med ett OSU.

Betrakta nu ett annat exempel på användning av urvalsmetod 1. Anta att undersökningen gäller lufttemperaturen på en viss plats utanför tätorten. Då ger var sjunde dag, oavsett om det blir söndagar eller andra veckodagar, ett bra urval som är väl utspritt över året. I det fallet är det systematiska urvalet bättre än ett OSU. Urvalet ligger väl i linje med en trend i ramen. I det här fallet är det systematiska urvalet mer effektivt än ett OSU. Man säger att det systematiska urvalet ger en *implicit stratifiering*, dvs. resultatet blir ungefär detsamma som om man hade stratifierat årets dagar i 52 veckor och dragit en dag varje vecka.

Anta att man vill jämföra effektiviteten i en skattning av ett medelvärde, baserad antingen på ett systematiskt urval eller på ett OSU. Ett sätt att göra detta är att beräkna kvoten mellan variansen för en skattning baserad på ett systematiskt urval och variansen för motsvarande skattning baserad på ett OSU. Den resulterande kvoten är den s.k. designeffekten (se avsnitt 6.1.3); ju lägre värde den har, desto högre är effektiviteten hos det systematiska urvalet. I exemplet med arbetstid blir kvoten för urvalsmetod 1 ungefär $n/(1-n/N)$, eller n om urvalsfraktionen är liten. För urvalsmetod 2 blir den 1, dvs. lika effektiv som OSU. Slutligen, i exemplet med lufttemperatur och användning av urvalsmetod 1, blir designeffekten mindre än 1. Om ramobjekten sorteras efter löpnummer, och om undersökningsvariabeln växer linjärt med löpnumren, blir designeffekten så liten som $1/n$ (se Rosén 2005, s. 97–98). Även om det här bygger på extrema exempel framgår att ramens ordning kan ha avgörande betydelse för effektiviteten hos ett systematiskt urval.

Om det är möjligt att sortera ramen i någon ordning som ungefärligen avspeglar en trend i undersökningsvariabeln, bör man alltså överväga att göra det inför dragning av ett systematiskt urval. Risken är dock att en periodicitet skapas i stället för en trend. Om t.ex. ett individregister sorteras efter region, och inom region efter ålder, så skapas en sorts periodicitet i ramen, eftersom åldrarna upprepas från yngst till äldst i varje region. Anta att regionerna är ungefär lika stora, och att det systematiska urvalets steglängd sammanfaller med regionernas storlek (eller en multipel därav). För en viss startpunkt kan exempelvis 25- och 35-åringar bli överrepresenterade i urvalet.

Exemplen ovan visar också att om urvalsfraktionen är n/N, finns det bara N/n möjliga systematiska urval att dra. Det var sju stycken i exemplen där var sjunde dag valdes ($d = 7$). Ramen kan därför delas in i sju grupper som kan uppfattas som kluster (se avsnitt 4.6). Ett systematiskt urval är därför en sorts klusterurval där endast ett kluster dras.

4.3.2 Användningsområden

Historiskt har den systematiska urvalsmetoden varit mycket viktig. Med dagens datoriserade urvalsramar och dragningsmetoder som lätt utförs med dator, har systematiska urval tappat i betydelse. Systematiska urval är ändå fortfarande mycket användbara i situationer då alla objekt i populationen kan nås trots att det saknas en ram. (Den är även praktisk när urvalsramen bara finns i form av en lista papper.) Exempel på situationer där ram saknas är *rörliga* (*mobila*) *populationer* såsom passagerare som kliver av en färja eller vattnet i en bäck. Andra exempel är träd i en skog och hushåll i ett land som saknar såväl befolknings- som adressregister. Se avsnitt 3.2.3 för frågor om ramförfarandet samt exempel 4.3.2.

Om ramens ordning inte har något samband med den ordning som det systematiska urvalet följer, kan ett systematiskt urval jämställas med ett OSU (jämför urvalsmetod 2 i avsnitt 4.3.1). Man kan t.ex. dra ett systematiskt urval ur en ram som är ordnad efter någon egenskap hos ramobjekten som har svag eller ingen koppling till undersökningsvariabeln. Då ligger ett systematiskt urval nära ett regelrätt OSU, men är ändå väl utspritt över populationen.

Systematiskt urval kan även vara ett bra alternativ till stratifierat urval i vissa fall. Om ramen innehåller en hjälpvariabel som är starkt korrelerad med undersökningsvariabeln, kan ett effektivt förfarande vara att dra ett systematiskt urval efter att ramen sorterats efter hjälpvariabeln. Exempelvis är hälsorelaterade variabler oftast starkt korrelerade med ålder. För en undersökning av befolkningens hälsa kan man därför dra ett systematiskt urval ur befolkningsregistret sorterat efter personnummer, dvs. efter födelsedatum.

4.3.3 Fördelar

Möjligt och enkelt att dra, även vid avsaknad av ram
Systematiskt urval är som sagt användbart i situationer då en ram saknas, men kan ibland även vara effektivt när ram finns. Ibland måste man dessutom överlåta den praktiska urvalsdragningen på någon som inte nödvändigtvis är så statistiskt bevandrad. Det kan handla om ett flerstegsurval där uppgiftslämnaren själv får göra urval, eftersom en samlad ram över undersökningsobjekten saknas. Man kanske drar skolor och ber varje skola göra ett systematiskt urval av klasser som de rapporterar in uppgifter om. Se även exempel 4.3.3 som är hämtat från SCB. I en undersökning där data samlas in via direkt observation kan observatörer instrueras att göra systematiska urval på plats ur en population som passerar förbi (exempelvis observera var tionde passerande cyklist på en cykelväg).

Ger urval med spridning
Ett urval över en geografisk yta kan dras med ett systematiskt urval i två dimensioner (se exempel 4.8.1 och Cochran 1977, avsnitt 8.13).

Systematiskt urval kan användas för implicit stratifiering, se exempel 4.3.1. Metoden är att föredra framför vanlig stratifiering i situationer där stratumen riskerar att bli alldeles för små.

Kan vara mycket effektivt
Systematiskt urval kan vara effektivare än stratifiering, genom att det bättre kan utnyttja informationen i en hjälpvariabel utan stratumindelningens förgrovning. I exemplet med hälsovariabler så kan vissa av variablerna ha ett starkt ålderssamband även inom åldersstrata. Detta samband kan det systematiska men inte det stratifierade urvalet nyttja.

4.3.4 Nackdelar

Problem att skatta variansen
Om man drar ett systematiskt urval ur en population kan variansen inte skattas väntevärdesriktigt. I de fall man kan anta att det systematiska urvalet är effektivare än ett OSU, kan variansen skattas som om det vore ett OSU. Ett sådant skattningsförfarande ger en överskattning av den sanna variansen. Det finns även andra metoder att skatta variansen – se Wolter (1985, kapitel 7) och Rosén (2005, avsnitt 9.3). Metoderna bygger på vissa antaganden och ger i allmänhet inte väntevärdesriktiga variansskattningar. En tredje möjlighet är att dra systematiskt urval med flera startpunkter. Då övergår man i klusterurval och väntevärdesriktiga variansskattningar kan beräknas. Se vidare Särndal et al. (1992, s. 84). Nackdelen där är dock att variansskattningen får stor varians, genom att den grundas på ett urval som inte är större än antalet startpunkter.

Så även om systematiskt urval i vissa fall kan ge mycket låg estimatorvarians, kvarstår problemet att få ett mått på variansen.

Mindre effektivt urval i vissa fall
En nackdel är att effektiviteten i systematiska urval kan bli dålig i fallet med periodicitet. I slutet på avsnitt 4.3.1 gavs ett fiktivt exempel som visar att periodicitet kan vara svår att upptäcka. Risken för periodicitet bör dock ibland kunna bedömas utifrån saklogiska överväganden. Cochrans råd är att stratifiera relativt fint och dra systematiskt inom strata. Det ”hackar sönder” eventuella periodiciteter (Cochran, 1977, punkt 2 på s. 229).

4.3.5 Exempel
Nedan ges några exempel på undersökningar där systematiskt urval används:

EXEMPEL 4.3.1 *Partisympatiundersökningarna*
I Partisympatiundersökningarna (PSU) består urvalet av tre delar, kallade *paneler*, som väljs från RTB. Varje panel ingår i tre på varandra följande undersökningar: vid ett givet undersökningstillfälle är alltså en tredjedel av urvalet med för första gången, en tredjedel för andra gången och återstoden för tredje och sista gången. Inför dragningen av en panel sorteras ramen i "kameralordning", dvs. i tur och ordning efter län, kommun, församling och fastighetsnummer. Från den sorterade ramen dras ett systematiskt urval, som alltså blir implicit stratifierat efter variablerna som används i sorteringen. □

EXEMPEL 4.3.2 *Inkommande besökare i Sverige*
I undersökningen ”Inkommande besökare i Sverige”, IBIS, intervjuas utresande vid ett antal utreseställen (flygplatser m.m.) i Sverige. Urvalet görs i två steg med urval av tidsperioder i första steget och urval av passerande i andra steget. Intervjuarna drar ett systematiskt urval i flödet av passerande resenärer. I praktiken kan dock problem uppstå som gör att man tvingas tumma på vilka objekt som undersöks – exempelvis kan flödet ibland vara så tätt att det är svårt för intervjuaren att upprätthålla den förutbestämda steglängden. □

EXEMPEL 4.3.3 *Varuflöde*n
I Varuflödesundersökningen (se exempel 3.2.2) skattas totaler av transporterade varors vikt och värde. Objekten i målpopulationen är varusändningar. I urvalet dras arbetsställen, som rapporterar om ankommande och avgående sändningar. Urvalet görs i tre steg, med urval av arbetsställen och mätveckor i de två första stegen. I det sista steget drar uppgiftslämnaren ett stratifierat systematiskt urval av varusändningar. Stratumen är ”avgående, inom län”, ”avgående, utom län” och ”ankommande”. □

4.3.6 Läs mer
Rosén (1991) ger en djupgående och teknisk beskrivning av varför variansskattningen blir besvärlig i systematiskt πps-urval.

Kalton (1991) skriver lättläst om urval ur mobila populationer.

4.4 Urval med varierande inklusionssannolikheter

4.4.1 Beskrivning av metoden

För urvalsmetoder som OSU och systematiskt urval gäller att första ordningens inklusionssannolikheter är lika för alla objekt. Varierande inklusionssannolikheter kan exempelvis uppstå om man använder sig av stratifiering i kombination med OSU med olika urvalsfraktioner i olika strata. Inom ett stratum har då alla objekt samma inklusionssannolikhet. I detta avsnitt betraktas metoder där objekten tillåts ha olika inklusionssannolikheter även inom ett stratum. Om inklusionssannolikheterna väljs på ett speciellt sätt kan en urvalsdesign teoretiskt bli optimal i den meningen att variansen för en specificerad estimator blir så liten som möjligt. Variansen för HT-estimatorn minimeras då inklusionssannolikheterna väljs proportionella mot undersökningsvariabelns värden, se Särndal et al. (1992, s. 88). För GREG-estimatorn (se avsnitt 2.3.1) uppnås minimum i stället om inklusionssannolikheterna är proportionella mot modellresidualernas varians, se Särndal et al. (1992, s. 452).

För HT-estimatorn är det idealt med inklusionssannolikheter som är proportionella mot undersökningsvariabelns värden. I en verklig undersökning finns dock inte tillgång till undersökningsvariabeln när inklusionssannolikheterna ska bestämmas. För att bestämma lämpliga inklusionssannolikheter kan dock möjligheten finnas att i stället använda värden på en hjälpvariabel x_k som är korrelerad med undersökningsvariabeln. Detta ger inklusionssannolikheterna

$$\pi_k = n\frac{x_k}{\sum_{k\in U} x_k} \qquad (4.4.1)$$

där n är den förväntade (eller önskade) urvalsstorleken. Urvalsmetoder där inklusionssannolikheterna på detta sätt väljs proportionella mot objektens ”storlek” kallas *pps*-urval eller *πps*-urval. Här följs benämningarna i Särndal et al. (1992), där *pps* används som beteckning för urval med inklusionssannolikheter *proportionella mot storlek* dragna *med* återläggning och *πps* för urval med inklusionssannolikheter proportionella mot storlek dragna *utan* återläggning. Variansskattningar under *pps*-urval är normalt enklare än under *πps*-urval. I flerstegsurval, där variansskattningarna riskerar att bli komplicerade, är därför *pps*-urval av särskilt intresse, trots att urval med återläggning i praktiken är mindre relevanta. Ibland används sålunda variansskattningsformler under *pps* trots att urvalet dragits med *πps*. Detta ger normalt en överskattning av variansen. Om urvalsfraktionen är liten är dragning med och utan återläggning ungefär samma sak, eftersom risken då är liten att objekt dras mer än en gång. I sådana fall är *pps*- och *πps*-urval likvärdiga, med god approximation. I återstoden av detta avsnitt behandlas endast *πps*-urval.

Det finns egenskaper som varje urvalsmetod helst bör ha. För diskussionen här räcker det att titta på följande:

- Urvalsmetoden ska ge en fix (fast) och förutbestämd urvalsstorlek n. Förutom det opraktiska i att inte veta urvalsstorleken, så leder slumpmässig urvalsstorlek till sämre precision, speciellt då n är litet.

- Variansskattningen ska vara rimligt enkel.
- Det ska vara enkelt att dra urvalet.

Det är viktigt att samordna urval bland annat i SCB:s företagsundersökningar (se kapitel 5). För samordning behövs följande egenskap:

- Det ska vara möjligt att samordna urval, över tid och mellan undersökningar, med hjälp av permanenta slumptal.

Slutligen krävs följande av en *πps*-metod:

- Inklusionssannolikheterna ska vara kända och proportionella mot ”storlek” (se ovan), åtminstone med god approximation.

I litteraturen finns ett stort antal *πps*-metoder beskrivna. Många av dem är dock beräkningsmässigt komplicerade och svåra att använda i praktiken; det gäller metoder med fix urvalsstorlek större än två och inklusionssannolikheter strikt proportionella mot storlek. I det följande beskrivs tre metoder: Pareto *πps*-urval, systematiskt *πps*-urval och PoMix-urval. Se t.ex. Särndal et al. (1992, avsnitt 3.6) och avsnitt 4.4.5 för beskrivningar av flera tänkbara *πps*-metoder.

Pareto *πps*-urval
Pareto *πps* uppfyller alla fem egenskaperna ovan. Ett Pareto *πps*-urval genereras enligt följande (se t.ex. Rosén, 2000b):

Pareto *πps*
Låt x_k vara ett positivt storleksmått för objekt $k = 1, 2,\ldots, N$, och låt n vara urvalsstorleken.

Låt λ_k vara de önskade inklusionssannolikheterna

$$\lambda_k = n \cdot \frac{x_k}{\sum_{k \in U} x_k}. \qquad (4.4.2)$$

Låt $R_1, R_2, \ldots, R_N$ vara slumptal som utgörs av oberoende observationer från en likformig fördelning på intervallet [0,1].

Beräkna variablerna Q_k enligt

$$Q_k = \frac{R_k \cdot (1-\lambda_k)}{\lambda_k \cdot (1-R_k)}. \qquad (4.4.3)$$

Urvalet består av de n objekt som har de minsta Q_k-värdena.

De inklusionssannolikheter, π_k, som man faktiskt får med det här dragningsschemat är approximativt lika med de önskade inklusionssannolikheterna, dvs. $\pi_k \approx \lambda_k$. Approximationen är mycket god (Rosén, 2000c). Vid punktskattning används inte π_k utan λ_k för att beräkna designvikter. I variansskattningen används också λ_k, se formel (4.2.13) eller Rosén (2000b).

Sannolikheter måste definitionsmässigt ligga mellan 0 och 1. Vid beräkning av λ_k kan det dock inträffa att vissa λ_k blir större än 1 för vissa objekt: i formel (4.4.2) finns inget som hindrar att detta sker för vissa n eller x_k. Ett

sätt att hantera problemet om det uppstår är att totalundersöka de objekt som har $\lambda_k \geq 1$ och beräkna nya λ_k för övriga objekt enligt följande:

$$\lambda_k = (n - n_{tot}) \cdot \frac{x_k}{\sum_{k \in U} x_k - \sum_{k \in U_{tot}} x_k} \qquad (4.4.4)$$

där n_{tot} är antalet objekt som har $\lambda_k \geq 1$, n - n_{tot} är antalet som har $\lambda_k < 1$ och $\sum_{k \in U_{tot}} x_k$ är summan av storleksmåtten för de objekt som totalundersöks. Om det efter omräkning finns ytterligare objekt med $\lambda_k \geq 1$ upprepas proceduren tills alla objekt som ingår i urvalsdelen har $\lambda_k < 1$.

Om en del objekt har mycket små värden på storleksmåttet x_k, kan detta leda till orimligt stora designvikter. I exempel 4.4.1 löstes detta problem genom att införa cut-off-gränser, så att de minsta företagen inte ingick i urvalet. Andra sätt att undvika extrema vikter är att modifiera storleksmåttet, t.ex. genom att dra kvadratroten ur det, eller att använda PoMix-urval (se nedan). Man kan också lägga små objekt i ett stratum ur vilket t.ex. ett OSU dras. Ur andra strata, som har större spridning i objektens storlek, kan ett *πps*-urval dras. I valet mellan HT-estimatorn och GREG-estimatorn kan man välja den senare, som i allmänhet är mindre känslig för extrema vikter.

Pareto *πps* tillhör en familj av urvalsmetoder som kallas *order πps sampling*, se Rosén (1997a, 1997b eller 2000b). Olika val av variablerna Q_k ger olika urvalsmetoder. Urval dras på samma sätt som för Pareto *πps*, nämligen genom att beräkna Q_k-värden och låta de n objekt som har de minsta Q_k-värdena utgöra urvalet. Se vidare avsnitt 4.4.5.

Ett exempel på en annan order *πps*-metod som används i några av SCB:s undersökningar är sekventiellt Poissonurval, se Ohlsson (1990, 1998). I denna typ av urval används Q_k-värdena

$$Q_k = \frac{R_k}{\lambda_k} \qquad (4.4.5)$$

där R_k och λ_k definieras på samma sätt som för Pareto *πps*.

Systematiskt *πps*-urval

I avsnitt 4.3 beskrevs systematiskt urval med lika inklusionssannolikheter. En generalisering av denna metod är att tillåta olika inklusionssannolikheter för olika objekt. Väljs inklusionssannolikheterna proportionella mot storlek kallas detta systematiskt *πps*.

Ett systematiskt *πps*-urval genereras på följande sätt (framställningen nedan är hämtad från Rosén 2005):

Systematiskt *πps*
Låt 1, 2, …, N vara objektens ordning i urvalsramen.

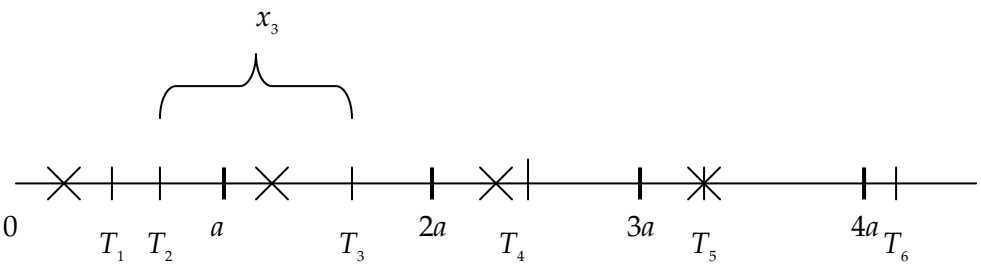
Beräkna de kumulerade storleksvärdena, $T_1 = x_1$, $T_2 = x_1 + x_2$, …, $T_N = x_1 + x_2 + \cdots + x_N = \sum_U x_k$.

Dela in intervallet $(0, T_N]$ i n delar av längden a $(= T_N / n)$.

Välj en startpunkt, b, slumpmässigt (med likformig fördelning) på intervallet $(0, a]$ och beräkna punkterna $b_1 = b$, $b_2 = b + a$, $b_3 = b + 2a$, …, $b_n = b + (n\text{-}1)\,a$.

Urvalet består av de objekt som är associerade till de storleksintervall som b-punkterna faller inom. Se figur 4.4.1.

Figur 4.4.1 Schematisk bild av punkter och intervall vid dragning av systematiskt *πps*-urval. De sex första objekten visas. Startpunkten *b* och de följande punkterna *b*+*a*, *b*+2*a* och *b*+3*a* är markerade med kryss. De utvalda objekten är de som har ordningsnummer 1, 3, 4 och 5



I beskrivningen ovan förutsätts att $n \cdot \left(x_k / \sum_{k \in U} x_k\right) < 1$. Om detta inte är uppfyllt kan flera punkter hamna i samma intervall, vilket leder till att samma objekt kan väljas mer än en gång. Om problemet uppstår används samma omräkningsförfarande som då λ_k blir större än 1 i Pareto *πps*-urval.

Systematiskt *πps*-urval är en beprövad urvalsmetod. I flera av SCB:s undersökningar har tidigare systematiskt *πps* använts som urvalsmetod, t.ex. i Gödselmedelsundersökningen, se exempel 4.4.2, men när Pareto *πps* utvecklades övergick man till den.

Systematiskt *πps* ger inklusionssannolikheter proportionella mot storlek och metoden är enkel att använda. Men precis som vid systematiskt urval med lika sannolikheter, se avsnitt 4.3, blir det problem med variansskattningen. En annan nackdel med systematiskt *πps* är att urvalssamordning via permanenta slumptal inte går att göra, se Rosén (2005, s. 122). Som nämndes ovan har Pareto *πps* inte dessa nackdelar.

En jämförelse mellan bland annat Pareto *πps* och systematiskt *πps* görs i Rosén (1997b), se även Rosén (2000b). Slutsatsen av den studien är att systematiskt *πps* med urvalsram ordnad efter storlek ger bättre precision än Pareto *πps* i vissa situationer, men att det inte går att säga vilken metod som är bäst eller vilken metod som bör väljas för en speciell undersökning.

PoMix-urval

Alla *πps*-metoder lider oundvikligen av risken att somliga inklusionssannolikheter blir mycket små. Med HT-estimatorn får man då mycket stora vikter. PoMix (Poisson Mixture) är en urvalsmetod som söker komma runt denna nackdel.

PoMix är liksom order *πps* en familj av olika urvalsmetoder. PoMix, med slumpmässig urvalsstorlek, beskrivs i Kröger et al. (1999). Av samma författare finns en artikel där idéer från order sampling används i urvalsdragningen, vilket bland annat leder till fix urvalsstorlek (Kröger et al., 2003). Det är den varianten som tas upp här.

PoMix

Låt x_k vara ett positivt storleksmått för objekt $k = 1, 2,\ldots, N$, låt n vara urvalsstorleken och f urvalsfraktionen n/N.

Låt λ_k vara definierad som i formel (4.4.2) och låt $\lambda_{\mathrm{mod},k}$ vara de önskade inklusionssannolikheterna

$$\lambda_{\mathrm{mod},k} = B + (1 - P)\lambda_k \qquad (\ 4.4.6\ )$$

där B är en parameter som uppfyller $0 \le B \le f$ och $P = B/f$.

Beräkna variablerna

$$\xi_k = \frac{R_k}{\lambda_{\mathrm{mod},k}} \qquad (\ 4.4.7\ )$$

där R_k är slumptal som utgörs av oberoende observationer från en likformig fördelning på intervallet [0,1].

Ett PoMix-urval med fix urvalsstorlek består av de n objekt som har de minsta ξ_k-värdena.

$\lambda_{\mathrm{mod},k}$ används för att beräkna designvikter i skattningen.

Sannolikheterna $\lambda_{\mathrm{mod},k}$ är som syns i (4.4.6) modifierade λ_k med en linjär transformation – härav indexet ”mod”. Om $B = 0$ förenklas (4.4.6) till (4.4.2).

Valet $B = f$ ger OSU. Som Holmberg (2003) har påpekat är spridningen av sannolikheterna $\lambda_{\mathrm{mod},k}$ störst då $B = 0$; ökar man B minskar spridningen. Vid det största värde B kan anta, B = f, är spridningen noll: vid OSU har alla objekt samma inklusionssannolikhet. Det den linjära transformationen med $B > 0$ gör, är att höja de små värdena på λ_k och minska de stora.

Genom att välja $B > 0$ kan man undvika extremt små värden på $\lambda_{\mathrm{mod},k}$ och därmed extremt stora vikter.

I Kröger et. al. (2003) ges även en annan variant, där man i stället för att använda ξ_k definierade enligt (4.4.7) använder variablerna

$$\xi_{2\,\mathrm{mod},k} = \frac{R_k \cdot \left(1 - \lambda_{\mathrm{mod},k}\right)}{\lambda_{\mathrm{mod},k} \cdot \left(1 - R_k\right)}. \qquad (\ 4.4.8\ )$$

Variablerna definieras alltså som i Pareto *πps*, se (4.4.3), men med de önskade inklusionssannolikheterna $\lambda_{\text{mod},k}$ enligt (4.4.6) i stället för (4.4.2). Kröger et. al. (2003) kallar detta för ”variant av Pareto *πps*” och på SCB kallas detta ibland för ParMix (”Pareto Mixture”).

För att kunna använda metoderna PoMix eller ParMix måste man välja ett värde på B. Den rekommendation som ges i Kröger et al. (2003) är att B bör väljas på intervallet $[0{,}2\,f,\ 0{,}5\,f]$. Rekommendationen är baserad på simuleringsstudier och är därmed beroende av dessas antaganden.

Samordning av urval med användande av de permanenta slumptalen R_k i formel (4.4.7) eller (4.4.8) går att göra.

För en given hjälpvariabel x är metoderna inte strikta *πps*-metoder. Men att först transformera x och sedan använda formel (4.4.2) är i princip detsamma som att först använda formel (4.4.2) på den ursprungliga hjälpvariabeln x och sedan transformera λ_k med hjälp av (4.4.6). Det går alltså att uppnå samma utjämnande effekt på λ_k genom att endera justera de små värdena på hjälpvariabeln eller att transformera λ_k med hjälp av (4.4.6). Den förra strategin används ofta inom SCB för *πps*-urval.

I undersökningen Företagens ekonomi, se exempel 4.4.1, övervägdes användning av PoMix-metoden, eftersom vikterna för små företag blev mycket stora. Genom att i stället införa en cut-off-gräns och ta bort de minsta företagen undveks problemet med stora vikter.

EXEMPEL 4.4.1 *Företagens ekonomi*
Företagens ekonomi, FEK, är en årlig företagsundersökning som är uppbyggd av fyra olika undersökningar med olika urvalsförfaranden. Som urvalsram används FDB kompletterat med data från administrativa källor.

De största företagen totalundersöks och får svara på en längre blankett. Små och medelstora företag urvalsundersöks och får en kortare blankett som förtrycks med uppgifter ur Standardiserade räkenskapsutdrag, SRU, från Skatteverket.

I en av undersökningarna samlas uppgifter in om intäkters fördelning på verksamheter och kostnaders fördelning på kostnadsslag. Urvalsramen stratifieras med avseende på branscher. Uppdelningen efter branscher är gjord för att möjliggöra redovisning och för att ta hänsyn till olika branschers intäkts- och kostnadsstruktur. Totalt används ca 170 branschstrata. Inom varje branschstratum används Pareto *πps* som urvalsmetod. Som hjälpvariabler i ramen finns företagets intäkter och kostnader enligt skattedeklarationen. Det storleksmått x_k som används är medelvärdet av dessa hjälpvariabler. Av bokföringsskäl brukar kostnader registreras som negativa värden, men här räknas de som positiva. De minsta företagen i varje branschstratum undersöks inte. Gränsen för cut-off bestäms på basis av storleksmåttet så att 90 procents täckning uppnås.

Urvalet dras i SAMU-systemet (se avsnitt 5.5) och samordnas över tid och med andra företagsundersökningar via användning av permanenta slumptal. Undersökningen beskrivs mer utförligt i avsnitt 7.4. □

EXEMPEL 4.4.2 *Gödselmedel i jordbruket*
Gödselmedelsundersökningen handlar om användningen av handelsgödsel och stallgödsel i jordbruket. Huvuddelen av undersökningen görs vartannat år. Målpopulationen för undersökningen 2007 utgörs av lantbruksföretag med mer än 400 standardtimmar (enligt typklassificeringssystemet för lantbruket) och med minst 5 hektar odling på åkermark eller stor djurbesättning. Lantbruksregistret 2006 används för att ta fram en urvalsram.

Totaler, medelvärden, andelar och förändringar av totaler av använd stallgödsel, handelsgödsel, växtnäringsämnen m.m. skattas. Redovisningsgrupper är län, produktionsområden, storleksgrupper (åkerareal), grödor, grödgrupper och djurslag.

Urvalet är ett tvåstegsurval med stratifierat Pareto πps av ca 3 700 lantbruksföretag i första steget. I andra steget väljs det största fältet av respektive gröda. Steg 2 återkommer vi till i exempel 4.9.2 i avsnitt 4.9.

Stratifieringen i steg 1 görs med avseende på produktionsområden och driftsinriktningar; totalt blir det 45 strata. Stratifieringen på produktionsområden avser redovisningsgrupper, och stratifieringen på driftsinriktning syftar till att homogenisera. Allokeringen av urvalet över strata görs proportionellt mot storleksmåttet för πps-urvalet.

Storleksmåttet som används för πps-urvalet är standardtimmar, vilket speglar företagets storlek med avseende på både grödarealer och djurhållning. För de 35 största företagen sätts inklusionssannolikheten till ett; dessa företag totalundersöks alltså. Eftersom antalet standardtimmar alltid överstiger 400, uppstår inga extremt små inklusionssannolikheter.

Urvalet samordnas med skördeundersökningen, se exempel 5.5.1, genom att samma slumptal används i urvalsdragningarna i båda undersökningarna. □

EXEMPEL 4.4.3 *Konsumentprisindex*
Före 1989 användes Poissonurval i KPI:s urval av butiker (för en allmän beskrivning av Poissonurval, se Särndal et al. 1992, avsnitt 3.6). Nackdelen med Poissonurval är att urvalsstorleken inte är förutbestämd. I det här fallet hade den slumpmässiga urvalsstorleken dessutom stor varians (se Ohlsson 1998). Detta var skälet till att urvalsmetoden byttes till sekventiellt Poisson. Som storleksmått används antal anställda plus ett, där den adderade ettan motsvarar ägarens arbetsinsats. □

EXEMPEL 4.4.4 *Skörd av spannmål, trindsäd och oljeväxter*
Genom undersökningen ”Skörd av spannmål, trindsäd och oljeväxter” (se exempel 3.4.3) insamlas årligen uppgifter om bärgade skördar. I målpopulationen ingår alla lantbruksföretag som brukar mer än 2 hektar åkermark och har odling av spannmål, trindsäd eller oljeväxter. Som ram används en delmängd av Statens jordbruksverks arealregister.

Urvalsstorleken har satts till ca 4 300 företag. Urvalsmetoden är stratifierat Pareto πps-urval, närmare bestämt ParMix-urval. Stratifieringen görs efter 101 s.k. skördeområden, som är avgränsade för att vara så homogena som möjligt med avseende på skördenivåer. Allokeringen beaktar arealerna av undersökningsgrödorna.

För πps-urvalet används ett storleksmått som tas fram genom att man väger samman de olika undersökningsgrödornas arealer. Vägningen görs så att

även ovanliga grödor ska bli representerade. Storleksmåttet varierar kraftigt, och man hade tidigare problem med mycket små inklusionssannolikheter, som ibland ledde till extrema bidrag till skattningarna av grödskördarna. Numera transformeras därför inklusionssannolikheterna enligt ParMix-metoden. ”Viddparametern” B sätts till 0,2 f, vilket betyder att man ”blandar in 20 procent OSU” i urvalsmetoden. Maximivärdena i ramen för designvikterna, dvs. inverserna av inklusionssannolikheterna, har på så vis kunnat sänkas från ca 400 till ca 120.

Positiv samordning görs med Gödselmedelsundersökningen – se exempel 5.5.1. □

4.4.2 Användningsområden

I avsnitt 4.2.5 under ”stratumgränser” beskrevs användning av kategoristrata i kombination med en finare indelning i storleksstrata. Ett skäl till ett sådant förfarande kan vara att kategoristrata motsvarar redovisningsgrupper, medan storleksstratifiering förväntas förbättra precisionen. Ett alternativ till storleksstratifiering inom kategoristrata är att använda någon πps-metod. Det främsta syftet är även nu att förbättra precisionen.

4.4.3 Fördelar

God precision
Den stora fördelen med πps-metoderna är att precisionen i skattningarna kan bli väsentligt bättre än med OSU.

Mindre arbetskrävande än storleksstratifiering
I exempel 4.4.1 används Pareto πps i varje branschstratum. Eftersom företagen varierar i storlek, skulle stratifiering med avseende på storlek i varje branschstratum vara nödvändig om en πps-metod inte hade valts. Att storleksstratifiera 170 branschstrata och att underhålla den stratifieringen kräver en hel del arbete. Om man gör en indelning i ett mycket stort antal strata, riskerar man vidare att en del strata blir för små.

Flerstegsurval
Vid flerstegsurval (se avsnitt 4.6) är πps-urval, med klusterstorleken som storleksmått, ofta lämpligt för att minska variansen från det första steget.

Självvägande i vissa fall
Det förekommer att populationsparametern som ska skattas är ett vägt medelvärde med vikter proportionella mot objektens storlek, enligt det i urvalsdragningen använda storleksmåttet. I detta fall reduceras HT-estimatorn till ett enkelt ovägt medelvärde, vilket innebär att urvalet blir självvägande med de fördelar detta ger (jämför avsnitt 4.6.4).

Exempel på ett sådant fall är konsumentprisindex (se exempel 4.4.3). Populationsparametern som ska skattas är en formel för prisindex, där observationer av priser i butiker ska vägas med vikter proportionella mot butikernas omsättning. Storleksmåttet i urvalsdragningen är butikens antal anställda plus ett. Med ett förenklat antagande att detta storleksmått är proportionellt mot omsättningen, så blir indexberäkningen självvägande. Den nämnda populationsparametern skattas då väntevärdesriktigt genom att indexformeln tillämpas på urvalets prisobservationer utan att dessa vägs.

4.4.4 Nackdelar

Kräver god hjälpinformation
Metoderna fordrar tillgång till tillförlitlig hjälpinformation. Om hjälpinformationen är av dålig kvalitet eller väljs på ett olämpligt sätt kan precisionen till och med bli sämre än för OSU.

Undersökningar med många variabler
Även om man finner ett storleksmått som samvarierar väl med en viss undersökningsvariabel, kan samvariationen med andra variabler vara svag. Då kan det vara lika bra att tillämpa ett stratifierat OSU i stället.

Kan ge extrema vikter
Risken finns att *πps*-metoder ger mycket små inklusionssannolikheter, som leder till extrema vikter i en HT-estimator. Olika sätt att hantera detta diskuterades i avsnitt 4.4.1.

Precision i redovisningsgrupper definierade av storlek
I undersökningen ”Företagens ekonomi”, exempel 4.4.1, finns problem med redovisning på storlek, eftersom urvalsstorleken i vissa storleksklasser inte blir tillräckligt stor.

Kan försvåra analys
Om man drar urval med varierande inklusionssannolikheter kan en statistisk analys av data bli mer komplicerad än vid OSU. Se vidare avsnitt 4.9.4.

4.4.5 Läs mer

Som redan nämnts i avsnitt 4.4.1 är Pareto *πps* ett specialfall av *order πps sampling*. Ett order *πps*-urval genereras på samma sätt som Pareto *πps*-urval, men variablerna Q definieras som

$$Q_k = \frac{H^{-1}(R_k)}{H^{-1}(\lambda_k)} \qquad (4.4.9)$$

där $H^{-1}(\cdot)$ är inversen till funktionen H. För att kunna dra ett urval måste H specificeras. Om $H(t) = t/(t+1),\ 0 \le t < \infty$, fås Pareto *πps*. Urvalsmetoden kallas så därför att $H(t) = t/(t+1),\ 0 \le t < \infty$, är fördelningsfunktionen för en standard-Paretofördelning.

Om H väljs som $H(t) = min(t, 1),\ 0 \le t < \infty$, fås

$$Q_k = \frac{R_k}{\lambda_k}, \qquad (4.4.10)$$

dvs. sekventiellt Poissonurval, se Ohlsson (1990, 1998). För olika val av H fås olika order *πps*-metoder.

Pareto *πps* är optimal bland order *πps*-metoderna i den meningen att den har den (asymptotiskt) minsta variansen.

Teorin för order sampling och Pareto *πps* introduceras i artiklarna Rosén (1997a, 1997b). En kortfattad beskrivning av metoderna och dess egenskaper ges i Rosén (2000b).

I Rosén och Lundqvist (1998) beskrivs hur bortfallskompensation kan göras då Pareto πps används. Några enkla bortfallsmodeller behandlas, och en sammanfattning av det arbetet finns även i Rosén (2000b).

Holmberg (2002) och Holmberg et al. (2002) skriver om optimala πps-urval för flera undersökningsvariabler.

I Särndal et al. (1992, avsnitt 3.6) finns en beskrivning av urval med varierande inklusionssannolikheter och speciellt πps-urval, men mycket har hänt inom området efter att boken kom ut, så framställningen där är något inaktuell.

Det finns i litteraturen en enormt stor mängd olika πps-urval föreslagna. Brewer och Haniff (1983) listar över fyrtio stycken. De allra flesta är nu omoderna. Thorburn (2006) diskuterar olika sätt att dra πps-urval och jämför med stratifierat OSU. Han förespråkar systematiskt πps-urval.

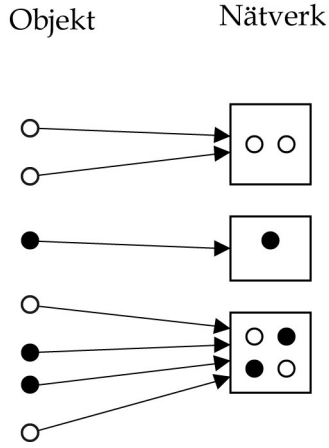
4.5 Nätverksurval

4.5.1 Beskrivning av metoden

När syftet är att undersöka grupper av objekt men det endast finns tillgång till en urvalsram över de enskilda objekten, kan nätverksurval användas. Exempelvis kan syftet vara att samla in uppgifter om hushåll, t.ex. hushållsinkomst eller bostadsyta per person inom hushållet. På SCB används nätverksurval främst för att ta fram hushållsstatistik. Ännu har SCB inget hushållsregister. Däremot finns RTB, varifrån urval av individer kan dras. Hushållen utgör grupper, eller *nätverk*, av individer, och varje individ tillhör precis ett nätverk, se figur 4.5.1.

Nätverksurval används i allmänhet i situationer då man saknar en relevant ram. Metoden används alltså för att lösa ett praktiskt problem.

Figur 4.5.1 Objekt och nätverk. En fylld ring representerar en utvald individ



Nätverksurval definieras på följande sätt. Dra ett urval av objekt med någon urvalsmetod, exempelvis OSU eller stratifierat OSU. För varje utvalt objekt identifieras vilket nätverk det tillhör. Detta nätverk betraktas som valt. Via urvalet av objekt fås på detta sätt ett urval av nätverk.

Nätverksurval kan ses som ett slags omvänt klusterurval; i nätverksurval dras enskilda objekt för att komma åt grupper av objekt, i klusterurval går man motsatt väg.

När man drar ett urval av individer och nätverken är hushåll, så kan flera individer i ett hushåll komma med i urvalet, vilket innebär att hushållet kommer med i urvalet flera gånger. I figur 4.5.1 är utvalda objekt markerade; ett av nätverken har replikerats i urvalet. Vid skattning kan man välja ett av två möjliga synsätt:

- Urvalet av nätverk omfattar endast de distinkta nätverk som dragits.
- Nätverk som dras flera gånger tillåts vara med flera gånger i urvalet.

I det första fallet är antalet utvalda nätverk slumpmässigt; i det andra är urvalsstorleken fix.

Nätverken varierar i regel i storlek, dvs. i antal objekt i nätverken. Därmed varierar även deras inklusionssannolikheter. Om objekten är individer och nätverken hushåll så har ett stort hushåll större sannolikhet att ingå i urvalet än ett litet hushåll. För skattning vid nätverksurval, se t.ex. Rosén (2005) eller Thompson (1992).

Det råder en viss begreppsförvirring i användningen av termen nätverksurval. I beskrivningen ovan ingår varje objekt i ett och endast ett nätverk. Det motsvarar fall 3, ”Många-till-ett”, i avsnitt 3.2.2. I litteraturen förekommer andra definitioner av nätverksurval där ett objekt tillåts ingå i mer än ett nätverk. Då hamnar man i fall 4, ”Många-till-många”, i samma avsnitt. Se Sirken (1998) för en överblick.

4.5.2 Exempel

Här ges ett exempel på nätverksurval från SCB, där hushållsstatistik ska redovisas. Eftersom ett hushållsregister inte finns tillgängligt, dras urval av individer. Utvalda individer tillfrågas om sammansättningen av sitt hushåll.

EXEMPEL 4.5.1 *Privatpersoners användning av datorer och Internet*
Rubricerade undersökning görs årligen; beskrivningen avser 2005. Huvudsyftet med undersökningen är att belysa tillgången till och användningen av informationsteknik bland privatpersoner i Sverige. Redovisning görs på kön, åldersgrupper samt utbildningsnivå. Även hushållens tillgång till t.ex. Internet är av intresse.

Målpopulationen är individer i åldern 16–74 år som är folkbokförda i Sverige, exklusive personer intagna på anstalt, långvarigt sjuka samt de som vistas utomlands.

Urvalet är en delmängd av urvalet till Arbetskraftsundersökningen, AKU. Urvalet till AKU dras som ett stratifierat urval av individer med OSU inom strata. RTB används för att ta fram urvalsramen.

För att möjliggöra skattning av hushållsparametrar samlas uppgifter om hushållens sammansättning in. Både individ- och hushållsstatistik redovisas. □

4.5.3 Läs mer

Utanför SCB används nätverksurval även till annat än hushållsstatistik, bland annat för att undersöka ovanliga företeelser. I en population av individer kan det t.ex. handla om en ovanlig sjukdom. Ett urval av individer dras; dessa svarar för sig själva och även för andra individer i sitt nätverk. Kalton och Anderson (1986) ger en lättläst översikt.

4.6 Klusterurval

4.6.1 Beskrivning av metoden

Tidigare avsnitt har mest handlat om direkturval, dvs. urval där man direkt drar de objekt man vill undersöka. I många situationer är det dock mer praktiskt att börja med att välja grupper av objekt, s.k. *kluster*. Läs om grupper av objekt även i avsnitt 4.2 (där i form av strata) och i avsnitt 4.7. Stratifierat urval och klusterurval har likheter, och det är därför inte helt ovanligt att de förväxlas. En enkel minnesregel för att skilja metoderna åt är att man gör urval *inom* strata men *av* kluster.

I grundläggande läroböcker i urvalsteori förutsätts ofta att klusterurval görs med OSU. Alla urvalsmetoder som behandlas i kapitel 4, och många flera metoder som det inte finns utrymme att ta upp i denna bok, kan dock användas för att dra kluster.

Ett alternativ till att totalundersöka utvalda kluster är att göra ett urval av objekt inom de utvalda klustren och begränsa datainsamlingen till dessa. Detta brukar kallas att man gör ett tvåstegsurval. Det finns inget i teorin som hindrar att man fortsätter att göra urval i tre eller flera steg. Med ett samlingsnamn kallas sådana urval för *flerstegsurval*. Hur många urvalssteg som det är lämpligt att använda i en enskild undersökning styrs framför allt av tillgång till ramar och av bedömningar om variabilitet inom kluster. Se vidare avsnitt 4.6.3.

4.6.2 Användningsområden

Många extra moment i urvalsmetoder används i syfte att förbättra precisionen i skattningar av olika populationsstorheter. Så är dock inte fallet med klusterurval – i själva verket innebär metoden i praktiken i allmänhet att precisionen försämras (jämfört med OSU av samma antal objekt). Klustren är ofta mer homogena än populationen i sin helhet. Att objekt inom samma kluster tenderar att vara mer lika varandra än objekt från skilda kluster, kallar man ibland för ”klustringseffekt”. Därtill har olika objekt vanligen olika sannolikhet att inkluderas i urvalet, vilket kan försvåra analys av urvalsdata. Analys diskuteras i avsnitt 4.9.4.

