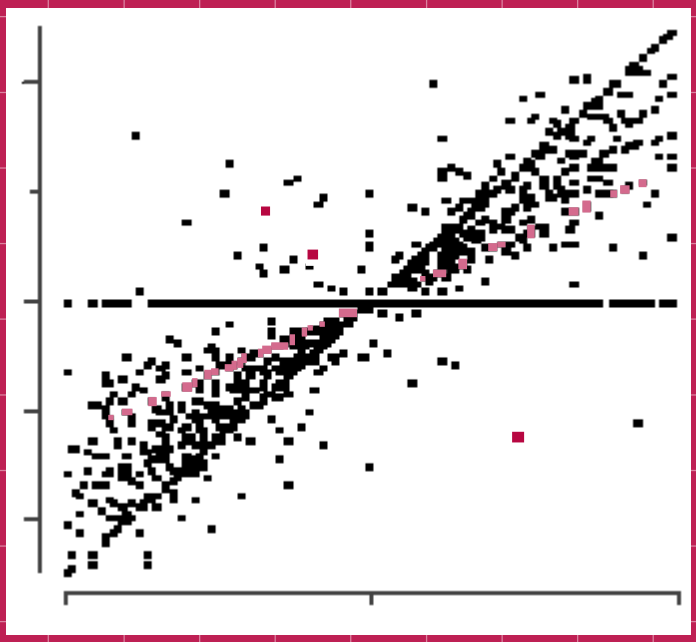
CBM 2002:1

SCB

Statistiska centralbyrån Statistics Sweden

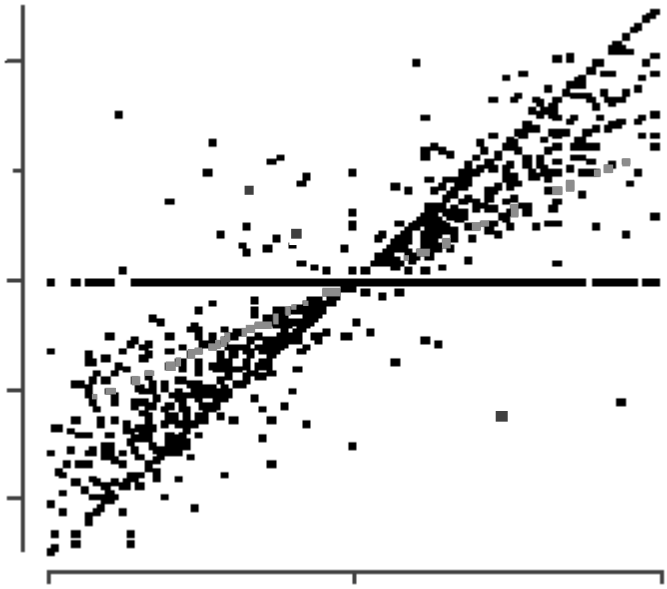
Guide till granskning



Leopold Granquist
Gunnar Arvidson
Catarina Elffors
Anders Norberg
Lars-Göran Lundell

CBM 2002:1

Guide till granskning



Leopold Granquist
Gunnar Arvidson
Catarina Elffors
Anders Norberg
Lars-Göran Lundell

För ytterligare information, kontakta:
Leopold Granquist telefon: 08-506 944 43
e-post: Leopold.Granquist@scb.se

eller
Gunnar Arvidson telefon: 019-17 65 59
e-post: Gunnar.Arvidson@scb.se

Catarina Elffors telefon: 08-506 947 98
e-post: Catarina.Elffors@scb.se

Anders Norberg telefon: 08-506 946 79
e-post: Anders.Norberg@scb.se

ISBN: 91-618-1138-6

Printed in Sweden
SCB-tryck, Örebro 2002.06

Generaldirektörens förord

Statistiska centralbyrån (SCB) utarbetar som en del i det systematiska kvalitetsarbetet rapporter rörande bästa kända metod på olika områden, s.k. Current Best Methods (CBM). ”Guide till granskning” är ett sådant CBM-dokument och har tagits fram inom ramen för SCB:s arbete med att utveckla granskningsarbetet. Det utgör en genomgripande uppdatering av CBM-dokumentet ”Granska effektivt”, som publicerades i mars 1997.

Guiden är avsedd för personer med ansvar för design och utveckling av statistiska produktionsprocesser vid SCB, andra myndigheter, organisationer och företag. Granskningens huvuduppgift att vara ett medel för kontinuerlig förbättring av hela undersökningen betonas. Generella grafiska programvaror rekommenderas komma till allmän användning.

Jag vill tacka den arbetsgrupp, som har utarbetat rapporten, och alla andra, som har sett till att den har tagits fram, för ett fint arbete.

Svante Öberg
Generaldirektör

Arbetsgruppens förord

I föregående CBM-dokument: ”Granska effektivt”, som publicerades i mars 1997, behandlades i avsnittet om grafisk granskning endast skräddarsydda programvaror för outputgranskning.

Nu finns det emellertid generella grafiska programvaror. De kan effektivt utnyttjas för nya outputgranskningsmetoder, i kontroller, utveckling och utvärdering av granskningskriterier och processer samt för identifiering av systematiska uppgiftslämnarfel – vilka är svåra att upptäcka med traditionella metoder. Detta genomsyrar hela dokumentet. Ett särskilt kapitel ägnas åt grafiska metoder, och flera kapitel har omarbetats grundligt med avseende på de nya möjligheter som tekniken medger. På detta område ligger SCB i frontlinjen, vilket bl.a. inneburit att experiment och utveckling har ingått i projektet.

Den internationellt accepterade nya synen på granskning – med tonvikten på att identifiera felkällor och problem i insamlings- och produktionsprocessen – är det genomgående temat i hela CBM-dokumentet. Föregående version var inriktad på att rationalisera granskningen, medan denna är orienterad mot att kontinuerligt förbättra kvaliteten i undersökningen. I det avseendet är dokumentet också framåtsyftande i meningen att det inte finns publicerade artiklar om erfarenheter av metoder, idéer osv.

Ett biskäl till beslutet om uppdatering var att dokumentet skulle förses med checklistor i större utsträckning och att form och layout skulle samordnas med övriga CBM-dokument.

Dokumentet har i huvudsak utarbetats av en arbetsgrupp bestående av Gunnar Arvidson, Catarina Elffors, Leopold Granquist (projektledare), Lars-Gösta Lundell och Anders Norberg. I inledningsskedet medverkade också Per Engström.

Synpunkter på dokumentets innehåll har vi fått via seminarier och framför allt av en läsgrupp som bestått av Stina Andersson, Heather Bergdahl och Roland Friberg.

Dokumentet har språkgranskats av Hans Berglund, och layouten har gjorts av Inga-Lill Kvist.

Innehållsförteckning

1 Principer

Vi utgår här från en kontinuerlig företagsundersökning (månad, kvartal, år), men det som sägs är i princip tillämpbart på alla typer av undersökningar.

Med **granskning** menar vi identifiering och åtgärdande av fel och outliers i individuella data som används för framställning av statistik. Ett huvudsyfte är att identifiera felkällor för senare åtgärder i undersöknings- och produktionsprocessen.

1.1 Granskningens syfte:

- Förbättra undersökningen = förhindra att fel uppkommer.
- Öka kvaliteten i ingående och utgående data på ett effektivt sätt.
- Bidra till kvalitetsbedömningar av statistiken.

1.2 Granskningens roll

- Identifiera felkällor i undersökningen, speciellt problem för uppgiftslämnaren att besvara frågorna
- Identifiera och åtgärda betydelsefulla fel

Ett upptäckt fel ska alltid ses som en representant för en felkälla eller ett problem med undersökningen.

Granskning ska ses som en del i arbetet att få hög slutlig kvalitet i statistiken.

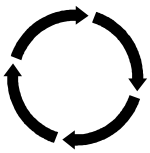
1.3 Faktorer för att uppnå hög slutlig kvalitet

- Processperspektiv
- God svarskvalitet
- Bra kontroller
- Systematisk insamling av data över felkällor

1.3.1 Processperspektivet: ”TQM - ständig förbättring” ska tillämpas på

- undersökningen
- granskningsprocessen.

Detta innebär att man ska:



- samla in processdata
- analysera
- åtgärda
- mäta effekterna, dvs. samla in processdata, analysera, åtgärda, mäta osv.

Ständig förbättring förutsätter ett bra processdatasystem
Processdata ska genereras och presenteras på ett sådant sätt att de snabbt kan analyseras – t.ex. genom pareto-diagram över

- omfattningen av betydelsefulla felkällor
- andelen felsignalerade, manuellt åtgärdade och automatiskt åtgärdade objekt
- andelen förändrade värden och effekten av förändringarna per variabel
- kontrollernas *träffsäkerhet* och effektivitet.

Processdata ska ge underlag för var och hur åtgärder ska vidtas för att förbättra undersökningen: blankettförbättring (mätteknik, frågor, layout); bortfallsreducering; granskning m.m.

Allt i insamlings- och produktionsprocessen ska vara enkelt att ändra!
Variabler, *kontroller* och *acceptansgränser* ska kunna ändras utan nämnvärt arbete och utan specialistkompetens. Organisationen bör tillhandahålla generella program och en genomtänkt systemutvecklingsmodell. De ska tillämpas för alla undersökningar utom där det kan beläggas att metoderna inte är tillämpliga.

SCB:s generella programvara för granskning heter GRETA. Programmet är särskilt flexibelt när det gäller ändringar i kontroller. I kombination med programmet AGDA kan acceptansgränser nästan helt automatiskt anpassas till de data som ska granskas. Processdata från granskningen genereras automatiskt på ett sätt som gör det möjligt att skapa generella rutiner för sammanställning och presentation av statistik över granskningen.

1.3.2 God svarskvalitet förutsätter att uppgiftslämnaren

- har kapacitet att ta fram data
- exakt förstår innebörden i frågor och definitioner
- är motiverad att lämna så korrekta uppgifter som möjligt
- ges instrument att undvika slarvfel.

Uppgifter bör vara lätt tillgängliga och ska helst finnas i uppgiftslämnarens informationssystem. Det är ofta bättre att samla in data med definitioner som är för uppgiftslämnaren än med dem som vore de ideala för undersökningen. Undvik begreppsförvirring! Uppgiftslämnarna ska förstå vad de ska svara på. Och vi måste veta vad svaret avser.

Hög kvalitet i indata kan uppnås om

- frågorna är anpassade till uppgiftslämnarnas kapacitet (uppgifter tillgängliga)
- definitioner och begrepp är lätt tillgängliga i anslutning till respektive fråga (förståelsen)
- uppgiftslämnaren uppmuntras – generellt och i anslutning till svårare frågor – att kontakta undersökningens *uppgiftslämnarservice* (förståelsen)
- uppgiftslämnaren får effektiv information om undersökningens betydelse och eventuellt belönas, t.ex. genom att uppgiftslämnaren får nyckeltal, resultat, publikationer o.d. (motivera)
- uppgiftslämnaren vägleds att kontrollera sina svar (inga slarvfel).

Krav på uppgiftslämnarservicen:

- hög tillgänglighet, dvs. ständig jour under dagtid
- telefonsvarare
- fax
- e-post
- kunnig personal, som ska
  - besvara frågorna
  - fråga efter andra eventuella problem
  - dokumentera och registrera frågor och problem!
- flexibilitet i form av längre jourtid och ökad bemanning när insamlingen är som intensivast.

1.3.3 Klassificering av fel

Granskningsprocessen ska effektivt identifiera och åtgärda betydelsefulla fel. En klassificering av fel och kontroller gör det lättare att uppnå hög effektivitet (identifiera de betydelsefulla felen med låg resursförbrukning).

Fel klassificeras i

Uppenbara fel: Kan identifieras säkert med enbart tillgång på data om det granskade objektet.
Misstänkta fel: Kontroller kan enbart ”säga” att ett variabelvärde är ”misstänkt”.

Misstänkta fel klassificeras i

Avvikelsefel: Värdet på testvariabeln är för stort eller för litet.
Definitionsfel: Exempel: Felaktigt svarsbeteende. Många uppgiftslämnare uppfattar en fråga eller underliggande definitioner på ett likartat men felaktigt sätt.

1.3.4 Bra kontroller – förutsättning för god effektivitet i granskningen

För att granskningen ska kunna höja kvaliteten fordras

- hög träffsäkerhet när det gäller att hitta betydelsefulla misstänkta fel
- hög kvalitet i verifieringsarbetet.

Kontroller och acceptansgränser för speciellt misstänkta fel måste ändras i takt med de allt snabbare förändringarna i näringsliv och samhälle.

Verifieringen av felsignaler måste utföras av välutbildad personal enligt välplanerade rutiner.

Problem

- Kontroller kan öka skevheten.
  Exempel: Om sannolikheten är mycket högre att identifiera fel som medför överskattning än att identifiera dem som medför underskattning. (Det finns exempel där acceptansgränserna inte täckte medianvärdet av värdena på kontrollfunktionen.)
- Kontroller hittar inte förekommande allvarliga systematiska fel.
  Exempel: *inliers* dvs. felaktiga värden som ligger innanför varje rimligt val av acceptansgränser.

- Kontrollsystemet kan orsaka *övergranskning*:
  - För många felsignaler kan fördröja publiceringen dvs. minskad tidskvalitet.
  - Stora enstaka fel drunknar i mängden av felmeddelanden.
  - Nya fel tillförs och uppgiftslämnarproblem döljs.
    Uppkommer genom felaktiga åtgärder för att objektet ska passera genom kontrollsystemet.
- Felsignalerade värden ändras felaktigt.

1.3.5 Systematisk insamling av data över felkällor

Kontrollsystemet kan bli effektivt om

- kontrollerna och kontrollsystemet utvärderas i ett autentiskt datamaterial
- data över granskningens effekter genereras kontinuerligt (processdatasystemet, se ovan)
- data över felkällor insamlas och analyseras
- datamaterialet analyseras för att identifiera eventuella *inliers*.

Med hjälp av granskningsprogrammet GRETA kan man se effekter av kontroller och acceptansgränser.

SAS/Insight ger möjligheter att identifiera *inliers* och andra förekommande systematiska fel samt är ett hjälpmedel när det gäller att finna bra kontroller för misstänkta fel och att avgöra att acceptansgränser ligger rätt.

Vår uppfattning

Processperspektivet och kontrollerna för identifiering av *inliers* torde öka kvaliteten väsentligt mer än vad traditionell granskning gör.

2 Vilka fel ska granskningen hitta?

I granskning utförs kontroller av data. En del kontroller kan fastslå att det finns fel i ett visst variabelvärde eller bland vissa variabelvärden. Andra kontroller kan enbart säga att variabelvärden ser misstänkta ut. Troligen finns det ett fel, men det ifrågasatta värdet kan också vara korrekt.

Värden som kontroller med säkerhet kan konstatera att de är fel kallar vi för *uppenbara fel*, medan de värden som identifieras på grundval av att de ser misstänkta ut och visar sig vara fel kallas för *misstänkta fel*.

Kategorin *uppenbara fel* är alltså reserverad för kontroller som med säkerhet identifierar att uppgifter saknas eller är felaktiga. Ett ”uppenbart” fel betyder emellertid inte alltid att det är uppenbart ens för en kunnig användare på ämnesområdet. Ett exempel är fel i identitetsvariabler. Sådana fel upptäcks dock alltid när de används i matchningar o.d. – genom det kaos de orsakar i produktionssystemet.

Data kan *misstänkas* vara felaktiga av i huvudsak två skäl: de avviker mycket från övriga data i samma stratum eller ser ut att vara exempel på ett visst felaktigt svarsbeteende. Vi kallar den första gruppen för *avvikelsefel* och den andra gruppen för *definitionsfel*. I engelskspråkig litteratur används begreppet *inliers* för den senare typen av fel, beroende på att sådana värden normalt accepteras av de kontroller som används för att avslöja avvikelsefel.

A*vvikelsefel* upptäcks genom att variabelvärdet eller värdet på testvariabeln är misstänkt litet eller stort i förhållande till vad som är förväntat för den del av populationen som objektet tillhör. Det avviker så mycket från övriga värden att det kan ha en oacceptabelt stor inverkan på skattningen.

Definitionsfel uppstår när många uppgiftslämnare uppfattar en fråga på likartade men felaktiga sätt eller när de (medvetet eller omedvetet) tillämpar andra definitioner än undersökningens. I företagsstatistik har ett sådant, systematiskt felaktigt, svarsbeteende sin grund i att uppgiftslämnaren hämtar data direkt från sitt informationssystem utan att ge akt på att definitionerna inte överensstämmer eller att han eller hon inte vill eller kan ta fram svar enligt undersökningens definitioner. Definitionsfelen är en typ av mätfel.

Vi har infört denna ”klassificering” av fel för att dessa utgör skilda granskningsproblem. För det första gör detta det enklare att inse hur man ska identifiera *avvikelsefel* på ett effektivt sätt. För det andra vill vi understryka att en väsentlig uppgift för granskningen är att hitta förekommande systematiska *definitionsfel*.

När det gäller avvikelsefelen, är det främst en fråga om att uppnå hög *träffsäkerhet* i kontrollerna. Just detta misslyckas många processer med, vilket framför allt medför onödiga kostnader som en direkt följd men kan även få men för kvaliteten, Granquist and Kovar 1997.

Definitionsfelen kan från kvalitetssynpunkt vara den allvarligaste feltypen, och de kan dessvärre vara synnerligen svåra att identifiera med enbart kontroller. Det behövs ytterligare åtgärder som t.ex. extra frågor i blanketten.

2.1 Uppenbara fel

Många – men nödvändigtvis inte alla – uppenbara fel måste elimineras i betydelsen att efter granskning ska de ifrågavarande felen inte finnas för något enskilt objekt. Det är undersökningens kvalitetsansvarige som måste ta ställning till vilka uppenbara fel som ska tas bort. Bedömningen ska göras utifrån produktionsprocessens krav och hur statistiken används och presenteras.

Dessutom är det inte enbart en kvalitetsfråga utan kan även vara en fråga om användares förtroende för statistiken. En användare som hittar ett i och för sig betydelselöst fel i ett enskilt objekt kan se det som bevis för slarv i granskningen och tänka att det då säkert finns många andra fel i statistiken. Detta gäller fel som en användare utan kännedom om objektets identitet kan fastslå som fel.

Några exempel på uppenbara fel av typen att de skulle kunna upptäckas även av användare med tillgång till ett avidentifierat material är:

- validitetsfel, dvs. det registrerade värdet tillhör inte variabelns värdeförråd (tillåtna värden)
- motsägelser, dvs. svar på en fråga motsäger svar på en annan eller andra frågor
- saknade värden på frågor där värden måste finnas t.ex. enligt undersökningens krav, svar på andra frågor, tillgänglig bakgrundsinformation osv.
- datastruktur- eller modellfel, dvs. definitionsmässiga samband mellan variabler satisfieras inte.
  Exempel: ”Antal män” ska vara mindre än eller lika med ”summa män och kvinnor”; svaret på en summavariabel ska överensstämma med summan av uppgifterna för delvariablerna.

Exempel på uppenbara fel som orsakar kaos eller ställer till med avsevärt trassel i granskning- och produktionsprocessen är:

- fel i identitetsvariabler, speciellt objektsidentiteter
- validitetsfel i variabler som anges i numeriska koder där antalet koder är stort, dvs. den angivna koden finns inte i kodförteckningen.

Speciellt är fel i identitetsvariabler allvarliga. Sådana fel innebär t.ex. att matchning mot bakgrundsdata och tidigare inhämtade data misslyckas.

Generellt är det relativt enkelt att identifiera uppenbara fel i betydelsen att konstruera kontroller för att upptäcka uppenbara fel enligt definitionen. De data som signaleras av kontrollerna för uppenbara fel ska praktiskt taget alltid innebära att det finns något fel i någon av de variabler som omfattas av kontrollen, men man vet inte genast vilken. Men i de fall man utifrån objektets data maskinellt kan identifiera det felaktiga värdet är det en fördel att till kontrollen koppla en maskinell ändringsrutin som ersätter det felaktiga värdet med ett bättre. En summeringskontroll kan t.ex. förses med en gräns under vilken en felaktig summering av uppgiftslämnaren accepteras och åtgärdas med att summavariabeln ändras till summan av delvariablernas värden.

Partiellt bortfall

Ett speciellt problem är partiellt bortfall, dvs. att uppgiftslämnaren har lämnat en fråga obesvarad trots att det finns uppgifter att redovisa. Partiellt bortfall måste praktiskt taget alltid åtgärdas, eftersom det annars medför underrapportering som

är större ju större det partiella bortfallet är. (Vi utgår från att man inte betraktar objektet som objektsbortfall när partiellt bortfall förekommer för ett relativt litet antal variabler.) Detta betyder att produktionsprocessen också måste kunna identifiera partiellt bortfall för alla variabler där bortfallet är av betydelse.

Dessvärre kan det för en del variabler vara svårt att få det partiella bortfallet att tillhöra kategorin uppenbara fel, dvs. att konstruera kontroller som med säkerhet kan påvisa att värden är utelämnade. Problemet är särskilt besvärligt för undersökningar där uppgiftslämnare normalt har data att redovisa enbart för en del av variablerna, t.ex. produktion, försäljning eller export avseende ett antal specificerade varor; sådda arealer eller skördar för ett antal specificerade grödor; djurbestånd för ett antal angivna djurkategorier osv.

Det gäller dock att söka få fram en så fullständig mängd kontroller som möjligt som med säkerhet kan påvisa att variabelvärde saknas. För detta krävs det ofta djupa ämneskunskaper och -erfarenheter. Men man måste nog också angripa problemet via blankettutformning och -anvisningar för att minska risken t.ex. att kräva att uppgiftslämnaren ska skriva in ”0” när inget finns att redovisa och ”kan ej” när en fråga inte kan besvaras. När det inte är uppenbart att en obesvarad fråga är partiellt bortfall, får man klassificera uppgiften som definitionsfel.

2.2 Misstänkta fel

För att misstänkta fel ska vara intressanta från kvalitetssynpunkt måste de ha betydelse på skattningsnivå. Det betyder antingen att ett enskilt (avvikelse-)fel påverkar skattningen eller att sammanlagda effekten av ett antal fel (avvikelsefel med samma tecken eller definitionsfel) gör det. Därför ska misstänkta fel ha effekter på skattningsnivå för att ingå i målet för granskningsoperationen.

2.2.1 Avvikelsefel

Signaler från kontroller inriktade på avvikelsefel innebär oftast enbart att det finns misstanke om fel. Dessa data måste alltså undersökas ytterligare för att man ska kunna fastställa om de är felaktiga eller inte. När man måste ta kontakt med uppgiftslämnaren för att utreda frågan blir det kostsamt för båda parter. Därför måste de maskinella kontrollerna ha hög träffsäkerhet i att identifiera avvikelsefel som är betydelsefulla för skattningarna.

Antalet betydande avvikelsefel för en variabel är relativt lågt i de flesta undersökningar. Detta är empiriskt bekräftat av den mångfald studier i olika former av traditionella granskningsprocesser som genomförts världen över (se t.ex. Granquist and Kovar 1997). De visar att i traditionell granskning med snäva acceptansgränser felsignaleras många observationer. Av dessa är det ofta inte mer än 30 procent som är felaktiga, och uppemot 80–90 procent av dessa upptäckta fel har ingen inverkan alls på skattningarna.

Många avvikelsefel uppstår på grund av slarv antingen hos uppgiftslämnaren eller vid dataregistreringen. De uppträder därför slumpartat, men felens riktning och storlek behöver därför inte vara ”likformigt fördelade”. Eftersom vissa slarvfel kan vara mycket stora, måste avvikelsekontroller ingå i varje granskningsprocess. Antalet fel av betydelse per variabel kan emellertid förväntas vara lågt. Detta bör

man ta hänsyn till när man utformar kontroller och acceptansgränser för avvikelsekontroller.

2.2.2 Definitionsfel

Tills för ett par år sedan har mycket litet intresse ägnats åt definitionsfel, trots att feltypen ur kvalitetssynpunkt mycket väl kan vara den som medför störst skevhet i skattningarna. Den internationella benämningen på feltypen, *inliers*, är dock egentligen ett snävare begrepp än definitionsfel. (Se DesJardins 1997 och Winkler 1997.)

Definitionsfel uppstår när en uppgiftslämnare har svårigheter att förstå innebörden av en fråga, saknar möjligheter att besvara den eller är omedveten om att definitionen på variabeln avviker från den definition uppgiftslämnaren tillämpar i sitt informationssystem. Om feltypen är frekvent för en variabel, kan felen i praktiken inte identifieras med avvikelsekontroller. Många uppgiftslämnare tillämpar ju samma felaktiga svarsbeteende. Observera att ovan omnämnda studier av granskningsprocesser också belägger att snäva acceptansgränser är en dålig metod när det gäller att hitta definitionsfel.

Vid programmet AM/FS (Företagsbaserad sysselsättning) har man sedan ett par år tillbaka systematiskt arbetat med att förebygga definitionsfel. Man besöker uppgiftslämnare ute på företag/organisationer som har ansvar för många urvalsobjekt i en given undersökning. Kontakter tas även med representanter från dataservicebyråer som levererar och ger stöd till programvara hos flera uppgiftslämnare. Vid dessa personliga besök klargör man definitionerna och reder ut ett stort antal missuppfattningar i fråga om centrala variabler för undersökningen. Exempelvis gjordes under hösten 2000 ett besök hos TietoEnator, där man tillsammans med systemerare och programmerare samt personal från kundsupport gick igenom undersökningen Kortperiodisk sysselsättningsstatistik för offentlig sektor (KSO). Man gick igenom variabel för variabel för att hitta potentiella fallgropar i löneprogrammet ”Respons” (en programvara som finns hos ett 30-tal kommuner).

Arbetet innebär att skapa insikt i att allvarliga definitionsfel förekommer, att identifiera felkällan och att finna åtgärder för att komma till rätta med felkällan.

I ”CBM: Fråga rätt” anges i exempel 9.4.1 på sidan 78 att SCB ändrade definitionen på en variabel (under året nedlagda kostnader) till den definition som 90 procent av uppgiftslämnarna alltid tillämpade (under året fakturerade kostnader). Därmed kunde man spara in det omfattande granskningsarbete som bestod i att undersöka vilken definition uppgiftslämnaren tillämpade och att ändra siffrorna för 90 procent av objekten. Omräkning till nedlagda kostnader, vilket användarna kräver, görs nu maskinellt.

Fråga rätt redovisar i exempel 9.4.2 på sidan 79 en ändring av undersökningsdesignen som minskade arbetet vid granskning både för SCB:s personal och för uppgiftslämnarna. En stor del av granskningsarbetet ägnades åt småföretagen, som hade betydande svårigheter att lämna så detaljerade uppgifter om sin verksamhet som efterfrågades i postenkäten. I en simuleringsstudie konstaterade man att arbetet med småföretagen endast marginellt påverkade resultatens tillförlitlighet. Metodstudier visade att man via modellberoende skattningar kunde skatta parametrar för småföretagen med en handfull av svaren i den ursprungliga 50-talet

frågor stora blanketten. Den förenklade blanketten för småföretagen medförde också en dramatisk ökning av svarsfrekvensen.

Ett par exempel från andra sidan Atlanten får ytterligare belysa problemen.

Exempel 1: Slakteriundersökning

I USA:s slakteriundersökning får de utvalda slakterierna veckovis ange bl.a. mängden slaktade djur och sammanlagd slaktvikt för varje förekommande djursort. Vid en översyn av granskningssystemet i samband med en övergång till ett pc-baserat produktionssystem, visade det sig att ett stort antal slakterier lämnade exakt eller nästan exakt samma uppgifter vecka efter vecka. Möjligen skulle man kunna acceptera att *antalet* slaktade djur inte varierar över tiden, åtminstone när det gäller de slakterier som går för full kapacitet. Men den sammanlagda slaktvikten kan omöjligen vara densamma vecka från vecka. Man konstruerade därför en kontroll av tidsserietyp som felsignalerade de slakterier där variationen i slaktvikt var för låg i förhållande till slakteriets tidigare rapporterade slaktvikter. (Se Mazur 1990.)

Kommentar: I detta fall lyckades man genom kontroller identifiera det felaktiga svarsbeteendet att i praktiken inte lämna uppgifter.

Exempel 2: Lönestatistik

Bureau of Labor Statistics (BLS) i USA producerar varje månad en lönestatistik för arbetare. Man genomför också regelbundna svarsanalysstudier. Lönebegreppet är uppbyggt av mer än 100-talet komponenter, s.k. löneslag, av vilka vissa ska räknas med, medan andra inte ska tas med. I en svarsanalysstudie undersöktes om uppgiftslämnarna inkluderade och exkluderade lönekomponenter enligt lönestatistikens definitioner. Man frågade också de uppgiftslämnare som inte följde anvisningarna om de i fortsättningen skulle kunna rapportera enligt undersökningens definitioner. I följande lönestatistik beräknade man estimaten för dem som ändrade beteende och jämförde resultaten med estimaten för en kontrollgrupp. Då fann man att för den viktigaste variabeln, Lönen för arbetare i direkt produktionsarbete, blev det i första gruppen 10.7 med standardavvikelsen 3.2 att jämföra med 1.6 för kontrollgruppen. (Se Werking m.fl. 1988.)

Kommentar: Eftersom det är fråga om en konsistent felrapportering över tiden, kan den inte upptäckas med kontroller som bygger på jämförelser över tiden. Ett sätt att lösa problemet med komponentuppbyggda variabler är att i blanketten bygga in kontrollfrågor, t.ex. på följande sätt: ”När du svarade på den här frågan tog du med … och utelämnade ….? Om inte, markera om du i fortsättningen kan redovisa uppgiften enligt våra anvisningar.”

Definitionsfel som tar sig uttryck i att uppgiftslämnaren inte besvarar en fråga (inte förstår den, inga data i informationssystemet, för mycket arbete att ta fram data) indikeras genom högt partiellt bortfall för variabeln. Ibland kan de identifieras i kontrollerna mot uppenbara fel. Här är uppföljningen av felsignalerna särskilt viktig, och det är granskarens uppgift att identifiera felorsaken, dvs. felkällorna. Data från övriga uppgiftslämnare kan här vara mycket osäkra, varför imputeringar som bygger på dessa data kan få mycket låg kvalitet.

Metoder för att hitta definitionsfel som beror på att uppgiftslämnaren ”tar vad han har” är:

- skräddarsydda kontroller som i exempel 1 ovan
- kontrollfrågor i blanketten som i exempel 2 ovan
- analys av data med syftet att identifiera felaktiga svarsmönster, t.ex. genom att använda SAS/INSIGHT (se kapitel 5).

Hela undersökningsprocessen ska genomsyras av att man identifierar felkällor av definitionsfeltyp och partiellt bortfall (som beroende på kontrolltyp kan hänföras till både uppenbara fel och definitionsfel, se avsnittet om uppenbara fel). Design av undersökningen, blanketter, uppgiftslämnarservice (se avsnitt 3.2.1), uppföljningen av felsignaler och evalvering är medel utöver kontroller i blanketten och i granskningsprocessen.

2.3 Referenser

DesJardins, D. (1997): Experiences With Introducing New Graphical Techniques for the Analysis of Census Data, Work Session on Statistical Data Editing, Prague, Working Paper No. 19.

Granquist, L and Kovar, J (1997): Editing of Survey Data: How much is enough? In L. Lyberg, P Biemer M. Collins, E. Leeuw, C. Dippo, N. Schwarz, D. Trewin (eds) Survey Measurement and Process Quality, Wiley, New York, 1997, pp. 415–436.

Mazur, C. (1990): Statistical Edit System for Livestock Slaughter Data, Staff Research Report No. SRB-90-01, Washington, DC: U.S. Department of Agriculture.

Statistiska centralbyrån (2001): Fråga rätt! – Utveckla, testa, utvärdera och förbättra blanketter. CBM-rapport juni 2001.

Werking, G., Tupek, A. and Clayton, R. (1988): CATI and Touchtone Self-Response Applications for Establishment Surveys, *Journal of Official Statistics*, **4**, pp. 349–362.

