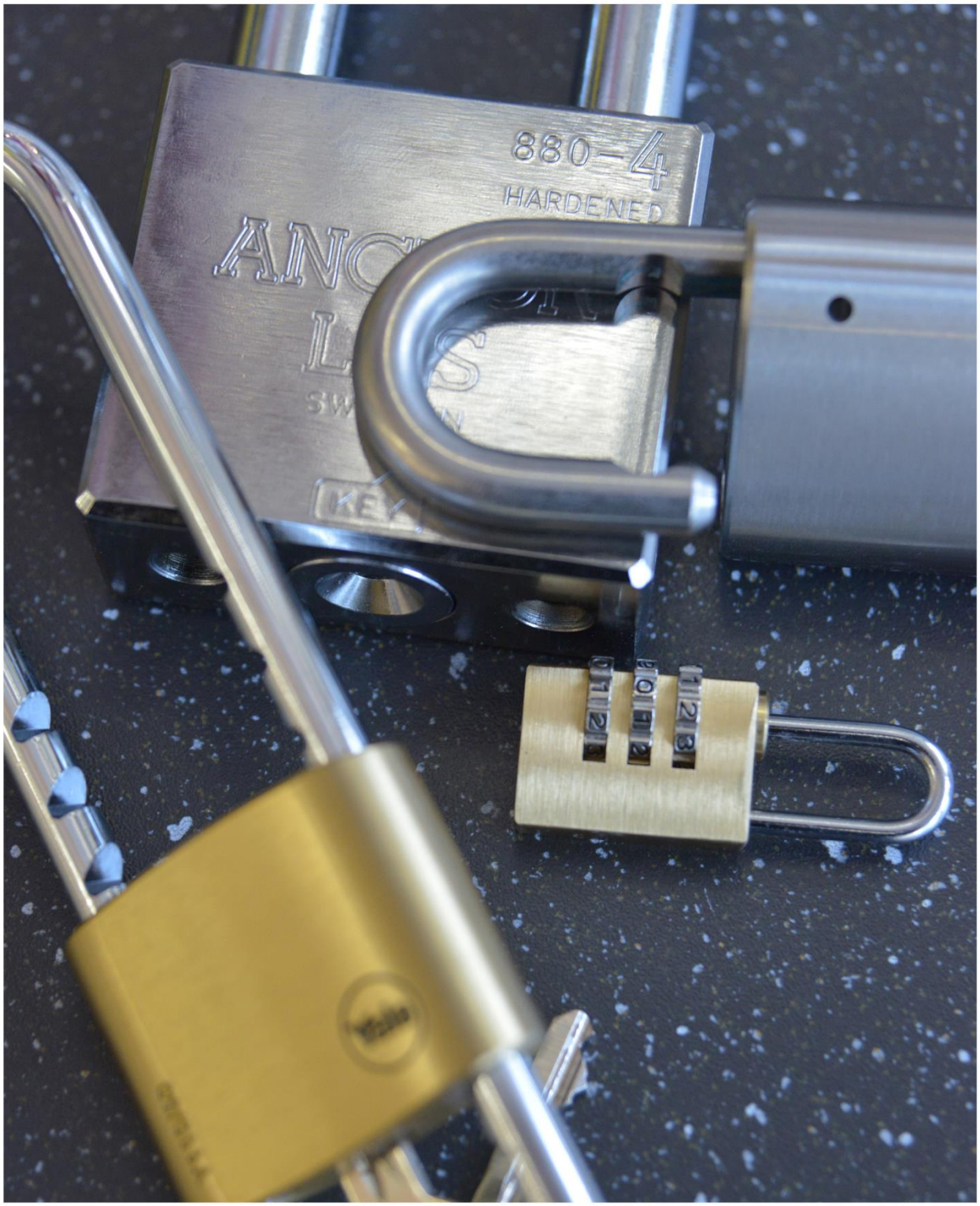
RÅDET FÖR DEN OFFICIELLA STATISTIKEN



Handbok i statistisk röjandekontroll

Sveriges officiella statistik

Förord

Denna handbok har tagits fram av Samarbetsgruppen för röjandekontroll inom Rådet för den officiella statistiken (ROS). Handboken är avsedd att vara en vägledning i statistisk röjandekontroll för statistikansvariga myndigheter i deras produktion och redovisning av officiell och annan statistik. Den kan även ge stöd vid utlämnande av mikrodata. Handboken ska bidra till att säkerställa sekretesskyddet för enskilda samtidigt som statistikredovisningens innehåll och kvalitet bibehålls så långt som möjligt.

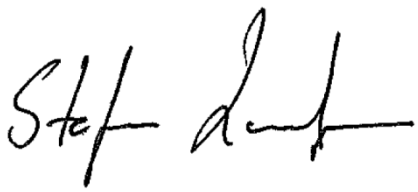
Samarbetsgruppen påbörjade, efter beslut av ROS, sitt arbete i februari 2013 med att kartlägga läget avseende röjandekontrollen hos de deltagande myndigheterna. Därefter vidtog planering och en längre skrivperiod. Delar av innehållet har hämtats från eller inspirerats av Statistiska centralbyråns *Handbok i statistisk röjandekontroll av tabeller* från 2010, skriven av Michael Carlson och Ingegerd Jansson. Inriktningen på den nya handboken förankrades vid ett seminarium med statistikansvariga myndigheter den 27 januari 2014.

Samarbetsgruppen har letts av Jörgen Brewitz (f.d. Svensson), Statistiska centralbyrån, och har i övrigt bestått av Anton Färnström, Brottsförebyggande rådet, Saadia Aitattaleb, Brottsförebyggande rådet, Malin Blomqvist (t.o.m. augusti 2013), Energimyndigheten, Klas Unger (fr.o.m. oktober 2013), Energimyndigheten, Jenny Johansson, Energimyndigheten, Olof Håkanson, Försäkringskassan, Jimmie Enhäll, Jordbruksverket, Henrik Sundström, Skolverket, Helena Svensson, Skolverket, Henrik Nordin, Socialstyrelsen, Charlotta Sandström (t.o.m. mars 2014), Socialstyrelsen, Mats Wiklund, Trafikanalys, Per Gillström, Universitetskanslersämbetet, Anders Sundström (t.o.m. juni 2013), ROS-sekretariatet, Cathy Krüger, ROS-sekretariatet, Emma Luukka (fr.o.m. april 2014), ROS-sekretariatet, och Michael Carlson, Stockholms universitet (konsult åt Statistiska centralbyrån). I gruppen har både statistisk och juridisk kompetens ingått.

Ett varmt tack riktas till alla som på olika sätt bidragit till framställningen av handboken. Ett särskilt tack riktas till några medarbetare på Statistiska centralbyrån – Martin Ribe, Ingegerd Jansson och Johan Strandman – som lämnat viktiga bidrag till texterna. Handboken fastställdes vid mötet med ROS den 20 februari 2015.

Samarbetsgruppen avslutar härmed sitt arbete. Handboken ska dock kunna revideras när det uppstår nya behov av rekommendationer, till exempel när vidareutvecklade metoder och it-verktyg tillkommer.

Stockholm i februari 2015



Stefan Lundgren

Ordförande i Rådet för den officiella statistiken

I oktober 2018 har uppdateringar gjorts på sidorna 26, 29, 50, 51 och 97-99 med anledning av att Europaparlamentets och rådets förordning (EU) 2016/679 av den 27 april 2016 om skydd för fysiska personer med avseende på behandling av personuppgifter och om det fria flödet av sådana uppgifter och om upphävande av direktiv 95/46/EG (allmän dataskyddsförordning) (EU:s dataskyddsförordning) började tillämpas den 25 maj 2018.

Innehåll

1 Inledning

1.1 Syfte och målgrupp

Denna handbok är avsedd att användas som vägledning vid statistisk röjandekontroll. Statistisk röjandekontroll är metoder för att se till att inte uppgifter om enskilda individer eller företag ska gå att utläsa ur redovisad statistik eller statistiska material. Röjandekontrollen behövs till följd av lagens krav och för statistikens kvalitet.

Målgruppen för handboken är i första hand de som arbetar med framställning eller spridning av statistik eller statistiska material och som behöver överväga eller tillämpa metoder för riskbedömning och skydd. Även andra kategorier av medarbetare kan ha nytta av handboken.

Handboken är avsedd främst för officiell statistik och övrig statistik som statistikansvariga myndigheter svarar för, men den kan vara till nytta även för annan statistik. Rekommendationerna i handboken är inte bindande, utan avsedda som vägledning för röjandekontroll av god kvalitet. Olika myndigheter kan behöva specificera mer detaljerade riktlinjer för sina specifika områden. Denna handbok avser att ge stöd för ett mellan och inom myndigheterna konsekvent synsätt på röjandekontroll i arbetet med statistik.

Handboken bygger på Statistiska centralbyråns interna *Handbok i statistisk röjandekontroll av tabeller* från 2010. Den har här utvidgats väsentligt och avpassats för vidare användningsområden.

Metoder för makro- och mikrodata

Utförligast i handboken behandlas metoder för *makrodata*, det vill säga aggregerade data. Det är i vid mening sammanräknade data, sådana som normalt redovisas i statistiktabeller, diagram eller kartor. Begreppet *tabeller* används i vid bemärkelse och avser aggregerade data som sammanställts på ett strukturerat sätt för olika redovisningsgrupper. De olika aggregat eller sammanräkningar som värdena i en tabell avser, även marginalsummor och liknande, kallas *celler*.

Handboken tar även upp metoder för *mikrodata*, data som inte är sammanräknade utan avser enskilda observationer för exempelvis individer eller företag.

1.2 Utgångspunkter

Meningen med röjandekontroll

Begreppet röjandekontroll används främst i betydelsen statistisk röjandekontroll. Förutom statistisk röjandekontroll finns också ”teknisk röjandekontroll”, som innebär att den tekniska säkerheten är så hög att obehöriga inte kan komma åt uppgifter som omfattas av sekretess, och ”administrativ röjandekontroll”, som innebär att uppgifterna skyddas genom efterlevnad av juridiska krav, till exempel genom en avgränsad statistikverksamhet inom en myndighet.

Statistisk röjandekontroll syftar till att hålla nere risken för att enskilda (individer eller företag) kan lida skada eller men genom att uppgifter om dem kan utläsas (röjas) i redovisningar av statistik eller relaterad information. Detta syfte ska uppnås utan att statistikredovisningens informationsinnehåll störs onödigt mycket. Röjandekontroll innebär därför en avvägning i form av ett optimeringsproblem: minimera åtgärdens inverkan på statistiken under villkoret att risken för röjande hålls acceptabelt låg.

Normalt går det inte att helt eliminera röjanderisken, men det är nödvändigt att hålla den på en acceptabelt låg nivå. Vad som är acceptabelt varierar och kan inte entydigt fastställas i generella kriterier. Åtgärder för att minska röjanderisken innebär i princip också att informationen i tabellen minskar, i och med att värden ändras eller tas bort.

Röjande i statistisk mening kan för denna handbok definieras enligt följande:

> Ett *röjande* föreligger när en utomstående med hjälp av statistiskt material, egen bakgrundskunskap och logiska slutledningar – med eller utan maskinell hjälp – får ny kunskap om en egenskap hos ett enskilt objekt i en population av individer, företag eller motsvarande.

Med *statistiskt material* ovan avses till exempel tabeller eller anonymiserade/avidentifierade (se begreppsförklaring i kapitel 13) mikrodatafiler.

Med *angripare* menas i det följande den som utnyttjar möjligheter till röjande. Det kan till exempel vara en statistikanvändare, någon som med uppsåt söker skyddad information eller någon som i ett annat ärende (utan uppsåt) råkar ta del av materialet. Bakgrundskunskapen hos angriparen kan variera från kännedom om en person till förfogande över ett helt register. Den erhållna kunskapen kan vara säker eller osäker, exakt eller diffus.

Definitionen av röjande ovan gäller enskilda objekt som individer eller företag. Det är inte fråga om att skydda något slags ”integritet” för hela befolkningsgrupper eller kategorier av företag. Om statistiken skulle visa att någon mindre tilltalande egenskap är oväntat frekvent i en bred befolkningsgrupp, så ska statistikredovisningen återge detta öppet, inte dölja det, för att tjäna sitt syfte.

Led i röjanden – identifiering och attribuering

Två led i röjanden är vad som kallas identifiering och attribuering. En *identifiering* sker när en angripare urskiljer vilket enskilt objekt i en population som ligger bakom en uppgift i ett statistikmaterial. En *attribuering* sker när en angripare utläser någon egenskap hos ett enskilt objekt i en population.

En identifiering kan ibland ses som ett röjande i sig, eller ge angriparen möjlighet att röja ett objekt genom attribuering. Röjande genom attribuering kan dock ske även på andra sätt, utan en identifiering innan. Identifiering och attribuering kommer att tas upp närmare i avsnitt 4.3 och 6.1.

Exempel: Om en statistiktabell visar att det bor exakt en x-åring i en mindre kommun Y, och angriparen från annat håll känner till namnet på kommunens enda x-åring, så kan angriparen identifiera personen bakom ettan i tabellen. Om en annan tabell visar att exakt en x-åring i kommunen har en årsinkomst på minst z kronor, så kan angriparen göra en attribuering och sluta sig till att den bekanta x-åringens årsinkomst är på minst z kronor.

Lagens krav

Utgångspunkten i denna handbok är att ett datamaterial inte får lämnas ut om det inte har genomgått en skadeprövning och sedan (vid behov) skyddats i enlighet med skadeprövningens utfall. Med utlämnande avses här offentliggörande, publicering eller leverans av statistik och även delning av mikrodata på begäran, med eller utan förbehåll. Myndigheter som producerar statistik är skyldiga enligt 24 kap. 8 § offentlighets- och sekretesslagen (2009:400) att upprätthålla sekretess för uppgifter som kan hänföras till en enskild. Som huvudregel gäller absolut sekretess.

I den bestämmelsen ingår dock ett antal undantag som anger lägen där uppgifter får lämnas ut om det står klart att så kan ske utan att någon enskild lider skada eller men, se vidare avsnitt 3.3. Till följd av lagens bestämmelser ska varje utlämnande föregås av en sådan

kontroll av materialet att det är säkerställt att det inte kan ge upphov till röjande som innebär skada eller men.