Varför använder man då klusterurval? Ja, oftast handlar det om att spara tid och pengar i samband med ramkonstruktion eller datainsamling:

- **Ramkonstruktion:** Ibland finns det ingen ram över alla objekt (observationsobjekt) i den population man vill undersöka. Däremot finns det en ram över grupper av objekt och det enklaste är då att använda den. Det kan vara skolor i en undersökning där man vill intervjua elever. Om det inte finns någon ram alls, utan man behöver skapa en, är det ofta enklare och billigare att skapa en ram över grupper än över enskilda objekt.

- **Datainsamling:** Det kan vara billigare och enklare att samla in data för objekt i utvalda kluster än för objekt valda helt slumpmässigt ur populationen. Det kan exempelvis handla om att minska resorna i samband med datainsamlingen.

Utöver ovanstående skäl kan det ibland finnas en efterfrågan på skattningar för enskilda kluster. Är det känt i förväg vilka kluster som är intressanta som redovisningsgrupper, dras de med sannolikheten ett. Det är heller inte ovanligt att det i efterhand framkommer intresse för särredovisningar av vissa slumpmässigt valda kluster. Eftersom urvalet inte utformats för särredovisning av slumpmässigt valda kluster i ett flerstegsurval kan precisionen bli dålig. Oavsett om klustren i urvalet totalundersöks eller om man använder sig av ett flerstegsförfarande så möjliggör klusterurvalet sådana skattningar.

På SCB är vi välförsedda med register, och det finns inte så ofta behov av att ta till klusterurval för att urvalsram saknas. En allt mindre andel av SCB:s data samlas in genom personliga besök, och härmed minskar betydelsen av intervjuarnas restider. Trots detta förekommer klusterurval på SCB – se följande exempel.

EXEMPEL 4.6.1 *Outhyrda lägenheter*
I skrivande stund finns inget register över alla landets hyreslägenheter. Därför tog man i Outhyrda bostadslägenheter i flerbostadshus (2005) vägen via ett urval av taxeringsenheter. (Urvalsundersökningen avser lägenheter i privatägda hus. Alla lägenheter i kommunala bostadsföretag undersöktes.) Ett register över taxeringsenheter finns nämligen, och varje taxeringsenhet kan betraktas som ett kluster av lägenheter. Data om lägenheterna samlades in genom kontakter med ägarna till utvalda taxeringsenheter. □

EXEMPEL 4.6.2 *Elevpanel*
Om man i Elevpanel 6 (2002) hade dragit ett slumpmässigt urval av elever från alla barn i rätt ålder, hade man behövt kontakta väldigt många barn (eller snarare deras föräldrar) runt om i landet. I stället använde man sig av ett tvåstegsförfarande med kommuner i steg ett och rektorsområden i steg två. Data om eleverna i utvalda rektorsområden samlades in via kontakter med de sistnämnda. Genom att göra ett klusterurval i stället för ett direkt elementurval blev datainsamlingen snabbare, enklare och billigare. □

EXEMPEL 4.6.3 *Varuflöden*
Det finns inget register över alla varusändningar till och från Sverige. Detta är huvudskälet till att man i Varuflödesundersökningen 2004–2005 (se exempel 3.2.2 och 4.3.3) först gjorde ett urval av arbetsställen, över vilka det *finns* en ram. De flesta arbetsställen tar emot och skickar ett mycket stort antal varusändningar under ett år. En möjlighet hade varit att undersöka alla utvalda arbetsställens varusändningar under året. Man valde dock i stället att för varje arbetsställe undersöka ett urval av varusändningar. Urvalet görs i tre steg: i steg 1 används ett stratifierat OSU av arbetsställen, i steg 2 dras 1–3 mätveckor inom kvartal, och i steg 3 genomförs ett stratifierat systematiskt urval i flödet av varusändningar. □

4.6.3 Klusterurval i ett eller flera steg?

När man fattar beslut om att använda klusterurval måste man samtidigt bestämma sig för om man ska dra klusterurvalet i ett eller flera steg. Några argument för att välja flerstegs klusterurval framför enstegs klusterurval (urval av kluster som totalundersöks) hittar vi i Kish (1965, s. 156):

- Först och främst kan det finnas naturliga kluster som är praktiska som urvalsobjekt men som är så stora att man inte har möjlighet att totalundersöka dem. Då kan man dra några stora kluster och sedan skapa mindre kluster inom dessa som är tillräckligt små för att kunna totalundersökas. I exempel 4.6.2 är det enkelt och naturligt att börja med att göra ett urval av kommuner. Det hade däremot varit besvärligt att samla in data om alla tredjeklassare i kommunerna i urvalet. Därför drog man i ett andra urvalssteg rektorsområden, som var lagom stora för att kunna totalundersökas.
- En annan aspekt är klustringseffekten, dvs. att objekten inom ett och samma kluster tenderar att likna varandra mer än objekten i populationen. Detta fenomen är orsaken till den precisionsförlust man riskerar att göra när man använder klusterurval i stället för direkt elementurval. Om man använder flerstegsurval kan man ha stora kluster i första steget. Detta dämpar förmodligen effekten av klustring, eftersom den tenderar att vara svagare i stora kluster. I exempel 4.6.3 kan varje arbetsställe ses som ett kluster av sina varusändningar. Man hade valet att antingen välja ett fåtal arbetsställen och totalundersöka dessa, eller välja flera arbetsställen och undersöka ett urval av deras varusändningar. När man valde det senare alternativet ökade man sannolikt precisionen i sina skattningar.
- Slutligen nämner Kish att det kan vara praktiskt besvärligt att totalundersöka ”kompakta” kluster, där observationer av olika objekt i klustret påverkas av varandra och får mätfel. Ett exempel är om klustren utgörs av hushållsmedlemmar. Det kan vara svårt att göra oberoende intervjuer av alla medlemmar i ett hushåll.

4.6.4 Faktorer att beakta

Vad kan man ytterligare behöva tänka på när man utformar ett klusterurval?

Val av kluster

Huvudargumentet för klusterurval är att det är enkelt och billigt. Detta bör också styra valet av kluster. Välj i första hand ”naturliga kluster”, för vilka en bra ram finns att tillgå. Om dessa kluster är för stora för att totalundersöka – överväg flerstegsurval. Om det finns flera klusterkandidater av lämplig storlek – välj dem som troligen är mest inbördes heterogena med avseende på den viktigaste undersökningsvariabeln.

Ofta är det praktiskt att göra ett tvåstegsurval där man i första urvalssteget väljer geografiska områden. Områdena betraktas då som kluster av människor, företag, taxeringsenheter, åkrar eller vad man annars ser som objekten i sin population. Inom varje kluster görs sedan ett urval av objekt som studeras. Exempel 4.6.2 visar en design med geografiska områden (kommuner) som urvalsenheter i första steget.

Självvägande urval
När man utformar ett klusterurval kan man vilja göra urvalet *självvägande*, dvs. så att alla objekt har samma sannolikhet att inkluderas i det slutliga urvalet. En självvägande design kan vara praktisk bland annat i planeringen av intervjuarbetet och i beräkningarna.

Nu ska visas hur ett självvägande tvåstegsurval kan dras. Framställningen, som baserar sig på Särndal et al. (1992, Remark 4.3.6), förutsätter att man i båda urvalsstegen använder sig av dragning utan återläggning.

Steg 1. Låt n_I beteckna önskad (förutbestämd) urvalsstorlek i steg 1. Anta att u_i är värdet på en variabel u för kluster i som beskriver hur stort klustret är, t.ex. summan av en hjälpvariabel för alla objekt i klustret. Summan av u_i -värdena för alla kluster i populationen betecknas t_u. Dra nu kluster med πps-urval (se avsnitt 4.4), dvs. med inklusionssannolikheterna

$$n_I \frac{u_i}{t_u}.$$

Steg 2. Dra sedan, från varje kluster i som valts ut i steg 1, ett OSU av objekt av storleken

$$n_i = \frac{N_i}{u_i}$$

där N_i är antalet objekt i kluster i. Inklusionssannolikheten i det här steget, betingat på att kluster i drogs i första steget, är $n_i / N_i = 1/u_i$.

Om man gör urvalet enligt steg 1–2 får alla objekt i urvalet samma inklusionssannolikhet, nämligen $(n_I u_i / t_u)(1/u_i) = n_I / t_u$. Urvalsstorlekarna inom kluster (n_i) kan variera. Man kan dock hitta specialfall där även n_i är lika. Anta att det för alla kluster i urvalet gäller att $u_i = kN_i$ där k är en konstant (anta med andra ord att u_i är proportionella mot N_i). Då är urvalsstorleken $n_i = 1/k$ för utvalda kluster.

Urvalsstorlekarna n_i kan bli för små eller för stora för att passa intervjuarbetet. Om så är fallet multipliceras alla u_i -värden med en lämplig konstant M så att $u_i' = Mu_i$ är den nya variabeln att användas i stället för u_i. M väljs så att urvalsstorlekarna $n_i = N_i / Mu_i$ blir av lämplig storlek. Konstanten M påverkar inte inklusionssannolikheterna, eftersom den i

$$n_I \frac{u_i}{t_u} = n_I \frac{Mu_i}{Mt_u}$$

kan förkortas bort.

Allokering över urvalssteg
Ett problem som skiljer flerstegsurval från urval i bara ett steg är att man behöver bestämma urvalsstorlekarna i varje urvalssteg; man behöver *allokera* urvalet över urvalsstegen. Det kan vara ett val mellan att dra många kluster i första steget och relativt få objekt inom valda kluster, eller

tvärtom. Om budgetrestriktioner eller andra restriktioner är svaga, är det bäst att välja många kluster och få objekt inom kluster. Det sprider ut urvalet bäst över populationen. (Utan restriktioner är det bäst att välja samtliga kluster. Då är det ett stratifierat urval; se minnesregeln i första stycket i avsnitt 4.6.1.)

4.6.5 Läs mer

Den ineffektivitet som klusterurval ofta har jämfört med OSU beror på att objekt inom samma kluster ofta liknar varandra. Intuitivt kan man förstå detta genom att tänka ungefär på följande sätt. Säg att ett objekt i målpopulationen observeras. Om nästa objekt som observeras är likt det första blir tillskottet av ny information måttligt. Ju mer homogena klustren är inbördes, desto flera kluster måste undersökas för att urvalet ska ge god precision. Särndal et al. (1992, avsnitt 3.4.3) inför och diskuterar begreppet homogenitetskoefficient i systematiska urval och i klusterurval. De tar upp det i ett avsnitt som handlar om systematiska urval, men ett systematiskt urval kan teoretiskt ses som ett klusterurval.

Att allokera det totala urvalet på ett optimalt sätt i flerstegsurval är inte enkelt. Under vissa förenklande antaganden finns uttryckliga lösningar till det matematiska optimeringsproblemet, men för att kunna använda dessa lösningar (formler) måste man ha stor förhandskunskap om olika populationsstorheter. I det allmänna fallet måste lösningarna beräknas numeriskt. Den som vill fördjupa sig i problemet hänvisas till Särndal et al. (1992, avsnitt 12.8) och Danielsson (1975, kapitel III).

4.7 Tvåfasurval

4.7.1 Beskrivning av metoden

Med *tvåfasurval* menas att man först drar ett förhållandevis stort urval ur populationen och samlar in basinformation, och sedan drar ett mindre urval ur det första större urvalet och samlar in mer detaljerad information. Informationen från den första fasen kan utnyttjas i designen av den andra fasens urval. Tvåfasurval är en generell metod som kan utformas på en mängd sinsemellan helt olika sätt. En detaljerad beskrivning av tvåfasurval finns i kapitel 9 i Särndal et al. (1992).

För att skatta parametrar utifrån ett tvåfasurval, kan den s.k. π^*-estimatorn utnyttjas, se Särndal et al. (1992, avsnitt 9.3).

I Cochrans (1977) och många andras framställning kallas tvåfasurval för *double sampling*. Cochran fokuserar på tvåfasurval där samma sorts objekt dras i båda faserna.

Särndal et al. (1992) behandlar tvåfasurval i ett vidare perspektiv. I deras bok är även tvåstegsurval ett specialfall av tvåfasurval. I ett tvåstegsurval dras först ett urval av kluster. Inom varje erhållet kluster, oberoende av vilka kluster som i övrigt erhållits, dras sedan ett urval av objekt som undersöks. I ett tvåstegsurval får inte utformningen av andra stegets urval utnyttja information som framkommit i första steget. Mer formellt ska följande egenskaper vara uppfyllda: *oberoende* i andra stegets urval och *invarians* i andra stegets urval. *Oberoende* innebär att urvalen i olika kluster dras oberoende av varandra. *Invarians* innebär att varje gång ett kluster

kommer med i första stegets urval så används exakt samma urvalsdesign i andrastegsurvalet. Detta betyder att urvalsdragningen i steg två inte får bero på vilka kluster som kom med i steg ett. Se vidare i Särndal et. al. (1992, avsnitt 4.3).

Flerfasurval – med urval i tre, fyra eller flera faser – är också tänkbara. Approximativt väntevärdesriktiga punktskattningar är då förhållandevis enkla att ta fram. Men tyvärr blir variansskattningarna vanligtvis mycket komplicerade; i nuläget saknas möjligheter att beräkna dem generellt. Ett allmänt gott råd är att fundera igenom variansformlerna innan urvalsdesignen bestäms. Det gäller i synnerhet vid flerfasurval.

Fallet då man drar flerfasurval med samma typ av objekt i varje fas och urvalsmetoden är OSU i varje fas är dock enkelt. Ett OSU som dras ur ett OSU, som i sin tur dras ur ett OSU, osv., kan betraktas som ett OSU i en fas. Varianserna kan då beräknas på vanligt sätt för OSU.

4.7.2 Användningsområden

Det finns flera skäl, av olika karaktär, till att använda tvåfasurval. I de exempel som tas upp här är objekten av samma typ i båda faserna.

Metoden är användbar när man saknar hjälpinformation i ramen. Informationen från ett stort förstafasurval kan då utnyttjas för stratifiering inför (det mindre) andrafasurvalet. I första fasen gör man en enkel undersökning med få frågor avseende möjliga stratifieringsvariabler. I andra fasen görs en mer fullständig undersökning, vilken avser de önskade undersökningsvariablerna. Stratumen inom förstafasurvalet, dvs. inför den andra fasens urval, används för att avgränsa redovisningsgrupper eller homogena grupper. Informationen från förstafasurvalet kan alternativt användas för att beräkna regressionsskattningar utifrån observationerna i andrafasurvalet och därmed höja precisionen i skattningarna.

Ett annat skäl till att utföra tvåfasurval är s.k. screening, dvs. att man vill filtrera ut vissa typer av objekt, t.ex. med en ovanlig egenskap, och sedan undersöka dessa noggrannare. För förstafasurvalet frågar man exempelvis om huruvida personen i fråga har fiskat under fjolåret. Bland dem som har fiskat dras sedan ett andrafasurval; sedan undersöks dessa individer med avseende på fångsternas storlek, utgifter för fisket e.d. (Om *alla* fiskare undersöks i en andra etapp, så är det inte ett tvåfasurval, utan ett vanligt enfasurval med en filterfråga om fiske och flera frågor för de individer som sållats ut.)

Tvåfasurval kan också bli aktuellt att överväga då det finns önskemål om att utnyttja uppgifter från en tidigare undersökning i en ny undersökning. Urvalet i den nya undersökningen dras då som ett underurval (andrafasurval) ur den tidigare undersökningens urval (förstafasurvalet). Urval i två faser leder dock till variation i två faser och därmed lägre precision i skattningarna. Man kan också få täckningsproblem, eftersom objekt kan ha upphört eller tillkommit sedan den tidigare undersökningen genomfördes. Därför bör man först kontrollera om det finns variabler som kan tas ur register i stället och som håller tillräcklig kvalitet för att brukas som bakgrundsvariabler eller liknande.

Det kan vara ogörligt att samla in detaljerad information för ett stort urval, beroende på att budgeten inte räcker eller att uppgiftslämnararbetet blir för stort. I denna situation kan tvåfasurval erbjuda en lösning. Den utförliga informationen samlas då in bara för ett underurval, som är mycket mindre än det ursprungliga urvalet.

Om kostnaden varierar mycket mellan olika frågor eller variabler i en undersökning, så kan tvåfasurval vara en bra metod. Exempelvis skulle man i en hälsoundersökning först kunna ställa bakgrundsfrågor om rökvanor e.d. i en postenkät, och sedan för ett underurval genomföra direkta mätningar av blodtryck och kolesterolnivå.

I de ovan beskrivna fallen (utom det sista) utförs huvudundersökningen med andrafasurvalet, medan förstafasurvalet ger värdefull extra information. Men huvudundersökningen kan i stället avse förstafasurvalet, medan andrafasurvalet nyttjas för evalvering e.d. enligt nedan.

Ett vanligt motiv till att nyttja tvåfasurval är nämligen att kunna genomföra en kvalitetsstudie på en del av urvalet. Det kan röra sig om en mätfelsstudie för direkt justering i skattningarna eller Hansen-Hurwitz bortfallsplan (se fallstudien i avsnitt 7.1). Vidare kan man dra s.k. interpenetrerande delurval, i syfte att belysa svarsvariationen mellan respektive ”inom” intervjuare. På så vis kan man också (approximativt) väntevärdesriktigt skatta de totala varianserna, inklusive urvals- *och* mätfelsvarians, se Särndal et al. (1992), avsnitt 16.10.

Begreppsmässigt kan de svarande i en urvalsundersökning i en fas ses som ett underurval ur det ursprungliga urvalet. Detta är alltså inte ett planerat, draget underurval, utan ett enligt en viss okänd svarsmekanism genererat underurval med ett slumpmässigt antal objekt. Teorin för tvåfasurval kan alltså användas när man räknar upp till populationsnivå i en urvalsundersökning med bortfall.

EXEMPEL 4.7.1 *Arbetskraftsundersökningar*
Utifrån det ordinarie urvalet i Arbetskraftsundersökningarna (AKU) får man via telefonintervju uppgifter om huruvida individerna har en viss egenskap, t.ex. att de är sysselsatta. Bland dessa dras sedan ett underurval. De som tillhör detta ”tilläggsurval” får sedan besvara ytterligare frågor i en särskild postenkät. Underurvalet dras med hjälp av panelurval. Samtliga sysselsatta som ingår i en rotationsgrupp i panelurvalet tillhör ”tilläggsurvalet”. Se även avsnitt 5.4.2 om panelurval och fallstudien om AKU i avsnitt 7.3! □

EXEMPEL 4.7.2 *Nyföretagande*
Ramen för Nyföretagarundersökningen utgörs av samtliga i Företagsdatabasen (FDB) nytillkomna juridiska enheter under ett år.

1) Varje månad skapas en ram av den senaste månadens nyregistrerade juridiska enheter. Alla enheter väljs ut och frågan ställs om enheten representerar nystartad verksamhet, till skillnad från överlåten verksamhet, byte av företagsform etc.
2) Efter det gångna året skapas en ram av de juridiska enheter som i månadsundersökningarna i 1) angett att enheten avser nystartad verksamhet. Ett stratifierat urval dras och ett stort antal andra frågor, som är ovidkommande i detta sammanhang ställs till företagen i urvalet.

3) Efter några år görs en uppföljningsundersökning genom ett underurval. Ramen utgörs av enheterna i urvalet i punkt 2).

Stratifieringen i uppföljningsundersökningen har ibland gjorts utifrån svar på andra frågor i den andra fasen, exempelvis om man fått starta eget-bidrag eller inte. □

4.7.3 Fördelar

Till fördelarna med tvåfasurval hör att man för undersökningar där man inte vet mycket om populationen, dvs. inte har någon hjälpinformation, kan höja precisionen avsevärt jämfört med OSU (genom att utnyttja förstafasurvalets data för stratifiering eller i skattningar) och att man kan allokera befintliga resurser effektivare.

4.7.4 Nackdelar

En nackdel med tvåfasurval är att undersökningen tar längre tid, eftersom resultaten från första fasen vanligen krävs för att kunna genomföra den andra fasen. En annan nackdel är att förstafasurvalet kan bli föråldrat, om populationen eller variabelvärdena ändras snabbt. Därtill blir kostnaden högre när data ska samlas in två gånger för en del av populationen. Vidare blir beräkningarna mer komplicerade. Precisionen sänks ofta på grund av att det förekommer variation i två faser.

4.7.5 Faktorer att beakta

Precisionen i skattningarna kan enligt ovan påverkas i båda riktningarna. Hjälpinformation från förstafasurvalet kan höja precisionen, medan urvalsvariationen i två faser sänker precisionen.

Bortfall i de två olika faserna kan leda till systematiska fel och sänkt precision. Ett ibland rimligt, men förenklat, angreppssätt är att betrakta de svarande i första fasen som ett OSU ur förstafasurvalet. Andrafasurvalet dras bland dessa svarande, och de svarande i andra fasen ses som ett OSU ur andrafasurvalet. Begreppsmässigt får man då egentligen ett fyrfasurval, där två av faserna utgörs av planerade urval och de övriga två beror på bortfallen, men beräkningsmässigt används tvåfasmetodik.

Vid stratifiering inför båda fasernas urval, är det lämpligt att stratifiera *inom* förstafasstrata inför andra fasen. Det går visserligen att låta de båda stratumindelningarna korsa varandra, men beräkningarna kompliceras i onödan.

Allokeringen – fördelningen av urvalsstorlekarna – mellan faserna i en tvåfasundersökning kan vara ganska komplicerad att ta fram. Om förhållandet mellan kostnaderna (för mätning och uppgiftslämnande) i andra respektive första fasen är högt, så har man inte råd med ett särskilt stort underurval. Om datainsamlingsmetoden är mycket dyrare i andra fasen, så väljs alltså underurvalsfraktionen förhållandevis liten. Det övergripande målet är ju att minimera variansen för en given kostnad, vilket naturligtvis försvåras av att man inte behöver specificera designen för den andra fasen förrän man sett resultaten från förstafasurvalet. Se Särndal et al. (1992), avsnitt 15.4.3, för allokering vid Hansen-Hurwitz bortfallsplan.

Tvåfasurval används för närvarande inte i särskilt stor utsträckning på SCB. Förekomst av bortfall i första fasen och täckningsproblem inför andra fasen försvårar användningen av tvåfasurval.

4.7.6 Läs mer

I Andersson (1994) utreds tvåfasurval i fallet med data som innehåller mätfel.

4.8 Urval ur kontinuerliga populationer

4.8.1 Beskrivning av metoden

Nästan all statistik som produceras av SCB avser *diskreta* populationer: deras objekt är uppräkneliga. Exempelvis består sådana populationer av individer, hushåll, företag, fastigheter eller organisationer. Denna handbok (och de flesta läroböcker i urvalsteori) handlar nästan uteslutande om urval från diskreta populationer. Det finns dock tillfällen då man vill genomföra urvalsundersökningar av *kontinuerliga* populationer:

- Man vill uppskatta andelen av markarealen i ett län som ett visst år är skogbevuxen. Populationen består av all mark i länet det aktuella året, och objekten i populationen är alla punkter på denna mark – ett oändligt antal.
- Man är intresserad av koncentrationen av kväve i ett vattendrag. Populationen består av allt vatten som passerar vattendraget under en given tidsperiod, och populationens objekt är alla punkter inom detta flöde – återigen ett oändligt antal.

Ramar för urval i kontinuerliga populationer (ramar för urval i ”rum och tid”) beskrivs närmare i avsnitt 3.2.3.

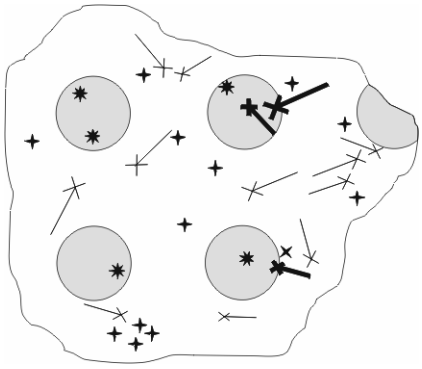
Flera exempel på naturliga kontinuerliga populationer är nederbörd och luft. I båda fallen kan exempelvis deras innehåll av vissa ämnen vara av intresse. Kontinuerliga populationer kan förstås också vara skapade av människohand. Inom trafikområdet görs undersökningar med syfte att skatta parametrar i stil med ”andel riksväg vars ytbeläggning är i ett visst skick”. Här består populationen av alla riksvägar och objekten är de oändligt många punkterna på dessa vägar.

För såväl diskreta som kontinuerliga populationer gäller att det i allmänhet är för dyrt och arbetskrävande att genomföra en totalundersökning. I stället samlar man in observationer för ett urval. När populationen är kontinuerlig brukar man ofta ”diskretisera” densamma, dvs. man delar in den kontinuerliga populationen i artificiella diskreta objekt. Man använder sig sedan av samma slags urvalsmetoder som för diskreta populationer:

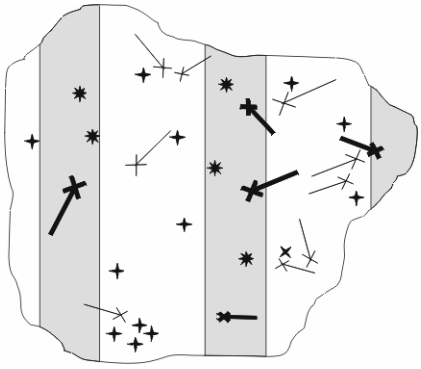
- Inom den skogliga statistiken används en rad olika ”objektiva” metoder för inventering av skogsmark (det finns också ”subjektiva” metoder som till stor del bygger på inventerarens bedömningar). I Ringvall et al. (2000) beskrivs exempelvis några olika metoder för att undersöka *lågor*, alltså liggande döda träd – se figur 4.8.1. Ett urval för undersökning av åkermark beskrivs i exempel 4.8.2.
- Om populationen består av vägar kan man dela in vägarna i delsträckor och undersöka ett urval av delsträckor. Denna ansats används i

Vägverkets urvalssystem för skattning av trafikarbete och årsmedeldygnstrafik för det statliga vägnätet. De statliga vägarna delas in i delsträckor eller *mätavsnitt* (totalt ca 20 000) som är så homogena som möjligt med avseende på trafikvolym. Varje avsnitt väljs ut för mätning med ett visst antal års mellanrum.

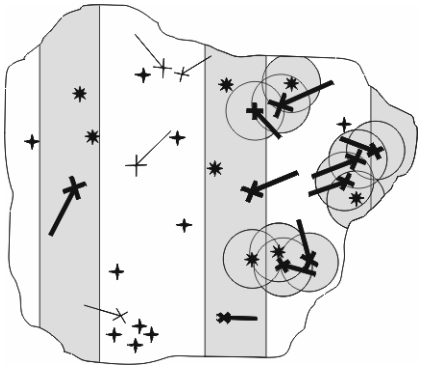
Figur 4.8.1 a–e Metoder för att inventera lågor. (Källa: Ringvall 2000; Ringvall et al 2000; figurer återgivna med tillstånd). Kors representerar stående, döda träd; stjärna träd som inventeras; ett streck är en låga där krysset är dess grövre del; ett grovt streck med kryss är en låga som inventerats



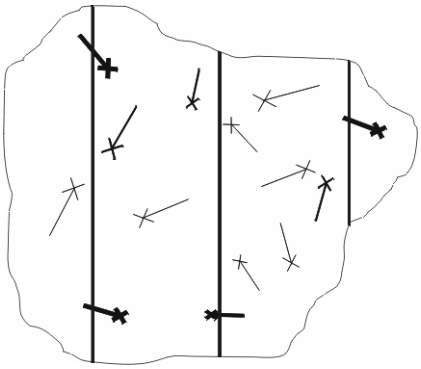
a) Cirkelinventering med systematiskt utlagda provytor. En låga inventeras då dess grövre del är belägen inom ytan



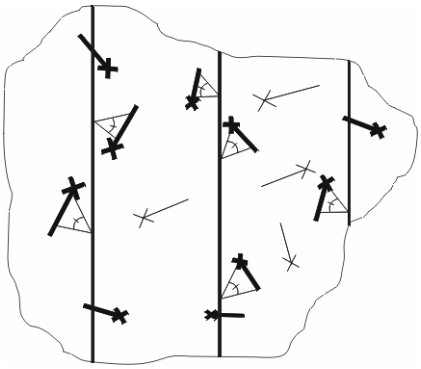
b) Bältesinventering med parallella bälten, utlagda med systematiskt avstånd. En låga mäts om dess grövre del är belägen inom bältet



c) Adaptiv klusterinventering. Då ett dött träd hittas inom bältet söker man vidare runt detta enligt en speciell procedur



d) Linjekorsningsinventering. De lågor som korsas av inventeringslinjerna räknas till urvalet



e) Inventering av lågor med relaskop från linjer. De lågor som fyller spaltöppningen på relaskopet räknas

I de flesta kontinuerliga populationer är närliggande objekt troligen korrelerade. Med anledning av detta vill man gärna försöka tvinga fram en spridning av sitt urval. Systematiskt urval är i detta sammanhang mycket användbart. Se även exempel 3.2.1. Det kan också vara nödvändigt att ta hänsyn till korrelationsstrukturer och eventuella trender i estimationen. Detaljer kring detta ligger utanför denna bok.

4.8.2 Exempel

Urval ur kontinuerliga populationer är inte så vanliga på SCB. De flesta exemplen finns inom miljö- och lantbruksområdet.

EXEMPEL 4.8.1 *LUCAS*
SCB deltog år 2001 och 2003 – tillsammans med en rad andra länder – i Eurostats pilotstudier av markanvändning: LUCAS (Land Use/Cover Area Frame Statistical Survey). Urvalet till LUCAS gjordes separat för varje land med ett enstegs klusterurval. Först drogs ett antal ekvidistanta koordinater med systematiskt urval. Punkterna bildade hörn i ett rutnät där rutorna var $18 \cdot 18$ km^2 stora. Runt varje utvald koordinat placerades tio nya koordinater i form av en rektangel. Runt var och en av dessa tio koordinater drogs en cirkel, tre meter i diameter, där markanvändningen observerades. Se vidare European communities (2003) och Bettio et al. (2002). □

EXEMPEL 4.8.2 *Jordprovsundersökningen*
Jordprovsundersökningen har genomförts av SCB 2001, 2003, 2005 och 2007, på uppdrag av Statens jordbruksverk. Data samlas in genom jordprover i ett urval av koordinatsatta punkter, belägna på jordbruksmark.

Urvalet gick till på följande sätt. Ett raster av punkter, med ett avstånd om tre kilometer mellan punkterna, lades ut över Sverige med systematiskt urval. Punkterna bildade alltså hörn i ett rutnät där rutorna var $3 \cdot 3$ km^2 stora.

De *block* (jordbruksfält med naturlig avgränsning) som visade sig innehålla en punkt identifierades. De punkter som tillhörde dessa block betraktades som valda i en första fas. I andra och sista fasen drogs ett stratifierat systematiskt urval av punkter från förstafasurvalet. För detaljer om urvalet, se Tångdahl (2006). □

4.8.3 Läs mer

Ett urvalsteoretiskt synsätt på undersökningar av kontinuerliga populationer inom miljöområdet presenteras i Brånvall et al. (odaterad).

Metoder för att undersöka sjöar och vattendrag behandlas utförligt på amerikanska miljöskyddsföreningens webbplats: www.epa.gov/nheerl/arm/.

4.9 Icke-sannolikhetsurval

4.9.1 Beskrivning av metoden

Med sannolikhetsurval menas ett urval där två villkor är uppfyllda:

1. varje objekt i rampopulationen har positiv sannolikhet att bli utvalt
2. för varje objekt i urvalet kan, åtminstone i princip, inklusionssannolikheten beräknas.

Det finns andra definitioner i litteraturen som är något annorlunda. Exempelvis Rosén (2005, s. 12) definierar det strängare.

I praktiken räcker det förstås inte att inklusionssannolikheterna är *möjliga* att beräkna, det måste vara rimligt lätt att göra det också.

Urval som inte uppfyller båda villkoren kallas för icke-sannolikhetsurval. Det är en rik flora av sinsemellan mycket olika metoder. I avsnitt 4.9.2 beskrivs närmare några av de vanligaste: *kvoturval*, *accesspaneler*, *bedömningsurval*, *deterministiska urval* samt *balanserade urval*. I 4.9.3 beskrivs några exempel på *modifierade sannolikhetsurval*.

Huvudmotiven till att använda sig av någon av de tre förstnämnda metoderna – kvoturval, accesspaneler eller bedömningsurval – brukar vara ungefär desamma: att samla in data snabbt, billigt och enkelt, och utan att behöva konstruera en urvalsram. Vinsterna är förföriska, och ska kontrasteras mot den förlust man lider av att frångå sannolikhetsurval: man mister möjligheten att utan modellantaganden generalisera sina resultat till hela populationen. Dessutom är det oklart hur osäkerheten i skattningarna ska beräknas. I praktiken leder detta till att statistik baserad på dessa typer av icke-sannolikhetsurval oftast redovisas utan vidhängande osäkerhetstal.

Balanserade urval är av helt annan karaktär. Metoden diskuteras kort nedan.

Ytterligare en form av icke-sannolikhetsurval är *cut-off-urval* (se t.ex. Särndal et al., 1992, avsnitt 14.4): en delmängd av populationen exkluderas avsiktligt från urvalsdragningen. Med definitionen ovan är cut-off-urval ett

icke-sannolikhetsurval, eftersom det första villkoret inte är uppfyllt. Alternativt kan man uppfatta cut-off-urval som sannolikhetsurval med avsiktlig undertäckning. SCB avstår i företagsundersökningar nästan alltid från att kontakta de små företagen, som i stället antingen försummas i beräkningen av skattningar eller tilldelas värden utifrån någon modell. Det finns flera anledningar till denna inskränkning av sannolikhetsurvalet. En av dem är att uppgiftslämnarkostnaderna är betungande för småföretag. Läs mer om cut-off-urval i avsnitt 6.2.4.

Huruvida sannolikhetsurval och icke-sannolikhetsurval är bra eller dåliga är en djup fråga som ibland kan bedömas olika bland metodstatistiker. Utöver den bättre vetenskapliga grunden för sannolikhetsurval finns ytterligare ett starkt argument som talar för sannolikhetsurval. Det är att den slumpmässiga processen garanterar att urvalet dragits opartiskt, utan godtyckliga överväganden om enskilda objekt. Detta är ett mycket viktigt argument för en producent av officiell statistik. Kvoturval, accesspaneler och bedömningsurval innehåller ett subjektivt moment av möjligt godtycke rörande enskilda objekt. Dessa metoder innebär därför en potentiell grund att ifrågasätta statistikproducentens oväld. Andra icke-slumpmässiga urval kan dras på ett objektivt, regelstyrt sätt.

Andra argument mot dessa typer av icke-sannolikhetsurval är de som nämndes ovan: svårigheten att generalisera resultaten till hela populationen och det faktum att det ofta är oklart hur osäkerheten i skattningarna ska beräknas. Emellertid har i praktiken även bortfall i sannolikhetsurval ungefär samma konsekvenser. Det skulle leda för långt att fördjupa sig i de här frågorna; se avsnitt 4.9.5 nedan.

En rekommendation är att icke-sannolikhetsurval ska undvikas i de flesta fall om

1. urvalet är huvudurval i undersökningen och inte bara del i en pilotundersökning e.d.
2. icke-slumpmässigt urval skulle innebära att valet av vilka objekt som ska komma med i urvalet blir beroende av subjektivt godtycke

   och
3. urvalet hade kunnat dras med ett sannolikhetsurval även om det inneburit större kostnader.

Man bör dock notera att det finns fall då sannolikhetsurval leder till dålig kvalitet. Det kan bli så t.ex. då ramen är otillräcklig.

4.9.2 Olika typer av icke-sannolikhetsurval

Kvoturval

Kvoturval är en metod med långa traditioner – för tidiga referenser, se exempelvis Stephan och McCarthy (1958, avsnitt 3.6) eller Kish (1965, avsnitt 13.7). I korthet ser metoden ut som följer. Man börjar med att identifiera viktiga delgrupper av populationen, och uppskatta deras storleksandelar eller *kvoter* i populationen. Uppskattningarna kanske baseras på andra undersökningar eller på registerdata. Därefter, i datainsamlingsfasen, rekryteras respondenter så att andelarna i urvalet tillhörande de olika delgrupperna överensstämmer rimligt med de skattade andelarna i populationen. Rent konkret rekryteras olika specificerade antal i exempelvis olika åldersgrupper. Eftersom bortfall ändå kan göra det svårt

att få exakt de avsedda proportionerna bland antalet svarande, förekommer i praktiken troligen att ambitionen ibland modifieras till att exempelvis få ett visst minsta antal svarande i varje grupp.

Ambitionen är förstås att åstadkomma ett urval som är så ”representativt” som möjligt för populationen som helhet, och representativiteten ska garanteras av att man ”kontrollerar för” vissa hjälpvariabler, vilket i sig är en god ambition. Men det blir inte ett slumpmässigt urval, eftersom inget av de två villkoren i föregående avsnitt är uppfyllt.

Kvoturval påminner i viss mån om stratifierat urval, och de kvoterade grupperna benämns ibland strata. Likheten är illusorisk, eftersom kvoturval saknar den vetenskapliga grunden för ett stratifierat sannolikhetsurval. En utförlig utredning av detta finns redan i Jerzy Neymans klassiska artikel från 1934 (Neyman, 1934).

Man skulle därför kunna tro att kvoturval i dag mest är av historiskt intresse. Faktum är dock att metoden har fått en renässans under senare decennier – framför allt inom marknadsundersökningsindustrin. Kvoturval används för undersökningar ”på stan”, utanför köpcentra och liknande. Även de moderna accesspanelerna (se nedan) kan under vissa förutsättningar betraktas som kvoturval. Förespråkare för kvoturval framhåller att man är garanterad svar från alla identifierade delgrupper, att man slipper kompensationsväga dessa svar i skattningarna och att metoden är billig och snabb.

I urval av varor för prismätning i konsumentprisindex gör bristen på ramar att man till vissa delar tvingas arbeta med något som liknar kvoturval (se vidare exempel 4.9.1). Internationellt kallas grupperna då ofta strata.

I slutet av 1990-talet tillfrågades Leslie Kish om hur han nu, trettio år efter publicerandet av sin bok, ser på kvoturval som metod. Följande utdrag kan sammanfatta hans svar: ”I am willing to make one broad generalization. When quota sampling is cheap (and fast) it is usually done poorly. When it is done better, it is not all that much cheaper *really* than efficient probability sampling” (Kish, 1998). Kvoturval analyseras teoretiskt av Deville (1991). Kvoturval används inte på SCB, utom möjligen under mycket speciella förhållanden.

Accesspaneler
Det har blivit allt vanligare, särskilt bland aktörer på den privata undersökningsmarknaden, att hålla sig med accesspaneler (även kallade accesspooler). En accesspanel består av en mängd personer som förklarat sig villiga att medverka i framtida undersökningar. Personerna kan bjudas in till panelen i samband med att de medverkar i en undersökning baserad på ett sannolikhetsurval, eller ha möjlighet att anmäla sig själva. Ibland talar man om de två fallen som *förrekryterade* respektive *självrekryterade* paneler (se exempelvis Stoop, 2004). Från panelen dras sedan urval för specifika undersökningsändamål, med eller utan sannolikhetsurval. Om man kontrollerar för vissa variabler i rekryteringen av panelen, eller i urvalet från densamma, åstadkommer man något som närmast liknar ett kvoturval. Accesspaneler av människor som speciellt är beredda att besvara elektroniska enkäter kallas *Internet-paneler* eller *webbpaneler*.

I USA är undersökningsföretaget Harris Interactive känt för sin stora självrekryterade webbpanel. Företaget har lanserat ett sätt att justera data

för att motverka urvalsskevhet – ”propensity score matching” – i en rad artiklar och konferenspresentationer (bland annat Bremer, Terhanian och Strange, 2004). I Sverige erbjuder företaget Synovate Temo sina kunder urval ur en förrekryterad webbpanel om ca 34 000 individer (Synovate Temo, 2006). Även om accesspaneler är vanligast inom den privata sektorn används de på sina håll även inom den officiella statistiken. Exempelvis använder sig den tyska statistiska centralbyrån av urval från en förrekryterad panel av hushåll – ”Haushalte Heute” – bland annat för en undersökning av inkomst- och levnadsförhållanden (Körner och Nimmergut, 2004; Nimmergut, 2005, 2006).

Generellt kan analytiska skevhetsjusteringar i efterhand inte anses motsvara sannolikhetsurval. Accesspaneler är dessutom särskilt problematiska, genom att de utöver ”vanlig” urvalsskevhet genom t.ex. självrekrytering också riskerar ytterligare skevhet genom s.k. utnötning eller erosion: att paneldeltagare med tiden faller ifrån av olika orsaker (se vidare avsnitt 5.2.2).

Bedömningsurval
Som namnet visar handlar bedömningsurval om att någon (kanske intervjuaren eller en expert) bestämmer vilka som ska ingå i urvalet genom att helt enkelt peka ut objekten. Det är vanligt i fokusgrupper och blankettester. Om urvalet i en urvalsundersökning dras med ett bedömningsurval, rör det sig troligen om en liten undersökning eller ett urval som sista steg i ett flerstegsurval, t.ex. ett bedömningsurval av en byxmodell i en klädbutik som i sin tur valts genom sannolikhetsurval.

I speciella fall kan det förekomma att ett bedömningsurval blir bättre än ett sannolikhetsurval. En sådan situation kan förekomma just vid urval av produkterbjudanden i en butik, där man följer prisförändringarna. Anta att man med bästa ambition drar ett sannolikhetsurval av produkterbjudanden i en butik för att följa priserna, och låter inklusionssannolikheten vara proportionell mot antalet hyllmeter e.d. som varan exponeras på. Då kan det lätt hända att det bland de produkterbjudanden man börjar följa blir en överrepresentation av kampanjvaror med tillfälligt nedsatt pris. Senare höjs priserna på dessa, vilket leder till bristande tillförlitlighet i måtten på prisförändringar.

Deterministiska urval
Icke-slumpmässiga urval kan vara regelstyrda. En sorts urval är att inkludera de n objekt som har störst värde på en hjälpvariabel som är korrelerad med undersökningsvariabeln. Det urvalet ger minst varians för en modellberoende kvotskattning om modellen är korrekt.

Balanserade urval
Ett sannolikhetsurval kan bli föga representativt genom ren otur. Det gäller särskilt mycket små urval och urval ur starkt heterogena populationer. En intuitivt tilltalande idé är att försöka tvinga fram ett representativt urval genom att åsidosätta kraven för sannolikhetsurval. Den idén ligger bakom kvoturval.

Balanserade urval liknar kvoturval i det att man kontrollerar för vissa variabler när man gör urvalet. Man strävar efter att åstadkomma ett urval som är sådant att när man beräknar kontrollvariablernas medelvärden i urvalet så sammanfaller dessa medelvärden ungefär med dem som gäller

för populationen. Ett sätt att åstadkomma detta är att dra många oberoende OSU och sedan välja det vars medelvärden mest liknar populationens. Detta enkla sätt räcker dock inte alltid. I avsnitt 4.9.5 finns referenser till djupare diskussioner av balanserade urval.

Man kan säga att balanserade urval är både en exakt specificering av begreppet ”representativt urval” och en objektiv metod att dra ett representativt urval. Förfarandet är på sätt och vis ett sannolikhetsurval (eftersom alla inklusionssannolikheter är positiva och kan beräknas genom simulering), men på sätt och vis inte (eftersom inklusionssannolikheterna knappast kan beräknas i praktiken). I det modellberoende angreppssätt som används i inferensen från balanserade urval anses dock inklusionssannolikheterna vara irrelevanta. Se vidare i avsnitt 4.9.4. Balanserade urval kan vara svåra att utforma och dra nytta av i fallet med många olikartade undersökningsvariabler.

4.9.3 Exempel

En urvalsdesign där sannolikhetsurval och icke-sannolikhetsurval kombineras kallas ibland för *modifierat sannolikhetsurval* (se exempelvis Statistics Canada, 2003, avsnitt 6.1.5). Det kan exempelvis röra sig om ett flerstegsurval där sannolikhetsurval används i alla steg utom det sista, i vilket man av praktiska skäl övergår till någon form av icke-sannolikhetsurval. Ett försvar för att det sista steget kan behandlas något styvmoderligt är att det ofta kan försummas i variansskattningen, se Särndal et al. (1992, avsnitt 4.3 och 4.6).