Winkler, W. (1997): Problems With Inliers, ECE, Work Session on Statistical Data Editing, Prague, Working Paper No. 22.

3 Var ska granskningen ske?

Data kan granskas i ett eller flera olika skeden av insamlings- och produktionsprocessen:

- **Uppgiftslämnargranskning**: Utförs av uppgiftslämnaren (t.ex. vid besvarande av postenkät eller elektronisk blankett) eller av uppgiftslämnare och intervjuare gemensamt i intervjuundersökningar.
- **Manuell förgranskning:** En manuell process som äger rum omedelbart före dataregistrering.
- **Dataregistreringsgranskning:** *Verifiering* av de uppenbara fel som dataregistreringsprogrammet identifierar.
- **Produktionsgranskning** eller **batch-granskning:** Verifiering av de variabler och objekt som flaggas av granskningsprogrammet. Oftast körs programmet för en mängd (batch) dataregistrerade objekt i taget, s.k. omgångar.
- **Outputgranskning:** Granskning när allt material är insamlat för kontroll av att inga stora misstag har gjorts i de tidigare processerna.

I postenkätundersökningar är det vanligt att alla dessa fem processer förekommer – även om såväl uppgiftslämnargranskning som manuell förgranskning och outputgranskning är mycket mindre processer än t.ex. dataregistrerings- och produktionsgranskningen. Frågan är hur man ska bygga upp en granskningsprocess på ett bra sätt och vilka delprocesser som ska ingå? Hur detta ska lösas för en specifik undersökning beror t.ex. på antalet undersökningsvariabler, hur komplicerade variablerna är, hur data insamlas, vilken kvaliteten är i insamlade data, vilken teknik och vilka programvaror som finns tillgängliga.

Varje undersökningsansvarig måste ta ställning till följande frågor:

- Vilken eller vilka delprocesser behövs egentligen?
- Är det verkligen nödvändigt med mer än en stor delprocess?
- Hur ska delprocesser samordnas för att man ska undvika bl.a. dubbelarbete och onödiga uppgiftslämnarkontakter?

Syftet med avsnitt 3.1 är att ge underlag och vägledning för att i varje enskild undersökning besvara dessa frågor. Rekommendationer för vad som bör tas med i de delprocesser som förekommer i en postenkätundersökning redovisas i 3.2. I avsnitt 3.3, om elektronisk datainsamling, diskuteras det som utmärker dessa metoder jämfört med postenkät. Slutligen berörs i avsnitt 3.4 granskning av registerdata.

3.1 Undvik flera stora delprocesser?

Datakvaliteten kan förväntas bli högre ju närmare uppgiftslämnandet eller insamlingen av data som granskning sker. Vid t.ex. prisinsamling i butiker och vid besöks- eller telefonintervjuer bör all granskning ske i butiken respektive under intervjun. Därför rekommenderar vi i 3.2 satsningar på uppgiftslämnarservice både när uppgiftslämnaren själv tar kontakt med SCB och när inkommande blanketter är så bristfälligt ifyllda att man måste återkontakta uppgiftslämnaren.

Dessutom ska man se till att uppgiftslämnaren instrueras att granska sina svar. Det senare kallar vi uppgiftslämnargranskning.

Vi framhåller att en fördel med granskning vid dataregistrering är att den ofta äger rum strax efter uppgiftslämnandet. Därmed menar vi att när uppgiftslämnarna behöver återkontaktas, har de uppgiftslämnandet i färskt minne. En annan fördel är att blanketten finns framför ögonen på den som granskar. Ingen tid krävs för att ta fram och sortera blanketter. Dessutom kan ”omgranskning” av ändringar eller nya värden ske omedelbart vid registreringen.

Men räcker det med endast dataregistreringsgranskning i postenkäter? Behövs det över huvud taget mer än en stor granskningsprocess? Fördelen med det är att man minskar risken för dubbelarbete. En annan fördel är att uppgiftslämnare inte kommer att kontaktas mer än en gång. Det främsta skälet till att vi kraftfullt varnar för omfattande manuell förgranskning (3.2.3) är just risken för att man vid efterföljande dataregistreringsgranskning gör om mycket av det som redan utförts (se t.ex. Linacre and Trewin 1989).

Observera även att granskningsprocessen som sådan har rationaliserats oerhört tack vare att datortekniken gjorde det möjligt att integrera registrering, kodning och granskning i pc-miljö (se t.ex. Bethlehem et al. 1989, Granquist 1992 och Pierzchala 1995). Internationellt uppmärk-sammade man denna möjlighet tidigt och utvecklade generella program-varor för registrering och samtidig granskning, t.ex. BLAISE (Holland), DC2 (Canada), IPS (Australien) och IMPS (USA). Vid SCB finns programmet Rode/pc.

Om man har flera stora delprocesser, måste man vidta åtgärder för att undvika det dubbelarbete och de upprepade återkontakter som kan uppstå till följd av att man har flera större processer. Man får då sätta och utnyttja *granskningskoder* i granskningsprogrammen samt planera verifieringsprocesserna noggrant.

Ett allvarligare problem kan vara att hela granskningen tar längre tid. Om granskningen fördröjer publiceringen av statistiken, måste kvalitetsförbättringen på grund av granskningen motivera den försämrade tidskvaliteten. Det gäller också att förvissa sig om att man faktiskt uppnår en kvalitetsförbättring. Återkontakter med uppgiftslämnaren kan t.ex. komma att tas så långt efter uppgiftslämnandet att man inte får högre svarskvalitet.

Man måste ha goda motiv för att t.ex. i postenkätfallet ha ytterligare en stor process utöver dataregistreringsgranskning. Skäl kan vara att angelägna kontroller inte kan utföras i dataregistreringsgranskning eller att kontroller kan göras mycket effektivare i t.ex. produktionsgranskning. Det är då fråga om kontroller som kräver data från andra insamlade objekt. Produktionsgranskning bör då reserveras enbart för sådana kontroller.

Outputgranskning bör alltid finnas, eftersom huvudsyftet är att säkerställa att inga stora fel har sluppit igenom tidigare utförd granskning. Här rekommenderar vi grafisk granskning (se kapitel 5).

Vid elektronisk datainsamling, som diskuteras i 3.3, bör man vara mycket försiktig med att låta processen följas av någon annan process än outputgranskning. Det är särskilt vid uppgiftslämnargranskning med elektroniska blanketter som en sådan situation kan uppstå. Huvudregeln är att granskningen vid SCB måste till-

föra något nytt, t.ex. kontroller som bedöms vara för komplicerade för uppgiftslämnaren. Men man bör absolut inte göra om det som uppgiftslämnaren redan har gjort.

Checklista för hela granskningsprocessen

- ❑ Ange de granskningsprocesser som kan vara aktuella:
  - ○ manuell uppgiftslämnargranskning
  - ○ manuell förgranskning
  - ○ dataregistreringsgranskning
  - ○ produktionsgranskning
  - ○ outputgranskning.
- ❑ Prioritera bland dessa.
- ❑ Begränsa antalet processer, men outputgranskning (helst i grafisk form) bör dock alltid finnas med.

3.2 Granskning av postenkäter

Vi ska här följa de olika faser som en blankett i en SCB-undersökning kan passera från det att den besvaras av uppgiftslämnaren tills objektets data bedöms vara färdiggranskade. Genom numrerade rubriker urskiljer vi i varje fas i produktionsprocessen de granskningsprocesser och aktiviteter/processer som syftar till att höja svarskvaliteten.

Bilagan till kapitel 3 presenterar ett flöde över granskningsprocessen i en postenkätundersökning.

Fas 1 – Svarsprocessen

När uppgiftslämnaren tar emot enkäten, blir den första åtgärden att sätta sig in i frågorna. Vad vill SCB ha svar på? Kan jag besvara frågorna? Var hittar jag uppgifterna?

I huvudsak är det blanketten med sina anvisningar som ska ge uppgiftslämnaren tillräcklig information om vilka uppgifter som begärs. Introduktionsbrev om undersökningens syfte och förekommande anvisningar om variabeldefinitioner hjälper också uppgiftslämnaren att förstå frågornas innebörd.

Nästa problem för uppgiftslämnaren är att undersöka: Vilka data kan hämtas direkt från informationssystemet (redovisningssystem, bokföring, kunskaper, minnesbilder med mera)? Vilka data behöver tas fram på annat sätt? Vilka data går inte att få fram?

När svårigheter uppstår, är det många uppgiftslämnare som kontaktar SCB. Frågorna kan gälla:

- om objektet faktiskt tillhör målpopulationen
- vilka definitioner som tillämpas
- vad frågorna innebär
- vad som ska göras när svaren inte kan hämtas direkt ur uppgiftslämnarens informationssystem, t.ex. på grund av skillnader i definitioner, referensperiod eller detaljeringsgrad.

Vi rekommenderar att uppgiftslämnaren på alla sätt uppmuntras att ta kontakt med undersökningens uppgiftslämnarservice. Uppmaningar att ta kontakt bör gärna

göras upprepade gånger i blanketten, introduktionsbrevet och de anvisningar som ofta bifogas.

3.2.1 Uppgiftslämnarservice

Uppgiftslämnarservicen måste ha hög kvalitet. Tillgängligheten ska vara hög – det ska alltid finnas någon som svarar. Erfaren och kunnig personal ska omsorgsfullt och uppmuntrande ta hand om uppgiftslämnaren. Uppgiftslämnaren ska få all den hjälp som han eller hon anser sig behöva.

Fördelar med god uppgiftslämnarservice är att

- ingående datakvalitet blir högre
- kunskaper om uppgiftslämnarens problem erhålls
- behovet av att senare i produktionsprocessen ta kontakt med uppgiftslämnaren minskar
- uppgiftslämnaren lär sig, vilket kan medföra högre kvalitet i svaren vid nästa undersökning.

Kontakterna med uppgiftslämnarna ska dokumenteras, så att man kan göra statistik över frågor och identifierade problem. Vissa uppgifter bör också återföras till registret eller databasen för att undvika att man tar kontakt med uppgiftslämnaren i frågor som man redan känner svaren på (en uppgift för systembyggaren).

Checklista för uppgiftslämnarservice

- ❑ Bemanning: kort eller lång tid av dygnet
  - ○ Telefonsvarare
  - ○ Fax
  - ○ E-post
- ❑ Se till att personalen är kunnig och erfaren.
- ❑ Skriv i blankett, introduktionsbrev och anvisningar på flera ställen att uppgiftslämnarservice finns och bör användas.
- ❑ Dokumentera samtliga samtal på en lista med åtminstone kolumnerna: tidpunkt; uppgiftslämnare; fråga; tidsåtgång.

3.2.2 Uppgiftslämnarens egen granskning

Många ambitiösa uppgiftslämnare kontrollerar på eget initiativ sina uppgifter när blankettens frågor besvaras. Det kan vara kontroller av att siffror skrivs korrekt; bedömning av att data är rimliga t.ex. i jämförelse med tidigare rapporteringar; kontroller av att summor stämmer, en delpost är mindre än summaposten osv.

Självklart blir svarskvaliteten högre, uppgiftslämnandet enklare och arbetet för SCB och uppgiftslämnare mindre, om blanketten uppmanar uppgiftslämnaren att göra sådana kontroller.

Följaktligen rekommenderar vi att granskningen integreras med uppgiftslämnandet i postenkätundersökningar. Det kan göras genom att blanketten ger instruktioner till uppgiftslämnaren, som t.ex. att

- summera delposterna och jämföra det erhållna resultatet med summaposten i bokföringen, vid misstämning kontrollera att ...
- kvoten mellan värdet för variabel x och variabel y bör ligga i intervallet (a,b), om inte: kontrollera att ...

Idén är att uppgiftslämnaren uppmärksammas på vanligt förekommande fel och samtidigt ges ytterligare information om vad som efterfrågas. Vi får på så sätt en styrning av den granskning som många uppgiftslämnare ändå gör, samtidigt som vi (förhoppningsvis) kan få många fler att granska sina egna svar.

Ett problem ligger i att blankettutrymmet i hög grad begränsar möjligheterna till sådana självkontroller. Dessutom ska inte sådana kontroller medföra något egentligt merarbete för uppgiftslämnarna, utan kontrollerna ska upplevas som något naturligt som ger värdefull hjälp i uppgiftslämnandet. Uppmaningarna får dock absolut inte vara av generell karaktär som t.ex. ”kontrollera noga svaren på samtliga frågor”, ”kontrollera att alla frågor är besvarade”, utan de ska vara enkla och konkreta.

Uppgiftslämnarna bör också i blanketten ges utrymme att förklara varför svar inte helt stämmer med vad som förväntas enligt anvisningarna, t.ex. när data ligger utanför ett angivet intervall. Det är också önskvärt att uppgiftslämnaren enkelt kan markera att en misstämning mot en blankettkontroll har beaktats, men att det lämnade svaret trots allt är korrekt.

Checklista för manuell uppgiftslämnargranskning

- ❑ Förtryck tidigare uppgifter på blanketten.
- ❑ Uppmana till summeringar och kontroll av dessa, t.ex.:
  - ○ Summor stämmer (t.ex. mot andra summor i blanketten eller mot summor i bokföringen).
  - ○ Delpost mindre än summan.
  - ○ Kvoten mellan värdet för variabel x och y bör ligga i (a,b).
- ❑ Se till att kontrollerna inte innebär merarbete för uppgiftslämnarna.
- ❑ Se till att kontrollerna uppfattas som meningsfulla.
- ❑ Ge utrymme för kommentarer på blanketten.

Fas 2 – Mottagningsprocessen

Vid ankomsten till SCB hamnar blanketten i mottagningsprocessen, som omfattar allt arbete som föregår dataregistreringen. Exempel:

- behandling av postreturer
- behandling av blanketter från uppgiftslämnare som inte längre tillhör (eller inte anser sig tillhöra) populationen
- behandling av blanketter som ställts till felaktig adress/uppgiftslämnare (t.ex. på grund av ägarbyte)
- uppdatering av utsändningsregistret.

I mottagningsprocessen sker också granskning, som vi kallar manuell förgranskning. Dess innehåll och omfattning ska utformas efter hur dataregistrering sker och vilket kodnings- och granskningsarbete som utförs vid dataregistreringen.

3.2.3 Manuell förgranskning

Förgranskningen har till uppgift att:

- kontrollera att objektet tillhör undersökningspopulationen
- pricka av inkomna svar, såvida inte detta redan gjorts eller görs i ett senare skede

- kontrollera att blanketten är tillräckligt ifylld för att den fortsatta behandlingen ska vara meningsfull
- utföra kodning, sortomvandlingar m.m., om detta inte kan göras vid dataregistreringen
- åtgärda uppenbara fel som granskaren omedelbart får ögonen på.

När blanketten är otillräckligt ifylld eller när grova fel förekommer i strategiska variabler, måste uppgiftslämnaren kontaktas. Det är då oftast fråga om att vägleda uppgiftslämnaren att fylla i blanketten. Det är alltså fråga om samma uppgiftslämnarservice som när uppgiftslämnaren själv tar kontakt (se 3.2.1). Skillnaden ligger i att uppgiftslämnaren måste motiveras och att man ska söka förutse dennes speciella problem. I synnerhet vid nya undersökningar är det viktigt att dessa kontakter dokumenteras.

Vi varnar generellt för omfattande förgranskning. Principen är att det ska räcka med en snabb blick på blanketten för att kontrollera att det är meningsfullt att låta blanketten fortsätta i processen.

Checklista för manuell förgranskning

- ❑ Kontrollera att objektet tillhör undersökningspopulationen.
- ❑ Pricka av inkomna svar.
- ❑ Kontrollera snabbt att blanketten ser ifylld ut.
- ❑ Koda.
- ❑ Gör erforderliga sortomvandlingar.
- ❑ Åtgärda uppenbara fel som du omedelbart får syn på.
- ❑ Undvik omfattande kontroller.
- ❑ Kontakta uppgiftslämnaren när blanketten är för dåligt ifylld eller ser ut att ha grova fel i strategiska variabler.
- ❑ Dokumentera samtal med uppgiftslämnarna samt rätta uppenbara fel.

Fas 3 – Dataregistrering

Blanketter som har godkänts i den manuella förgranskningen ska därefter dataregistreras. Detta kan göras manuellt eller maskinellt, med s.k. skanning. Det senare blir allt vanligare i takt med att skanningtekniken utvecklas.

Vid skanning skapas en elektronisk bild av varje blankett. Bilderna tolkas till data. Dessa kontrolleras med i huvudsak validitetskontroller. Det som inte kan tolkas och det som tolkas osäkert felmarkeras, liksom de värden som inte tillhör variabelns värdeförråd. Felmarkerade objekt presenteras variabelvis för operatören för åtgärd.

SCB:s skanningutrustning finns vid BV/ENK (Enkäter) i Örebro, och skanningen utförs av specialister. Även om utrustningen medger mer avancerade kontroller av data än enkla validiteskontroller, kan personalen normalt inte verifiera flaggningar. Men med den kapacitet SCB:s datanät har i dag kan filer från skanningen överföras till ämnesprogrammen för att verifieras där. Kontakta den skanningsansvarige, som i februari 2002 är Anis Kovacevic, för information och tillvarata den potential till granskning som skanningutrustningen utgör.

Registrering som integreras med granskning kallar vi dataregistreringsgranskning. Denna kan utföras interaktivt – antingen variabelvis: *variabelvis dataregistreringsgranskning* eller objektvis: *objektvis dataregistreringsgranskning*.

3.2.4 Dataregistreringsgranskning

Vid variabelvis dataregistreringsgranskning stoppas registreringen så fort ett variabelvärde underkänns i någon kontroll. Registreringen kan fortsätta först när det flaggade värdet behandlats, dvs. godkänts, åtgärdats eller ”markerats” för senare verifiering. Markeringen ”görs” när operatören fortsätter inmatningen, antingen genom att en kod automatiskt påförs variabeln och objektet eller genom att den felkod som genererats vid flaggningen kvarstår.

Vid registreringen verifieras vissa flaggningar omedelbart, medan de som tar längre tid verifieras när hela objektet har registrerats eller när ett antal objekt är klara – beroende på hur personalen finner det lämpligt att arbeta.

Vid objektvis dataregistreringsgranskning utförs granskningen först när alla värden för ett objekt har registrerats.

Fördelar med dataregistreringsgranskning är att

- granskningen kan ske nära uppgiftslämnandet
- registreringsfel och vissa andra typer av fel kan åtgärdas omedelbart och orsakar därmed inga problem i den fortsatta processen
- blanketterna är omedelbart tillgängliga för verifiering av felsignaler
- blanketterna kan arkiveras mycket enkelt
- kodning, sortomvandlingar och beräkningar av nya variabler, t.ex. summor, kan utföras automatiskt.

Den enda begränsningen i dataregistreringsgranskning är att kontrollers acceptansgränser måste vara helt specificerade innan den första blanketten behandlas. I dataregistreringsgranskning kan man därför utföra alla kontroller för identifiering av uppenbara fel samt de kontroller mot misstänkta fel där acceptansgränser inte beror av data från hur andra objekt har svarat.

Ett krav på dataregistreringsgranskningen bör vara att all information som uppgiftslämnaren har tillfört blanketten i form av kommentarer, upplysningar med mera registreras i syfte att pappersblanketten inte ska behöva tas fram i ett senare skede av produktionsprocessen.

Checklista för dataregistreringsgranskning

- ❑ Kontrollera identiteter.
- ❑ Anpassa graden av tolkningssäkerhet vid skanning.
- ❑ Lägg in kontroller för giltiga värden per variabel.
- ❑ Avgör om verifiering ska göras
  - ○ löpande
  - ○ i efterhand
  - ○ kombinerat.
- ❑ Registrera kommentarer, upplysningar m.m. från uppgiftslämnarna.

Fas 4 – Traditionell granskning

Efter dataregistrering och den granskning som utförts vid dataregistrering, granskas vår blankett tillsammans med andra registrerade objekt i en s.k. produktionsomgång. En omgång består av de objekt som registrerats sedan föregående omgång. I äldre system, där man inte tillämpar interaktiv uppdatering, ingår i varje omgång även omgranskning av de objekt som flaggats och som fått variabelvärden ändrade i tidigare omgångar. Vi kallar tills vidare denna granskningsfas för *produktionsgranskning*.

3.2.5 Produktionsgranskning

Med produktionsgranskning avses den granskning som utförs på registrerade data för inkomna objekt, omgång för omgång tills insamlingen avslutats. Observera att identiteter måste ha kontrollerats innan produktionsgranskningsprogrammet kan köras. Normalt utförs granskningen i flera omgångar i takt med blankettinströmningen. När undersökningen går ut på att bearbeta administrativa register eller material som samlats in av andra, omfattar produktionsgranskningen alla objekt på en gång.

För varje objekt som underkänns i någon av kontrollerna genereras ett *felmeddelande* som omfattar alla underkända variabelvärden samt vidtagna maskinella åtgärder. Meddelandena presenteras för granskningspersonalen – antingen i form av pappersutskrifter eller som interaktiva skärmbilder.

En utförlig beskrivning av processen ges i kapitel 9. I kapitel 6 beskrivs generering av felmeddelanden och verifieringsprocessen.

Produktionsgranskning omfattar kontroller mot såväl uppenbara som misstänkta fel samt automatiska åtgärdanden av vissa typer av uppenbara fel. Jämfört med processen med granskning i samband med dataregistreringen kan produktionsgranskningen innehålla en mer sofistikerad typ av kontroller mot misstänkta fel. Det är kontroller vars acceptansgränser beror av data från övriga objekt i undersökningen (kapitel 4) som kan utföras i produktionsgranskning men inte vid dataregistreringsgranskning.

Checklista för produktionsgranskning

- ❑ Presentera fel, flaggade värden och maskinellt ändrade värden för varje objekt för sig. Inom objektet ska felen osv. komma i samma ordning som variablerna gör på blanketten.
- ❑ Formulera felen och flaggningarna så, att granskaren får ett bra underlag för verifieringsarbetet.
- ❑ Bestäm hur felen respektive flaggningarna av misstänkta värden ska presenteras för granskaren:
  - ○ skärmbilder, en för varje felsignalerat objekt för interaktiv uppdatering, vilket vi rekommenderar

  eller
  - ○ papperslista med felsignaler samlade för varje felsignalerat objekt i blankettens variabelordning.

Fas 5 – Kvalitetssäkring

När insamlingen avbryts och allt material har granskats, framställs tabeller. Vid tabelleringen kontrolleras tabellceller, oftast genom jämförelser med motsvarande tabellvärden för föregående period. Denna slutgranskning kallar vi outputgranskning.

3.2.6 Outputgranskning

Outputgranskningens syfte är att kontrollera att inga allvarliga fel finns kvar i materialet. Kontroller utförs därför vanligen på aggregerade i stället för på enskilda variabelvärden. Granskning görs alltså primärt på tabellcellsnivå. Men sedan gäller det att identifiera det eller de objekt som gör att det aggregerade värdet blivit misstänkt. Det gör man genom att utföra kontroller för alla objekt som ingår i det misstänkta aggregatet. Granskningen görs alltså i två steg.

- Steg 1: identifiering av misstänkta tabellceller.
- Steg 2: identifiering av de mest misstänkta objekten inom den misstänkta tabellcellen.

Outputgranskning kan numera också göras som grafisk slutgranskning av mikrodata med EDA-metoder – med t.ex. programvaran SAS/Insight (se kapitel 5 och 8).

Checklista för outputgranskning

- ❑ Välj metod:
  - ○ tabellcellsgranskning
  - ○ granskning med EDA-metoder.

Tabellcellsgranskning

- ❑ Identifiera misstänkta tabellceller genom
  - ○ jämförelse med tidigare resultat
  - ○ jämförelse med andra celler
  - ○ jämförelser med motsvarande aggregat från andra undersökningar.
- ❑ Identifiera misstänkta objekt.

3.3 Elektronisk datainsamling

Vid SCB förekommer datorstödd datainsamling i form av elektroniska blanketter, datoriserade intervjuer (WinDATI) och s.k. tonvalsinsamlingar (Touchtone Data Entry, TDE). Elektroniska blanketter har fram till 2000 mestadels bestått av Excel-blanketter som distribueras och återsänds ifyllda via diskett eller e-post. Uppgiftsinsamling via webben finns för ett tiotal undersökningar. Exempel är:

- Elenergiundersökningen (MR/EN), där 250 elenergiföretag i år (2000) erbjuds möjligheten att lämna uppgifter via webben
- Räddningstjänststatistik (BV/BE), Hakim Sjöström
- Verksamhetsstatistik Svenska Kyrkan (ES/OE), Håkan Johansson
- Grundskolestatistik (AM/S), Kristina Lekeborn
- Forskningsbibliotek och ytterligare tre undersökningar på BV, Kjell Tambour
- Forskning och utveckling inom universitet och högskolesektorn (ES/FOI), Peter Skatt
- Intrastat (ES/UH), Jan Sävenborg

Rapporter finns från insamlingen vid AM/S och ES/OE. För den senare, se Helena Karlsson (2001).

År 2000 startade U/MET det s.k. indataprojektet som syftar till att samordna utvecklingen av applikationer för datainsamling. Stor vikt har lagts vid säkerhetsfrågor, verktyg samt metodfrågor vid blandad insamling. Projektet arbetar även med frågor kring standardisering av administrativ datainsamling via det för myndigheterna gemensamma Spridnings- och hämtningssystemet – SHS. SCB har medverkat i ett internationellt utvecklingsarbete, TELER-projektet, med målet att elektroniskt överföra data från uppgiftslämnarnas informationssystem till den statistiska myndigheten. Projektet har visat att kostnaderna för uppgiftslämnandet kan minska betydligt om de statistiska uppgifterna kan föras över till en s.k. elektronisk blankett direkt ur bokföringen och därefter överföras till statistikproducenten. För att vidareutveckla denna ansats måste ytterligare studier av bokföringssystem och SCB-undersökningar genomföras.

Weeks (1992) diskuterar i en översikt de möjligheter som öppnas vid användning av olika verktyg för datorstödda insamlingsmetoder.

3.3.1 Elektroniska blanketter

Elektroniska blanketter innebär att uppgiftslämnare via diskett eller elektronisk kommunikation, t.ex. webben och e-post, får ett programpaket bestående av blankett (med anvisningar, definitioner, föregående rapporterade värden med mera) samt ett granskningsprogram, dvs. ett dataregistreringsprogram med förprogrammerade kontroller, felmeddelande- och uppdateringsrutiner osv. Ett sådant program för granskning vid dataregistreringen måste vara betydligt mer genomarbetat när det gäller utformning av felmeddelandetexter och annan information än vad programmen för dataregistrering av postenkäter brukar vara. Detta registreringsprogram kan också användas i registreringen och kontrollen av de pappersblanketter som kommer från uppgiftslämnare som föredrar att rapportera på vanligt sätt, och kan då hanteras även av ny personal eller i utbildning av personal.

Generellt ställer elektroniska blanketter större krav på design av kontrollerna än vad som krävs vid dataregistreringsgranskning av postenkäter. Kontrollerna måste upplevas som väsentliga av uppgiftslämnarna, dvs. träffsäkerheten i kontrollerna måste vara hög. Dessutom bör alla kontroller som behövs i granskningen finnas med. Någon mer granskning ska inte behövas utöver outputgranskning. Det är en fördel om granskningen här kan göras variabelvis, så att programmet stoppas när en felsignal initieras och det flaggade värdet kan verifieras omedelbart.

Checklista för granskning vid elektronisk datainsamling
Se till att:

- ❑ felsignaler når uppgiftslämnarna just då de är i färd med att besvara frågorna och har tillgång till det underlag som behövs. Alla slarvfel (t.ex. registrerings-, summerings- och konsistensfel) som tar sig uttryck i felsignaler kan därmed åtgärdas direkt av uppgiftslämnarna eller av programmet.
- ❑ summeringar, andra sammanställningar, sortomvandlingar och härledda variabler beräknas av programmet och utgör därmed ett stöd för uppgifts-

lämnandet och verifieringen av felsignaler (jämför ovan självgranskningen i postenkätundersökningar)

- ❑ anvisningar, definitioner på variabler dyker upp på skärmen när uppgiftslämnarna kommer till aktuell fråga eller när de begär det via att trycka på en hjälptangent
- ❑ bakgrundsinformation, registrerade värden från tidigare rapporteringar med mera tillhandahålls uppgiftslämnarna vid besvarandet av varje fråga
- ❑ felmeddelandena förser uppgiftslämnarna med tänkbara orsaker till flaggningen, vilket dels underlättar verifieringen, dels ger möjligheter till automatisk registrering av felorsaker
- ❑ uppgiftslämnarna ges möjligheter att enkelt förmedla upplysningar. De ska kunna ge information till de särskilda omständigheter som gör att deras svar av programmet betecknas som ”misstänkta” eller varför de inte kan besvara frågor. Vidare ska de kunna ange om de följt anvisningarna om vilka komponenter som ska ingå i variabeln.
- ❑ processen ger information om huruvida felsignalerade uppgifter verkligen har kontrollerats (verifierats) av uppgiftslämnarna eller inte
- ❑ registreringar av tidsåtgång, ändringar, orsaker till ändringar, förklaringar till felsignaler, uppgifter om uppgiftslämnarkapacitet med mera görs automatiskt.

Problem eller förutsättningar:

- Uppgiftslämnarna måste ha tillgång till pc och enkelt kunna öppna den elektroniska blanketten.
- Vid diskettinsamling måste vid periodiska undersökningar individuella disketter framställas för varje uppgiftslämnare.
- Disketthanteringen utgör ett visst administrativt problem jämfört med andra elektroniska insamlingsmetoder – större ju fler uppgiftslämnare som ingår i undersökningen.

För mer detaljer om elektroniska blanketter, se Granquist, Hanaeus och Nilsson (1994) och Blom, Irebäck (2000).

3.3.2 Tonvalsinsamling

Tonvalsinsamling, Touch-tone Data Entry (TDE), innebär att uppgiftslämnarna via telefonens knappsats besvarar frågor på en blankett till en telefondator när som helst på dygnet.

Tekniken kan enbart användas i undersökningar med högst 5–10 variabler.

Bureau of Labor Statistics i USA har ca 15 års positiva erfarenheter av tekniken, som är billigare än både postenkäter och datorstödda telefonintervjuer (se Werking m.fl. 1988). I Sverige används tekniken av banker, försäkringsföretag, postorderföretag med flera. Vid SCB har tekniken använts bl.a. i socialbidragsstatistiken (Granquist 1994), och den används bl.a. i de undersökningar som anges i tablån:

Undersökning	Ansvarig	Objekt	Periodicitet	Urvalsstorlek	Antal variabler	Metod
Partihandel och andra tjänsteföretag	Daniel Lennartsson	Partihandelsföretag och andra tjänsteföretag	Kvartalvis	4 000	1	TDE
Detaljandelsundersökning	Daniel Lennartsson	Detaljhandelsföretag	Månatlig	3 500	1	TDE
Prisindex i producent- och importled (PPI)	Mats Haglund, Jan Lesins	Företag	Månad	1 200	1–25 prisnoteringar, ca 3 i snitt	TDE 65 procent, blankett 30 procent, övrigt 5 procent

Dataregistrering och granskning utförs av uppgiftslämnarna. Inknappade siffror motläses av telefondatorn, som utför validitets- och konsistenskontroller – varvid registrerings-, sort- och validitetsfel undviks redan vid källan. Sådana jämförelsekontroller måste vara enkla och helt naturliga för uppgiftslämnarna. Jämförelsetal mot inknappade eller tidigare rapporterade data meddelas omedelbart till uppgiftslämnarna, som då kan ta ställning till om lämnade data är rimliga eller inte. En erfarenhet från socialbidragsstatistiken är att uppgiftslämnarna reagerar på mindre förändringar än de gränser som tillämpades för motsvarande kontroller när data insamlades via postenkäter. Kontrollerna kan alltså göras parameterfria. Ytterligare en fördel är att man kan låta uppgiftslämnarna bekräfta att lämnade data är korrekta samt redovisa om vissa komponenter ingår eller är exkluderade i frågor. I likhet med elektroniska blanketter kan allt som sker i insamlingsprocessen automatiskt loggas, räknas och sammanställas.