Begreppen skada och men används av lagstiftaren för att beskriva de olägenheter som är avgörande för om publicering kan ske. Med skada avses ekonomisk skada och med men avses integritetskränkningar av olika slag.

Internationella regler

Statistiksekretessen stöds även i Riktlinjer för europeisk statistik (Rådet för den officiella statistiken, 2013). Riktlinjerna för europeisk statistik (även kallade uppförandekod) gäller enligt EU-förordningen (nr 223/2009) om den europeiska statistiken. Denna statistik är förkortat sagt statistik som behövs för EU:s verksamhet, och den styrs genom *Det europeiska statistiksystemet*, som är partnerskapet mellan de europeiska och nationella statistikorganen. Den femte av de femton principer som presenteras i riktlinjerna lyder: *Absoluta garantier ges avseende uppgiftslämnarnas (hushålls, företags, förvaltningars och andra respondenters) integritet samt att uppgifterna behandlas konfidentiellt och enbart används för statistikändamål*. Skrivningen om absoluta garantier får tolkas som att absolut sekretess gäller eller att det för undantagen ska stå klart att uppgifter kan offentliggöras utan att någon enskild lider skada eller men.

FN:s grundläggande principer för officiell statistik innefattar principen att data som insamlas av statistikmyndigheter för statistikproduktion, oavsett om de avser fysiska eller juridiska personer, ska vara sekretessbelagda och enbart användas för statistiska ändamål.

Viktigt för kvalitet

Sekretessen och röjandekontrollen är också viktiga för statistikens kvalitet. De ger nämligen uppgiftslämnare (de som lämnar uppgifter om sig själva eller om andra) en grund för förtroende för att uppgifterna skyddas, vilket ger förutsättningar för god svarsfrekvens och god svarskvalitet.

Om uppgiftslämnare skulle misstänka att känsliga uppgifter kan komma att spridas eller användas på andra sätt än vad som uppgetts, finns det en risk att svaren uteblir eller inte blir uppriktiga. Detta skulle kunna gå ut över hela den officiella statistiken. Att absolut sekretess är huvudregeln ger statistikansvariga myndigheter ett utrymme att ha ett restriktivare skydd än vad undantagen öppnar för, och det kan vara till fördel för statistikens kvalitet. Restriktiv tillämpning av undantagen är i linje med uttalanden i förarbetena till lagen.

Röjandekontrollen borgar också för kvaliteten i fråga om statistikens tillgänglighet för användare, genom att den skapar förutsättningar för statistikens spridning.

1.3 Innehållet i denna handbok

De följande kapitlen avhandlar olika aspekter av röjandekontroll, med avseende på arbetsgång, förutsättningar, metoder, verktyg och råd för typiska fall:

- Kapitel 2 beskriver röjandekontroll som en process och visar schematiskt hur en kontroll utförs.
- Kapitel 3 behandlar de grundläggande juridiska förutsättningarna.
- Kapitel 4 behandlar de huvudsakliga metodmässiga förutsättningarna, till exempel typ av tabeller, variabler och objekt, som spelar en viktig roll vid skadeprövningen och i valet av metoder.
- Kapitel 5 ger sammanfattande beskrivningar av olika metoder för röjanderisk-bedömning.
- Kapitel 6 behandlar bedömning av risk för skada eller men.

- Kapitel 7 behandlar metoder för skyddande av makrodata, främst tabeller.
- Kapitel 8 behandlar möjligheten att begära samtycke till att efterge sekretess.
- Kapitel 9 behandlar metoder för bedömning av informationsförlust.
- Kapitel 10 behandlar röjandekontroll av mikrodata, till skillnad från kapitel 4–9 som huvudsakligen ägnas åt makrodata.
- Kapitel 11 behandlar it-verktyg för röjandekontroll.
- Kapitel 12 ger handledning för olika kombinationer av grundläggande tabelltyper och för olika andra förutsättningar, och kompletterar med enklare exempel.
- Kapitel 13–15 innehåller begreppsförklaringar, en svensk-engelsk ordlista och en referenslista.

Denna handbok är inte uttömmande på så sätt att den som ansvarar för röjandekontroll klarar sig helt utan annan dokumentation och litteratur. För att undvika redundans har alltför detaljerade beskrivningar valts bort, särskilt när andra källor med fördel kan användas. De statistiska beskrivningarna har i allmänhet hållits relativt enkla. Kapitel 5, 7, 9 och 10 fordrar dock viss förtrogenhet med statistisk metodik.

En viktig källa är Eurostats *Handbook on Statistical Disclosure Control* (se Statistics Netherlands, 2010), ett omfattande dokument som tagits fram i en rad projekt finansierade av Eurostat där även Statistiska centralbyrån medverkat. I den ges uttömmande beskrivningar av många relevanta metoder, men handbokens omfattning, statistiskt-tekniska karaktär och språk gör att den i praktiken delvis kan vara svår att använda direkt i det dagliga arbetet. Boken *Statistical Disclosure Control* (se Hundepool m.fl., 2012) bygger på handboken och rekommenderas för närmare information.

Den föreliggande handboken är tänkt att fungera som en brygga mellan produktionen och Eurostats handbok. Läsare som behöver fördjupade beskrivningar av metoder hänvisas till Eurostats handbok eller till annan relevant litteratur.

Referenser till relevant litteratur finns inlagda i den löpande texten, och en referenslista finns i kapitel 15.

2 Röjandekontrollsprocessen

I detta kapitel beskrivs röjandekontroll som en process. De olika ingående momenten och begreppen behandlas och förklaras översiktligt, som en introduktion till närmare beskrivningar i följande kapitel.

2.1 Processen och dess moment

Röjandekontroll av tabeller är en sammansatt process som kan delas upp i olika moment enligt följande:

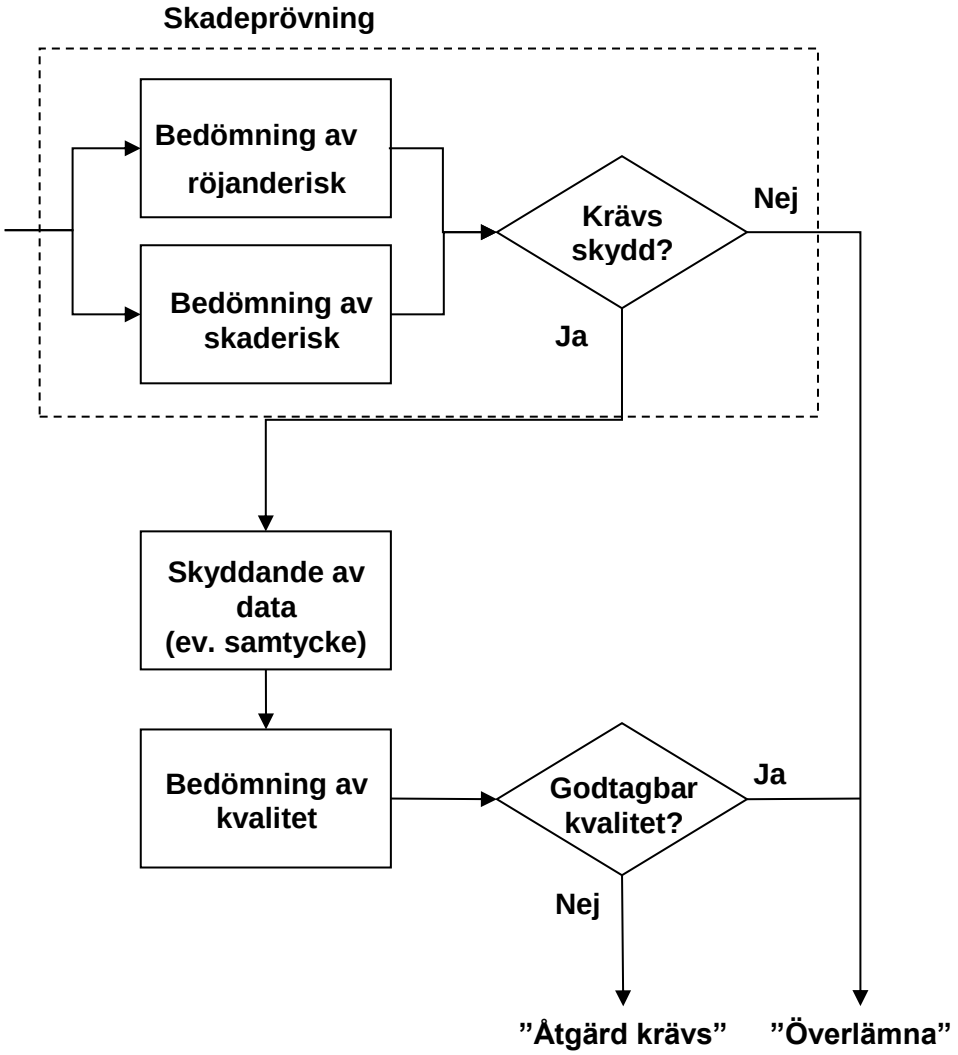
- Skadeprövning av tabell.
  - Bedömning av röjanderisk.
  - Bedömning av skaderisk (risk för skada eller men).
- Skydd av tabell.
- Bedömning av kvalitet.

Röjandekontrollsprocessen tillämpas för alla statistikmaterial som bygger på uppgifter om personliga eller ekonomiska förhållanden för enskilda, det vill säga fysiska och juridiska personer. En förenklad bild av röjandekontrollsprocessen ges i figur 2.1 på nästa sida. Röjandekontroll av mikrodata behandlas i kapitel 10.

Skadeprövningen ska ge en bild av dels de risker för röjande som finns i en given tabell, dels de skador och men som kan följa om ett röjande skulle ske. När behovet av skydd är konstaterat ska tabellen behandlas för att skyddas, med så god kvalitet som möjligt i den skyddade tabellen. Motsvarande gäller för mikrodatamaterial.

Momenten kan utföras tillsammans i en process där val i ett led ibland kan påverka beslut i ett senare led. I praktiken kan kontrollen ofta utföras som en iterativ (upprepad) process som prövar sig fram och kan backa i olika steg tills en tillfredsställande lösning har nåtts. Dessutom ingår hela denna process som en komponent i en större process för statistikproduktionen, som även den kan innehålla iterativa inslag.

Utförandet av en röjandekontroll ska naturligtvis avpassas efter de förutsättningar som gäller för statistiken i fråga och dess användningsområden. Det är viktigt att utforma metodiken för röjandekontrollen i grova drag i ett planeringsskede innan tabellerna börjar tas fram. En förberedande beskrivning av hur röjandekontrollen ska gå till utgör en grund för kommande produktionstillfällen, då riskbedömning och skyddande utförs enligt den design som valts för statistikprodukten. All statistik som publiceras eller lämnas ut ska genomgå en skadeprövning och sedan (vid behov) skyddas i enlighet med skadeprövningens utfall.



Figur 2.1 Röjandekontroll av tabeller eller mikrodata: processen och dess olika moment

Resultatet av röjandekontrollsprocessen dokumenteras för officiell statistik på vanligt sätt i dokumentationsverktyget *Beskrivning av statistiken*. De som svarar för röjandekontrollen ska vara restriktiva i extern och intern kommunikation med detaljer om röjandekontrollsprocessen, såsom valda värden på styrparametrar (inställningar av metodval) till programvaran, för att inte underlätta röjanden. Sådan information omfattas av sekretess, eftersom ett utlämnande av sådan information kan innebära detsamma som utlämnande av sekretesskyddade uppgifter om enskilda. Se även avsnitt 5.10. För restriktivt internt bruk behövs dock utförlig dokumentation av gjorda val och planer för att se över metodiken.

2.2 Skadeprövning av tabell

Skadeprövningen består av två delar som ofta hanteras tillsammans. För det första gäller det att bedöma *sannolikheten för att uppgifter kan röjas* om ett enskilt objekt, det vill säga en individ, ett hushåll, ett företag eller en organisation. För det andra ska det bedömas om de uppgifter som kan röjas och hänföras till en enskild *medför skada eller men* för denne.

Även om sannolikheten för röjande i en tabell bedöms vara obetydlig, kan skada eller men för en enskild bli betydande om ett röjande ändå skulle inträffa.

Begreppen inom röjandekontroll svarar delvis men inte exakt mot dem i annan riskanalys, såsom Risk Management. Även där bedöms sannolikheter för olika oönskade händelser och dessas konsekvenser. Risken är där den förväntade skadan, uppfattad som sannolikheten för en händelse multiplicerad med konsekvensen av händelsen, eventuellt summerat över olika händelser.

Inom röjandekontroll står begreppet röjanderisk för sannolikheten att ett röjande kan ske, och begreppet skaderisk står för konsekvensen av ett röjande. Den ”sammantagna skaderisken” är den förväntade skadan och ska bedömas med hänsyn till dels sannolikheten för ett röjande, dels omfattningen eller storleken av skada eller men vid ett röjande. Riskkalkylen går inte att formalisera fullt ut, utan i praktiken blir det fråga om bedömningar. Ju större den sammantagna risken är, desto mer omfattande åtgärder kan behövas för att det ska kunna anses stå klart att det inte uppstår skada eller men.

Ett exempel kan vara en tabell där det framgår att någon har vaccinerats mot influensa och en annan tabell där det framgår att någon har en allvarlig sjukdom. Sannolikheten för att någon identifieras kan vara lika stor i båda tabellerna, men konsekvenserna genom det men som uppkommer till följd av röjandet skiljer sig troligen åt, och den senare tabellen kan därför behöva mer omfattande skyddsåtgärder.

Det går inte att fixera en generell turordning för skadeprövningens två delar. I somliga fall är det naturligt att först bedöma skaderisken vid ett röjande innan det går att bestämma vilka parametervärden som ska sättas för bedömningen av röjanderisken. I andra fall går det inte att bedöma skaderisken förrän det har klarlagts vilka objekt som riskerar att röjas. Exempelvis storleken på orten där ett företag är beläget kan inverka på både sannolikheten för och konsekvenserna av ett röjande, och de båda aspekterna ska behandlas samstämmigt.