Modifierat sannolikhetsurval förekommer i flera fall på SCB – se exempel 4.9.1–4.9.3 nedan. I exempel 4.9.1 kombineras sannolikhetsurval med ett bedömningsurval. I exempel 4.9.2 respektive 4.9.3 görs icke-sannolikhetsurval i det sista steget; urvalet liknar cut-off-urval och riskerar att leda till systematiska fel.

EXEMPEL 4.9.1 *Konsumentprisindex*
I en av delundersökningarna inom Konsumentprisindex samlar intervjuare varje månad in prisuppgifter för ett urval av produkterbjudanden. Det gäller exempelvis insamlingen av priser inom varugruppen "verktyg, elartiklar och trädgårdsredskap". Målpopulationen utgörs av alla produkter som erbjuds till försäljning till privata konsumenter under mittveckan i månaden.

Urvalet är tvådimensionellt: i den ena dimensionen dras ett sannolikhetsurval av butiker, och i den andra dimensionen konstrueras ett antal "representantvaror" som får representera varugruppen. Representantvarorna är avgränsade genom att material, storlek, effekt m.m. specificeras med enstaka värden eller intervall. För kombinationer av valda butiker och definierade representantvaror väljer intervjuare i slutsteget vid besök ut ett eller ett antal produkterbjudanden, huvudsakligen enligt regeln "mest sålda", vars priser noteras. Urval av produkterbjudanden görs dels i december för alla nya butiker och representantvaror i urvalen, dels varje gång som ett tidigare produkterbjudande inte längre finns. □

EXEMPEL 4.9.2 *Gödselmedel i jordbruket*
I Gödselmedelsundersökningen görs ett tvåstegsurval, där steg 1 är ett stratifierat Pareto πps-urval (se exempel 4.4.2) av lantbruksföretag. I

steg 2 har, för varje utvalt företag och varje gröda (eller grödgrupp), gödselgivan på det största fältet observerats.

Eventuellt kan ett sannolikhetsurval åstadkommas genom att alla fält eller ett sannolikhetsurval av fält observeras. Det blir alltså en avvägning mellan den billiga, enkla metoden som riskerar att leda till ett systematiskt fel och den dyrare, kanske praktiskt svårgenomförbara metoden med sannolikhetsurval. □

EXEMPEL 4.9.3 *Båtliv*
I undersökningen ”Båtliv 2004” gjordes först ett stratifierat OSU av individer. För de utvalda individer i vars hushåll det visade sig finnas flera än k båtar tillämpades ett andra urvalssteg. För vissa undersökningsvariabler användes $k = 2$. Man valde då ut två av utvalda hushålls alla båtar. För andra undersökningsvariabler användes $k = 5$. I dessa fall observerades utvalda hushålls fem *största* båtar. □

4.9.4 Designbaserad och modellbaserad inferens

Olika synsätt
Frågan om huruvida det är bra eller dåligt med slumpmässiga respektive icke-slumpmässiga urval hör samman med vilken typ av resonemang statistiker applicerar på sina urvalsdata. Här ska skissas två vanliga huvudtyper av resonemang.

Urval från ändliga populationer skiljer sig från många andra områden inom statistisk teori och praktik i det att målpopulationen är konkret. Parametrarna som ska skattas är oftast knutna till populationen på ett konkret sätt, t.ex. medelvärdet av undersökningsvariabelns värde för alla objekt i målpopulationen. Dessa värden skulle i princip kunna erhållas med en totalundersökning. Man kan ha en föreställning av att värdena finns där och kan tas reda på med kontakt- och mätprocedurer, som – om de vore ideala – skulle ge felfria observationer. Med det synsättet är slumpen i urvalet den väsentliga källan till osäkerhet, medan variabelvärdena ses som givna konstanter. Variansen för en skattning $\hat{\theta}$ definieras som:

$$V\left(\hat{\theta}\right)=\sum_{s\in\mathcal{S}} p(s)\left\{\hat{\theta}(s)-E\left(\hat{\theta}\right)\right\}^2$$

där $p(s)$ är sannolikheten att dra urvalet s, $\mathcal{S}$ är den mängd av urval som är möjliga att dra och $E\left(\hat{\theta}\right)$ är väntevärdet för $\hat{\theta}$, se Särndal et al. (1992, s. 40). Denna definition leder fram till variansformler som är vanliga i produktion av officiell statistik. Förfarandet och sättet att tänka brukar kallas *designbaserad inferens*.

Jämför med en undersökning av skotillverkning där man vill veta vilka kombinationer av lim och sultyper som håller bäst. Den intressanta populationen består då inte av en ändlig hög av skor, utan den omfattar alla skor som man kommer att tillverka med befintligt produktionssätt. I det fallet är det rimligt att anpassa en modell som beskriver hållbarheten. Exempelvis kan man använda sig av en variansanalysmodell som beskriver vidhäftningsförmågan som en funktion av lim och sultyp. I den modellen är man intresserad av modellparametrarna, som visar hur mycket olika sorters lim, sulor och eventuellt deras samverkan betyder för

vidhäftningsförmågan. Hur urvalet drogs kommer aldrig in i resonemanget här annat än indirekt i modellformuleringen.

Men även vid urval ur ändliga populationer kan man anpassa en modell till urvalsdata och med den räkna ut predikterade värden för ej utvalda objekt. En totalskattning blir då helt enkelt summan av observerade och predikterade värden. Då känns det mest naturligt att den väsentliga källan till osäkerhet (återigen med ideala kontakt- och mätprocedurer) är själva modellen varmed de predikterade värdena beräknades. De icke-observerade värdena på undersökningsvariabeln ses nu som slumpvariabler. Skattningen blir beroende av vilken modell statistikern valt, och risken för systematiska fel blir stor. Skattningsproceduren brukar kallas *modellberoende inferens*. Risken för systematiska fel minskas väsentligt om urvalet har dragits balanserat (se avsnitt 4.9.2). Därför hör modellberoende inferens ihop med balanserade urval. Denna urvalsmetod hör inte ihop med designbaserad inferens, eftersom det är opraktiskt att räkna ut inklusionssannolikheterna (se dock den s.k. kubmetoden i avsnitt 4.9.5).

Om det är rätt att se modellen eller mängden av alla möjliga urval som källan till osäkerhet, har debatterats fram och tillbaka sedan början på 1970-talet. I ett ofta refererat exempel visade Hansen et al. (1983) på risken för bias med modellberoende estimation. Valliant et al. (2000, s. 85–90) tillbakavisar resonemanget. I avsnitt 1.5 i samma bok diskuterar de om slumpmässiga urval är något som behövs i urvalsundersökningar. Även Brewer (1994, 1999) och Brewer (2002, särskilt s. 18–28 och 77–78) funderar på om principen att dra slumpmässiga urval är helig, och i så fall, varför.

Fred Smith har skrivit många trevliga och resonerande artiklar om urvalsteorins fundament. Sviten Smith (1983, 1994, 1997, 2001) ger också en personlig utveckling av hans tänkande.

Särndal et al. (1992) kombinerar slumpmässiga urval med modelltänkande, men det är fortfarande designbaserad inferens. Särndal och Lundström (2005) använder hjälpinformation utan att stödja sig på explicita modeller.

Proceduren att dra ett slumpmässigt urval och låta det följas av designbaserad inferens leder ibland till en del logiska konstigheter. Holt och Smith (1979) visar att det med en poststratifierad estimator blir helt orimligt att i designbaserad inferens ta hänsyn till *alla* urval som är möjliga att dra. Slutsatserna måste begränsas till en delmängd av de möjliga urvalen. Smith (1983) påpekar att det för andra estimatorer är oklart vilken den delmängden är. Basu (1971) och Brewer (1999) visar att HT-estimatorn förutsätter ett representativt urval, annars kan skattningarna bli hur konstiga som helst. Hedlin et al. (2001) analyserar ett reellt exempel på ett fall där en designbaserad GREG-estimator bryter samman på grund av en dålig modell. Den annars ofta framförda poängen med designbaserad inferens är att den är okänslig för val av modell, se t.ex. Särndal et al. (1992, avsnitt 6.7).

Ett argument mot slumpmässiga urval är att bortfallet förstör de planerade inklusionssannolikheterna. Om det formella urvalsförfarandet följt av en svarsprocess tillsammans ses som urvalsmekanismen så är i stort sett alla SCB:s urval icke-slumpmässiga. I praktiken arbetar man i estimationen med antagandet att bortfallet är slumpmässigt givet hjälpvariablers värden. Detta antagande innebär helt säkert ett visst fel, men i många fall torde bortfallsgruppen inte vara extrem till sin karaktär.

Analys av data som kommer från en urvalsundersökning
Frågan om inklusionssannolikheterna ska tas med i beräkningarna eller inte, behöver besvaras även vid analys av data som kommer från en urvalsundersökning. Med ”analys” avser vi här främst statistiska beräkningar där intresset gäller modellparametrar snarare än parametrar i den ändliga populationen. Jämför exemplet med skotillverkningen ovan. Det är den synen på analys som bland andra Thomson (1997, avsnitt 6.1) har. Det motsvarar ”case 4” i Särndal et al. (1992, avsnitt 13.6). Om urvalet dragits med en metod som ger olika inklusionssannolikheter ger de, om de tas med i beräkningarna, olika vikt för olika observationer. Den främsta skälet till att i analys av surveydata införa vikter som beror av inklusionssannolikheterna är att sådana gör beräkningarna mer robusta mot mindre modellfel. Denna fråga har diskuterats ingående i litteraturen, se avsnitt 4.9.5 för referenser.

Om slutsatserna avser parametrar i den ändliga populationen tas vikterna med i en designbaserad ansats. Då används metoder som beskrivs i t.ex. Särndal et al. (1992, avsnitt 13.2-13.5). Situationen motsvarar ”case 1” i avsnitt 13.6 i samma bok.

En annan fråga vid analys av surveydata och en modellberoende ansats är om ett objekts värden kan antas vara statistiskt oberoende av andra objekts värden. (Vid en designbaserad ansats ses objektens variabelvärden som givna konstanter, inte som värden från slumpvariabler, och kan alltså inte vara beroende.) Om objekten är dragna med OSU från en stor population är det ofta lämpligt att anta att objektens värden är statistiskt oberoende, och vanliga statistiska metoder kan användas. Ett snedvridande bortfall kan dock göra analysen mer komplicerad. Vid klusterurval kan värdena inte antas vara oberoende, eftersom objekt inom samma kluster har en tendens att likna varandra. En ”klustereffekt” behöver då införas i analysen, t.ex. via en flernivåmodell. Analysmetoder ligger utanför ramen för denna handbok. Se vidare avsnitt 4.9.5.

4.9.5 Läs mer

Balanserade urval som följs av modellbaserad estimation beskrivs och diskuteras ingående i Valliant et al. (2000). Deville och Tillé (2004) presenterar en ny metod, *kubmetoden* (the cube method), varmed balanserade sannolikhetsurval med kända inklusionssannolikheter kan dras. Deville och Tillé (2004) ansluter sig till den designbaserade traditionen som är det vanliga sättet att se på inferens även på SCB. Man kan t.ex. dra ett πps-urval med fix urvalsstorlek med storleksmåttet x, och samtidigt få det ungefärligt balanserat på ytterligare hjälpvariabler. Ett utförligt exempel finns i Tillé (2006). Möjligheten att dra samordnade urval med kubmetoden verkar dock begränsad, se Tillé och Favre (2004). Chauvet och Tillé (2005 och 2006) presenterar en algoritm för dragningsförfarandet. Ett SAS-macro finns tillgängligt på INSEE:s webbplats, med dokumentation på franska (La macro SAS Cube d’échantillonnage équilibré).

Det finns en omfattade litteratur kring ämnet analys av data från urval med varierande inklusionssannolikheter. En introduktion finns i Lohr (1999, kapitel 10–11). Även Thompson (1997, kapitel 6) är kortfattad men mer teoretisk. Särndal et al. (1992, kapitel 13) diskuterar analys, främst från ett designbaserat perspektiv. Väsentligt mer omfattande är Chambers och Skinner (2003) och Skinner et al. (1989). Nordberg (1989) presenterar en ansats som är delvis modell- och delvis designbaserad.

5 Samordning av urval och urvalsramar

Kapitel 5 handlar om metoder för samordning av urval och urvalsramar. Med *samordning* menas att man organiserar urval och urvalsramar för ett gemensamt syfte. Det kan handla om att samordna versioner av samma urvalsundersökning över tiden, eller om att samordna olika urvalsundersökningar avseende (ungefär) samma population.

Syftet med samordning kan vara att sprida uppgiftslämnarbördan, få jämförbarhet mellan skattningar, uppnå säkrare skattningar av förändringar över tiden och samutnyttja vissa variabelvärden mellan undersökningar.

5.1 Viktiga begrepp

Om hela eller (stora) delar av ett urval behålls i ett annat urval, talar man om *positiv samordning*. Motsatsen är *negativ samordning* och innebär att man försöker få ett urval som har så få gemensamma objekt som möjligt med ett annat urval. De båda urvalen blir i dessa samordningsfall beroende, vilket påverkar varianserna vid förändringsskattningar.

När man diskuterar upprepade undersökningar, finns det behov av benämningar på en undersökning eller skattningar som bara avser *en* tidpunkt. En undersökning som belyser ett tidsutsnitt kallas för en *tvärsnittsundersökning* (engelska *cross-sectional survey*). Detta är den vanligaste formen av studie. Dess svaghet är att en sådan studie inte kan särskilja tidsrelaterade skillnader, såsom ålder, annat än genom att retrospektiva frågor ställs till respondenterna. Motsatsen är en längdsnittsstudie, eller som den vanligen kallas, en *longitudinell undersökning*. Här följer man ett antal objekt över tiden och gör upprepade mätningar på samma individ. På så vis kan man studera en företeelses utveckling under en längre tid. Termen används ibland som en allmän term, ibland för att beteckna endast panelundersökningar.

Med en *panelundersökning* menas en metod att med hjälp av upprepade mätningar av samma objekt studera stabilitet och förändring av olika variabler och parametrar. Panelen utgörs av ett urval ur en population. I en panelstudie görs en positiv urvalssamordning. Man kan ha fasta paneler eller roterande paneler; se vidare avsnitt 5.4.2. Panelundersökningen är alltså en longitudinell teknik. En annan sådan är *kohortundersökningen*, då man följer en grupp objekt med någon gemensam egenskap – en kohort. Ofta underförstås att den gemensamma egenskapen är ålder. I Elevpanelerna görs exempelvis urval av vissa årskurser av grundskoleelever, vilka följs genom skolsystemet och ut i arbetslivet.

Kohortundersökningar görs bland annat i medicinska sammanhang och kan handla om att studera långtidseffekter av expositioner för farliga ämnen. I praktiken önskar man i regel jämföra den exponerade kohorten med en icke exponerad sådan. Även i demografin talar man om kohorter,

exempelvis om en födelsekohort, som består av personer födda under samma kalenderperiod, t.ex. ett visst år.

Ett annat begrepp som förekommer vid samordning av urval är *basurval* (engelska *master sample*). Med detta avses ett statistiskt urval som man avser att använda vid flera tillfällen i framtiden. Då sparar man in förnyad urvalsdragning ur populationen och reducerar kostnader för mätningar, t.ex. besöksintervjuer. Ett basurval är ofta stort och används som ram för delurval eller urval i ett andra steg. Vidare avser ett basurval ofta geografiska områden, som ju förändras långsamt med tiden. SCB använde på 1950- och 1960-talen ett basurval för hushållsundersökningar. Basurval är vanliga i andra länder.

5.2 Fördelar och nackdelar med samordning över tid och mellan undersökningar

De flesta större undersökningar som görs på SCB är återkommande och producerar skattningar vid givna, återkommande tidpunkter.

Utöver skattningar av nivåer vill man i allmänhet också skatta förändringar mellan olika tidpunkter. Förändringar kan i princip vara endera *netto-* eller *bruttoförändringar*. I undersökningar av t.ex. sysselsättning och arbetslöshet är det värdefullt att kunna skatta hur många personer som rört sig i vardera riktningen mellan tillstånden sysselsatt och arbetslös, dvs. bruttoförändringen till skillnad från nettoförändringen. Man kan också se det som att man skattar flöden. Bruttoflödesskattningar i Arbetskraftsundersökningarna (AKU) visar t.ex. att under 2003 rörde sig 4,2 procent från att inte tillhöra arbetskraften till att bli arbetslösa, medan 18 procent rörde sig i andra riktningen. Skattningarna är genomsnitt av skattade kvartalsförändringar, se Statistiska centralbyrån (2005).

Att samordna urval innebär att urvalen är beroende. Dras urvalen oberoende kan överlappningen bli stor eller liten beroende på slumpen och på hur stora urvalen är i förhållande till populationen. I företagsundersökningar får man ofta stor överlappning även om urvalen är oberoende. Det beror på att inklusionssannolikheterna ofta är stora för stora och medelstora företag.

5.2.1 Fördelar

Vill man skatta bruttoförändringar är det nödvändigt att dra urval som *överlappar* varandra över tid, åtminstone i någon mån.

Om man endast vill skatta nettoförändringar är det inte alldeles nödvändigt att samordna återkommande undersökningar över tiden. Man har alltid möjlighet att helt enkelt dra nya, oberoende urval vid varje tidpunkt av intresse. Men förändringsskattningarnas varians blir i allmänhet större, ibland betydligt större, vid oberoende urval än om skattningarna baseras på beroende, överlappande urval. Ett undantag är om objektens förändring eller brist på förändring är oberoende av ursprungstillståndet, vilket sällan eller aldrig är fallet i praktiken. Givet att, exempelvis, en person var sysselsatt vid första intervjun, så är sannolikheten hög att hon eller han är sysselsatt även vid nästa intervju tre månader senare.

Vid positiv samordning finns vanligen en positiv korrelation mellan nivåskattningarna vid två tidpunkter. Därmed blir variansen för förändringsskattningen mindre än vid oberoende urval.

En annan fördel med överlappande urval är att informationen från det urval som dras först kan användas vid granskningen av data i det andra urvalet.

5.2.2 Nackdelar

Ett problem med att dra urval som överlappar över tid är att de objekt som bibehålls mellan undersökningsomgångar utsätts för en större uppgiftslämnarbörda än vad som vore fallet om de skulle medverka vid endast ett tillfälle. Detta riskerar i sin tur att successivt minska viljan att medverka i undersökningen, och bortfallet tenderar att öka med panelens ålder. Bortfallet kallas ibland för *panelutnötning* (engelska *panel attrition*).

Ett annat problem belyses av studier som visar att antalet gånger som ett objekt ingår i en undersökning kan påverka hur uppgiftslämnarna för dessa objekt svarar på vissa frågor. Detta fenomen kallas för *paneleffekt* (engelska *panel conditioning*). I en intervjuundersökning av individer kan variationen i svaren bero på att minnet av tidigare intervjuer påverkar hur individen svarar, eller på att intervjuerna inte genomförs på samma sätt vid olika tillfällen. Man riskerar att introducera ett systematiskt mätfel (en mätfelsbias) i skattningarna. Se avsnitt 5.2.4 för artiklar som beskriver mätfelsbias i panelundersökningar.

5.2.3 Samordning mellan undersökningar

Argumenten för och emot samordning *mellan undersökningar* liknar dem vid samordning över tid inom en upprepad undersökning. Negativ samordning sprider uppgiftslämnarbördan och minskar därmed belastningen på vissa, enskilda uppgiftslämnare. Positiv samordning förbättrar möjligheten till skattningar som kräver data ur alla de undersökningar som är positivt samordnade.

5.2.4 Läs mer

Lättlästa översiktsartiklar om urvalssamordning i tiden är Duncan och Kalton (1987), Kalton och Citro (1993) samt Binder (1998).

Bortfall och mätfel i panelundersökningar behandlas i Kalton et al. (1989) och O'Muircheartaigh (1989), mätfelsbias även i Bailar (1975).

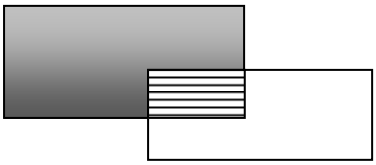
Holt och Skinner (1989) innehåller en introduktion om skattning av bruttoförändringar. Bassi et al. (1998) ger estimatorer för bruttoförändringar baserade på data som antas ha vissa mätfel.

5.3 Målpopulation under förändring

Att definiera målpopulationen är mer komplicerat vid skattning över tid än vid skattning av parametrar som avser en enda tidpunkt. Ett typfall är en parameter som inbegriper två tidpunkter, t.ex. förändringen i totalen från tidpunkt t till tidpunkt $t + 1$. I figur 5.3.1 visas målpopulationerna vid t och $t + 1$ samt deras snittmängd. Om man skattar nettoförändringen är det vanliga att man först skattar en total eller ett medelvärde för målpopula-

tionen vid *t* och motsvarande för målpopulationen vid *t* + 1. Sedan beräknar man differensen eller kvoten.

Figur 5.3.1 Populationer vid tidpunkt *t* (grå plus streckad) och *t* +1 (vit plus streckad). Snittmängden utgörs av de objekt som existerar vid båda tidpunkterna



5.3.1 En nettoförändrings beståndsdelar

I detta avsnitt visas hur en nettoförändring kan brytas ned i beståndsdelar: förändring för konstant population och förändring som beror av populationsförändring. Stoffet är hämtat från Holt och Skinner (1989). Anta att *D* redovisningsgrupper är intressanta, t.ex. *D* regioner. De får i formlerna index $d = 1, 2, \ldots, D$. Då kan man uttrycka en nettoförändring i medelvärdet i populationen mellan tidpunkt *t* (index 1 nedan) och tidpunkt *t*+ 1 (index 2) som

$$\begin{aligned}\bar{y}_2 - \bar{y}_1 = &\sum_d W_{1d}\left(\bar{y}_{2d} - \bar{y}_{1d}\right) \\ &+ \sum_d \left(W_{2d} - W_{1d}\right)\bar{y}_{1d} \\ &+ \sum_d \left(W_{2d} - W_{1d}\right)\left(\bar{y}_{2d} - \bar{y}_{1d}\right) \qquad (5.3.1)\end{aligned}$$

där $W_{1d} = N_{1d}/N_1$ och $\bar{y}_{1d}$ är den andel och det medelvärde som hör till redovisningsgrupp *d* vid tidpunkt 1. $W_{2d} = N_{2d}/N_2$ och $\bar{y}_{2d}$ är samma storheter, fast vid tidpunkt 2. Lägg märke till att formeln endast innehåller populationsparametrar. Våra fullständiga beteckningar hade egentligen krävt *U* i index för $\bar{y}_1$ och $\bar{y}_2$ men det hade blivit omständligt. Formeln har inget att göra med hur förändring skattas, vilket är ett annat problem. Holt och Skinner (1989) visar hur de olika termerna kan skattas. Den första termen uttrycker förändringen för konstant population. Den gäller alltså en population där inga har fötts eller dött och som dessutom har samma objekt i alla redovisningsgrupper. Den andra termen står för den förändring som enbart beror på att populationen förändrats, antingen om objekt har lämnat eller tillkommit, eller om de har bytt redovisningsgrupp. Den tredje termen kan ses som en samspelseffekt (interaktion).

Formeln ovan är egentligen bara en omskrivning av

$$\bar{y}_2 - \bar{y}_1 = \sum_d W_{2d}\bar{y}_{2d} - \sum_d W_{1d}\bar{y}_{1d},$$

där ”onödiga” termer som tar ut varandra lagts till i högerledet. För att få en formel för differensen mellan två totaler, skriv om ovanstående till

$$N_2\bar{y}_2 - N_1\bar{y}_1 = \sum_d N_{2d}\bar{y}_{2d} - \sum_d N_{1d}\bar{y}_{1d}$$

Den ”utvecklade” men ekvivalenta formeln för differensen mellan två totaler blir

$$t_2 - t_1 = \sum_d N_{1d}\left(\bar{y}_{2d} - \bar{y}_{1d}\right) + \sum_d \left(N_{2d} - N_{1d}\right)\bar{y}_{1d} + \sum_d \left(N_{2d} - N_{1d}\right)\left(\bar{y}_{2d} - \bar{y}_{1d}\right) \qquad (5.3.2)$$

där $t_1 = N_1\bar{y}_1$ och $t_2 = N_2\bar{y}_2$.

Holt och Skinner (1989) ger ett exempel på dödsfall i prostatacancer för män i åldern 45–75 år. Antalet dödsfall i England och Wales steg från perioden 1950–54 till perioden 1980–84 med 3876 fall. Denna siffra sönderdelas i 2201 för term 1, 1322 för term 2 och slutligen 353 för term 3. En stor del av förändringen beror alltså på att det under 1980-talet fanns flera män i målgruppen än under 1950-talet.

Formlerna ovan är viktiga eftersom de inte bara hjälper en att förstå och presentera en förändrings beståndsdelar, utan även att förstå vilka beståndsdelar som är intressanta och därmed hur målpopulationen bör definieras. Är man främst intresserad av den första termen i (5.3.1) eller (5.3.2) är det den konstanta målpopulationen. Då följer man en panel av objekt över tid. Man drar ett urval, en panel, vid den första tidpunkten, observerar objekten (med enkät eller annan metod) och observerar dem igen vid den andra tidpunkten. Är man intresserad av summan av de tre termerna skulle man kunna dra två oberoende urval, men det finns metoder som är mer effektiva, se vidare avsnitt 5.4.

Att bara skatta förändring som baseras på de objekt som finns vid båda tidpunkterna och som dessutom finns i samma redovisningsgrupp missar förändringar som beror på förändringar i populationen. Det finns ändå situationer då det är lämpligt att fokusera på förändringar i konstant population. Den kan t.ex. vara bra om man är mer intresserad av en panel av objekt än hela populationen. Anta att man utvärderar en metod att minska uppgiftslämnarbördan genom att tillämpa den på 100 objekt för att se om deras börda minskar. Ett sådant försök säger en del om hur bra metoden fungerar, men säger föga om huruvida bördan skulle minska i populationen om metoden applicerades på alla objekt.

Det vanliga sättet att se på en nettoförändring torde vara att skattningen vid tidpunkt *t* avser alla objekt som fanns då och som hörde till målpopulationen, dvs. det gråa och det streckade fältet i figur 5.3.1, och motsvarande vid tidpunkt *t* + 1, då skattningen avser det streckade och det vita fältet. Notera att det synsättet medför att det inte finns något enkelt samband mellan den ”totala” nettoförändringen $\bar{y}_2 - \bar{y}_1$, och nettoförändringarna inom redovisningsgrupperna, alltså $\bar{y}_{2d} - \bar{y}_{1d}$ för grupperna $d = 1, 2, \ldots, D$. Sambandet beror på såväl vikterna W_{1d} och W_{2d} som förändringarna i medelvärdena. Se vidare i avsnitt 5.3.3 för ett par referenser.

5.3.2 Definition av målpopulation som ändras över tid

Föregående stycke visar att det inte är självklart vilka målpopulationer som är intressanta vid skattningar över tid. Det allra enklaste sättet att skatta bruttoflöden är att dela in målpopulationen vid tidpunkt t i de intressanta klasserna och motsvarande för målpopulationen vid tidpunkt $t + 1$. Exempelvis kan man vara intresserad av klasserna arbetslösa, sysselsatta och utanför arbetskraften. Sedan skattar man andelen som hör till vardera av de 3x3 klasserna. Om urvalet inte är en fast panel är det dock inte så enkelt. Se avsnitt 5.2.4 för referenser som handlar om bruttoskattningar.

För att skatta bruttoflöden räcker det således inte att bara definiera en målpopulation för tidpunkt t och en annan för tidpunkt $t + 1$. Man måste även ta ställning till hur de objekt som fanns vid t ska räknas vid $t + 1$. Se Folsom et al. (1989) för olika sätt att definiera en longitudinell population.

5.3.3 Läs mer

Populationen är inte nödvändigtvis konstant bara för att den består att de objekt som finns vid båda tidpunkterna. För det krävs även att de stannar kvar inom samma redovisningsgrupp. Om de inte gör det, kan en effekt uppstå som kallas Simpsons paradox. Även om förändringen $\bar{y}_2 - \bar{y}_1$ på populationsnivå är positiv kan ändå förändringen *inom varje redovisningsgrupp* vara negativ. Wainer (1986) ger ett exempel och en lättläst förklaring av hur Simpsons paradox uppstår. Paik (1985) ger en grafisk beskrivning av Simpsons paradox.

5.4 Vilka metoder finns för urvalssamordning?

5.4.1 Oberoende urval

För att skatta en förändring över tid dras ett urval vid den första tidpunkten och ett ytterligare urval vid den andra tidpunkten. Om urvalen dras oberoende av varandra finns ingen kontroll över den del som finns i båda urvalen (snittmängden). Förändringen kan skattas ändå, men man får i regel en sämre precision.

Den främsta fördelen med att göra oberoende urval är enkelheten: standardmetoder kan användas för både urval och skattningar.

5.4.2 Panelurval

Vad är panelurval?

Termen panelurval används med något olika betydelser. Alla går ut på att samma objekt eller delvis samma objekt observeras mer än en gång. I strikt mening undersöks samma objekt i panelurval. Om man byter ut en del av urvalet från den ena tidpunkten till den andra talar man i mer strikt mening om *roterande urval* eller *roterande panelurval*. Men ofta kallas alla varianter för *panelurval*. För att framhäva att objekten inte byts ut använder man ibland termerna *fast panel* eller *fix panel*.

Om målpopulationen förändras med tiden, bör urvalet förnyas vid varje ny urvalsdragning. Det främsta skälet är att göra det möjligt att dra slutsatser om den del i målpopulationen som inte fanns vid tidigare tillfällen. Det finns även andra skäl som har med mätfel att göra (se avsnitt 5.2.2). Om

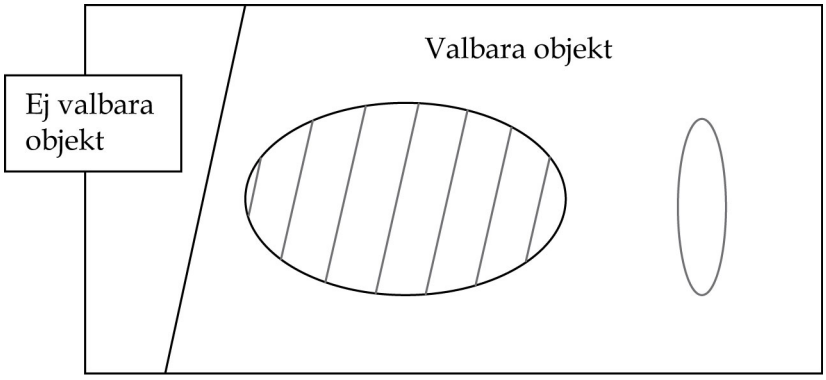
det är opraktiskt att förnya urvalet vid varje ny undersökningsomgång, bör urvalet förnyas åtminstone ibland.

Hur stor bör rotationsandelen vara?
I avsnitt 5.2 diskuterades olika förhållanden i en upprepad undersökning som talar för att undersöka samma objekt vid upprepade tillfällen och de krav som talar för att byta objekt. Det är svårt att hitta en optimal kompromiss mellan dessa motstående krav. Cochran (1977, s. 351–355) räknar på varianser för nivå- och förändringsskattningar under olika antaganden om korrelation över tid. I hans exempel är målpopulationen oföränderlig. Han försummar även bortfall och mätfel. I detta förenklade fall finner han att om det är viktigt att skatta såväl förändring som nivå är det lämpligt att byta mellan 1/3 och 1/5 av urvalet. Om effekter av typen uttröttning och panel-effekt kan befaras bli starka, bör en större andel av urvalet bytas ut inför varje omgång. Om det å andra sidan krävs en initial investering att kontakta uppgiftslämnaren eller att ”lära upp” denna, kan det i stället vara lämpligt att behålla en större andel av urvalet. Det är vanligt inom SCB att sträva efter att behålla en större andel än 4/5 av urvalet (jämför exemplet i avsnitt 7.3).

Tekniker för panelurval
Det går att dra panelurval med olika tekniker. Ett urval med någon lämplig urvalsmetod dras vid tidpunkt t. Vid nästa tidpunkt, t +1, dras ett underurval ur det urval som drogs vid t. Dessutom dras ett urval från den del av ramen som inte ingick i urvalet som drogs vid t.

En teknik att utföra detta är att dela in hela urvalet vid t i grupper, rotationsgrupper. Dessa tilldelas ordningstal slumpmässigt. Vid $t + 1$ samlas uppgifter från alla objekt i grupperna förutom den med högst ordningstal. Denna grupp ”roteras ut” och hamnar i en ”ej valbar” grupp, se en schematisk beskrivning i figur 5.4.1. Där finns även objekt som har roterats ut vid tidigare omgångar. Ramen består alltså av tre mängder: de som tillhör aktiva rotationsgrupper, de som inte är valbara och övriga. Vid $t + 1$ dras även ett urval ur de övriga, av samma storlek som en rotationsgrupp.

Figur 5.4.1 En population ur vilken ett panelurval med ett antal rotationsgrupper samt en ny rotationsgrupp dras



Det är inte ovanligt hos andra statistikproducenter än SCB att man har en viloperiod mitt i sekvensen av panelurval. Data från objekten i en rotationsgrupp hämtas in i ett antal omgångar. Sedan får de vila ett par omgångar och kommer därefter med igen. Slutligen roteras gruppen ut. Skälet är att minska effekterna av sådana fenomen som paneltrötthet. Det har hävdats

(Hidiroglou och Srinath 1993, s. 399) att viloperioder är olämpliga i företagsundersökningar, eftersom företag använder fasta rutiner för sitt uppgiftslämnande. En viloperiod skulle därför riskera att öka uppgiftslämnarbördan.

I Arbetskraftsundersökningarna (AKU) försöker man intervjua varje utvald person vid åtta tillfällen, en gång per kvartal. Varje månad tas kontakt med personer som ingår i en rotationsgrupp. Grupperna är dragna oberoende av varandra. Skälet till att man låter det gå tre månader mellan intervjuerna är att förändringar från kvartal till kvartal är prioriterade framför månadsförändringar. I det här fallet beror inte ”viloperioderna” på oro över alltför stor uppgiftslämnarbörda.

Nya objekt måste spridas ut på ett sådant sätt att de inte över- eller underrepresenteras i någon rotationsgrupp eller i den ej valbara gruppen.

I avsnitt 7.3 beskrivs AKU:s paneldesign kortfattat.

Vad har panelurval för urvalsdesign?
Ett panelurval har, precis som för andra urvalsmetoder, en urvalsdesign som beror på hur urvalet är draget. Ett sätt är att dela in hela populationen i kluster och dra ett förutbestämt antal kluster som får utgöra rotationsgrupper. Det är dock vanligt att det vid tidpunkt *t* dras ett urval med någon fastställd design, t.ex. OSU. Ett underurval dras ur detta vid tidpunkt *t*+1 och ytterligare ett ur den del av populationen som inte drogs vid *t*. Om underurvalet designas med hänsyn till utfallet av urvalet vid *t*, är det ett tvåfasurval.

Differensen kan skattas som den enkla skillnaden mellan skattningarna vid *t* +1 och *t*, $\hat{\theta}_{t+1} - \hat{\theta}_t$. Variansen blir

$$V\left(\hat{\theta}_{t+1} - \hat{\theta}_t\right) = V\left(\hat{\theta}_{t+1}\right) + V\left(\hat{\theta}_t\right) - 2Cov\left(\hat{\theta}_{t+1}, \hat{\theta}_t\right) \qquad (5.4.1)$$

Om $\hat{\theta}_{t+1} - \hat{\theta}_t$ är skillnaden mellan två urvalsmedelvärden vid OSU med försummad ändlighetskorrektion, blir variansen

$$V\left(\bar{y}_{t+1} - \bar{y}_t\right) = S_{t+1}^2 / n_{t+1} + S_t^2 / n_t - 2\rho S_{t+1} S_t \, n_m / n_{t+1} n_t \qquad (5.4.2)$$

där S_t^2 och n_t är populationsvariansen och urvalsstorleken vid *t*, ρ är korrelationen mellan y_t och y_{t+1}, och n_m är storleken på det urval som är gemensamt för tidpunkterna. Hidiroglou et al. (1995, s. 498) visar hur kovarianstermen kan skattas för en differens av totaler för ett stratifierat OSU, där skattningen tar hänsyn till att objekt försvunnit och tillkommit under perioden.

En typ av skattning som ger lägre varians än den enkla differensen mellan skattningarna vid *t* +1 och *t*, $\hat{\theta}_{t+1} - \hat{\theta}_t$, är en sammansatt skattning. Skattningsformler finns i Särndal et al. (1992, avsnitt 9.9). Den skattningen blir enklast att förstå om man ser urvalet som ett tvåfasurval.

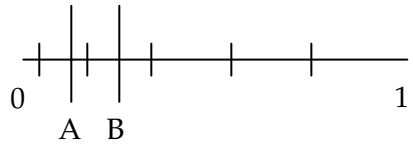
5.4.3 Urval med permanenta slumptal

Att dra urval med hjälp av *permanenta slumptal* är en elegant och effektiv metod (Atmer, Thulin och Bäcklund 1975). Brewer (2002) hävdar att SCB var först i världen med den metoden. Metoden kallas ibland *JALES* efter upphovsmännen Johan Atmer och Lars-Erik Sjöberg. Den har senare kommit att användas i åtskilliga länder.

Varje objekt tilldelas ett slumptal när det läggs in i urvalsramen. Slumptalen är observationer från en likformig fördelning på intervallet 0 till 1. Objektet behåller sitt slumptal under hela sin livslängd. Dock kan man vilja rotera objekten, något som behandlas nedan.

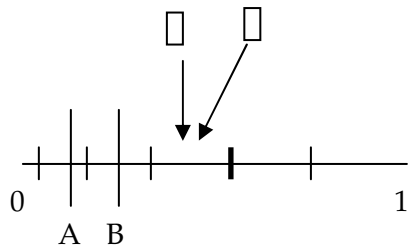
I exemplet i figur 5.4.2 består ramen av fem objekt som har placerats ut på en linje som går från 0 till 1. Vid tidpunkt t dras ett urval. Som startpunkt väljs punkt A. Om urvalsstorleken är tre, inkluderas de tre objekten till höger om A i urvalet. Punkten A kan väljas slumpmässigt. Normalt gör man dock inte det, utan punkten väljs så att man får lämplig samordning mellan undersökningar och mellan omgångar inom samma undersökning. Det går att bevisa att ett urval som dras med permanenta slumptal på det sätt som beskrivs här, är ett OSU oavsett hur startpunkten väljs (Ohlsson 1992). Vid tidpunkt $t + 1$ dras ett nytt urval. Vill man att det nya urvalet ska vara positivt samordnat med det första urvalet, tas t.ex. B som startpunkt.

Figur 5.4.2 Två urval dras från en linje där fem ramobjekt är inlagda med permanenta slumptal. Objektens position markeras med lodräta sträck. Det första urvalet dras med början från A, det andra med början från B



SCB:s implementering av tekniken med permanenta slumptal medför en skillnad mot panelurval. I figur 5.4.3 dras ett urval om tre objekt med start till höger om A. Nästa urval innefattar även det tre objekt, men med start från B. Emellertid tillkommer två nya objekt just innan det andra urvalet dras. De nya objekten får nya slumptal. De råkar landa på det ställe som figuren anger. Då består det andra urvalet av det objekt som redan förut låg omedelbart till höger om B (detta objekt ingick även i det första urvalet) samt de två följande objekten, som är de två nya. Det objekt som i figuren är ritat med en fetare linje kommer däremot inte med i det andra urvalet, trots att det ingick i det första urvalet. Det ”knuffas ut” av de nytillkomna. Detta ger en extra slumpvariation som inte finns i panelurval. Nordberg (2000) visar hur variansen kan beräknas i den här situationen.

Figur 5.4.3 Efter det att det första urvalet dragits med början från A, får två nya ramobjekt sådana slumptal att de kommer att ingå i urvalet som dras med början från B



Samordning av undersökningar går relativt smidigt med permanenta slumptal. Varje undersökning får sin startpunkt på linjen med permanenta slumptal. De som ska samordnas positivt får samma startpunkter eller startpunkter nära varandra så att deras urval överlappar varandra.

Som alltid i upprepade undersökningar finns det behov av att byta ut en del av urvalet mellan omgångarna. I dragning med permanenta slumptal finns två alternativ. Antingen flyttas startpunkten något mellan omgångarna eller så roteras slumptalen. T.ex. kan man lägga till en liten konstant till alla slumptal. De som befinner sig nära 1 och får slumptal större än 1 roteras ”cirkulärt”: man drar bort 1 från alla slumptal som är större än 1.

Samordningen kräver inte att urvalen dras med samma urvalsmetod. Det går att samordna två undersökningar som t.ex. använder olika stratumindelning. Samordningen blir dock inte lika effektiv. Se vidare i Ohlsson (1995, s. 156–157). Även πps-urval kan enkelt dras med permanenta slumptal; se avsnitt 4.4.

5.4.4 Jämförelse mellan panelurval och urval med permanenta slumptal

Panelurval och urval med permanenta slumptal fyller delvis samma syften. Panelurval är lämpliga i undersökningar där det behövs överlappning över tid, men där en samordning med åtskilliga andra undersökningar är oväsentlig. Det senare är nämligen svårt eller omöjligt med panelurval. Panelurval passar bättre om populationen är mycket stor än om den är liten.

Urval med permanenta slumptal klarar positiv och negativ samordning på ett enkelt sätt. Vid tvåstegs- eller flerstegsurval kan dock samordningen bli komplicerad.

5.4.5 Läs mer

Hidiroglou, Choudhry och Lavallée (1991) föreslår en teknik att dra panelurval som de hävdar är enklare än den som beskrivs ovan i avsnitt 5.4.2. Med deras teknik delas hela ramen in i rotationsgrupper. Tekniken har också fördelen att det även går att styra hur länge olika objekt *inte* blir utvalda. De ger även skattningsformler och regler för hur nya objekt ska föras in i rotationsgrupperna.

Vid dragningen av det andra urvalet i en sekvens av två delvis överlappande urval kan värdena på undersökningsvariabeln som observerades i det första urvalet utnyttjas. T.ex. kan det första urvalet dras med OSU. I nästa omgång kan ett underurval ur det första urvalet dras med πps-urval, där undersökningsvariabeln utgör storleksmått. Arnab (1991) diskuterar och ger skattningsformler för olika sätt att göra detta, med målsättning att skatta totalen vid tidpunkt t +1.

Ohlsson (1995) ger en pedagogisk beskrivning av olika urvalsmetoder som bygger på permanenta slumptal. Ernst, Valliant och Casady (2000) löser vissa ytterligare problem med tekniken som uppkommer när objekt tillkommer och tas bort ur ramen.

5.5 Hur samordnar SCB sina företagsundersökningar?

På SCB kan samordning av urval mellan olika företagsundersökningar eller samordning över tid inom en företagsundersökning göras i *SAMU-systemet*, SAMordnad Urvalsdragning. SAMU-systemet är integrerat med FDB, Företagsdatabasen. SAMU är ett system för urvalsdragning ur FDB. Samordningen mellan och inom undersökningar styrs med JALES-tekniken med permanenta slumptal. Tekniken beskrevs i avsnitt 5.4.3.

De urvalsmetoder som kan användas i SAMU är OSU, stratifierat OSU och Pareto πps-urval.

Inte alla SCB:s företagsundersökningar använder SAMU. Ett exempel på företagsundersökningar utanför SAMU är undersökningar där populationen består av lantbruksföretag. Då skapas urvalsramarna från Lantbruksregistret (LBR) i stället för från FDB. LBR innehåller betydligt mer detaljerade uppgifter om lantbruket än FDB.

Andra skäl till att undersökningar inte läggs in i SAMU är att urvalsmetoden som används inte finns implementerad i SAMU eller att företag som ska ingå i undersökningen inte finns med i SAMU, se avsnitt 5.5.1.

Avsnitt 5.5 handlar främst om samordning av företagsurval för undersökningar som använder SAMU-systemet. Det finns företagsundersökningar på SCB som är samordnade utan att använda SAMU. T.ex. är Gödselmedelsundersökningen och en skördeundersökning sinsemellan samordnade genom permanenta slumptal; se avsnitt 5.5.5.