3.3.3 Datorstödda intervjuer

Datorstödda intervjuer innebär att en intervjuare på en dator registrerar svar fråga för fråga – antingen vid egen utförd datainsamling, t.ex. prisinsamling i en butik, eller i en personlig intervju via telefon eller besök. Felmeddelanden kommer upp på skärmen så fort något svar ifrågasätts av de inprogrammerade kontrollerna. I intervjusituationen förmedlar intervjuaren meddelandet till intervjupersonen för verifiering. Ur kontrollprocessynpunkt gäller i övrigt samma riktlinjer som vid elektroniska blanketter, frånsett att felmeddelandetexten inte kräver samma omsorg. Notera särskilt att kontrollerna också här måste ha hög träffsäkerhet och vara uttömmande, så att behov av återkontakter elimineras.

3.4 Registerdata

Register är ett samlingsbegrepp för flera olika typer av förteckningar. Förteckningarna kan vara heltäckande eller avse en delmängd av en population (stor eller liten). Registret kan bygga på återkommande inventeringar och kompletteras med förändringar, anmälda eller observerade, under mellanliggande perioder. Fastighetstaxeringsregistret och lantbruksregistret är två exempel på denna typ av

register. Register kan också vara rent händelsebaserade. Dödsorsaksregistret är ett exempel på detta.

Register som upprättas av en annan myndighet eller organisation benämns ofta administrativa register. Administrativa register och även vissa externa databaser är primärt upprättade för andra ändamål än statistik.

En kopia på ett administrativt register omvandlas till ett statistiskt register med hjälp av registerstatistiskt metodarbete. Ett led i det registerstatistiska metodarbetet är granskning. Med vår terminologi är det fråga om outputgranskning i syfte att identifiera kvarvarande fel efter den granskning som registerhållaren har utfört. Denna granskning kan ha brister speciellt i fråga om de variabler som inte är av primärt intresse för de ändamål som registret ska uppfylla. SCB kan dock inte kontakta uppgiftslämnaren för att verifiera misstänkta fel. Granskningen får därför begränsas till att man utarbetar en rapport över förekomst av uppenbara och misstänkta potentiellt betydande fel och om möjligt över förekomst av systematiska fel (felkällor) i registret. Uppenbara fel kan i vissa fall ersättas med korrekta värden, nämligen när det är uppenbart vilket fel uppgiftslämnaren har gjort. Rapporten kan ses som en kvalitetsdeklaration av registret för dem som ska använda det. Grafiska metoder för outputgranskning rekommenderas (se t.ex. kapitel 5, som behandlar grafisk granskning). Granskning av registerdata är ett problem som ännu inte aktualiserats internationellt. Något systematiskt utvecklingsarbete har inte heller gjorts vid SCB.

Syftet med granskning av registerdata är att

- upptäcka fel
  - uppenbara fel
  - misstänkta värden (statistiska outliers)
- identifiera felkällor
- kvalitetsdeklarera registret för statistisk användning.

Kontroll av registerdata är en återkommande process som ska genomföras vid varje större uppdatering av registret.

Skillnaden mellan eget insamlat material och externa databaser är att man i de senare inte har någon kontroll över datainsamlingen. I stora drag kan man känna till vilka uppgifter som samlas in och ungefär när datainsamlingen äger rum. Dock saknar man ofta mer detaljerad information om exakt vilka uppgifter som samlas in och hur insamlat material granskas. Det är som regel svårt att från statistisk synpunkt påverka datainsamlingen, speciellt av variabler som inte är av primärt intresse för ägaren till databasen. Kontakt med databasägaren med t.ex. den ovan nämnda rapporten som underlag är i stort den enda möjligheten att på sikt kunna påverka registerkvaliteten och kontrollprocessens omfattning.

Uppdatering av SCB:s register med hjälp av externa databaser kan ske på i princip två sätt. Antingen görs uppdateringen med hjälp av registrerade förändringar eller också görs den med hjälp av nya kopior av den externa databasen. Dataöverföring av externa databaser sker i dag på elektronisk väg. I många fall upprättas ett parallellt register på SCB, en kopia, där uppgifterna från de externa databaserna samlas.

3.4.1 Kontrollprocess

I Lindström (1999) behandlas kvalitetssäkring av register. Man kan i allmänhet inte påverka (”förbättra”) kvaliteten i inkommande data.

Indirekta metoder används för kontroll av inkommet material: konsistenskontroller mellan variabler, att klassificeringsvariabler antar godkända värden osv. Outputgranskning för olika redovisningsgrupper ger möjligheter att undersöka om antalet objekt är rimligt eller om fördelningen av antalet objekt på olika redovisningsgrupper är rimligt. På motsvarande sätt kan man utnyttja outputgranskning för att kontrollera aggregat på redovisningsgrupper, såsom medelvärden och totaler, för några centrala variabler. Jämförelsematerial är tidigare produktionsomgångar eller andra register. Information från databasägaren om version, period och antal observationer måste naturligtvis utnyttjas.

Även om man hittar fel, är det i praktiken inte möjligt att verifiera felen. Det vi rekommenderar är att man i en rapport dokumenterar vilka kontroller som utförts, vilka variabler som kontrollerats samt ger en beskrivning av misstänkta värden (med hjälp av t.ex. kontrollfunktion, intervall). Rapporten ska överlämnas till databasägaren för kännedom. Den bör om möjligt innehålla en systematiskt uppställd felstatistik som underlag för diskussioner av prioriterade åtgärder.

I register med många variabler är det inte rimligt att kontrollera alla variabler utan endast dem som är nödvändiga för officiell statistik och vanliga återkommande tillämpningar, t.ex. urvalsramar. Om registret ska användas för andra tillämpningar, måste man överväga om man behöver göra ytterligare kontroller för att se i vilken utsträckning registret kan användas för tillämpningen i fråga.

En kontrollprocess bör innehålla följande moment:

- kontroll av att dataöverföringen fungerar
- kontroll av att period och version är korrekt
- kontroll av att alla objekt finns med
- variabelkontroll av för SCB intressanta variabler
- kontroll av identitetsvariabler
- kontroll av att variabelvärden är rimliga (konsistenskontroller, godkända värden)
- frekvensberäkningar för vanliga indelningsvariabler (outputgranskning).

En sammanställning av vilka kontroller som genomförts och resultatet av kontrollerna ska skickas till databasägaren samt dokumenteras i den interna kvalitetsdeklarationen. Även fältobservationer av misstänkta fel, när registret används som t.ex. urvalsram, ska dokumenteras. Denna dokumentation lämnas till den registeransvarige på SCB.

Ett problem med kontroll av och kvalitet på statistiska register sammanställda med hjälp av en eller flera externa databaser kan bero på ”envägskommunikation”. (A) är ägaren till en extern databas. Mottagaren (B) på SCB är den registeransvarige. Användaren (C) av registret är en person på eller utanför SCB. Man kan förutsätta att alla tre kontrollerar registret måhända med olika utgångspunkter. Om kontakterna mellan de tre olika instanserna är begränsad och envägs (A→B→C), är det hög risk att onödigt merarbete uppstår. För att undvika onödigt arbete krävs samarbete mellan A, B och C. Detta kan man t.ex. åstadkomma med rapporter över

uppenbara eller misstänkta fel. Återkommande och samordnade evalveringsstudier kan förbättra kvaliteten på sikt.

3.4.2 Process- och kontrollvariabler

Speciellt frekvens- och antalsberäkningar är enkla processvariabler med hjälp av vilka man bland annat kan upptäcka om alla objekt kommit med. Tidsserier i form av frekvenser för respektive aggregat för centrala variabler är ett annat exempel på kontrollvariabler. Genom att matcha mot andra register, t.ex. en tidigare version, kan man hitta avvikare. Ofta är det tillräckligt med jämförelser på aggregerad nivå, men ibland behöver jämförelser göras på objektsnivå för enstaka variabler. En sammanfattning av dessa kontroller ska infogas i den dokumentation som överlämnas till databasägaren.

Kontroller kan genomföras både på objektsnivå och på aggregerad nivå (redovisningsgrupper). Outputgranskning, kontroll på aggregerad nivå, kan utföras t.ex. grafiskt med hjälp av SAS/Insight och jämförelse av tabellceller med hjälp av Excel.

3.4.3 Dokumentation av kontrollprocessen

Alla kontroller och kontakter med databasägaren ska dokumenteras. Uppdatera proceduren med aktuella acceptansgränser. Inkludera även fältobservationer av uppenbara fel när registret används som t.ex. urvalsram. Stora register måste dokumenteras i metadata med tydlig information om vilka variabler som kontrollerats.

3.5 Referenser

Bethlehem, J.G., A. J. Hundepool, M. H. Schuerhoff, and L. F. M. Vermeulen (1989), *”BLAISE 2.0 An Introduction”*, Vorburg, The Netherlands: Central Bureau of Statistics, February 1989

Blom, E. och Irebäck Hans (2000): Utveckling av elektronisk datafångst 2001 – 2003. PM 2000-01-31.

Granquist L. (1992): ”Granskningsprocessen i omvandling genom integrering av databeredningsarbetet”, GRANSK-PM NR 28, 1992–05–18.

Granquist L., C.-G. Hanaeus och G. Nilsson (1994): ”Elektronisk distribution och insamling av data – Erfarenheter och generella slutsatser från ett försök”, Rapport 30 juni 1994

Granquist, L. (1994): ”Tonvalsinsamling i Socialbidragsstatistiken”, Slutrapport 30 juni 1994

Karlsson, Helena (2001): ” Insamling av verksamhetsstatistik”, Utvärdering 2001–04–09

Linacre, S. J., and D. J. Trewin (1989), ”Evaluation of Errors and Appropriate Resource Allocation in Economic Collections”, *Proceedings of the US Bureau of the Census Fifth Annual Research Conference*, U. S. Department of Commerse, Bureau of the Census, March 19–22, 1989, pp. 197–209.

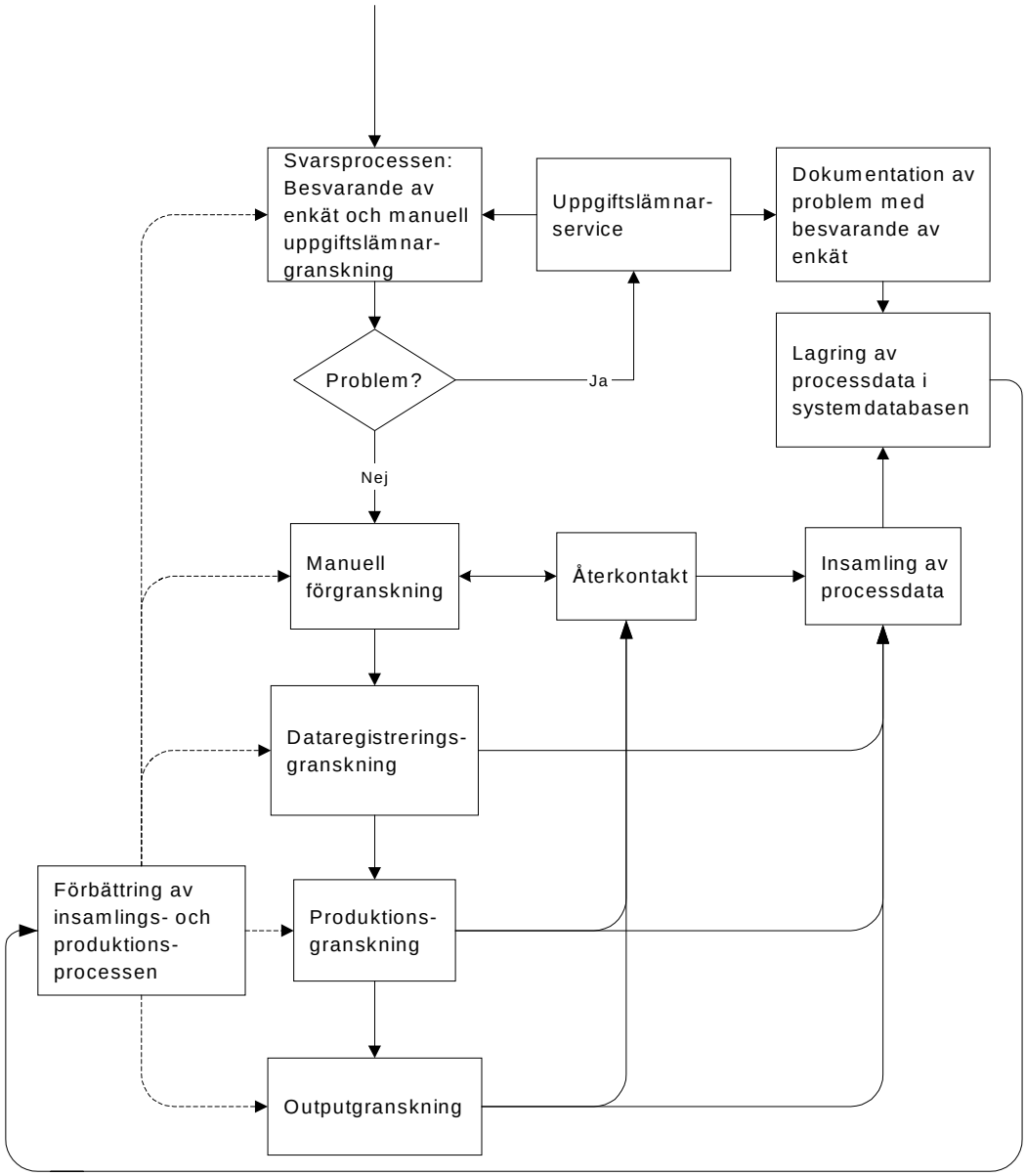
Lindström, Håkan L. (1999); Kvalitetssäkring i register för statistikproduktion med administrativt underlag. Rapport från Registerprojektet, November 1999, SCB.

Pierzchala, M. (1995), ”Editing Systems and Software”, in B.G. Cox, D.A. Binder, N. Chinnappa, A. Christianson, M.J. Colledge and P.S. Kott (eds.) *Business Survey Methods*, New York: Wiley, pp. 425–441.

Weeks, M. F. (1992), ”Computer-Assisted Survey Information Collection: A Review of CASIC Methods and Their Implications for Survey Operations”, *Journal of Official Statistics*, Vol. 8, No.4, 1992, pp. 445–465.

Werking, G., A. Tupek, and R. Clayton (1988), ”CATI and Touchtone Self-Response Applications for Establishment Surveys”, *Journal of Official Statistics*, Vol. 4, No.4, 1988, pp. 349–362.

Bilaga kapitel 3
Granskningsprocessen i en postenkätundersökning



4 Kontrollmetoder

I kapitel 2 klassificeras fel i uppenbara och misstänkta fel. Misstänkta fel delas i sin tur in i avvikelsefel och definitionsfel. Klassificeringen motiveras med att feltyperna representerar skilda granskningsproblem.

I det här kapitlet diskuterar vi underlag för formulering av kontroller i allmänhet och metoder för att effektivisera avvikelsekontroller i synnerhet. Metoder för identifiering av definitionsfel tas upp i kapitlet om grafisk granskning.

Avsnitten om avvikelsekontroller tillhandahåller generella råd och rekommendationer, redovisar några utvalda metoder för selektiv granskning och beskriver i detalj en metod för effektivisering av kvotkontrollen.

Checklista för kontrollmetoder

- ❑ Klassificera möjliga fel i uppenbara och misstänkta fel.
- ❑ Skilj mellan avvikelse- och definitionsfel.
- ❑ Begränsa och samordna kontrollerna för misstänkta fel.
- ❑ Anpassa kontrollernas acceptansgränser till aktuella data.
- ❑ Använd grafiska metoder i outputgranskningen, men också i produktionsgranskning.
- ❑ Sammanställ statistik över kontrollers träffsäkerhet.
- ❑ Sikta mot minst 60 procent träffsäkerhet.

4.1 Underlag för formulering av kontroller

Utifrån kunskaper inom sakområdet och goda insikter i de informationsbehov som undersökningen ska tillgodose kan man konstruera kontroller för att identifiera i synnerhet uppenbara fel.

Källor för utformning av kontroller är dessutom

- blanketten
- definitioner och anvisningar för undersökningsvariabler
- information i urvalsram och utsändningsregister
- dataanalys (EDA) med hjälp av t.ex. SAS/Insight på data från undersökningen
- processdata och övriga erfarenheter.

4.1.1 Några speciella problem

Vissa variabler – särskilt i den ekonomiska statistiken – måste av skatterättsliga eller juridiska orsaker eller av jämförbarhetsskäl följa definitioner som fastställs och revideras av särskilda nationella och internationella organ. Begreppet lön innehåller t.ex. cirka 200 komponenter. I en statistisk undersökning kan många av dessa komponenter vara irrelevanta och får inte räknas in i variabeln lön för att tillgodose undersökningens syften. Vilka komponenter som ska inkluderas och vilka som ska exkluderas måste klart framgå av variabeldefinitionen eller av anvisningar till uppgiftslämnarna.

Problemet är att få reda på om uppgiftslämnarna verkligen följer anvisningarna. För sådana typer av variabler ska det övervägas om man i stället ska fråga efter samtliga komponenter och maskinellt beräkna undersökningsvariabeln genom additioner och subtraktioner.

Ett alternativ är att enbart fråga efter de komponenter som ska vara med och antingen fråga efter summan eller maskinellt beräkna summan som en undersökningsvariabel. Man kan också i blanketten ställa frågor av typen:

”När du svarade på frågan *nnnn* räknade du då in komponent *bbbb*, komponent *cccc*,…, och såg till att värden på komponent *xxxx*, komponent *yyyy* ,…, inte finns med i de lämnade siffrorna. Om du inte gjorde detta, ange varför och ange också om du i framtiden skulle kunna rapportera enligt våra önskemål.”

Sådan information kan användas i skattningen – men framför allt vid förbättringar av undersökningen.

Ett liknande situation uppstår när en variabel (summavariabel) är uppdelad på ett antal delvariabler, som alla är viktiga undersökningsvariabler. Svaret på summavariabeln ska då vara lika med summan av delvariablernas värden. Vid granskningen söker man alltid att imputera värden på delvariablerna när svar endast finns på summavariabeln och att i övrigt säkerställa att summan aritmetiskt stämmer med summavariabelns värde. Kontroller för att se till att värdena på delvariablerna summerar sig till värdet på summavariabeln kallas *balanskontroller*.

Det är viktigt att ta fram statistik över behovet av imputeringar för delvariablerna och över frekvensen förändringar av summavariabeln. Stort behov av imputeringar tyder på att uppgiftslämnarna har svårt att ge värden på delvariablerna. Många ändringar i summavariabeln indikerar att vi har att göra med en allvarlig felkälla.

Ett alternativ som bl.a. föreslagits av Linacre and Trewin (1989) är att bara fråga efter delvariablerna och maskinellt beräkna summavariabeln.

En annan möjlighet är att föra in t.ex. följande anvisning och uppgiftslämnarkontroll:

> ”Ange delvärdena. Beräkna summan av dess och jämför med motsvarande uppgift i bokföringen innan du för in värdet på summaraden.”

Båda fallen illustrerar att man måste integrera utformningen av mätinstrument och datainsamling med konstruktionen av kontroller.

4.1.2 Blanketten

Blanketten utgör huvudunderlaget för formulering av kontroller, i synnerhet mot uppenbara fel.

Kontroller kan skapas genom att man systematiskt går igenom blanketten och dess variabler på följande sätt: För varje variabel noterar man vilka värden som är tillåtna. Därefter anger man för variabel nummer 1 vilka värden som är tillåtna i kombination med variabel 2, 3, 4 osv. Sedan gör man samma sak för variabel 1 och 2 tillsammans med variabel 3, 4 osv. Eftersom varje enskild variabel endast är relaterad till högst en handfull andra variabler, är proceduren inte särskilt arbetskrävande. Men det är ett effektivt sätt att se till att man garderar sig mot alla möjliga konsistens- och strukturfel.

Kompletterande källor för identifiering av speciellt partiellt bortfall är informationen i urvalsramen eller utsändningsregistret samt data från tidigare perioder och eventuella andra undersökningar.

4.1.3 Studier i datamaterialet

Kontroller mot systematiska fel, definitionsfel och avvikelsefel bygger i första hand på ämneskunskaper. Ett viktigt komplement till ämneskunskaperna är att man studerar materialet (dataanalys) när man skapar kontroller. Den primära uppgiften är att

- testa kontroller
- finna lämpliga acceptansgränser för kontrollerna
- finna på förhand okända misstänkta svarsmönster, för kontroller mot definitionsfel.

Vi rekommenderar att man använder SAS/Insight. I denna programvara finns det stöd för att skapa testvariabler. Man kan även göra transformationer av testvariabler, vilket underlättar val av acceptansgränser. Effekten av valda acceptansgränser kan lätt avläsas.

Dessutom gör programvaran det möjligt att identifiera misstänkt felaktiga svarsmönster, dvs. systematiska svarsfel. Detta ska utnyttjas för att skapa kontroller mot definitionsfel.

4.1.4 Processdata och erfarenheter

Alla potentiellt betydelsefulla felkällor kan inte förutses. Nya felkällor uppstår på grund av teknisk och administrativ utveckling samt genom förändringar av olika slag i samhället. Förändringarna i datamaterialet aktualiserar nya eller förändrade kontroller samt val av acceptansgränser.

Det finns två möjligheter att identifiera behov av förändring av kontroller och justering av acceptansgränser:

- processdata
- analys av effekter av granskning i data med hjälp av SAS/Insight.

Processdata används för att identifiera felkällor och justering av parametrar i acceptansgränserna. Processdata är en central och viktig information om insamlings- och produktionsprocessen (se kapitel 5 och 6). Man kan säga att processdata är en systematisk sammanställning av erfarenheter.

SAS/Insight används för att identifiera felaktiga svarsmönster, dvs. hitta och kvantifiera möjliga definitionsfel, för att analysera och kontrollera acceptansgränser, för att analysera processdata och för att studera effekten av granskningen.

4.2 Generella tips för effektiva avvikelsekontroller

Övergripande metoder för identifiering av avvikande observationer är att:

- skapa effektiva kontroller som är väl anpassade till undersökningens data (Det är en fördel om kontrollerna kan utformas så att verifieringsarbetet kan prioriteras till de potentiellt största felen.)
- begränsa antalet möjliga kontroller per variabel till de mest effektiva

- prioritera verifieringsarbetet, t.ex. med poängfunktioner, till de potentiellt viktigaste objekten.

Dessa förslag kan med fördel tillämpas samtidigt, vilket illustreras av Engström (1995).

4.2.1 Effektiva kontroller

Kontroller för misstänkta avvikelsefel ska

- ha minst 60 procent träffsäkerhet
- hitta de flesta felen, åtminstone fel med betydande effekt på undersökningens resultat.

SAS/Insight är ett hjälpmedel för att studera kontrollernas egenskaper med avseende på

- symmetri
- acceptansgränser
- gruppering.

Av kapitel 2 framgår att det förväntade antalet avvikelsefel per variabel är lågt, troligen någon enstaka procent av antalet observationer. I kombination med målet att varje kontroll ska ha hög träffsäkerhet, minst 60 procent, innebär det att avvikelsekontroller inte ska felsignalera mer än högst 3–4 procent av de rapporterade variabeluppgifterna.

Det är en fördel om testvariabeln har en någorlunda symmetrisk fördelning, eftersom man då slipper en parameter i acceptansområdet.

4.2.2 Kontroller mot avvikelsefel

En kontroll för misstänkta fel består av en *testvariabel* och ett *acceptansområde*. I kontrollen kan det också ingå ett villkor som avgör vilka objekt som ska kontrolleras.

Exempel på testvariabler (kontroller) som används i KLP

Genomsnittlig timlön = {Lönesumma}/{Arbetade timmar}

Arbetade timmar per anställd = {Arbetade timmar}/{Antal anställda}

Relativ förändring = {Lönesumma månad t}/{Lönesumma månad t-1}

Regeln som avgör om ett objekt ska felsignaleras eller inte benämns acceptansgränser. Området som gränserna omsluter kallas acceptansområde. Granskningen kan ofta bli mer effektiv om man delar in populationen i grupper.

Exempel på acceptansområde

Antag att en kontroll med testvariabeln ”Genomsnittlig timlön” har acceptansområdet (50,120). Ett företag felsignaleras då om värdet på ”Genomsnittlig timlön” är mindre än 50 eller större än 120.

Ämnesmässiga fakta och framför allt empiriska studier på data från tidigare undersökningar (med hjälp av analysprogram, som t.ex. SAS/Insight) ger ofta en god uppfattning om kontroll, acceptansgränser och lämpliga grupperingar. När gruppering är aktuell, anpassas acceptansgränserna till respektive grupp.

Exempel på gruppering av populationen

Beroende på löneavtal och lönestrukturer har testvariabeln ”Genomsnittlig timlön” olika egenskaper inom olika näringsgrenar, när det gäller både nivå och variation. Det är därför rimligt att skräddarsy acceptansgränser för ”Genomsnittlig timlön” för varje näringsgren.

4.2.3 Begränsning av antalet kontroller

Man skulle kunna tro att man, för att gardera sig mot olika tänkbara fel, borde konstruera många kontroller för varje variabel. Emellertid måste man vara uppmärksam på att de enskilda kontrollerna faktiskt bidrar till att identifiera fel med rimlig träffsäkerhet. Det finns många exempel på att de fel som ”extra”-kontroller hittar även upptäcks av en annan kontroll av variabeln. Det medför att ”extra”-kontrollen enbart ökar antalet onödiga felsignaler. Därför måste man för varje variabel välja ut de kontroller som effektivast identifierar fel – helst genom att man utvärderar på data från undersökningen. En eller högst två kontroller för en variabel är i allmänhet tillräckligt.

Exempel på reducering av antalet kontroller

I en lönestatistik kan för kontroll av lönesumma och antal arbetade timmar följande fyra kontroller vara tänkbara:

{lönesumma}/{lönesumma föregående period}

{arbetade timmar}/{arbetade timmar föregående period}

{beräknade timlönen}/{beräknade timlönen föregående period}

{lönesumman}/{arbetade timma}

Vid en studie fann vi att för kontroll av såväl lönesumma som antal arbetade timmar var det tillräckligt med lönesumma dividerat med antal timmar. De tre övriga kontrollerna bidrog inte till identifiering av felaktiga observationer utan ökade enbart andelen onödigt flaggade observationer.

4.2.4 Prioritering av objekt genom poängfunktioner

En framgångsrik metod att begränsa verifieringen till de misstänkta observationer som kan ha betydelse på skattningarna är att använda s.k. poängfunktioner (engelska: *score functions*). Metoden innebär att varje objekt för vilket minst ett variabelvärde felsignaleras tilldelas poäng, som beror på objektets urvalsvikt, variabelns betydelse, avvikelsen från ett riktvärde osv. Därefter verifierar man enbart de objekt vars poäng överstiger ett värde som är fastställt på förhand.

4.3 Utformning av kontroller

4.3.1 Testvariabler

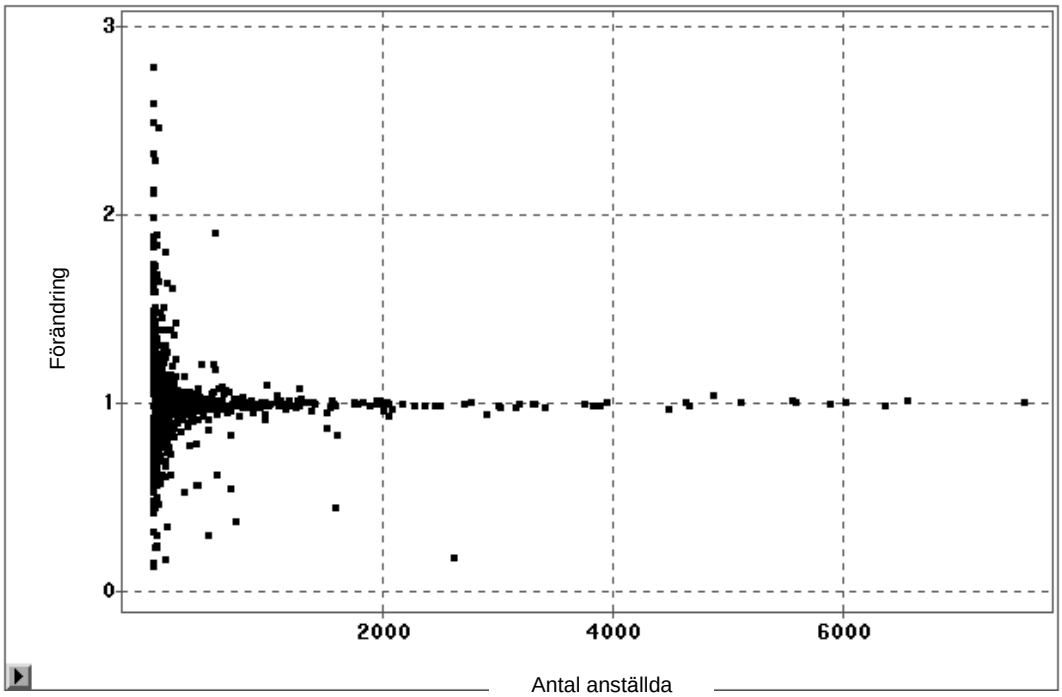
Testvariabeln är antingen en insamlad variabel eller ett aritmetiskt uttryck baserat på undersökningens variabler.

Som påpekats i 4.2.1 Effektiva kontroller, underlättar det om testvaribeln har en symmetrisk fördelning. Medlet för att åstadkomma symmetri är att transformera testvariabeln. I företagsundersökningar med deras ofta sneda fördelningar är logaritmering i allmänhet en bra metod.

I företagsundersökningar används ofta ”Relativ förändring”. Detta är en problematisk variabel om man använder den utan modifieringar. Den naturliga variationen i en ren förändringskvot är av naturliga skäl betydligt större för små företag än för

stora företag, vilket illustreras av diagrammet nedan. Ett alternativ är att som testvariabel använda differensen för små företag och förändringskvoten för stora företag. Men då måste man fastställa gränser dels för differensen och kvoten, dels för när differensen respektive förändringskvoten ska användas.

Diagram 1
Relativ förändring av antalet anställda februari till mars 2000



Diagrammet visar den *relativa förändringen av antal anställda från februari till mars 2000* plottat mot *antal anställda i februari*. Av diagrammet framgår tydligt att variationen i förändringen beror på företagets storlek (mätt i antal anställda). Om samma acceptansgränser används för samtliga företag, riskerar man dels att få många onödiga återkontakter, dels att stora fel inte upptäcks genom att acceptansområdet tvingas bli alltför brett för stora företag. (Källa: Lönestatistik privat sektor)

En metod för att lösa nämnda problem som i experiment och praktisk användning världen över (inklusive inom SCB) visat sig framgångsrik, är att man konstruerar en testvariabel som innehåller både ett storleksmått (eller den absoluta differensen) och den relativa förändringen. Det finns olika utformningar av den idén. Den som vi rekommenderar är Hidiroglou-Berthelots metod, se bilagan till kapitel 4, främst därför att det är den som vi har erfarenheter av. Metoden kan med fördel användas för andra typer av kvoter än förändringskvoter, se van de Pol (1994).

4.3.2 Acceptansområde

Acceptansgränser ska bestämmas på huvudsakligen följande grunder:

- ämnesmässiga fakta
- empiriska fakta erhållna genom
  - studier av testvariabelns egenskaper inom olika grupperingar av undersökningsenheter (t.ex. näringsgren, företagsstorlek eller region). Använd *lådagram*.

- erfarenheter av hur kontrollen har fungerat i tidigare undersökningar eller i en genomförd pilotundersökning. Det är träffsäkerheten som man ska studera.
- studier av säsongsmässiga variationer (gäller kortperiodiska undersökningar).

Ett generellt hjälpmedel är grafisk analys med hjälp av SAS/Insight.