De två delarna av skadeprövningen ska därför normalt utföras parallellt och samordnat. Enligt huvudregeln i 24 kap. 8 § offentlighets- och sekretesslagen är statistiksekretessen absolut för uppgifter om enskildas personliga och ekonomiska förhållanden som kan hänföras till den enskilde. I bestämmelsen görs dock fyra undantag från huvudregeln som kan tillämpas om det står klart att uppgifter kan röjas utan att den enskilde eller någon närstående till denne lider skada eller men (se avsnitt 3.3). Detta är vad skadeprövningen har att utgå från.

Det kommer an på den statistikansvariga myndigheten att ta ställning till troliga röjandescenarier och principer för skadeprövning, med överväganden och beslut på ansvarig nivå enligt delegeringsordning. Enskilda handläggare ska inte oplanerat behöva stå för övergripande bedömningar av sannolikheter och konsekvenser.

Verksamheten i företag och organisationer är delvis öppen för insyn, bland annat genom årsberättelser. Det kan kanske vara lätt att tro att ett företag inte lider skada om uppgifter som redan är offentliggjorda röjs i en statistisk tabell, av det skälet att uppgifterna är allmänt kända. Detta är emellertid inte ett godtagbart resonemang, för statistiksekretessen gäller ändå. Dessutom skulle ett publicerande av sådana uppgifter i ett statistiskt sammanhang innebära att uppgifterna exponeras på ett nytt sätt och kan komma att kopplas ihop med andra uppgifter på ett sätt som kan innebära skada. Skadeprövningen behövs därför även om uppgifterna tidigare är offentliggjorda. Röjandekontrollen är också viktig för förtroendet för statistikproducenterna.

Bedömning av röjanderisk

Sannolikheten att en uppgift om en enskild går att röja är i regel inte möjlig att beräkna med någon matematisk teori. I stället arbetas med bedömda kriterier för att identifiera celler där

sannolikheten för röjande är alltför betydande. Röjandesannolikheten kan i regel inte garanteras vara i strängaste mening obefintlig, men den ska kunna bedömas vara så ringa att den är betydelselös.

Gränsen för vad som är betydelselöst operationaliseras i bedömda kriterier. Ett exempel som visar idén är följande. Enligt ett kriterium, som kallas tröskelvärdesregeln med celltröskelvärdet *k* (se avsnitt 5.1), betraktas sannolikheten för röjande som betydelselös om antalet populationsobjekt som ett statistikvärde grundas på är minst *k*. Ett offentliggörande av en tabell med endast betydelselöst små röjanderisker kan inte betraktas som ett utlämnande av uppgifter om enskilda.

Olika tabellkonstruktioner och förutsättningar i övrigt ger olika röjandescenarier, det vill säga olika sätt som ett röjande skulle kunna ske på. Dessa förutsättningar är avgörande vid valet av metod för identifiering av riskceller. En röjanderiskbedömning ska beakta

- tabelltyp, allmänna förutsättningar och beskrivning av riskscenarier
- val av kriterier för riskbedömningen och val av parametervärden.

Röjanderiskbedömningen genererar utdata i form av

- en tabell med markering av identifierade riskceller
- en sammanfattande beskrivning av resultatet, till exempel antalet eller andelen riskceller och berörda objekt.

Geografiska förhållanden ska uppmärksammas, då sådana uppgifter underlättar identifiering av uppgiftslämnare. Uppgifter om populationens storlek och sammansättning samt antalet observationer (totalräknat eller från urval) är också viktiga i den totala bedömningen av röjanderisken. Metoder för bedömning av röjanderisk beskrivs i kapitel 5.

Bedömning av skaderisk

Utöver bedömningen av röjanderiskerna och röjandescenarierna, ska en kvalitativ bedömning också göras av om ett röjande kan medföra skada eller men för de berörda objekten. Med skada avses ekonomisk skada, med men avses integritetskränkningar av olika slag. Benämningen skaderisk innefattar både risk för skada och risk för men.

Det gäller att bedöma om skaderisken är betydelselös eller inte.

Skaderisken är betydelselös om det står klart att uppgiften kan röjas utan att den enskilde som uppgiften avser eller någon närstående till denne lider skada eller men (se vidare avsnitt 3.3).

Frågor att ta hänsyn till är bland annat:

- Vilka uppgifter riskerar att röjas? Är det integritetskänsliga data?
- Vilken detaljeringsgrad har uppgifterna som riskerar att röjas? Är värdena exakta i någon mening eller approximativa?
- Tidsaspekter: Är det gamla eller färska data?
- Vem avser uppgifterna som riskerar att röjas?
- Ska uppgifterna publiceras så att många lätt kan ta del av dem, eller ska de lämnas ut till enstaka användare? Vad är i så fall mottagarens syfte? Hur kommer uppgifterna att skyddas hos mottagaren? Kan ett förbehåll hindra att uppgifterna röjs av mottagaren?

Bedömning av skaderisk diskuteras mer ingående i kapitel 6.

2.3 Skydd av tabell

Tabellkonstruktionen och förutsättningarna i övrigt enligt avsnitt 2.2 ovan styr även valet av skyddsmetoder. Beslut behöver tas om val av metod(er) för hur en tabell ska skyddas och om värden på styrparametrar för den valda metoden.

Ett genomförande av denna delprocess genererar minst

- en ny skyddad tabell
- en sammanfattande beskrivning av resultatet, till exempel antalet eller andelen berörda celler och objekt.

Metoder för att skydda tabeller beskrivs i kapitel 7.

En alternativ möjlighet som kan minska eller eliminera behovet av tabellskydd för vissa typer av undersökningar är att den enskilde genom samtycke efterger sekretessen, se kapitel 8.

2.4 Bedömning av kvalitet

Åtgärder för att minska risken för skada eller men medför i princip också att informationen i tabellen minskar, genom att den skyddade tabellen avviker från den ursprungliga. Detta behandlas i kapitel 9. Tabellen ska därför bedömas i fråga om kvalitet utifrån begreppen i Meddelanden i samordningsfrågor för Sveriges officiella statistik, MIS 2001:1, se Statistiska centralbyrån (2001a). Därutöver kan det finnas krav från kunden. Exempel på kvalitetsaspekter som kan påverkas av röjandekontrollen är

- innehåll, att de variabler och redovisningsgrupper som efterfrågas finns med utan alltför besvärande luckor
- tillförlitlighet, främst att precisionen inte försämras alltför mycket
- jämförbarhet och samanvändbarhet, att skyddet utförs på ett kontrollerat sätt och inte omöjliggör jämförelser över tid eller mellan undersökningar
- tillgänglighet, att statistiken kan spridas utan hinder av röjanderisk
- innehållsspecifika bedömningar av hur röjandekontrollen påverkar hur kundens önskemål uppfylls.

Olika mått på ändringarna i den överlämnade tabellen i förhållande till den oskyddade tabellen kan enkelt dokumenteras. Exempel på olika aspekter är

- antalet eller andelen påverkade celler
- antalet eller andelen påverkade objekt
- den redovisade variabeln summerad över samtliga påverkade celler.

Från början ska läggas fast vilka kvalitetskriterier som ska tillämpas. En sammanfattande beskrivning av röjandekontrollens effekter på slutprodukten ska ingå i kvalitetsdeklarationen.

2.5 Överlämnande eller vidare åtgärd

Efter en utförd röjandekontroll finns två möjliga utfall. Har kontrollen fungerat som avsett med en acceptabel kvalitet som följd, kan tabellen överlämnas till nästa steg i produktionsprocessen. Om skyddsmetoden däremot medfört en kvalitetsförlust som inte är godtagbar behöver ytterligare åtgärder vidtas.

Sådana åtgärder kan vara mer eller mindre omfattande. Det kan till exempel handla om en iterativ process där nya parametervärden för den valda skyddsmetoden väljs eller att alternativa skyddsmetoder prövas. Ibland behövs mer omfattande åtgärder, som en större modifiering av den ursprungliga tabellplanen (varvid en ny skadeprövning ska göras) eller överläggningar med kunden/uppdragsgivaren.

Det är viktigt att ha strategier för att kunna hantera situationer när de angivna kriterierna med avseende på kvalitet och röjanderisk inte kan uppnås.

För material som levereras till Eurostat kan det finnas krav på hur röjanderiskerna ska beräknas och vilket skydd som ska tillämpas. Se vidare avsnitt 3.5.

3 Juridiska förutsättningar

Detta kapitel innehåller en beskrivning av bestämmelsen i offentlighets- och sekretesslagen som reglerar den sekretess som gäller för uppgifter avseende en enskilds personliga eller ekonomiska förhållanden vid framställning av statistik. Kapitlet inleds med en allmän genomgång av sekretessbestämmelsen, varefter beskrivningar av huvudregeln och undantagen följer. Dessutom presenteras några rättsfall för att ge en bild av hur den aktuella bestämmelsen tillämpats av domstolarna. I det sista avsnittet belyses en del internationella aspekter.

3.1 Sekretessbestämmelsen i 24 kap. 8 § offentlighets- och sekretesslagen

Allmänna kommentarer

De grundläggande bestämmelserna om allmänna handlingar finns i en av Sveriges grundlagar – tryckfrihetsförordningen (1949:105). Bestämmelser om i vilken utsträckning allmänna handlingar kan omfattas av sekretess finns i offentlighets- och sekretesslagen (2009:400). En kommentar till lagen ges i Lenberg m.fl. (2010). Med sekretess avses ett förbud att röja en uppgift, vare sig det sker muntligen, genom utlämnande av en allmän handling eller på något annat sätt, till exempel genom publicering av en detaljerad tabell.

Vid en begäran om utlämnande av allmänna handlingar ska en ansvarig handläggare först ta ställning till om det som begärs utlämnat är en allmän handling eller en uppgift i en sådan handling. Om så är fallet ska handläggaren göra en sekretessprövning, det vill säga ta ställning till om uppgifterna omfattas av en bestämmelse om sekretess i offentlighets- och sekretesslagen och, i sådant fall, om uppgifterna kan lämnas ut. Formellt är det myndigheten i fråga som ska göra dessa ställningstaganden, men detta ansvar kan i praktiken vara delegerat till en handläggare enligt en delegeringsordning. Reglerna ovan gäller bland annat uppgifter som lämnas ut via mikrodatafiler och uppgifter som offentliggörs via statistiska tabeller, databaser, diagram, kartor eller analyser.

Att röja en uppgift som är sekretessbelagd är straffbart enligt 20 kap. 3 § brottsbalken (1962:700). Som brott mot tystnadsplikt betraktas röjande av en uppgift som är hemlig enligt offentlighets- och sekretesslagen, annan förordning eller enligt förordnande eller förbehåll som har meddelats med stöd av offentlighets- och sekretesslagen eller annan författning. Straffbestämmelsen riktar sig alltså mot var och en som har tystnadsplikt enligt lag eller annan författning.

Vidare får inte uppgifter i den officiella statistiken sammanföras med andra uppgifter i syfte att utröna en enskilds identitet. Detta följer av 6 § lagen (2001:99) om den officiella statistiken (jämför avsnitt 3.3 nedan, inledningen). Det är straffbart att bryta mot förbudet enligt 26 § i samma lag, men i ringa fall ska ansvar inte utdömas. Det krävs någon form av manipulativt beteende eller andra klart klandervärda åtgärder[1] för att handlingen ska vara straffbar.

Skaderekvisit

Flera bestämmelser i offentlighets- och sekretesslagen är utformade på ett sådant sätt att vissa förutsättningar måste finnas för att uppgiften ska vara hemlig. En bedömning ska då göras med hänsyn till den risk för skada eller men som utlämnandet kan medföra i det

[1] Prop. 2000/01:27, s. 60.

specifika fallet. Dessa sekretessbestämmelser innehåller rekvisit som anger sekretessens styrka, så kallad skaderekvisit.

Ett skaderekvisit innebär att sekretessen gäller under förutsättning att någon viss angiven risk för skada eller men uppstår om uppgiften lämnas ut. Det finns två huvudtyper av skaderekvisit, raka och omvända.

Det *raka* skaderekvisitet uttrycker att huvudregeln är offentlighet och att uppgifterna får lämnas ut. Detta skaderekvisit uttrycks vanligen som att sekretess gäller i en viss verksamhet för en viss uppgift ”om det kan antas att den enskilde [som uppgiften avser] eller någon närstående till denne lider skada eller men om uppgiften röjs”.

Det *omvända* skaderekvisitet utgår däremot från sekretess som huvudregel. Ett sådant skaderekvisit brukar formuleras så att sekretess gäller i viss verksamhet för viss uppgift ”om det inte står klart att uppgiften kan röjas utan att den enskilde eller någon närstående till denne lider skada eller men”.

I en del sekretessbestämmelser ställs det inte upp några särskilda villkor för att sekretessen ska gälla för de uppgifter och i det sammanhang som beskrivs där. Det brukar kallas för att sekretessen är absolut.

Sekretessbestämmelsen

Sekretessbestämmelsen avseende uppgifter inom statistikverksamhet i 24 kap. 8 § offentlighets- och sekretesslagen lyder som följer:

Sekretess gäller i sådan särskild verksamhet hos en myndighet som avser framställning av statistik för uppgift som avser en enskilds personliga eller ekonomiska förhållanden och som kan hänföras till den enskilde.

Motsvarande sekretess gäller i annan jämförbar undersökning som utförs av Riksrevisionen, av riksdagsförvaltningen, av Statskontoret eller inom det statliga kommittéväsendet. Detsamma gäller annan jämförbar undersökning som utförs av någon annan myndighet i den utsträckning regeringen meddelar föreskrifter om det.

Uppgift som behövs för forsknings- eller statistikändamål och uppgift som inte genom namn, annan identitetsbeteckning eller liknande förhållande är direkt hänförlig till den enskilde får dock lämnas ut, om det står klart att uppgiften kan röjas utan att den enskilde eller någon närstående till denne lider skada eller men. Detsamma gäller en uppgift som avser en avliden och som rör dödsorsak eller dödsdatum, om uppgiften behövs i ett nationellt eller regionalt kvalitetsregister enligt patientdatalagen (2008:355).

För uppgift i en allmän handling gäller sekretessen i högst sjuttio år, om uppgiften avser en enskilds personliga förhållanden, och annars i högst tjugo år.