5.5.1 Urvalsramar i SAMU

FDB underhålls kontinuerligt av SCB. I SAMU används fyra olika versioner av FDB varje år. SAMU-systemet hämtar in nya versioner av FDB i mars, maj, augusti och november. Alla undersökningar som använder SAMU utgår från någon av dessa versioner av FDB vid ramframställning (se avsnitt 3.4 för mer detaljer om ramframställning). Det förekommer att urval dras oftare än en gång per år och då kan exempelvis senast tillgängliga version av FDB i SAMU användas. Det är dock en hel del arbete med att byta urval: nya uppgiftslämnare ska kontaktas och man ska lära upp dem. Dessutom kan bortfallet bli större första gången man använder ett nytt urval.

Inte hela FDB ingår i en SAMU-version, utan det är endast de företag som är klassade som verksamma samt företag som inte är verksamma men som är av stor ekonomisk betydelse som ingår i SAMU. Ibland kan det vara av intresse att ha med även andra företag som av olika skäl inte kan tas med i SAMU. Om större delen av ramen finns med i SAMU-systemet kan man dra urval från den delen i SAMU och sedan komplettera SAMU-urvalet med ett urval av de företag som inte finns med i SAMU. Företagspopulationen är snabbt föränderlig: nya företag bildas, företag läggs ned eller går ihop med andra företag, och företag splittras. Om två undersökningar ska samordnas med syftet att göra statistiken jämförbar är det därför viktigt att utgå från samma version av FDB vid ramframställning och urvalsdragning, se även avsnitt 3.4.2.

5.5.2 Samordning av urval över tid i SAMU

Som nämndes i avsnitt 5.2 är det viktigt vid skattning av förändring i upprepade undersökningar att överlappningen av urvalen är stor. Men eftersom företagspopulationen förändras snabbt är det viktigt att uppdatera urvalet med jämna mellanrum. I SAMU sköts det genom att urvalen dras från en uppdaterad urvalsram, baserad på en ny version av FDB.

5.5.3 Samordning av urval mellan undersökningar i SAMU

Alla undersökningar som ingår i SAMU ingår i ett av fem ”block”. Undersökningarna i ett block är positivt samordnade genom att samma startpunkt och riktning används. Med riktning menas åt vilket håll från startpunkten man går, åt höger eller vänster, då urvalet ska dras. Jämför figur 5.4.2 där riktningen är höger. För två undersökningar som ligger i olika block används olika startpunkter och ibland även olika riktningar, och dessa blir då negativt samordnade. Graden av negativ samordning beror på hur nära startpunkterna ligger varandra, på riktningarna i de två blocken och på hur stora inklusionssannolikheterna är.

Om stratifiering görs på en alltför detaljerad nivå medför det ofta att inklusionssannolikheterna blir relativt stora. Då tenderar företagen att ingå i många undersökningar. Ofta stratifieras företagsundersökningar åtminstone med avseende på bransch och storlek, där branscher utgör redovisningsgrupper. Om precisionen ska vara bra i alla redovisningsgrupper så leder det till att i branscher med få företag väljs en stor andel av företagen ut, vilket i sin tur leder till att dessa företag får en stor uppgiftslämnarbörda. Ett sätt att komma runt detta är att ha olika precisionskrav i olika redovisningsgrupper.

När en ny undersökning ska läggas in i SAMU måste man avgöra vilka andra undersökningar som den nya undersökningen ska samordnas positivt med. Det bestämmer blocktillhörighet. Om det inte finns några krav på positiv samordning kan inplaceringen ske med hänsyn till uppgiftslämnarbördan. I Statistiska centralbyrån (2003) finns en beskrivning av blocken när det gäller startpunkter, riktningar och vilka undersökningar som ingick i de olika blocken då rapporten skrevs. Beskrivningen av vilka undersökningar som ingår är något inaktuell, eftersom antalet undersökningar som är inlagda i SAMU har ökat. Startpunkterna och riktningarna gäller dock fortfarande.

5.5.4 Rotation i SAMU

För att sprida uppgiftslämnarbördan över tiden och för att förnya urvalet används rotation i SAMU. Rotationen görs genom att de ”permanenta” slumptalen roteras. Den lösningen är tekniskt enklare än att flytta på startpunkterna och ger samma effekt, se avsnitt 5.4.3. Rotationen av slumptal görs en gång per år. Ca 20 procent av företagen får därvid sina slumptal roterade. I praktiken innebär detta dock inte att ca 20 procent av urvalet byts ut om ett nytt urval dras en gång per år. Rotationen av slumptal sköts i SAMU-systemet, och det går inte att själv välja hur stor andel som ska roteras i en specifik undersökning.

Hur fort ett enskilt företag roterar ut ur en undersökning beror inte bara på rotationen av slumptalen. Det beror även på populationsförändringar, dvs. hur länge det enskilda företaget består och hur länge det stannar kvar i den

bransch och storleksklass det tillhör. Populationsförändringar går inte att kontrollera och om många nya företag bildas mellan två urvalsdragningar kan rotationen bli betydande. Rotationen beror även på hur stora inklusionssannolikheterna är och förändringar i urvalsdesignen. Om t.ex. stratifiering görs med avseende på storlek mätt i antal anställda och företaget växer något, så kan det hamna i en större storleksklass där inklusionssannolikheten är större, och därmed kan det bli kvar i urvalet längre.

I företagsundersökningar av ekonomiska variabler väljs oftast urval bland de storleksmässigt, mätt i t.ex. antal anställda eller omsättning, små och medelstora företagen, medan de stora företagen totalundersöks. I de flesta företagsundersökningar varierar inklusionssannolikheterna kraftigt med företagens storlek. För de mindre företagen är de ofta små, medan de för de medelstora kan vara betydligt större, på grund av att dessa företag dels är mer betydelsefulla för skattningarna, dels är färre. Detta gör att rotationen oftast bara fungerar fullt ut bland de små företagen. De medelstora kan bli kvar alltför länge. De stora företagen får alltid ingå så länge de är stora, så bland dem blir rotationen mycket liten.

5.5.5 Exempel

Här ges ett exempel på urvalssamordning utanför SAMU-systemet. Ytterligare exempel på positiv samordning finns i avsnitt 7.2 och 7.3. Ett exempel på negativ samordning ges i avsnitt 7.5.

EXEMPEL 5.5.1 *Gödselmedel och skörd av spannmål m.m.*
För Gödselmedelsundersökningen (se exempel 4.4.2) dras ett stratifierat Pareto πps-urval, med standardtimmar som storleksmått. För skördeundersökningen avseende spannmål, trindsäd och oljeväxter (se exempel 3.4.3) dras ett stratifierat Pareto πps-urval av typen ParMix, med sammanvägda grödarealer som storleksmått.

Urvalen samordnas genom att de slumptal som genererats för skördeundersökningen används i urvalsdragningen för Gödselmedelsundersökningen. För de företag (huvudsakligen djurgårdar) i urvalsramen för Gödselmedelsundersökningen som inte ingår i ramen för skördeundersökningen saknas därmed slumptal. Därför genereras slumptal för dessa, enligt en rektangelfördelning i intervallet mellan 0 och 1. Slumptalen används sedan för Pareto πps-urvalen inom stratumen för Gödselmedelsundersökningen. □

5.5.6 Läs mer

I Ohlsson (1992) ges en beskrivning av de metoder som används i SAMU.

6 Övergripande urvalsfrågor

Tidigare kapitel har behandlat ramar, vissa kända urvalsmetoder och samordning av ramar och urval. Detta kapitel diskuterar mer övergripande urvalsfrågor.

6.1 Hur stort ska urvalet vara?

I avsnitt 2.2.4 berördes de viktigaste faktorerna bakom kraven på urvalsstorlek: önskad precision, kostnad, bortfall och övertäckning samt uppgiftslämnarbörda. Här diskuteras dessa faktorer djupare.

I planeringen kan det vara bra att prova några olika ansatser för att se om urvalsstorleken blir ungefär densamma eller om den är känslig för olika ansatser. *Känsligheten* kan åskådliggöras i en graf med t.ex. urvalsstorlek efter y-axeln och olika antaganden om populationsvarians efter x-axeln.

Facit på hur stort urval man verkligen behöver får man inte förrän undersökningen är genomförd.

6.1.1 Precision

Vid planering av nya undersökningar finns inte alltid kvantifierade *precisionskrav*. För att få fram precisionskrav diskuterar man statistikens avsedda användning med beställaren eller intressenterna av statistiken. När man diskuterar precision för kvantitativa variabler kan man använda sig av flera tänkbara mått. Anta att man vill skatta totalen t_y för en kvantitativ variabel *y* med estimatorn $\hat{t}_y$. Några tänkbara **precisionsmått** är

- *medelfelet* för $\hat{t}_y$, eller standardavvikelsen, $\sqrt{V(\hat{t}_y)}$
- *relativa medelfelet* för $\hat{t}_y$, $CV(\hat{t}_y) = \sqrt{V(\hat{t}_y)} \big/ E(\hat{t}_y)$
  (Särndal et al., 1992, s. 42)
- *osäkerhetsmarginalen*, eller felmarginalen, för $\hat{t}_y$, $1{,}96\sqrt{V(\hat{t}_y)}$
- *relativa osäkerhetsmarginalen* för $\hat{t}_y$, $r = 1{,}96 \times CV(\hat{t}_y)$
  (Statistiska centralbyrån, 2001a, s. 27)
- *absoluta felet* för $\hat{t}_y$, $\left| t_y - \hat{t}_y \right|$.

Det relativa medelfelet kallas ibland även för variationskoefficienten. Här reserverar vi dock termen *variationskoefficient* för populationsparametern, se formel (6.1.2). I Statistiska centralbyrån (2001a, s. 27) avses med relativt medelfel kvantiteten $CV(\hat{t}_y) = \sqrt{\hat{V}(\hat{t}_y)} \big/ \hat{t}_y$. Om $\hat{t}_y$ är en HT-estimator är väntevärdet $E(\hat{t}_y)$ exakt lika med t_y. För de flesta estimatorer som används inom SCB gäller att $E(\hat{t}_y) \approx t_y$.

Man kan visa att den relativa osäkerhetsmarginalen blir densamma oavsett om man skattar en total eller ett medelvärde, givet att populationsstorleken N är känd.

Vilket precisionsmått man vill använda sig av när man bestämmer urvalsstorleken beror på sammanhanget. Om man exempelvis i en undersökning med flera redovisningsgrupper vill ha samma precision för alla grupper bör man använda sig av medelfelet eller osäkerhetsmarginalen (separat för varje grupp). Här i avsnitt 6.1 används i allmänhet den relativa osäkerhetsmarginalen, r.

De av precisionsmåtten ovan som troligen är lättast för icke-statistiker att föreställa sig i planeringen av undersökningen är den relativa osäkerhetsmarginalen och det absoluta felet. I SAMU-systemet (se avsnitt 5.5) används det relativa medelfelet som precisionsmått för att bestämma urvalsstorleken.

Efter diskussion brukar de användare som har ett tydligt syfte med statistiken kunna specificera en relativ osäkerhetsmarginal eller ett absolut fel, låt vara preliminärt och med viss approximation. Efter beräkningar av urvalsstorlekar visar det sig ofta att den önskade precisionen leder till en alltför stor urvalsstorlek, som är alltför kostsam. Då diskuterar man precisionskraven igen på en mer realistisk nivå. Ett par tänkbara lösningar är att

- sänka precisionskraven för planerade redovisningsgrupper
- dra ned på detaljeringsnivån i redovisningsgrupperna.

Andra möjligheter, som ligger utanför val av urvalsmetod, inkluderar byte till en billigare datainsamlingsmetod eller att överväga cut-off-urval i kombination med modellbaserade skattningar, se avsnitt 6.2.4.

Se avsnitt 6.1.4 för exempel på hur det kan gå till i praktiken att beräkna urvalsstorlekar.

Relativa osäkerhetsmarginalen härleds från ett 95 procents konfidensintervall. Konfidensgraden 95 procent gäller med god approximation om urvalsstorleken är tillräckligt stor. I Cochran (1977, s. 42) finns en mycket grov tumregel för minsta val av urvalsstorlek n som krävs för att approximationen ska vara giltig om urvalet dragits med OSU:

$$n > 25G_1^2, \qquad (6.1.1)$$

där $G_1 = \sum_{k=1}^{N} \left(y_k - \bar{y}_U\right)^3 / NS_{yU}^3$ är Fishers mått på hur *sned* fördelningen för y är, S_{yU} är standardavvikelsen för y i populationen U, $\bar{y}_U$ är populationsmedelvärdet och N är populationsstorleken. För andra tänkbara tumregler för val av n, se referenserna i avsnitt 6.1.13.

EXEMPEL 6.1.1 *Snedhet*
Vad kan tumregeln i formel (6.1.1) tänkas ge för urvalsstorlekar? Betrakta OSU dragna ur FDB. Här avgränsas FDB-versionen som användes i SAMU november 2004 till företag som är registrerade som aktiebolag. För aktiebolag med minst 5 och högst 99 anställda är snedheten 2,7 för antal anställda och 39 för omsättning. Snedheten 39 leder enligt tumregeln till urvalsstorleken n = 28 025 för ett OSU. I storleksklassen 20–49 anställda är snedheten för antal anställda mindre än 1. För denna storleksklass räcker det med

n = 25 för att skattningens fördelning kan approximeras med en normalfördelning. □

6.1.2 Kostnad och resursallokering

Undersökningens budget kan grovt delas in i kostnader för insamling av data respektive andra kostnader. Insamling kan i sin tur delas in i

- huvudinsamling

och eventuellt även

- bortfallsuppföljning
- mätfelsstudie och bortfallsstudie, t.ex. genom underurval (se avsnitt 7.1)
- särskild insamling av objekt som på grund av bristande ramtäckning inte kom med i huvudinsamlingen.

Problemet att allokera resurser över de olika momenten i punktsatserna ovan ligger utanför ramen för denna skrift. Mer allmänt handlar det om att hitta en lämplig balans mellan bias och varians. Biasen kan komma från bortfall, mätfel och ramfel. En stor redovisningsgrupp med stort urval har liten varians, varför biasen tenderar att vara den största källan till bristande tillförlitlighet. För små redovisningsgrupper med små urval är variansen ofta den största källan till bristande tillförlitlighet. Vår diskussion handlar fortsättningsvis om kostnaderna för huvudinsamlingen.

Den budget som finns för huvudinsamlingen kan vanligen översättas till en ungefärlig största urvalsstorlek, givet en viss datainsamlingsmetod. Då går det att prova sig fram med olika designer och estimatorer för att få en uppfattning om hur hög precisionen blir med den budgeten.

Avsnitt 7.1 innehåller ett exempel på en undersökning som innehöll tilläggsurval i form av underurval för både en bortfallsstudie och en mätfelsstudie. I det fallet bedömdes det vara kostnadseffektivt att styra resurser till dessa moment, eftersom bortfallet och mätfelet annars hade kunnat orsaka systematiska fel.

6.1.3 Variationskoefficient och designeffekt

Begreppen variationskoefficient och designeffekt är praktiska att använda vid urvalsbestämning.

Begreppet variationskoefficient för *y*

I det följande används *variationskoefficienten* av y i populationen U, nämligen

$$cv_{yU} = \frac{S_{yU}}{\bar{y}_U} \qquad (6.1.2)$$

där S_{yU} är populationsstandardavvikelsen och $\bar{y}_U$ är medelvärdet i populationen (Särndal et al., 1992, s. 64). ”Populationen” kan vara hela målpopulationen eller en redovisningsgrupp. Fördelen med att normera med $\bar{y}_U$ är att variationskoefficienterna för olika variabler kan jämföras. Som tidigare nämnts kallas även relativa medelfelet för variationskoefficienten, se avsnitt 6.1.1.

Variationskoefficienten för y kan skattas utifrån data från föregående undersökningsomgång eller med hjälp av en registervariabel som används som substitut. Se även avsnitt 6.1.10 för ytterligare råd om hur man kan bilda sig en uppfattning om dess storlek. Särndal et al. (1992, s. 55) skriver att den är mellan en halv och ett för ”många variabler och många populationer”. Men för variabler med extremt sned fördelning, t.ex. vissa ekonomiska variabler i företagspopulationer, är variationskoefficienten ofta mycket större, se exempel 6.1.2 nedan.

EXEMPEL 6.1.2 *Variationskoefficienter i företags- och individstatistik*
Här följer några exempel på variationskoefficienter beräknade enligt formel (6.1.2):

Ett exempel från företagsstatistiken: Betrakta samma population som i exempel 6.1.1. För registervariabeln antal anställda var variationskoefficienten ca 16. Det extremt höga värdet beror på att ca 35 procent av företagen inte har några anställda alls, medan en mindre andel har många anställda. Färre än 0,1 procent har flera än 500 anställda, men dessa företag höjer variationskoefficienten avsevärt. För företag med minst 5 och högst 99 anställda var variationskoefficienten 1,0 för antal anställda. För omsättning var variationskoefficienten 3,5 för denna grupp av företag.

Ett exempel från individstatistiken är löneinkomst under 2005 för individer äldre än 17 år och bosatta i Sverige. Löneinkomsten är en härledd variabel som kommer från Inkomst- och taxeringsregistret. För hela populationen var variationskoefficienten för denna variabel ca 1,7. Uppdelat på män och kvinnor var variationskoefficienten fortfarande 1,7 för män medan den var 1,3 för kvinnor. Uppdelat på åldersklasserna 18–29, 30–65 och 65+ var variationskoefficienten för löneinkomst 1, 1,3 respektive 7,5. Uppdelat slutligen på kommun befann sig alla variationskoefficienter inom intervallet 0,3–3,8. □

Begreppet designeffekt
Designeffekt (DEFF) definieras som

$$\mathrm{DEFF} = \frac{\text{skattningens varians under aktuell urvalsmetod}}{\text{skattningens varians under OSU}} \qquad (6.1.3)$$

Designeffekten används för jämförelser av varianser för olika urvalsmetoder. Ofta är DEFF ungefär densamma för πps-urval som för stratifierat OSU: i storleksordningen 0,5 (alltså 50 procent lägre varians än för vanligt OSU). En så pass kraftig variansreduktion kan man bara hoppas på om populationen är ganska heterogen och stratifieringsvariablerna har stark korrelation med undersökningsvariablerna. Vinsten med stratifiering blir betydligt mindre för urval ur relativt homogena populationer, men även i det fallet kan man räkna med att designeffekten är aningen mindre än 1, kanske 0,95. Det här är naturligtvis grova siffror, vilka tas upp här endast för att ge läsaren en känsla för vad som är rimliga storleksordningar. Se Cochran (1977, s. 101–103) för exempel på både stora och små variansminskningar vid stratifierat urval jämfört med OSU. För en djupare diskussion av val mellan urvalsmetoder, se avsnitt 6.2.

Klusterurval är mindre effektivt än OSU. Om den genomsnittliga klusterstorleken är mindre än 100 är DEFF nästan alltid mellan ett och sex, se vidare Kish (1987, s. 41–42). Man kan alltså i detta fall räkna med att DEFF

för ett klusterurval är något större än ett, ibland upp till två eller tre och någon gång upp till sex (se Kish 1987, s. 41–42, och även Särndal et al. 1992, s. 492, för en beräkningsformel). Designeffekter som är större än sex kan förekomma; det beror ofta på något av följande:

- förekomst av outliers
- att något har gått allvarligt fel i undersökningen.

6.1.4 Olika steg för bestämning av urvalsstorlek

Det kan rekommenderas att bestämning av urvalsstorlek görs i följande fyra steg:

Steg 1: Ta reda på *vilka skattningar* som undersökningen ska producera. Säg att den ska producera totalskattningar för variablerna y_1 och y_2. Dessutom ska den ge jämförelser av totalerna för y_1 och y_2 sinsemellan och var för sig över tid. Det är alltså tre sorters skattningar för två variabler. Dessutom finns redovisningsgrupper att ta hänsyn till. På planeringsstadiet är redovisningsgrupperna i princip av två sorter: de som består av objekt vars tillhörighet till en viss redovisningsgrupp går att identifiera innan urvalet dragits (åtminstone med hög sannolikhet) och övriga redovisningsgrupper, som alltså inte går att identifiera i förväg. Den förra typen kan man och bör man planera för i urvalsdesignen, den senare är svårare att göra något åt. Emellertid kan och bör man fundera på tänkbara utfall, dvs. olika tänkbara sammansättningar av det erhållna urvalet och hur olika dessa kan uppfylla framtida behov. Finns exempelvis risken att man får ett urval innehållande få eller inga individer med ett visst medborgarskap, samtidigt som man förväntar sig framtida önskemål om särredovisningar för denna grupp?

Steg 2: Försök få fram restriktioner vad gäller *kostnader* (såväl producentkostnader som uppgiftslämnarkostnader) och önskemål om *precision*. Diskutera precisionskrav med kunden eller beställaren i enlighet med kapitel 8. Dessa begränsningar och önskemål kan i ett senare skede visa sig stå i konflikt med varandra.

Steg 3: Bestäm urvalsstorleken med hänsyn till hur mycket som *i förväg är känt* om det som undersökningen ska belysa. Några olika situationer beskrivs nedan:

- Undersökningen har gjorts förut under liknande förhållanden …
  - … och ska upprepas på samma sätt men med förändrade precisionskrav eller med en annan estimator (se avsnitt 6.1.5 och 6.1.6)
  - … men en annan urvalsmetod ska användas (se avsnitt 6.1.7)
  - … men det finns önskemål om nya redovisningsgrupper (se avsnitt 6.1.8)
  - … men det finns önskemål om skattning av ytterligare parametrar (se avsnitt 6.1.9).
- Undersökningen har antingen inte gjorts förut eller har genomförts under väsentligt annorlunda förhållanden. Även här är ovanstående fyra situationer tillämpliga, men då ska de exempelvis uppfattas som att allt utom precisionskrav är bestämt, eller att allt utom estimator är bestämt. Det förutsätts på flera ställen att det finns en uppfattning om vad cv_{yU} har för storleksordning. Det förutsätts vidare ofta att de

relativa storlekarna av variationskoefficienter i redovisningsgrupper, cv_{yU_d}, är ungefärligen kända. I en helt ny undersökning kan bedömningar göras enligt avsnitt 6.1.10. Se också avsnitt 6.1.6. Eftersom undersökningen är ny bör man tillåta sig större marginaler än vad man annars hade haft. Det innebär att man hellre låter urvalsstorleken bli för stor än för liten.

Steg 4: Steg 1–3 ger ett grepp om den urvalsstorlek som användningen av *en* variabel kräver. I praktiken har man i allmänhet *många undersökningsvariabler* och vill ofta *skatta mer än en parameter*. Då får man försöka identifiera vilka skattningar som är viktigast eller som ställer stora eller speciella krav på urvalsmetod och urvalsstorlek. Avsnitt 6.1.5–6.1.9 ger vägledning i några olika situationer.

6.1.5 Nytt urval med annat precisionskrav eller annan estimator

Här antas att det enda som ändrats, eller det enda som inte är bestämt, är precisionskrav eller estimator. Nedan gås igenom hur man beräknar ny urvalsstorlek för olika urvalsmetoder och estimatorer. Här antas att bortfall inte förekommer; se avsnitt 6.1.11 för en diskussion om förväntat bortfall i urvalsplaneringen.

OSU och HT-estimatorn av en total eller ett medelvärde

Anta att man formulerat ett nytt krav på den relativa osäkerhetsmarginalen *r* för en total eller ett medelvärde. Genom att lösa ut *n* ur formeln $r = 1{,}96 \times CV\left(\hat{t}_y\right)$ får vi

$$n = \left(\frac{1{,}96 \times cv_{yU}}{r}\right)^2 \Bigg/ \left(1 + \frac{1}{N}\left(\frac{1{,}96 \times cv_{yU}}{r}\right)^2\right) \qquad (6.1.4)$$

Om urvalsstorleken är liten i förhållande till populationsstorleken, dvs. om n/N är liten, förenklas formel (6.1.4) till

$$n = \left(\frac{1{,}96 \times cv_{yU}}{r}\right)^2 = \frac{1{,}96^2 \times cv_{yU}^2}{r^2} \qquad (6.1.5)$$

Om urvalsfraktionen n/N är liten och *r* ska halveras, behöver *n* nästan fyrfaldigas. I många företagsundersökningar är n/N inte försumbar och då blir den erforderliga ökningen av urvalet något mindre.

OSU och HT-estimatorn av en andel

Formel (6.1.4) gäller i princip även då *y* är en kategorivariabel. Man utnyttjar då att $S_{yU}^2 \approx P(1-P)$, där *P* är andelen objekt i populationen med den sökta egenskapen. När det gäller andelar är det praktiskt att använda det absoluta felet i planeringsfasen. Variansen för skattningen av andelen med en viss egenskap, *p*, blir densamma som för skattningen av andelen utan egenskapen, 1–*p*. Det relativa medelfelet för *p* och 1–*p* blir däremot olika. Anta exempelvis att $P = 0{,}9$ och $V(p) = 0{,}02$. Då gäller att

$$CV(p) = \frac{\sqrt{V(p)}}{P} = \frac{0{,}02}{0{,}9} \approx 0{,}022 \text{ och}$$

$$CV(1-p) = \frac{\sqrt{V(p)}}{1-P} = \frac{0{,}02}{0{,}1} = 0{,}2\,.$$

Att ställa olika precisionskrav då man skattar *P* och 1–*P* kan upplevas som orimligt och därför är relativa medelfelet inte lämpligt att använda för andelar.

Som precisionskrav då man vill skatta en andel kan i stället föreslås att man använder det absoluta felet, $|P-p|$, där *p* är andelen i urvalet med aktuell egenskap. Den önskvärda precisionen kan exempelvis vara ± 2 procentenheter.

Man kan visa att det är samma sak att använda det absoluta felet och osäkerhetsmarginalen *d* för att bestämma urvalsstorleken, se Cochran (1977, s. 75). Om man löser ut *n* ur formeln $d = 1{,}96 \times \sqrt{V(p)}$ får man:

$$n = \frac{1{,}96^2 P(1-P)}{d^2} \Bigg/ \left(1 + \frac{1}{N}\left(\frac{1{,}96^2 P(1-P)}{d^2} - 1\right)\right) \qquad (\ 6.1.6\ )$$

enligt Cochran (1977, formel (4.1)).

Den exakta populationsvariansen för en kategorivariabel är $S_{yU}^2 = (N/(N-1))P(1-P)$, se formlerna (4.2.6)–(4.2.8). Genom att använda approximationen $S_{yU}^2 \approx P(1-P)$ förenklar sig (6.1.6) till följande formel, vilket är den som används i praktiken.

$$n = \frac{1{,}96^2 P(1-P)}{d^2} \Bigg/ \left(1 + \frac{1}{N}\frac{1{,}96^2 P(1-P)}{d^2}\right) \qquad (\ 6.1.7\ )$$

Återigen, om urvalsstorleken är liten i förhållande till populationsstorleken, dvs. om n/N är liten, förenklas formel (6.1.7) till

$$n = \frac{1{,}96^2 P(1-P)}{d^2} \qquad (\ 6.1.8\ )$$

som i sin tur är detsamma som

$$n = \frac{P(1-P)}{V(p)} \qquad (\ 6.1.9\ )$$

Nu är $P(1-P)$ som mest 1/4. En övre begränsning för den urvalsstorlek som behövs är därför

$$n = \frac{1}{4V(p)} \qquad (\ 6.1.10\ )$$

Formel (6.1.10) kan jämföras med motsvarigheten för en kvantitativ variabel, som kan härledas ur formel (6.1.5):

$$n = \frac{N^2 S_{yU}^2}{V(\hat{t}_y)} \qquad (6.1.11)$$

där S_{yU}^2 i princip kan anta alla positiva värden. Eftersom S_{yU}^2 ofta är avsevärt större än 1/4 är en slutsats att kvantitativa variabler ofta kräver större urvalsstorlek än kategorivariabler.

Stratifierat OSU och HT-estimatorn av en total
De olika stegen för att avgöra en urvalsstorlek för stratifierat OSU då man skattar en total är följande:

1. Beräkna största acceptabla varians $V(\hat{t}_y)$ som följer av precisionskravet *r*:
   $$V(\hat{t}_y) = \frac{(r \cdot t_y)^2}{1{,}96^2}, \qquad (6.1.12)$$
   där totalen t_y skattas (se avsnitt 6.1.10).
2. Utför stratifiering i enlighet med avsnitt 4.2.
3. Beräkna
   $$n = \frac{\left(\sum_{h=1}^{H} N_h S_{yU_h} \right)^2}{V(\hat{t}_y) + \sum_{h=1}^{H} N_h S_{yU_h}^2} \qquad (6.1.13)$$
   Populationsstandardavvikelserna S_{yU_h} i strata måste uppskattas. Även detta problem behandlas i avsnitt 6.1.10.

Om undersökningen avser ett medelvärde och inte en total, dividerar man täljare och nämnare i formel (6.1.13) med N^2 och får

$$n = \frac{\left(\sum_{h=1}^{H} W_h S_{yU_h} \right)^2}{V(\bar{y}) + \frac{1}{N} \sum_{h=1}^{H} W_h S_{yU_h}^2} \qquad (6.1.14)$$

där $W_h = N_h / N$ och $V(\bar{y})$ är variansen av den skattning av medelvärdet i populationen som man tillämpar. Om man t.ex. skattar medelvärdet som $\hat{t}_y / N$, där *N* är den kända populationsstorleken, räknas variansen som $V(\hat{t}_y) / N^2$. Formel (6.1.14) motsvarar Cochrans (1977, s. 98) formel (5.25) om varje objekt kostar lika mycket att undersöka. Formeln gäller vid optimal allokering med samma kostnad för att undersöka ett objekt oavsett stratum, när man vill minimera kostnaden för ett givet precisionskrav *r*. Om det som ska skattas är en andel, byts $\bar{y}$ mot urvalsandelen *p* och $S_{yU_h}^2$ mot $P_h(1 - P_h)$ (jämför formel (4.2.8)).

Cochran ger även en formel för fallet då man vill minimera variansen givet en fast kostnad.

Ofta vill man bestämma en lämplig urvalsstorlek för en redovisningsgrupp. I figur 6.1.1 skär en redovisningsgrupp över strata. I detta fall brukar antalet objekt i redovisningsgruppen vara okänt. I det fallet brukar man inte räkna ut en lämplig urvalsstorlek, eftersom det ändå inte går att dra ett urval vars täckning sammanfaller med redovisningsgruppen.

I fall då redovisningsgruppen består av en union av strata är antalet objekt i redovisningsgruppen känt. Det första steget är att modifiera undersökningsvariabeln för varje objekt till

$$y_{dk} = \begin{cases} y_k \text{, om objekt k tillhör redovisningsgruppen d} \\ 0 \quad \text{annars} \end{cases}$$

Sedan beräknas populationsvariansen för den modifierade undersökningsvariabeln y_d som

$$S^2_{y_dU} = \frac{1}{N-1}\sum_U \left(y_{dk} - \bar{y}_{dU}\right)^2 \qquad (6.1.15)$$

med samma beräkningsformel som (4.2.6), fast med *y* bytt mot y_d. För strata görs samma beräkning inom varje stratum:

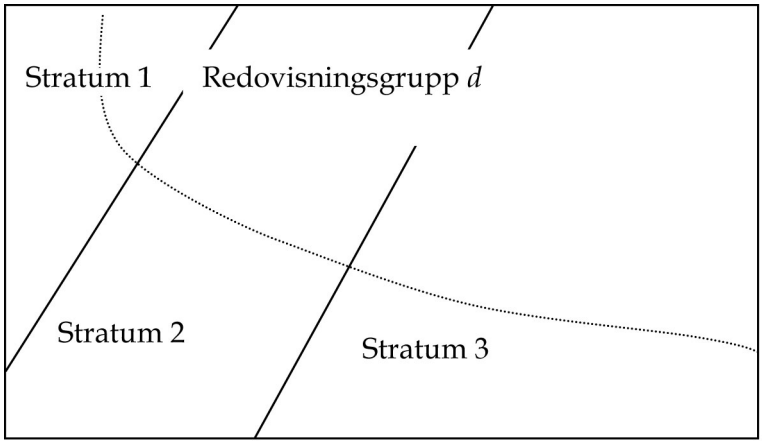
$$S^2_{y_dU_h} = \frac{1}{N_h-1}\sum_{U_h} \left(y_{dk} - \bar{y}_{dU_h}\right)^2 \qquad (6.1.16)$$

För en total i en redovisningsgrupp kan formel (6.1.13) tillämpas med denna populationsvarians och $V\left(\hat{t}_y\right)$ ersatt av motsvarande varians för en totalskattning för redovisningsgruppen. För medelvärden i redovisningsgrupper divideras nämnare och täljare i högerledet i formel (6.1.13) med N_d^2, där N_d är antalet objekt i redovisningsgruppen. Detta ger

$$n = \frac{\frac{1}{N_d^2}\left(\sum_{h=1}^{H} N_h S_{y_dU_h}\right)^2}{V\left(\bar{y}_d\right) + \frac{1}{N_d^2}\sum_{h=1}^{H} N_h S^2_{y_dU_h}} \qquad (6.1.17)$$

där $V\left(\bar{y}_d\right)$ är variansen för en skattning av medelvärdet i redovisningsgruppen. Exempel 6.1.3 beskriver en tillämpning.

Figur 6.1.1 En redovisningsgrupp *d* som skär över stratum 1–3



EXEMPEL 6.1.3 *IT-användning i företag*
Huvudsyftet med undersökningen av IT-användningen i företag är att skatta andelen företag med vissa egenskaper. Ett exempel är andelen företag som tagit emot beställningar via Internet. Andelar skattas för hela populationen samt för olika redovisningsgrupper definierade av företagens bransch eller storleksklasser mätt i antal anställda. I undersökningen skattas även totaler såsom total e-handel i kronor, men andelar anses vara de viktigaste parametrarna, och val av urvalsdesign och bestämning av urvalsstorlek görs i första hand med hänsyn till att andelarna ska skattas med bra precision.

Urvalet dras som ett stratifierat urval med OSU inom strata. Urvalsramen stratifieras med avseende på branschgrupper och storleksklasser, se tabell 6.1.1. Indelningen i branschgrupper och storleksklasser görs med hänsyn till redovisningsgrupperna. Stratifieringsvariablerna, företagens bransch och storleksklass, definierad av antal anställda, hämtas från urvalsramen som tas fram från den version av FDB som används i mars-omgången av SAMU, se avsnitt 3.4.2.

Tabell 6.1.1 Stratifiering med avseende på branschgrupper och storleks-klasser

	Storleksklasser, antal anställda						
Bransch-grupper (SNI)	10–19	20–49	50–99	100–199	200–249	250–499	500–
15–22							
23–25							
26–28							
29–37							
⋮							
72							
92.1–92.2							
93							

Här beskrivs hur urvalsstorleken i redovisningsgrupperna bestämdes och allokerades till storleksstrata, dvs. storleksklasserna.

Redovisningsgrupper utgörs av såväl storleksklasser som branschgrupper. Beräkningar av urvalsstorlekar görs uteslutande på branschgrupper, eftersom dessa utgör en finare indelning av populationen än vad storleksklasser gör.

Varje branschgrupp som utgör en redovisningsgrupp betraktas i urvalsplaneringen som en population. Precisionen bestäms genom att osäkerhetsmarginalen för en redovisningsgrupp sätts lika med ett värde,

$$1{,}96 \times \sqrt{V(p_d)} = \alpha_d \qquad (6.1.18)$$

där α_d är önskad precision i redovisningsgruppen *d*. Precisionen kan väljas olika i olika redovisningsgrupper men detta har inte tillämpats. I 2006 års undersökning valdes precisionen $\alpha_d = 0{,}07$ för alla redovisningsgrupper. Undersökningens budget hade stor inverkan på valet av precision och därmed urvalsstorlek.

Nu ger (6.1.18) att

$$V(p_d) = \left(\frac{\alpha_d}{1{,}96} \right)^2 . \qquad (6.1.19)$$

Med $\alpha_d = 0{,}07$ blir variansen 0,0013. Kostnaden att undersöka ett företag antas vara lika i alla strata. Urvalsstorleken beräknas med formel (6.1.17) som här blir

$$n_d = \frac{\left(\sum_{h=1}^{H} W_h S_h \right)^2}{V(p_d) + \frac{1}{N_d} \sum_{h=1}^{H} W_h S_h^2} \qquad (6.1.20)$$

där $W_h = \frac{N_h}{N_d}$, N_d är antalet objekt i redovisningsgrupp *d* och S_h^2 är stratumvariansen. N_d räknas som det antal företag i urvalsramen som enligt FDB tillhör den branschgrupp som utgör redovisningsgruppen. Eftersom stratumvarianserna i det här fallet avser andelar tillämpas en stratifierad motsvarighet till formel (4.2.7)

$$S_h^2 = \frac{N_h}{N_h - 1} P_h (1 - P_h) \qquad (6.1.21)$$

där P_h är den okända stratumproportionen. Eftersom undersökningen gjorts flera gånger tidigare skulle man kunna skatta P_h utifrån data från t.ex. föregående års undersökning. I stället för att använda föregående års data för att skatta P_h, valdes $P_h = 0{,}5$, dvs. det värde för vilket S_h^2 maximeras. Anledningen till detta val var att ett antal olika andelar skattas och att dessa inte är lika samt att osäkerheten i skattningarna också kan vara stor.

Redovisningsgrupperna definierade av branschgrupper är stratifierade med avseende på storleksklasser. De tre största storleksklasserna i varje branschgruppsstratum totalundersöks. Alla företag, i ingående branscher, med minst 200 anställda ingår i undersökningen. Dessa företag dras bort från urvalsstorleken n_d som ges av formel (6.1.20). Låt n_d' beteckna den kvarvarande urvalsstorleken som kan utnyttjas för urvalsundersökta strata. Formel (4.2.4) för x-optimal allokering tillämpas för att allokera den totala urvalsstorleken på de i redovisningsgruppen ingående stratumen. Urvalsstorleken i stratum *h* ges av

$$n_h = n_d' \cdot \frac{N_h \cdot S_h}{\sum_{h=1}^{H} N_h \cdot S_h} \qquad (6.1.22)$$

S_h fås av formel (6.1.21) med $P_h = 0{,}5$, och summeringen i nämnaren görs över de i redovisningsgruppen ingående urvalsundersökta stratumen. Med dessa val reduceras (6.1.22) till något som i praktiken ligger mycket nära proportionell allokering

$$n_h = n_d \cdot \frac{\sqrt{\frac{N_h^3}{N_h - 1}}}{\sum_{h=1}^{H} \sqrt{\frac{N_h^3}{N_h - 1}}}. \qquad (6.1.23)$$

Som redan nämnts är det inte bara andelar som skattas, utan även totaler som t.ex. total e-handel i kronor. I bestämningen av urvalsstorlek och allokeringen antas att andelar utgör de viktigaste parametrarna, men eftersom de största företagen, mätt i antal anställda, totalundersöks så tas viss hänsyn även till att totaler ska skattas. De stora företagen har i regel de stora värdena vad gäller t.ex. e-handel, och dessa är viktiga för skattning av totaler. Vissa justeringar i de framräknade urvalsstorlekarna gjordes i 2006 års undersökning. Problem med att redovisa en av storleksklasserna hade förekommit i tidigare undersökningar och därför ökade man urvalsstorleken i den storleksklassen. □

6.1.6 Andra urvalsmetoder och estimatorer

Formel (6.1.13) gäller även för GREG-estimatorn, med rätt val av populationsvarianser inom strata. Se avsnitt 4.2.7 under ”Varianter på optimal allokering” för hur S_{xU_h} ska modifieras för att passa en GREG-estimator.

För andra estimatorer och urvalsmetoder kan man försöka lösa ut *n* ur aktuell variansformel, såsom det är gjort i formel (6.1.13). Om det visar sig krångligt att lösa ut *n* kan man i stället sätta in olika *n* i den variansskattningsformel som användes i senaste produktionsomgången. Det *n* som hamnar närmast precisionsönskemålet gäller.

Eftersom syftet är att få grepp om urvalsstorleken behövs bara en ganska grov variansestimator. För t.ex. ett flerstegsurval räcker det att räkna på variansen som om det första urvalssteget var det enda och låtsas som om klustren som dragits i första steget är de elementära objekten. Man använder formler för motsvarande enstegsurval; om klustren drogs med t.ex.

stratifierat OSU kan man räkna som om urvalet var ett enstegs stratifierat OSU. Det leder till en liten underskattning av variansen. Att det fungerar bra i praktiken beror på att en variansskattning som baseras enbart på det första urvalssteget ändå implicit inkluderar osäkerhet som kommer från de följande stegen. Eftersom det är en vanlig men inte helt korrekt föreställning att ”variansen från det andra steget och följande steg inte betyder något” ges här en genomgång.

Det framgår av Särndal et al. (1992, formlerna 4.3.12 och 4.3.14, s. 137) att den första termen i den väntevärdesriktiga variansskattningen för tvåstegsurval ger en överskattning av bidraget från första steget. Uttryckt i formler är variansen i princip

$$V_{2st}\left(\hat{t}_\pi\right) = V_{PSU} + V_{SSU}$$

där $V_{2st}\left(\hat{t}_\pi\right)$ är variansen för tvåstegsurval och V_{PSU} och V_{SSU} är dess första respektive andra term, och

$$\hat{V}_{2st}\left(\hat{t}_\pi\right) = \hat{V}_{PSU} + \hat{V}_{SSU}$$

är motsvarande skattningar. Då är $\hat{V}_{PSU}$ en överskattning av V_{PSU}. Särndal et al. (1992, s. 140) visar att den underskattning av variansen $V_{2st}\left(\hat{t}_\pi\right)$ man får genom att skatta denna med $\hat{V}_{PSU}$ ofta är betydelselös; formel (4.3.19) i samma bok visar storleken av underskattningen relativt variansen för tvåstegsurval. Om alla kluster dras med lika sannolikhet, som här betecknas med π_I, förenklas formeln till

$$\frac{B\left(\hat{V}_{PSU}\right)}{V_{2st}\left(\hat{t}_\pi\right)} = \frac{\pi_I}{\frac{V_{PSU}}{V_{SSU}} + 1},$$

där $B\left(\hat{V}_{PSU}\right)$ är underskattningen (Särndal et al. 1992, s. 137). Även om $V_{PSU} = V_{SSU}$, vilket sällan händer i praktiken, blir inte underskattningen större än $\pi_I/2$. Andra förenklade variansestimatorer diskuteras i Särndal et al. (1992, avsnitt 4.6).

6.1.7 Nytt urval med annan urvalsmetod

Att byta urvalsmetod kan innebära att det behövs information om sådant som inte kan hämtas ur föregående produktionsomgång. Anta t.ex. att ett OSU drogs förra gången, men den här gången vill man av ekonomiska skäl dra ett stratifierat OSU. Då känner man cv_{yU} från förra gången men behöver cv_{yU_h} för varje stratum; dessa cv_{yU_h} är då okända. Det finns olika trick att ta till. Det enklaste är att helt enkelt gissa. Stratifiering har troligen gjort att cv_{yU_h} är mindre än cv_{yU}; man ansätter något värde som får gälla för cv_{yU_h} i alla strata.

Ett annat sätt är att använda cv_{xU_h} i stället, där *x* bygger på någon variabel som är tillgänglig i ramen eller känd på något annat sätt. Att ta stratifieringsvariabeln precis som den är leder dock till en ”glädjekalkyl”. Inom

strata är ofta variationen med avseende på stratifieringsvariabeln liten; det är så strata är konstruerade. Variationen med avseende på undersökningsvariabeln kommer att vara större. Det framräknade urvalet kommer att bli orealistiskt litet. I stället kan man använda en annan variabel för att räkna ut precisionen (se avsnitt 6.1.10). Då undviker man glädjekalkylen, men det blir oundvikligen en del gissningar i beräkningarna. Som redan nämnts är det lämpligt att göra någon form av känslighetsanalys när man arbetar med bedömningar. I det här fallet provar man lämpligen några olika värden för att se om slutsatserna blir ungefär desamma.

En alternativ ansats är att använda sig av designeffekten, DEFF. Se avsnitt 6.1.3 för en diskussion av DEFF. Anta att ett OSU användes förra gången och att man planerar ett urval som inte är ett OSU, med oförändrat precisionskrav:

1. Bestäm ett rimligt värde på DEFF för dels den nya urvalsmetoden, dels den gamla. Kalla dessa DEFF (ny) och DEFF (gammal).
2. Utgå från det *n* som användes förra gången.
3. Multiplicera *n* med DEFF (ny)/DEFF (gammal). Resultatet är ett riktmärke för en lämplig storlek på det nya urvalet.

Se Kish (1965, avsnitt 8.2) för en vidare diskussion av denna omräkningsmetod.