Acceptansgränser ska för varje enskild kontroll anpassas till data från tidigare undersökningar och från den aktuella undersökningen. De ska inte sättas subjektivt, vilket ofta leder till asymmetrisk granskning som i sin tur leder till bias. Alltför restriktiva regler för acceptans leder till *övergranskning*. Det innebär en stor mängd felmeddelanden. En effekt blir att de granskare som tar hand om felmeddelandena tappar förtroendet för maskinens granskning och därför tänjer gränserna efter egna uppfattningar om vad som kan vara viktigt att följa upp. En annan effekt är att enstaka fel kan tappas bort i mängden, och de kan vara stora.

Acceptansgränser ska alltså sättas utifrån de data som ska granskas och baseras på fakta. Man kan utnyttja den grafiska metoden men också ämneskunskap. Detta kan åstadkommas genom att man låter gränserna bero på parametrar som exempelvis median och kvartilavstånd, beräknade från data i den aktuella produktionscykeln. Eftersom ogranskade data används för att bestämma acceptansgränserna, bör man inte använda parametrar som påverkas kraftigt av outliers (t.ex. medelvärde och standardavvikelse). Ett sådant tillvägagångssätt tillämpas t.ex. i Hidiroglou-Berthelot-metoden.

En enkel men kraftfull metod är att acceptansgränserna sätts lika med ({1:a kvartilen}–{k * kvartilavståndet},{3:e kvartilen}+{k * kvartilavståndet}) där *k* är en konstant som bestämmer bredden på acceptansområdet. Det gäller att finna lämpliga utgångsvärden på denna konstant och därefter följa upp kontrollernas egenskaper när det gäller *träffsäkerhet* och betydelsen av hittade fel vid varje undersökningstillfälle.

Lämplig bredd på acceptansområdet
Studier har visat, se Anderson (1989), att lämpliga acceptansgränser kan fås med k=3. Det går inte säga att detta gäller generellt, men det kan kanske ändå gälla som riktmärke. Ta gärna hjälp av lådagram i SAS/Insight för att bestämma acceptansområde och följa upp kontrollens träffsäkerhet.

Vårt råd bygger på arbetet kring utvecklingen av aggregatgranskningskontroller (se t.ex. Wickberg 1990) är att acceptansgränser bör sättas så långt från den stora massan av observationer men så nära den första outliern som möjligt (Granquist 1990).

I de experiment med Hidiroglou-Berthelot-metoden (bilagan till kapitel 4) som redovisas i Höglund (1994), var de bästa värdena på de konstanter som styr bredden på acceptansområdet: C = 41 och U = 0,4. (U har att göra med objektens storleksfördelning.) Detta gällde för alla de kontroller i den statistik (dåvarande order- och leveransstatistiken) som studerades. Grovt motsvarar detta Andersons (1989) resultat.

Wickberg (1990), Andersson (1989) och Höglund (1994) har visat att k=3 är ett bra utgångsvärde. K-värdet justeras när man studerar processdata. Observera att vi i 4.2.1 Effektiva kontroller, har sagt att andelen felsignaler inte ska vara högre än 4 procent.

Vi rekommenderar att man testar olika k-värden. Empiriska resultat från tidigare eller liknande undersökningar tillsammans med ämneskunskap kan vara vägledande. Lådagram i SAS/Insight erbjuder ett utmärkt stöd för att både välja olika k-värden och testa träffsäkerheten i kontrollen. Om man inte har tillgång till någon hjälpinformation, föreslår vi att man börjar med låga k-värden, dvs. snäva intervall, för att genom ökade värden på k öka intervallbredden.

Ett alternativt sätt att få fram lämpliga värden på förekommande parametrar/konstanter är att utgå från data som granskats med smalare acceptansområden än nödvändigt. Särskilt rekommenderas det att nya undersökningar och pilotundersökningar alltid använder smala, men ändå någorlunda rimliga, acceptansområden för att man ska få många återkontakter och därigenom kunna skaffa sig information om felkällor och uppgiftslämnarproblem.

När uppgiftsinsamlingen (i vidare mening) har avslutats, beräknas träffsäkerheten för olika intervall enligt exemplet nedan. Sist analyserar man resultaten för att bestämma ”bästa” acceptansgränser (helst uttryckta i parametrar och konstanter). Vår uppfattning är att man bör sikta mot en träffsäkerhet på åtminstone 60 procent och sträva mot uppemot 80 procent.

Exempel på metod att få fram lämpliga acceptansgränser

> Antag att för testvariabeln ”Genomsnittlig timlön” används acceptansgränserna (70,100) som snävast tänkbara intervall. Ett effektivare acceptansintervall kan då beräknas genom analys av träffsäkerheter från simuleringar i materialet med exempelvis intervallen (30,180), (30+k,180–l),…,(70,100) med k = 10 och l = 20.

En alternativ metod för att testa nya kontroller och acceptansgränser beskrivs i avsnitt 8.3.

4.3.3 Gruppering

Om för vida acceptansgränser tillämpas över hela populationen, medför detta att man inte upptäcker fel som är betydelsefulla för undersökningen. Timlönen för t.ex. ett städföretag kan vara en tiondel av genomsnittstimlönen för ett datakonsultföretag. För att man effektivt ska kunna urskilja misstänkta observationer, är det nödvändigt att acceptansgränser för kontroller mot avvikelsefel tar hänsyn till stora skillnader i medelvärden och varians för olika branscher. Men det tjänar inget till att göra en alltför detaljerad indelning av populationen.

En utmärkt metod är att med SAS/Insight framställa lådagram över testvariabelns fördelningar för att få underlag för en gruppering. Lådagram ger en visuell bild av fördelningars medianer och spridning. I SAS/Insight kan man få alla lådagram för undergrupper till en grupp (t.ex. hela populationen) utskriven i ett diagram. Med blotta ögat ser man i ett sådant diagram enkelt om indelningar behöver göras och i så fall vilka (se kapitel 5).

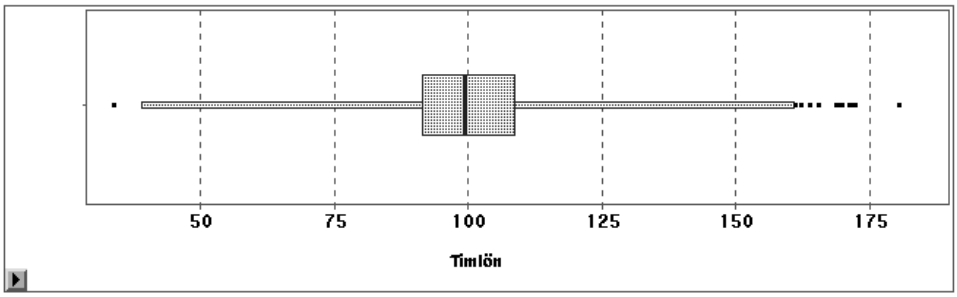
För att testa om grupperingen tillför något i granskningsarbetet börjar man med att finna acceptansgränser för hela materialet utan gruppering. Nästa steg är att under-

söka om en uppdelning av materialet ger väsentligt annorlunda acceptansgränser för olika delgrupper. Vissa klassificeringsvariabler, som t.ex. SNI-kod, är hierarkiskt uppbyggda, vilket gör det möjligt att testa gruppering från grov indelning till allt finare indelning. Två exempel får illustrera tekniken. Det första exemplet – hämtat från en lönestatistisk undersökning – bygger på granskat material. I det andra exemplet – prisförändringar på grönsaker i KPI – har några stora fel inplanterats.

Exempel på att finna acceptansgränser. Lönestatistik.

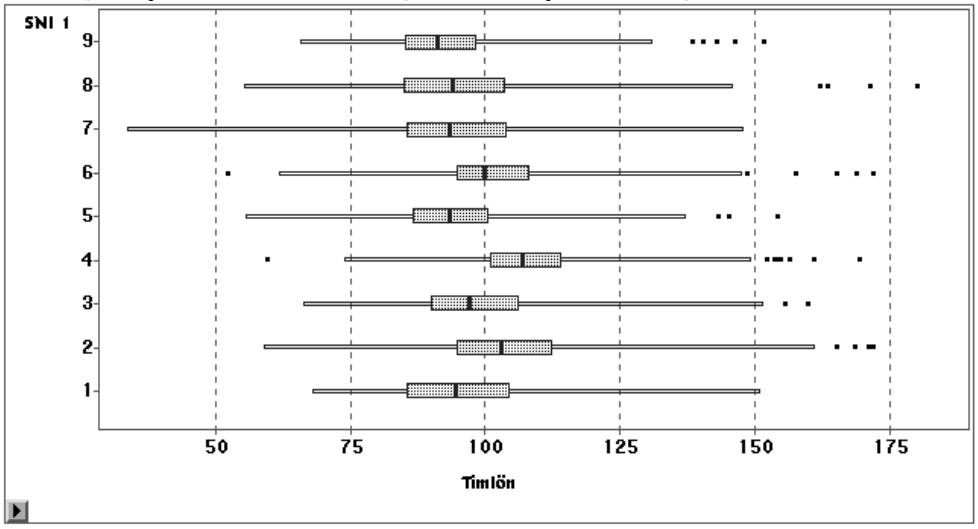
En av granskningsvariablerna i en lönestatistisk undersökning är timlön, dvs. lön per arbetade timmar. Acceptansgränserna (k=3) bestäms enklast med hjälp av ett lådagram i SAS/Insight. Utnyttja referenslinjerna för att ”bestämma” acceptansgränserna. Om man betraktar hela materialet, kan acceptansgränserna fastställas till ungefär (40,160) kronor per timme.

Diagram 2
Timlön, lön per arbetade timmar, i hela populationen



De ovan valda acceptansgränserna leder sannolikt till övergranskning. Vi försöker därför inledningsvis med en relativt grov gruppering av materialet. Vi väljer **bransch**, SNI-kod på ensiffernivå. Acceptansgränserna varierar med SNI-kod (diagram 3).

Diagram 3
Timlön, lön per arbetade timmar, fördelade på SNI-kod, ensiffernivå



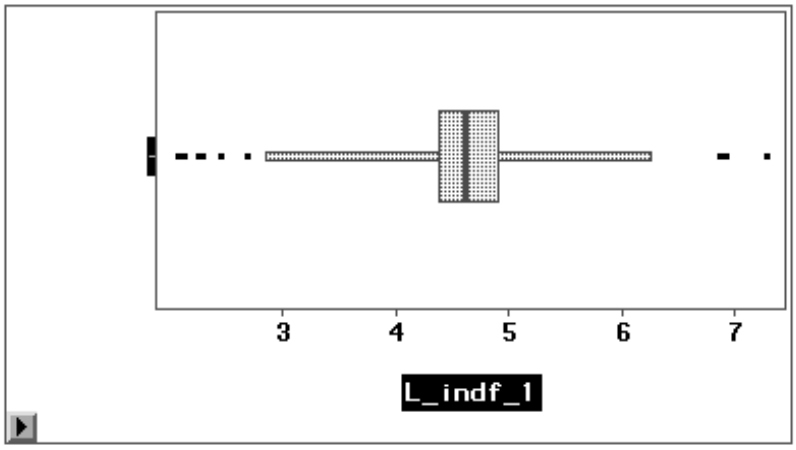
För de flesta branscher på ensiffernivå är troligen acceptansgränserna (40,160) bra utgångsvärden, utom möjligen för bransch 2 och 7. Ytterligare uppdelning av branscherna på t.ex. tvåsiffernivå kan ge bättre anpassade acceptansområden. Utnyttja ämneskunskaper i kombination med reglerna i 4.2.1 Effektiva kontroller.

Utnyttja t.ex. referenslinjerna och börja utifrån och gå mot mitten. Observera att man kan välja hur täta referenslinjerna ska vara.

Exempel på att finna acceptansgränser. Grönsaker i KPI.

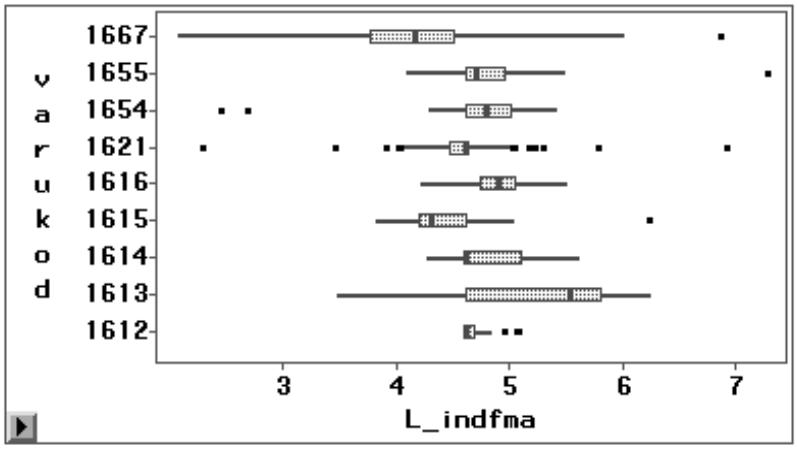
En kontroll i KPI bygger på månadsvisa prisförändringar, här uttryckt som ett index, och är logaritmerad. (Värdet 100 motsvarar i den naturliga logaritmen 4,6). För varugruppen grönsaker är det naturligt med ibland relativt stora prisförändringar, bland annat beroende på säsongvariationer. I den första bilden betraktar vi därför prisförändringen för hela undergruppen grönsaker.

Diagram 4
KPI, grönsaker. Prisförändring



Undergruppen grönsaker representeras av flera varor. Vi delar därför upp gruppen med avseende på representantvarorna.

Diagram 5
KPI, grönsaker fördelat på representantvaror



Om man jämför med exemplet ovan, finns det i KPI-exemplet större skillnader mellan acceptansområdena för undergrupperna jämfört med huvudgruppen. Acceptansområdet ligger mellan knappt 3 och drygt 6 för hela gruppen, vilket t.ex. ska jämföras med ungefär (4.2, 5.5) för varugruppen 1654. Lägg också märke till att för varugruppen 1621 skulle vi inte upptäcka några avvikare.

Ett annat men kanske inte lika effektivt sätt är att grunda indelningar av populationen på ämneskunskaper och erfarenheter helst i form av ett processdatasystem (se kapitel 7). Både processdata och erfarenheter är bra underlag till förbättringar.

4.4 Selektiv granskning – metoder för prioritering av objekt eller variabelvärden för verifiering

Selektiv granskning kallas kontrollprocesser som används för att skilja ut (och eventuellt åtgärda) endast de felsignalerade objekt vars eventuella fel har stor presumtiv inverkan på skattningarna för någon av de felsignalerade variablerna. Det är alltså ett sätt att tillgodose kravet på kontroller att de inte bara ska ha hög träffsäkerhet utan även att felen ska ha betydelse på skattningsnivå (se 4.2.2). Eftersom den här tekniken kräver skattningar och andra data från en tidigare undersökning, innebär detta i praktiken att den i första hand är intressant för periodiska undersökningar.

Tekniken kallades tidigare allmänt för makrogranskning, en benämning som fortfarande används. En översikt av sådana metoder ges i Granquist (1994). Här beskrivs idéerna bakom top-downmetoden och grafisk granskning.

Top-down är en allmänt använd teknik i granskning. Den innebär att granskning och eller verifiering utförs i prioriteringsordning, där prioritet bestäms efter potentiell effekt på skattningsnivå eller värde på testvariabel. Efter varje ändring av data beräknas nya skattningar. Granskningen upphör när skattningarna inte längre påverkas. Top-down används med fördel i outputgranskning (Anderson 1989) men kan mycket väl tillämpas i produktionsgranskning (Granquist 1994).

Vid grafisk granskning är det granskaren och effekterna av verifieringen som ”bestämmer” acceptansgränserna. Tekniken är då att utifrån den grafiska representationen av data börja med de värsta och sluta när verifieringen inte längre får effekter på skattningarna.

Selektiv granskning kan även ses som en process där de felsignalerade objekt väljs ut som kräver en omsorgsfull verifiering – medan de övriga objektens felsignaler antingen lämnas därhän, tas om hand av ett automatiskt imputeringsprogram eller utsätts för endast enkel och framför allt resurssnål verifiering, dvs. utan återkontakter. Vid Statistics Canada används selektiv granskning i form av en poängfunktion, DIFF, (Latouche och Berthelot 1992) för att välja ut objekt för verifiering, medan de övriga objekten tas om hand av det generella helautomatiska granskningsprogrammet GEIS. Lawrence och Mc Kenzie (2000) ger en utförlig beskrivning av selektiv granskning (i artikeln kallad signifikant granskning) med en annan typ av poängfunktion än funktionen DIFF.

En översikt av metoder för selektiv granskning med hjälp av poängfunktioner ges i Farwell och Raine (2000).

4.5 Hur idén med poängfunktioner kan implementeras

En realisering av idén med poängfunktioner kräver metoder för beräkning av

- lokal poäng
- global poäng
- kritiskt värde.

Se Engström (1995) för en beskrivning av ett experiment som utförts vid SCB.

4.5.1 Lokal poäng

Utgångspunkten för den lokala poängen är att den ska vara högre ju mer det potentiella felet i det flaggade värdet påverkar skattningen på lägsta nivå. Men vi vet ju inte om det flaggade värdet är felaktigt, hur stort det eventuella felet är eller effekten på statistikuppgiften för aggregatet. Beräkningen måste således bygga på antaganden och approximationer. Valet av metod beror därmed på vad som förefaller mest realistiskt för den egna undersökningen. Några idéer som tillämpas som uppskattning av det potentiella felet är:

- absoluta differensen mellan det ogranskade värdet och motsvarande värde vid föregående undersökning (se DIFF nedan)
- absoluta avståndet till det beräknade medelvärdet av förändringskvoten (medelvärdet från föregående undersökning multiplicerat med en antagen utvecklingstrend)
- absoluta avståndet till närmaste acceptansgränslinje.

Som approximation för skattningen används i DIFF och andra metoder motsvarande statistikuppgift vid föregående undersökning.

Den lokala poängen kan också modifieras genom multiplikation med en faktor som avspeglar variabelns relativa betydelse (DIFF), genom vägning med medelfelet (Engström 1995), genom division med någon aggregatberoende parameter för att öka sannolikheten att aggregat med få objekt verifieras m.m.

4.5.2 Global poäng

Global poäng kan bestämmas exempelvis genom:

- summering av de lokala poängen, som i DIFF
- högsta lokala poängen, som i Lawrence och McDavitt (1994)
- någon form av genomsnittspoäng (ingen tillämpning rapporterad).

4.5.3 Kritiska värdet

Kritiska värdet kan bestämmas med hjälp av granskade och ogranskade data från en undersökningsomgång där selektiv granskning inte tillämpats eller med tekniker som beskrivs i kapitel 8 eller Latouche och Berthelot (1992).

Det kritiska värdet ska ständigt övervakas. Ett medel för detta är de indikatorer som föreslås i kapitel 6.

4.6 Erfarenheter

Den metod som vi rekommenderar här beskrivs i detalj i bilaga kapitel 4. Det är poängfunktionen DIFF som används av Statistics Canada. Vid utvecklingsarbetet

med poängfunktioner fann man att DIFF var det mest intressanta av de alternativ som prövades. Observera att det är ett flexibelt instrument, som kan modifieras och förbättras på många sätt för att passa den egna undersökningens speciella problem och krav. Anpassningen kan göras på samma sätt som vid bestämning av kritiska värdet (se ovan).

4.7 Referenser

Anderson, K. (1989), ”Enhancing Clerical Cost-Effectiveness in the Average Weekly Earnings”, Belconnen, Australia: Australian Bureau of Statistics, Statistical Services Branch, November 1989.

Engström, P. and C. Ängsved (1994), ”A Description of a Graphical Macro Editing Application.” Conference of European Statisticians, Work Session on Statistical Data Editing, working paper No. 35, Cork 1994.

Engström, P. (1995), A study on using selective editing in the Swedish survey on wages and employment in industry. Conference of European Statisticians, Work Session on Statistical Data Editing, room paper No. 11, Athens 1995.

Farewell, K. and Raine M. (2001), “Some Current Approaches to Editing in the ABS”, *ICES II – Proceedings of the Second International Conference on Establishment Surveys, Invited Papers.* American Statistical Association 2001, pp 529–538.

Granquist L. 1990, ”A Review of Some Macro-Editing Methods for Rationalizing the Editing Process”, *Proceedings of Statistics Canada Symposium 90, Measurement and Improvement of Data Quality*, October 1990.

Granquist L. (1994), ”Macro-editing—A Review of Some Methods for Rationalizing the Editing of Survey Data” in *Conference of European Statisticians, Statistical Standards and Studies-No. 44*, 111–126.

Höglund, E. (1994), “ Macroediting in Statistics—The Hidiroglou-Berthelot Method (Statistical Edits) ” in *Conference of European Statisticians, Statistical Standards and Studies-No. 44*, 127–136.

Latouche, M., and Berthelot. J.-M. (1992), ”Use of a score function to prioritize and limit recontacts in business surveys”, *Journal of official Statistics*, Vol. 8, No. 3, 1992, pp. 389–400.

Lawrence, D., and McDavitt, C. (1994), ”Significance Editing in the Australian Survey of Average Weekly Earnings,” *Journal of Official Statistics*, **10**, pp. 437–447.

Lawrence, D., and McKenzie, R. (2000), “The General Application of Significance Editing,” *Journal of Official Statistics*, **16**, pp. 243–253.

Linacre, S. J., and D. J. Trewin (1989), "Evaluation of Errors and Appropriate Resource Allocation in Economic Collections," *Proceedings of the US Bureau of the Census Fifth Annual Research Conference*, U. S. Department of Commerse, Bureau of the Census, March 19–22, 1989, pp. 197–209.

van de Pol, F. (1995), Selective editing in the Netherlands Annual Construction Survey. Conference of European Statisticians, Work Session on Statistical Data Editing, working paper No. 26, Athens 1995

Wickberg I. (1990), ”Makrogranskning med Hidiroglou-Berthelots metod och med aggregatmetoden utvärderad mot produktionsgranskningen av månadsstatistiken över sysselsättning och löner under november 1987 för industriarbetare”, GRANSK-PM nr 20, 1990–06–25.

Bilaga kapitel 4
Hidiroglou-Berthelots metod (HB-metoden)

Antag att variabeln $X_i(t)$ ska kontrolleras, där i är objekt och t är period. Vanlig testvariabel är förändringskvoten:

$$R_i = \frac{X_i(t)}{X_i(t-1)} \qquad (1)$$

HB-metoden innebär att testvariabeln ovan transformeras samt att acceptansintervallet för denna transformerade testvariabel beräknas utifrån de data som ska granskas. Intervallet har egenskaperna att det är robust mot felaktiga data och okänsligt mot den naturliga variationen i förändringskvoter (se exempel 4.4). Dessa egenskaper åstadkoms genom att robusta parametrar, som median och kvartilavstånd, används i acceptansgränsintervallen och genom att testvariabeln transformeras två gånger. Första transformationen syftar till att få intervallet approximativt symmetriskt för att man ska bli kvitt en parameter. Den andra transformationen syftar till att göra acceptansintervallet känsligare för numeriskt stora värden på variabeln, vilket sker genom att den symmetritransformerade förändringskvoten multipliceras med ett storleksmått.

Symmetritransformationen

Låt

$$S_i = \begin{cases} 1 - \frac{R_{MEDIAN}}{R_i}, & 0 < R_i < R_{MEDIAN} \\ \frac{R_i}{R_{MEDIAN}} - 1, & R_i \geq R_{MEDIAN} \end{cases} \qquad (2)$$

där R_{MEDIAN} är medianvärdet för R_i

Storlekstransformationen

$$E_i = S_i * \left(MAX\left(X_i(t-1), X_i(t)\right)\right)^U, \quad 0 \leq U \leq 1 \qquad (3)$$

Acceptansgränserna

Acceptansgränser bildas på följande sätt:

Låt

$$D_{Q1} = MAX\left(E_{MEDIAN} - E_{Q1}, \left|A * E_{MEDIAN}\right|\right)$$

$$D_{Q3} = MAX\left(E_{Q3} - E_{MEDIAN}, \left|A * E_{MEDIAN}\right|\right)$$

där index Q1 och Q3 står för undre och övre kvartilen, och A är en konstant som normalt sätts till 0,05.

Syftet med $A * E_{MEDIAN}$ är att undvika svårigheter när kvartilavstånden är mycket små. Det vill säga när E_i är klustrade kring ett värde, vilket innebär att små avvikelser kan bli klassificerade som outliers.

Undre gräns:

$$E_{MEDIAN} - C * D_{Q1} \qquad (4)$$

Övre gräns:

$$E_{MEDIAN} + C * D_{Q3} \qquad (5)$$

Parametervärdena C och U kan bestämmas utifrån tester på insamlade data. Här kan det vara mycket lämpligt att använda samma typ av diagram som används på sidan 120 i Granquist (1994) för att fastställa bra värden på konstanterna C och U ovan. En interaktiv grafisk applikation vore lämplig, i vilken man direkt skulle se hur acceptansgränserna ändras för olika värden på konstanterna.

Poängfunktionen DIFF

Poängfunktionen DIFF har utvecklats av Latouche och Berthelot (1992). Metoden används av Statistics Canada m.fl.

Låt

$y_{i,k,t}$ vara värdet för objekt i (i = 1, … ,I) och variabel k (k = 1, … ,K) vid tidpunkt t

$y'_{i,k,t-1}$ vara värdet för objekt i (i = 1, … ,I) och variabel k (k = 1, … ,K) vid tidpunkt t-1

$w_{i,t}$ vara vägningstalet för objekt i (i = 1, … ,I) vid tidpunkt t

P_k vara relativ betydelse av variabel k

$$Z_{i,k,t} = \begin{cases} 1 & \text{om variabel k felsignaliseras av en eller flera kontroller} \\ 0 & \text{annars} \end{cases}$$

$\hat{Y}_{d,k,t-1}$ vara en skattning av totalen från redovisningsgrupp d (där undersökningsenhet i ingår) för variabel k, vid tiden t-1.

Poängfunktionen DIFF ges då av;

$$DIFF_{i,,t} = \sum_{k=1}^{K} \frac{w_{i,t} \left| y_{i,k,t} - y'_{i,k,t-1} \right| Z_{i,k,t} P_k}{\hat{Y}_{d,k,t-1}} \qquad (6)$$

Om $DIFF_{i,,t}$ är större eller lika med något kritiskt värde, sker verifiering.

Poängfunktionen är framför allt avsedd för kontroller mot misstänkta fel i kontinuerliga variabler. Men man kan även ta med uppenbara fel, förutsatt att man har ett bra och heltäckande system för automatimputeringar.

5 Grafisk granskning

I detta kapitel ska vi klargöra vad grafisk granskning är samt redovisa för- och nackdelar, tillämpningsområden och begränsningar. Interaktiv grafisk granskning exemplifieras med både input- och outputgranskning.

5.1 Bakgrund

Människor har lättare för att läsa och analysera bilder, än siffror och bokstäver. Därför har man länge funderat på om det går att använda grafik vid granskning. John Tukey lanserade i början på 1970-talet EDA, Exploratory Data Analysis. Hans ansatser byggde på manuella rutiner, där man relativt snabbt konstruerade grafer. Den tekniska utvecklingen under de senaste decennierna har visat att idéerna om grafisk granskning av data i statistiska undersökningar mycket väl kunde realiseras.

I slutet av 1980-talet startades de första projekten för att utveckla grafiska granskningssystem – och några år senare projekt för att ta fram generella programvaror för EDA. Många EDA-metoder är dels utmärkta granskningsmetoder, dels hjälpmedel när det gäller att utveckla metoder och instrument för att ta fram effektiva acceptansgränser för traditionella granskningsprogram.

Ett brännande problem för SCB de senaste åren har varit missöden med kvarvarande fel i data som förorsakat fel i publicerad statistik. Dessa brister i kvalitetssäkringen kan dock snabbt och enkelt undanröjas med hjälp av grafisk granskning – något som följande fiktiva exempel illustrerar.

Exempel på en situation

Antag att vi har en löneundersökning där variablerna total lönesumma och antal arbetade timmar är två av många variabler. Variablerna kontrolleras med kontrollen lön/timme. Den genomsnittliga timlönen är 100 kr med standardavvikelsen 20 kr.

I sista minuten före slutbearbetning kommer en blankett in. Man beslutar att den ska granskas manuellt, därför att en ytterligare granskningskörning tar för lång tid, är för dyr, eller medför annat oundvikligt men i det här sammanhanget onödigt arbete.

Vid den manuella granskningen upptäcker man inte att timlönen för företaget felaktigt har angivits till 1200 kronor. Företaget är stort eller har stort uppräkningstal, vilket medför att felet uppmärksammas i pressen. Punkten 1200 hade legat skyhögt över alla andra i ett punktdiagram med lönesumma och arbetstimmar som y-respektive x-axel eller hade utgjort en extrempunkt i ett lådagram av variabeln löner/timmar.

På någon eller några minuter hade man kunnat hitta felet med hjälp av grafik och därmed undvikit fadäsen!

5.2 Vad är interaktiv grafisk granskning?

Med interaktiv grafisk granskning menar vi inte bara samarbete mellan datoranvändare och dator utan framför allt ”kommunikation” mellan bilder och mellan bild och datablad.

Praktiskt betyder interaktiv grafisk granskning att:

- visa alla objekt som punkter eller staplar i diagram efter en eller flera variabler
- markera misstänkta objekt, varvid identitet och variabelvärden visas på skärmen för verifiering
- avläsa effekterna av ändringar och avsluta granskningen när effekterna ser ut att sakna betydelse för skattningar.

Interaktiv grafisk granskning är en kraftfull metod för att hitta avvikande värden, men även för att hitta egendomliga mönster. Det finns programvaror som är utmärkta för detta. De är enkla att implementera och använda. Metoden ger också god överblick av datamaterialet och bidrar dessutom till ökad ämneskunskap.

En granskning kan initieras på några få minuter, och diagram kan sedan tas fram på en eller annan sekund – vilket är ovärderligt när man ska granska enstaka objekt som kommer in i sista minuten.

Med den ständigt ökande datorkraften i våra persondatorer och programvaror som SAS-Insight kommer grafisk interaktiv granskning att bli ett allt starkare alternativ till granskning med inprogrammerade kontroller även för stora undersökningar. För måttligt stora undersökningar torde grafisk granskning redan nu vara betydligt effektivare än traditionell granskning (se avsnittet 5.6).

Standardiserat uttag av grafer
Grafisk granskning används ofta flexibelt (ad-hoc). I en produktionssituation kan detta vara mindre önskvärt, i synnerhet om granskningen utförs av flera personer. Granskningen blir beroende av den enskilda personens förmåga att se till att granskningen blir fullständig och är också emot principen vid SCB att minska variationen i angreppssätt. I stort löses problemet med ett program som dels startar SAS/Insight, dels producerar på förhand bestämda grafer.

5.3 Användning

Interaktiv grafisk granskning används i produktion i:

- produktionsgranskning
- Nya Zeelands ekonomiska undersökningar med ”gred”
- outputgranskning (se exemplet i avsnitt 5.6)
- SCB:s kortperiodiska statistik lönestatistik (KSP) sedan 1995
- SCB:s kortperiodiska lönestatistik (KLP) sedan 1996
- USA:s löne- och sysselsättningsstatistik med ”Aries” sedan 1993.

Gred är ett generellt grafiskt granskningssystem, medan systemen för outputgranskning är skräddarsydda applikationer.

System liknande SCB-applikationerna (se avsnitt 5.6), som fått mycket stor uppmärksamhet, kan utvecklas på några veckor. Aries har ökat produktiviteten med 100 procent i USA:s federala löpande löne- och sysselsättningsstatistik genom att datalistor ersatts med grafer. Produktivitetsökningen kan uttryckas med följande travestering av ett gammalt ordspråk:

”En grafisk bild säger mer än hundra datalistor.”

5.4 Introduktion till grafisk granskning

Grafisk granskning klarar merparten av vad traditionell granskning gör. Dessutom kan man med hjälp av t.ex. plottningar hitta mönster som avslöjar skevheter i materialet. Sådana är svårare att upptäcka med traditionell granskning.