Bestämmelsen fick sin nuvarande utformning den 1 augusti 2014.[2]

Meddelarfrihet

I den svenska offentlighetsprincipen enligt grundlagarna ingår meddelarfriheten, som gör det möjligt för befattningshavare i offentlig tjänst att utan straff lämna normalt sekretessbelagda uppgifter för publicering i tryckt skrift eller för offentliggörande i radioprogram eller tekniska upptagningar. Meddelarfriheten ger en rätt att lämna *uppgifter*, men inte lämna ut sekretessbelagda *handlingar*. Rätten att meddela och offentliggöra uppgifter är inskränkt avseende uppgifter som omfattas av sekretess, inklusive statistiksekretess enligt 24 kap. 8 § offentlighets- och sekretesslagen. Det betyder att om

[2] Prop. 2013/14:162, s. 1.

uppgiften skyddas av bestämmelsen om statistiksekretess har sekretessen företräde framför reglerna om meddelarfrihet. Uppgiften får då inte lämnas till exempelvis massmedia i syfte att publiceras.[3]

3.2 Huvudregeln i sekretessbestämmelsen

Absolut sekretess

Den första meningen i 24 kap. 8 § offentlighets- och sekretesslagen innehåller inget skaderekvisit som anger sekretessens styrka. Sålunda är huvudregeln att uppgifterna omfattas av så kallad absolut sekretess, vilket medför att myndigheten inte ska göra en prövning av om någon lider skada eller men. Det finns dock fyra undantag från huvudregeln, se nedan i avsnitt 3.3.

När bestämmelsen infördes i dåvarande sekretesslagen (1980:100) framhölls bland annat att ett av skälen till att uppgifterna skulle omfattas av starkt sekretesskydd var att intresset av att skydda enskildas integritet ansågs väga tyngre än allmänhetens rätt till insyn i uppgifter som lämnats av enskilda för statistikändamål.

De följande underavsnitten förklarar begreppen som nämns i huvudregeln.

Särskild verksamhet som avser framställning av statistik

I bestämmelsen anges det sammanhang för vilket uppgifter kan omfattas av sekretess. Det framgår att sekretess gäller för uppgifter i särskild verksamhet som avser framställning av statistik. Det saknar betydelse varifrån uppgifterna kommer eller hur de har kommit till myndigheten. Vad som avgör är i stället i vilken verksamhet uppgifterna förekommer. Den verksamhet som avses är myndigheternas egen statistikproduktion och uppgiftslämnande till andra myndigheters produktion av statistik. Verksamheten kan vara organiserad som en egen enhet eller liknande. Det ska vara sådan statistikverksamhet som är allmänt utredande, utan anknytning till något särskilt ärende. Vad som avses med ärende är inte särskilt definierat. I kommentarer till förvaltningslagen sägs att det är naturligt att till ”ärende” hänföra främst sådant som utmynnar i någon form av beslut från myndighetens sida i förhållande till enskild, se Hellners och Malmqvist (2010). Det viktiga är att verksamheten är avgränsad från annan verksamhet. Det innebär att samma uppgift kan omfattas av sekretess när den finns i en verksamhet som framställer statistik inom en myndighet, men vara offentlig när den finns i en annan verksamhet inom samma myndighet.

Ett typexempel på särskild verksamhet är sådan verksamhet som avser framställning av officiell statistik, som ska finnas för allmän information, utredningsverksamhet och forskning.

Vid bedömning av om uppgifterna förekommer i särskild verksamhet för statistik-framställning måste en helhetsbedömning av verksamheten göras. Om den särskilda verksamhetens huvudsakliga syfte är att framställa statistik är 24 kap. 8 § offentlighets- och sekretesslagen tillämplig. Bestämmelsen är inte tillämplig i en verksamhet som har något annat än statistikframställning som sitt främsta syfte, även om det i verksamheten skulle förekomma insamling eller bearbetning av uppgifter med hjälp av statistiska metoder.[4] Vägledning kan hämtas från uttalanden i förarbetena som säger att det inte ska vara fråga om ren driftsstatistik eller statistik som framställs för att användas som underlag för beslut i visst ärende.

[3] 24 kap. 9 § offentlighets- och sekretesslagen.

[4] Prop. 2013/14:162, s. 9.

Annan jämförbar undersökning

Vissa undersökningar som genomförs utanför en särskild verksamhet för framställning av statistik omfattas också av statistiksekretess. En skillnad här är att sekretessens räckvidd inte gäller i hela den verksamhet där undersökningen görs, utan är begränsad till de uppgifter som kan hänföras till själva undersökningen. En sådan undersökning benämns i offentlighets- och sekretesslagen som ”annan jämförbar undersökning”. Bestämmelser om vilka undersökningar som omfattas av statistiksekretess, förutom de som utförs av Riksrevisionen, riksdagsförvaltningen, Statskontoret eller inom det statliga kommittéväsendet, finns i 7 § offentlighets- och sekretessförordningen (2009:641).[5]

Uppgift som avser en enskilds personliga eller ekonomiska förhållanden och som kan hänföras till den enskilde

Med begreppet enskild avses både fysiska och juridiska personer. Även avlidna personer[6] omfattas av begreppet, men däremot inte myndigheter. Kommuner och landsting ingår exempelvis inte i begreppet.

Endast fysiska personer kan ha uppgifter som avser personliga förhållanden. Uppgifter om ekonomiska förhållanden kan avse både fysiska och juridiska personer.

För att uppgifterna ska omfattas av sekretess enligt den ifrågavarande bestämmelsen måste de kunna hänföras, direkt eller indirekt, till en enskild. Av uttrycket ”direkt eller indirekt” följer att uppgifter anses kunna hänföras till en enskild inte bara genom sådant som namn eller personnummer, utan även på andra sätt såsom via registreringsnummer på ett fordon som ägs av den enskilde. Inom statistisk röjandekontroll används begreppen identifiering och attribuering i detta sammanhang, se förklaringar i avsnitt 1.2, 4.3 och 6.1 eller i kapitel 13.

3.3 Undantagen i sekretessbestämmelsen

Enligt huvudregeln i sekretessbestämmelsen är sekretessen absolut, det vill säga den gäller oberoende av om det kan antas att den enskilde lider skada eller men om uppgiften lämnas ut. Det betyder att uppgifterna inte i något fall får lämnas ut. För tabeller innebär detta att sannolikheten för röjande ska bedömas vara betydelselös (se förklaring i avsnitt 2.2).

Statistiken får alltså inte vara så detaljerad att det utan vidare går att bakvägsidentifiera enskilda. Med *bakvägsidentifiering* menas indirekt identifiering genom kombination med andra uppgifter. I enlighet med vad som nämnts ovan (avsnitt 3.1, underavsnittet ”Allmänna kommentarer”) är bakvägsidentifiering från officiell statistik förbjuden enligt 6 § lagen om den officiella statistiken (2001:99). Även om detta förbud också lägger ett ansvar på användarna av statistiken, så är det ändå nödvändigt att förebygga risken i möjlig mån.

Paragrafen med sekretessbestämmelsen ger fyra undantag från huvudregeln. Dessa undantag avser

1. uppgifter som behövs för forskningsändamål
2. uppgifter som behövs för statistikändamål
3. uppgifter som inte genom namn, annan identitetsbeteckning eller liknande förhållande är direkt hänförliga till den enskilde
4. uppgifter som rör dödsorsak eller dödsdatum och behövs i ett nationellt eller regionalt kvalitetsregister.

[5] Prop. 2013/14:162, s. 10.

[6] Begreppet enskild omfattar även avliden, se prop. 1979/80:2, del A, s. 264.

För uppgifter som omfattas av något av undantagen till regeln om absolut sekretess gäller ett omvänt skaderekvisit, det vill säga uppgifter får lämnas ut om det står klart att den enskilde som avses eller någon till denne närstående inte lider skada eller men av utlämnandet. Det omvända skaderekvisitet utgår från sekretess som huvudregel. Det är bara om det med en viss grad av sannolikhet kan utgås från att det är ofarligt att lämna ut en viss uppgift som sekretessen viker.[7] Undantagen bör enligt förarbetena till lagen tillämpas restriktivt. Det innebär i praktiken att uppgifter i de allra flesta fall skyddas av den absoluta sekretessen. Vid minsta tvekan om undantagens tillämplighet ska alltså uppgifter inte lämnas ut.

Forskningsändamål

För att tillgodose forskningens behov har möjligheten att lämna ut uppgifter som behövs för forskningsändamål införts. Det ändamål för vilket uppgifterna begärs ut ska alltså vara forskning. Med uppgifter avses både mikro- och makrodata. För det fall uppgifterna lämnas till en myndighets forskningsverksamhet överförs sekretessen dit, om inte uppgifterna där redan omfattas av en annan sekretessbestämmelse till skydd för samma intresse.[8]

Om en myndighet i sin forskningsverksamhet får en sekretessreglerad uppgift från en annan myndighet enligt 11 kap. 3 § offentlighets- och sekretesslagen, ska myndigheten tillämpa den sekretessbestämmelsen som gäller för uppgiften hos den utlämnande myndigheten. Sekretesskyddet hos den mottagande myndigheten utgör i det fallet inte någon skaderisk i sig. I vissa fall kan det dock finnas en primär sekretessbestämmelse som hos den mottagande myndigheten gäller före den överförda sekretessen, 11 kap. 8 § offentlighets- och sekretesslagen. Här kan det vara viktigt att vara uppmärksam så att uppgifterna på detta sätt inte får ett sämre skydd än vad som är avsett. Om utlämnandet medför att uppgifterna kommer att omfattas av ett svagare sekretesskydd kan det i sig innebära en risk för skada, inte för att själva användningen av uppgifterna hos mottagaren medför skada utan för att det svagare sekretesskyddet hos mottagaren kan medföra att uppgifterna lämnas vidare på ett sätt som medför skada.[9]

Utöver att fastställa ändamålet för utlämnandet är det alltså också viktigt att säkerställa vem utlämnandet sker till och vilket skydd uppgifterna får hos mottagaren. Lämnas uppgifterna till en privat forskningsaktör är det normalt lämpligt att utlämnandet sker med ett förbehåll som inskränker mottagarens rätt att lämna uppgifterna vidare eller hur dessa får utnyttjas, enligt 10 kap. 14 § offentlighets- och sekretesslagen. Huruvida ett förbehåll kan användas och hur det kan utformas får utredas och avgöras i varje specifikt ärende. Se vidare avsnitt 6.3.

För att bedöma om det är fråga om forskningsändamål kan vägledning[10] hämtas från 2 § lagen (2003:460) om etikprövning, där forskning definieras som vetenskapligt experimentellt eller teoretiskt arbete för att inhämta ny kunskap och utvecklingsarbete på vetenskaplig grund, dock inte sådant arbete som utförs inom ramen för högskoleutbildning på grundnivå eller avancerad nivå. Vid tillämpning av detta undantag ska åtskillnad göras mellan forskning å ena sidan och uppföljning, utvärdering och kvalitetssäkring å andra sidan.[11] För avgränsningen av vad som avses med forskning i förhållande till uppföljning och utvärdering kan vidare viss vägledning hämtas från lagen (1998:543) om

[7] Prop. 1979/80:2, del A, s. 79.

[8] Jämför 11 kap. 3 § och 8 § offentlighets- och sekretesslagen.

[9] Prop. 2013/14:162, s. 12.

[10] Regeringsrättens årsbok 2004, ref. 9.

[11] Regeringsrättens årsbok 2004, ref. 9.

hälsodataregister. Förarbetena definierar här uppföljning enligt följande. *Uppföljning* avser att fortlöpande och regelbundet mäta och beskriva behov, verksamheter och resursåtgång angivet i termer av till exempel behovstäckning, produktivitet och nyckeltal. Uppföljning syftar till att ge en översiktlig bild av verksamhetens utveckling och att fungera som en signal för avvikelser som bör beaktas. *Utvärdering* avser analys och värdering av kvalitet, effektivitet och resultat hos en verksamhet i förhållande till de mål som bestämts för denna.[12] Vid varje enskilt fall får bedömas, utifrån vad den som begär ut uppgifter anger, om uppgifterna ska användas i forskning.

Statistikändamål

Utöver undantaget avseende forskningsändamål infördes genom en lagändring år 1995 ett andra undantag från huvudregeln. Undantaget avser uppgifter som behövs för statistikändamål. Med uppgifter avses både mikro- och makrodata. Syftet med lagändringen var att säkra tillgången till uppgifter för framställning av statistik, främst den officiella statistiken.[13] I förarbetena till lagändringen betonades bland annat att möjligheten till utlämnande borde tillämpas mycket restriktivt, såsom varit fallet vid utlämnande till forskningsändamål.[14] Förutsättningarna för att lämna ut uppgifter för statistikändamål är desamma avseende bedömningen av risken för skada eller men som för forskningsändamål. I samband med att undantaget infördes påpekade regeringen i propositionen att ett utlämnande enligt detta undantag i realiteten endast blir aktuellt vid utlämnande till myndigheter som också tillämpar statistiksekretess.[15] I praktiken behöver därför den utlämnande myndigheten undersöka vilket sekretesskydd uppgifterna skulle få hos den mottagande myndigheten. En förutsättning för att en mottagande myndighet ska kunna tillämpa statistiksekretess är att den har en sådan särskild statistikverksamhet som avses i 24 kap. 8 § offentlighets- och sekretesslagen eller att den bedriver en därmed jämförbar undersökning enligt samma bestämmelse (se avsnitt 3.2).

Ett exempel på tillämpning av undantaget för statistikändamål är att Brottsförebyggande rådet har gett Statistiska centralbyrån uppgifter om ungdomar i åldern 15–17 år som har lagförts för ett brott under ett kalenderår. Uppgifterna samkörs med andra uppgifter från Statistiska centralbyrån inom ett uppdrag att beräkna indikatorer för att följa barns levnadsvillkor. Redovisning sker i en databas hos Barnombudsmannen.

Det kan noteras att det i artikel 3 i EU-förordningen (223/2009) om den europeiska statistiken finns en definition av *användning för statistiska ändamål*. Begreppet definieras där som användning uteslutande för att utveckla och framställa statistiska resultat och analyser.

Uppgifter som inte är direkt hänförliga till en enskild

Förutom de ovan beskrivna undantagen från den absoluta sekretessen undantas också uppgifter som inte genom namn, annan identitetsbeteckning eller liknande förhållande är direkt hänförliga till den enskilde. Med liknande förhållande avses telefonnummer, anställningsnummer, registreringsnummer för fordon, fastighetsbeteckning eller dylikt.