6.1.8 Nytt urval med andra redovisningsgrupper

Betrakta först fallet då redovisningsgruppens objekt kan identifieras i ramen. Då kan formel (6.1.4) utnyttjas för en enstaka redovisningsgrupp. I det fallet avser cv_{yU} variationskoefficienten för *y* inom redovisningsgrupp *d*. Beteckna den cv_{yU_d}. För att få en uppfattning om storleken på cv_{yU_d} kan en variabel i ramen användas som substitut för *y* (om en lämplig registervariabel finns). Man approximerar alltså variationskoefficienten för *y* med dito för registervariabeln *x*. Liksom i föregående avsnitt finns risk för en glädjekalkyl om redovisningsgrupperna sammanfaller på ett ungefär med stratumen, se vidare i avsnitt 6.1.10.

Det går att säga något allmänt om cv_{yU_d} i förhållande till cv_{yU}, även om alla sådana uttalanden med nödvändighet blir osäkra och ganska svepande:

- Om redovisningsgruppen är slumpmässigt sammansatt blir cv_{yU_d} i genomsnitt lika med cv_{yU}.

  ”I genomsnitt” ska här tolkas som genomsnittet i en tänkt lång serie av redovisningsgrupper. I praktiken kan en slumpmässigt sammansatt redovisningsgrupp ha mindre eller större cv_{yU_d} än cv_{yU}.

- Om kunskapen om cv_{yU_d} är osäker kan man anta att den är lika stor som variationskoefficienten för hela populationen, dvs. $cv_{yU_d} \approx cv_{yU}$.

Variationskoefficienter är dock ofta något mindre för redovisningsgrupper än för hela populationen, varför antagandet ovan tenderar att ge en försiktig uppskattning. Jämför med de variationskoefficienter för variabeln inkomst som gavs i exempel 6.1.2.

- I vissa fall vet man att en redovisningsgrupp är mycket mer homogen än populationen i sin helhet, dvs. $cv_{yU_d} < cv_{yU}$. Då får man försöka gissa hur mycket mindre cv_{yU_d} är än cv_{yU}. Det kan gälla t.ex. för *storleksklasser* i företagsundersökningar; se exempel 6.1.2.

Anta att önskemålet är att skattningar ska tas fram för *två* nya redovisningsgrupper samt för hela populationen. För hela populationen har man kommit fram till att urvalsstorleken *n* behövs. För varje redovisningsgrupp krävs det också urvalsstorleken *n*, om $cv_{yU_d} \approx cv_{yU}$ och precisionskravet *r* är lika högt för alla tre skattningarna. Anta att kravet på relativ precision *r* är detsamma i alla redovisningsgrupper och att objekten inte är väsentligt mer lika varandra inom respektive redovisningsgrupp än mellan grupperna. Då behövs det, generellt, i runda tal *k* gånger så stort urval för *k* redovisningsgrupper som för *en* redovisningsgrupp.

Om det inte går att identifiera en redovisningsgrupp i förväg, men man vet ungefär hur stor proportion av populationen som hör till den, kan man uppdatera urvalsstorleken till att bli $n = n_0 / P_d$, där n_0 är lämplig storlek på ett urval som ska ge skattningar för hela populationen och P_d är andelen av populationen som hör till aktuell redovisningsgrupp. Om P_d är liten blir alltså *n* mycket stor. Alternativt, om *n* inte görs tillräckligt stor, kommer risken finnas att det blir alltför få utvalda objekt som hör till *d*.

Går det att identifiera redovisningsgruppen i förväg, exempelvis genom att bilda en indikator för tillhörighet, ska man vanligtvis göra det och stratifiera efter denna så långt det är möjligt. Se även exempel 4.2.2.

6.1.9 Nytt urval med andra statistiska mått

Skattning av differens för godtycklig urvalsmetod
Betrakta en differens som statistiskt mått. Variansen för en differens mellan skattningarna av två *y*-totaler, $\hat{t}_1$ för redovisningsgrupp 1 och $\hat{t}_2$ för redovisningsgrupp 2, är

$$V(\hat{t}_1 - \hat{t}_2) = V(\hat{t}_1) + V(\hat{t}_2) - 2Cov(\hat{t}_1, \hat{t}_2). \qquad (6.1.24)$$

Om $\hat{t}_1$ och $\hat{t}_2$ är oberoende försvinner kovarianstermen. Två situationer då detta är uppfyllt är följande:

- Skattningarna baseras på delurval som är dragna ur var sitt stratum.
- Urvalen är dragna vid två olika tidpunkter oberoende av varandra (se vidare kapitel 5.4).

Anta att man tidigare funnit att urvalsstorleken *n* ger precision *r* för vardera $\hat{t}_1$ och $\hat{t}_2$. Nu vill användaren ha precision *r* även för differensen. Då blir den nödvändiga urvalsstorleken 2*n*. I praktiken kan det vara orealistiskt från kostnadssynpunkt att dubblera urvalet. Då tvingas man diskutera sänkningar av precisionskraven med användaren.

Om $\hat{t}_1$ och $\hat{t}_2$ däremot baseras på delurval som kan överlappa är skattningarna beroende och kovariansen är skild från noll. Överlappande delurval kan uppstå

- vid överlappande redovisningsgrupper
- om redovisningsgrupper korsar strata – se figur 6.1.1.

I praktiken är kovariansen nästan alltid positiv, men ofta liten om överlappningen är liten. I planeringssyfte kan man försiktigtvis sätta

$$V(\hat{t}_1 - \hat{t}_2) = V(\hat{t}_1) + V(\hat{t}_2) \qquad (6.1.25)$$

(Kish 1987, s. 40–41). Principen är att inte hämta ut vinster i form av en stor kovarians innan man vet att den är stor. Även i det här fallet blir den erforderliga urvalsstorleken alltså 2*n* om man inte är beredd att pruta på precisionskraven.

EXEMPEL 6.1.4 *Djurräkningen*
Från Djurräkningen (se exempel 3.2.3, 4.2.1 och 4.2.7) önskades skattningar av djurantal för tre uppsättningar av redovisningsgrupper: län, produktionsområden och stödområden. Endast länsgrupper var beaktade i stratifieringen. Redovisningsgrupperna korsade alltså i många fall stratumgränserna, och de olika uppsättningarna av redovisningsgrupper var överlappande. Skattningarna gjordes med en regressionsansats och ledde preliminärt till olika rikssummor utifrån de olika uppsättningarna av redovisningsgrupper. Estimaten för produktions- och stödområdena justerades därför proportionellt för att få konsistens mellan alla redovisningsgrupperna. □

Skattning av differens för klusterurval
För klusterurval kan variansen snävas in med följande empiriskt grundade formel (Kish 1987, s. 43):

$$\frac{S_{U_1}^2}{n_1} + \frac{S_{U_2}^2}{n_2} < V(\hat{t}_1 - \hat{t}_2) < V(\hat{t}_1) + V(\hat{t}_2), \qquad (6.1.26)$$

där $S_{U_1}^2$ är populationsvariansen för de objekt som hör till redovisningsgrupp 1 och n_1 är antal objekt i urvalet som klassas till den gruppen (motsvarande för redovisningsgrupp 2).

Skattning av andra statistiska mått än differenser
För det mesta kräver totaler större urvalsstorlek än andelar. För medelvärden som skattas som kvoter av totaler – och andra kvotskattningar – räcker det ofta med mindre urvalsstorlek än för totaler. Se t.ex. Särndal et al. (1992, s. 250) för ett exakt villkor för att det ska vara uppfyllt under OSU.

6.1.10 Att skatta populationsvarianser och -totaler för allokering och bestämning av urvalsstorlek

I många av de formler som blir aktuella i urvalsplanering finns populationsvariansen S_{yU}^2 och populationstotalen t_y. Dessa gäller undersökningsvariabeln *y* och är därför okända i planeringsstadiet. I det följande gås igenom metoder som kan ge en förhandsuppfattning om storleken av

dessa parametrar. Ibland behövs även variationskoefficienten $cv_{yU} = S_{yU} / \bar{y}_U$. I många SCB-undersökningar är populationsstorleken *N* känd, åtminstone på ett ungefär. Därigenom kan $\bar{y}_U$ skattas genom $\hat{\bar{y}}_U = \hat{t}_y / N$. I de fall *N* är okänd kan $\bar{y}_U$ gissas genom metoder under rubriken ”Utgå från en tänkt fördelning” nedan.

Populationsvariansen kan behövas för redovisningsgrupper. Det finns några allmänna råd om hur man kan gissa variationskoefficienten för redovisningsgrupper i avsnitt 6.1.8.

Här behandlas några olika situationer:

- **Undersökningen har gjorts förut.** Om det finns ”bra” skattningar av S_{yU}^2 och t_y från en tidigare undersökningsomgång, används dessa. Skattningarna bör vara relevanta för aktuell målpopulation. Dessutom bör skattningar av t_y ha liten eller måttlig varians. En skattning av S_{yU}^2 har ofta stor varians, eftersom värdena på variabeln tas i kvadrat i beräkningen av S_{yU}^2 . Om populationsvariansen behöver skattas i ett stratifierat urval, kan inte estimatorn $s_{yU}^2 = \sum_s \left(y_k - \bar{y}_s\right)^2 / \left(n-1\right)$ användas, eftersom objekten troligen har valts med olika inklusionssannolikheter i olika strata. I stället kan estimatorer i Särndal et al. (1992, avsnitt 5.9) tillämpas.
- **Det finns en hjälpvariabel i ramen.** Om det finns en hjälpvariabel *x* i urvalsramen som man tror är en god approximation av *y*, kan S_{xU}^2 och t_x användas i stället för S_{yU}^2 och t_y . Hjälpvariabeln *x* fungerar som en ”ställföreträdande” undersökningsvariabel. Fördelen med detta alternativ jämfört med föregående är att S_{xU}^2 och t_x inte har någon urvalsosäkerhet. Korrelationen är dock ofta sämre än om man använder skattningar från en tidigare undersökningsomgång. Finns det mer än en hjälpvariabel, går det ibland bra att konstruera en ny variabel genom att t.ex. ta summan eller medelvärdet av två hjälpvariabler (jämför exemplet i avsnitt 7.4).
- **Det finns ingen god approximation av vare sig totalen t_y eller enskilda *y*-värden.** Man använder trots allt den bästa approximation som finns tillgänglig. I t.ex. formel (6.1.13) riskerar man att underskatta $V\left(\hat{t}_y\right)$ och därigenom ansätta en för liten urvalsstorlek. Om man tror att man riskerar att underskatta $V\left(\hat{t}_y\right)$ kan man höja precisionskravet *r* för att kompensera för underskattningen.
- **Det finns mycket liten kunskap om undersökningsvariabeln.** I det här fallet utgår man oftast från budgeten som ger en uppfattning om den maximala urvalsstorleken (se avsnitt 6.1.2). Beställaren bör få en förvarning om vilken precision undersökningen kan tänkas ge. Detta kan åstadkommas på följande sätt. Man kan gissa populationsvariansen (se under rubriken ”Utgå från en tänkt fördelning” nedan) och skatta $V\left(\hat{t}_y\right)$ utifrån populationsvariansen. Om t.ex. urvalsmetoden OSU är i åtanke, skattas $V\left(\hat{t}_y\right)$ med

$$\hat{V}\left(\hat{t}_y\right) = N^2\left(1-\frac{n}{N}\right)\frac{\hat{S}_{yU}^2}{n},$$

där $\hat{S}_{yU}^2$ är en skattning eller gissning av S_{yU}^2.

Även då det finns ganska god förhandskunskap om undersökningsvariabeln kan den behöva uppdateras. Se exempel 6.1.7.

I samtliga fall ovan finns risk för att skattningarna av S_{yU}^2 och t_y får bristfällig kvalitet. Det finns anledning att kompensera för detta enligt försiktighetsprincipen. Vid bestämning av urvalsstorleken höjs kraven något för att kompensera. Vid allokering bör alltid de beräknade urvalsstorlekarna gås igenom manuellt och eventuellt justeras.

En pilotundersökning bör också övervägas i samtliga fall ovan. Behovet av en pilotunderökning är förstås större ju mindre förhandsinformation det finns. Se vidare avsnitt 6.5.

Praktiska problem i beräkningarna
Det finns ytterligare överväganden i beräkningarna när man utgår från en hjälpvariabel:

- Beräkningen av S_{xU}^2 eller dess stratumvisa motsvarighet $S_{xU_h}^2$ är känslig för stora, avvikande värden. Sådana bör tas bort i beräkningen. Annars kan S_{xU}^2 bli extremt stor, vilket leder till ett onödigt stort urval.
- Ofta saknas en del värden på den variabel *x* som ska användas i beräkningen. I sådant fall finns det olika alternativ att överväga:
  - Imputera *x*-värden.
  - Lägg objekten i ett särskilt stratum.
  - Imputera $S_{xU_h}^2$ för vissa strata i stället för att imputera *x*-värden. Det sistnämnda kan ske genom att helt enkelt ”låna” ett $S_{xU_h}^2$ från ett annat stratum eller genom att en variansanalysmodell anpassas till de strata där $S_{xU_h}^2$ föreligger och sedan används för att skatta ett värde på $S_{xU_h}^2$ för övriga strata.

Samma variabel i stratifiering och allokering
Ofta används samma variabel i såväl stratifiering som allokering. Det är inte oproblematiskt, som följande två exempel visar. Det första är artificiellt, det andra är hämtat från den ekonomiska statistiken.

EXEMPEL 6.1.5 *Glädjekalkyl*
Vi tänker oss en population som består av åtta objekt. Vi ska dra ett stratifierat OSU med *n* = 4. Vi har registervariabeln *x* och ska undersöka *y*. Av tabell 6.1.2 framgår att på basis av *x* kan populationen delas in i två homogena strata med fyra objekt i varje. Med stratifiering grupperar vi objekt som är relativt lika med avseende på stratifieringsvariabeln. Som framhölls i avsnitt 4.2 är homogenisering ett viktigt mål vid storleksstratifiering. Men i de flesta fall är spridningen större med avseende på undersökningsvariabeln, som inte är känd på planeringsstadiet. I tabell 6.1.3 syns att populationsvariansen är betydligt större för *y* än *x* inom strata.

Tabell 6.1.2 Ramens utseende samt den okända undersökningsvariabeln *y*

Stratum	X	Y
1	1	1,5
1	1	2,5
1	1,5	0,5
1	1,5	3
2	10	8,5
2	10	10,5
2	10,5	12
2	12	9
Total	47,5	47,5

Tabell 6.1.3 Populationsvarianser inom stratum

Stratum	$S^2_{xU_h}$	$S^2_{yU_h}$
1	0,08	1,23
2	0,90	2,50

Med x-optimal allokering av urvalsstorleken $n = 4$ enligt formel (4.2.2) fås $n_1 = 0{,}93$ för stratum 1 och $n_2 = 3{,}07$ för stratum 2. Dessa urvalsstorlekar avrundas. Spridningen med avseende på x är särskilt liten inom stratum 1 och urvalsstorleken blir liten där. Med Neymanallokering baserat på undersökningsvariabeln y blir $n_1 = 1{,}65$ och $n_2 = 2{,}35$. Det här fiktiva exemplet visar på faran med att använda samma hjälpvariabel i såväl stratifiering som allokering.

Anta i stället att ett precisionskrav är givet och det gäller att beräkna en urvalsstorlek som ska uppfylla den precisionen. Säg att den relativa osäkerhetsmarginalen $r = 1{,}96 \times CV\left(\hat{t}_y\right)$ ska sättas till 0,06, dvs. att man efter avslutad undersökning ska kunna leverera en totalskattning som har precisionen ± 6 procent. Beräkning av urvalsstorleken på basis av x med formel (6.1.4) med detta precisionskrav ger $n = 4{,}1$. Men samma beräkning baserad på undersökningsvariabeln y blir $n = 6{,}8$. Eftersom populationsvarianserna inom strata för x är för små, blir den skattade urvalsstorleken för liten. Den relativa osäkerhetsmarginalen blir inte önskade 0,06 utan 0,15. Om vi i vår stratifiering är framgångsrika vad beträffar homogenisering, kan vi alltså göra oss skyldiga till en glädjekalkyl om vi beräknar urvalsstorlek enbart på basis av stratifieringsvariabeln. □

EXEMPEL 6.1.6 *Utrikeshandel med tjänster*
(Fortsättning på exempel 4.2.2.) Branschgrupper och den av SCB skapade variabeln ”aktör” definierar kategoristrata. Som storleksvariabel för stratifiering inom kategoristrata används summan av omsättning och tjänsteexport i momsregistret för senast tillgängliga tolvmånadersperiod. En urvalsstorlek bestäms i varje branschgrupp och för varje värde på aktör. Urvalet allokeras till storleksstrata med Neymanallokering. Som allokeringsvariabel hade man kunnat tänka sig att använda stratifieringsvariabeln. Emellertid används omsättning från en tidigare period än i stratifieringen. En anledning är att stratifieringen går väldigt långt i homogenisering inom vissa strata där S_{xU_h} är liten. Med x-optimal allokering får man då alltför små urvalsstorlekar. Att i stället använda omsättning från en tidigare period som allokeringsvariabel ger erfarenhets-

mässigt bättre allokering. Den allokering formlerna ger granskas även manuellt. □

Utgå från en tänkt fördelning
Om man vet hur fördelningen av en variabel ser ut i princip, kan man beräkna eller gissa dess medelvärde och varians. Figur 6.1.2 visar en provkarta över fördelningar som en variabel kan ha. I det följande ges några kommentarer till de fall som är vanligast på SCB:

- Ett välkänt exempel är fall A, en dikotom variabel som har variansen $S_{yU}^2 \approx P(1-P) = PQ$, se formel (4.2.8), och medelvärdet P.
- Fall B är en diskret variabel med fem kategorier och lika sannolikhet för varje kategori. Om kategorierna antar värdena 0–4 (alternativt har etiketterats på detta sätt) är variansen 2. Allmänt för kategorierna 0–k är variansen $k^2/12+k/6$ och medelvärdet $k/2$. För k = 1, dvs. då man har endast två kategorier, sammanfaller fall A och B om $P = 1/2$.
- Ett läroboksexempel är den likformiga fördelningen i fall C. Denna har variansen $k^2/12$ och medelvärdet $k/2$, där k är bredden (som är normerad till 1 i figuren).
- Triangelfördelningen i fall F har variansen $k^2/18$ och väntevärdet $k/3$.
- Fall G–J är blandade fördelningar där varje objekt med sannolikhet P tas från den ena fördelningen och med sannolikhet $Q = 1–P$ från den andra. Känner man varians och medelvärde för båda fördelningarna kan man räkna ut variansen för den blandade fördelningen med formeln

$$V(Y) = PV(Y|\text{hämtas från fördelning A}) + QV(Y|\text{hämtas från fördelning B}) + PQ\left[E(Y|\text{hämtas från A}) - E(Y|\text{hämtas från B})\right]^2. \qquad (6.1.27)$$

  En härledning av formel (6.1.27) finns i Bilaga A. Populationsvariansen S_{yU}^2 kan betraktas som en skattning av modellvariansen $V(Y)$. Se ett exempel på uträkning i exempel 6.1.7.
- Fall I är en likformig fördelning med ”spik i nollan”, dvs. en blandad fördelning där i genomsnitt P objekt får värdet noll och Q objekt får värde hämtat från den likformiga fördelningen. Den blandade fördelningen har variansen

$$S_{yU}^2 = \frac{k^2 Q}{12}(1+3P). \qquad (6.1.28)$$

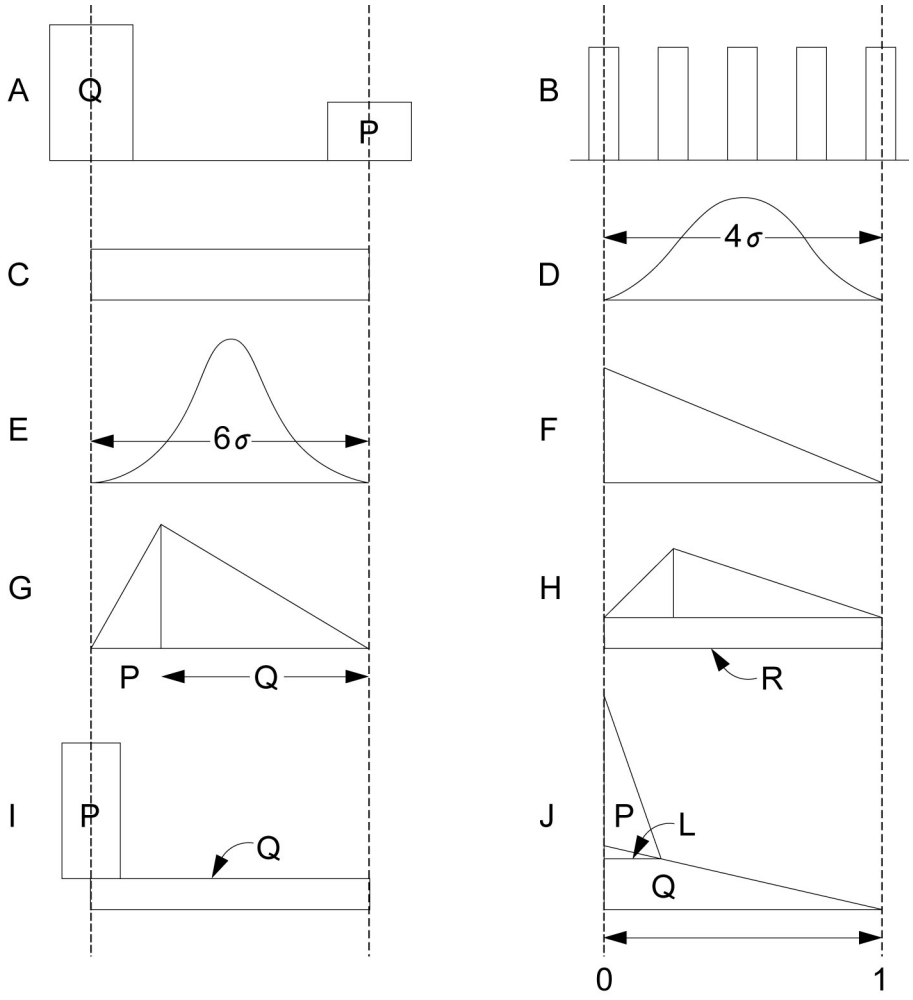
  Medelvärdet är $kQ/2$.
- Fall J är vanligt i företagsstatistik. Fördelningen har populationsvariansen

$$S_{yU}^2 = \frac{k^2}{18}\left[Q + PL^2 + 2PQ(1-L^2)\right]. \qquad (6.1.29)$$

  Medelvärdet är $k(Q+PL)/3$.

Lägg märke till att variansen är störst för de jämnaste fördelningarna. Sålunda har C större varians än F, som har större varians än I.

Figur 6.1.2 Principskisser över några vanliga, ändliga fördelningar. Från Kish (1965, s. 262) efter omarbetning



Hur kan man använda de här stiliserade fördelningarna? Man kan ofta förstå ungefär hur fördelningen borde se ut, även om variabeln är okänd. Även om man inte vet vilket fall som passar bäst, kan man misstänka att den verkliga fördelningen ligger mellan två av de stiliserade fördelningarna. Exempelvis borde antal vakanser hos företag och organisationer se ut ungefär som I, men ha mindre varians än I. Se följande exempel.

EXEMPEL 6.1.7 *Antal vakanser*
Variabeln antal lediga jobb hos arbetsställen har en ”spik i nollan” som representerar arbetsställen som inte har lediga jobb, dvs. fält P i fall I. Hos de arbetsställen som har lediga jobb, kan antalet vakanser tänkas vara fördelat ungefär som fält Q i fall J. Variansen för en sådan blandad fördelning kan räknas ut med formel (6.1.27) och fall F. Anta att 95 procent av företagen saknar lediga jobb. Väntevärde och varians inom den gruppen är noll. Anta vidare att de som har lediga jobb utgör en triangelfördelning som går från 0 till 10. Den har väntevärde 10/3 och variansen 100/18 enligt fall F. Då är den blandade fördelningens varians enligt formel (6.1.27)

$$V(Y) = 0{,}95 \cdot 0 + 0{,}05 \cdot 100/18 + 0{,}05 \cdot 0{,}95\left[0 - 10/3\right]^2 \approx 1{,}9 \approx S_{yU}^2 .$$

Väntevärdet för den blandade fördelningen är

$$\bar{y}_U = 0{,}95 \cdot 0 + 0{,}05 \cdot 10/3 = 0{,}17.$$

Säg att man vill dra ett OSU och få en medelvärdesskattning av typen $0{,}17 \pm 0{,}1$, dvs. dubbla medelfelet ska vara ett tiondels ledigt jobb, alltså $2\sqrt{V(\bar{y}_U)} = 0{,}1$. Eftersom $r = 2\sqrt{V(\bar{y}_U)}/\bar{y}_U$, sätts $r\bar{y}_U = 0{,}1$. Formel (6.1.5) ger

$$n = \left(\frac{2S_{yU}}{r\bar{y}_U} \right)^2 = 760.$$

I det här fallet kan en mer tillförlitlig uppgift på S_{yU}^2 fås från SCB:s vakansstatistik. Trots att den uppgiften finns kan man ändå ha nytta av de stiliserade fördelningarna. Det kan t.ex. finnas anledning att tro att andelen arbetsställen med vakanser har ökat sedan den senaste undersökningsomgången. Då går det att räkna om S_{yU}^2 med antaganden om aktuella *P* och *Q*. □

6.1.11 Bortfall och övertäckning

Bortfall har två effekter: dels blir det ”egentliga” urvalet mindre än vad det var tänkt, dels kan de svarande utgöra ett skevt urval. Den senare effekten lämnas därhän i det här sammanhanget. Som nämndes i avsnitt 6.1.2 kan det vara bra att i urvalsplaneringen avsätta resurser för en bortfallsstudie. (För justering av bortfallets snedvridande effekt i estimationen, se t.ex. Särndal och Lundström, 2005, för kalibreringsansatsen.)

Bortfallet reducerar den ”egentliga” urvalsstorleken. Man kan göra ett antagande om hur stort det sammanlagda objektbortfallet och partiella bortfallet blir för en viss, viktig undersökningsvariabel. Anta att man misstänker att det sammanlagda bortfallet kommer att bli 20 procent. Då dividerar man *n* med 0,8 för att få en justerad urvalsstorlek.

Övertäckning har också effekten att det ”egentliga” urvalet blir mindre. Däremot slipper man den eventuellt snedvridande effekten. Man kan ta hänsyn till övertäckningen genom att sätta undersökningsvariabelns värde till noll för objekt som utgör övertäckning. Ett annat sätt att se på övertäckning är att redovisa dessa objekt i en redovisningsgrupp som inte är av intresse.

6.1.12 Uppgiftslämnarbörda

Hur uppgiftslämnarbördan påverkar kvaliteten i statistiska undersökningar är inte särskilt väl utrett. Men det är knappast någon djärvare förmodan att i de fall uppgiftslämnarna finner det förenat med stora kostnader och stort besvär för dem att lämna begärda uppgifter, så kan det lätt leda till högt bortfall och stora mätfel. Men det är inte självklart; Stanley McCarthy et al. (2006) hittar inget empiriskt stöd för att stor uppgiftslämnarbörda leder till ökad bortfallsbenägenhet. Tre situationer då man kan misstänka att stor uppgiftslämnarbörda leder till kvalitetsproblem är vid

- lång enkät
- panelundersökningar (se avsnitt 5.4.2)
- undersökningar av företag och organisationer.

Det sistnämnda fallet kan kort kommenteras, då uppgiftslämnarkostnaderna kan bli särskilt betungande för företag. Dels beror detta på att vissa undersökningar kan vara betungande i sig, med begäran om uppgifter som det kan vara arbetskrävande att ta fram. Dels beror det på att mängden undersökningar som de har att delta i kan vara stor, utöver uppgiftskrav från myndigheter för annat än statistik. Det är välbekant att uppgiftslämnarbördan i relation till företagens resurser är särskilt betungande för småföretag, även om de större företagen kan lägga ned mer tid på uppgiftslämnandet mätt i arbetstimmar per företag (Hedlin et al. 2005). Det finns därför ytterligare ett skäl att inte låta bestämningen av urvalsstorlek enligt föregående avsnitt bli alltför mekanisk. Om man kommer fram till att man ska göra stora urval bland små företag, kan det vara lämpligt att justera urvalsstorlekarna så att de små företagen drabbas mindre hårt.

6.1.13 Läs mer

Cochran (1977, avsnitt 2.15) behandlar giltigheten av konfidensintervall i urvalsundersökningar. I avsnitt 6.1.1 återgavs i formel (6.1.1) en tumregel av Cochran för minsta urvalsstorlek n för att normalapproximationen ska vara giltig vid bildande av konfidensintervall. I Dalén (1986) diskuteras andra tänkbara tumregler för val av n och hur dessa kan användas i praktiken. Rao et al. (2003) ger flera tumregler för normalapproximationens giltighet och referenser till artiklar om dessa.

Lessler och Kalsbeek (1992, kapitel 12) sätter upp en allmän totalfelsmodell, där hänsyn tas till urval, mätfel och bortfall. Modellen, som är rent teoretisk, kan användas som ett hjälpmedel för att allokera resurser mellan olika felkällor i en undersökning. Den fordrar dock att man har mycket kunskap om de olika felen: i modellen förekommer olika parametrar som måste skattas.

Petri (2005) ger en översikt av problemet att allokera resurserna i en survey.

6.2 Hur väljer man mellan olika urvalsmetoder?

I det här avsnittet diskuteras beslut som fattas efter det att beställaren har formulerat krav på vad undersökningen ska uppfylla men innan den mer detaljerade planeringen startar. I praktiken är dessa överväganden ofta avgjorda på förhand, men även när så är fallet kan det ibland vara nyttigt att ta ett steg tillbaka och se efter vilka och hur starka motiveringar som ligger bakom.

6.2.1 Kvantitativ eller kvalitativ undersökning?

En statistisk undersökning är *kvantitativ*. I en del sammanhang är det bra att ersätta eller komplettera en kvantitativ studie med en *kvalitativ* studie. För att det ska vara värt att använda kvantitativ metodik och göra en statistisk undersökning måste samtliga kvalitetskomponenter vara under tillräcklig kontroll. Några av de mest relevanta kvalitetskomponenterna i detta sammanhang är

- statistiska mått
- redovisningsgrupper (och då i synnerhet deras storlek)
- tillförlitlighet
- aktualitet

(se vidare Statistiska centralbyrån, 2001a).

Andra faktorer att beakta är statistikproducentens kostnader, som bland annat är en funktion av vilken datainsamlingsmetod som väljs, och uppgiftslämnarbördan.

Om inte kvalitetskomponenter, kostnader och uppgiftslämnarbörda är under kontroll kan det vara bättre att göra en kvalitativ studie. Man kan dock inte få fram kvantitativa resultat ur en kvalitativ studie.

Ofta vill man ändå ha en kvantitativ undersökning, men har anledning att hysa farhågor om huruvida det går att nå viktiga kvalitetskrav, t.ex. acceptabel tillförlitlighet. Då bör någon form av pilotundersökning genomföras först (se avsnitt 6.5).

En kvalitativ studie är även bättre än en kvantitativ då informationsbehovet inte kan operationaliseras i ett frågeformulär eller annan mätmetod som kan utföras på samma sätt för alla uppgiftslämnare. Vissa frågeställningar kan vara för komplexa för en frågeblankett. Val av metod styrs av frågeställningen och perspektivet.

För en pedagogisk diskussion av kvalitativa och kvantitativa undersökningar, se Ribe (1995).

Kvalitativa och kvantitativa metoder ska inte ses som konkurrerande utan som kompletterande. För en undersökning i ett outforskat ämne är kvalitativa metoder lämpliga i planeringsfasen för att ge en djupare förståelse för hur uppgiftslämnarna uppfattar och avgränsar fenomenet som ska undersökas. I ett fall genomfördes fokusgrupper med nyanlända invandrare som genomgått ett introduktionsprogram på Stockholm stads invandrarförvaltning. Syftet var att undersöka hur de hade uppfattat introduktionsprogrammet, se Boynton (2004b). En skriftlig, svenskspråkig enkät hade prövats med samma syfte men inte fungerat. I det här fallet ersatte fokusgrupper en enkät. I andra fall kan någon form av kvalitativ studie vara lämplig som förberedelse av en ny enkät. I både nya och upprepade undersökningar utnyttjas kvalitativa metoder även för att förstå svarsprocessen och för att justera frågor och formulär därefter. Ett exempel finns i Boynton och Mujagic (2002). Metoderna kan även användas efter avslutad undersökning för att förstå resultaten från en kvantitativ studie. I ett fall ville Socialstyrelsen veta vad som låg bakom svaren på en fråga om vårdsökande som hade avstått från vård på grund av kostnaderna (Boynton 2004a). Även studier av användning av statistik har genomförts av SCB med hjälp av kvalitativa metoder, se t.ex. Boynton (2003).

Valet mellan kvantitativ och kvalitativ metod ligger nära valet mellan sannolikhetsurval och icke-sannolikhetsurval. För det senare valet, se avsnitt 4.9. Se avsnitt 6.2.7 för ytterligare referenser till metoder för kvalitativa undersökningar.

6.2.2 Om kvantitativ undersökning: registerbaserad undersökning eller survey?

Ibland ställs man inför valet mellan att göra en registerbaserad undersökning eller en survey. Med en survey menas här en undersökning med egen datainsamling, oavsett om det är en total- eller urvalsundersökning. För en detaljerad definition av begreppet survey, se Lessler och Kalsbeek (1992, s. 1–2). De utmärkande dragen hos en registerbaserad undersökning diskuteras i Wallgren och Wallgren (2007, avsnitt 1.4).

Valet mellan registerstatistik och att göra en undersökning med egen datainsamling är ett val av datakälla. Det valet föreligger naturligtvis endast i det fall då ett register med de efterfrågade uppgifterna existerar, åtminstone i någon version. I ett mycket långsiktigt arbete kan man dock tänka sig att bygga upp nya register.

För- och nackdelar med registerbaserade undersökningar visavi undersökningar med egen datainsamling diskuteras i Wallgren och Wallgren (2004, s. 20, och 2007, kapitel 1). Fördelarna med egen datainsamling är att

- aktuella resultat kan tas fram snabbt
- insamlingsmetod och observationsobjekt kan väljas flexibelt
- variablerna ibland kan mätas med hög kvalitet.

Nackdelarna med egen datainsamling är att

- bortfallet kan bli högt
- mätfelet kan bli högt
- det blir låg kvalitet på skattningar för små redovisningsgrupper (gäller urval)
- den är dyr.

Fördelarna med registerbaserade undersökningar är att

- det inte blir någon extra uppgiftslämnarbörda
- kostnaderna blir låga för statistikproducenten (när systemet är uppbyggt)
- man har goda möjligheter till longitudinella studier
- skattningar kan tas fram för små redovisningsgrupper
- mätfelet ofta är litet, eftersom uppgiftslämnarna enligt Wallgren och Wallgren (2004) ”svarar omsorgsfullt på administrativt viktiga frågor”. I vissa administrativa system har uppgiftslämnarna juridiskt ansvar för att lämna korrekta uppgifter.

Nackdelarna med registerbaserade undersökningar är att

- uppgifterna kan vara inaktuella
- man är bunden till en viss typ av observationsobjekt
- variabler som är mindre viktiga i det administrativa systemet kan ha lägre kvalitet.

Tabell 6.2.1 Jämförelse mellan tre typer av undersökningar. Stjärnor representerar en stark sida hos typen av undersökning, med antalet stjärnor som gradering. Frågetecken betyder att det varierar från fall till fall om det är en stark eller svag sida. Från Kish (1979) efter översättning och omarbetning

	Egen datainsamling		
	Ja		Nej
Egenskaper	Urvals-under-sökning	Total-under-sökning	Register-statistik
Flexibel; möjligt att ställa många och relativt komplicerade frågor	***		
Tillförlitlig (främst måttligt mätfel) och relevant	*		?
Jämförelsevis billig	*		***
Snabb, säsonganpassad, kan utföras när rätt tillfälle erbjuds eller efterfrågan finns, kan repeteras för snabba förändringsskattningar	**		*
Inget urvalsfel; möjligt att producera skattningar för små redovisningsgrupper		**	*
Möjligt att få god täckning och litet bortfall		*	?
Möjligt att definiera målpopulationen och välja definition av observationsobjekt	**	*	

Kish (1979 och 1987, s. 140–142) jämför registerbaserade undersökningar, totalundersökningar och urvalsundersökningar, se tabell 6.2.1. Tabellen lyfter fram den kompletterande karaktären hos de tre typerna av undersökningar. Skälet till att han har en stjärna för totalundersökningar på täckning och bortfall är att poängen med att utföra en totalundersökning ofta är just att få in uppgifter från ”alla” i målpopulationen. Därför dimensioneras ofta resurserna för totalundersökningar så att det målet nås.

6.2.3 Om survey: total- eller urvalsundersökning?

I litteraturen underförstås ofta att en totalundersökning är en storskalig undersökning, t.ex. en folkräkning. Men SCB utför många olika totalundersökningar, varav en stor andel är småskaliga. Ett exempel på det senare är målpopulationen Sveriges kommuner, som undersöks med såväl urvals- som totalundersökningar. Se exempel 6.2.1 för en urvalsundersökning av kommuner.

För att kunna diskutera totalundersökningar bör man göra skillnad på ”stora” och ”små” totalundersökningar. De förra utmärks av lång planeringstid och ett mycket stort antal undersökta objekt. Ofta kan man endast samla in ett fåtal variabler i den sortens undersökningar, på grund av krav på att hålla nere både uppgiftslämnarbörda och producentkostnader. En del totalundersökningar är mycket småskaliga; de kan bestå av ett så litet antal objekt som kostar så lite att undersöka att det inte lönar sig att dra ett urval. Dessa diskuteras inte här.

En av de främsta fördelarna med totalundersökningar är att skattningar kan tas fram för små redovisningsgrupper med god noggrannhet, åtminstone om felkällor såsom bortfall är obetydliga. En annan viktig fördel är att det insamlade datamaterialet kan användas som ram för framtida undersökningar.

Viktiga skäl för urvalsundersökningar är att de kostar mindre, att uppgiftslämnarbördan är mindre och att de går fortare att göra än en (stor) totalundersökning.

Man kan se valet mellan total- och urvalsundersökning som en fråga om resursallokering (avsnitt 6.1.2).

Valet bör inte göras av slentrian. Det finns estimatorer för små redovisningsgrupper som bör övervägas (Rao 2003). Vill man kunna dra underurval ur en totalundersökning, kan man kanske genomföra en urvalsundersökning och dra underurval med tvåfasurval. Även möjligheter att använda registerbaserade undersökningar i stället för totalundersökningar eller som komplement till urvalsundersökningar bör övervägas. Se exempel 6.2.1 för en undersökning där man gått från totalundersökning till urvalsundersökning.

6.2.4 Om survey: införande av en cut-off-gräns?

I många undersökningar finns en vilja att fokusera på en viss del av målpopulationen. Ett extremfall är att inte alls undersöka en del av målpopulationen. Den del av målpopulationen som man inte avser att undersöka avskiljs med en *cut-off-gräns*. Man inför på detta sätt en avsiktlig undertäckning. Här kallas fortsättningsvis denna undertäckning för *nollpopulationen*. Med ”noll” åsyftas att inklusionssannolikheterna är noll för alla objekt i mängden. För nollpopulationen gör man en modellbaserad skattning. Att ignorera bidraget därifrån är också en sorts modellbaserad skattning.

De tre vanligaste skälen till att inte undersöka objekten i nollpopulationen är följande:

- Det vore för dyrt och besvärligt att undersöka dem.
- Uppgiftslämnarbördan skulle bli för stor för dessa objekt om man undersökte dem.
- De skulle inte bidra så mycket till skattningarna om de undersöktes.

På SCB används cut-off-gränsdragning främst då fördelningen för den population man vill undersöka är mycket sned, med många små objekt och färre stora. I företagsundersökningar inkluderas ofta de minsta företagen i en nollpopulation. I många andra länder utgör geografiskt avlägsna eller svårtillgängliga områden nollpopulationer.

Det går inte att ge något generellt råd om var cut-off-gränsen ska dras, eller ens säga något precist om lämpligheten i förfarandet i stort; det beror på hur väl den modellbaserade skattningen går att göra och vilken bias man riskerar. Se vidare Särndal et al. (1992, avsnitt 14.4).

EXEMPEL 6.2.1 *Kvartalsutfall för kommuner*
Undersökningen kvartalsutfall för kommuner var tidigare en totalundersökning över cut-off-gränsen 30 000 invånare. Denna gräns fanns för att hålla uppgiftslämnarbördan nere. Undersökningen omfattade 81 kommuner. Ett Pareto πps-urval med antal invånare som storleksmått infördes och cut-off-gränsen sänktes till 15 000 invånare. Den nya urvalsstorleken var 100 kommuner. Biasen minskade. Att dra ett slumpmässigt urval i stället för att göra en totalundersökning medför ett urvalsfel. Bias och urvalsfel kan sammanfattas med måttet medelkvadratavvikelse, som är varians plus bias i kvadrat. Medelkvadratfelet var mindre för den nya urvalsdesignen än den gamla. □

6.2.5 Om urvalsundersökning: vilken design ska väljas?

Om ram- och populationsobjekt sammanfaller är det möjligt att dra direkturval. Fyra metoder för direkturval är vanliga inom SCB: OSU, stratifierat urval, systematiskt urval och πps-urval. Dessa diskuteras var för sig i avsnitt 4.1–4.4. I det här avsnittet jämför vi metoderna.

Om ram- och populationsobjekten skiljer sig åt, drar man inom SCB vanligen nätverksurval eller klusterurval. Avsnitt 3.2.2 visar grafiskt hur nätverks- och klusterurval förhåller sig till varandra.

Även tvåfasurval förekommer inom SCB. Tvåfasurval är en mycket generell urvalsmetod som kan användas i en mängd olika fall. Dessa täcker både situationer då ram- och populationsobjekt skiljer sig och situationer då de sammanfaller. Tvåfasurval tas upp i avsnitt 4.7.

Att ram- och populationsobjekt skiljer sig åt är i allmänhet inte önskvärt, eftersom man då tvingas till urvalsmetoder som vanligen har större varians än direkturval.

För kontinuerliga populationer finns inga egentliga ram- eller populationsobjekt (se avsnitt 4.8). Även här är det populationens natur som ger typen av urvalsmetod. Klusterurval, nätverksurval och urval ur kontinuerliga populationer drivs alltså fram mer av verkligheten än av producentens önskemål.

Även då det skulle vara möjligt att dra direkturval förekommer det att ett klusterurval dras. Anledningen är att det blir billigare. Inom SCB är det dock ovanligt att klusterurval dras i situationer då direkturval är möjligt.

Nedan diskuteras valet mellan fyra metoder för direkturval. Som har framhållits i avsnitt 4.2 finns det två sorters strata: kategoristrata och storleksstrata. Den förra typen av strata kan kombineras med nästan alla urvalsmetoder. Här fokuseras på storleksstrata.

Valet av urvalsdesign tar på olika sätt hänsyn till hur målpopulationen ser ut och hur dess struktur återspeglas i ramen. I klusterurval utnyttjas ofta naturliga, hierarkiskt ordnade grupper i populationen, t.ex. elever, klasser, skolår, skolor och kommuner. Även vid indelning i kategoristrata utnyttjas populationens naturliga gruppering. En fråga i planeringen av ett urval är

om det finns delar av målpopulationen som skiljer sig från övriga delar av populationen. Det kan t.ex. vara skillnader i homogenitet eller nivåer med avseende på undersökningsvariabler eller svarsbenägenhet. En annan typ av struktur i målpopulationen som kan utnyttjas i urvalsdesignen är samband mellan hjälp- och undersökningsvariabler.

Det är värt att notera att den mängd av urval som kan dras med OSU innehåller som delmängder alla de urval som kan dras med de tre andra metoderna för direkturval. Det man gör med stratifierat urval, systematiskt urval eller *πps*-urval är alltså att kraftigt begränsa mängden av urval som är möjliga under OSU. Med andra ord utnyttjas den information som finns i hjälpvariabler till att skala bort urval som tros vara ineffektiva.

Om undersökningsvariablerna är kvantitativa och om det finns en kvantitativ hjälpvariabel som korrelerar med undersökningsvariablerna är stratifierat OSU eller stratifierat *πps*-urval att föredra framför systematiskt urval eller OSU.