Den grafiska granskningen bygger på bilder som plottningar och stapeldiagram. Graferna beskriver en eller flera variabler i datasetet. Om flera grafer visas samtidigt, blir granskningen flerdimensionell. Bilderna samvarierar på så sätt att om man markerar ett område i en graf, blir de samtidigt markerade i övriga grafer och i databladet. Detta ger en god bild av hur variabler är relaterade till varandra. När man studerar graferna, kan man med fördel utnyttja verktygen färger och symboler för att markera olika grupper av objekt som ska jämföras.

Verktyg

- Överblicka datamaterialet med punktdiagrammatris.
- Koda om partiellt bortfall (*missing value*) till ett avvikande numeriskt värde.
- Transformera om variabeler med sned fördelning, t.ex. med logaritm eller kvadratrot.
- Prova olika typer av diagramtyper:
  - histogram/stapeldiagram (eng.:*histogram/bar chart*)
  - lådagram (eng.: *box plot*)
  - punktdiagram (eng.: *scatter plot*)
  - tredimensionellt punktdiagram (eng.: *rotation plot*)
  - linjediagram (eng.: *line plot*).
- Använd färger och olika symboler för att visa t.ex. delgrupper.
- Kombinera diagramtyper i samma skärmbild.
- Utnyttja interaktionen mellan flera olika diagram och datablad.
- Gruppera materialet.
- Jämför mot likartat datamaterial, t.ex. från föregående undersökning.
- Inkludera uppräkningstal i analysen vid urvalsundersökning.
- Dela vid behov upp stora datamaterial.

Idéer, tips

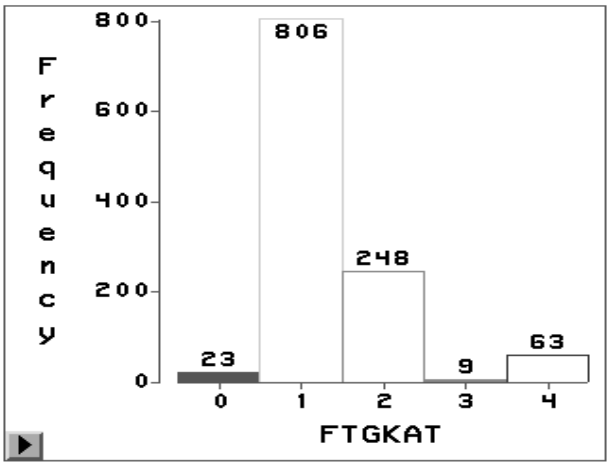
- För att finna uppenbara fel – använd stapeldiagram.
- För att bestämma acceptansgränser – använd lådagram.
- För att hitta avvikelsefel – använd punktdiagram eller lådagram.
- För att hitta inliers – använd punktdiagram.
- För outputgranskning – använd lådagram.

5.5 Exempel

Med hjälp av några exempel med tonvikt på feltyperna uppenbara fel, avvikelsefel och definitionsfel (se kapitel 2) illustreras grafisk granskning. Dessutom visas ett exempel på hur grafisk granskning kan användas som ett komplement till traditionell granskning när man ska välja acceptansgränser. I anknytning till exemplen ges kortfattade beskrivningar av de två grundläggande graferna, punktdiagram och lådagram.

5.5.1 Uppenbara fel

Diagram 1
Illustration av uppenbara fel. Giltiga värden för kod är 1–4



Med stapeldiagram kan man avslöja felaktiga koder eller om en kod saknas. I figur 1 är det 23 objekt som saknar kod för företagskategori (FTGKAT=0). När stapeln med den ogiltiga koden markeras, markeras även posterna i databladet, varefter man kan ta ut dessa för vidare behandling

Om man dubbelklickar på stapeln, visas ett fönster med de 23 posterna med samtliga variabelvärden.

Genom att komplettera stapeldiagrammet (diagram 1) med ytterligare grafer, som histogram, lådagram och punktdiagram, kan man identifiera de felkodade objekten. Om man markerar objekten med en avvikande färg, syns de tydligare i diagram och datablad.

En kontroll av icke giltiga värden kan på detta sätt ofta göras snabbt och enkelt för ett antal variabler samtidigt.

5.5.2 Misstänkta fel – avvikelsefel

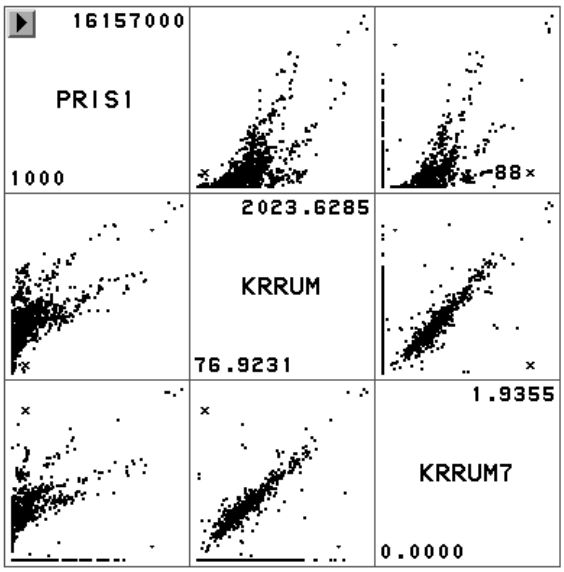
Punktdiagrammatris

Ett *punktdiagram* illustrerar samband mellan två eller tre variabler. En *punktdiagrammatris* (diagram 2) är flera tvådimensionella punktdiagram i en och samma bild och består av parvisa kombinationer av variabler. Punktdiagrammatrisen är ett effektivt verktyg när man vill studera sambandet mellan flera variabler (se SAS/ Insight kapitel 5, *Exploring Data in Two Dimensions*, och kapitel 35, *Scatter Plots*).

I diagonalen, där variabelnamnen står, kan man avläsa maximi- och minimivärden. Punktdiagrammatrisen är symmetrisk: diagrammen till vänster om diagonalen med variabelnamnen är en spegelbild av diagrammen till höger om diagonalen. Det är därför tillräckligt att studera diagrammen till vänster eller höger om diagonalen.

Data i detta exempel är hämtade från en region i inkvarteringsstatistiken, där bland annat uppgifter om intäkter och antal belagda rum samlas in för ett antal hotellanläggningar i olika regioner. I en punktdiagrammatris, diagram 2, studeras sambandet mellan variablerna intäkt, intäkt per belagt rum och intäkt per belagt rum föregående år.

Diagram 2
Inkvarteringsstatistiken. Månadsuppgifter



Variabler:

PRIS1
logiintäkt

KRRUM
intäkt per belagt rum

KRRUM7
intäkt per belagt rum föregående år

När man markerar en punkt i en av rutorna i diagrammet, visas identiteten för objektet (här företaget) samtidigt som man kan se var observationerna finns i de övriga rutorna. Ett dubbelklick på en punkt ger samtliga uppgifter för företaget. I exemplet har vi markerat en punkt i rutan uppe till höger. Objektet visar sig ha identiteten ’88’. Den har markerats med ”×” och får då automatiskt samma markering i övriga diagram och även i databladet.

Lådagram

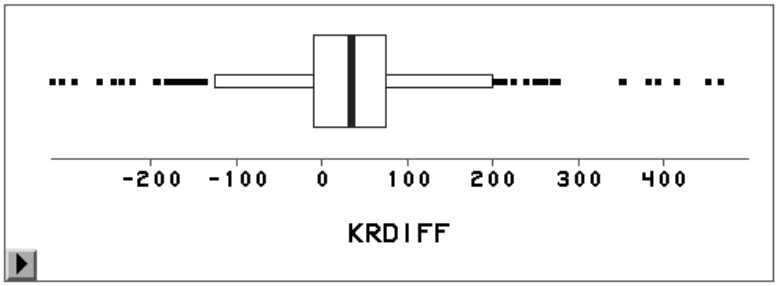
Ett *lådagram* används för att bland annat illustrera spridningen i ett material (se Wallgren m.fl. Statistikens bilder, kapitlet Att visa variation). Diagram 3 och 4 är lådagram. Lådagrammet kan indelas i beståndsdelarna låda, morrhår och extremvärden. Lådan begränsas av den undre och övre kvartilen. Följaktligen ligger 50 procent av fördelningen i lådan. Strecket inne i lådan är medianen.

Morrhåren är strecken på vardera sidan om lådan. Morrhårens längd är en valfri konstant (k) multiplicerat med kvartilavvikelsen (se SAS/Insight, kapitel 33 Box Plots and Mosaic Plots, avsnitt Method). Standardvärdet för konstanten k är 1,5. Om inte det passar, kan man välja ett annat k-värde. Det är det egna valet, baserat på erfarenhet, som styr!

Punkterna till höger och vänster om morrhåren är extremvärden.

En jämförelse mot tidigare uppgifter kan man göra genom att ta differensen eller kvoten mellan nya och gamla värdet (KRDIFF). I ett lådagram (diagram 3) framträder objekt med stora skillnader tydligt. En bra strategi vid granskning är att börja från ytterkanterna och gå inåt mot morrhåren.

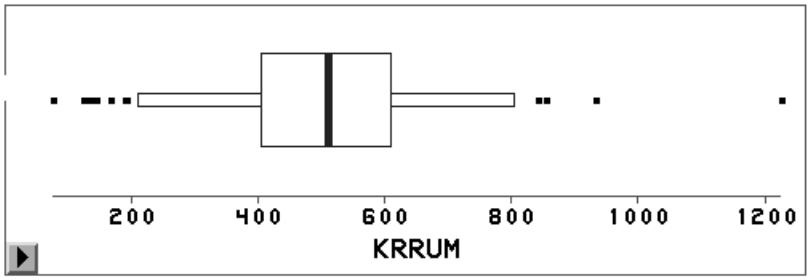
Diagram 3
Inkvarteringsstatistiken. Genomsnittlig skillnad i intäkt per rum jämfört med föregående år



Lådagram kan användas för att snabbt bestämma *acceptansgränser* om man avser att på traditionellt sätt ta fram misstänkta poster för verifiering.

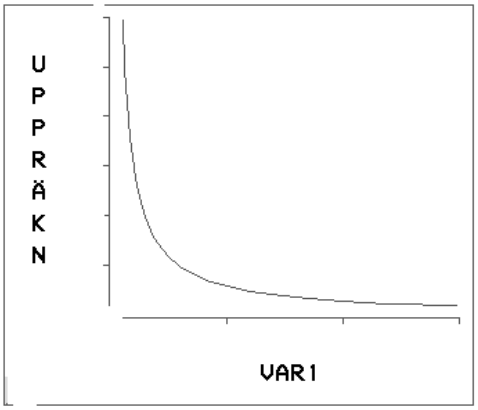
I den stora lådan ligger 50 procent av fördelningen, dvs. hälften av hotellen tog mellan 400 kr och 600 kr per natt. Den stora svärmen av observationer ligger mellan 200 kr och 800 kr. Sätter man acceptansgränserna 200 kr och 800 kr, får man ut endast de extrema värdena, vilket ofta räcker (se avvikelsefel).

Diagram 4
Inkvarteringsstatistiken. Fördelning för intäkt per belagda rum



I en **urvalsundersökning** kan man, för att ta hänsyn till uppräkningstal vid granskningen, även ta fram en graf med uppräkningstal som Y-variabel och undersökningsvariablerna som X-variabler.

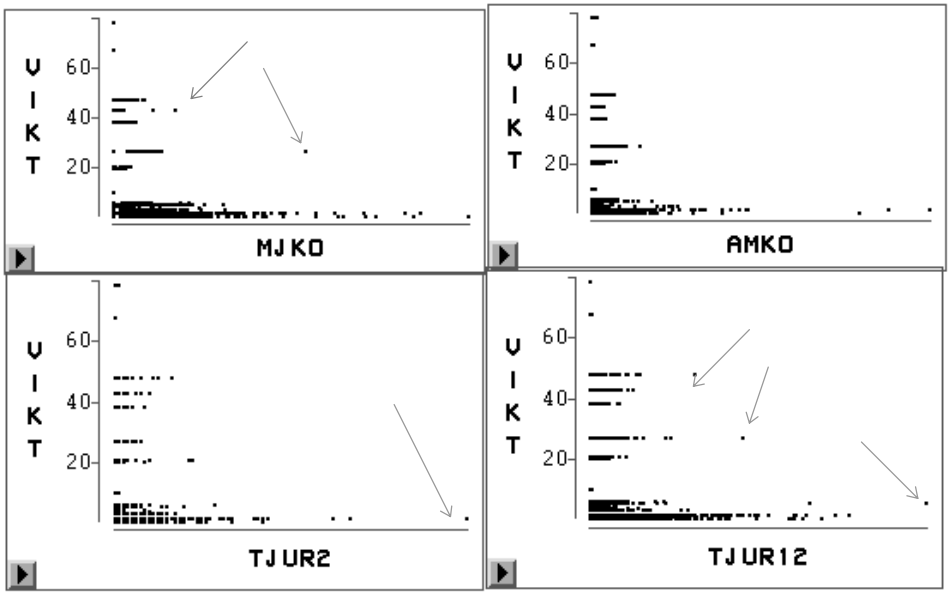
Diagram 5
Inflytande på skattning. Alla punkter på kurvan har samma inflytande



Produkten av variabelvärdet och uppräkningstalet bestämmer observationens tyngd i skattningen. I diagram 5 är produkten $w \cdot VAR1$, där w är uppräkningstalet, konstant på kurvan. Misstänkta värden ligger till höger om kurvan.

Misstänkta observationer hittar man genom att kombinera diagrammen 5 och 6. Tekniken exemplifieras på data från djurräkningen. Vi har i grafen manuellt ritat in pilar till de observationer som vi kan tänka oss att vi behöver undersöka närmare. Genom att man kan ta fram fler variabler samtidigt, blir denna process ofta ganska snabb. Variabeln vikt är uppräkningstal och övriga variabler är antalet djur av fyra olika djurslag.

Diagram 6
Djurundersökningen. Misstänkta fel



Som alternativ kan granskning i urvalsundersökningar göras direkt på uppräknade data, dvs. variabelvärden multiplicerade med uppräkningstal.

5.5.3 Misstänkta fel – definitionsfel

Definitionsfel är systematiska svarsfel (se kapitel 2), vilka praktiskt taget är omöjliga att hitta med traditionell granskning. Här kommer poängen med EDA enligt Tukey:

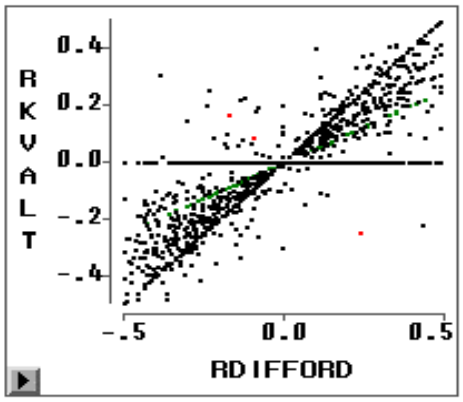
> ”The true value of a graph is found when it forces us to see something we could not see before.”

Idén är att finna mönster som indikerar förekomst av oförutsedda definitionsfel i ett datamaterial.

Exemplet från konsumentprisindex, KPI, nedan visar hur man med SAS/Insight kan identifiera felaktiga ”svarsbeteenden”.

Exempel: Svarsbeteende, KPI

Diagram 7
KPI. Kvalitetsvärdering vid byte av vara



Varje månad samlas prisuppgifter in för ett urval av varor i ett urval av butiker. Prisuppgifterna utgör en del av underlaget för beräkningen av KPI. Index beräknas ett år i taget genom att man jämför priser under aktuell månad med priser i december året innan. Från månad till månad försvinner en del produkter från butikernas sortiment och ersätts då med urval av nya produkter.

Intervjuarna, som sköter insamlingen, ska värdera skillnaden i kvalitet mellan den gamla och den nya produkten. Värderingen ska huvudsakligen göras utifrån intervjuarens egna bedömningar. Det är inte tillåtet att som regel sätta kvalitetsskillnaden lika med prisskillnaden, bara om det är motiverat.

I detta diagram plottas kvalitetsvärderingen (RKVALT) mot prisskillnaden (RDIFFORD), båda ställda i relation till prisnivå. Värdemängderna har begränsats till intervallet [-0,5;+0,5].

Här ser man tydligt att kvalitetsvärderingen 0 är vanlig (horisontell punktsamling), och det är tillåtet. Det är också vanligt att kvalitetsskillnaden är lika med prisskillnaden (diagonal linje $y = x$), vilket generellt är otillåtet. Man kan också se ett litet antal punkter på en linje $y = -x$, vilket antyder att kvalitetsvärderingen kan ha fått fel tecken, t.ex. vid dataregistreringen.

Det går också att skönja att det förekommer att kvalitetsvärderingen sätts lika med halva prisskillnaden (en diagonal linje $2 \cdot y = x$). Förekomsten av dessa mönster är intressant att konstatera, och de otillåtna beteendena hos intervjuarna bör åtgärdas med utbildning och information. Däremot är det inte meningsfullt att för de enskilda dataposterna försöka hitta ”rätt” värde, eftersom varje enskild datapost kan vara ”korrekt” samtidigt som massförekomsten tyder på felaktigt förfarande.

5.6 Grafisk aggregatgranskning

Aggregatgranskning innebär att man först kontrollerar aggregerade data (makronivå) och därefter identifierar de objekt i misstänkta aggregat som sannolikt bidragit mest till att aggregatet betecknats som misstänkt (mikronivå).

I arbetsställeundersökningen Kortperiodisk sysselsättningsstatistik (KS) används sedan 1995 ett system för grafisk outputgranskning. Systemet granskar för närvarande endast variabeln ”Antal anställda”. Ungefär samma typ av system används också sedan 1996 i Kortperiodisk lönestatistik inom privat sektor (KLP).

I KS system visar makronivån skattningar och relativ förändring av antal anställda per näringsgren och län mot föregående månad/kvartal, alternativt motsvarande månad/kvartal föregående år. Dessutom visas relativ förändring av populationsvariansen i syfte att identifiera aggregat där stora fel tar ut varandra så att aggregatvärdet verkar vara korrekt. Förändringar markeras med olika färger beroende på deras storlek.

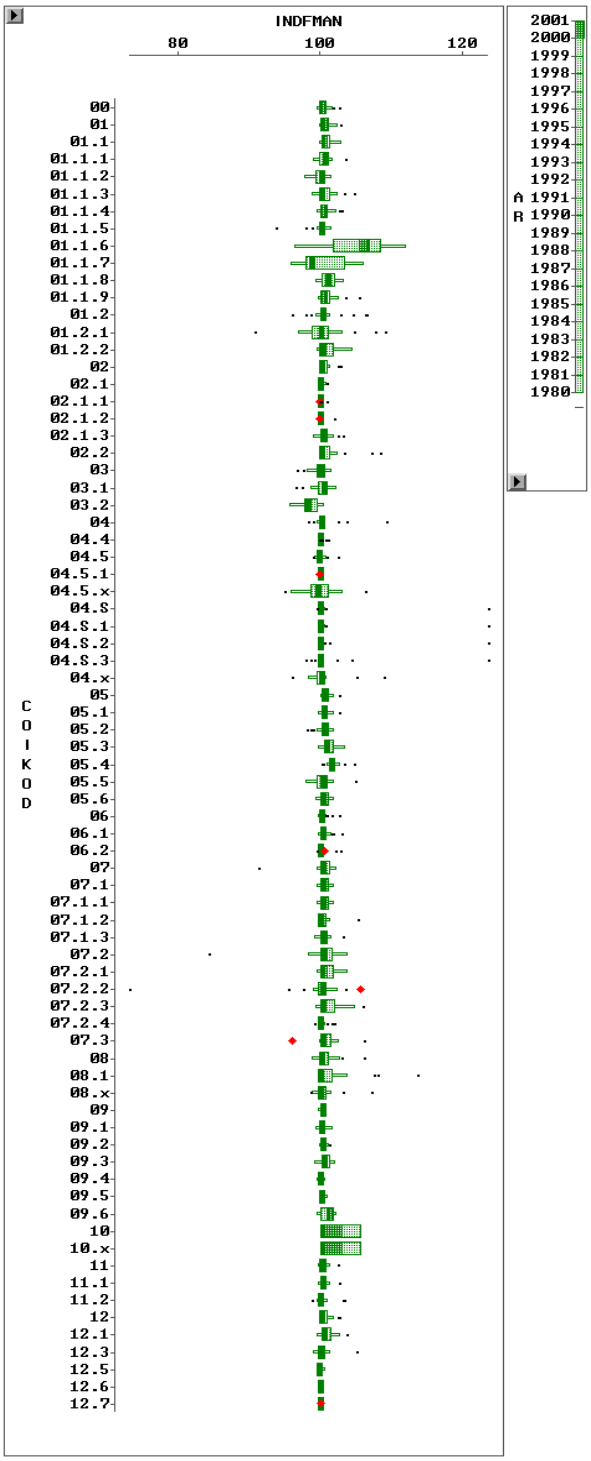
Några erfarenheter av systemet:

- Allmänt – misstänkta data upptäcks mycket snabbare än med traditionella papperslistor.
- Misstänkta skattningar på makronivån upptäcks snabbt genom färgmarkeringarna.
- Överblick fås över samtliga mikrodata för en misstänkt skattning i samma bild.
- Avvikande observationer med stor påverkan kan lätt identifieras och åtgärdas.
- Speciella mönster, som exempelvis stor variation eller speciella kluster (avslöjar eventuellt vissa systematiska fel), kan upptäckas med spridningsdiagrammet.

Ytterligare exempel på grafisk outputgranskning visas i exemplet nedan i diagram 8.

Exempel: Outputgranskning i KPI

Diagram 8
Outputgranskning av KPI. Förändringen, mellan januari och februari, på varugruppsnivå de senaste 21 åren



KPI beräknas genom en hierarkisk aggregering av priser upp till en total för hela den privata konsumtionen. På den högsta nivån används den internationella klassificeringen COICOP med 12 avdelningar.

Diagrammet här intill ger en snabbkoll av resultaten. Varje lådagram visar fördelningen av prisförändringarna från januari till februari för åren 1980–2000, dvs. 21 år.

Kod 00 avser KPI totalt. 01–12 avser avdelningar som Livsmedel, Alkohol & tobak.

Diagram 8 visar aggregatgranskning för februari månad år 2000. Om resultatet enligt lådagrammets definition är att betrakta som ”avvikande” markerar vi detta med ◆.

Här finns två tydliga fall. 07.2.2 avser drivmedel, och för februari 2000 har SCB uppmätt den största februariändringen under de senaste 21 åren (OPEC har begränsat leveranserna alltsedan vintern 1999 för att hålla uppe priserna på råolja). Koden 07.3 avser kollektiva transporter, och prissänkningen beror till stor del på ”rabatten” med 150 kronor som Stockholms Lokaltrafik lämnade som ersättning för problem med pendeltåg och tunnelbana under vintern år 2000.

Observera att acceptansområdet för 01.1.6, frukt, inte bör vara symmetriskt kring 100. Detta var utgångspunkten i den ursprungliga granskningen.

5.7 SAS/Insight: Hur stora dataset?

Grafisk granskning bygger på interaktivt samspel mellan grafer, datablad och enskilda observationer.

När datamängden ökar, blir punktdiagram gyttriga och svårlästa. Enskilda objekt kan man inte identifiera genom att klicka på en punkt, eftersom man då markerar en hel svärm av punkter. Det finns hjälpmedel för att hantera detta: minska punktstorleken och zoomning. Läsbarheten i histogram, lådagram med flera påverkas inte direkt av antalet observationer. Däremot bestämmer internminne och processorkapacitet hur snabbt det går att arbeta. Bildskärmen är ofta den praktiska begränsningen.

Det finns i princip tre faktorer som har betydelse för grafisk granskning med SAS/Insight:

- datamängden, antal poster och antal variabler
- dataformatet
- organisationen av data.

Det är omöjligt att ge generella rekommendationer för hur stora dataset får vara. Analys av 100 000 poster kan fungera bra i en tillämpning, medan det i en annan kan vara trögt att arbeta med mindre än halva antalet poster. Man måste pröva sig fram.

Vid större datamängder rekommenderar vi att man arbetar enligt någon av följande metoder:

- minska antalet variabler i analysen
- dela upp datamängden i undergrupper
- analysera ett urval av datamängden, t.ex. zoomning (leta efter mönster).

Överblicken över hela datamängden går förlorad, men man vinner i smidighet.

5.8 Kompetens

Man kan indela användarna av grafiska analysinstrument i tre olika kategorier

1. användare som granskar enbart enligt ett på förhand uppgjort schema
2. användare som granskar enligt plan, men även utför fördjupade analyser
3. avancerade användare

En förutsättning för den första gruppen är att kunna hantera analysverktyget, SAS/Insight. Detta kräver datorvana men inte nödvändigtvis kunskaper i SAS-programmering.

De som utför fördjupade analyser (grupp 2) behöver dessutom elementära kunskaper i EDA för att kunna analysera mönster, dvs. att identifiera fel med hjälp av mönster som är karaktäristiska för felen. Det fordras ibland vissa kunskaper i SAS-programmering.

Den avancerade användaren behöver goda kunskaper i EDA, SAS/Insight och SAS-programmering.

5.9 Sammanfattning av interaktiv grafisk granskning

5.9.1 Fördelar

Granskaren kan lätt se vilka datapunkter som kan vara avvikelsefel. Acceptansgränser behöver inte beräknas och uppdateras. Granskningen anpassas direkt till det aktuella materialet utan antaganden. Man får en god överblick över materialet och ökar kunskaperna om materialet och ämnet.

Metoden löser följande problem som traditionell granskning ofta har:

- I förväg satta parametervärden byggda på antaganden från föregående undersökning blir inaktuella och medför onödiga felsignaler samt snedvrider granskningen. (Det finns exempel där acceptansgränser inte har täckt medianvärdet.)
- Outliers i materialet kan medföra att gränserna snedvrids eller blir för vida när parametervärdena beräknas ur det material som ska granskas.
- Kontroller blir dåliga på grund av att antaganden om datastrukturer med mera inte håller till följd av förändringar i populationen.
- Införa nya kontroller i granskningsprogrammet (vilket kan vara tids- och resurskrävande).
- Objekt felsignaleras ständigt i varje undersökningsomgång.
- Komplicerade kontroller kan vara svåra att förstå.
- Svårt att få överblick av datamaterialet, ökad insikt i materialet och problemen med datainsamlingen, samt ökad ämneskunskap.
- Höga initiala IT-kostnader, programmering osv. (gäller speciellt engångsundersökningar).

5.9.2 Problem

Om man använder grafisk granskning som enda metod, kan man få följande problem – speciellt vid stora datamängder med många objekt eller variabler.

- Det kan vara svårt att få överblick över alla problem i ett objekt. En och samma uppgiftslämnare kan behöva kontaktas flera gånger. Dessutom kan det bli svårt att få granskningen fullständig för alla variabler.
- Det kan vara svårt att få fram processdata och dokumentation över granskningen.
- Förändringar i felbeteendet kan påverka granskningen (gäller all granskning).
- Det kan vara svårare att få granskningen reproducerbar.

5.9.3 När kan grafisk interaktiv granskning tillämpas?

Små datamängder	Passar bra för grafisk granskning, såväl produktionsgranskning som outputgranskning.
Medelstora datamängder	Traditionell granskning kompletteras med grafisk outputgranskning.
Stora datamängder	Traditionell granskning kan kompletteras med grafisk för delar av populationen.
När man inte har tid med flexibilitet (månadsstatistik)	Traditionell granskning eller alternativt en skräddarsydd grafisk applikation.
När flexibilitet är viktig	Grafisk granskning.
Granskningsomfattning liten	Grafisk granskning.
Granskningsomfattning större	Traditionell granskning eller skräddarsydd applikation för grafisk granskning.

5.9.4 Rekommendationer

Använd:

- interaktiv grafisk granskning i outputgranskning i alla undersökningar
- grafisk granskning i mindre undersökningar
- grafisk granskning i engångsundersökningar
- grafiska metoder som hjälpmedel för att utveckla granskningsprocesser.

5.10 Referenser

Houston G., and A. G. Bruce (1993), ”gred: Interactive Graphical Editing for Business Surveys”, *Journal of Official Statistics*. Vol. 9, No. 1, 1993, pp. *81–90.*

SAS/Insight, User’s Guide (1999), version 8. SAS Institute Inc. Manualen finns även elektroniskt tillgänglig via <Help> <Books and Training> <SAS OnlineDoc> <SAS/INSIGHT User's Guide>

Tukey, J.L., (1977) *Exploratory Data Analysis,* Addison-Wesley Publishing Company, 1977.

Wallgren, Anders, Wallgren, Britt, Persson, Rolf, Jorner, Ulf, Haaland, Jan-Aage (1996): Statistikens bilder – att skapa diagram, Statistiska centralbyrån

6 Vad ska göras vid felsignal?

Insamling av data om produktionsprocessen och sammanställning till processdata är ett område som man först nyligen har börjat utveckla – det gäller både internationellt och vid SCB. I avsnittet 6.6 kan vi följaktligen endast ange exempel på vilka data som kan samlas in. Endast en referens ligger till grund för vårt resonemang, nämligen Engström (1997). Inom de närmaste åren kommer dock säkerligen studier att redovisas som kan ligga till grund för riktlinjer och rekommendationer på området.

Förslaget till åtgärdskoder är ett utkast och ska alltså tills vidare ses som idéer. De innehåller dock allt som behövs för sammanställningen till de processdata som föreslås i kapitel 7.

I detta kapitel beskrivs aktiviteter som är förknippade med att granskningsprogrammet signalerar variabler inom objekt för **manuell** behandling, som här kallas *verifiering*. Granskningsprogram kan även signalera observationer/variabler för **maskinell** åtgärd, vilken omedelbart utförs av programmet. Detta kallas här *automatimputering* och berörs kortfattat i avsnittet om imputering.

Verifieringsprocessen har följande syften:

- kontrollera och åtgärda felsignalerade data
- registrera *granskningsstatus* och utfört arbete
- ge uppgiftslämnarna förståelse för vad de ska ge svar på
- samla in och registrera data över felorsaker, uppgiftslämnarproblem med mera.

6.1 Felmeddelanden

Med *felmeddelande* avses här den information som dataprogrammet ger om de objekt och variabler som flaggas som fel eller misstänkta fel. Syftet med felmeddelandet är att:

- ge tillräcklig information för rationell verifiering av felsignaler.
- utgöra ett underlag för data om insamlings- och produktionsprocessen.

6.1.1 Innehåll

Checklista för innehållet i ett felmeddelande

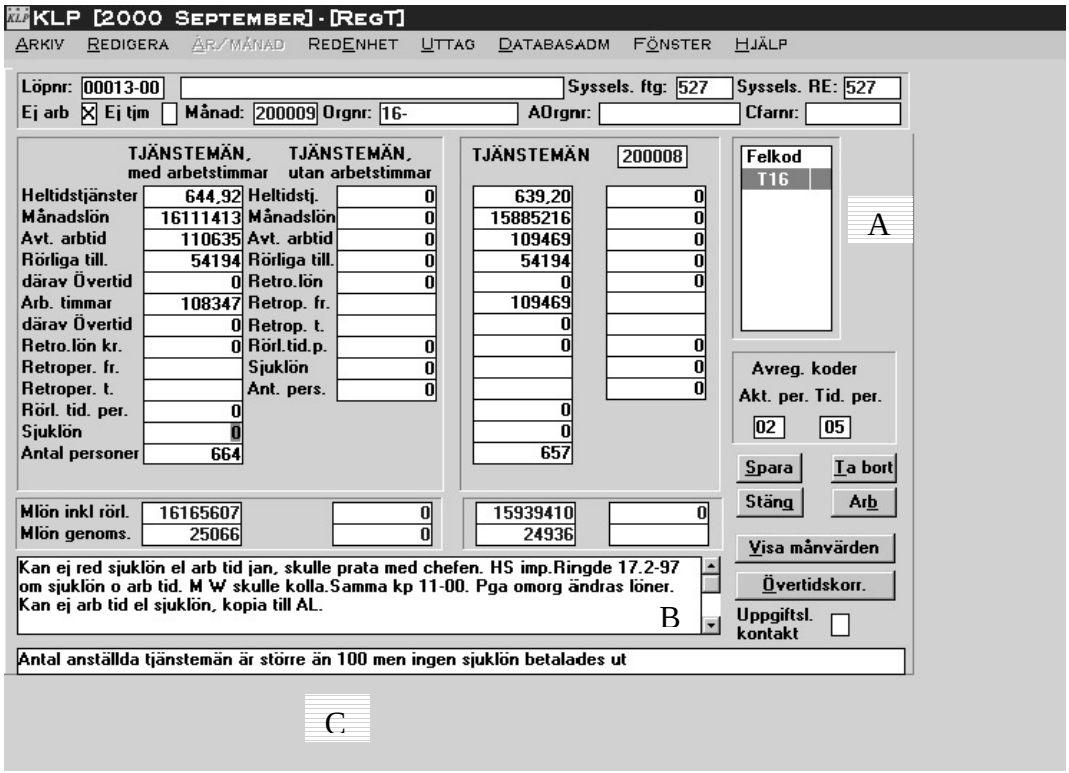
- ❑ objektidentitet
- ❑ felsignalerad variabel
- ❑ kontroll som medförde felsignalen
- ❑ verbal beskrivning av kontrollen
- ❑ data om objektet som bedöms nödvändiga för verifieringen.