Enligt förarbetena tillkom detta tredje undantag med anledning av behovet av detaljerad statistik.[16] Det är alltså tillåtet att publicera detaljerad statistik, även med röjanderisk (men

[12] Prop. 1997/98:108, s. 49.

[13] Prop. 2013/14:162, s. 11.

[14] Prop. 1994/95:200, s. 38.

[15] Prop. 1994/95:200, s. 28, 38 och 57.

[16] Prop. 1979/80:2, del A, s. 264.

bara med uppgifter som inte är direkt hänförliga till den enskilde), om det står klart att uppgifterna kan röjas utan att den enskilde som avses eller någon närstående till denne lider skada eller men.

Undantaget är även tillämpligt för utlämnande av anonymiserade (se begreppsförklaring i kapitel 13) mikrodata, som alltså kan lämnas ut för skilda ändamål, till exempel utredningar av effektivitet i myndigheters verksamhet, om det står klart att utlämnandet kan ske utan risk för skada eller men.

Uppgifter som rör dödsorsak eller dödsdatum

Genom en lagändring som trädde i kraft den 1 juli 2008 tillkom ett fjärde undantag från huvudregeln om absolut sekretess. Detta undantag gäller endast för uppgifter som avser en avliden och som rör dödsorsak eller dödsdatum. Sådana uppgifter får lämnas ut till ett nationellt eller regionalt kvalitetsregister som avses i patientdatalagen (2008:355) om det står klart att uppgifterna kan röjas utan att den enskilde eller någon närstående till denne lider skada eller men. Syftet är att Socialstyrelsen, som ansvarig för dödsorsaksregistret, ska kunna lämna ut uppgifter från detta till de nämnda registren.

Förutsättning för utlämnande enligt undantagen – begreppen skada och men

En förutsättning för utlämnande enligt ovanstående fyra undantag är att det står klart att uppgifterna kan röjas utan att den enskilde eller någon närstående lider skada eller men. Ett sådant så kallat omvänt skaderekvisit innebär att det är presumtion för sekretess, det vill säga att sekretess gäller tills det är visat att skada eller men inte kan uppstå. I förarbetena till 24 kap. 8 § offentlighets- och sekretesslagen anges att eftersom uppgifter som omfattas av statistiksekretess har ett starkt sekretesskydd genom den absoluta sekretessen innebär ett utlämnande som medför att uppgifterna kommer att omfattas av ett svagare sekretesskydd i sig en risk för skada. Även om användningen av uppgifterna hos mottagaren inte medför skada kan det svagare sekretesskyddet hos mottagaren medföra att uppgifterna lämnas vidare på ett sätt som medför skada.[17] Det betyder att den som gör bedömningen av om en uppgift kan lämnas ut har ett begränsat utrymme för sin bedömning. I praktiken innebär det att uppgiften inte kan lämnas ut utan kännedom om mottagarens identitet och avsikter med uppgifterna.[18] Se vidare avsnitt 6.3. Med begreppet ”står klart” menas att den som lämnar ut uppgifterna har gjort bedömningen att enskild inte lider skada eller men om uppgifterna blir kända.[19]

Att det ska stå klart att uppgiften kan röjas utan att någon lider skada eller men lägger ett stort ansvar på den utlämnande myndigheten. Myndigheten ska kunna motivera varför uppgiften är möjlig att lämna ut – finns det någon grund för tvekan vid bedömningen av risk för skada eller men ska uppgiften inte lämnas ut.

Skada eller men kan uppstå från åtgärder som i regel upplevs som en påfallande nackdel för den berörde även om åtgärden i sig är rättsenlig. Med begreppet *skada* avses ekonomisk skada, medan begreppet *men* avser integritetskränkningar av olika slag. Av det följer att en

[17] Prop. 2013/14:162, s. 12.

[18] Prop. 1979/80:2, del A, s. 82.

[19] I prop. 1979/80:2 har detta formulerats enligt följande: ”Endast om man med en viss grad av sannolikhet kan utgå från att det är ofarligt att lämna ut en viss uppgift viker sekretessen” (prop. 1979/80:2, del A, s. 79) och ”känna god säkerhet för att röjande inte leder till skada eller men” (prop. 1979/80:2, del A, s. 157). Se även von Essen, *Biobanksforskning – forskares möjligheter att få tillgång till vävnadmaterial och personuppgifter*, Förvaltningsrättslig tidskrift 2003, s. 197–214, särskilt s. 204.

juridisk person bara kan drabbas av skada, medan en fysisk person kan drabbas av såväl skada som men.

Med integritetskränkningar avses till exempel att någon blir utsatt för andras missaktning om personliga förhållanden blir kända, men det kan också röra sig om spridning av uppgifter som den enskilde upplever som obehagliga att andra känner till. Utgångspunkten vid bedömning av risken för men är den subjektiva upplevelsen hos den som riskerar att drabbas. Den bedömningen kan dock behöva korrigeras med utgångspunkt i värderingar i samhället generellt. Det betyder att även om en person i det specifika fallet upplever sig drabbas av men, så innebär det inte säkert att det bedöms som men vid en sekretessprövning.[20] Att hänsyn ska tas till gängse värderingar i samhället gör också att bedömningen av vad som är men kan behöva ändras över tid.

Enbart möjligheten för en angripare att identifiera en enskild bakom en statistikuppgift behöver inte generellt innebära men genom integritetskränkning, men kan göra det under vissa omständigheter (se avsnitt 6.1).

Vad som ska anses vara men får därför ytterst avgöras i det specifika fallet mot bakgrund av uppgifternas karaktär och omständigheterna i övrigt, se vidare kapitel 6. Ledning kan också fås från rättspraxis.

Vem som är att betrakta som *närstående* finns inte definierat i offentlighets- och sekretesslagen. I likhet med vad som gäller för begreppet men kan detta också komma att förändras över tid.[21] Innebörden av begreppet närstående har behandlats av regeringsrätten, som ansåg att en väninna till en avliden person inte var en närstående.[22]

3.4 Exempel på rättsfall

I det följande ges några exempel på rättsfall. De är inte på något sätt avsedda att ge en heltäckande bild av praxis på området, men kan klargöra hur sekretessbestämmelsen har tillämpats när det gäller utlämnande av uppgifter.

Personliga förhållanden med mera

Högsta förvaltningsdomstolen tog 2012 upp ett mål där en journalist begärde att få ta del av uppgift om vilka tingsrätter som motsvarade siffrorna 1–53 i en tabell om utdömda påföljder för unga lagöverträdare som fanns i en rapport från Brottsförebyggande rådet. Uppgifterna omfattades enligt Brottsförebyggande rådet av sekretess med stöd av 24 kap. 8 § offentlighets- och sekretesslagen och 7 § offentlighets- och sekretessförordningen (2009:641).

De som i detta fall skulle beröras av statistiksekretessen var dels domare som tjänstgjorde vid de aktuella tingsrätterna, dels ungdomar som dömts till påföljd för brott.

Uppgift om namnet på enskilda domstolar skulle kunna ge information om vilka påföljder enskilda domare dömt ut. En offentlig befattningshavares åtgärder i tjänsten bedömdes dock inte röra dennes personliga eller ekonomiska förhållanden. På denna grund kunde uppgifterna därför inte omfattas av sekretess enligt bestämmelsen i 24 kap. 8 § offentlighets- och sekretesslagen.

Uppgift om namnet på tingsrätten kunde däremot vara indirekt hänförlig till enskilda dömda ungdomar och därmed omfattas av statistiksekretessens tillämpningsområde. Vid

[20] Regeringsrättens årsbok 1994, ref. 91.

[21] Prop. 1979/80:2, del A, s. 168.

[22] Regeringsrättens årsbok 2009, ref. 17.

prövningen ansågs emellertid att uppgifterna om tingsrätternas namn kunde lämnas ut. I bedömningen beaktade domstolen dels att journalistens syfte var att granska enskilda domstolars rättstillämpning och inte att undersöka vem som dömts till viss påföljd, dels att det inte gällde någon sekretess på annan grund för de uppgifter om identiteten på dömda ungdomar som möjligen skulle kunna uppenbaras vid ytterligare efterforskningar efter ett utlämnande. Uppgift om vilka tingsrätter som motsvarade siffrorna i tabellen skulle därför lämnas ut till journalisten.[23]

Uppföljning av vård utgör inte forskning

En läkare som begärt ut uppgifter från Socialstyrelsens dödsorsaksregister för uppföljning av viss medicinsk behandling fick avslag på sin begäran då uppgifterna skulle användas för kvalitetsuppföljning och inte för forskning. Då det i flera sammanhang gjorts åtskillnad mellan forskning å ena sidan och kvalitetssäkring, uppföljning och utvärdering å andra sidan, bör denna åtskillnad upprätthållas även vid tillämpningen av sekretesslagstiftningen. Eftersom läkaren själv uppgett att uppgifterna skulle användas för uppföljning, kunde dessa inte lämnas ut med stöd av forskningsundantaget. Omständigheten att uppgifterna kunde komma att användas för forskning längre fram ändrade inte bedömningen.[24]

Domsdatum för sexualbrottsmål kan inte lämnas ut utan risk för men

Klagande begärde hos Brottsförebyggande rådet ut domsdatum för tingsrättsdomar i sexualbrottsmål. Kammarrätten avslog överklagandet. Domstolen fann att domsdatum utgör uppgift som avser enskilds personliga förhållanden och som kan hänföras till enskild och att uppgiften därmed omfattas av sekretess enligt 9 kap. 4 § sekretesslagen (1980:100)[25]. Vidare fann domstolen att domsdatum är en uppgift som inte är direkt hänförlig till enskild och därför får lämnas ut om det står klart att detta kan ske utan att den som uppgiften rör eller någon honom närstående lider skada eller men. Enligt domstolens mening stod det inte klart att uppgifterna kunde lämnas ut utan risk för skada eller men.[26]

Utlämnande av statistik baserad på ett fåtal elever

Skolverket fick en begäran om att lämna ut fullständig statistik om slutbetyg på skolnivå, eftersom uppgifter som baseras på färre än tio elever i cellen inte publiceras av sekretesskäl. Enligt Skolverkets bedömning står det inte klart att dessa uppgifter, som i vissa fall bedöms indirekt hänförliga till enskilda, kan lämnas ut utan att eleven eller någon närstående lider skada eller men. I och med det avslog Skolverket begäran om utlämnande av data. Kammarrätten bedömde dock att de begärda uppgifterna varken direkt eller indirekt kunde hänföras till enskilda elever. Sekretess gällde därför inte.[27]

Skolverket fick därefter en begäran gällande motsvarande statistik för ett senare läsår. Denna statistik var inte publicerad vid den första begäran. Skolverket gjorde samma bedömning som vid den tidigare begäran om utlämnande och lämnade inte ut uppgifterna. I beslutet beskrev Skolverket att resultaten vid små skolor exempelvis kan gälla en ensam elev i årskurs 9. Med kunskap om vid vilken skola eleven finns är det i dessa fall möjligt att utläsa elevens resultat avseende slutbetyget. Beslutet överklagades. I yttrande till kammarrätten poängterade Skolverket att ett syfte med statistiksekretessen är att säkra

[23] Högsta förvaltningsdomstolen 2012, ref. 64.

[24] Regeringsrättens årsbok 2004, ref. 9.

[25] Denna bestämmelse motsvarade nuvarande 24 kap. 8 § offentlighets- och sekretesslagen (2009:400).

[26] Kammarrätten i Stockholm, 2002, mål nr 7975-02.

[27] Kammarrätten i Stockholm, mål nr 3233-13, dom 2013-07-19.

kvaliteten i den slutliga statistiken. Genom att bevara enskildas förtroende för sekretessen genom hela produktionsprocessen tillförsäkras kvaliteten på statistiken. Kammarrätten ansåg att det inte stod klart att begärda uppgifter kunde röjas utan att den enskilde eller någon närstående till denne lider men. Överklagandet avslogs därför.[28]

Indirekt utsläppsstatistik

Malmö kommun begärde ut statistik från Statistiska centralbyrån, utifrån vilken det var möjligt att göra antaganden och beräkningar om utsläpp. Begäran avslogs, vilket överklagades. Kammarrätten bedömde att de begärda uppgifterna omfattas av sekretess enligt 24 kap. 8 § offentlighets- och sekretesslagen. Frågan i målet var om det fanns skäl att bryta sekretessen med stöd av 10 kap. 5 § första stycket offentlighets- och sekretesslagen. Sekretessen kan brytas om det efter en intresseavvägning framstår som uppenbart att de begärda uppgifterna har sådan betydelse från miljösynpunkt att intresset av allmän kännedom om uppgifterna har företräde framför det intresse som sekretessen ska skydda. Kammarrätten gjorde bedömningen att de begärda uppgifterna inte kunde anses utgöra uppgifter om utsläpp i miljön i den mening som avses i 10 kap. 5 § offentlighets- och sekretesslagen. Den omständigheten att det utifrån uppgifterna var möjligt att göra antaganden och beräkningar avseende mängden utsläpp som har genererats till följd av bränsleförbrukningen medförde ingen annan bedömning.[29] Malmö kommun hade därför inte rätt att ta del av uppgifterna.

Uppgifter ur PISA-undersökning

Målet gällde en journalist som av Mittuniversitetet begärde att få ut en förteckning över vilka skolor som deltagit i PISA 2012. Syftet med begäran var enligt journalisten att granska urvalsprocessen för PISA-undersökningen och inte att ”hänga ut enskilda elever i en tidningsartikel”. Journalisten hänvisade till pressetiska regler och att den tidning som han arbetade för var noga med att följa den nu upphävda personuppgiftslagen.

Mittuniversitet avslog journalistens begäran. Nyckeln med skolnamn skulle göra det möjligt att med begränsad arbetsinsats matcha såväl rektorer som elever mot databaserna med känsliga personuppgifter och känsliga personliga ställningstaganden. Enligt Mittuniversitetet förelåg en uppenbar risk för skada för både de rektorer och de elever som besvarat enkäten.