OSU eller stratifierat OSU
I avsnitt 4.2.3 diskuteras fyra skäl för stratifiering: redovisningsgrupper, homogenitet, ämnesskäl och administrativa skäl. Omvänt, OSU är ett alternativ bland annat då inget av de fyra nämnda skälen är särskilt starkt. I det scenariot finns sällan användbar hjälpinformation i ramen. OSU kan vara ett alternativ även då stark hjälpinformation finns. Det inträffar om undersökningen ska uppfylla en mängd olika syften och det inte går att hitta någon bra designkompromiss för dem. I stället används hjälpinformationen i estimatorn. Det är annars ofta fördelaktigt att så långt som det går använda hjälpinformation, även samma hjälpinformation, i såväl design som estimation. Ett annat skäl till att använda OSU kan vara att insamlade data huvudsakligen ska användas för analys (se vidare 4.1.3).

Vid skattningar som avser redovisningsgrupper kan det hända att ett OSU ger högre precision än ett stratifierat urval. Detta kan inträffa då stratifieringen och allokeringen inte är anpassade efter redovisningsgrupperna. Se även exempel 4.2.9.

Ur rent praktisk synvinkel är det värt att notera att stratifiering av en population som innehåller åtskilliga redovisningsgrupper och många storleksstrata kräver en viss arbetsinsats.

OSU eller systematiskt urval
Som konstaterades i avsnitt 4.3.1 kan ett systematiskt urval vara mycket mer effektivt än ett OSU, dvs. ha betydligt högre precision. Där jämfördes variansen för en enkel medelvärdesskattning under antagandet att ramen är sorterad på något fördelaktigt sätt. Jämförelsen haltar dock något, eftersom hjälpinformationen inte används i estimatorn i något av fallen. En annan jämförelse vore OSU mot systematiskt urval med en estimator som i båda fallen använder tillgänglig hjälpinformation. Då blir skillnaden i effektivitet väsentligt mindre.

Ett systematiskt urval kan även vara mindre effektivt än ett OSU. I praktisk utövning är det svårt eller omöjligt att veta vilken situation man befinner sig i. Det beror på att det systematiska urvalets precision är svår att skatta (den går överhuvudtaget inte att skatta väntevärdesriktigt). Således kommer värderingen av det systematiska urvalet till slut an på ens intuition

eller kännedom om populationens struktur. Även bland andra Rosén (1996, s. 24–25) kommer fram till det. Den egenheten hos systematiska urval, som den urvalsmetoden inte delar med någon annan metod som bygger på sannolikhetsurval, kan möjligen förklara varför just systematiska urval verkar framkalla mer debatt mellan statistiker än de andra typerna av sannolikhetsurval som nämns i kapitel 4.

Stratifierat OSU eller *πps*-urval
Erfarenheten säger att *πps*-urval ofta ger aningen bättre precision än stratifierat OSU då man har ett bra storleksmått som är väl korrelerat med undersökningsvariabeln. I undersökningar med flera undersökningsvariabler, där sambandet mellan storleksmått och de olika undersökningsvariablerna varierar, är skillnaden i precision mellan urvalsmetoderna liten, i synnerhet om hjälpinformationen utnyttjas även i estimatorn (Holmberg 2003).

Arbetsinsatsen är i allmänhet större för stratifierat urval än för *πps*-urval. Det gäller åtminstone då man har många redovisningsgrupper och potentiella stratifieringsvariabler.

En vanlig metod för bortfallskompensation är att ändra vikterna. Om hjälpinformation saknas är det vanligt att vid OSU använda rak uppräkning, se avsnitt 2.3.3. Motsvarigheten vid *πps*-urval är att räkna om vikterna med den realiserade urvalsstorleken (dvs. antal svarande) som den ursprungliga. Vid förekomst av bortfall ändras därför vikterna för alla objekt i ett *πps*-urval. Vid stratifierat urval ändras vikten endast stratumvis.

Objekt ändrar storlek över tid. Vid storleksstratifierat urval kan därmed objekt byta stratum mellan två urvalsdragningar. Samordningen av urval över tid störs för det enskilda företaget. Vid *πps*-urval ger inte en mindre ändring i storlek lika drastiska resultat.

6.2.6 Om urvalsundersökning: undersökning av ovanliga företeelser?

När man talar om undersökningar av ovanliga företeelser så tänker man sig i allmänhet någon av följande situationer:

- En rampopulation innefattar en liten mängd objekt med en sökt egenskap. Objekten med den sökta egenskapen utgör målpopulationen för undersökningen.

  En variant är att man är intresserad av objekt med olika egenskaper, som då tillhör olika, små redovisningsgrupper.
- Undersökningen avser en målpopulation där alla objekt är relevanta men utspridda i den fysiska miljön och därför sällsynta.

  Objekten av intresse kan exempelvis vara djur eller växter av någon sällsynt art.

I båda fallen gäller att man inte har möjlighet att på förhand identifiera eller lokalisera objekten i målpopulationen, utan detta måste göras i samband med datainsamlingen. Målpopulationen i det första fallet brukar på engelska kallas *hidden population*, eftersom den är dold i ramen (eller ram saknas). I det följande kallas den *gömd population*.

Anta att man vill kartlägga fritidsfiskets omfattning i Sverige. Det finns inget register över fritidsfiskare, utan man är hänvisad till att använda befolkningsregistret som urvalsram och sedan fråga varje utvald individ om fiskevanor. Detta är ett exempel på en undersökning av en gömd population.

Hur liten ska en målpopulation vara för att betraktas som en gömd population? Purcell och Kish (1979) samt Kish (1987, s. 36) ger tumregler för ovanliga målpopulationers relativa storlek (i förhållande till rampopulationen) och vad storleken betyder för valet av urvalsmetod. Deras storleksklassificering är följande (fritt översatt från engelskan):

- Betydande: mer än 1/10 av rampopulationen.
- Mindre betydande: mellan 1/100 och en 1/10 av rampopulationen.
- Obetydlig: mellan 1/10 000 och en 1/100 av rampopulationen.
- Mycket liten: mindre än 1/10 000 av rampopulationen.

Vad kan göras om den population som ska undersökas är gömd?

Alternativ 1 Gör en urvalsundersökning med en särskild urvalsmetod

- *Använd stratifierat urval med varierande urvalsfraktioner*
  Strategin bygger på att man har förhandsinformation om frekvensen hos de relevanta objekten i olika strata.

- *Använd tvåfasurval*
  Vid tvåfasurval dras ett större urval med främsta syfte att identifiera de ovanliga eller sällsynta objekten; detta kallas ibland för ”screening”. Ett urval av dessa objekt kan undersökas närmare i den andra fasen (se avsnitt 4.7.2).

- *Använd nätverksurval*
  Nätverksurval kan användas för att undersöka ovanliga företeelser; se avsnitt 4.5.3.

- *Använd ”adaptive sampling” eller ”inverse sampling”*
  ”Adaptive sampling” går i princip ut på följande. Man drar ett urval och observerar objekten i detta. Om minst ett av objekten tillhör målpopulationen så drar man ett tilläggsurval ”i närheten” av det relevanta objektet. Med ”närhet” avses ofta – men inte alltid – geografisk närhet. Rocco (2003) ger en metod där man med en viss procedur drar tilläggsurval även om man inte hittar något relevant objekt i sitt initiala urval. Rocco ger också många referenser till området ”adaptive sampling”. För att vara effektiv kräver urvalsmetoden att objekten har en tendens att förekomma nära varandra. Félix-Medina och Thompson (2004) utökar ”adaptive sampling” till att inkludera möjlighet till tvåfasurval. I andra fasen kan man använda t.ex. de observationer som gjorts i första fasen.

  Med *”inverse sampling”* drar man successivt objekt efter objekt slumpmässigt ur hela populationen tills något kriterium uppfyllts, t.ex. att man observerat något förutbestämt antal objekt i målpopulationen. Se Salehi och Seber (2004) för en överblick.

- *Använd multipla ramar*
  Gör en sammanställning av vilka relevanta ramar som finns tillgängliga och gör urval ur dessa (se avsnitt 3.4.5). Se exempel 3.4.4.

Alternativ 2 Avstå ifrån att göra en urvalsundersökning
Om hjälpinformationen är knapphändig, är det mycket svårt att genomföra en urvalsundersökning som avser en gömd population. Framför allt blir undersökningen inte särskilt kostnadseffektiv. Säg att ett OSU av storleken n = 20 000 dras ur ramen. Om målpopulationen är ”mycket liten”, enligt definitionen ovan, kommer urvalet med 95 procents säkerhet att innehålla högst fyra relevanta objekt. En kvalitativ studie kan vara ett bättre alternativ (se avsnitt 6.2.1). I denna situation fungerar inte metoderna i denna handbok.

Alternativ 3 Gör en totalundersökning
Som alternativ till urval vid undersökningar av gömda populationer, kan man överväga att koppla på några frågor till en totalundersökning inom samma ämnesområde.

Alternativ 4 Gör en registerbaserad undersökning
Ibland kan det finnas någon registervariabel som är besläktad med undersökningsvariabeln, och som kan användas i dess ställe. Kunden eller beställaren måste då göra en avvägning mellan olika kvalitetskrav, i första hand relevans och tillförlitlighet.

I flera av ovanstående strategier brukar man använda filterfrågor när objekten är individer eller organisationer. Det innebär att man helt enkelt frågar respondenterna om de har en viss egenskap, t.ex. om de har ägnat sig åt fritidsfiske under fjolåret. De som har egenskapen utgör den gömda populationen och undersöks vidare. Se även fallstudien i avsnitt 7.1!

6.2.7 Läs mer
Kjaer Jensen (1994), Taylor och Bogdan (1998) och Ritchie och Lewis (2005) är introduktionsböcker i ämnet kvalitativa metoder. Även Holme och Solvang (1997) ger en introduktion och diskuterar starka och svaga sidor hos kvalitativa respektive kvantitativa metoder. Wibeck (2000) är framför allt en praktisk handledning i hur man planerar för, utformar och arbetar med fokusgrupper. Kvale (1997) är mer teoretisk; den diskuterar vad som händer i en djupintervju. Den ger även många praktiska råd om hur en bra djupintervju genomförs.

Kalton och Anderson (1986) skriver lättläst om surveyer av sällsynta händelser.

6.3 Hur beaktas tidsaspekten?
Det här avsnittet behandlar olika frågor med tidsanknytning som aktualiseras i urvalsundersökningar. Till dessa frågor hör val av tidpunkter: dels för ramframställning, dels för urvalsdragning. I återkommande undersökningar tillkommer frågan avseende hur ofta urvalet ska göras om.

Tidsaspekten berörs även i kapitel 3 om urvalsramar och i kapitel 5 om samordning över tiden.

6.3.1 Val av tidpunkter för ramframställning och urvalsdragning
Referenstiden för en undersökning avser den tidpunkt eller den tidsperiod som objekt och variabler hänför sig till, se Statistiska centralbyrån (2001a). Exempel på populationer med tydliga referenstider är ”alla personer som

är bosatta i Sverige ett visst datum" eller "alla företag som varit verksamma under ett visst år". På SCB används ofta ett befintligt register som utgångspunkt för ramframställning. Då är valet av *version* av registret betydelsefullt just med tanke på objektens referenstid. Huvudregeln är att välja den registerversion som bäst avspeglar den population som man är intresserad av. Detta påverkas i sin tur av när registret uppdateras med avseende på objekt och variabler. Det är dock, på grund av säsongsvariation, inte alltid som den senaste registerversionen bäst avspeglar populationen. Ytterligare en aspekt på valet av registerversion är eventuell samordning med andra undersökningar (se avsnitt 3.4.2).

6.3.2 Hur ofta behöver man dra urval?

Det är inte helt självklart att man måste byta urval ens i återkommande undersökningar. I analyser av urvalsdata kan det vara av intresse att följa en mängd objekt över tiden, vilket skulle tala för att behålla samma urval. Ofta finns dock även andra syften med en undersökning, såsom att skatta parametrar för en aktuell tidsperiod eller tidpunkt. Skattningar baserade på ett föråldrat urval riskerar att få för dålig kvalitet för att vara användbara. Utöver detta problem, som hänger samman med att populationen förändras, löper man risken att "slita ut" sina uppgiftslämnare:

- **Populationens förändringstakt**

  Om populationen förändrar sig snabbt, vilket exempelvis gäller för företagspopulationer, riskerar man att få stora täckningsfel i sina skattningar om man inte förnyar urvalet ofta. För stabila populationer kan man vänta längre med att byta urval.

- **Utmattningseffekter**

  Ett argument för att byta urval är att utvalda objekt efter en tid riskerar att tröttna på att vara med i undersökningen. Bristande motivation kan leda till att de börjar lämna uppgifter av dålig kvalitet, eller att de väljer att helt enkelt avstå ifrån att medverka i undersökningen.

Ibland finns ingen möjlighet att ta fram en ny, uppdaterad urvalsram när som helst, varför det inte är någon idé att dra urval oftare än t.ex. en gång per år. Det kan handla om att de register som används för framställning av urvalsramen endast uppdateras en gång per år. I företagsundersökningar som handlar om ekonomiska variabler måste en del objekt alltid ingå, nämligen de med mycket stora värden på efterfrågade variabler. Detta är ett exempel på att man inte alltid har möjlighet att undvika utmattningseffekter genom urvalsbyten. I sådana fall får man arbeta med olika motivationshöjande åtgärder.

I *årsundersökningar* är målet att undersöka en population ett givet år. De flesta populationer förändras så pass mycket under ett år att det är rimligt att dra ett nytt urval inför varje undersökningsomgång, dvs. en gång om året. I *månads- och kvartalsundersökningar* kan urval dras inför varje undersökningsomgång, dvs. varje månad eller varje kvartal, eller för flera undersökningsomgångar, exempelvis en gång per år eller två gånger per år. I vissa av SCB:s *korttidsundersökningar* dras urval två eller flera gånger per år, medan det i andra dras urval endast en gång per år, se exempel 6.3.1–6.3.3.

6.3.3 Korttidsundersökningars speciella villkor

I en månads- eller kvartalsundersökning där samma urval används t.ex. under ett helt år ökar täckningsproblemen med tiden. I populationer som förändras snabbt blir täckningsproblemen ofta märkbara vid periodens slut. Övertäckning kan i många fall identifieras och korrigeras för i skattningarna. Ibland kan det dock vara svårt att avgöra vilka objekt som är övertäckning och vilka som är bortfall. Undertäckning är oftast betydligt svårare att hantera. Olika sätt att minska effekten av täckningsproblemen är att

1. dra urval oftare (från en uppdaterad urvalsram)
2. korrigera för täckningsproblem i skattningarna
3. göra tilläggsurval
4. lägga till identifierad undertäckning.

Att byta urval oftare kan låta som ett enkelt sätt att minska täckningsproblemen, men det medför också en hel del arbete. Nya uppgiftslämnare måste kontaktas och ”läras upp”, och bortfallet blir i regel större då urvalet används för första gången. Mätfelen kan också bli större med nya uppgiftslämnare. Urvalsbyten innebär också att ett objekt åker in och ut ur urvalet oftare. För uppgiftslämnare kan det vara enklare att ingå i ett urval en längre period, för att sedan slippa ingå en längre period, än att perioderna är kortare.

Att korrigera för över- och undertäckning i skattningen är ibland möjligt. I undersökningen ”Omsättning inom detaljhandeln” görs försök att korrigera för undertäckning genom att kalibrera, se exempel 6.3.2. Om det går att korrigera för undertäckning i skattningen med hjälp av t.ex. registerinformation, så är detta en lösning som ofta är att föredra. Problemet är att det i de flesta fall inte finns tillräckligt bra hjälpinformation att tillgå. Korttidsundersökningar görs för att få en så aktuell statistik som möjligt, och registeruppgifter, som skulle vara möjliga att använda som hjälpinformation, är i dessa sammanhang oftast för gamla när de väl finns tillgängliga.

Tilläggsurval syftar till att undersöka den del av populationen som inte kunde nås vid urvalstillfället, dvs. den del som annars tillhör undertäckningen. Det kan t.ex. handla om företag som är nya eller individer som bosatt sig i Sverige.

Det förekommer att man bildar ett stratum med identifierad undertäckning som totalundersöks. I undersökningen ”Investeringsenkäterna” fångas ”nya” stora investerare upp under året och dessa totalundersöks. Endast en del av undertäckningen kan på detta sätt fångas upp. I lantbruksstatistiken fångas nybildade företag upp, se exempel 6.3.3.

EXEMPEL 6.3.1 *Kortperiodisk sysselsättningsstatistik*
Kortperiodisk sysselsättningsstatistik är en kvartalsundersökning. Objekten är arbetsställen. De parametrar som skattas är totala antalet anställda och förändring av antal anställda från föregående kvartal. Resultaten redovisas uppdelade på sektor, bransch och storleksklass. Urval till undersökningen dras i SAMU. Urval dras två gånger per år, i mars avseende kvartal 2 och 3, och i augusti avseende kvartal 4 och 1 nästa år. Urvalen dras från en ram som tas fram från en uppdaterad version av FDB. □

EXEMPEL 6.3.2 *Omsättning inom detaljhandeln*
Undersökningen ”Omsättning inom detaljhandeln” är en månadsundersökning. Den enda variabel som samlas in är omsättning. Urvalet dras en gång per år, i SAMU-systemet, och urvalet för år t dras i mars år t. Samma urval används i tolv månader, och eftersom företagspopulationer förändras snabbt uppstår täckningsproblem, åtminstone i slutet av året. Ett projekt pågår som syftar till att reducera den bias som uppkommer på grund av detta. En kombinerad kvotestimator med årsomsättning som hjälpvariabel har använts för att skatta total omsättning i redovisningsgrupperna som är definierade av branscher. Genom att använda en mer aktuell version av omsättning som hjälpinformation i skattningen kan den bias som uppkommer på grund av att urvalet blir föråldrat åtminstone delvis reduceras, se Lindblom och Nordberg (2004). □

EXEMPEL 6.3.3 *Lantbruksregistret*
Från Lantbruksregistret (LBR), som tidigare konstruerades av SCB men numera tas fram av Statens jordbruksverk, framställs statistik över arealer av olika grödor, antal husdjur av olika slag, m.m. Vissa uppgifter fås via urvalsundersökning. Lantbrukare som slutar med eller delar upp verksamheten anmodas att ange detta och att ange om någon har tagit över verksamheten samt vem detta är.

När SCB framställde LBR kontaktades den nya brukaren och klarlades vilka objekt (lantbruksföretag) som var nya. Alla kända nya objekt, dvs. den identifierade undertäckningen, lades i ett eget ”ny-stratum” och totalundersöktes. Numera används en något modifierad rutin för att hantera nya lantbruksföretag. □

6.4 Hur används hjälpinformationen?

Med hjälpinformation menas här den kännedom som finns om objekten i populationen (eller urvalet) innan eller efter att urvalet dras. Den senare informationen kan erhållas genom t.ex. en registeruppdatering som skett efter det att urvalet drogs. *Hjälpvariabler* är registervariabler som kan utgöra en ”hjälp” till att förbättra urvalsdesign och skattning. För att vara till nytta ska hjälpvariablernas värden helst finnas tillgängliga för hela populationen. En *hjälpvektor* består av en eller flera hjälpvariabler. En *hjälptotal* är en känd summa av variabelvärdena för en hjälpvariabel. Termen *hjälpinformation* åsyftar hjälpvektorn och hjälptotalerna.

I avsnitt 2.2.3 och 3.2.4 nämndes nyttan av hjälpinformation. Ju mer kraftfull hjälpinformation man har, desto större möjligheter har man att utforma en effektiv urvalsdesign. OSU och vissa former av systematiskt urval är de enda av de urvalsmetoder som beskrivs i kapitel 4 som inte använder hjälpinformation.

En relevant hjälpvariabel är en variabel som kan förmodas spegla redovisningsgrupper eller ha något annat samband med det man vill undersöka. En hjälpvariabel som är bra att använda i urvalsdesignen bidrar med något av följande:

1. Information om vilken redovisningsgrupp ett objekt troligen tillhör.
2. En ”förhandsuppfattning” om målvariablernas värden.

3. En uppfattning av risken att det utvalda objektet kommer att utgöra bortfall i undersökningen.
4. En uppfattning om kostnaden för att samla in information från ett observationsobjekt.

Hjälpinformation kan ofta användas i såväl urvalsdesign som estimation.

I allmänhet har en undersökning flera målvariabler och flera redovisningsgrupper. En viss hjälpvariabel kan ha ett starkt samband med en viss målvariabel, men knappast något samband alls med en annan. Valet av hjälpvariabler kan vara relativt svårt, eftersom man behöver prioritera mellan olika målvariabler och tillhörande parametrar.

Såsom nämndes i avsnitt 3.2.4, kan hjälpinformationen avse olika typer av variabler:

- kategorivariabler (t.ex. region och kön, eller bransch i företagsstatistik), som kan användas för att dela in populationen i strata eller redovisningsgrupper.
- storleksvariabler, som kan användas för stratifiering i storleksgrupper eller för att ta fram storleksmått för πps-urval.

En komponent av hjälpinformationens kvalitet är korrelationen mellan hjälp- och målvariabler. Styrkan i sambandet säger mycket om hjälpinformationens relevans och aktualitet. Relevansen avser hur ”besläktad” hjälpvariabeln är med målvariabeln.

Det gäller att utnyttja så mycket registerinformation som möjligt utan att gå till överdrift. Man bör noga eftersöka relevant information från olika register och lägga in informationen i ramen. Det innebär att man tänker efter om det finns information ”någonstans där ute” som kan ha bäring på punkt 1–4 ovan. Därefter ska man bedöma styrkan i den identifierade hjälpinformationen och försöka hitta de ”starkaste” hjälpvariablerna. Om undersökningen har gjorts tidigare, kan man beräkna korrelationer mellan mål- och hjälpvariabler (för kvantitativa variabler). I fallet med kvalitativa variabler kan man beräkna andelar med olika egenskaper enligt mål- och hjälpvariablerna. Vidare är det lämpligt att göra bortfallsanalyser, för att se skillnaden i svarsfrekvens för olika värden på hjälpvariablerna, t.ex. olika åldersgrupper eller inkomstklasser för individer. Exempel på hur ett sådant resonemang förs i praktiken finns i Lundström och Särndal (2001, kapitel 3).

Det är vanligt att man använder fjolårets registeruppgifter som hjälpinformation. Man får avgöra från fall till fall om uppgifterna är tillräckligt relevanta och korrekta för att vara användbara. Fel i hjälpinformationen kan leda till dåliga skattningar.

EXEMPEL 6.4.1 *Djurräkningen*
Djurräkningen (se exempel 3.2.3, 4.2.1, 4.2.7 och 6.1.4) avsåg antalet nötkreatur, svin, får och höns i juni 2001. Ramen innehöll djurantalen i augusti 2000 för de flesta lantbruksföretagen. Korrelationen mellan antalet av ett visst djurslag 2001 och 2000 var hög. Inför urvalet stratifierades rampopulationen efter storleksklasser för de olika djurslagen och efter en regional indelning. Urvalsstorleken var 15 100 företag, utav 77 000 i ramen. Antalet strata var 98, varav 40 totalundersöktes. Estimationen utfördes med generaliserade regressionsskattningar (GREG-estimatorn). För respektive

djurslag i en viss redovisningsgrupp (region) utnyttjades enbart antalet djur och antalet företag med djurslaget i redovisningsgruppen enligt ramen. Därmed fick ett och samma företag olika vikter för olika djurslag och för olika kategorier av regioner. Hjälpinformationen utnyttjades alltså både i designen och i estimationen, och ledde till en väsentligt förbättrad precision i skattningarna av antalet nötkreatur, svin, får och höns. □

Det är naturligt att ställa sig frågan om huruvida man ska **använda hjälpinformationen i urvalsdesignen eller i estimationen eller i båda dessa moment**. Detta beror på vilka felkällor man vill bemästra; nedan beskrivs fem fall:

Systematiska bortfallsfel (bortfallsbias) kan reduceras genom att man stratifierar efter homogenitet i svarsbenägenhet, dvs. så att svarsfrekvenserna kan förväntas bli olika i olika strata, men relativt lika mellan grupper inom strata. Alternativt kan man utnyttja hjälpinformationen i estimationen, genom att kalibrera med hjälpvariabler som speglar svarssannolikheterna för olika grupper (se vidare Lundström och Särndal, 2001). Detta är särskilt värdefullt om man inte har tillräckligt stort urval för att kunna stratifiera även efter svarsbenägenheten.

Genom att utnyttja *sambandet mellan hjälp- och målvariabler* kan man höja precisionen (*minska urvalsfelet*) i skattningarna. I urvalsdesignen kan detta framför allt ske genom stratifiering efter storleksgrupper eller genom πps-urval, där ett storleksmått utnyttjas. I estimationen kan detta göras genom att utnyttja storleksvariabeln som en hjälpvariabel, t.ex. i en kalibreringsestimator. Det är flexiblare att använda sambanden i estimationen än i designen, eftersom man ibland kan skräddarsy skattningarna för olika parametrar genom att utnyttja olika viktsystem (se exempel 6.4.1). Det finns dock inga teoretiska hinder för att använda samma hjälpinformation i både urvalsdesign och estimation. Vissa precisionsvinster kan ibland uppnås genom att utnyttja informationen även i estimationen.

Som nämndes i avsnitt 4.2 kan *urvalsfelet också minskas* genom att man *identifierar redovisningsgrupper* genom hjälpinformationen. I urvalsdesignen görs detta genom att man stratifierar efter viktiga redovisningsgrupper och allokerar tillräckligt många urvalsobjekt till varje stratum. I estimationen kan man införliva hjälpvariabler som identifierar viktiga redovisningsgrupper. Hjälpinformationen bör primärt användas redan i designen, eftersom det endast är där man kan se till att man får tillräckligt med observationer i redovisningsgrupperna.

Ett fjärde fall är att man vill *uppnå konsistens* mellan å ena sidan skattningarna från en urvalsundersökning och å andra sidan uppgifterna från ett register eller en totalundersökning. Detta kan bara åstadkommas i estimationen, nämligen genom att man kalibrerar mot de summor (hjälptotaler) som man vill uppnå konsistens med. Det finns ett egenvärde i konsistens, men konsistens kan också ge lägre bortfalls-, täcknings- och urvalsfel.

Täckningsfel kan reduceras genom att man förbättrar ramens kvalitet. Därtill kan man i vissa fall i estimationen minska täckningsfelen genom att kalibrera med hjälpvariabler som identifierar målpopulationen, t.ex. genom en uppdaterad version av ramen eller möjligen genom en säkrare summa från en annan undersökning eller ett annat register. Se vidare

Lundström och Särndal (2001, kapitel 11) eller Särndal och Lundström (2005, kapitel 14).

En fråga som kan uppstå är hur man ska utnyttja kvantitativ hjälpinformation. Bör man **klassindela hjälpvariabeln eller inte**, i urvalsdesignen respektive estimationen? I urvalsdesignen motsvarar klassindelning att man stratifierar efter storleksgrupper, medan motsatsen är att man t.ex. drar ett πps-urval efter storleksmåttet.

En fördel med att klassindela är att man minskar risken för extrema uppräknade värden i estimationen. Användningen av hjälpinformationen blir robustare. En nackdel med att klassindela är att man inte utnyttjar hela hjälpinformationen för variabeln i fråga, utan reducerar den. (Om man har flera ungefär lika viktiga variabler, kan man dock ändå inte utnyttja hela hjälpinformationen för var och en av variablerna.)

En ofta lämplig metod är att klassindela, dvs. stratifiera, i urvalsdesignen och sedan använda den kvantitativa variabeln oförändrad som hjälpvariabel i estimationen. Om man i hjälpvektorn lägger in en indikator för stratum och värdet för den kvantitativa variabeln, erhålls en enkel regressionsskattning inom varje stratum.

6.5 Hur utformas en pilotundersökning?

Det finns inte mycket litteratur om generella metoder för *pilotundersökningar* (även kallade *pilotstudier* eller *provundersökningar*). Det beror dels på att pilotundersökningar har varit ett försummat forskningsområde, dels på att pilotundersökningar snarare är en bred flora av olika studier än en speciell sorts undersökning. Även terminologin är skiftande. Biemer och Lyberg (2003, s. 365) föreslår följande termer:

Pilotundersökning: En urvalsundersökning (survey) som görs i god tid före huvudundersökningen med syfte att förbättra denna. Den har en urvalsmetod som är så vald att aktuella frågor belyses på bästa sätt. Den syftar oftast till kvantitativa resultat.

Test (engelska *pretest)*: Mindre studier som görs före huvudundersökningen för att t.ex. testa blanketten. De är ofta kvalitativa (se avsnitt 6.2.1 om kvalitativa studier).

Feasibility test: Tester som görs då det finns anledning att undersöka om en metod är genomförbar. Man kanske vill testa en helt ny datainsamlingsmetod. På svenska skulle *feasibility test* kunna kallas genomförbarhetstest.

Generalrepetition (genrep): Den främsta skillnaden mellan genrep och pilotundersökning är att genrepet görs tätt inpå huvudundersökningen som en realistisk miniatyr av denna. Om inget oväntat händer kan dess data användas tillsammans med huvudundersökningens.

Gränserna mellan dessa typer av studier eller aktiviteter är inte skarpa.

Hur en pilotundersökning utformas beror bland annat på budget, tidsram, vad som ska undersökas och vilken typ av information som ska komma fram, t.ex. vilken tillförlitlighet som krävs. Sannolikhetsurval behöver inte vara nödvändigt för denna typ av undersökningar. Biemer och Lyberg (2003, s. 366) listar tjugo områden som har undersökts på senare tid med

pilotundersökningar, bland annat uppgiftslämnarbörda, effektivitet av metoder som nedbringar bortfall och storlek av populationsvarians.

Om pilot- och huvudundersökningen ligger nära varandra i tiden och variabeldefinitionerna är tillräckligt lika, kan man överväga att använda data från pilotundersökningen i de slutliga skattningarna. Då kan det vara lämpligt att låta de objekt som pilotundersökningen omfattade bilda ett totalundersökningsstratum (se avsnitt 4.2).

Se avsnitt 7.5 för ett exempel på en pilotundersökning. En potentiell stratifieringsvariabel, tillstånd för utrikes körning, undersöktes med en särskild enkät som gick till 70 företag i transportbranschen.

6.6 Hur kontrolleras om urvalet är korrekt draget?

Innan blanketter e.d. skickas ut, bör man för säkerhets skull kontrollera att urvalet är korrekt draget. Följande kontroller är värda att utföra:

- En första kontroll är att summera inklusionssannolikheterna för alla ramobjekt som har positiv inklusionssannolikhet. För direkturval som dras med OSU och urval med varierande inklusionssannolikheter på något av de sätt som beskrivs i avsnitt 4.4 ska inklusionssannolikheterna summera sig till urvalsstorleken (Särndal et al. 1992, s. 38). För stratifierade direkturval ska inklusionssannolikheterna summera sig till urvalsstorleken inom strata. För flerstegsurval (se avsnitt 4.6) ska inklusionssannolikheterna i det första urvalssteget summera sig till antalet utvalda kluster.
- Om det finns någon variabel i ramen kan man jämföra medelvärdet av denna över alla objekt i urvalsramen med samma medelvärde för ramobjekten i det dragna urvalet. Dessa medelvärden bör ungefär sammanfalla under OSU. För stratifierat OSU bör medelvärdena inom strata för ram respektive urval ungefär överensstämma. För urval med varierande inklusionssannolikheter (se avsnitt 4.4) måste medelvärdet först räknas ut med designvikter för att kunna jämföras med motsvarande medelvärde i ramen.
- Om kalibrering ska användas i estimationen, ska kalibrerade skattningar av totaler med avseende på hjälpvariabler stämma exakt med motsvarande ramtotaler (Särndal et al. 1992, Remark 6.5.1 på s. 234).

7 Fallstudier

Kapitel 7 innehåller fem större exempel, hämtade från undersökningar som genomförts av SCB: olyckor i lantbruket, hushållens ekonomi, arbetskraftsundersökningarna, företagens ekonomi samt inrikes och utrikes trafik med svenska lastbilar. Fokus ligger på urvalsmetoderna.

7.1 Olycksfall i lantbruket

En undersökning om olycksfall i lantbruket under 2004 genomfördes av SCB. Dessförinnan gjordes en liknande studie avseende 1987. Detta exempel beskriver 2004 års undersökning. Se även Svensson (2007). Beställare av undersökningen var Sveriges Lantbruksuniversitet.

De parametrar som skulle skattas var antal olycksfall och olycksföretag inom områdena jordbruk, trädgård, skogsbruk och annan näringsverksamhet samt antal olycksfall uppdelat på olika kategorier av olyckor, m.m.

7.1.1 Urvalsram

Undersökningens målpopulation bestod av lantbruksföretag med minst 2,1 hektar åker eller stor djurhållning eller stor trädgårdsodling. De lantbruk som hamnade under gränsen ansågs försumbara. För att ta fram en urvalsram användes 2003 års version av Lantbruksregistret. Delmängder av detta register utgör ramar för de flesta lantbruksundersökningar i Sverige. Efter sambearbetning med Registret över totalbefolkningen (RTB) togs avlidna brukare bort ur ramen. Det totala antalet företag i ramen blev ca 67 100.

7.1.2 Urvalsmetoder

Urvalet drogs som ett stratifierat urval med Pareto πps inom strata. Urvalsstorleken var 7000 lantbruksföretag. Som hjälpinformation i designen användes data från den s.k. typologin för lantbruket. I typologin uttrycks arbetsbehov i standardtimmar, som tas fram med hjälp av modellberäkningar, varefter också en driftsinriktning bestäms. Här kalkylerades en olycksrisk för företaget genom att arbetsbehov 2003 inom olika driftsgrenar vägdes samman. Vägningen kunde göras utifrån resultaten i 1987 års olycksfallsundersökning.

Stratifieringen gjordes med avseende på olycksrisk (två timklasser), förekomst av uppgift om annan näringsverksamhet, och driftsinriktningar (fem stycken). Totalt blev det, efter sammanslagningar av planerade strata, tio slutligt valda strata, varav ett stratum (med de största lantbruken) totalundersöktes.

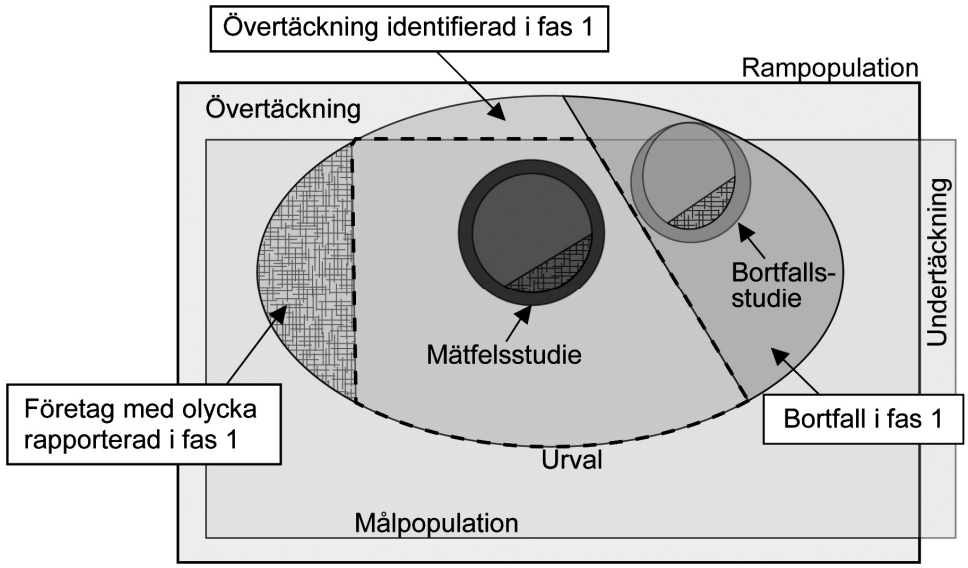
Som storleksmått i urvalsdragningen inom strata med Pareto πps användes olycksrisken. Detta mått utnyttjades även dessförinnan, för allokering av urvalsstorleken över strata. Olycksrisken avspeglade den viktigaste variabeln i undersökningen, nämligen om huruvida företaget hade haft någon olycka eller inte under 2004. Måttet antog aldrig extremt låga värden, varför risken för instabila skattningar till följd av stora vikter var liten.

En postenkät sändes ut till urvalet av lantbruksföretag. Därefter skickades två skriftliga påminnelser. För de lantbruk som angav minst en olycka, gjordes en utförligare intervju per telefon.

Sedan genomfördes uppföljningsstudier med hjälp av telefonintervjuer. Tvåfasmetodik användes för dessa studier av bortfallsfel och mätfel samt justering av dessa. Urvalet delades in i tre strata (inom första fasens strata): de som svarat och rapporterat olycka, de som svarat och angivit att ingen olycka förekommit, och de som inte alls svarat på enkäten. I det sistnämnda stratumet drogs ett underurval av 400 företag för en bortfallsstudie enligt Hansen-Hurwitz bortfallsplan – se Särndal et al. (1992), avsnitt 15.4.3. Bland dem som angivit att ingen olycka förekommit drogs också ett underurval av 400 företag; syftet var att justera för att olyckorna hade underrapporterats i postenkäten.

Figur 7.1.1 visar översiktligt de olika objektmängderna i olycksfallsundersökningen. De mönstrade fälten representerar företag med minst ett olycksfall. Cirklarna inuti den stora ovalen visar underurvalen till mätfelsstudien respektive bortfallsstudien. De innersta ovalerna står för de svarande i dessa kvalitetsstudier.

Figur 7.1.1 Olycksfallsundersökning för lantbruket



Svarsandelarna i kvalitetsstudierna blev så höga som 91 procent för bortfallsstudien respektive 99 procent för mätfelsstudien. Resultaten av studierna utnyttjades i skattningarna genom att den s.k. π^*-estimatorn användes, se Särndal et al. (1992), avsnitt 9.3 och 15.4.3. Statistikens kvalitet höjdes alltså genom att den stora undersökningens fördelar (hög precision) kombinerades med den lilla kvalitetsundersökningens fördelar (reducerade systematiska mätfel och bortfallsfel). Effekten visade sig i denna undersökning vara marginell för bortfallsfelet, men betydande för mätfelet: andelen företag med olyckor justerades från 4 till 7 procent genom mätfelsstudien. I retrospektiva studier av förekomster av ovanliga händelser finns det en risk för underrapportering; mätfelsstudien bedöms ha minskat underrapporteringen.

7.2 Hushållens ekonomi

Undersökningen ”Hushållens ekonomi”, HEK, är en årlig undersökning. Syftet med undersökningen är att kartlägga den disponibla inkomstens fördelning bland olika hushåll, belysa inkomststrukturen samt beskriva boendet och boendeutgifterna för hushåll i olika upplåtelseformer. Redovisningsgrupper är huvudsakligen hushållstyp, kön och ålder. Många olika typer av statistiska mått skattas: totaler, medelvärden, medianer och Ginikoefficienter.

Målpopulationen för undersökningen utgörs av de hushåll som fanns i landet den 31 december referensåret. Hushållen består av personer som enligt gällande lagar och förordningar skulle ha varit folkbokförda i landet. I redovisningen av inkomststatistiken används helårshushåll. Helårshushåll utgörs av de personer som var folkbokförda vid såväl årets början som dess slut och vars hushåll har en disponibel inkomst skild från noll.

Skattningar tas fram för såväl individer som hushåll.

7.2.1 Urvalsram

Undersökningens målpopulation består av alla hushåll. Det finns dock för närvarande inget register över dessa. För att hantera detta definieras en population bestående av de individer som är 18 år eller äldre och som bott i landet under aktuellt år. För denna population finns en bra urvalsram, nämligen Registret över totalbefolkningen (RTB). RTB bedöms vara en bra ram med litet täckningsfel för populationen av individer. När man gör ett urval av individer väljer man samtidigt hushåll som utvalda individer tillhör. Urvalsmetoden för hushåll är alltså nätverksurval – se avsnitt 4.5.

7.2.2 Urvalsmetoder

Urvalsramen stratifieras i fyra strata: personer som bor i bostadsrätt, personer över 75 år, övriga levande personer som bor kvar i landet samt personer som utvandrat eller avlidit under året. År 2005 drogs ca 17 000 individer, med OSU inom varje stratum.

Bostadsrätt finns med som ett kriterium för stratumtillhörighet, eftersom bostadsrättsinnehavare är en viktig redovisningsgrupp som inte säkert hade täckts med tillräckligt stort urval utan stratifiering.

Tilläggsurval görs ibland ur särskilda delpopulationer av RTB om en kund efterfrågar särskilda redovisningsgrupper.

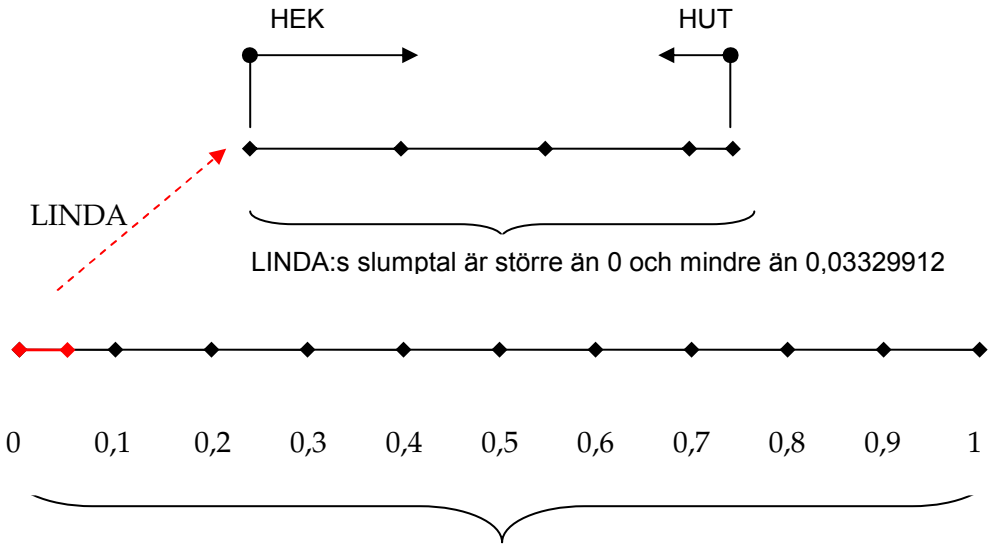
Det primära är att beskriva inkomstfördelningen i samhället över lag och mellan olika grupper med olika deskriptiva mått. Vissa av de variabler som undersöks har sneda fördelningar – inte minst olika inkomstvariabler. Detta leder till att skattningar för vissa redovisningsgrupper riskerar att få mycket stora medelfel. För att kompensera för detta görs tilläggsurval av individer och deras hushåll med stora kapitalvinster och/eller kapitalförluster. Variabeln kapitalvinster/-förluster korrelerar positivt med variabler som hör till gruppen inkomster av kapital.

7.2.3 Samordning av urval och urvalsramar

För att möjliggöra longitudinella studier av vissa variabler är både HEK-urvalet och HUT-urvalet positivt samordnade med den urvalsundersökning som kallas Longitudinella inkomstdatabasen (LINDA). (HUT står för ”Hushållens utgifter”.) Samordningen görs med hjälp av permanenta slumptal (se avsnitt 5.4.3). Alla individer i RTB tilldelas ett permanent slumptal. Slumptalen är likformigt fördelade mellan noll och ett. Alla individer med ett permanent slumptal i intervallet (0, 0,03329912) omfattas av LINDA-urvalet, vilket resulterar i en slumpmässig urvalsstorlek. Nytillkomna individer får också ett permanent slumptal, och de som faller inom LINDA:s intervall ingår i urvalet. För att minimera uppgiftslämnarbördan är HEK och HUT sinsemellan maximalt negativt samordnade.