Data som kan underlätta verifieringen är värden på vissa andra variabler, tidigare rapporterade värden på den felsignalerade variabeln och värden på härledda variabler, t.ex. förändringstal.

När en variabel felsignaleras måste en *felkod* sättas som visar vilket kriterienummer som förorsakade felsignaleringen.

Nedan följer ett exempel från Kortperiodisk lönestatistik, privat sektor (KLP), där dock allt det som vi rekommenderar inte finns med.

Diagram 1
Exempel på granskningsbild med felkod (A), problemet i klartext (C), felorsak och åtgärd i klartext (B). Kortperiodisk lönestatistik, privat sektor (KLP)



6.1.2 Generering av felmeddelande

Granskningen genomförs vanligen för ett objekt i taget, variabel för variabel eller kontroll för kontroll. Så fort variabelvärden underkänns i en kontroll, ska variablerna (fälten) förses med felsignal och felkoder. Felmeddelanden kan genereras på två sätt:

- omedelbart (variabelvisa felmeddelanden, vilket framför allt gäller dataregistreringsgranskning)
- efter att datorprogrammet har granskat hela objektet (objektvisa felmeddelanden).

Variabelvisa felmeddelanden. Granskaren registrerar nya värden på berörda variabler, anger orsaken till ändringen (se 6.6) och den maskinella granskningen startas om. Alla kontroller som innehåller de variabler som ändrats görs då om. Om det felsignalerade värdet accepteras, registrerar granskaren detta genom att ändra åtgärdskoden. Om felsignalen inte snabbt kan lösas, markerar operatören/granskaren värdet med en felkod och fortsätter granskningen.

Objektvisa meddelanden. När ett objekt gått igenom alla maskinella kontroller, skrivs det ut ett felmeddelande med alla felsignalerade variabelvärden och tillhörande koder (vid dataregistreringsgranskning: alla återstående variabelvärden och koder). Meddelandet ska i förekommande fall även omfatta uppgifter om utförda

imputeringar. Granskaren eller arbetsgruppen avgör om objektmeddelandet ska verifieras omedelbart eller först när ett antal meddelanden har samlats på hög.

Rutinen för felmeddelanden ska utformas så, att felsignalerade variabelvärden med tillhörande felkoder enkelt kan sammanställas för eventuell ytterligare bearbetning vid granskningen, som t.ex. automatimputering (se ovan).

Den produktansvarige ska utforma kontrollerna och bestämma ordningsföljden mellan dem samt svara för kodsättningen och den verbala beskrivningen med tillhörande information.

6.2 Verifiering av felsignaler

Det primära syftet med verifiering av felsignalerade variabler är att utreda om det felsignalerade variabelvärdet kan accepteras, är felaktigt eller är en outlier.

Om värdet är felaktigt, ska utredningen även leda till att man får fram ett acceptabelt värde.

Outliers ska egentligen behandlas i estimeringsprocessen, men diskuteras kortfattat nedan i avsnitt 6.4.

Verifieringen består i att få fram hur det ifrågasatta värdet ska åtgärdas – utifrån underlag (felmeddelande, blankett m.m.), ämneskunskaper, erfarenheter eller kunskaper om objektet eller genom kontakt med uppgiftslämnaren. Från processynpunkt är frågan här hur man ska organisera tillgänglig information om objektet, så att verifieringen kan göras så effektivt som möjligt.

6.2.1 Underlag

Underlag är först och främst felmeddelandet och registreringsunderlaget, vilket i postenkäter är den ifyllda blanketten. Registreringsfel, sortfel, förskjutningsfel, vissa konsistensfel m.fl. uppenbara (slarv-)fel kan lätt identifieras genom jämförelser mot registreringsunderlaget. I övrigt är det svaren på övriga frågor, registeruppgifter, tidigare rapporterade uppgifter samt tillskriven information (meddelanden från uppgiftslämnaren) som kan ge nödvändiga ledtrådar. Om sådana data inte finns på felmeddelandet, måste uppgifterna vara lättillgängliga, helst online – t.ex. genom en inskannad bild av den ifyllda blanketten.

Vid periodiska undersökningar kan tidigare rapporterade uppgifter för relevanta variabler finnas med på felmeddelandet. I annat fall ska detta underlag finnas lätt tillgängligt, helst online. Från granskningssynpunkt är det en fördel om uppgifter från föregående period är förtryckta på blanketten. Detta ger dessutom ett bra stöd för uppgiftslämnaren.

Vid verifiering ska man också utnyttja sådana data om det felsignalerade objektet som finns i tillgängliga register eller i andra undersökningar.

Om underlaget är otillräckligt, ska uppgiftslämnaren kontaktas där det är motiverat och möjligt. Återkontakterna har dessutom till uppgift att samla in data om förklaringar till felsignaler, orsaker till fel och uppgiftslämnarens problem att ge korrekta svar på undersökningens frågor. (Se vidare avsnitt 6.3.)

6.2.2 Kompetenskrav

Det är ofta svårt att verifiera felmeddelanden. Visserligen finns det i alla undersökningar felsignaler som lätt kan verifieras, men för många felsignaler krävs det omfattande kunskaper i ämnet och om undersökningen. Detta gäller särskilt när uppgiftslämnaren ska kontaktas. Det krävs fortlöpande utbildning i ämnet, erfarenhetsutbyte och en tillräckligt kvalificerad bakgrund.

6.3 Återkontakter

Syften med att kontakta uppgiftslämnaren är att:

- hjälpa uppgiftslämnaren att lämna korrekta svar inte bara vid undersökningstillfället utan även i framtiden
- verifiera felsignaler
- inhämta kunskaper om data, noggrannhet i data, felorsaker och uppgiftslämnarproblem.

Kontakter är kostsamma för både producent och uppgiftslämnare, varför varje beslut om återkontakt måste vara välgrundat. En fråga man måste ställa sig vid beslutet om återkontakt är om det är realistiskt att man kan få högre kvalitet.

Återkontakter måste ge betydligt mer än enbart att man verifierar objektets alla olösta felsignaler. Man ska på ett systematiskt sätt samla in felorsaker samt synpunkter på undersökningen – både när det gäller de efterfrågade uppgifterna (t.ex. uppgiftslämnarens möjlighet att lämna uppgifter) och i fråga om blankettens utformning. Åtgärder och felorsaker kodas på lämpligt sätt (se avsnitt 6.6).

Kontakter måste tas så nära uppgiftslämnandet som möjligt. Om lång tid har förflutit mellan uppgiftslämnandet och återkontakten, har uppgiftslämnaren troligen varken tid eller vilja att åter sätta sig in i såväl uppgiftslämnandet som sitt eget informationssystem. Dessutom kan det vara svårare att få tillgång till data.

Vid engångsundersökningar och vid de första omgångarna av en ny periodisk undersökning är det befogat med många återkontakter. Efter hand som uppgiftslämnaren lär sig undersökningen, blanketten förbättras och granskarnas erfarenheter och kunskap ökar, blir det alltmer de potentiella felens betydelse som ska avgöra frekvensen av återkontakter.

6.4 Hantering av outliers

O*utliers* är korrekta data som uppräknade skiljer sig så mycket i absolut storlek från övriga uppräknade värden i sin redovisningsgrupp, att de aktualiserar frågor om förändringar i estimationen eller kommentarer i redovisningen av undersökningens resultat. Detta gäller speciellt urvalsundersökningar.

Outliers behandlas vanligen med:

- reducering av deras inverkan för att undvika dels missvisande information om målpopulationen, dels störningar i jämförbarheten
- kommentarer i publicerade skattningar.

6.5 Imputering

Ett felsignalerat och felaktigt värde kan imputeras, dvs. ersättas med ett acceptabelt värde efter fastställda regler utan kontakt med uppgiftslämnaren. Imputering

kan göras antingen genom *manuell imputering* med hjälp av jämförelser mot värden från tidigare rapportering, information i urvalsramen (registret) eller mot statistikuppgifter i andra undersökningar. Saknade eller orimliga värden kan även ersättas med ett acceptabelt värde maskinellt enligt fastställda regler, s.k. *automatisk imputering*. Källorna som används för att hitta ersättningsvärden kan vara desamma för såväl manuell som automatisk imputering.

Automatisk imputering kan göras vid partiellt bortfall och eventuellt för bortfallsobjekt. Även i samband med verifieringen kan automatisk imputering användas när felsignalerade variabler ska ersättas med acceptabla värden (t.ex. vid orimliga värden). Huruvida en felsignalerad variabel ska imputeras automatiskt eller åtgärdas på annat sätt beror ofta på objektets/felets betydelse för den slutliga skattningen.

En fördel med automatisk imputering framför manuell imputering är att åtgärdade variabler behandlas på ett likformigt sätt. Risken vid manuell imputering kan vara att bedömningarna blir subjektiva. Dock finns det objekt och undersökningar där det inte är möjligt att använda automatisk imputering.

Imputering är ett stort område och tas inte vidare upp här. Det viktiga är att den imputering som görs följs upp och dokumenteras väl, t.ex. genom att man tillämpar åtgärdskoder som visar hur variabeln har behandlats. Detta beskrivs i följande avsnitt.

6.6 Insamling av information om granskningen

Vidtagna åtgärder, orsaker till fel m.m. ska insamlas systematiskt och dataregistreras som underlag till statistik över granskningsprocessen (se vidare kapitel 7) och för att undvika dubbelarbete.

Checklista för verifiering av felsignal

- ❑ hur felsignalerade data har åtgärdats
- ❑ vilka orsakerna är till identifierade fel
- ❑ huruvida svar utgör uppskattningar eller är hämtade från uppgiftslämnarens informationssystem
- ❑ vad avgivna svar egentligen står för (relevansen i erhållna svar)
- ❑ vilka problem uppgiftslämnaren har att besvara frågorna
- ❑ hur stor resursåtgången är, i synnerhet när det gäller återkontakter.

Hur felsignalerade data har åtgärdats, kodas lämpligen enligt ett kodsystem utarbetat av den produktansvarige.

En *åtgärdskod* som visar vilken/vilka åtgärder som har genomförts ska definieras för varje enskild felkod.

Exempel på åtgärder:

- Variabelvärde felsignalerat (men inte verifierat).
- Variabelvärde godkänt utan ändring och **utan kontakt** med uppgiftslämnaren.
- Variabelvärde godkänt utan ändring **efter kontakt** med uppgiftslämnaren.
- Åtgärdas senare (t.ex. ett viktigt objekt där uppgiftslämnaren inte har kunnat nås).

- Variabelvärde korrekt men har klassats som outlier och åtgärdas enligt bestämda regler (se 6.4)
- Variabelvärdet ändrat **utan** återkontakt (t.ex. överföringsfel).
- Variabelvärdet ändrat **med** kontakt (primärdata felaktigt) – (ett korrekt värde har givits av uppgiftslämnaren).
- Variabelvärdet ändrat vid återkontakt (primärdata felaktigt) – (ett uppskattat värde har givits av uppgiftslämnaren).
- Värde manuellt imputerat (primärdata felaktigt/saknas).
- Värde automatiskt imputerat (primärdata felaktigt/saknas).
- Variabelvärde saknas.
- Objektet tillhör övertäckningen.
- Objektet utgör bortfall.

Vid insamlingen av information om granskningsprocessen är det viktigt att skaffa information om orsaker till att fel uppstått. Informationen från återkontakt eller blankett används för kodning av *felorsaker* till identifierade fel.

Exempel på orsaker till fel:

- fel vid skanning eller dataregistrering
- missuppfattning (definitionsfel) hos uppgiftslämnaren (föranleder eventuellt åtgärder i blanketten)
- svårigheter för uppgiftslämnaren att lämna efterfrågad uppgift.

Det kan tyckas vara orimligt resurskrävande att felorsaks- och åtgärdskoda varje variabel eller att registrera tidsåtgången vid återkontakter. Men för att man ska känna till och kunna kontrollera processen, krävs kodning och framställning av processtatistik. Ofta är det dock inte möjligt att felorsakskoda alla felsignaler eller att göra det vid varje undersökningsomgång. Man kan då t.ex. välja ut vissa grupper av inkommande objekt, välja ut vissa variabler, göra kodningar endast för vissa undersökningsomgångar och så vidare. Det gäller att i början koncentrera sig på de variabler, delpopulationer osv. som är viktiga och som man känner till är mest problematiska. Hur det ska göras får man i början pröva sig fram till.

Resursåtgången, uttryckt i t.ex. tid, är en viktig faktor i planeringen och styrningen av en undersökning. Vissa aktiviteter kan kontrolleras med hjälp av åtgärdskoder och *indikatorer* (se vidare kapitel 7), t.ex. andelen ändringar, antalet återkontakter eller andelen manuellt imputerade variabelvärden. Andra insatser, såsom tiden som använts för återkontakter (inklusive tiden för försök till återkontakt, dvs. då uppgiftslämnaren inte har kunnat nås), kräver ett system för att resursåtgången ska kunna registreras.

6.7 Referens

Engström, P. (1997): A Small Study on Using Editing Process Data for Evaluation of the European Structure of Earnings Survey. Working paper No. 19, UN/ECE Work Session on Statistical data Eding, Prague.

7 Processdata

Checklista

- ❑ Producera processdata fortlöpande, dvs. automatiskt.
- ❑ Presentera processdata
  - ○ grafiskt
  - ○ enligt paretoprincip.
- ❑ Analysera processdata för att
  - ○ på ett tidigt stadium ge besked om något har inträffat som eventuellt är väsentligt för granskningen
  - ○ snabbt finna orsaken till det inträffade och se om det finns möjligheter att rätta till de eventuella problemen under pågående process
  - ○ ge information om kontrollers träffsäkerhet och effektivitet.

Som nämndes i inledningen av kapitel 6, är processdata ett utvecklingsområde som är i sitt inledningsskede såväl internationellt som vid SCB. Vid SCB har en förstudie gjorts inom det s.k. processdataprojektet (2000) som bland annat diskuterar processdata för granskning.

Ett ramarbete, Nordbotten (2000), utgör grund för detta kapitel. De förslag på indikatorer som anges har accepterats av den internationella gruppen ”Work Session on Statistical Data Editing” (leds av FN:s Economic Commission for Europe) som en plattform att arbeta vidare på. Martin (2000) framför förslag som är i enlighet med nedanstående.

Utan data om processen är det omöjligt att leva upp till principen: ständig förbättring av hela undersökningen. Syftena med att producera statistik över granskningsprocessen är mer konkret uttryckt att:

- övervaka och styra processen under pågående produktion
- få underlag till förbättring av undersökningen
- mäta effekter av processförändringar
- få underlag för analys och kvalitetsredovisningar.

Här redovisas de indikatorer som en granskningsprocess åtminstone måste generera för att syftena ovan ska tillgodoses.

Indikatorerna ska produceras fortlöpande, dvs. automatiskt under och efter varje produktionscykel. För att indikatorerna i praktiken ska komma till faktisk användning, måste de presenteras lättåtkomligt och attraktivt. Presentationen bör därför vara grafisk och organiserad enligt paretoprincipen.

Indikatorerna presenteras i två nivåer. I den översta, översiktliga indikatorer (avsnitt 7.1.1), pekas på eventuella problem i processen och visas också hur mycket arbete som utförs i olika delar av processen. I den andra nivån, avsnitt 7.1.2, presenteras indikatorer relaterade till variabler. De ska användas till att identifiera förekommande problem med variabler eller kontroller.

I avsnitt 7.2 ges en bild av hur processövervakningen enligt modellen skulle kunna gå till.

7.1 Indikatorer

Indikatorerna bör ange hur mycket granskningsarbete som har utförts, hur stor andel av kontrollerna som resulterat i ändringar av variabelvärden, hur många variabelvärden som ändrats osv. När man ska analysera effektiviteten i granskningsprocesser är det emellertid inte tillräckligt med indikatorer. Man behöver också data om storleken av ändringar, imputeringar, rättningar m.m. Den informationen samlas lämpligen in i samband med underlaget för indikatorframställningen. Hur sådana data kan användas anges i kapitel 8, Effekter av granskning.

7.1.1 Objektrelaterade indikatorer

I indikatorerna har täljaren och i vissa fall nämnaren ett egenintresse för den produktansvarige och ska naturligtvis skrivas ut explicit.

Andelen felsignalerade objekt ges av:

$$O1 = \frac{Antal\ \ felsignalerade\ objekt}{Antal\ objekt\ som\ genomlöpt\ granskningskontroller}$$

Andelen ändrade objekt ges av:

$$O2 = \frac{Antal\ ändrade\ objekt}{Antal\ objekt\ som\ genomlöpt\ granskningskontroller}$$

Andelen återkontaktade objekt ges av:

$$O3 = \frac{Antal\ återkontakter}{Antal\ objekt\ som\ genomlöpt\ granskningskontroller}$$

Andelen ändrade objekt efter återkontakter ges av:

$$O4 = \frac{Antal\ ändrade\ objekt\ efter\ återkontakt}{Antal\ återkontaktade\ objekt}$$

Ett litet värde på O4 jämfört med värdet för O1 antyder t.ex. att granskningsprocessen är ineffektiv.

Vid återkontakter kan man misslyckas med att få acceptabla data från uppgiftslämnaren. Om då imputeringar av saknade och felaktiga data har utförts, kan indikatorn *Andelen imputerade av återkontaktade objekt* belysa betydelsen av det problemet:

$$O5 = \frac{Antal\ återkontaktade\ objekt\ \ för\ vilka\ en\ variabel\ imputerats}{Antal\ återkontaktade\ objekt}$$

7.1.2 Variabelrelaterade indikatorer

De variabelrelaterade indikatorerna identifierar variabler eller kontroller som eventuellt orsakar problem i processen. Indikatorerna bör presenteras storleksordnade (t.ex. i paretodiagram), som i det fiktiva exempel som redovisas i 7.2. I annat fall drunknar lätt intressanta fall i undersökningar där det finns många variabler och kontroller.

Andelen felsignaler per variabel ges av:

$$V1 = \frac{Antal\ \ felsignalerade\ objekt\ \ för\ variabel\ X}{Antal\ objekt\ som\ genomlöpt\ kontroller\ \ för\ variabel\ X}$$

Andelen felsignaler per kontroll ges av:

$$V2 = \frac{Antal\ \ felsignalerade\ objekt\ av\ kontroll\ K\ \ för\ variabel\ X}{Antal\ objekt\ som\ genomlöpt\ kontroll\ K\ \ för\ variabel\ X}$$

Andelen ändringar per variabel ges av:

$$V3 = \frac{Antal\ ändringar\ av\ variabel\ X}{Antal\ objekt\ med\ värde\ på\ variabel\ X}$$

Träffsäkerheten per variabel ges av:

$$V4 = \frac{Antal\ objekt\ med\ ändring\ \ för\ variabel\ X}{Antal\ \ felsignalerade\ objekt\ \ för\ variabel\ X}$$

Man kommer mycket långt med dessa indikatorer. Om man emellertid vill gå vidare, är det mycket enkelt att bygga ut systemet till att omfatta de speciella företeelser i undersökningen som man vill få belysta. Felorsakskodar man vissa eller alla variabler, kan man t.ex. lägga till indikatorn

$$V5 = \frac{Antal\ \ felsignaleringar\ \ för\ variabel\ X\ \ av\ orsak\ O}{Antal\ \ felsignaleringar\ variabel\ X}$$

och om man imputerar när återkontakter inte medför att man får acceptabla data, kan man införa följande indikator:

$$V6 = \frac{Antal\ imputeringar\ av\ variabel\ X\ \ för\ återkontaktade\ objekt}{Antal\ återkontaktade\ objekt}$$

Metadataprojektets delprojekt om processdata kommer förhoppningsvis att resultera i någon form av standard för indikatorer.

7.2 Hur övervakning med hjälp av indikatorer skulle kunna gå till

En primär uppgift för indikatorer är att på ett tidigt stadium ge besked om att något har inträffat som eventuellt är väsentligt för granskningen. Det kan vara att acceptansgränser eller parametrar i kontroller fått helt felaktiga värden, t.ex. på grund av registreringsfel eller felaktiga antaganden i acceptansgränssättningen, eller också att stora reella förändringar i populationen eller omvärlden inträffat sedan föregående undersökning eller sedan den undersökning på vilken acceptansgränserna bygger. Tecken på detta kan vara att antalet felsignalerade objekt är väsentligt större eller mindre än vid föregående undersökningar.

En andra uppgift är att snabbt finna orsaken till det inträffade och se om det finns möjligheter att rätta till de eventuella problemen under pågående process.

En tredje uppgift är att ge information om kontrollers träffsäkerhet och effektivitet både för att man ska kunna se till att de är på tillräcklig hög nivå och för att man ska kunna bedöma var det bäst kan löna sig att sätta in åtgärder i såväl insamlings- som granskningsprocessen. Detta kan vara ett viktigt inslag i evalvering av granskningsprocessen, vilket diskuteras i kapitel 8.

Eftersom alla data för indikatorframställningen måste tas fram under processen, är det naturligt att oavsett syfte behandla indikatorerna i ett sammanhang. Vår rekommendation är att man under själva granskningsprocessen producerar en databas eller fil med alla data om vad som händer med originaldata i processen, dvs. felsignaler, åtgärdskoder osv., inklusive originaldata och färdiggranskade data. Det bör göras på ett sådant sätt att dessa data kan sammanställas till indikatorer och för de metoder för evalvering som i kapitel 8 betecknas som differensmetoder. Allt sådant behöver inte tas fram för alla objekt, variabler och kontroller för varje undersökning. Men dessa data ska finnas, så att man när som helst ska kunna göra effektivitetsstudier.

Nedan anger vi i form av ett fiktivt exempel hur övervakning med indikatorer skulle kunna gå till. Exemplet bygger på Engström (1996).

När andelen felsignalerade objekt ökar eller minskar för mycket eller när de felsignalerade objekten är för många eller för få, söker man hitta anledningen till den oväntade förändringen. Man försöker då lokalisera felsignalerade objekt efter någon viktig indelningsvariabel, undersökningsvariabel och kontroll. Andelen felsignalerade objekt sorteras i fallande ordning (paretoprincipen). Felsökningen illustreras med hjälp av diagram.

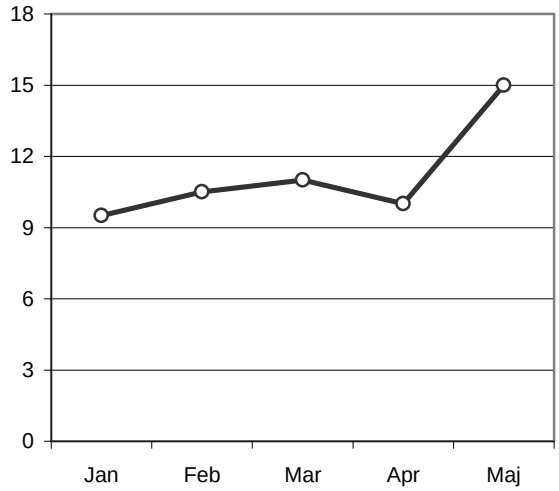
Fiktivt exempel på övervakningssystem

I en månatlig företagsundersökning felsignalerar granskningssystemet cirka 10 procent av de inkomna objekten.

Andelen felsignalerade objekt ges av indikator O1 och redovisas i ett styrdiagram som kan nås via undersökningens systemapplikation. Det bedöms att andelen felsignalerade objekt får variera med ± 2 procent.

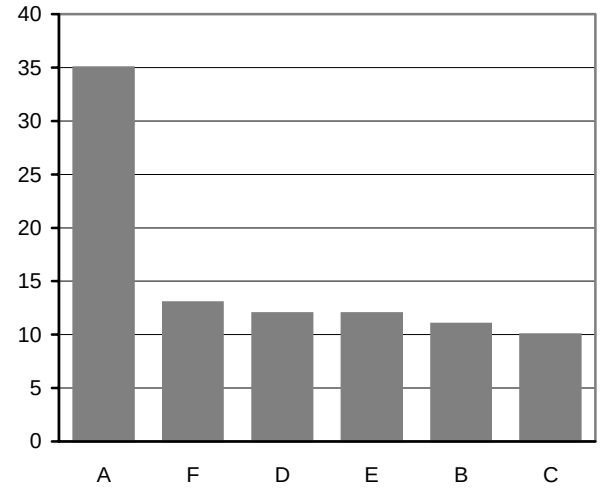
Enligt den beslutsregel som tillämpas undersöker man processen närmare när andelen felsignalerade objekt ligger utanför acceptansgränserna.

Diagram 1
Andel felsignalerade objekt för undersökningsperioderna januari–maj



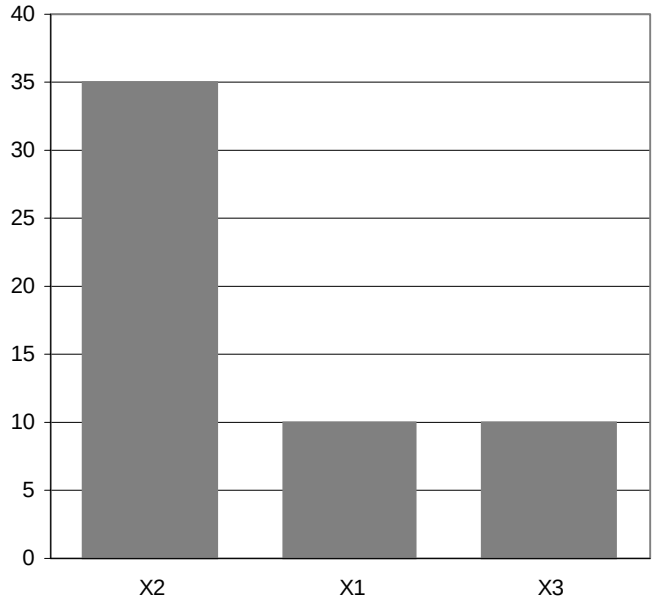
Eftersom andelen felsignalerade objekt för den aktuella undersökningsperioden bedöms vara för stort, fortsätter man med undersöka felsignalerade objekt med avseende på variabler och kontroller. Bransch är då en viktig indelningsvariabel.

Diagram 2
Andel felsignalerade objekt per bransch. Sortering i fallande ordning, paretodiagram



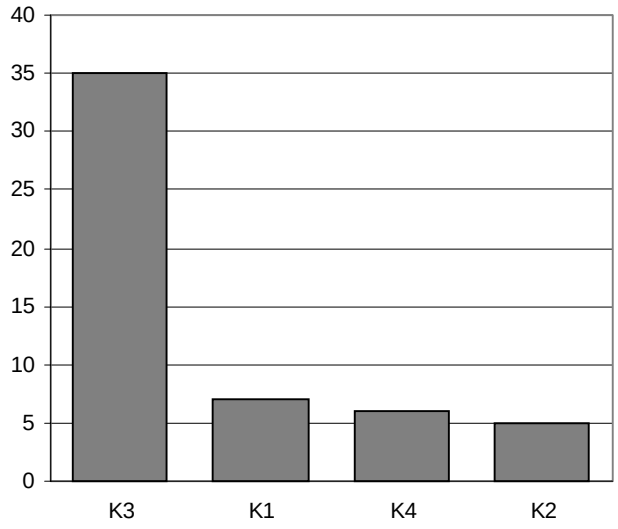
Nästa åtgärd är att undersöka andelen felsignalerade objekt per variabel, indikator V1.

Diagram 3
Andel felsignalerade objekt per variabel



Ofta används flera kontroller i granskningsprocessen för att identifiera fel. Nästa steg är därför att undersöka hur många felmarkeringar som de olika kontrollerna ger.

Diagram 4
Andel felsignaler per kontroll. Sortering i fallande skala efter andel felsignaler per kontroll

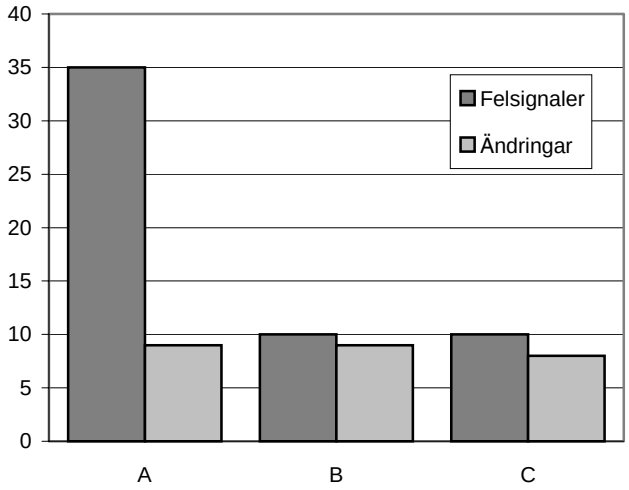


En väsentlig och återkommande uppgift är att utvärdera granskningsprocessen. Vi fortsätter därför med det fiktiva exemplet för att illustrera ett sätt att analysera processen.

Exempel på analys av granskningsprocessen med hjälp av processdata

När man ska **analysera** processen, är andelen ändringar av stor betydelse för kontrollernas träffsäkerhet. I diagram 5 har andelen felsignalerade objekt per variabel utökats med andelen ändrade objekt per variabel (indikator V4).

Diagram 5
Underlag för analys av processen. Andel felsignaler och ändringar per variabel. Sortering i fallande skala efter andel felsignaler per variabel



Träffsäkerheten är endast 25 procent för kontroll K3 (diagram 4) när det gäller variabel X2. Det är många felsignaler och låg träffsäkerhet, varför kontrollen (K3) bör effektiviseras. Skulle däremot träffsäkerheten vara hög, är det problem med variabeln.

7.3 Referenser

Engström, P. (1996): Monitoring the Editing Process. Working paper No. 9, UN/ECE Work Session on Statistical Data Editing, Voorburg.

Martin, C. (2000): Meta Data – An Aid to Managing the Edit and Imputation Process. Working Paper No. 6, UN/ECE Work Session on Statistical Data Editing, Cardiff.

Nordbotten, S. (2000): Evaluating Efficency of Statistical Data Editing: General Framework. United Nations, Geneva, 2000.

Processdataprojektet (2000): Processdata för processförbättring – Slutrapport från förstudie i Processdataprojektet. Slutrapport 2000–12–15.

Work Session on Statistical Data Editing: Länk till webbplats, http://www.unece.org/stats/

8 Mätning av effekter av granskning

I detta kapitel föreslår vi några enkla metoder som ger information om de ändringar som görs vid granskning. Metoderna bygger inte på processdata utan enbart på data före och efter den granskningsprocess som man vill studera. Det betyder att de alltid kan användas där man kan spara ogranskade data. Metoderna ger svar på frågor som:

- Hur ser fördelningen av ändringarna ut?
- Hur påverkar granskningen skattningarna?
- Vilken omfattning och betydelse har vissa feltyper?
- Är ändringarna koncentrerade till vissa undergrupper, t.ex. vissa branscher?

Svaren på sådana frågor kan användas när man ska bedöma om det kan finnas effektivare kontroller, när man vill målinrikta kontroller på vissa feltyper, förbättra frågeformuleringar och anvisningar generellt eller för vissa undergrupper i populationen osv.