Kammarrätten prövade begäran utifrån bestämmelserna i 24 kap. 8 § offentlighets- och sekretesslagen och 7 § offentlighets- och sekretessförordningen, som säger att internationella skolundersökningar som görs av Skolverket utgör sådan särskild verksamhet som avses i 24 kap. 8 § offentlighets- och sekretesslagen. Uppgifterna som journalisten begärde ut om vilka skolor som deltagit i PISA-undersökningen kunde inte genom namn, annan identitetsbeteckning eller liknande förhållande direkt hänföras till någon enskild person. För dessa uppgifter gällde således sekretess med ett omvänt skaderekvisit. En uppgift om en skolas namn skulle i förening med andra uppgifter kunna leda till att en enskild person som deltagit i undersökningen gick att identifiera. Kammarrätten bedömde därför, även med beaktande av det journalisten angett som stöd för sin begäran, att det inte står klart att uppgiften om vilka skolor som deltagit i PISA-undersökningen kunde lämnas ut utan att en enskild eller dennes närstående lider skada eller men. Överklagandet avslogs. Domen[30] har vunnit laga kraft.

[28] Kammarrätten i Stockholm, mål nr 6608-13, dom 2013-12-16.

[29] Kammarrätten i Stockholm, mål nr 1031-14, dom 2014-07-10.

[30] Kammarrätten i Sundsvall, mål nr 498-14, dom 2014-09-03.

3.5 Internationella aspekter

Statistik är en internationell verksamhet, både som vetenskap och som infrastruktur för informationsförsörjning i samhället. Statistikens villkor och ambitioner i omfattning, kvalitet, datakällor och annat styrs delvis genom internationella organ, främst EU och FN. Internationella regler om skydd för uppgiftslämnarnas integritet finns sedan länge. Som beskrivs i avsnitt 1.2 stöds statistiksekretessen såväl i FN:s grundläggande principer för officiell statistik som i EU:s statistikförordning 223/2009 och Riktlinjer för europeisk statistik.

När det gäller svensk lagstiftning står följande i 8 kap. 3 § offentlighets- och sekretesslagen:

En uppgift för vilken sekretess gäller enligt denna lag får inte röjas för en utländsk myndighet eller en mellanfolklig organisation[31]*, om inte*

1. utlämnande sker i enlighet med särskild föreskrift i lag eller förordning, eller
2. uppgiften i motsvarande fall skulle få lämnas ut till en svensk myndighet och det enligt den utlämnande myndighetens prövning står klart att det är förenligt med svenska intressen att uppgiften lämnas till den utländska myndigheten eller den mellanfolkliga organisationen.

Svensk lagstiftning är alltså i grunden restriktiv när det gäller internationella utlämnanden. Paragrafen ovan reglerar när uppgifter får röjas till en utländsk myndighet eller en mellanfolklig organisation, men berör inte frågan om röjande till enskilda utbildningsanordnare (icke-statliga lärosäten) eller andra enskilda.

Det första undantaget avseende internationella utlämnanden

Det första undantaget i texten ovan avser utlämnande enligt särskild föreskrift i lag eller förordning. Ett grundläggande regelverk vad gäller europeisk statistik är EU:s statistikförordning 223/2009. Som förordning från EU är den bindande och direkt tillämplig i alla medlemsstater och kan åberopas likt en nationell lag. Sekretessbelagda uppgifter får alltså sändas från Sverige till kommissionen, vanligen till EU:s statistikkontor Eurostat, eller till Europeiska centralbanken *om överföringen är nödvändig för att på ett effektivt sätt utveckla, framställa och sprida europeisk statistik eller för att förbättra den europeiska statistikens kvalitet* (från förordningens artikel 21).

Det allmänna syftet med EU-förordningen är att stärka samarbetet och samordningen mellan de myndigheter som medverkar till att bland annat framställa och sprida europeisk statistik. Förordningens artikel 20–26 berör statistisk konfidentialitet; begreppet kan tolkas som i praktiken likvärdigt med statistiksekretess i svensk mening. I dessa artiklar regleras utöver överföring av konfidentiella uppgifter (artikel 21) bland annat även skyddet för dessa uppgifter inom Eurostat (artikel 22).

Utöver statistikförordningen finns för olika statistikområden särskilda EU-förordningar. Ibland ges detaljerade regler, till exempel om tröskelvärden i röjandekontroll, i sådana förordningar. Kompletterande regler kan vara utfärdade av EU-kommissionen och dess organ såsom Eurostat, och eventuellt även överenskomna informellt genom ”gentlemen’s agreement” med myndighetschefer i medlemsländerna eller vid arbetsmöten och dylikt. Det är viktigt att för varje statistikmaterial klarlägga vilka regler som gäller och vara tydlig i kommunikationen med Eurostat.

[31] En mellanfolklig organisation är detsamma som en internationell organisation, det vill säga en organisation vars verksamhet eller medlemsbas går över nationsgränser.

Flera svenska myndigheter sänder tabelldata (makrodata) till Eurostat och andra internationella organ med hjälp av Eurostats verktyg Edamis. Det förekommer även överföring av mikrodata, vanligen krypterade, via Edamis. Överföring via e-post förekommer i undantagsfall, men bör undvikas eftersom det är förenat med större säkerhetsrisker än med överföring via Edamis. Efterfrågan på mikrodata i internationella sammanhang kan komma att öka.

Leveranser av mikrodata

Leveranser av mikrodata från medlemsländer till EU bygger vanligen på att Eurostat genomför all röjandekontroll. Mikrodata kan anonymiseras eller avidentifieras innan de sänds till Eurostat, men är i övrigt statistiskt oskyddade. Länderna kan dock beskriva vilka riskceller som kan uppstå vid aggregering. Enligt artikel 22 hanterar Eurostat sedan dessa data med sekretess och genomför röjandekontroll (vid behov) inför publiceringen av statistiken.

Leveranser av makrodata

Leveranser av makrodata till EU förutsätter att den aktuella myndigheten i medlemslandet har genomfört en röjandekontroll. Makrodata (tabeller) kan sändas till Eurostat på några alternativa sätt:

- Data från identifierade riskceller tas inte med i leveransen, det vill säga inga sekretessmarkerade data ingår. (I somliga fall, till exempel för en del urvalsbaserade individundersökningar, finns inga riskceller, varför alla data kan levereras.)
- Alla data är med i leveransen, men riskceller är sekretessmarkerade.
    - Sekretessmarkeringarna förklaras, till exempel för primär- respektive sekundärundertryckning.
    - Sekretessmarkeringarna förklaras inte.

Vid Eurostats publicering tas inte riskcellerna med. Eurostat röjandekontrollerar de EU-aggregat som tillkommer i publiceringen. Det första alternativet ovan, där riskceller inte levereras, innebär att inga EU-aggregat kan tas fram och leder därför till informationsförlust på EU-nivå. För det andra alternativet ovan gäller att ju mer information Eurostat har att tillgå om sekretessmarkeringar med mera, desto effektivare kan informationsförlusten minimeras vid skyddandet. Eurostat tillämpar i princip minst lika stränga regler, till exempel nivåer på tröskelvärden, för röjandekontroll som de enskilda länderna. De data som sänts från medlemsländer kan inte byta status när det gäller vad som är konfidentiellt (belagt med sekretess) och inte. Icke konfidentiella data förutsätts publicerade på nationell nivå, och då kan inte Eurostat använda dessa för sekundärundertryckning, utan får använda EU-aggregat för detta. För information om Eurostats egna rekommendationer om röjande-kontroll, se European Commission, Eurostat (2014). Som it-verktyg vid röjandekontrollen används bland annat τ-ARGUS, se vidare avsnitt 11.1.

Det finns särskilda regler i artikel 21 i EU:s statistikförordning om vidarebefordran av data från Eurostat till Europeiska centralbanken. Eurostat kan även sända vidare mikro- eller makrodata till andra internationella organ, men först efter överenskommelse med medlemslandet i fråga. Från exempelvis Statistiska centralbyrån förekommer leverans av data direkt eller indirekt till bland annat FN, Internationella valutafonden, Världsbanken, OECD, International Energy Agency (IEA) och Bank for International Settlements (BIS).

Det andra undantaget avseende internationella utlämnanden

Det andra undantaget i den ovan citerade texten (se inledningen av detta avsnitt) blir aktuellt att pröva om det första undantaget inte kan åberopas. Det andra undantaget gäller om uppgiften i motsvarande fall skulle få utlämnas till svensk myndighet och det står klart

att det är förenligt med svenska intressen att uppgiften lämnas till den utländska myndigheten eller den mellanfolkliga organisationen (avser alltså inte enskilda utbildningsanordnare). En sekretessprövning ska göras på vanligt sätt för ett utlämnande. Det ska dessutom stå klart att uppgiftslämnandet är förenligt med svenska intressen. Vid behov ska samråd ske med Utrikesdepartementet (se Lenberg m.fl. 2010, 8:3:1).

Vid ett utlämnande till exempelvis ett utländskt universitet, har den utlämnande myndigheten att pröva om uppgifterna kan skyddas av sekretess hos mottagaren, antingen genom det andra landets interna regler eller genom en förbindelse om sekretess (konfidentialitet). Det måste också beaktas att det är förbjudet att lämna ut personuppgifter till ett tredje land, det vill säga land utanför EU, som inte har en skyddsnivå som motsvarar den inom EU. Detta framgår av artikel 44-45 i EU:s dataskyddsförordning. Vidare gäller enligt 21 kap. 7 § offentlighets- och sekretesslagen sekretess för personuppgift, om det kan antas att uppgiften efter ett utlämnande kommer att behandlas i strid med EU:s dataskyddsförordning, i den ursprungliga lydelsen, eller lagen (2018:218) med kompletterande bestämmelser till EU:s dataskyddsförordning.

Forskares tillgång till mikrodata

Tillgång till sekretessbelagda uppgifter i vetenskapligt syfte behandlas i EU:s statistikförordning 223/2009, artikel 23. Bland annat framgår att Eurostat inte kan lämna ut uppgifter som överförts från en nationell myndighet till forskare utan uttryckligt tillstånd från den myndigheten. Ytterligare föreskrifter finns i EU:s förordning 831/2002.

Från Eurostat har det funnits en vision på lång sikt om europeiska forskares tillgång till mikrodata från hela EU och om en fri överföring av statistiska data mellan EU:s medlemsländer, som skulle öppna för en långtgående arbetsfördelning mellan statistikmyndigheter i olika länder. Det har då talats om en Schengen-ansats med något liknande fri rörlighet för data inom EU. EU:s statistikförordning 223/2009 behöver då ändras för att detta skulle kunna förverkligas. Inriktningen nu i det europeiska statistiksamarbetet när det gäller forskares tillgång till mikrodata är snarare ett mindre långtgående samarbete, där utlämnande görs efter prövning i varje enskilt fall. Beträffande fri överföring av data mellan EU:s medlemsländer för statistikproduktion, pågår ett pilotprojekt för att testa möjligheterna inom handelsstatistiken. För att upprätthålla sekretessen vid utlämnande av mikrodata till forskare kan det komma att behövas olika insatser för att skydda mikrodata mot röjande. Se vidare kapitel 10 om röjandekontroll av mikrodata.

4 Metodmässiga förutsättningar

Detta kapitel visar vilka metodmässiga förutsättningar som kan behöva beaktas vid röjandekontroll. Kapitlet beskriver olika typer av objekt, variabler samt tabeller, diagram och kartor, och dessas betydelse för skadeprövning och skydd av datamaterial. Förhållningssätt anges för tabeller som baseras på uppgifter från ett totalräknat material och för tabeller som inte gör det utan baseras på urvalsdata. Olika scenarier beskrivs för utformning av till exempel frekvens- och magnitudtabeller. Dessutom behandlas betydelsen av olika osäkerhetskällor såsom bortfall.

Metodmässiga förutsättningar för länkade (sammankopplade) tabeller redovisas. För enskilda tabeller som bedöms säkra med avseende på röjanderisken kan det ändå vara möjligt att röja information om enskilda objekt om två eller flera tabeller länkas.

4.1 Objekttyp

En *objekttyp* är ett slag av objekt som förekommer i statistik och statistikframställning. Till exempel är ”individer” som begrepp en objekttyp, medan varje individ för sig är ett objekt. På samma sätt är ”företag” som begrepp en objekttyp, medan varje företag för sig är ett objekt. Andra exempel på objekttyper är hushåll, arbetsställen, fastigheter, lantbruksföretag och händelser såsom varutransporter och anmälda brott.

I arbetet med röjandekontroll är det primärt enskilda fysiska och juridiska personer som ska skyddas, alltså objekttyperna individer respektive företag/organisationer. Även andra objekttyper kan behöva skydd genom att objekt kan vara relaterade till andra objekt, såsom beskrivs i nästa underavsnitt.

I en röjandekontroll är det viktigt att ha med sig tydliga definitioner av populationen (eller populationerna) och de objekt som ingår där. Objekttypen har betydelse för valet av metoder för riskbedömning och skyddande.

Karakteristiskt för populationer av företag är att objekten, företagen, kan vara av relativt olika storlek sinsemellan. De största företagen i en bransch kan vara få till antalet och ”sticka ut” så att de lätt röjs i statistiktabeller. Det kan delvis vara enklare att bedöma skaderisken för företag än för individer. Det är främst bedömningen av men som kan vara svår. Användning av samtycke till att efterge sekretess är främst relevant i företagsundersökningar, se vidare kapitel 8.

Relaterade objekt

Om det finns kopplingar mellan olika objekt kan de kallas *relaterade objekt*. I statistik kan det gå att utläsa information om en annan typ av objekt än den som statistiken avser. Exempelvis kan lönestatistik över individer leda till röjande av företag där de arbetar.

Genom att objekt kan vara relaterade så kan även objekt av andra typer än individer och företag behöva skyddas, för att inte enskilda individer eller företag ska röjas. Exempel: I hamnstatistik skulle dels säljare och köpare av varor, dels fraktföretag och hamnföretag kunna identifieras. Betygsstatistik för skolor kan avslöja enskilda elever; exempelvis röjer uppgiften att 0 elever har betyg A i ett visst ämne på en skola att en enskild individ inte har betyget A.