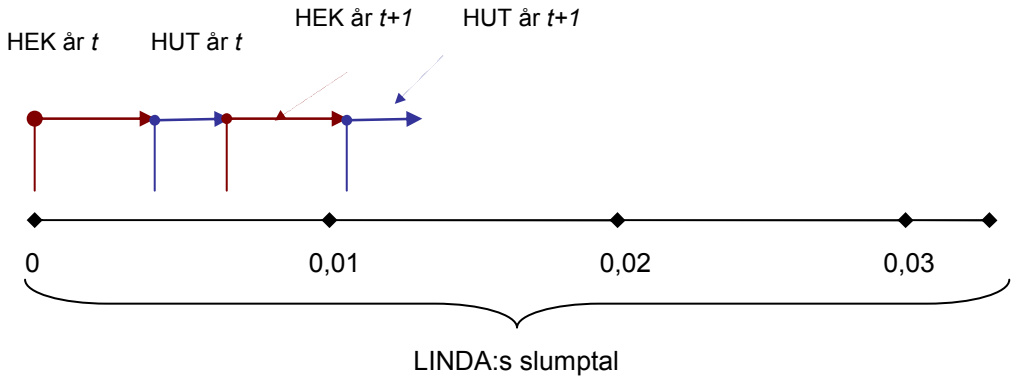
I figur 7.2.1 visas den positiva samordningen mellan LINDA och HEK samt LINDA och HUT och den negativa samordningen mellan HEK och HUT.

Figur 7.2.1 Schematisk beskrivning av samordningen av undersökningarna LINDA, HEK och HUT



Under 2006 överlappade de preliminära urvalen för HEK och HUT något inom vissa grupper. För att minimera risken för att detta upprepas, kommer den negativa samordningen mellan HEK och HUT att göras om. Undersökningarna kommer att dras sekventiellt efter varandra, se figur 7.2.2. Bilderna och förslaget till ny samordning är hämtade från Petric (2007).

Figur 7.2.2 Schematisk beskrivning av samordningen av undersökningarna LINDA, HEK och HUT från och med 2007



7.3 Arbetskraftsundersökningarna

Arbetskraftsundersökningarna (AKU) har som många större undersökningar flera syften. Det som har haft stor betydelse för valet av urvalsdesign är kravet på nivå- och förändringsskattningar avseende kvartal, eftersom AKU är ett väsentligt underlag till nationalräkenskaperna. Även månatliga förändringar är viktiga, men lägre prioriterade än kvartalsförändringar. Huvudparametrar är antal sysselsatta, antal arbetslösa och antal arbetade timmar. Nettoförändringarna i huvudstorheterna och andra storheter skattas för en mängd redovisningsgrupper. Även bruttoförändringar skattas. Urvalet omfattar ca 21 500 individer varje månad.

Eurostat efterfrågar även data som gör det möjligt att skatta vissa parametrar som är definierade för hushåll.

7.3.1 Urvalsram

Målpopulationen är personer i åldern 15–74 år som är folkbokförda i Sverige. Urvalsramen skapas från Registret över totalbefolkningen (RTB) och bedöms ha litet täckningsfel för populationen av individer.

7.3.2 Urvalsmetoder

Urvalsmetoden är stratifierat OSU. Stratifieringen görs genom korsklassificering av kön, län med storstadsområdena som tre extra områden samt åldersgrupperna 15, 16–64 och 65–74 år. Sammanlagt ger denna stratifiering 144 kategoristrata. Storleksstratifiering används inte. Tidigare stratifierade man även efter sysselsatt/ej sysselsatt enligt sysselsättningsregistret. Den stratifieringsvariabeln utgick eftersom det totala antalet strata blev stort. Dessutom används den variabeln och ytterligare hjälpinformation från sysselsättningsregistret i estimationen.

Urvalet allokeras till strata som ingår i åldergruppen 16–64 år med proportionell allokering. Även strata inom de två andra åldersgrupperna allokeras proportionellt mot storleken, men med lägre proportionalitetskonstant.

För att inte belasta urvalspersoner för hårt undantas utvalda personer om de har varit med i AKU någon gång de senaste fem åren. Det händer ganska sällan att individer råkar komma med i två urval med så kort mellanrum.

Även hushåll inkluderas i urvalet med nätverksurval. Ett hushåll är utvalt om någon hushållsmedlem är det (se avsnitt 4.5). I estimationen räknas hushållets inklusionssannolikhet som $M\pi_k$ där *M* är antal valbara hushållsmedlemmar i det valda hushållet och π_k är individ *k*:s inklusionssannolikhet. Eurostat får inversen av $M\pi_k$ som hushållsvikt multiplicerad med en kalibreringsvikt.

7.3.3 Övergripande urvalsfrågor

Nyligen övergick man från systematiskt urval till att dra obundet slumpmässigt urval (OSU) inom strata.

Urvalsstorleken styrs framför allt av undersökningens budget. Eftersom det finns efterfrågan på skattningar för små redovisningsgrupper, skulle ett större urval kunna motiveras. Under en tid var urvalsstorleken betydligt mindre för att frigöra resurser till en omläggning av AKU.

7.3.4 Samordning av urval och urvalsramar

För att förbättra precisionen i förändringsskattningarna används paneler, där varje uppgiftslämnare blir kontaktad åtta gånger. Eftersom det är kvartalsförändringar som har högst prioritet kontaktas varje utvald individ en gång per kvartal. Urvalspersonerna stannar alltså kvar i undersökningen i ungefär två år.

Rotationsgraden är således en åttondel, dvs. var tredje månad återkontaktas sju åttondelar medan en åttondel tackas för sin medverkan. Det är mindre än det intervall för rotationsgrad som Cochran föreslår, se avsnitt 5.4.2. Det finns flera skäl till att AKU har en relativt långsam rotation av urvalspersoner. Ett är att förändringsskattningar prioriteras framför nivåskattningar. Ett annat är att det är praktiskt och leder till lågt bortfall att inte rotera för snabbt. I AKU är det som svårast att få en hög svarsandel i en ny panel (kallas rotationsgrupp i AKU-sammanhang). När det inledande spårnings- och eventuellt övertalningsarbetet är avklarat, är kraven på intervjuarna något lägre.

I tabell 7.3.1 visas urvalsplanen för rotationsgrupperna för en svit av kvartal. I t.ex. januari år *t* väljs slumpmässigt en grupp individer som bildar en rotationsgrupp. Den betecknas i tabellen med $A_{t,1}$. Första intervjun sker i januari år *t*, nästa i april år *t*, och så vidare fram till oktober år *t*+1, då åttonde och sista intervjun görs. En rotationsgrupp förekommer alltså i tabellen längs en diagonal.

Det framgår av raden för januari år *t* att åtta rotationsgrupper kontaktas för intervju. Den som då kontaktas för åttonde gången är två år ”gammal”; med andra ord, den valdes år *t*-2. Varje månad kontaktas alltså rotationsgrupper som är mellan 0 och 2 år ”gamla”.

I mellanliggande månader, som för tydlighetens skull inte visas i tabellen, finns två andra sviter av kvartal. T.ex. väljs i februari år *t* en rotationsgrupp som skulle kunna betecknas $B_{t,1}$. Den följer samma mönster som $A_{t,1}$ fast med en månads förskjutning. Skälet att en rotationsgrupp kontaktas igen efter ett kvartal och inte redan nästa månad är att skattningar av förändringar kvartal till kvartal prioriteras framför korttidsförändringar.

Tabell 7.3.1 Urvalsplanen för rotationsgrupperna för en svit av kvartal

	Vilken gång i ordningen som rotationsgruppen kontaktas							
Aktuell månad	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Jan., *t*–1	$A_{t-1,1}$	$A_{t-2,4}$	$A_{t-2,3}$	…			…	$A_{t-3,2}$
April, *t*–1	$A_{t-1,2}$	$A_{t-1,1}$	$A_{t-2,4}$	…			…	$A_{t-3,3}$
Juli, *t*–1	$A_{t-1,3}$	$A_{t-1,2}$	$A_{t-1,1}$	…	⋱		…	$A_{t-3,4}$
Okt., *t*–1	$A_{t-1,4}$	$A_{t-1,3}$	$A_{t-1,2}$	…		⋱	…	$A_{t-2,1}$
Jan., *t*	$A_{t,1}$	$A_{t-1,4}$	$A_{t-1,3}$	…			…	$A_{t-2,2}$
April, *t*	$A_{t,2}$	$A_{t,1}$	$A_{t-1,4}$	…			…	$A_{t-2,3}$
Juli, *t*	$A_{t,3}$	$A_{t,2}$	…				…	$A_{t-2,4}$
Okt., *t*	$A_{t,4}$	$A_{t,3}$	…				…	$A_{t-1,1}$
Jan., *t*+1	$A_{t+1,1}$	$A_{t,4}$	…				…	$A_{t-1,2}$
April, *t*+1	$A_{t+1,2}$	…					…	$A_{t-1,3}$

Not: $A_{t,i}$ = rotationsgrupp *i* som valdes slumpmässigt år *t*

En fyllig praktisk beskrivning av AKU:s urvalsförfarande finns i Mirza och Hörngren (2002). Där finns också en formelbeskrivning av estimatorerna för nivå- och förändringsskattningar. Här beskrivs urvalsdragningen kortfattat. I början på varje år dras ett kompletteringsurval omfattande ca 32 000 individer ur RTB. Detta ska försörja året som kommer med nya urvalspersoner, dvs. en ny rotationsgrupp varje månad. Kompletteringsurvalet delas därför slumpmässigt in i tolv grupper. Eftersom det dröjer nästan ett år innan den sista av dessa intervjuas, dras kompletteringsurvalet ur gruppen 14–75 år. De som vid intervjun ännu inte uppnått 15 års ålder kontaktas inte, men stannar kvar i rotationsgruppen för att kunna kontaktas senare.

Tidigare drogs även ett annat kompletteringsurval ur gruppen nytillkomna i Sverige. Denna undertäckning justeras numera endast genom kalibrering i skattningsförfarandet.

Hjälpinformationen aktualiseras för samtliga intervjupersoner fortlöpande. Alla dessa personers data finns i en ”urvalsbank” som uppdateras.

Vardera av de tolv delgrupperna i kompletteringsurvalet delas vidare in i slumpceller. Det gör det praktiskt att dra slumpmässiga underurval ur AKU-urvalet för tilläggsundersökningar. Sådana görs ofta.

7.4 Företagens ekonomi

7.4.1 Inledning

Det övergripande syftet med Företagens ekonomi, FEK, är att förse olika användare med underlag om utveckling och struktur med avseende på branscher och regioner. Statistiken ger underlag för beräkningar av långa tidsserier, vilka utnyttjas för att i olika sammanhang belysa utvecklingen i det svenska näringslivet. Statistiken används också som underlag för analyser av kostnadsläge, produktivitetsutveckling och branschers ekonomiska utveckling. Ett av de viktigaste användningsområdena är som underlag för beräkning av national- och finansräkenskaper. Med hjälp av statistiken kan företagssektorns bidrag till bruttonationalprodukt, sparande, investeringar och förändringar av finansiella tillgångar och skulder räknas fram.

Förutom administrativt material från Skatteverket (standardiserade räkenskapsutdrag, SRU) baseras statistiken på i huvudsak fyra olika undersökningar. Urvalsförfarandet för var och en av dessa delundersökningar beskrivs separat under avsnitt 7.4.3 nedan.

De viktigaste parametrarna i FEK är aggregat av följande variabler: förädlingsvärde, produktionsvärde, förbrukningsvärde, intäkter och deras fördelning på verksamheter, kostnader och deras fördelning på kostnadsslag, investeringar och deras fördelning på objektslag samt aktier och deras fördelning på noterade och onoterade aktier.

De viktigaste redovisningsgrupperna är olika aggregat av bransch. Redovisning på storleksgrupper definierade av antal anställda görs också.

7.4.2 Urvalsramar

Målpopulationen för FEK utgörs av alla företagsenheter, FE, och verksamhetsenheter, VE, och lokala verksamhetsenheter, LVE, som dessa enligt respektive objektsdefinition kan delas upp i, som bedrivit näringsverksamhet under referensåret. Statistiken från FEK ska avse kalenderår. Den *gemensamma* rampopulationen för FEK år t, utgörs av samtliga icke-finansiella FE (med tillhörande VE och LVE) som enligt novemberversionen av SAMU år t var verksamma. SAMU-ramen kompletteras även med ett mindre antal enheter som inte är verksamma enligt aktivitetsbegreppet i FDB, men som haft betydande aktieinnehav. I de olika delundersökningarna hanteras objekten i ramen på delvis olika sätt, vilket beskrivs under avsnitt 7.4.3 nedan.

7.4.3 Urvalsmetoder för de fyra delundersökningarna

De statistiska storheterna av intresse i FEK är i huvudsak av två typer, nämligen *nivåer* och *specifikationer.* Med nivåer avses här olika populationstotaler som t.ex. nettoomsättning, investeringar och kostnader. Till var och en av dessa populationstotaler finns en eller flera *specifikationer* knutna. Exempel på sådana specifikationer är nettoomsättningens fördelning på produkter och investeringarnas fördelning på objektslag. Själva grundidén

med FEK är att man genom att kombinera administrativt material (i huvudsak SRU) med direktinsamling från de största företagen i princip har tillgång till nivåerna på objektnivå för samtliga enheter som ingår i FEK-populationen. Detta innebär att skattningen av populationstotalerna av typen nivåer inte är behäftade med något urvalsfel. Däremot finns ett bortfallsfel, då SRU-materialet inte är fullständigt. Nedan beskrivs undersökningen av de största företagen och de tre specifikationsundersökningarna var och en för sig.

(i) Fullständig blankett
För de största företagen används inte SRU som källa för FEK. I stället får dessa företag en s.k. fullständig blankett, som dels omfattar motsvarande informationsmängd som för övriga företag hämtas från SRU, dels omfattar de tre delundersökningarna SpecRR, SpecA och SpecI.

Följande grupper av företag får en fullständig blankett:

- *Företag med flera verksamhetsenheter.* Samtliga företag som har delats upp i mer än en verksamhetsenhet enligt FDB.
- *Vissa sammansatta företagsenheter.* Sammansatta företagsenheter som utgör koncernenhet eller annan typ av konsolidering, t.ex. del av koncern.
- *Företag med ständigt saknade/underkända SRU:er.* Till denna grupp förs affärsverk, stora stiftelser, stora ideella och ekonomiska föreningar (juridisk form 51–72) och stora utländska filialer (sektor 130).
- *Övriga stora företag.* Följande grupper av företag kommer att ingå här:
  - Företag tillhörande branscherna A–D (enligt SNI2002-nomenklaturen) med minst 500 anställda.
  - Företag tillhörande branscherna E–Q med ett produktionsvärde som överstiger 1 miljard kronor eller vars produktionsvärde utgör en andel av branschgruppens produktionsvärde överstigande 10 procent.
- *Stora företag (minst 200 anställda) med brutet räkenskapsår med bokslutsmånad maj–augusti.*
- *Stora företag som vid leverans från Skatteverket upptäcks ha en underkänd SRU med utsändning i oktober.*

Totalt omfattar urvalet avseende den fullständiga blanketten ca 600 företag.

SpecRR
Syftet med delundersökningen SpecRR är att möjliggöra skattningar av intäkternas fördelning på verksamheter och kostnadernas fördelning på kostnadsslag. Urvalsstorleken för FEK 2004 uppgick till ca 13 000 FE/VE. Stratifiering sker efter FE/VE-bransch. Totalt användes ca 170 olika branschstrata. Grunden för uppdelning på branschstrata är att möjliggöra redovisning efter nationalräkenskapernas (NR:s) s.k. årsbranscher. Hänsyn tas också till branschernas olika intäkts- och kostnadsstruktur. Allokeringen mellan branscher görs med målsättningen att erhålla samma absoluta precision i alla branschstrata. Dock är det så att stora branscher (mätt i ekonomisk aktivitet) prioriteras framför mindre branscher. Allokeringen beskrivs i avsnitt 7.4.4. Inom varje branschstratum dras sedan ett Pareto πps-urval. Som hjälpvariabler i ramen finns företagets intäkter och kostnader enligt skattedeklarationen. Det storleksmått x_k som används är

medelvärdet av dessa hjälpvariabler. Av bokföringsskäl brukar kostnader registreras som negativa värden, men här räknas de som positiva.

På basis av storleksmåttets fördelning inom branschstratum görs en cut-off så att de minsta företagen som inte behövs för att nå 90 procents täckning ges inklusionssannolikheten noll. Antalsmässigt innebär detta förfarande att drygt två tredjedelar av företagen hamnar under cut-off-gränsen.

SpecA
Främsta syftet med SpecA är att få fram ett marknadsvärde för onoterade aktier i den icke-finansiella företagssektorn. Det finns ett krav från NR som gör det nödvändigt att dela upp aktieinnehavet mellan noterade, onoterade, svenska och utländska aktier. Även en uppdelning mellan koncern och andra aktier måste göras för att kunna beräkna marknadsvärdet på det onoterade aktieinnehavet. Däremot finns inga krav på nedbrytning av resultatet på redovisningsgrupper, något som gör att urvalsstorleken kan hållas nere. Urvalet dras som ett stratifierat OSU av företagsenheter (FE). Stratifieringen görs med avseende på variabeln ”Summa aktier och andelar” från SRU. Företag med ett innehav av aktier och andelar på mer än 100 miljoner kronor totalundersöks. Mellan övriga storleksstrata sker fördelningen av urvalsstorleken med hjälp av Neymanallokering. Den totala urvalsstorleken uppgår till ca 600 företag.

SpecI
Syftet med delundersökningen SpecI är att ta fram skattningar avseende investeringarnas fördelning på objektslag och region. I praktiken är detta inte en utan snarare tre olika delundersökningar. Med hjälp av SRU-materialet görs investeringsberäkningar så att fördelningar av brutto- respektive nettoinvesteringar kan erhållas på Maskiner och inventarier samt Fastigheter och markanläggningar. Framför allt NR behöver dock ytterligare fördelningar, nämligen följande:

- Bruttoinvesteringar i fastigheter, vilka fördelas på ny-, till- och ombyggnad. Samtliga företag som enligt det administrativa materialet har gjort fastighetsinvesteringar på minst 5 miljoner kronor ingår i detta urval. Totalt för 2003 omfattades drygt 2000 företag. För ca 600 av dessa var det dock möjligt att hämta önskade uppgifter från SCB:s korttidsundersökning avseende investeringar (Investeringsenkäten) i stället för att direktinsamla uppgifterna.
- Nettoinvesteringar i maskiner och inventarier ska vidarespecificeras på flygplan, fartyg och rälsfordon. Företag i utpekade branscher med minst 5 miljoner kronor i maskiner och inventarier undersöks. Gruppen omfattar knappt 100 företag.
- Regional fördelning av industriföretagens investeringar. För att klara kraven på regional redovisning av industriföretagens investeringar ombeds industriföretag med lokala verksamhetsenheter (LVE) i minst två kommuner och med en investering i maskiner och inventarier på minst 5 miljoner kronor att fördela sina investeringar per kommun. Totalt omfattas 300 företag, varav ca 100 företags uppgifter kan hämtas från Investeringsenkäten.

SpecI-urvalet är med andra ord ett cut-off-urval där företagen ovanför cut-off-gränsen totalundersöks.

7.4.4 Allokering mellan branscher i delundersökningen SpecRR

Allokeringen mellan branscher görs på basis av olika typer av totaler, dvs. urvalsstorleken i en bransch *b* har räknats ut som:

$$n_b = n_{tot} \frac{t_b}{\sum_{k=1}^{B} t_k} \qquad (7.4.1)$$

där n_{tot} är en total urvalsstorlek som bestämts och t_b är totalen av någon eller några variabler i branschen *b*, och där nämnaren är summan av denna total över alla branscherna (*1,…, b,…, B*). Om man har som målsättning att alla skattningar ska ha ungefär samma varians oavsett bransch är det inte säkert att (7.4.1) fungerar så bra, då detta mått inte tar någon hänsyn till variationen inom branscher.

I syfte att åstadkomma relativt jämnstora varianser beslöts att i stället testa följande iterativa procedur för bestämning av den branschvisa urvalsstorleken n_b:

$$n_b^{(k)} = n \frac{n_b^{(k-1)} V_b^{(k-1)}}{\sum_j n_j^{(k-1)} V_j^{(k-1)}} \qquad (7.4.2)$$

där *k* betecknar iteration *k*. Som startvärde (dvs. vid *k=0*) på n_b används det värde som fås från (7.4.1).

Här följer en förklaring till valet av iterationsförfarandet (7.4.2):

Låt $n_b^{(0)}$ vara det n_b-värde som fås ur relationen (7.4.1) ovan.

Låt $V_b^{(0)}, b = 1,2,...,B$ beteckna de varianser per bransch som fås om $n_b^{(0)}, b = 1,2,...,B$, är urvalsstorlek i respektive bransch. Man kan förvänta sig en ganska stor variation mellan varianserna $V_b^{(0)}, b = 1,2,...,B$.

Man söker en ny uppsättning urvalsstorlekar $n_b^{(1)}, b = 1,2,...,B$ som, under bivillkoret $\sum_j n_j^{(1)} = n$, ger jämnstora varianser:

$$V_1^{(1)} = V_2^{(1)} = ... = V_B^{(1)} = V. \qquad (7.4.3)$$

Följande (troligen ganska grova) approximation gjordes:

$$V_b \approx \frac{A_b}{n_b}. \qquad (7.4.4)$$

Här antas med andra ord att den varians som fås utifrån Pareto πps-urvalet är omvänt proportionell mot urvalets storlek.

Genom att kombinera (7.4.3) och (7.4.4) fås följande relationer:

$$n_b^{(1)} \approx \frac{A_b}{V_b^{(1)}} = \frac{A_b}{V}, \qquad (7.4.5)$$

$$n = \sum_j n_j^{(1)} \approx \frac{\sum_j A_j}{V} \quad . \qquad (7.4.6)$$

Från (7.4.5) och (7.4.6) följer:

$$n_b^{(1)} \approx n \frac{A_b}{\sum_j A_j} \qquad (7.4.7)$$

Genom att applicera (7.4.4) på $n_b^{(0)}$ och $V_b^{(0)}$, dvs. $A_b \approx n_b^{(0)} V_b^{(0)}$ fås slutligen ur (7.4.7) följande relation som är precis iterationsformeln (7.4.2) ovan för $k = 1$:

$$n_b^{(1)} = n \frac{n_b^{(0)} V_b^{(0)}}{\sum_j n_j^{(0)} V_j^{(0)}} \quad . \qquad (7.4.8)$$

7.5 Inrikes och utrikes trafik med svenska lastbilar

7.5.1 Inledning

Följande exempel avseende undersökningen *Inrikes och utrikes trafik med svenska lastbilar* (SLIT) visar hur kvaliteten i skattningarna av viktiga parametrar kan förbättras genom modifiering av stratifiering och urvalsallokering samtidigt som urvalsstorleken är oförändrad. Problemet i lastbilsstatistiken uppstod när kvalitetskraven på utrikestransporterna skärptes omkring år 2002 med anledning av en successiv EU-anpassning och det bedömdes som orealistiskt att utöka urvalsstorleken. Lösningen på kvalitetsproblemet blev att ersätta tidigare använda stratifieringsvariabler och att ändra stratumindelning och urvalsallokering. Metodöversynen rapporterades i Eriksson et al. (2003).

Undersökningar av varutransporter på väg med lastbil har gjorts av SCB sedan 1972, på senare år på uppdrag av SIKA (Statens institut för kommunikationsanalys). Mellan 1972 och 1999 gjordes undersökningar av inrikes transporter med svenska lastbilar. Statistiken över inrikes transporter kompletterades med motsvarande statistik över utrikes transporter med svenska lastbilar, med anledning av den EU-anpassning av statistiken som följde av Sveriges medlemskap i EU 1995. Mellan 1995 och 1999 gjordes det alltså två parallella undersökningar av varutransporter på väg: inrikes respektive utrikes transporter med svenska lastbilar.

Från 1999 gäller en ny EU-förordning om statistikrapportering av varutransporter på väg, och de båda tidigare undersökningarna av inrikes respektive utrikes varutransporter ersattes av undersökningen *Inrikes och utrikes trafik med svenska lastbilar* (SLIT) från och med år 2000.

Viktiga undersökningsvariabler är körsträcka och sändningens vikt samt den härledda variabeln tonkilometer, som är produkten av vikt och körsträcka. En ”sändning” är ett bestämt varuparti som transporteras från ett pålastningsställe till ett avlastningsställe. Begreppet körning definieras i Eriksson et al. (2003). I praktiken sammanfaller ofta sändning och körning.

En årlig undersökning omfattar fyra separata kvartalsundersökningar med redovisning av kvartalsstatistik. Statistik för hela året tas fram efter sammanläggning av statistik för kvartalen. Det vanligaste statistiska måttet är summan (totalen).

7.5.2 Urvalsram

Målpopulationen utgörs av alla lastbilar, eller dragbilar, med vissa (de flesta) karosserikoder. Lastbilarna ska vidare vara högst 30 år gamla och ha en maximilastvikt på minst 3,5 ton samt vara i trafik under referensperioden. Alternativt kan man se målpopulationen som mängden av sändningar och körningar, grupperade till veckor för alla lastbilar enligt ovan.

Rampopulationen består av lastbilar som uppfyller ovanstående kriterier och är registrerade i Vägtrafikregistret (Vägverkets fordonsregister) och vars ägare finns med i FDB. Ramen färdigställs 1–2 månader före kvartalet i fråga. Se även beskrivningen nedan av steg 2 i urvalet.

För att få bedriva internationell godstransport på väg i yrkesmässig trafik krävs tillstånd från Länsstyrelsen. Uppgift om att t.ex. en åkare har ett antal tillstånd påförs ramen och används i stratifieringen. Körsträcka är en annan hjälpvariabel som numera används i stratifieringen. Från AB Svensk Bilprovning hämtas nämligen uppgifter om mätarställningar, vilka sedan används för att beräkna körsträckor.

7.5.3 Urvalsmetoder

En lastbil medverkar i undersökningen en vecka och rapporterar under den veckan alla transporter, dvs. alla sändningar och alla körningar. Urvalet görs i två steg. Urvalsobjekt i första steget är lastbil, i andra steget är det vecka.

Mellan 2000 och 2002 tillämpades den urvalsdesign som redovisas i tabell 7.5.1, med stratifiering av lastbilar efter

- om bilens ägare har tillstånd för utrikes transporter eller inte
- lastbilar, tankbilar respektive bankebilar (timmerbilar)
- yrkesmässig trafik respektive firmatrafik
- län respektive länsgrupper
- lastbilarnas maximilastvikt (med olika antal storleksstrata i de tre storstadslänen och i övriga län).

Den mest övergripande stratifieringsnivån är efter tillstånd för utrikes transporter, i två huvudstrata: inrikes och utrikes. Samtliga stratifieringsvariabler i punktlistan ovan, utom maximilastvikten, kontrollerar urvalsstorlekar i viktiga redovisningsgrupper. Särskilda strata för tank- och bankebilar har dessutom den funktionen att de avgränsar grupper av lastbilar som kör ett stort antal tonkilometer. Lastbilarnas maximilastvikt är en variabel som används för storleksstratifiering, eftersom antal utförda tonkilometer samvarierar med lastbilens lastvikt.

Från och med 2003 förbättrades precisionen i skattningarna av de viktigaste parametrarna (transporterad godsmängd och transportarbete) genom revidering av stratifieringen och urvalsallokeringen. Detta beskrivs nedan. Stratifiering och urvalsallokering reviderades med syftet att uppnå

en så kostnadseffektiv undersökning som möjligt, utan utökning av urvalets storlek. Den förändrade stratifieringen och urvalsallokeringen innebar att urvalet styrdes så att fler utrikestransporter kunde observeras.

Från och med 2003 tillämpas den urvalsdesign som redovisas i tabell 7.5.2. De mest väsentliga skillnaderna i stratifieringen av lastbilar var följande:

- Körsträcka infördes. Det är ännu en variabel för storleksstratifiering. Baserat på en studie av korrelationer bedömdes att båda maximilastvikt och körsträcka gör nytta som stratifieringsvariabler. Stratumgränser bestämdes med ett sökförfarande, där varianser beräknades för olika stratumgränser. De som inte hade någon registrerad körsträcka lades i ett eget stratum. Stratifiering på maximilastvikt har behållits för gruppen lastbilar med korta årliga körsträckor, beroende på att bilarna i den kategorin visserligen gör korta körningar men ofta med stor godsmängd.
- Liksom tidigare bildar tank- respektive bankebilar egna strata inom huvudstratumet ”inrikes trafik”. Nu stratifieras dessa strata ytterligare endast med avseende på körsträcka.
- Yrkesmässig trafik respektive firmatrafik togs bort som stratifieringsvariabel, eftersom firmatrafiken minskat kraftigt sedan 1992 då denna (typ av) undersökning startades. En kontroll av precision för redovisningsgrupper efter uppdelning i yrkesmässig trafik och firmatrafik visade att sådan statistik skulle kunna tas fram ändå.
- Stratumet med bilar med ägare som har tillstånd för utrikes trafik stratifierades ytterligare efter antalet tillstånd. Idén var att ägare med många tillstånd kör särskilt mycket utrikes. Detta antagande testades med en pilotundersökning. Se Eriksson et al. (2003, avsnitt 3.2).
- I den regionala stratifieringen ersatte åtta NUTS 2-regioner län i inrikestransporterna, med syftet att reducera antalet strata bland vanliga lastbilar. Dock fick Gotland vara ett eget stratum även i fortsättningen, så att redovisningsgruppen Gotland skulle täckas med ett tillräckligt stort urval.

Urvalsstorleken är oförändrat 3 000 lastbilar per kvartal, med 1 500 lastbilar fördelade på vardera huvudstratumet inrikes respektive utrikes trafik.

Inom inrikes trafik allokeras urvalet på grupper av bilar, med 85,6 procent vanliga lastbilar, 6,1 procent tankbilar och 8,3 procent bankebilar. Proportionerna bestämdes efter ett omfattande sökförfarande, vars principer rapporteras i Rosén och Zamani (1993).

Inom stratumen tankbilar och bankebilar görs en proportionell allokering. Neymanallokering (x-optimal allokering) testades utan att precisionen i skattningarna förbättrades.

Stratumet ”vanliga lastbilar”, dvs. bilar som inte är tank- eller bankebilar, stratifieras ytterligare på finare nivå med ovanstående stratifieringsvariabler (se Eriksson et al. 2003, diagram 1). Olika tänkbara allokeringsvariabler som hade bäring på olika undersökningsvariabler testades. För varje allokeringsvariabel användes x-optimal allokering. Som kompromiss söktes bland vägningar av de urvalsstorlekar som man fann för olika allokeringsvariabler. Slutligen valdes medelvärdet av de bästa allokeringarna med avseende på tonkilometer respektive ton. Denna lösning framstod som mest robust med avseende på samtliga undersökningsvariabler.

En minsta urvalsstorlek tillämpas. Den är satt till 10 lastbilar. Om x-optimal allokering faller ut så att urvalsstorleken blir mindre än 10, sätts denna till 10 och allokeringen görs om för övriga strata.

Stratifiering och allokering för utrikesstratumet beskrivs kort i Eriksson et al. (2003).

Urvalet är som sagt ett tvåstegsurval. Det är dock inte ett standardurval av någon av de typer som beskrivs i kapitel 4, av följande två skäl: Urvalet av lastbilar styrs på ett visst sätt för att minska bördan för enskilda bilar. Exempelvis kan inte en bil som valts ett visst kvartal väljas på nytt samma år. (Undantag görs dock för lastbilar i körgrupp 2 i utrikesstratumet, vilka vanligtvis kommer med två gånger om året, men inte två kvartal i rad.) Vidare sprids urvalet av veckor ut över kvartalet för utvalda bilar. Urvalet kan dock approximeras med en välkänd urvalsmetod, vilket visades i Rosén och Zamani (1993). För att göra detta uppfattas s.k. lastbilsveckor, där varje lastbil har 13 veckor i ett kvartal, som objekt. Man kan se det som en rektangel med populationens lastbilar längs ena dimensionen och 13 veckor längs den andra. En lastbilsvecka är en ruta i rektangeln. Som strata uppfattas strata av lastbilar (enligt ovan) korsklassificerade med 13 veckor i kvartalet. Approximationen säger att i denna del av rektangeln dras ett OSU. Urvalsmetoden approximeras sålunda med stratifierat OSU.

För varje utvald lastbilsvecka görs en totalundersökning av körningar/sändningar.

7.5.4 Övergripande urvalsfrågor

Målpopulationen, som den definieras i EU-direktiv, är lastbilar med maximilastvikt på minst 3,5 ton. I tidigare versioner av undersökningen urvalsundersöktes även bilar i intervallet 2,0–3,5 ton. I en studie av Rosén och Zamani (1993) visades att bilar i detta intervall bidrog endast försumbart till totalskattningarna, dvs. ett cut-off-urval kunde införas.

Precisionskraven från Eurostat förtydligades. Kraven innebar högst 5 procents relativ osäkerhetsmarginal för inrikestransporterna och högst 10 procents relativ osäkerhetsmarginal för utrikestransporterna på årsstatistiken. Precisionskraven klarades åren 2000–2006 när det gäller årsstatistiken över inrikestransporterna. För utrikestrafiken uppfylldes kraven för körda kilometer 2001–2006 och för tonkilometer 2002–2006, men inte något år för godsmängden.

Tabell 7.5.1 Urvalsdesign i undersökningen *Inrikes och utrikes trafik med svenska lastbilar* 2000–2002

3 000 lastbilar per kvartal med uppgiftsinsamling för ”körning” (som är transport hel körsträcka eller del av körsträcka).

Inrikes yrkesmässig trafik Svenskregistrerade lastbilar, maximilastvikt minst 3,5 ton; ägaren saknar tillstånd för internationell yrkesmässig trafik; inkl. Vägverkets bilar (1 500 lastbilar, 50 procent)

Lastbilar (0,5 x 0,856 x 3000 bilar)

Yrkesmässig trafik (0,5 x 0,556 x 3000 bilar)

Regioner: län 01–25 (21 st.)

maximilastvikt 3,5–12,5 ton (utom Gotland och storstadslän)
maximilastvikt 12,5– ton (utom Gotland och storstadslän)

maximilastvikt 3,5–8 ton (i tre storstadslän)
maximilastvikt 8–13 ton (i tre storstadslän)
maximilastvikt 13– ton (i tre storstadslän)

Firmabilstrafik (0,5 x 0,300 x 3000 bilar)

Regioner: län 01–25 (21 st.)

maximilastvikt 3,5–11,5 ton (utom Gotland och storstadslän)
maximilastvikt 11,5– ton (utom Gotland och storstadslän)

maximilastvikt 3,5–7,5 ton (i tre storstadslän)
maximilastvikt 7,5–12,5 ton (i tre storstadslän)
maximilastvikt 12,5– ton (i tre storstadslän)

Tankbilar (0,5 x 0,061 x 3000 bilar)

Yrkesmässig trafik (0,5 x 0,045 x 3000)

Firmabilstrafik (0,5 x 0,016 x 3000)

Bankebilar (0,5 x 0,083 x 3000 bilar)

Yrkesmässig trafik (0,5 x 0,059 x 3000)

Firmabilstrafik (0,5 x 0,024 x 3000)

Utrikes yrkesmässig trafik Svenskregistrerade lastbilar, maximilastvikt minst 3,5 ton; ägaren innehar tillstånd för internationell yrkesmässig trafik; exkl. Vägverkets bilar (1 500 lastbilar, 50 procent)

Lastbilar Körgrupp 1 (högst 15 tillstånd för internationell trafik eller minst 16 tillstånd och mindre än 80 procent utrikes körning)

Regioner 1–6 (länsgruppering)

Lastbilar Körgrupp 2 (minst 16 tillstånd för internationell trafik)

Regioner 1–6 (länsgruppering)

Tabell 7.5.2 Urvalsdesign i undersökningen *Inrikes och utrikes trafik med svenska lastbilar* från och med 2003

3 000 lastbilar per kvartal med uppgiftsinsamling för ”sändning” (som är varuparti, hel last eller del av last, transporterad hel körsträcka eller del av körsträcka).

Inrikes yrkesmässig trafik Svenskregistrerade lastbilar, maximilastvikt minst 3,5 ton; ägaren saknar tillstånd för internationell yrkesmässig trafik; inkl. Vägverkets bilar (1 500 lastbilar, 50 procent)

Lastbilar (0,5 x 0,856 x 3 000 bilar)

Körsträcka, uppgift saknas

NUTS 2-regioner SE01–SE08, exkl. Gotland

Körsträcka 0–3 499 mil

maximilastvikt högst 13 ton
maximilastvikt minst 13 ton

Körsträcka 3 500–7 999 mil

Körsträcka 8 000– mil

Region Gotland

Körsträcka 0–3 999 mil
Körsträcka 4 000– mil

Tankbilar (0,5 x 0,061 x 3 000 bilar)

Körsträcka 0–6 499 mil
Körsträcka 6 500–9 999 mil
Körsträcka 10 000– mil
Körsträcka, uppgift saknas

Bankebilar (0,5 x 0,083 x 3 000 bilar)

Körsträcka 0–13 499 mil
Körsträcka 13 500–17 999 mil
Körsträcka 18 000– mil
Körsträcka, uppgift saknas

Utrikes yrkesmässig trafik Svenskregistrerade lastbilar, maximilastvikt minst 3,5 ton; ägaren innehar tillstånd för internationell yrkesmässig trafik; exkl. Vägverkets bilar (1 500 lastbilar, 50 procent)

Lastbilar Körgrupp 1 (högst 15 tillstånd för internationell trafik eller minst 16 tillstånd och mindre än 80 procent utrikes körning)

Körsträcka, uppgift saknas

Regioner 1–6 (länsgruppering)

Körsträcka 0–10 999 mil
Körsträcka 11 000– mil

Lastbilar Körgrupp 2 (minst 16 tillstånd för internationell trafik och minst 80 procent utrikes körning)

---

8 Den statistiska konsultens roll i planeringen av en urvalsundersökning

Undersökningens syfte är det som styr dess uppläggning. De flesta undersökningar har flera syften, som kan ställa olika, ibland motstridiga krav. Dessutom ingår de flesta undersökningar i mer eller mindre formaliserade system, som gör att en undersökning inte kan betraktas som en isolerad enhet. Kish (1988) listar olika syften som en statistisk undersökning kan ha. Han diskuterar också konflikter mellan olika syften.

Här är meningen att lyfta fram konsultens roll i planeringen av en undersökning. Inom SCB spelar metodstatistiker ofta rollen som konsult i såväl internt som externt arbete. Konsultens roll och konsultens färdigheter är något som ibland underskattas i organisationen och även i statistikutbildningen. Även om enbart urval och skattning ska utföras inom organisationen, behöver syftet med undersökningen i dess helhet klarläggas. En bra konsult blir man genom lärande i arbetet. Men man kan förbättra sina färdigheter som konsult även genom litteraturstudier. Derr (1999) är en bra bok om statistisk konsultation som tar upp både statistiska moment och konsultationsteknik.

Den första uppgiften för den statistiska konsulten är att diskutera undersökningens syften med beställaren eller kunden. Syftena behöver översättas till kvalitetskrav. Erfarenhetsmässigt har kunden ofta svårt att precisera sådana krav. Konsultens roll är att hjälpa kunden att formulera kvalitetskrav i form av krav på olika kvalitetskomponenter, se avsnitt 9.3. Efter diskussion med kunden formulerar konsulten undersökningen i statistiska termer. En översikt av termer gavs i kapitel 2. Kunden och konsulten hjälps åt att bryta ned det övergripande syftet med undersökningen i beståndsdelar. Dessa kan sedan ligga till grund för planering av undersökningen. Se vidare avsnitt 9.2 för specificering av en undersöknings syften.

Konsulten måste identifiera relationer mellan kvalitetskrav och urvalsmetod. En vedertagen relation är den mellan precisionskrav och urvalsstorlek, som behandlas i avsnitt 6.1. I de fall relationerna innehåller okända storheter, t.ex. en populationsvarians, är det konsultens roll att hitta metoder för att bedöma eller skatta dessa. Cochran (1977, s. 74–75) diskuterar den principiella beslutsprocessen vid bestämning av urvalsstorlek.

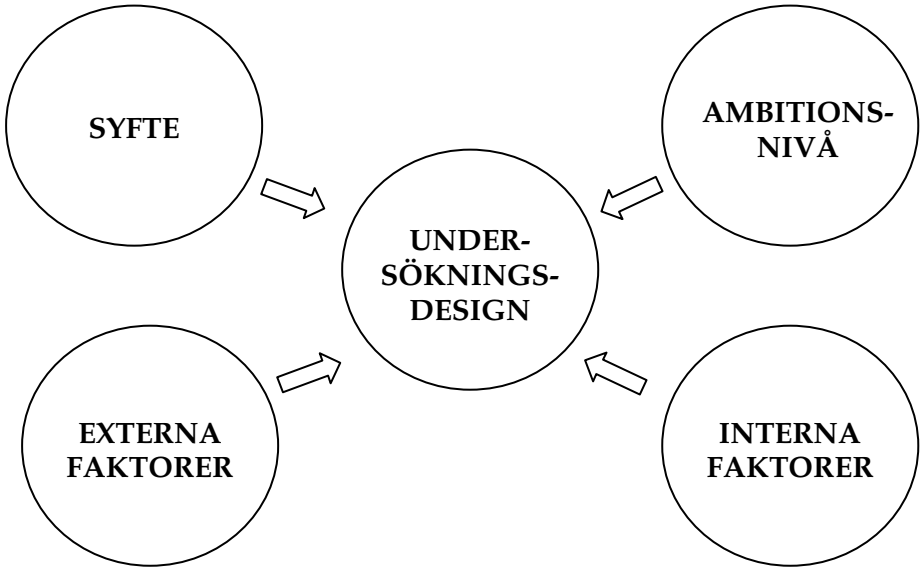
I de flesta undersökningar önskas skattningar för olika redovisningsgrupper. Då behöver man i princip driva resonemanget om urvalsmetod separat för varje redovisningsgrupp. Detsamma gäller om man har flera olika variabler, för vilka skattningar önskas.

9 Checklista för urvalsmetoder

9.1 Inledning

Det är lätt att beskriva den perfekta undersökningen: den ger relevanta svar på utpekade frågeställningar, är sparsam i sin resursanvändning, levererar resultat i tid, osv. Problemet är att nå dit. Omsorgsfull planering är ingen garanti för framgång, men chanserna ökar betydligt. Det här kapitlet identifierar faktorer som är betydelsefulla i planeringen av en urvalsundersökning. Läsanvisningar ges till avsnitt där djupare diskussioner förs. Det är att rekommendera att man i planeringen utgår ifrån undersökningens syfte och ambitionsnivå samt dess externa och interna förutsättningar. Dessa faktorer behandlas i tur och ordning i checklisteform i avsnitt 9.2–9.5. En checklista för själva utformningen av undersökningen finns i avsnitt 9.6. Observera att vår framställning är mycket förenklad. I praktiken är undersökningsplanering ett iterativt förfarande där olika beslut påverkar varandra på ett komplicerat sätt.

Figur 9.1.1 Utformning av en undersökning



9.2 Syfte

Syftet med en undersökning bör formuleras i nära samarbete med dess användare eller beställare. Initialt bör man diskutera på en ganska övergripande nivå, för att sedan gå in på detaljerna.

Övergripande syfte

- Vilket är sakproblemet?
- Vilken är intressepopulationen? (Se avsnitt 3.1.2.)
- Vilka är undersökningsobjekten? (Se avsnitt 2.1.)
- Vilka är de viktigaste undersökningsvariablerna? (Se avsnitt 2.1.)
- Vilka statistiska mått ska skattas? (Se avsnitt 2.3.1.)

Syften med undersökningen mer detaljerat

- Kan sakproblemet operationaliseras i en begränsad mängd precisa frågor? (Om så inte är fallet, bör en kvalitativ studie genomföras i stället för en urvalsundersökning, se avsnitt 6.2.1.)
- Vilken är målpopulationen? (Se avsnitt 3.1.2.)
- Hur ser eventuella redovisningsgrupper ut?
  - Hur definieras de?
  - Vad ska de användas till?
- Hur ser undersökningsvariablerna ut?
  - Hur definieras de?
  - Är de kvantitativa eller kategoriska? (Har betydelse för urvalsstorleken, se avsnitt 6.1.5.)
  - Vilka är deras referensperioder eller -tidpunkter? (Se avsnitt 3.4.2.)
  - Hur prioriterar beställaren de olika variablerna sinsemellan? (Se avsnitt 2.1.)
- Hur ser de statistiska måtten ut mer i detalj?
  - Ställer de särskilda krav på urvalsdesignen? (Se avsnitt 6.1.4-6.1.6.)
  - Gäller de jämförelser mellan grupper? (Påverkar urvalsdesignen och urvalsstorleken, se avsnitt 6.1.9.)
  - Involverar de brutto- eller nettoförändringar över tid? (Se kapitel 5.)