En förutsättning är att man i en datafil har granskade och ogranskade värden för alla variabler man vill undersöka.

I avsnitt 8.1 och 8.2 behandlas brutto- och nettoeffekterna av ändringen på skattningen. När det gäller få och stora avvikelser, vill man inte dra slutsatsen att de tar ut varandra i nästa undersökning bara för att de råkade göra det denna gång. Därför analyserar man absolutbeloppen av ändringarna (brutto). Men för de många små ändringar som delvis tar ut varandra kan man anta att det kommer att inträffa även i följande undersökningar. Därför analyserar man ändringarna med tecken (netto).

D*ifferensstudiemetoden* har använts i flera undersökningar på SCB, t.ex. finansstatistiken (Wahlström 1990 och Forsman 1991 b), industristatistiken (Hedlin 1992), lönestatistiken över kommunalt anställda (Lindell 1995 a) och lönestatistiken över landstingsanställda (Lindell 1995 b). Ursprungligen utvecklades metoden vid Bureau of the Census och modifierades av Forsman (1991 b).

En mer arbetskrävande metod är *feltypsmetoden*, som utvecklades av Forsman (1991 a) på data från dåvarande finansstatistiken. Den förutsätter att man i granskningsprocessen noterar ändringar i den fysiska blanketten, varefter ämnesexperter klassificerar ändringarna på ett antal feltyper.

Vi rekommenderar att man genomför analysen med en *grafisk ansats* – gärna med hjälp av programvaran SAS/Insight. Anledningen är att analyserna av differenserna mellan granskat och ogranskat då kan göras betydligt mer omfattande och djupgående, samtidigt som arbetet underlättas.

Rena evalverings- eller utvärderingsmetoder faller utanför ramen för denna CBM-rapport. För sådana metoder hänvisar vi till Granquist (1997). Den rapporten är en översikt av ett antal metoder för evalvering av granskningsprocesser, beskrivna i litteraturen fram till 1996. Metoderna är resurskrävande och i många fall sofistikerade. Några innebär regelrätta evalveringar av kvaliteten i processen och i obser-

vationsregister. Internationellt utvecklar man år 2002 nya metoder för att beräkna effekter av granskning (*Work session on statistical data-editing*).

Data från industristatistiken 1990 har använts för att exemplifiera metoderna i detta kapitel. Industristatistiken 1990 var en totalundersökning (*cut off*) som omfattar många variabler. Det totala antalet objekt var 5 540.

När metoderna tillämpas på urvalsundersökningar, måste effekten av granskningen mätas på uppräknade tal.

8.1 Differensmetoder

Det finns många exempel på att nya metoder eller utvidgning av acceptansområdena har kunnat rationalisera granskningen väsentligt med bibehållen kvalitet på undersökningen (Granquist, Kovar 1997). Det beror på att den tidigare metoden har haft låg *träffsäkerhet* (se kapitel 4), men även på att de ursprungliga metoderna har resulterat i många *små ändringar med obetydliga effekter på skattningarna*. Vad som i detta sammanhang ska betraktas som obetydligt beror på precisionskraven och på effekterna av övriga felkällor. Var gränsen går mellan betydelsefulla och betydelselösa fel är en bedömningsfråga. En riktlinje kan dock vara att gränsen sätts så, att effekten av ändringarna under gränsen utgör högst en tiondel av urvalsfelet för estimaten på någorlunda detaljerad nivå. Syftet med differensmetoderna är att undersöka om processen kan effektiviseras.

Vi ger exempel på två metoder där man utnyttjar skillnaden mellan granskat och ogranskat värde. I det ena fallet undersöker man differensernas andel av den totala differensen. I det andra fallet jämförs ändringarna med skattningen.

8.1.1 Differensernas andelar av den totala differensen

Genom denna beräkning får man bl.a. information om ett fåtal stora ändringar svarar mot en mycket hög andel av den totala ändringen. För detaljer hänvisar vi till Wahlström (1990).

Gör så här:

Välj en eller flera variabler som ska ingå i studien.

- Beräkna för varje objekt absolutbeloppet av differensen mellan det granskade och det ogranskade värdet.
- Rangordna differenserna i fallande storleksordning. För urvalsundersökningar multiplicerar man absolutbeloppen med inverterade inklusionssannolikheten innan man sorterar efter storlek.
- Beräkna ett mått på hur stor andel av summan av (de absoluta) differenserna som kan hänföras till de **j** största differenserna. Beräkna måttet för varje **j**.
- Rita ett diagram, dvs. plotta måttet enligt punkt 4 för varje **j** såsom i diagram 1 nedan.

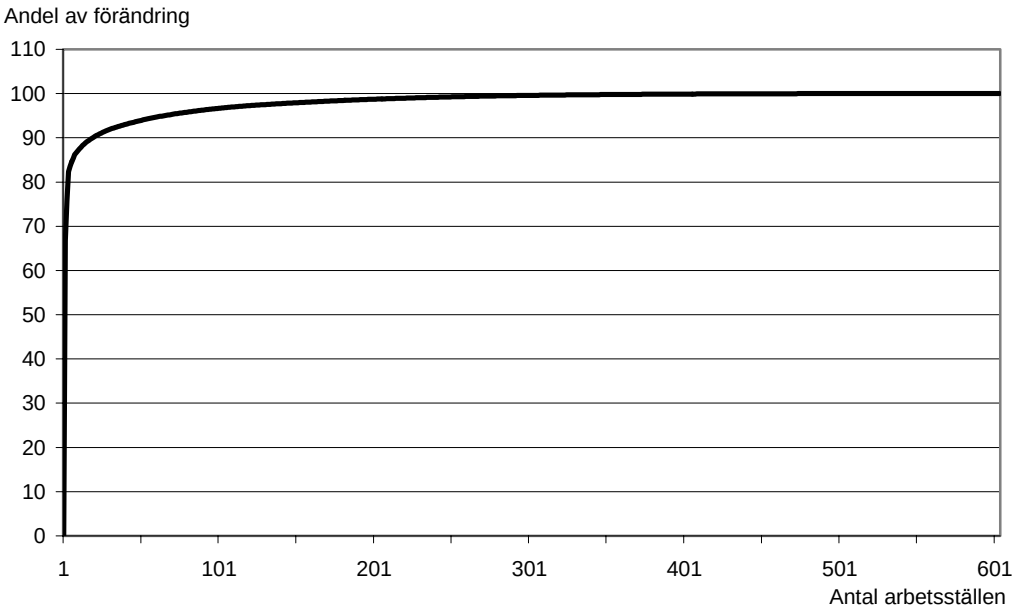
Exempel: Industristatistikens granskningsprocess

Studien av industristatistikens granskningsprocess visar bland annat att, för variabeln industriproduktion, 3 procent av antalet differenser svarade för ca 90 procent av den totala värdemässiga differensen.

Tablå 1
Data som används i diagram 1. Tablån är sorterad i fallande skala efter absolutvärdet av avvikelsen mellan granskat och ogranskat värde. (Andel av förändringen = 0 är ett tekniskt värde för att figuren ska bli snyggare.) Här har vi bara tagit med de nio största avvikelserna.

ogranskad	granskad	granskad-ogranskad	|granskad-ogranskad|	kum(|granskad-ogranskad|)	Andel av förändring	Antal arbetsställen
					0	
31797622	92084	-31705538	31705538	31705538	66.79716314	1
4234634	4234	-4230400	4230400	35935938	75.70976128	2
3162410	17339	-3145071	3145071	39081009	82.33579048	3
601194	0	-601194	601194	39682203	83.60238478	4
613236	156444	-456792	456792	40138995	84.56475324	5
556574	156574	-400000	400000	40538995	85.40747242	6
403330	30847	-372483	372483	40911478	86.19221885	7
241151	24151	-217000	217000	41128478	86.64939401	8
0	208769	208769	208769	41337247	87.08922812	9

Diagram 1
Industristatistiken 1990. Variabeln industriproduktion



Kurvans form beror på att det finns några enstaka mycket stora förändringar.

De utan jämförelse största differenserna i studien av finansstatistiken (Forsman 1991a) berodde på de så kallade 1000-felen, dvs. fel som orsakades av att observationerna angivits i fel enhet, t.ex. i kronor i stället för i tusentals kronor.

Gör alltid studier på lägre nivåer än totalnivån. På så sätt kan man dessutom få veta om granskningsarbetet är koncentrerat till vissa redovisningsgrupper, t.ex. branscher. Om det förekommer många felaktiga svar från företag i vissa branscher, kan det bero på att dessa företag har speciella svårigheter att lämna de efterfrågade uppgifterna. I så fall har metoden avslöjat en felkälla, och då är det blanketter och anvisningar som man ska förbättra.

8.1.2 Differensernas successiva inverkan på skattningen

Genom denna beräkning får man reda på om många differenser tillsammans är obetydliga jämfört med skattningen. För detaljer hänvisas till Forsman (1991 b).

Gör så här:

- Välj en eller flera variabler som ska ingå i studien.
- Beräkna för varje objekt differensen mellan det granskade och det ogranskade värdet.
- Rangordna differenserna i fallande skala (efter absolutbeloppet av differensen). Multiplicera differenserna med inverterade inklusionssannolikheten när det gäller urvalsundersökningar.
- Beräkna ett mått på hur nära man ligger den publicerade skattningen om man endast tar hänsyn till de **j** största differenserna. Beräkna måttet, med hänsyn tagen till differensens tecken för varje **j**.
- Rita ett diagram, dvs. plotta måttet enligt punkt 4 för varje **j**.

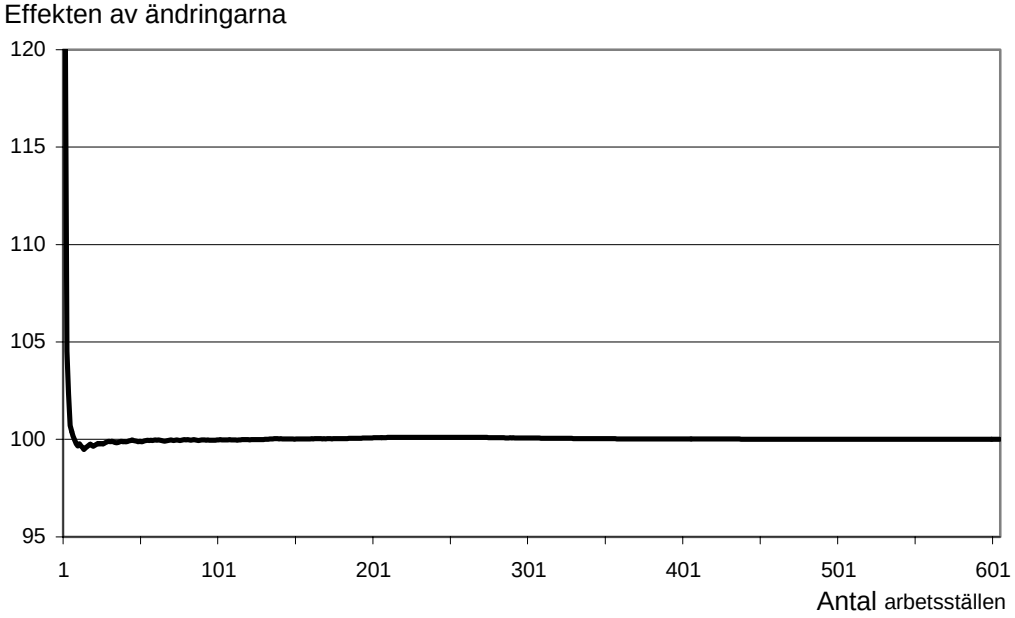
Exempel: industristatistikens granskningsprocess

Studien av industristatistikens granskningsprocess (Hedlin 1992) visar att, för variabeln industriproduktion, de 95 procent minsta differenserna sammanlagt motsvarade mindre än 1 procent av skattningen.

Tablå 2
Data som används i diagram 2. Tablån är sorterad i fallande skala efter absolutvärdet av avvikelsen mellan granskat och ogranskat värde. Här har vi bara tagit med de nio största avvikelserna. (Se även tablå 1)

ogranskad	granskad	granskad-ogranskad	|granskad-ogranskad|	Effekten av ändringarna	Antal arbetsställen
				120.870366	
31 797 622	92 084	-31 705 538	31 705 538	104.516199	1
4 234 634	4 234	-4 230 400	4 230 400	102.334098	2
3 162 410	17 339	-3 145 071	3 145 071	100.711826	3
601 194	0	-601 194	601 194	100.401722	4
613 236	156 444	-456 792	456 792	100.166102	5
556 574	156 574	-400 000	400 000	99.959776	6
403 330	30 847	-372 483	372 483	99.767644	7
241 151	24 151	-217 000	217 000	99.655713	8
0	208 769	208 769	208 769	99.763399	9

Diagram 2
Industriproduktionen, 1990, variabeln industriproduktion



Figuren visar inverkan av enskilda förändringar på totalskattningen.

Skattningen beräknas dels med hjälp av det granskade materialet, $\hat{Y}_g$, dels med hjälp av $\hat{Y}_g'$.

Där $\hat{Y}_g'$ är den skattning man får när man successivt, efter differensernas storlek, ersätter ogranskat med granskade uppgifter.

Kvoten $100 * \frac{\hat{Y}_g'}{\hat{Y}_g}$ uttrycker förändringen mellan skattningarna i procent. Förändringarna är sorterade i fallande ordning och plottade mot antal arbetsställen (x-axeln).

8.2 Feltypsmetoden

I en feltypsstudie studeras granskningens effekter, uppdelade på olika typer av fel i det ogranskade materialet. Feltyperna klassificeras innan analysen genomförs (se Forsman 1991 a).

Studien syftar till att sortera ut de feltyper som är mest intressanta från kostnads- och kvalitetssynpunkt. Om man känner till omfattningen av olika typer av fel i det ogranskade materialet, har man ett utmärkt underlag för att konstruera effektiva kontroller för identifiering av dessa fel (målinriktning av kontroller). Men man ska framför allt ändra blanketten, anvisningarna osv. för att minska risken för att sådana fel uppstår (eliminera felkällan).

Här beskrivs metoden vid urvalsbaserad klassificering av åtgärderna vid granskning.

Gör så här:

- Välj en eller flera variabler som ska ingå i studien.
- Klassificera för varje variabel de fel som upptäcks vid granskningen. Det finns inget generellt sätt att göra klassificeringen på, utan den måste för varje undersökning bygga på ämneskunskaper. Man bör dock inte samtidigt ta med mer än högst fyra till sex feltyper.
- Dra ett urval av objekt som ska undersökas. Urvalet bör vara så stort att åtminstone 100 objekt som ändrats kan förväntas komma med. Om man t.ex. av erfarenhet vet att ca 25 procent av objekten ändras för variabeln i samband med granskningen, bör urvalsstorleken vara minst 400.
- Klassificera felen i samband med granskningen.
- Beräkna antalet fel av respektive feltyp.
- Skatta felens effekter på totalskattningen.

8.3 Differensstudier med SAS/Insight

Differensstudier kan med fördel genomföras med hjälp av SAS/Insight. Den här analysen blir mer flexibel än de som redovisas i 8.1 ovan, genom att man bland annat simultant kan följa flera variabler. Förutsättningen är att data är organiserade i en SAS-tabell (se bilagan).

Fördelningen av differensen mellan granskat och ogranskat värde kan åskådliggöras med hjälp av lådagram eller stapeldiagram (histogram). Effekten på skattningen beräknas med hjälp av formel 3 i bilagan till detta kapitel. Om det finns enstaka mycket stora avvikelser mellan granskat och ogranskat värde, kan dock lådagrammet bli otydligt. I detta fall fungerar histogram bättre. Genom att transformera skalan av ändringarna, t.ex. 10-logaritmering, kan man få en tydligare bild av både storleksordningen och effekten av ändringarna fördelade på olika intervall.

Man kan välja mellan två metoder som båda bygger på analys av skillnaden mellan granskat och ogranskat värde.

I den första metoden betraktas hela fördelningen för differenserna med hjälp av ett histogram. Genom att ändra staplarnas bredd kan man avläsa effekten av ändringarna i olika intervall.

Idén med den andra metoden går ut på att skapa successiva restmängder. Man tar successivt bort effekterna av ändringarna i fallande storlek för att se om och i så fall när vi kommer till en punkt då ändringarna saknar betydelse i förhållande till skattningen.

Det är enkelt att genomföra en analys med hjälp av SAS/Insight. I bilagan beskrivs i detalj hur man gör.

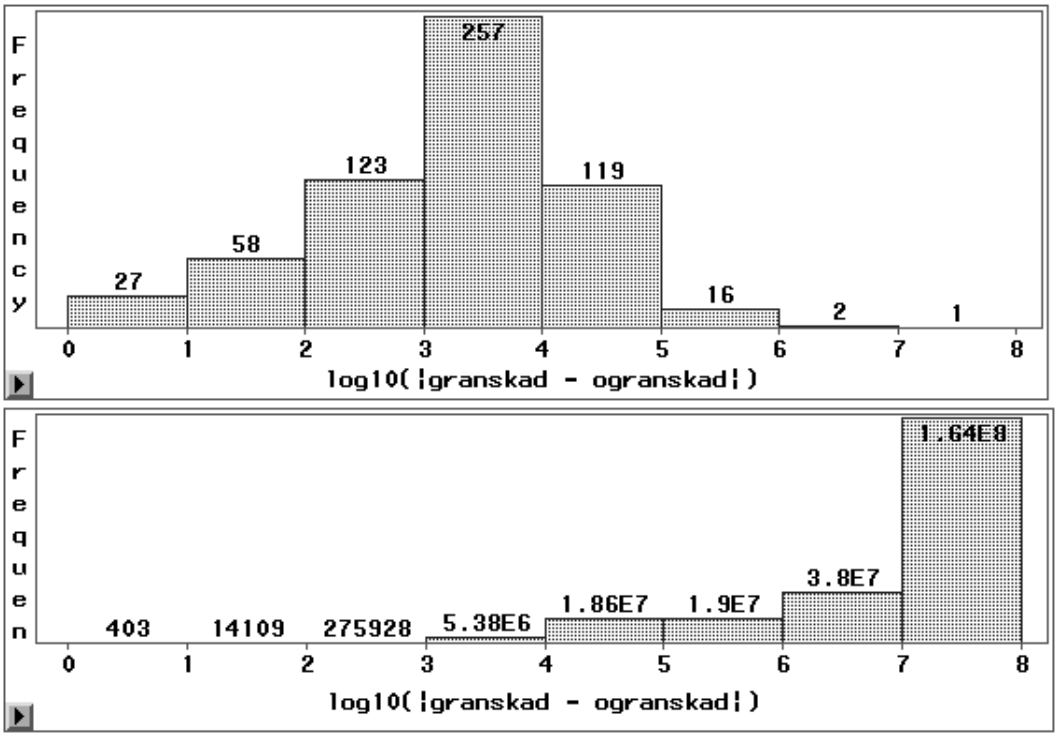
Nedanstående exempel från industristatistiken 1990 illustrerar tekniken. För en van användare tar analysen cirka en timme att genomföra.

Exempel: Industristatistikens granskningsprocess

> Vi studerar skillnaden mellan granskad och ogranskad uppgift för omsättningen. Rita ett histogram för fördelningen med avseende på såväl antal observationer som effekten på skattningen (se bilagan formel 3).

Skapa absolutbeloppet av skillnaden mellan granskat och ogranskat värde. Dölj observationer där ingen förändring gjorts (granskad = ogranskad). Bilda ett histogram för variabeln **|granskat – ogranskat|** med avseende på såväl antal observationer som effekten på skattningen.

Diagram 3
Industristatistiken 1990. Skillnad mellan granskad och ogranskad uppgift på variabeln omsättning. ^{10}log-skala, dvs. 1 motsvarar 10, 2 100, 3 1000 etc



De numeriskt tre största ändringarna svarar mot $(3.8*10^7 + 1.64*10^8)*10^{-9} \approx 20.2$ procent av skattningen.

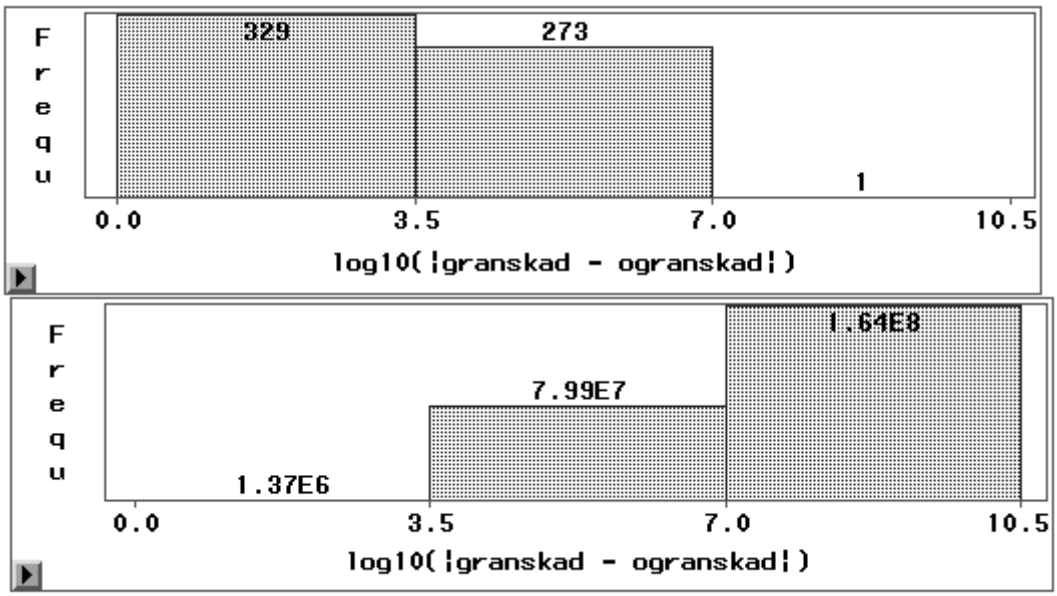
Med hjälp av de två histogrammen kan vi dels beräkna antalet objekt i olika intervall, dels uppskatta effekten av granskningen i förhållande till skattningen. Kom ihåg att i det undre histogrammet dividera med skalfaktorn, dvs. här 10^9, för att få rätt storleksordning på beräkningen.

Metod 1
Vi utgår från diagram 3 (de logaritmerade absoluta differenserna) och ändrar staplarnas bredd. Syftet är att studera ändringarnas effekt på skattningen (undre diagrammet). Vi är speciellt intresserade av de många små ändringarna och dess effekt på skattningen.

Antag att vi i denna undersökning kan ignorera rättningar på en sammanlagd effekt av maximalt en promille av skattningen, som i det här fallet visar sig vara vid stapelbredden 3.5 (diagram 4). Man måste alltså prova sig fram.

I en urvalsundersökning kan en nettoeffekt (bias) negligeras som är mindre än en tiondel av standardavvikelsen i skattningen (Särndal, Swensson, Wretman, 1992).

Diagram 4
Industristatistiken, 1990. Effekten av rättningar

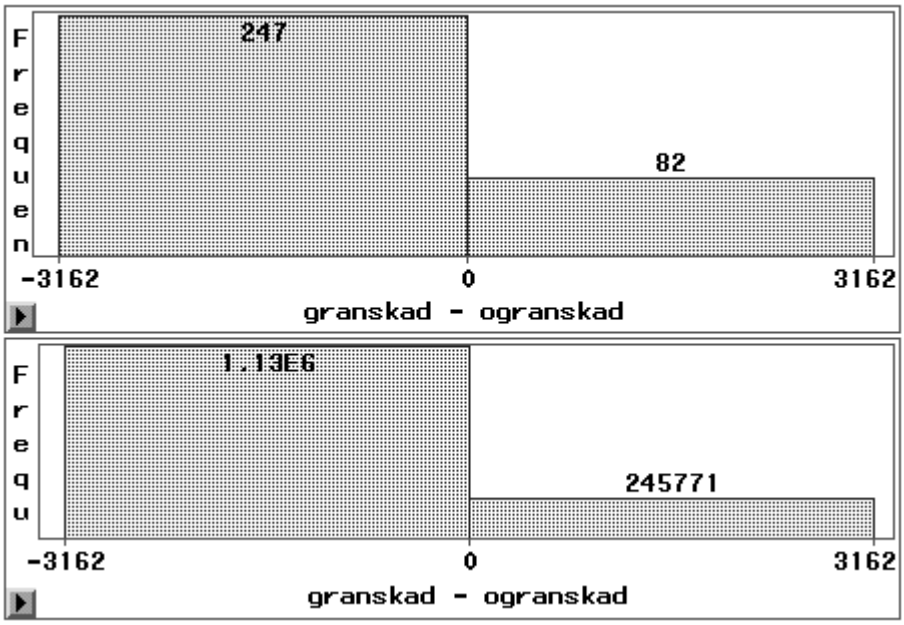


Enligt diagram 4 finns 329 ändringar som är numeriskt mindre än $10^{3.5} \approx 3\,162$. Deras sammanlagda bruttoeffekt på skattningen är $1.37*10^{6}*10^{-9} = 1.37$ promille.

Vid statistikredovisning är det nettofelet som är intressant. Många små fel tenderar att ta ut varandra, varvid nettofelet blir betydligt mindre än bruttofelet.

När vi skall beräkna nettoeffekten av ändringarna återgår vi till den ursprungliga undersökningsvariabeln **granskad – ogranskad**. Markera stapeln med de 329 observationerna i diagram 4 och välj ut dessa till en ny SAS-tabell med hjälp av <extract> i databladets dialogruta. Bilda ett nytt histogram för denna datamängd (diagram 5). Justera skalan så att de båda staplarna blir lika breda.

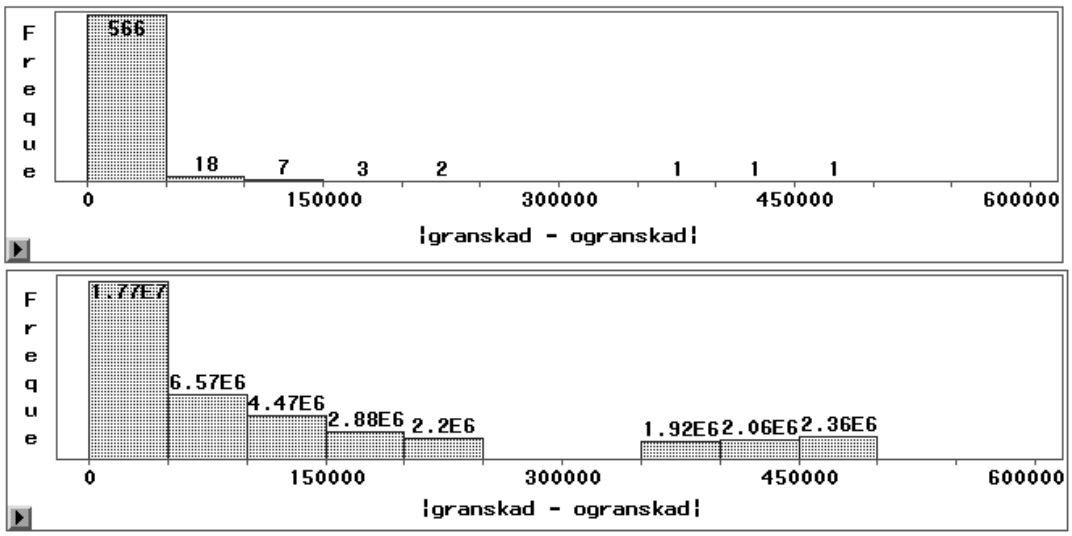
Diagram 5
Industristatistiken, 1990, fortsättning. Nettoeffekten (undre diagrammet) är skillnaden mellan staplarnas höjd så när som på den inverterade skalfaktorn



De 329 ändringarna (övre delen av diagram 5) fördelar sig på 247 negativa och 82 positiva och detta svarar mot en nettoeffekt (beräknad med hjälp av under delen av diagram 5) på skattningen som är $(1.13*10^{6} - 245771)*10^{-9} \approx 0.88$ promille.

Metod 2 Successiva restmängder
Rita ett histogram för absolutvärdet av differensen. Diagram 6 nedan visar resultatet när vi dolt de mest extrema differenserna.

Diagram 6
Industristatistiken 1990. Skillnad mellan granskat och ogranskat värde oavsett tecken. Siffrorna ovanför staplarna i det övre histogrammet markerar antalet observationer. Siffrorna ovanför staplarna[1] i det undre histogrammet visar, för givna intervall, effekten av granskningen i förhållande till skattningen så när som på en skalfaktor



Fortsätt att dölja differenser från höger till vänster, dvs. i första hand 1+1+1+2+3+7+18=33 stycken. Ta dessutom bort dem från beräkningen. Vi är härmed nere på differenser som är mindre än 50 000. Dessa kvarvarande 566 differensers sammanlagda effekt är 1.77 procent på skattningen. Ovan sade vi att vi i denna undersökning kan ignorera rättningar på en sammanlagd effekt av maximalt en promille av skattningen. Fortsätt därför att dölja observationer tills man uppnår detta resultat.

Avsluta med ett diagram som visar nettoeffekten av ändringarna (se diagram 5).

I detta exempel görs ingen uppdelning på redovisningsgrupper. Vi rekommenderar att analysen görs även för viktigare redovisningsgrupper.

Vi har i detta exempel visat att 329/603 ≈ 55 procent av de rättade värdena tillsammans svarade mot mindre än en tusendel av den totala skattningen. Här borde man kunna arbeta fram nya kontroller, t.ex. genom att vidga acceptansgränser. Målsättningen med de nya kontrollerna är att de inte ska felsignalera observationer där effekten på skattningen är obetydlig. Med en sådan åtgärd skulle vi kunna spara en stor del av granskningsarbetet. Vi vet också att risken för övergranskning är stor – större ju fler observationer som felsignaleras. Övergranskning innebär bland annat att fel som tidigare inte fanns i materialet kommer in i detta.

8.4 Numerisk metod

Ovanstående beräkningar kan också göras rent numeriskt. Man missar visserligen den information och känsla för materialet som man får från de successivt framtagna diagrammen. I gengäld leder metoden snabbt och direkt till målet, som är att få en uppfattning om var gränsen går mellan betydelsefulla och negligerbara ändringar. Vi vill således finna ett värde d sådant att sammantagna effekten av

[1] 10-potenser uttrycks i SAS med bokstaven E åtföljd av en siffra. E6 betyder således 10^6 (1 000 000). E-2 betyder 10^{-2} (0.01).

ändringar absolut mindre än d inte skulle haft någon praktisk betydelse på estimaten i det material man studerar.

$$\left|granskat\ värde - original\ värde\right| \begin{cases} > d & så\ är \quad d = betydelsefull\ ändring \\ \leq d & så\ är \quad d = betydelselös\ ändring \end{cases}$$

Jämför därefter kvoten mellan skattningen beräknad med endast ändringar större än d och materialet med alla ändringar medtagna.

$$\hat{Y} = \sum_{granskad} w_i x(g)_i + \sum_{originaluppgifter} w_i x(og)_i \qquad \text{vissa ändringar}$$

$$\hat{X} = \sum_{i} w_i x(g)_i \qquad \text{alla ändringar}$$

$$kvoten = \frac{\hat{Y}}{\hat{X}}$$

Vi använder även i det här fallet data från Industristatistiken 1990 för att exemplifiera våra beräkningar. Ett val av de fem gränserna d = ∞, 1 000 000, 100 000, 10 000 och 3 000 skulle ha resulterat i tablå 1. Där ∞ symboliserar helt ogranskat material.

Tablå 1
Industristatistiken 1990. Effekten av granskningen på skattningen

	d				
	∞	1 000 000	100 000	10 000	3 000
Kvot	1.208704	1.007118	0.996902	1.000281	1.000831

Tablån visar effekten av ändringarna på totalskattningen – precis som i exemplet i 8.3 Differensstudier med SAS/Insight. Liksom i exemplet i föregående avsnitt kan man dra slutsatsen att ”små” förändringar saknar betydelse på skattningen. Det här behöver dock inte vara fallet om man gör samma jämförelse på redovisningsgrupper. Vi rekommenderar därför att man även i det här fallet åtminstone genomför bearbetningar på viktiga (centrala) redovisningsvariabler.