Ytterligare exempel: Uppgifter om anmälda brott kan om de är detaljerade avslöja misstänkta gärningsmän och brottsoffer. Försäkringskassan publicerar statistik över tandvårdsåtgärder på patienter. Publicerad statistik är till exempel antal besök och antal åtgärder men även medianpriser, vilket säger något om tandläkarna.

Blandade objekttyper

Statistiken kan befatta sig med objektmängder som är sammansatta av olika objekttyper. Detta ger *blandade objekttyper*. Bilinnehav i bilregistret utgör ett exempel på blandade objekttyper, med både företag och individer som bilägare.

Hierarkiska objekt

För *hierarkiska objekt* kan varje objekt föras till en grupp (ett kluster) av flera objekt i en hierarkisk struktur med två eller flera nivåer. Gruppindelningen kan definieras efter andra kriterier än tabellens bakgrundsvariabler och redovisningsgrupper. Några exempel är arbetsställen som kan grupperas inom företag och skolelever som kan grupperas inom skolor i skoldistrikt. Detta medför att olika objekt som tillhör samma grupp kan finnas utspridda över flera olika celler i tabellen. Detta kan utgöra en ökad risk för röjande via koalitioner (samarbetsgrupper). Finns behov att skydda objekt på flera hierarkiska nivåer kan röjandekontroll tillämpas på alla de nivåerna.

Ett exempel av detta slag är så kallade *holdings*, såsom koncerner med dotterföretag i olika län. En marginalcell för riket kan då ha flera bidragsgivare (företag) från samma koncern. Om då röjanderisken bedöms utifrån hur dominerande de två största företagen är, kan bedömningen bli missvisande om båda företagen hör till samma koncern. Koncernen kan då eventuellt röja det tredje största företaget. Dessutom kan det vara risk att koncernen röjs. Bättre kan vara att utgå från koncernerna som objekt då röjanderisken ska bedömas, även om det kan bli arbetskrävande att uppdaterat hänföra rätt företag till rätt koncern.

4.2 Totalräknade data eller urvalsdata

Förutsättningarna skiljer sig delvis mellan å ena sidan tabeller som är baserade på uppgifter från totalräknade material, till exempel registerdata, och å andra sidan tabeller som inte är det.

Om tabellen inte är baserad på ett totalräknat material utan på ett urval så ger det i regel en mindre risk för röjande. Detta eftersom användaren normalt inte vet vilka objekt i populationen som ingår i urvalet. På så sätt ställer totalräknade och urvalsbaserade data något olika krav på röjanderiskbedömning och skyddsmetoder.

Här säger ”urval” inte något om urvalsmekanismen eller huruvida det är ett sannolikhetsurval.

I många urvalsundersökningar av företag förekommer totalundersökta strata. Exempelvis kan tänkas att uppgifter samlas in från alla företag med fler än 500 anställda plus ett slumpurval av övriga företag. Risken för röjande i den totalundersökta delen kan då inte förutsättas minska av att även urval ingår.

En vanlig form av icke-sannolikhetsurval är så kallade cut-off-urval. För ett sådant urval exkluderas avsiktligt en delmängd av populationen från urvalsdragningen. Förfarandet är vanligt för företagsundersökningar där de minsta företagen inte behöver lämna uppgifter. Även denna typ av urval kan i princip anses minska röjanderisken, när en kvalificerad bedömning av de aktuella förutsättningarna stödjer detta.

4.3 Variabler

I det följande beskrivs några olika variabeltyper med olika betydelse för röjanderiskerna.

Identifierare

Uppgifter som är väsentligt särskiljande, inte nödvändigtvis unikt, för objekt i populationen kallas direkt identifierande eller *identifierare*. Dit hör uppgifter som person- eller organisationsnummer, namn, adress, telefonnummer och lägenhetsnummer.

Nyckelvariabler

Uppgifter som inte är identifierare, men som ändå kan användas för att hänföra uppgifter till objekt, kallas indirekt identifierande eller *nyckelvariabler*. Nyckelvariablerna utgörs vanligen av lättillgänglig och allmänt känd information och används ofta i en tabell som bakgrundsvariabler, det vill säga de är med och spänner upp tabellen (kategoriserar rader och kolumner i tabellen) och definierar tabellens celler. Exempel på vanligt förekommande nyckelvariabler är ålder, kön och bostadsort, men vilka variabler som kan tjäna som nyckelvariabler varierar mellan undersökningar.

Identifiering står för att på något sätt urskilja vilket enskilt objekt i en population som ligger bakom en uppgift i ett statistikmaterial. En identifiering kan ske på olika sätt, såsom direkt med identifierare, eller indirekt med nyckelvariabler eller andra uppgifter.

Nycklar

En kombination av flera nyckelvariabler som en angripare kan använda för att hänföra uppgifter till objekt kallas här *nyckel*. Detta kan exempelvis ske genom att objektet hänförs till en redovisningsgrupp eller cell i tabellen. Även om nyckelns variabler inte var för sig riskerar att röja objektet, så kan kombinationen av variabler röja ett objekt.

Målvariabler

Målvariabler definieras här som de variabler som inte är identifierare eller nyckelvariabler men som utgör målinformationen för den som försöker röja information om enskilda objekt (angriparen). Det behöver inte nödvändigtvis vara en variabel som i någon mening är känslig, utan det räcker med en variabel som utgör ny information för angriparen och som inte används vid själva identifieringen av enskilda objekt.

En variabel kan vara indirekt identifierande i ett sammanhang och målvariabel i ett annat. Ett exempel på det är inkomst, som kan vara målvariabel i en tabell men som klassindelad även kan användas som bakgrundsvariabel i en annan tabell.

Att utläsa en egenskap (ett värde på en målvariabel) för ett enskilt objekt i en population kallas *attribuering*. En attribuering kan ske utifrån en identifiering eller på annat sätt.

Uppräkningstal och andra tekniska variabler

Uppräkningstal eller designvikter är tekniska variabler som kan ingå i mikromaterial för att möjliggöra uppräkning till statistiktabeller. Sådana och likartade variabler kan medföra ökad risk för röjande i avidentifierade eller anonymiserade mikromaterial (vilket behandlas i kapitel 10). De kan nämligen bland annat visa urvalssannolikheter, eller ge information om nyckelvariabler som de är härledda ur. Till exempel, en urvalssannolikhet lika med 1 kan låta ana ett företag som är bland de största i sin bransch.

Även i statistiktabeller kan uppräkningstalen vara förenade med röjanderisk, såsom i små redovisningsgrupper där det ur tabellen eventuellt kan gå att gissa uppräkningstal som är lika stora för flera objekt i populationen.

Variabelegenskaper

I röjandekontrollen behöver i princip tabellens alla variabler gås igenom med avseende på dels den roll (till exempel som nyckelvariabler) de spelar i möjliga röjanden, dels deras

egenskaper såsom datatyp och klassindelningar. Några situationer som är vanligt förekommande och påkallar särskilda hänsyn tas upp nedan.

Varierande skaderisk

Kategoriska målvariabler kan ha en varierande skaderisk. Risken för skada eller men vid ett röjande kan variera mellan de kategorier eller klasser som svarar mot värdena på variabeln. Exempel: För en variabel som beskriver sjukfrånvaro medför uppgiften ”ingen frånvaro” troligen en mindre risk för skada eller men om den röjs än uppgifter om andra kategorier för grad av sjukfrånvaro.

Hierarkiska klassifikationer

Hierarkiska klassifikationer har speciell problematik. Det gäller till exempel Standard för svensk näringsgrensindelning (SNI) och Standard för svensk produktindelning efter näringsgren (SPIN). Det kan vara möjligt att dra slutsatser om en klass på någon siffernivå med ledning av aggregerade värden på en högre nivå. När hierarkiska klassifikationer ingår i tabellen ska varje överordnad nivå i hierarkin betraktas som en känd marginal (sammanslagning av de underordnade klasserna) som kan användas för att härleda cellvärden på en lägre nivå. Om de högre nivåerna inte publiceras i samma tabell ska bedömningarna ändå utgå från att de kan vara kända för användaren, exempelvis för att de används i någon annan publicerad tabell.

Geografiska variabler

Geografiska variabler om belägenhet såsom kommun behöver uppmärksammas eftersom de i regel underlättar identifiering av enskilda objekt i populationen. Det kan vara enklare för angriparen att hitta och isolera ett geografiskt område och där finna och identifiera enskilda populationsobjekt än att isolera en delmängd av populationen baserat på en annan egenskap såsom inkomst.

Ofta redovisas de geografiska uppdelningarna i en hierarkisk struktur, till exempel län, kommun och församling. Vidare finns samma geografiska indelningar ofta i annan statistik. Vägledande i bedömningen av geografiska variabler är bland annat storleken på de åsyftade geografiska områdena, men det går inte att ange en exakt gräns för hur små geografiska områden som kan redovisas, utan skadeprövningen beror även på andra faktorer.

4.4 Typ av tabeller, diagram eller kartor

Frekvenstabeller

En grundläggande uppdelning görs mellan huvudsakligen två typer av tabeller. Tabelltyperna medför olika röjandescenarier och därmed även olika metoder för riskbedömning och tabellskydd.

Den första typen är frekvenstabeller, vilket är tabeller som för cellerna redovisar antalet objekt som faller i respektive cell. Frekvenstabeller redovisar antalsfördelningar för värden på kategorivariabler.

Ofta redovisas andelar (relativa frekvenser) i stället för antal (absoluta frekvenser). I frekvenstabeller med andelar jämte antalsuppgifter i marginalerna går det att återfå antalen i cellerna. Det blir alltså samma situation som med antal i cellerna.

Magnitudtabeller

Den andra typen av tabeller är magnitudtabeller (kvantitativa tabeller). Det är tabeller som i princip spänns upp av kategorivariabler och där cellerna redovisar värdena på en statistisk storhet, såsom summan av en kvantitativ variabel över de objekt som ingår i cellen.

Magnitudtabeller grundas alltså främst på kvantitativa data. Exempel på magnitudtabeller är vidare tabeller som redovisar medelvärden eller kvoter mellan summor.

Andra varianter av magnitudtabeller förekommer. En variant är tabeller för variabler som kan anta negativa värden (till exempel vinster i företag), vilket medför speciella problem då andra riskscenarier gäller jämfört med icke-negativa variabler. Röjanderisken kan möjligen vara lägre för variabler som kan anta negativa värden, men det är kanske ofta inte så mycket att lita på, särskilt om negativa värden är undantagsmässiga, som för företagsvinster.

En annan variant är tabeller som visar fördelningar av exempelvis inkomster genom sådana statistiska mått som medianer, kvartiler och percentiler.

Magnitudtabeller med index eller förändringstal

Tabeller med index eller förändringstal är också magnitudtabeller. Ett *index* är ett jämförelsetal som i princip anger kvoten mellan två värden på en statistisk storhet. Exempelvis anger industriproduktionsindex (förenklat) kvoten mellan volymen för industriproduktionen ett aktuellt år och motsvarande volym ett basår, även nedbrutet på branscher. Med *förändringstal* avses här (oftast relativa) förändringen i ett indexvärde eller en total (summa, totalsumma) mellan två tidpunkter. Förändringstalet uttrycks vanligen i procent och kan anta negativa värden. Vid publicering av förändringstal kan ett skydd mot röjande behövas. Detta är särskilt påkallat för företagsstatistik, där branscher kan domineras av ett eller några få företag, vilka dessutom kan ha dragits med sannolikhet 1 i undersökningen.

Diagram

Statistik som presenteras i diagram bygger på tabelluppgifter. Röjandekontroll för ett diagram kan därför utgå från röjandekontroll för den tabell som utgör underlag för diagrammet. Ibland kan ett diagram ge mindre precis information än siffrorna i en tabell, så att röjanderisken kan anses lägre. Å andra sidan kan presentationen i ett diagram underlätta spontana röjanden. Residualplottar, det vill säga diagram över resttermer från till exempel regressionsanalyser, ska hanteras med hänsyn till att de avslöjar mikrodata.

Kartor

Statistik över ett geografiskt område kan presenteras antingen som en tabell eller som en karta. Röjandekontroll av kartor innebär därför oftast i princip samma slags uppgift som röjandekontroll av tabeller. Kartor möjliggör en flexiblare indelning av områden jämfört med tabeller. Vid redovisning på låg regional nivå kan röjanderisken vara avsevärd. Se även beskrivningen av diagram ovan och av geografiska variabler i avsnitt 4.3.

4.5 Länkade tabeller

Länkade tabeller

Med *länkade* (sammankopplade) tabeller avses en uppsättning av två eller flera tabeller som genererats ur samma population (via en mikrodatafil eller en ”supertabell”). Med *semilänkade* tabeller avses en uppsättning av två eller flera tabeller som genererats ur liknande populationer, till exempel från olika år. Här ingår även tabeller med longitudinella data, som paneldata, där samma objekt ofta finns med vid flera tidpunkter.

Det kan inträffa att tabellerna var för sig bedöms som säkra – i betydelsen att ingen av dem innehåller några riskceller – men för den som har tillgång till hela uppsättningen av tabeller är det ändå möjligt att röja information om enskilda objekt.

Problemet bottnar i att det finns en överlappning av information mellan de länkade eller semilänkade tabellerna, till exempel genom gemensamma variabler och undersökta objekt. Om samma detaljerade förspalt används i flera tabeller kan det ge ökade möjligheter till identifiering. De metoder som används för tabeller var för sig riskerar i sådana fall att inte ge ett tillräckligt skydd. Ett alternativ är att utgå ifrån ”supertabellen” och utföra röjandekontrollen på denna, se även avsnitt 5.9. För en djupare genomgång, se Statistics Netherlands (2010, s. 138) och Willenborg och de Waal (2000, s. 17).

Differentiering

Differentiering innebär att flera tabeller som skiljer sig lite åt och avser samma population kan avslöja information om objekt genom att det går att räkna ut differensen mellan tabellerna. Om skillnaden mellan två tabeller endast rör en individ blir differentiering ett problem, en möjlighet till röjande, genom att tabellerna är i en mening länkade (som beskrevs i det föregående underavsnittet).

Ett tänkt exempel på möjligt röjande genom differentiering kan vara att en angripare subtraherar en tabell avseende total förvärvsinkomst för åldersgruppen 65–68 år från motsvarande tabell för åldersgruppen 65–69 år. Om då endast en enda 69-årig individ finns med i tabellen kan förvärvsinkomsten för denna läsas ut direkt.