Ett bekvämt sätt att sammanställa kraven på statistiska mått, redovisningsgrupper och variabler är i form av en eller flera tabellplaner, se avsnitt 2.1. Det ger överblick och kan hjälpa till att identifiera tänkbara svårigheter och brister.

9.3 Ambitionsnivå

- Hur stor är budgeten? Hur ska resurserna allokeras? (Se avsnitt 6.1.2.)
- Vilka kvalitetskrav finns? De faktorer som framför allt påverkas av urvalsplanering är följande:
  - Vilken precision krävs för att uppfylla syftena, och vilken är den förmodade precisionen för varje skattning? (Påverkar urvalsstorleken, se avsnitt 6.1.5.)
  - Hur fort ska undersökningen göras?
  - (Se också avsnitt 6.2.1 för en diskussion om kvantitativa och kvalitativa undersökningar.)
- Ska några specialstudier genomföras för att mäta eller justera för bortfallsfel, mätfel e.d.? (Se avsnitt 4.7 och 7.1.)
- Om det finns redovisningsgrupper, hur många och hur stora är de, dvs. hur fin är indelningen? (Har betydelse för urvalsstorleken, se avsnitt 6.1.8.)
- Vilka regler bör följas?
  - För officiell statistik: se webbplatsen http://www.scb.se/templates/Listning2____47329.asp för länkar till regelverket för Sveriges officiella statistik (SOS).

  - För annan statistik än officiell statistik: vilka av kriterierna för tillräcklig kvalitet i officiell statistik är tillämpliga? Se Statistiska centralbyrån (2006, s. 11–14).

Ett viktigt kvalitetskrav är att statistiken har god tillförlitlighet, inte minst i form av god precision i skattningarna. De precisionskrav som finns kan infogas i tabellplanerna. Det är dock praktiskt att göra arbetet i två steg: att först konstruera en tabellplan med statistiska mått, sedan en plan med precisionskrav för dessa mått.

9.4 Externa faktorer

- Vilka naturliga strukturer innehåller målpopulationen? (Bör eventuellt återspeglas i urvalsdesignen, se avsnitt 6.2.5.)
- Hur heterogen är målpopulationen? (Påverkar urvalsstorleken, se avsnitt 6.1.5.)
- Vilka uppgiftskällor kan tillhandahålla uppgifter om objekten? (Se avsnitt 3.1.1.)
- Vilka är de tänkbara kontaktvägarna till uppgiftskällorna? Känner man till några särskilda svårigheter med att kontakta uppgiftskällorna och få uppgifter från dem?
- Behövs en pilotundersökning för att ta reda på mer om populationens sammansättning och om uppgiftskällorna? (Se avsnitt 6.5.)
- Vilka krav finns på att begränsa uppgiftslämnarbörda och -kostnad (särskilt för undersökningar som avser företag och andra organisationer).

9.5 Interna faktorer

- Vilka användbara register finns att tillgå? Kan de sammanfogas och användas till registerbaserad statistik (6.2.2)? Om så inte är fallet, kan de sammanfogas och omformas till en urvalsram ur vilken kontaktvägar till uppgiftskällorna kan bestämmas (kapitel 3)?
- Vilken sorts objekt finns i ramen och hur förhåller de sig till populationsobjekten? (Kan ha betydelse för val av urvalsmetod, se avsnitt 3.2.2.)
- Finns det hjälpinformation? (Hjälpinformation kan eventuellt användas i såväl urvalsdesign som skattningar, se kapitel 4 och avsnitt 6.4.)
- I de fall det finns redovisningsgrupper:
  - Kan redovisningsgrupperna identifieras i urvalsramen? (Detta påverkar undersökningens kvalitet, se avsnitt 6.1.8.)
  - Ska redovisningsgrupperna få utgöra strata? (Se avsnitt 4.2.5 om stratumkonstruktion samt avsnitt 6.1.8 om bestämning av urvalsstorlek.)
- Om en intervjuundersökning planeras, hur ligger fältarbetsperioden i förhållande till intervjuarkårens beläggning?

9.6 Undersökningsdesign

Utifrån de förutsättningar som kommer fram enligt avsnitt 9.1–9.5 fattas beslut om hur undersökningen ska utformas.

- Vilket slags undersökning ska göras, urvals- eller totalundersökning, registerbaserad undersökning eller kvalitativ studie? (Se avsnitt 6.2.1–6.2.3.)

Det förutsätts i det följande att en urvals- eller totalundersökning ska göras med datainsamling i (undersökningsplanerarens) egen regi.

- Vilken datainsamlingsmetod ska användas? (Se Biemer och Lyberg, 2003, kapitel 6.)
- I de fall avsikten är att göra en urvalsundersökning som ska löpa över tid (se kapitel 5):
  - Ska urvalsmetoden ta hänsyn till tidsaspekten (t.ex. samordnade urval)?
  - Hur hanteras objekt som föds, dör eller på annat sätt förändras?
- Vilken urvalsmetod ska användas, givet all information ovan? (Se avsnitt 6.2.5 och kapitel 4.)
- Hur kontrolleras att urvalet är korrekt draget? (Se avsnitt 6.6.)
- Hur hanteras bortfall, mätfel och ramfel?
  - Hur förebyggs bortfall i undersökningen? (Se Japec et al., 1997.)
  - Hur hanteras bortfall genom urval och estimation? (Se Särndal och Lundström, 2005.)
  - Hur korrigeras för misstänkta mätfel? (Se Biemer och Lyberg, 2003.)
  - Hur korrigerar vi för undertäckning? (Se Särndal och Lundström, 2005.)
  - Hur fördelas resurser mellan åtgärder som förebygger respektive justerar för olika typer av fel? (Se avsnitt 6.1.2.)
- Hur och när ska uppgiftskällan kontaktas?

10 Böcker om urvalsteori

Vi har försökt att gradera böckerna på en skala från ”lätt” till ”svår”. ”Lätt” betyder att boken kan läsas med måttliga förkunskaper om man lägger ned en rimlig arbetsinsats. Vi tror att de böcker som vi angett som lätta upplevs av de flesta universitetsstudenter med viss formelvana som lätta om de ingår i en statistikkurs på ungefär andraterminsnivå.

Foreman (1991) ger en introduktion till urvalsundersökningar. Foreman var metodchef på Australian Bureau of Statistics. Innehållet är i stort sett en delmängd av Cochran (1977), men mer modernt. Boken innehåller även ett översiktligt kapitel om planering av urvalsundersökningar. Foremans bok är välskriven och helt igenom praktiskt inriktad. Den innehåller ett minimum av formler.
Lätt. 497 sidor.

Brewer (2002) skiljer sig i framställningen kraftigt från andra böcker om urvalsundersökningar. I större delen av boken samtalar två noviser om tänkbara lösningar på ett urvals- och estimationsproblem: nämligen att skatta totalvikten av en flock elefanter. I boken finns ett avsnitt om skattning när urvalet innehåller outliers. Vi avstår från att försöka ange om boken är lätt eller svår. I sin stil är den lätt, rentav kåserande, men det är djupa tankar som den tar upp.
316 sidor.

Lohr (1999) används som kurslitteratur på många håll i världen. Den tar upp grundläggande teori för urvalsundersökningar. Jämfört med Särndal et al. (1992), se nedan, innehåller Lohr färre generella resultat. Det gör att Lohr får mindre räckvidd men blir enklare att läsa. Varje urvalsmetod diskuteras även kortfattat ur en modellberoende synvinkel (jämför med Valliant et al. (2000), vars huvudperspektiv är modellberoende). Boken innehåller två kapitel om enklare analys av surveydata.
Lätt–medelsvår. 494 sidor.

Kish (1965, 1987) innehåller enormt många praktiska råd. Författaren öser ur sin oerhört omfattande erfarenhet av urvalsundersökningar. Framställningen lider i båda böckerna av viss brist på struktur och kan fresta på läsarens tålamod.
Lätt–medelsvår. Kish (1965) har 643 sidor, och Kish (1987) har 266 sidor.

Rosén (2005) är ett kurskompendium. Det har använts som kurslitteratur i matematisk statistik i Uppsala och Stockholm för en första kurs i urvalsundersökningar. Framställningen är föredömligt strukturerad och tydlig. Kompendiet kräver vissa kunskaper i matematik, men inte nödvändigtvis mer än en eller två terminer i matematik på universitetsnivå. Med sin effektiva framställning förmedlar kompendiet ganska mycket stoff på ett relativt litet antal sidor. Kompendiet kan både användas som en introduktion till ämnet och som fördjupning för den som läst t.ex. Cochran (1977) eller Foreman (1991). Det har fortfarande några tomma avsnitt, men det stör inte läsningen.
Medelsvår. 184 sidor.

Cochran (1977) är i dag något föråldrad men fortfarande mycket läsvärd. Den har under många år varit standardboken för urvalsundersökningar. Boken innehåller mängder av praktiska råd, som man kanske bäst uppskattar efter några års praktiskt arbete som metodstatistiker. Den är inte lika välstrukturerad som Särndal et al. (1992), vilket främst beror på att den sammanhållande teori som den senare boken använder och förmedlar inte fanns då Cochran skrev sin bok. Många läsare kan finna bokens sätt att förmedla teori något otydligt och besvärligt att förstå.
Medelsvår. 428 sidor.

Särndal et al. (1992) är i dag standardboken över stora delar av världen i lite mer avancerad teori för urvalsundersökningar. Den är också standardboken i modellassisterad inferens, som är det vanliga på stora statistikbyråer i världen. Med sitt omfång och sin plats i metodstatistikers medvetande gör den skäl för uttrycket ”bibel”. Även om boken är pedagogisk och inte formellt förutsätter några förkunskaper i urvalsundersökningar, måste läsaren räkna med en rejäl ingångströskel att ta sig över, särskilt om förkunskaperna är måttliga. Boken kräver även vissa kunskaper i matematik, eftersom många av resultaten presenteras i generell form. Den typ av matematik som används är till övervägande delen algebra. Om läsaren kompletterar med Särndal och Lundström (2005) har hon eller han en väldigt bra teoretisk grund för arbete med urval och estimation.
Medelsvår–svår. 694 sidor.

Valliant et al. (2000) håller sig helt till modellberoende inferens. Boken propagerar för att urval ska dras balanserat och att estimation ska grundas på en modell där det som anses vara slumpmässigt ges av modellen och inte av urvalet. Varje kapitel innehåller ett avsnitt som polemiserar mot modellassisterad och designbaserad inferens (dvs. det som vi inom SCB håller oss till för det mesta). För en metodstatistiker på SCB utmanar boken (inom SCB) rådande föreställningar. I likhet med Brewer (2002) finns ett avsnitt om estimation när urvalet innehåller outliers. Till skillnad från många andra böcker pekar den ut problem som i dagsläget saknar en bra lösning. Boken är välskriven och framställningarna är tydliga. Den innehåller många exempel där resonemang illustreras grafiskt. I notationen och beräkningarna har matriser en framträdande plats.
Medelsvår–svår. 504 sidor.

Att läsa Thompson (1997) ger mycket, men kräver eftertanke på nästan varje sida. Den vänder sig till läsare som redan kan mycket om urvalsundersökningar och som vill ha teoretisk fördjupning. Bland de ämnen som den tar upp och som inte finns med i många andra böcker om urvalsundersökningar, märks ett kapitel om analys av surveydata och ett om urval i tid och rum.
Svår. 305 sidor.

Bilaga A: Härledning av populations-variansen för en blandad fördelning (formel 6.1.27)

Låt *X* vara en binär stokastisk variabel som antar värdet 0 med sannolikhet p_c och värdet 1 med sannolikhet $p_d = 1 - p_c$. Låt vidare *Z* vara en stokastisk variabel med den blandade fördelningen:

fördelning C om *X*=0
fördelning D om *X*=1,

där C och D är godtyckliga fördelningar.

Nu ska härledas ett uttryck för variansen av *Z*. Använd formeln

$$V(U) = EV(U|Y) + VE(U|Y) \tag{1}$$

som gäller för alla stokastiska variabler *U* och *Y*.

Per definition gäller för varje *U* och *Y*

$$V(U|Y) = E(U^2|Y) - [E(U|Y)]^2 . \tag{2}$$

Om *Y* är en binär stokastisk variabel gäller

$$E(U) = EE(U|Y) = p_c E(U|Y=0) + p_d E(U|Y=1) \tag{3}$$

där $P(Y=0) = p_c$ och $P(Y=1) = p_d$.

Från (3):

$$E(Z^2) = EE(Z^2|X) = p_c E(Z^2|X=0) + p_d E(Z^2|X=1) \tag{4}$$

Genom att sätta $U = [E(Z|X)]^2$ och *Y* = *X* i (3), fås även

$$E[E(Z|X)]^2 = p_c [E(Z|X=0)]^2 + p_d [E(Z|X=1)]^2 \tag{5}$$

Från (2):

$$EV(Z|X) = EE(Z^2|X) - E[E(Z|X)]^2 \tag{6}$$

Med (4) och (5) insatta i (6) får man ut den första termen i (1):

$$\begin{aligned} &EV(Z|X) \\ &= p_c E(Z^2|X=0) + p_d E(Z^2|X=1) \\ &- p_c [E(Z|X=0)]^2 - p_d [E(Z|X=1)]^2 \end{aligned} \tag{7}$$

Enligt definitionen av varians är

$$VE(Z|X) = E\left[E(Z|X)\right]^2 - \left[EE(Z|X)\right]^2 \quad (8)$$

Genom att sätta in (3) och (5) i (8) blir den andra termen i (1):

$$\begin{aligned} & VE(Z|X) \\ &= p_c\left[E(Z|X=0)\right]^2 + p_d\left[E(Z|X=1)\right]^2 \\ &\quad - \left\{p_c E(Z|X=0) + p_d E(Z|X=1)\right\}^2 \end{aligned} \quad (9)$$

Summan av (7) och (9) blir

$$\begin{aligned} & V(Z) \\ &= p_c E(Z^2|X=0) + p_d E(Z^2|X=1) \\ &\quad - \left\{p_c E(Z|X=0) + p_d E(Z|X=1)\right\}^2 \end{aligned} \quad (10)$$

Notera att eftersom $p_d = 1 - p_c$:

$$\begin{aligned} & \left\{p_c E(Z|X=0) + p_d E(Z|X=1)\right\}^2 \\ &= p_c^2\left[E(Z|X=0)\right]^2 + p_d^2\left[E(Z|X=1)\right]^2 + 2p_c p_d E(Z|X=0)E(Z|X=1) \\ &= -p_c p_d\left[E(Z|X=0) - E(Z|X=1)\right]^2 \\ &\quad + p_c\left[E(Z|X=0)\right]^2 + p_d\left[E(Z|X=1)\right]^2 \end{aligned} \quad (11)$$

Ur (2) fås $p_c V(Z|X=0) = p_c E(Z^2|X=0) - p_c\left[E(Z|X=0)\right]^2$ och motsvarande för $X = 1$. Från (10) och (11) får vi då att

$$\begin{aligned} V(Z) &= p_c V(Z|X=0) + p_d V(Z|X=1) \\ &\quad + p_c p_d\left[E(Z|X=0) - E(Z|X=1)\right]^2 \end{aligned} \quad (12)$$

Formlerna (10) och (12) är två ekvivalenta uttryck för variansen av *Z*. Eftersom (10) innehåller ett väntevärde av Z^2 kan (12) vara mer praktisk att använda.

Populationsvariansen S_{yU}^2 kan betraktas som en skattning av modell-variansen $V(Y)$ under en superpopulationsmodell.

Referenslista

Andersson, C. (1994). On the use of two-phase sampling where data contains misclassification and measurement errors. *Acta Universitatis Upsaliensis*, Studia Statistica Upsaliensia 3, Uppsala.

Andersson, C., Lindström, H.L. och Polfeldt, T. (1999). *Att mäta statistikens kvalitet.* R&D Report 1999:3. Statistiska centralbyrån.

Arnab, R. (1991). Sampling on Two Occasions. Estimation of a Population Total. *Survey Methodology*, 24, 185–192.

Atmer, J., Thulin, G. och Bäcklund, S. (1975). Samordning av urval med JALES-metoden, *Statistisk Tidskrift*, 13, 443–450.

Bailar, B. A. (1975). The Effects of Rotation Group Bias on Estimates from Panel Surveys. *Journal of the American Statistical Association*, 70, 23–30.

Bankier, M.D. (1988). Power Allocations: Determining Sample Sizes for Subnational Areas. *The American Statistician*, 42, 174–177.

Bassi, F., Torelli, N. och Trivellato, U. (1998). Data and Modelling Strategies in Estimating Labour Force Gross Flows Affected by Classification Errors. *Survey Methodology*, 24, 109–122.

Basu, D. (1971). An Essay on the Logical Foundations of Survey Sampling. I *Foundations of Statistical Inference* (red. V.P. Godambe och D.A. Sprott). Toronto: Holt, Rinehart and Winston, 203–242.

Bettio, M., Delincé, J., Bruyas, P., Croi, W. och Eiden, G. (2002). Area frame surveys: Aim, Principals and Operational Surveys. *EC report EUR 20521*, Eurostat, 12–27.

Biemer, P.P. och Lyberg, L.E. (2003). *Introduction to Survey Quality.* New York: Wiley.

Binder, D.A. (1998). Longitudinal Surveys: Why are these Surveys Different from all other Surveys? *Survey Methodology*, 24, 101–108.

Boynton, I.M. (2003). *AKU:s resultatredovisning och publiceringsformer: fokusgrupper med användare*. PM 2003-11-21. Statistiska centralbyrån.

Boynton, I.M. (2004a). What's Behind an Answer? [cd-rom] *Proceedings, European Conference on Quality and Methodology in Official Statistics (Q2004)*. Mainz, 24–26 maj.

Boynton, I.M. (2004b). *Introduktion för invandrare vid Stockholms invandrarmottagning: fokusgrupper med invandrare.* PM 2004-06-23. Mättekniska laboratoriet, Statistiska centralbyrån.

Boynton, I.M. och Mujagic, H. (2002). *Gymnasieungdomars studieintresse: test av introduktionsbrev och enkät.* PM 2002-12-30. Statistiska centralbyrån.

Bremer, J., Terhanian, G. och Strange, P. (2004). *Propensity Score Matching as a Bias Correction Method for Internet-Based Studies.* Presentation på Joint Statistical Meetings, Toronto, 8-12 augusti.

Brewer, K.R.W. (1994). Survey Sampling Inference: Some Past Perspectives and Present Prospects. *Pakistan Journal of Statistics*, 10, 213–233.

Brewer, K.R.W. (1999). Design-Based or Prediction-Based Inference? Stratified Random vs Stratified Balanced Sampling. *International Statistical Review*, 67, 35–47.

Brewer, K.R.W. (2002). *Combined Survey Sampling Inference: Weighing Basu's Elephants*. London: Arnold.

Brewer, K.R.W. och Haniff, M. (1983). *Sampling with Unequal Probabilities*. New York: Springer-Verlag.

Brånvall, G., Mark-Berglund, T., Polfeldt, T. och Vorwerk, P. (odaterad). *Handledning i miljöstatistik för miljöövervakning.* PM. Statistiska centralbyrån.

Chambers, R. L. och Skinner, C. J. (red.) (2003). *Analysis of Survey Data.* Chichester: Wiley.

Chauvet, C. och Tillé, Y. (2005). *New SAS Macros for Balancing Samples, nr 323* [online]. Institute National de la Statistique et des Études Économiques (INSEE). [citerad 2007-09-21]. Tillgänglig som: <http://jms.insee.fr/site/files/documents/2005/>.

Chauvet, C. och Tillé, Y. (2006). A Fast Algorithm for Balanced Sampling. *Computational Statistics*, 21, 53–62.

Cochran, W.G. (1977). *Sampling Techniques*, 3rd ed. New York: Wiley.

Cox, B.G., Binder, D.A., Chinnappa, B.N., Christianson, A., Colledge, M.J. och Kott, P.S. (red.) (1995). *Business Survey Methods.* New York: Wiley.

Dalén, J. (1986). Sampling from Finite Populations: Actual Coverage Probabilities for Confidence Intervals on the Population Mean. *Journal of Official Statistics*, 2, 13–24.

Dalenius, T. (1957). *Sampling in Sweden.* Stockholm: Almquist och Wiksell.

Dalenius, T. (1985). *Elements of Survey Sampling.* Stockholm: Sarec.

Danielsson, S. (1975). *Optimal allokering vid vissa klasser av urvalsförfaranden*. Doktorsavhandling, Linköpings högskola.

Derr, J. (1999). *Statistical Consulting: A Guide to Effective Communication.* Pacific Grove, CA: Duxbury.

Deville, J.-C. (1991). A Theory of Quota Surveys. *Survey Methodology*, 17, 163–181.

Deville, J.-C. och Tillé, Y. (2004). Efficient Balanced Sampling: the Cube Method. *Biometrika*, 91, 893–912.

Drew, J.D., Bélanger, Y. och Foy, P. (1985). Stratification in the Canadian Labour Force Survey. *Survey Methodology*, 11, 95–110.

Duncan, G. J. och Kalton, G. (1987). Issues of Design and Analysis of Surveys Across Time. *International Statistical Review*, 55, 97–117.

Ekman, G. (1959). An Approximation Useful in Univariate Stratification. *The Annals of Mathematical Statistics*, 30, 219–229.

Eriksson, J., Paulson, P.A. och Rosén, B. (2003). *Översyn av undersökningen Inrikes och utrikes trafik med svenska lastbilar.* R&D Report 2003:1. Statistiska centralbyrån.

Ernst, L.R., Valliant, R. och Casady, R.J. (2000). Permanent and Collocated Random Number Sampling and the Coverage of Births and Deaths. *Journal of Official Statistics*, 16, 211–228.

European communities (2003). *The Lucas Survey – European Statisticians Monitor Territory*. Updated edition – June 2003.

Fabrizi, E. och Trivisano, C. (2006). Efficient Stratification Based on Non-Parametric Regression Methods. *Journal of Official Statistics*, 23, 35–50.

Félix-Medina, M.H. och Thompson, S.K. (2004). Adaptive Cluster Double Sampling. *Biometrika*, 91, 877–891.

Folsom, R., LaVange, L. och Williams, R.L. (1989). A Probability Sampling Perspective on Panel Data Analysis. I *Panel Surveys* (red. D. Kasprzyk, G.J. Duncan, G. Kalton och M.P. Singh). New York: Wiley, 108–138.

Ford, B.L. och Tortora, R.D. (1978). A Consulting Aid to Sample Design. *Biometrics*, 34, 299–304.

Foreman, E.K. (1991). *Survey Sampling Principles.* New York: Marcel Dekker.

Granquist, L., Arvidson, G., Elffors, C., Norberg, A. och Lundell, L.-G. (2002). *Guide till granskning*. CBM 2002:1. Statistiska centralbyrån.

Gunning, P. och Horgan, J.M. (2004). A New Algorithm for the Construction of Stratum Boundaries in Skewed Populations. *Survey Methodology*, 30, 159–166.

Hansen, M. H., Hurwitz, W. N., Marks, E. S. och Mauldin, W. P. (1951). Response Errors in Surveys. *Journal of the American Statistical Association*, 46, 147–190.

Hansen, M.H, Madow, W.G. och Tepping, B.J. (1983). An Evaluation of Model-Dependent and Probability Sampling Inferences in Sample Surveys. *Journal of the American Statistical Association*, 78, 776–793.

Hedlin, D. (1998). *On the Stratification of Highly Skewed Populations.* R&D Report 1998:3. Statistiska centralbyrån.

Hedlin, D. (2000). A Procedure for Stratification by an Extended Ekman Rule. *Journal of Official Statistics*, 16, 15–29.

Hedlin, D., Dale, T., Haraldsen, G. och Jones, J. (red.) (2005). *Developing Methods for Assessing Perceived Response Burden.* A joint report of Statistics Sweden, Statistics Norway and the UK Office for National Statistics.

Hedlin, D., Falvey, H., Chambers, R. och Kokic, P. (2001). Does the Model Matter for GREG Estimation? A Business Survey Example. *Journal of Official Statistics*, 17, 527–544.

Hidiroglou, M.A., Choudhry, G.H. och Lavallée, P. (1991). A Sampling and Estimation Methodology for Sub-Annual Business Surveys. *Survey Methodology*, 17, 195–210.

Hidiroglou, M.A. och Laniel, N. (2001). Sampling and Estimation Issues for Annual and Sub-Annual Canadian Business Surveys. *International Statistical Review*, 69, 487–504.

Hidiroglou, M.A. och Srinath K.P. (1993). Problems Associated with Designing Sub-Annual Business Surveys. *Journal of Business and Economic Statistics*, 11, 397–405.

Hidiroglou, M.A., Särndal, C.-E. och Binder, D.A. (1995). Weighting and Estimation in Business Surveys. I *Business Survey Methods* (red. B. Cox, D. Binder, N. Chinnappa, A. Christianson, M. Colledge och P. Kott). New York: Wiley, 477–502.

Holmberg, A. (2002). *On the Choice of Sampling Design under GREG Estimation in Multiparameter Surveys*. R&D Report 2002:1. Statistiska centralbyrån.

Holmberg, A. (2003). *Essays on Model Assisted Survey Planning*. Acta Universitatis Upsaliensis. Uppsala universitet.

Holmberg, A., Flisberg, P. och Rönnqvist M. (2002). *On the Choice of Sampling Design in Business Surveys with Several Important Study Variables*. R&D Report 2002:3. Statistiska centralbyrån.

Holme, I.M. och Solvang, B.K. (1997). *Forskningsmetodik: om kvalitativa och kvantitativa metoder.* Lund: Studentlitteratur.

Holt, D. och Skinner C.J. (1989). Components of Change in Repeated Surveys. *International Statistical Review*, 57, 1–18.

Holt, D. och Smith, T.M.F. (1979). Post Stratification. *Journal of the Royal Statistical Society, Series A*, 142, 33–46.

Horgan, J.M. (2006). Stratification of Skewed Populations: A Review. *International Statistical Review*, 74, 67–76.

Japec, L, Ahtiainen, A., Hörngren, J., Lindén, H., Lyberg, L. och Nilsson, P. (1997). *Minska bortfallet*. Statistiska centralbyrån.

Kalton, G. (1991). Sampling Flows of Mobile Human Populations. *Survey Methodology*, 17, 183–194.

Kalton, G. och Anderson, D.W. (1986). Sampling Rare Populations. *Journal of the Royal Statistical Society, Series A*, 149, 65–82.

Kalton, G. och Citro, C.F. (1993). Panel Surveys: Adding the Fourth Dimension. *Survey Methodology*, 19, 205–215.

Kalton, G., Kasprzyk, D. och McMillen, D.B. (1989). Nonsampling Errors in Panel Surveys. I *Panel Surveys* (red. D. Kasprzyk, G. J. Duncan, G. Kalton, M.P. Singh). New York: Wiley, 249–270.

Kish, L. (1965). *Survey Sampling.* New York: Wiley.

Kish, L. (1987). *Statistical Design for Research.* New York: Wiley.

Kish, L. (1988). Multipurpose Sample Designs. *Survey Methodology*, 14, 19–32.

Kish, L. (1998). Quota Sampling: Old Plus New Thought [online]. [citerad: 2006-05-26]. Tillgänglig som: <http://www.websm.org/uploadi/editor/1132389572kish.doc>.

Kjaer Jensen, M. (1994). *Kvalitativa metoder.* Lund: Studentlitteratur.

Kröger, H., Särndal, C.E. och Teikari, I. (1999). Poisson Mixture Sampling: A Family of Designs for Coordinated Selection Using Permanent Random Numbers. *Survey Methodology*, 25, 3–11.

Kröger, H., Särndal, C.E. and Teikari, I. (2003). Poisson Mixture Sampling Combined with Order Sampling. *Journal of Official Statistics*, 19, 59–70.

Kvale, S. (1997). *Den kvalitativa forskningsintervjun.* Lund: Studentlitteratur.

Körner, T. och Nimmergut, A. (2004). Using an Access Panel as a Sampling Frame for Voluntary Household Surveys. *Statistical Journal of the United Nations*, 21, 33–52.

Lavallée, P. (1988). Two-Way Optimal Stratification Using Dynamic Programming. *American Statistical Association, Proceedings of the section on survey research methods*, 646–651.

Lavallée, P. och Hidiroglou, M.A. (1988). On the Stratification of Skewed Populations. *Survey Methodology*, 14, 33–43.

Lessler, J.T. och Kalsbeek, W.D. (1992). *Nonsampling Error in Surveys*. New York: Wiley.

Lindblom, A. och Nordberg, L. (2004). *On adjustment for coverage problems in short-term business surveys* [cd-rom]. *Proceedings, European Conference on Quality and Methodology in Official Statistics (Q2004)*. Mainz, 24–26 maj.

Lohr, S. (1999). *Sampling: Design and Analysis.* Pacific Grove, CA: Duxbury.

Lu, W. och Sitter, R.R. (2002). Multi-way Stratification by Linear Programming Made Practical. *Survey Methodology*, 28, 199–207.

Lundström, S. och Särndal, C.-E. (2001). *Estimation in the Presence of Nonresponse and Frame Imperfections.* Statistiska centralbyrån.

Mecatti, F. (2007). A Single Frame Multiplicity Estimator for Multiple Frame Surveys. *Survey Methodology*, 33, 151–157.

Mirza, H. och Hörngren, J. (2002). *The Sampling and the Estimation Procedure in the Swedish Labour Force Survey.* R&D Report 2002:4. Statistiska centralbyrån.

Neyman, J. (1934). On the Two Different Aspects of the Representative Method: the Method of Stratified Sampling and the Method of Purposive Selection. *Journal of the Royal Statistical Society*, 97, 558–606.

Nimmergut, A. (2005). *The Permanent Sample in the FSO Germany.* Presentation på International Household Survey Nonresponse Workshop, Tällberg, 28–31 augusti.

Nimmergut, A. (2006). *Data Quality Assessment of Low Response Rate Situations.* Presentation på European Conference on Quality in Survey Statistics (Q2006), Cardiff, 24–26 april.

Nordberg, L. (1989). Generalized Linear Modeling of Sample Survey Data. *Journal of Official Statistics*, 5, 223–239.

Nordberg, L. (2000). On Variance Estimation for Measures of Change when Samples are Coordinated by the Use of Permanent Random Numbers. *Journal of Official Statistics*, 16, 363–378.

Ohlsson, E. (1990). *Sequential Poisson Sampling from a Business Register and its Application to the Swedish Consumer Price index.* R&D Report 1990:6. Statistiska centralbyrån.

Ohlsson, E. (1992). *SAMU – The System for Co-Ordination of Samples from the Business Register at Statistics Sweden: A Methodological Description.* R&D Report 1992:18. Statistiska centralbyrån.

Ohlsson, E. (1995). Coordination of Samples Using Permanent Random Numbers. I *Business Survey Methods* (red. B. Cox, D. Binder, N. Chinnappa, A. Christianson, M. Colledge och P. Kott). New York: Wiley, 153–169.

Ohlsson, E. (1998). Sequential Poisson Sampling. *Journal of Official Statistics*, 14, 149–162.

O'Muircheartaigh, C. (1989). Sources of Nonsampling Error: Discussion. I *Panel Surveys* (red. D. Kasprzyk, G.J. Duncan, G. Kalton, M.P. Singh). New York: Wiley, 271–288.

Paik, M. (1985). A Graphic Representation of a Three-Way Contingency Table: Simpson's Paradox and Correlation. *The American Statistician*, 39, 53–55.

Petri, C. (2005). *Om resursallokeringsproblemet i en survey* [online]. Magisteruppsats. Statistiska centralbyrån. [Citerad 2007-10-15]. Tillgänglig som: <http://sps01/Intranet/Document%20Library/Stöd/Statistiskt%20metodarbete/Metodrapporter/Om%20resursallokeringsproblemet%20i%20en%20survey.pdf>.

Petric, M. (2007). *Samordning av urval mellan undersökningarna LINDA, HEK och HUT*. Dokumentation, februari 2007. Statistiska centralbyrån.

Purcell, N.J. och Kish, L. (1979). Estimation for Small Domains. *Biometrics*, 35, 365–384.

Rao, J.N.K, (2003). *Small Area Estimation.* New York: Wiley.

Rao, J.N.K, Jocelyn, W. och Hidiroglou, M.A. (2003). Confidence Interval Coverage Properties for Regression Estimators in Uni-Phase and Two-Phase Sampling. *Journal of Official Statistics*, 19, 17–30.

Ribe, M. (1995). Vem behöver siffror? [online]. *Välfärdsbulletinen*, 25, nr 4, 20–21. [citerad]. Tillgänglig under "Fokusområden\Metodinformation\Statistikskolan" i: <www.scb.se>.

Ringvall, A. (2000). *Assessment of sparse populations in forest inventory. Development and evaluation of probabiltiy sampling methods*. Acta Universitatis Agriculturae Sueciae, Silvestria 151. Swedish University of Agricultural Sciences, Umeå.

Ringvall, A., Fridman, J., Lämås, T. och Ståhl, G. (2000). Inventering av död ved – några objektiva inventeringsmetoder. *Fakta Skog*, nr 1. Sveriges Lantbruks-universitet.

Ritchie, J. och Lewis, J. (red.) (2005). *Qualitative Research Practice: A Guide for Social Science Students and Researchers.* Thousand Oaks, CA: Sage.

Rivest, L.-P. (2002). A Generalization of the Lavallée and Hidiroglou Algorithm for Stratification in Business Surveys. *Survey Methodology*, 28, 191–198.

Rocco, E. (2003). Constrained Inverse Adaptive Cluster Sampling. *Journal of Official Statistics*, 19, 45–57.

Rosén, B. (1986). *Om HINK-undersökningens design- och allokeringsproblematik.* R&D Report 29, Statistiska centralbyrån.

Rosén, B. (1991). *Variance estimation for systematic pps-sampling.* R&D Report 1991:15. Statistiska centralbyrån.

Rosén, B. (1996). *On Sampling with Probability Proportional to Size.* R&D Report 1996:1. Statistiska centralbyrån.

Rosén, B. (1997a). Asymptotic Theory for Order Sampling. *Journal of Statistical Planning and Inference*, 62, 135–158.

Rosén, B. (1997b). On sampling with Probability Proportional to Size. *Journal of Statistical Planning and Inference*, 62, 159–191.

Rosén, B. (2000a). *Generalized Regression Estimation and Pareto πps.* R&D Report 2000:5. Statistiska centralbyrån.

Rosén, B. (2000b). *A User's Guide to Pareto πps Sampling.* R&D Report 2000:6. Statistiska centralbyrån.

Rosén, B. (2000c). On Inclusion Probabilities Order πps Sampling. *Journal of Statistical Planning and Inference*, 90, 117–143.

Rosén, B. (2005). *Teori och praktik för urvalsundersökningar.* Kurskompendium. Matematisk statistik, Uppsala universitet.

Rosén, B. och Lundqvist, P. (1998). *Estimation from Order πps Samples with Non-Response.* R&D Report 1998:5. Statistiska centralbyrån.

Rosén, B. och Zamani, M. (1993). *Översyn av lastbilstransporter i Sverige (UVAV).* R&D Report 1993:2. Statistiska centralbyrån.

Roshwalb, A. och Wright R.L. (1991). Using Information in Addition to Book Value in Sample Designs for Inventory Cost Estimation. *The Accounting Review*, 66, 348–360.

Salehi, M.M. och Seber, G.A.F (2004). A General Inverse Sampling Scheme and its Applications to Adaptive Cluster Sampling. *Australian and New Zeeland Journal of Statistics*, 46, 483–494.

Sedlmeier, P. och Gigerenzer, G. (1997). Intuitions About Sample Size: The Empirical Law of Large Numbers. *Journal of Behavioral Decision Making*, 10, 33–51.

Sigman, R. och Monsour, N. (1995). Selecting samples from list frames of businesses. I *Business Survey Methods* (red. B. Cox, D., Binder, N. Chinnappa, A. Christianson, M. Colledge och P. Kott). New York: Wiley.

Sirken, M.G. (1998). A Short History of Network Sampling [online]. Proceedings of the Survey Methods Research Section, American Statistical Association, 1–6. [citerad: 2006-09-04]. Tillgänglig som: <www.amstat.org/sections/SRMS/proceedings/papers/1998_001.pdf>.

Skinner, C.J., Holmes, D.H. och Holt, D. (1994). Multiple Frame Sampling for Multivariate Stratification. *International Statistical Review*, 62, 333–347. New York: Wiley.

Skinner, C.J., Holt, D. och Smith, T.M.F. (red.) (1989). *Analysis of Complex Surveys.* Chichester: Wiley.

Smith, T.M.F. (1983). Comment. Diskussion av "An Evaluation of Model-Dependent and Probability Sampling Inferences in Sample Surveys" av Hansen, Madow and Tepping. *Journal of the American Statistical Association*, 78, 801–802.

Smith, T.M.F. (1994). Sample Surveys 1975–1990. An Age of Reconciliation? *International Statistical Review*, 62, 5–19.

Smith, T.M.F. (1997). Social Surveys and Social Science. *Canadian Journal of Statistics*, 25, 23–44.

Smith, T.M.F. (2001). Biometrika Centenary: Sample Surveys. *Biometrika*, 88, 167–194.

Stanley McCarthy, J., Beckler, D.G. och Qualey, S.M. (2006). An Analysis of the Relationship Between Survey Burden and Nonresponse: If We Bother Them More, Are They Less Cooperative? *Journal of Official Statistics*, 22, 97–112.

Statistics Canada (2003). *Survey Methods and Practices.* ISBN 0-660-19050-8.

Statistiska centralbyrån (2001a). *Kvalitetsbegrepp och riktlinjer för kvalitetsdeklaration av officiell statistik.* Meddelanden i samordningsfrågor för Sveriges officiella statistik, MIS 2001:1.

Statistiska centralbyrån (2001b). *Standard för institutionell sektorindelning 2000, INSEKT 2000.* Meddelanden i samordningsfrågor för Sveriges officiella statistik, MIS 2001:2.

Statistiska centralbyrån (2002). *Företagsenheten i den ekonomiska statistiken.* Bakgrundsfakta till ekonomisk statistik, 2002:3.

Statistiska centralbyrån (2003). *SAMU: The system for co-ordination of frame populations and samples from the Business Register at Statistics Sweden.* Background facts on Economic Statistics, 2003:3.

Statistiska centralbyrån (2005). *Flow Statistics from the Swedish Labour Force Survey.* Background facts on Labour and Education Statistics, 2005:5.

Statistiska centralbyrån (2006). *Tillräcklig kvalitet och kriterier för officiell statistik.* Riktlinjer från Rådet för den officiella statistiken, 2006:1.

Stephan, F. F. och McCarthy, P. J. (1958). *Sampling Opinions.* New York: Wiley.

Stoop, I. (2004). Access Pools as a Solution to the Nonresponse Problem? [online]. Presentation på *15th International Workshop on Survey Nonresponse*, Maastricht, 23–25 augusti. [citerad 2006-05-26]. Tillgänglig som: <http://cbs-rndtest.vb.cbs.nl/workshop-nonresponse>.

Svensson, J. (2007). Measurement bias adjustment in the Swedish farm accidents survey [cdrom]. *Proceedings of the Third International Conference on Establishment Surveys* (ICES-III), Montréal, 18–21 juni.

Synovate Temo (2006). *Webbpanelen* [online]. [Citerad 2006-05-26]. Tillgänglig som: <http://www.temo.se/Templates/Page____162.aspx>.

Särndal, C-E. och Lundström, S. (2005). *Estimation in Surveys with Nonresponse.* New York: Wiley.

Särndal, C.-E., Swensson, B. och Wretman J. (1992). *Model Assisted Survey Sampling.* New York: Springer-Verlag.

Taylor, S.J. och Bogdan, R. (1998). *Introduction to Qualitative Research Methods: A Guidebook and Resource*, 3rd ed. New York: Wiley.

Thompson, M.E. (1997). *Theory of Sample Surveys.* London: Chapman & Hall.

Thompson, S. K. (1992). *Sampling*. New York. Wiley.

Thorburn, D. (2006). *CBM Urval: synpunkter vid Vetenskapliga Rådets möte på Steningevik, den 30 mars 2006*. Stockholms universitet.

Tillé, Y. (2006). *Sampling Algorithms.* New York: Springer-Verlag.

Tillé, Y. och Favre, A.-C. (2004). Coordination, Combinaton and Extension of Balanced Samples. *Biometrika*, 91, 913–927.

Tångdahl, S. (2006). *Urval av koordinater för jord- och grödprovtagning.* PM. Statistiska centralbyrån.

Valliant, R., Dorfman, A.H. och Royall, R.M. (2000). *Finite Population Sampling and Inference: A Prediction Approach*. New York: Wiley.

Wainer, H. (1986). Minority Contributions to the SAT Score Turnaround: An Example of Simpson's Paradox. *Journal of Educational Statistics*, 11, 239–244.

Wallgren, A. och Wallgren, B. (2004). *Registerstatistik – administrativa data för statistiska syften.* R&D Report 2004:2. Statistiska centralbyrån.

Wallgren, A. och Wallgren, B. (2007). *Register-based Statistics – Administrative Data for Statistical Purposes.* New York: Wiley.

Wibeck, V. (2000). *Fokusgrupper: om fokuserade gruppintervjuer som undersöknings-metod.* Lund: Studentlitteratur.

Wolter, K. (1985). *Introduction to Variance Estimation.* New York: Springer-Verlag.

In English

Summary

This handbook aims at bridging the gap between sample survey theory and sampling in practice. The focus is on sampling design rather than estimation and other issues in survey methodology.

The handbook is primarily directed at survey methodologists looking for an introduction or a closer insight into the practical side of sampling design. A prerequisite is a traditional undergraduate course in sampling theory such as Scheaffer, R.L., Mendelhall, W., and Ott, R.L. (2005), Elementary Survey Sampling, Duxbury, or comparable.

Following an introductory Chapter 1, an overview of survey planning is given in Chapter 2. Chapter 3 addresses differences between a register and a sampling frame. Here we discuss population of interest, target and frame populations and the timing of the sampling frame construction.

Chapter 4 contains guidelines for methods such as simple random sampling, stratification and sampling in two or more stages or phases. Section 4.4 on probability proportional to size (πps) sampling covers systematic πps sampling as well as PoMix and Pareto πps sampling methods. We also bring up non-probability sampling methods and sampling from continuous populations.

Chapter 5 deals with coordination of samples, frames and surveys. Here we describe panel sampling as well as coordination with permanent random numbers (PRN). Section 5.3 addresses the issue of components of change in repeated surveys. Section 5.5 is devoted to the SAMU, a PRN business survey coordination system used at Statistics Sweden.

In Chapter 6 we consider the choice of sampling method and determination of sample size, including the closely linked issue of population variance estimation. We also discuss the relative merits of sample surveys, censuses, register-based surveys and qualitative studies.

The handbook includes a variety of examples taken from surveys conducted by Statistics Sweden. Chapter 7, in particular, is devoted to five surveys where methods and ideas described in previous chapters, are exhibited in some detail. Chapters 8–10 contain some brief advice on statistical consulting, a checklist for survey design and short summaries of books on sampling methods and theory. References give hints on further reading.

For additional information, see "Inquiries" in the beginning of the handbook.

Sakregister

Urvalsmetoder är centrala i statistiska undersökningar. Det finns åtskilliga läroböcker i ämnet. Syftet med denna handbok, Urval – från teori till praktik, är att överbrygga det av många upplevda gapet mellan teori och praktik när det gäller urvalsmetodik. Boken är avsedd som vägledning vid urvalsplanering för sådana statistiska undersökningar som utförs vid Statistiska centralbyrån.

ISSN 1654-7268 (online)
ISSN 1654-7276 (print)
ISBN 978-91-618-1412-1 (print)

Publikationstjänsten:
E-post: publ@scb.se, tfn: 019-17 68 00, fax: 019-17 64 44. Postadress: 701 89 Örebro.

Information och bibliotek: E-post: information@scb.se, tfn: 08-506 948 01, fax: 08-506 948 99.
Försäljning över disk, besöksadress: Biblioteket, Karlavägen 100, Stockholm.

Publication services:
E-mail: publ@scb.se, phone: +46 19 17 68 00, fax: +46 19 17 64 44. Address: SE-701 89 Örebro.

Information and library: E-mail: information@scb.se, phone: +46 8 506 948 01, fax: +46 8 506 948 99
Over-the-counter sales: Statistics Sweden, Library, Karlavägen 100, Stockholm, Sweden.

www.scb.se