8.5 Referenser

Engström, P. (1995), ”A Study on Using Selective Editing in the Swedish Survey on Wages and Employment in Industry”, Room paper No. 11, presented at the Conference of European Statisticians, Work Session on Statistical Data Editing, Athens, Greece, November 6–9, 1995

Forsman, G. (1991a), ”Olika felorsakers betydelse för granskningskostnaderna och skattningarnas kvalitet: en fallstudie på SCB:s finansstatistik”, GRANSK-PM Nr 22, 1991–02–05

Forsman, G. (1991b), ”Handledning för effektstudier av granskning”, GRANSK-PM NR 24, 91–08–25

Granquist, L (1996): ”An Overview of Methods of Evaluating Editing Processes”, A Contribution to Economic Commission for Europe, (1996): Statistical Data Editing: Methods and Techniques, Volume No. 2, United Nations New York and Geneva 1997.

Granquist, L and Kovar, J (1997): Editing of Survey Data: How much is enough? In L. Lyberg, P Biemer M. Collins, E. Leeuw, C. Dippo, N. Schwarz, D. Trewin (eds) Survey Measurement and Process Quality, Wiley, New York, 1997, pp. 415–436.

Hedlin, D. (1992), ”Jämförelse av granskade och ogranskade data i industristatistiken”, GRANSK-PM NR 29, 1992–09–07

Lindell, K. (1995 a), ”Evalveringsstudie av granskningsprocessen i lönestatistiken över kommunalt anställda”, Bakgrundsfakta till Arbetsmarknads- och Utbildningsstatistiken 1995:7.

Lindell, K. (1995 b), ”Evalveringsstudie av granskningsprocessen i lönestatistiken över landstingsanställda”, Bakgrundsfakta till Arbetsmarknads- och utbildningsstatistiken 1995:8

SAS/Insight, User’s Guide (1999), version 8. SAS Institute Inc. Manualen finns även elektroniskt tillgänglig via <Help> <Books and Training> <SAS OnlineDoc> <SAS/INSIGHT User's Guide>

Särndal, C-E, Swensson, B och Wretman, J: Model Assisted Survey Sampling, Springer, New York, 1992, pp 164–165.

Wahlström, C. (1990), ”Granskningens effekter – En studie av SCB:s Finansstatistik, F-Metod Nr 27, 1990–02–26.

Bilaga kapitel 8

Att skapa en SAS-tabell från en extern databas kan göras i SAS på två sätt

1. <File> <Import Data> (Excel, Access, kommaseparerad fil etc.)
2. Från SYBASE menyvägen eller med hjälp av följande program

```
libname mysyblib sybase
     server = SERVER
     database = DATABAS
     user = USERID
     dbprompt = yes
     defer = No;
data work.nytab;
     set mysyblib.SYBTAB
          (keep = var1 var2 var3
          rename = (var1=nyvar1 var2=nyvar2));
run;
```

där SERVER är namnet på servern
DATABAS är namnet på databasen
USERID är användaridentiteten
Var1, var2 etc. är variabelnamn i databasen. Vill man ändra namnen görs det i rename-satsen.

Detaljerad beskrivning av SAS/Insight-analysen

Manualen SAS/Insight, boken eller online-versionen, är mer utförlig än nedanstående beskrivning.

Välj en variabel.
Komplettera SAS-tabellen med skattningen[2] beräknad på den granskade variabeln. Vid en urvalsundersökning använder vi uppräknade värden.

Bilda differensen[3] mellan granskad och ogranskad uppgift för vald variabel, dvs.

$$granskad - ogranskad \qquad (1)$$

Beräkna även skattningen med hjälp av de granskade värdena.

Använd analysverktyget Distribution.

Komplettera SAS-tabellen med en ny variabel som nu är en konstant (skattningen enligt ovan).

Använd Fill values i databladets dialogruta.

Dölj de objekt som inte ändrats i granskningen (*granskat = ogranskat*) och ta bort dem från beräkningen.

Använd variabelverktyget (<Edit><Observations><Hide in graphs>).
Använd variabelverktyget (<Edit><Observations><Exclude in Calculations>).

Ta fram ett histogram (se ref Online-versionen Exploring Data in One Dimension, Box Plots) över differenserna.

[2] Enkla skattningar, såsom medelvärde och total, kan beräknas i SAS/Insight.
[3] Differenser och andra transformationer, se ref Online-versionen Transforming Varibles.

Skapa variabeln absolutvärdet av differensen, dvs.

$$\left|granskad - ogranskad\right| \qquad (2)$$

Använd variabelverktyget (<Edit><Variable>[funktion]) och skapa variabeln (funktion: abs(Y)).

Bilda variabeln effekt

$$effekt = \frac{\frac{1}{\pi} \times \left|granskad - ogranskad\right|}{skattningen} \times skalfaktor \qquad (3)$$

där π är urvalssannolikheten

Skalfaktorn, en tiopotens, bestäms så att variabeln effekt blir större än ett, 1, för alla objekt. SAS/Insight beaktar nämligen endast heltalsdelen vid vägningen i diagramverktyget histogram.

Variabeln enligt formel (3) måste skapas i flera steg, t.ex.

$$steg\_1 = \frac{1}{\pi} \times \left|granskad - ogranskad\right| \qquad (3´)$$

$$steg\_2 = \frac{steg\_1}{skattningen} = \frac{\frac{1}{\pi} \times \left|granskad - ogranskad\right|}{skattningen} \qquad (3´´)$$

Och slutligen det tredje steget som ger formel (3)

$$effekt = steg\_2 \times skalfaktor = \frac{\frac{1}{\pi} \times \left|granskad - ogranskad\right|}{skattningen} \times skalfaktor$$

Kommentar:
Det finns flera sätt att bestämma skalfaktorn. Det enklaste är att räkna antalet heltalssiffror i skattningen. Skalfaktorn blir tio upphöjt till antalet heltalssiffror. Ett annat sätt är att sortera SAS-tabellen i fallande skala efter variabeln, v

$$v = \frac{\frac{1}{\pi} \times \left|granskad - ogranskad\right|}{skattningen} \qquad (3´)$$

Välj skalfaktorn som det tal, tiopotens, som gör att det minsta (översta) värdet blir större än ett (1).

10-logaritmera, variabeln (2). (Det finns även alternativa trasformationer som kan vara möjliga att använda.)

Ta fram ett histogram för det transformerade värdet av variabeln (2)

Kommentar:
Histogrammet visar fördelningen för en variabel, i det här fallet absolutvärdet av differenserna. Stapelns höjd anger *antalet* objekt i delintervallet. För att få reda på hur stor effekt granskningen har på skattningen, behöver man ett vägt histogram. Stapelns höjd i *det histogrammet* visar effekten av granskningens inverkan på skattningen.

Observera att transformationer såsom logaritmering, kvadratrot osv. kan göras efter att man tagit fram histogrammet. Markera variabeln och använd därefter variabelverktyget (<Edit><Variable>[funktion]). Välj funktion. Variabelnamn (Name) och klartext (Label) genereras automatiskt, men det kan vara en fördel att åtminstone skriva in en egen klartext.

Gör plats för ett diagram intill det första diagrammet genom att rita en ruta under histogrammet.

Ta fram ett histogram för 10-logaritmen av variabeln (2) i den markerade rutan. Ange effekt (3) som frekvens. Detta visar *effekten* oavsett tecken.

Ta fram lämplig intervallbredd på följande sätt:

- Markera variabeln (x-axeln) i diagrammet.
- Öppna dialogrutan (►) längst ned till vänster i ett av diagrammen.
- Ändra intervallbredden tills en lämplig detaljeringsnivå uppnås. (Fördelningen ändras, dvs. blir mer eller mindre detaljerad.)

Skalan på axeln över frekvenser i det uppräknade histogrammet kan också behöva modifieras för att staplarna ska bli tydligare (högre)

- Markera variabeln, Frequency.
- Öppna dialogrutan (►) längst ned till vänster.
- Ändra maxvärdet till värdet på den högsta stalpen.

  Kommentar:
  I det här fallet är det inte nödvändigt att dölja och ta bort objekt. Logaritmen för oförändrade värden, *granskat = ogranskat* , är minus oändligheten. Av SAS uppfattas detta värde som missing values (uppgift saknas). Den (bakomliggande) procedur som genererar histogrammet bortser från missing values.

  Histogrammen visar samtliga förändrade värden i storleksordning (logskalan), ”storleksklass”. De stora förändringarna ligger till höger och mindre till vänster. Det första diagrammet (ovägt) visar antalet ändringar för gällande storleksklass. Det andra (vägt) visar effekten av ändringarna på skattningen.

Bruttoeffekt
Metod 1
Studera fördelningen av ändringarna genom att ändrar staplarnas bredd (”ticks” i diagrammets dialogruta). Bruttoeffekten av ändringarna på skattningen avläses som höjden på stapeln.

Metod 2
Skapa den första restmängden:

- Markera ett område från höger till vänster i ett av diagrammen.
- Dölj observationerna.
- Ta bort observationerna från beräkningen.

Fortsätt att skapa successiva restmängder genom att *dölja* objekt i diagrammen och *ta bort* dem från beräkningen.

Nettoeffekt
Nettoeffekten beräknas med hjälp av de observationer vars ändringar bedöms ha liten inverkan på skattningen. De ändringar som sammanlagt bedöms ha ringa inverkan på skattningen är de observationer som uppfyller villkoret

$$|granskat - ogranskat| \leq d$$

där d är gränsen.

Metod 1
Bilda en SAS-tabell med dessa observationer.

- Markera stapeln i diagrammet.
- Gå till databladet.
- Välj <extract> i databladets dialogruta (►)

Återgå till den ursprungliga undersökningsvariabeln **granskad – ogranskad.** Bilda ett nytt histogram för denna datamängd (diagram 5). Justera skalan så att de båda staplarna blir lika breda. Nettoeffekten är differensen av staplarnas höjd (undre diagrammet). Antalet positiva och negativa ändringar är staplarnas höjd i det övre diagrammet.

Metod 2
Fortsätt att dölja och ta bort observationer (se Bruttoeffekt Metod 2) till dess att villkoret

$$|granskat - ogranskat| > d$$

är uppfyllt. Denna restmängd används för att beräkna nettoeffekten.

Återgå till den ursprungliga undersökningsvariabeln **granskad – ogranskad.** Bilda ett nytt histogram för denna datamängd (diagram 5). Justera skalan så att de båda staplarna blir lika breda. Nettoeffekten är differensen av staplarnas höjd (undre diagrammet). Antalet positiva och negativa ändringar är staplarnas höjd i det övre diagrammet.

9 IT-miljön och Greta

I detta kapitel beskrivs granskning med hjälp av SCB:s rekommenderade program, Greta.

I huvudsak omfattar granskningsprocessen delprocesserna dataregistrering, granskning och rättning. Så stor del av granskningen som möjligt bör ske vid dataregistrering/skanning. Kontroll av misstänkta fel ska här bara ske för de kontroller som inte fordrar uppdatering av acceptansgränser med aktuella data.

Fel och misstänkta data ska presenteras så, att verifiering underlättas och uppdatering kan göras interaktivt.

9.1 Granskningsprocessens krav på IT-system

Krav på granskningsprogram:

- Fel och misstänkta data ska presenteras på ett sådant sätt att
  - verifiering underlättas
  - uppdatering kan göras interaktivt.
- Det ska vara enkelt för en användare att själv
  - ändra kontroller och acceptansgränser
  - lägga till nya kontroller
  - ta bort kontroller
  - använda hjälpinformation
  - utforma felmeddelanden
  - koda felkällor och uppgiftslämnarkapacitet.
- Omedelbar omgranskning ska kunna göras efter ändringar.
- Systemet ska generera processinformation och information om felorsaker.
- Lättillgänglig processtatistik ska kunna tas fram för övervakning och analys.
- Det ska vara en flexibel databaslösning.

En väl fungerande granskningsprocess måste vara flexibel. Detta betyder att det måste vara enkelt för en användare att utan hjälp från IT-specialister ändra kontroller och gränsvärden, lägga till nya kontroller och ta bort kontroller.

Granskningsprocessen ska generera processdata och data om felorsaker, vilka man använder för att utveckla undersökningen, t.ex. ändra blankettdesign och variabeldefinitioner. Detta ställer krav på bland annat flexibel databaslösning.

9.2 Generell programvara eller specialprogrammering

Grundregel:

- Använd helst generell programvara.
- Skräddarsy program bara om prestanda och funktionalitet blir mycket bättre.

SCB och andra statistikbyråer strävar efter att minska behovet av specialprogrammering av varje applikation genom att så långt som möjligt tillhandahålla generella program för varje uppgift. Fördelarna med generell programvara står främst

att finna i mindre utvecklingsinsats för den enskilda produkten samt billigare underhåll och drift av det färdiga systemet. Utvecklingsarbetet kan fokuseras på det generella programmet i vilket nya metoder kan implementeras och därmed komma hela organisationen till del omedelbart. Det finns naturligtvis fall där en generell programvara inte är lämplig. Då måste specialprogram tas fram. Den enda fördelen med skräddarsydda program är att man i vissa fall kan optimera prestanda och funktionalitet. Det måste finnas starka skäl för att specialskriva ett granskningsprogram.

9.3 Typer av granskningsprocesser

De granskningsprocesser som beskrivs i kapitel 3 har varierande grad av IT-stöd på SCB i form av generella programvaror.

9.3.1 Uppgiftslämnargranskning

Uppgiftslämnaren får ett elektroniskt formulär på diskett, via e-post eller med uppkoppling till SCB:s webbplats. Allt eftersom de efterfrågade uppgifterna fylls i, utlöses kontroller och hoppinstruktioner. Uppgiftslämnaren kan också avsluta med att initiera en granskning av samtliga uppgifter och kommentera funna ”avvikelser” innan uppgifterna skickas till SCB. Formuläret kan också innehålla bakgrundsinformation som kan utnyttjas för granskning och återrapportering.

TDE (Touchtone Data Entry) är ett system som är lämpligt när uppgiftslämnaren ska återkommande besvara ett mycket begränsat antal frågor. Uppgiftslämnaren kontaktar då själv SCB via telefon och lämnar med hjälp av knappsatsen ett fåtal uppgifter, vilka kontrolleras omedelbart och uppgiftslämnaren ges möjlighet att korrigera felaktiga svar.

Dessa system ger i dag inget generellt stöd för granskning.

9.3.2 Granskning vid dataregistrering

Skanning är den metod som rekommenderas för dataregistrering av pappersblanketter. Vid skanning lagras blanketten som en bild i en databas för senare analys. Det finns ett antal färdiga kontrollfunktioner i SCB:s nuvarande system att utnyttja vid granskning, t.ex. checksifferkontroll, kontroll mot intervall av giltiga värden. Därtill kommer möjligheten att utnyttja VBA-kod (VBA, Visual Basic for Applications) för att programmera mer komplicerade granskningsvillkor.

SCB:s erfarenheter visar på stora fördelar med att använda skanningteknik för dataregistrering och granskning. Den stora fördelen är att man omedelbart kan verifiera mot blanketten. Stora besparingar kan göras genom att blanketthanteringen och den manuella dataregistreringen nästan helt elimineras. Nackdelarna är få. Skannern kan missa något enstaka papper vid inläsningen. Bilderna av blanketterna kräver stort diskutrymme, men detta är ett problem som med tiden blir allt mindre.

För datainsamling via intervju använder SCB sedan många år DATI-systemet (numera WinDATI), som kan användas för såväl besöks- som telefonintervjuer. All granskning bör ske när intervjuaren lägger in uppgiften i datorn.

För traditionell dataregistrering av pappersblanketter finns programvaran RODE/PC. Förutom kontroll av dataregistreringsfel medger systemet att granskningskontroller läggs in för de fält som ska registreras.

9.3.3 Produktionsgranskning

För produktionsgranskning i databasmiljö rekommenderas det generella programmet Greta (se 9.5). Vi avråder från att man skriver egna granskningsprogram.

9.3.4 Outputgranskning

Grafisk granskning används antingen som ett alternativ till produktionsgranskning eller som ett komplement. Pc-miljön i kombination med klient-servertekniken är idealisk för detta.

Ett exempel är ett system vid AM-avdelningen som har fått många positiva omdömen och bedömts möjligt att generalisera.

SCB utreder och utvecklar metoder för att använda SAS/Insight för i första hand outputgranskning. Se kapitel 5.

9.4 Komponenter i ett granskningsprogram

Ett granskningsprogram ska stödja:

- *Kontroller*: inklusive anpassning av acceptansgränser.
- Hjälpinformation
  - definitioner och instruktioner: gäller främst uppgiftslämnargranskning men är också utmärkt som hjälpmedel åt granskaren.
- Felsignalering.
- Felmeddelanden.
- *Uppdatering* (interaktiv) med omedelbar omgranskning.
- *Kodning* av felkällor och uppgiftslämnarkapacitet.
- Generering av *processdata* (loggningar).
- Framtagning av lättillgänglig *processtatistik för* övervakning och analys.

9.5 Standardprogrammet Greta

Greta är SCB:s rekommenderade system för satsvis granskning av data som lagras i databaser i klient-servermiljö. Det är ett system där användaren får möjlighet att skapa och underhålla sina granskningskontroller. All granskning sker i servern med hjälp av SQL.

Syftet med Greta är att:

- tillhandahålla en standardiserad modell för hur granskningsarbetet ska gå till i klient-servermiljön
- ge produktansvariga möjlighet att själva, utan stöd från IT-personal, bygga upp och underhålla granskningsrutiner
- minska personberoendet
- skapa processinformation som senare kan analyseras.

Den typiska användaren är en produktionsstatistiker i såväl återkommande undersökningar som engångsundersökningar (uppdrag). Med Greta kan användaren lätt lägga till, ta bort eller förändra kontroller.

Det bör understrykas att Greta passar bra för uppdrag och engångsundersökningar, eftersom det är lätt för en uppdragsansvarig att konstruera kontrollerna och interaktivt rätta upp fel helt utan programmering. Den processinformation som Greta skapar i dessa fall kan komma till användning i liknande uppdrag.

9.5.1 Möjligheter med Greta

Med Greta bygger användaren upp kontroller med nyckelord på svenska via menyval och inbyggda verktyg. Kontrollsystemet kan omedelbart testas ut i datamaterialet genom att visa:

- totala antalet felsignalerade variabelvärden
- totala antalet felsignalerade objekt
- antalet objekt fördelade efter antalet felsignaler för objekten (i fallande ordning).

Inför produktionsstarten kan man försäkra sig om att granskningen ser ut att bli rimlig genom att ovanstående statistik automatiskt presenteras för användaren.

Styrning och uppföljning

Produktionen kan följas och styras genom att Greta presenterar statistik över antalet objekt fördelade efter felstatus, dvs. hur många objekt som återstår att granska, hur många som väntar med status ”under utredning” osv.

Anpassning till de data som ska granskas

Acceptansgränser för kontroller för misstänkta fel (ex: kvotkontroller X/Y) kan göras på aggregerade nivåer efter användningens behov. Användaren specificerar alltså själv nivåer (t ex: efter SNI på ensiffer-, tvåsiffernivå osv.). För varje kontroll kan man sedan automatiskt anpassa gränserna till de data som ska granskas genom att t.ex. använda Agda för att räkna fram och lägga in parametervärden för kontrollen i en Greta-tabell.

Selektiv granskning / makrogranskning

I detta CBM rekommenderas Hidiroglou-Berthelots metod och en poängfunktion (score-function) för prioritering av objekt för manuell verifiering (uppföljning), s.k. selektiv granskning. Det krävs viss SQL-kunskap för att implementera selektiv granskning i Greta.

Automaträttningar och imputeringar

För att man ska kunna göra automaträttningar, även kallade deterministiska rättningar, krävs dels att kontroller kan skrivas så att felet identifieras, dels att ett nytt värde kan härledas, t.ex. vid summeringsfel. Vissa typer av imputeringar för partiellt bortfall kan också göras – t.ex. med värdet från föregående period. Ändrade och nya värden kommer att granskas om kontrollerna för identifiering av fel och partiellt bortfall placeras först på kontrollistan.

Felmeddelanden

Greta sammanställer flaggningarna till objektvisa felmeddelanden som presenteras på skärmen. Felmeddelandena kan sorteras på olika sätt, t.ex. efter bransch eller region, beroende på hur man vill organisera den manuella verifieringen. Man kan

även få meddelandena sorterade efter antalet flaggningar, t.ex. så att man kan börja med att verifiera de värsta objekten först. Med en poängfunktion implementerad bör man kunna ta ut objekten i prioritetsordning, dvs. efter objektens poäng.

Orsaken till en flaggning anges genom felkoden, beskrivning av kontrollen i klartext, angivande av kontrollens värde med mera.

Interaktiv uppdatering med omgranskning
Man ändrar eller godkänner direkt på skärmen samt inför åtgärds- och orsakskoder. När alla flaggningar är åtgärdade, ska man begära omgranskning av just det objektet om man har ändrat eller lagt till något värde. Genom omgranskningen kontrollerar man ändringarna och får fram om ändringar medför att andra kontroller utlöses.

Man kan dessutom ångra en felaktig knapptryckning, avbryta verifiering av objekt för att vid ett senare tillfälle fortsätta med objektet, markera variabel och objekt för utredning, markera outliers, dvs. korrekta men starkt avvikande värden som måste beaktas vid estimationen, m.m.

Top-down-granskning
Greta ger möjlighet till en form av top-down-granskning genom att det är möjligt att sortera objekten eller kontrollerna efter antal felsignaler.

Processtatistik
Ett centralt begrepp i Greta är *felradstabellen*, som, förutom att vara underlag för felmeddelanden, utgör underlag för indikatorer och statistik över insamlings- och produktionsprocessen.

Framställning av processdata som underlag för kontinuerlig förbättring av processer är grundpelaren i TQM-filosofin och för den moderna synen på granskningen. I den är det identifiering av felkällor som ska stå i fokus, och sådana kunskaper ska användas till åtgärder att förhindra att fel uppstår.

Det är möjligt att med hjälp av SQL-kodning ta fram processdata från felradstabellen.

9.5.2 Gretas svagheter
I dag finns version 2.75 tillgänglig i nätet. Där har inte alla funktioner som ursprungligen planerades kunnat införas.

Dataregistrering
Greta är mindre lämpligt att använda vid dataregistrering, eftersom programmet är konstruerat för att granska hela objekt och inte enskilda variabler. Man kan alltså inte arbeta interaktivt på så sätt att man får en felsignal när markören lämnar ett fält som innehåller ett fel. Om man behöver denna funktion, bör man använda Rode PC.

Outputgranskning
Greta är inte lämpligt för outputgranskning i betydelsen att man först utför kontroller på aggregat, t.ex. jämförelser med motsvarande tabellcell för perioden innan; och därefter söker identifiera vilket/vilka objekt som bidragit mest till förändringen. För denna funktion kan t.ex. SAS Insight användas.

Grafisk granskning
Greta ger inget direkt stöd för grafisk granskning. Det är dock relativt enkelt att göra en grafiskt baserad applikation som utnyttjar den information som Greta skapar.

Omaka objekt
Kontroller som innebär matchning mellan flera tabeller kommer inte att ge felsignal om ett objekt saknas i någon av de ingående tabellerna.

> Exempel: Om granskningen bl.a. avser kontroll mot samma objekt föregående period, kommer ett objekt som inte förekom tidigare period endast att granskas ”inom posten”. Inget meddelande kommer att visa att objektet saknas föregående period. Nyckelordet ”förekommer i” är reserverat för denna typ av kontroll, men har inte implementerats i nuvarande version. Detta problem står högst på önskelistan över förbättring av Greta.

Beroende kontroller
Flödesvillkor (”if … then”, ”select … case”) kan inte uttryckas direkt i Gretas språk. I stället kan man i de flesta fall uppnå motsvarande funktionalitet genom att använda funktionen ”Beroende av” i ett eller flera led. Metoden kan dock snabbt bli osmidig och svår att underhålla.

> Exempel: att uttrycka ”om variabel A < x ska variabel B > y, men om variabel A > x ska variabel B > z” kräver totalt fyra kontroller, varav två beroende.

Matchning mellan databaser
Alla tabeller som ingår i en granskning (såväl granskat data som eventuella hjälptabeller) måste av praktiska skäl finnas i en och samma databas.

Radorienterade tabeller
Greta är utvecklad för att granska data som lagras på traditionellt sätt i en relationsdatabas, dvs. där en (eller flera) kolumner identifierar objektet och övriga kolumner representerar variabelvärden: ”identitet”, ”var1”, ”var2”, ”var3”, … Greta klarar endast vissa mycket begränsade kontroller när variabler i stället lagrats på s.k. radorienterat sätt, dvs. där en rad i databastabellen i princip kan beskrivas med tre kolumner: ”identitet”, ”kod”, ”värde” och därmed en kolumn inte unikt identifierar en variabel.

9.5.3 Vidareutveckling av Greta
Den mest angelägna vidareutvecklingen av Greta är en generell statistikmodul för presentation av processtatistik.

Ordlista

Ord	Förklaring
Acceptansgränser	Gränser som avgör vilka observationer som ska godkännas respektive felsignaleras.
Aggregatgranskning	Kontroller utförs på aggregerade data.
Avvikelsefel	Värdet på testvariabeln är för stort eller för litet i förhållande till acceptansgränserna.
Balanskontroll	Kontroll för att se till att värdena på delvariablerna summerar sig till värdet på summavariabeln.
Batch-granskning	Se *produktionsgranskning*.
Dataregistreringsgranskning	Identifiering av uppenbara fel vid dataregistrering.
Datastruktur- eller modellfel	Definitionsmässiga samband mellan variabler satisfieras inte. Specialfall: *balanskontroll*.
Definitionsfel	Många uppgiftslämnare uppfattar en fråga eller underliggande definitioner på ett likartat men felaktigt sätt.
EDA, Exploratory Data Analysis	Utforskning av data med hjälp av diagram.
Felkod	Beteckning på den kontroll som signalerade ett misstänkt eller uppenbart felaktigt värde.
Felmeddelande	Den information som granskningssystemet ger om objekt och variabler som har felsignalerats eller automatiskt åtgärdats.
Felrad	Teknisk term i granskningsprogrammet Greta (kapitel 9).
Felsignal	Markering av att ett variabelvärde saknas eller har underkänts av de maskinella kontrollerna. Kan även användas om objekt och avser då att objektet har minst ett felsignalerat variabelvärde.
Feltyp	Beteckning eller benämning av visst slag av fel som förekommer i en datainsamling, t.ex. registrerings- eller sortfel.

Ord	Förklaring
Flagga	Synonym till *felsignal*.
Grafisk granskning	Grafisk granskning av data beskrivs i form av olika diagram. Här behandlas interaktiv granskning, vilket innebär kommunikation mellan diagram och mellan diagram och datablad.
Granskning	Identifiering och åtgärdande av fel och outliers i individuella data som används för framställning av statistik. Ett huvudsyfte är att identifiera felkällor för senare åtgärder i undersöknings- och produktionsprocessen.
Granskningskod	Synonym till *åtgärdskod*.
Granskningsstatus	Anger på objektsnivå om objektet har granskats av programmet, är godkänt eller ska åtgärdas.
Imputering	Ett felsignalerat och felaktigt värde eller saknat värde ersätts med ett acceptabelt värde efter fastställda regler utan kontakt med uppgiftslämnaren.
Inlier	Felaktiga värden som ligger innanför varje rimligt val av acceptansgränser för avvikelsefel.
Interaktiv grafisk granskning	Se *grafisk granskning*.
Konsistensfel	Svar på en fråga motsäger svar på en annan eller andra frågor.
Kontroll	Algoritm att urskilja felaktiga, misstänkta eller saknade värden.
Lådagram	Diagram som visar en variabels fördelning, bl.a. med kvartiler och extremvärden.
Makrogranskning	Se *selektiv granskning*.
Manuell förgranskning	Granskning av postenkäter före registrering.
Mikrogranskning	Identifierar och åtgärdar fel i individuella data.
Misstänkt värde	Värde utanför ett acceptansområde, explicit eller implicit, som med stor sannolikhet är felaktigt.

Ord	Förklaring
Objektvis dataregistrerings-granskning	Granskningen utförs när alla värden för ett objekt har registrerats.
Outlier	Korrekta data som uppräknade skiljer sig så mycket i absolut storlek från övriga uppräknade värden i sin redovisningsgrupp, att de aktualiserar frågor om förändringar i estimationen eller kommentarer i redovisningen av undersökningens resultat.
Outputgranskning	Granskning när allt material är insamlat för kontroll av att inga stora misstag har gjorts i de tidigare utförda processerna. Utförs i allmänhet på aggregerade data.
Paretodiagram	Stapeldiagram som presenteras i fallande ordning.
Partiellt bortfall	Obesvarad fråga trots att objektet har data att redovisa.
Processdata	Data om produktionsprocessen, felkällor och problem i undersökningen och granskningsprocessens effektivitet.
Processtatistik	Statistik över felsignaler, åtgärdade data, orsaker till fel, uppgiftslämnarproblem med mera.
Produktionsgranskning	Granskning som utförs på mikrodata efter det att en viss mängd objekt finns.
Selektiv granskning	Metod för prioritering av objekt eller variabelvärden för verifiering.
Strukturfel	Se *datastruktur- eller modellfel*.
Testvariabel	En insamlad variabel eller ett aritmetiskt uttryck av undersökningens variabler, som används i kontroller.
Top-down	Granskning/verifiering utförs i prioriteringsordning med prioritet ”det värsta först.”
Träffsäkerhet	Andel av felsignalerade observationer som resulterar i att data ändras.
Uppenbara fel	Kan identifieras med säkerhet enbart med tillgång till data om det granskade objektet.

Ord	Förklaring
Uppgiftslämnargranskning	Utförs av uppgiftslämnaren (t.ex. vid besvarande av postenkät eller elektronisk blankett) eller av uppgiftslämnare och intervjuare gemensamt i intervjuundersökningar.
Uppgiftslämnarkapacitet	Uppgiftslämnaren kan inte besvara frågan (med rimlig arbetsinsats).
Uppgiftslämnarservice	Statistikproducentens stöd till uppgiftslämnaren i rapporteringsarbetet.
Validitetsfel	Det registrerade värdet tillhör inte variabelns värdeförråd (tillåtna värden).
Variabelvis dataregistreringsgranskning	Registreringen stoppas för åtgärd så fort ett variabelvärde underkänns.
Verifiering	Manuell utredning av felsignalerade variabelvärden i syfte att fastställa om de kan accepteras, är felaktiga eller utgör outliers. Fel och outliers ska åtgärdas och orsaker till fel ska registreras.
Återkontakt	Kontakt med uppgiftslämnaren i syfte att verifiera felsignaler, dvs. få korrekta uppgifter, finna ut orsaker till fel, förklaringar till felsignaler samt att identifiera uppgiftslämnarens problem med att besvara frågor.
Åtgärdskod	Anger vilken/vilka åtgärder som genomförts för ett variabelvärde. Exempel: Felsignalerat men ej verifierat; godkänts utan ändring och utan kontakt med uppgiftslämnaren; ändrats utan kontakt med uppgiftslämnaren; ändrats efter kontakt med uppgiftslämnaren. Synonym till *granskningskod*.
Övergranskning	Innebär att det i ett visst skede i processen tillkommer fler fel än dem som åtgärdas eller att resursåtgången inte står i rimlig proportion till kvalitetsförbättringen.

ISBN 91-618-1138-6

Statistikpublikationer kan beställas från SCB, Publikationstjänsten, 701 89 ÖREBRO, e-post: publ@scb.se, telefon: 019-17 68 00, fax: 019-17 64 44. De kan också köpas genom bokhandeln eller direkt hos SCB, Karlavägen 100 i Stockholm. Aktuell publicering redovisas på vår webbplats (www.scb.se). Ytterligare hjälp ges av SCB:s Bibliotek och Information. e-post: information@scb.se, telefon: 08-506 948 01, fax: 08-506 948 99.

www.scb.se

SCB-Tryck, Örebro 2002.06