Risken för differentiering är särskilt stor vid specialbeställningar av statistik eller om statistiken läggs in i ett flexibelt tabelluttagssystem, där användare kan skapa egna tabeller. För skydd mot differentiering, se avsnitt 7.4. Ett enkelt skydd mot viss differentiering är att undvika att publicera olika överlappande kategoriseringar av samma nyckelvariabler för samma population, till exempel olika överlappande geografiska indelningar. Indelningarna ska i stället vara hierarkiska (jämför med underavsnittet ”Variabelegenskaper” i avsnitt 4.3).

Flexibla tabelluttag från statistiska databaser

Ökade röjanderisker på grund av länkade tabeller eller differentiering är ett särskilt tydligt problem vid samanvändning av olika tabeller, till exempel vid flexibla tabelluttag från statistiska databaser. Databaserna ifråga består ofta av finfördelade makrodata; databaser med mikrodata är också tänkbara. Om användarna själva får välja sina tabeller, kan en svår röjandeproblematik uppstå. För skyddsmetoder i dessa fall, se avsnitt 7.4 eller kapitel 10.

4.6 Bortfall

Bortfallets effekter på röjanderisken

Allmänt medför bortfall att röjanderiskerna kan minska något. Bortfallet leder i sig till en informationsförlust, dels genom att svarsmängden är mindre än urvalet och därmed kräver högre uppräkningstal vid skattningen, dels genom att skevhet kan uppstå om de svarande skiljer sig från dem som inte svarar. Om uppgifter saknas för målvariabeln blir attribueringen mindre säker. Om uppgifter för någon eller flera nyckelvariabler saknas blir identifieringen mindre säker. Stora bortfall kan möjligen medföra att röjanderisken och behovet av skydd minskar, men det synsättet kan vara vanskligt. Ett kontrollerad skydd mot röjande kan ge mindre informationsförlust än ett bortfall, som ju i sig är okontrollerat. I bedömningen av röjanderisken kan hänsyn tas till bortfallsandelen, om det är fråga om objektbortfall eller partiellt bortfall, vilka variabler som uppgiften saknas för, hur bortfallet redovisas, och så vidare. Beroende på hur tabellen har definierats kan exempelvis partiellt bortfall redovisas under en egen klass i tabellen, vilket kan ge ökad röjanderisk.

Med hänsyn till röjanderisken kan det vara lämpligt att i tabellerna redovisa ”Uppgift kan ej anges” enbart i klump som en sammanslagen svarskategori. Då delas inte upp efter vad som

är orsaken till ej angiven uppgift, såsom svarsbortfall, undertryckning för röjandeskydd (beskrivs i avsnitt 7.2), otillräckligt underlag, eller något annat. Men detta kan vara en avvägning mellan å ena sidan bra röjandeskydd, och å andra sidan den eventuella nyttan för dataanalyser av en sådan uppdelning i tabellen.

Kalibrering

Kalibrering och liknande metoder för att kompensera för bortfallsskevhet medför att nya eller modifierade vikter erhålls för de enskilda objekten. Oavsett om det är vikterna i svarsmängden från ett totalräknat material eller från urvalsdata som kalibrerats, kan metoderna för urvalsdata med vikter användas. Se avsnitt 5.5 i denna handbok och s. 129–132 i Statistics Netherlands (2010).

Imputering

Imputering innebär ersättande av saknade värden i en datamängd med nya värden som kan antas ligga nära de sanna värdena. Imputerade värden ska markeras i mikrodata. En imputering leder till att ett fel som beror på imputeringsmodellen påförs de objekt för vilka uppgiften saknas, vilket kan medföra mindre säkra röjanden och då mindre röjanderisker. Imputeringen medför att datamaterialet ändå kan betraktas som komplett, så att de vanliga metoderna för röjandekontroll bör tillämpas.

4.7 Frekvenstabeller med många små tal

Små frekvenser behöver inte betyda samma röjanderisk i alla typer av tabeller. Det kan exempelvis spela stor roll vilken geografisk nivå som redovisas. Små antal har större röjanderisk på lägre geografiska nivåer som kommuner, stadsdelar eller SAMS[32]-områden, än de har på högre nivåer som län eller hela landet. Detta främst genom att lokaliseringen till små områden är mest utpekande, och även genom att angripare kan ha god kännedom om lokala förhållanden.

Det är dock inte lämpligt att ha detta som en absolut regel, utan hänsyn får tas till övriga omständigheter. Det kan finnas skäl till att se små frekvenser som riskabla även på länsnivå och för hela landet. Gotlands län är så litet att det kan behandlas på samma sätt som kommuner vid bedömning av röjanderisker, och omvänt kan Stockholm, Göteborg och Malmö behandlas på samma sätt som län.

Är tabellen meningsfull?

Då en myndighet redovisar tabeller med många små tal kan det också vara lämpligt att lyfta frågan om vad tabellerna syftar till. Meningsfull beskrivande statistik eller statistisk analys förutsätter i de flesta fall att det finns grupper av objekt att beskriva eller analysera. Om många tabellceller innehåller små antal kan tabellen få karaktären av mikrodata eller data från bokföring eller från en personallängd. I ett sådant läge kan det finnas skäl att se över hur tabellen är uppbyggd för att minska antalet celler som innehåller små tal. Det behov som statistikanvändare har gällande små tal kan, liksom behovet av mikrodata, tillgodoses via specialbeställningar.

Statistikvärden är dessutom i regel störda av osäkerhetskällor, vilket gör att små tal kan vara alltför missvisande för att redovisas. Tabeller som är mycket glesa, det vill säga har många tomma celler (nollvärden) eller många celler med små frekvenser, ger knappast heller relevant information. Sådana tabeller ger inte någon bra överblick över det som ska beskrivas och är därmed normalt inte så användbara för analys. Även om informationen i

[32] Small Areas for Market Statistics (SAMS) är en indelning i omkring 9 000 områden som bygger på kommunernas delområden i de större kommunerna och på valdistrikt i de mindre.

snäv bemärkelse minskar vid aggregering av till exempel redovisningsgrupper för att få en mindre gles tabell, kan informationen i ett vidare perspektiv sägas bli bättre genom att tabellerna ger en effektivare överblick.

4.8 Övriga förutsättningar

Osäkerhet i data

Osäkerhet i objektens individuella uppgifter, till exempel olika typer av *mätfel*, medför viss osäkerhet i uppgifter som röjs. På motsvarande sätt som med bortfall medför mätfel i målvariabler att attribueringen blir mindre säker, och mätfel i någon eller flera nyckelvariabler att identifieringen blir mindre säker. *Avrundning* av andra skäl än för att skydda mot röjande i mikrodata eller av cellvärden försvårar också röjande. Även *granskning* kan inverka på osäkerheten. Misstänkta fel ändras ofta i granskningsprocessen, men det går inte att veta om det slutliga variabelvärdet är korrekt.

Härledning av variabelvärden, till exempel beräkning av disponibel inkomst, kan leda till en minskad röjanderisk jämfört med risken för originalvariablerna. Motsvarande gäller för modellberäknade variabelvärden. Strängt taget förutsätts då att inte mycket detaljerad information finns tillgänglig om hur härledningen eller modellberäkningen har gjorts. Ett exempel på *modellberäkning* är skattning av förbrukningen av växttillgängligt kväve från stallgödsel utifrån dels modellfaktorer avseende kväveinnehållet i olika slags stallgödsel, dels uppgifter från telefonintervjuer om spridd kvantitet stallgödsel. Det är svårare för en angripare att röja en enskild lantbrukare utifrån uppgiften om mängd spritt kväve än utifrån uppgiften om mängd spridd stallgödsel.

Det kan alltså finnas skäl att ta hänsyn till olika felkällor och osäkerhet om enskilda objekts variabelvärden vid en röjandekontroll, men med det är förstås inte sagt att en planlöst ökad osäkerhet är någon lämplig väg för att skydda tabeller. Däremot finns skyddsmetoder som bygger på att osäkerhet förs in i data på ett välkontrollerat sätt (beskrivs i avsnitt 7.3–7.4). Se vidare om totalfelsaspekter nedan.

Möjligen kan det se ut som att lagen skulle tillåta att ”osanna” data röjs, och dit skulle till exempel imputerade data eller data med stora mätfel kunna räknas. Detta kan dock gå att ifrågasätta, eventuellt utifrån andra lagar. Oavsett detta är det för statistikens kvalitet angeläget att skydda alla data, så att uppgiftslämnarna kan känna gott förtroende. Felaktiga uppgifter som röjs kan också de bli till olägenhet för berörda enskilda. Förtroendet för kvaliteten i stort riskerar att ifrågasättas och statistiken kan hamna i dålig dager.

Totalfelsaspekter

Inom totalfelsparadigmet (se Groves, 2004) identifieras i princip alla källor till osäkerhet (fel): urval, täckning, mätning, bortfall och bearbetning med mera. Ambitionen är att minska påverkan av respektive felkälla, eller åtminstone beskriva påverkan. Den osäkerhet som tillförs vid skyddande av data kan betraktas som ett bearbetningsfel. Ett tydligt exempel på införande av extra osäkerhet är skyddsmetoder som arbetar med addition av brus (beskrivs i avsnitt 7.4 och 10.3). Felkomponenten som röjandeskyddet tillför ska ses i ljuset av att röjandeskyddet å andra sidan förbättrar statistikens kvalitet i fråga om tillgängligheten för användare, genom att den skapar förutsättningar för statistikens spridning. En lättnad med röjandeskydd som felkälla är att osäkerheten kan tillföras under kontrollerade former.

Vidare kan det gå att beakta totalfelet i röjandekontrollsprocessen. Om original-datamaterialet bedöms innehålla många och stora felkomponenter kan det möjligen vara

försvarligt att skydda data mindre än annars. Detta kan gälla statistiktabeller och mikrodata från urvalsundersökningar utom vid stor urvalsandel.

5 Metoder för bedömning av röjanderisk

I detta kapitel presenteras några av de vanligaste metoderna som finns att tillgå för att undersöka om röjanderisk föreligger. Det finns inte några givna regler för vilken metod som ska användas, utan den som svarar för röjandekontrollen har att avväga olika faktorer.

Röjandescenarier är ett sätt att närma sig valet av metod för identifiering av röjanderisk. Det går att ställa sig frågor om vilka som kan tänkas försöka eller råka läsa ut skyddad information ur statistik som publiceras.

Det kan finnas olika typer av angripare: allt från enskilda personer med begränsade resurser till journalister som försöker visa att det går att identifiera någon individ, och koalitioner (samarbetsgrupper) av företag med avancerad teknisk utrustning. Utifrån en uppfattning om tänkbara angripare kan valet av metod för bedömning av röjanderisk avpassas och skyddsnivån kan göras strängare eller mer tillåtande.

Olika statistiska mått kan behöva olika typer av bedömningar. Detta har redan berörts i avsnitt 4.4. Exempelvis är regressionskoefficienter, varianser och typvärden i regel svåra att utnyttja för att få ut information om de underliggande objekten. Däremot är frekvenser, magnituder (summor med mera), maximi- och minimiuppgifter och residualer (resttermer från analysberäkningar) exempel på uppgifter med större risk för röjande. Se vidare Brandt m.fl. (2010).

Statistikpublicering i databaser eller länkade tabeller är något som ger särskilda förutsättningar att beakta i metodvalet och designen för röjanderiskbedömningen. Detta har redan berörts i avsnitt 4.5. Eftersom det är svårt att omedelbart se alla sätt på vilka olika statistiska uppgifter kan förhålla sig till varandra, kompliceras här valet av metod. En väg kan då vara att tillämpa en restriktivare skyddsnivå för att säkerställa att skyddade uppgifter inte röjs.

5.1 Tröskelvärdesregeln

Publicering av frekvenstabeller föranleder bedömning av röjanderisker. Exempelvis kan tabeller publiceras över antalet elever med olika betyg i olika skolor. Det gäller då att säkerställa att ingen enskild elevs betyg röjs. Ett annat exempel är att en forskare begär ut en detaljerad frekvenstabell med uppgifter om hur många personer som misstänkts för grov stöld under olika år i olika kommuner. Även här gäller det att bedöma risken för röjande av enskilda (individer, företag eller motsvarande).

I de fallen kan det vara lämpligt att tillämpa *tröskelvärdesregeln*, som är den kanske mest använda metoden för att bedöma röjanderisk i individbaserad officiell statistik.

Grundprincip och användningsområden

Grundprincipen för tröskelvärdesregeln är att en cell i en tabell anses som riskcell om antalet observerade objekt i cellen är mindre än ett valt tröskelvärde. Typiska tröskelvärden kan vara 3, 5 eller 10, men tröskelvärdet måste vara minst 3. (Om tröskelvärdet 2 accepterades skulle celler med bara två objekt publiceras, vilket skulle kunna innebära att de två bidragsgivarna kan avslöja varandra.)

I *frekvenstabeller* är tröskelvärdesregeln ofta användbar för att skydda celler med låga frekvenser (låga antal). I *magnitudtabeller* kan den ge förstärkt skydd mot röjande via

koalitioner. Även i tabeller med magnitudvariabler som kan anta negativa värden kan tröskelvärdesregeln användas för att ge ett grundskydd.

Kvalitetsskäl snarare än röjandeskäl kan ibland motivera tröskelvärden, så att ett minsta antal observationer krävs för att skattningen ska redovisas. Ett sådant tröskelvärde kan vara lämpligt att kombinera med ett tak för det skattade medelfelet.

Skydd mot koalitioner

Om en totalsumma över investeringarna i tre företag finns publicerad, så kan två företag tillsammans räkna ut det tredje företagets investeringar (som skillnaden mellan den publicerade totalsumman och tal som de två företagen känner). Om den ansvariga för röjandekontrollen bedömer att risken för koalitioner behöver förekommas, så kan det motivera att tröskelvärdet för antalet bidragsgivare till cellerna höjs. På så sätt försvåras angreppen från aktörer i koalition med varandra.

Gruppröjande

Även celler med stora frekvenser kan vara riskceller, och det är därför oftast inte tillräckligt att använda en regel som endast identifierar celler med små frekvenser.

Ett viktigt exempel är att alla objekt i en cell kan röjas när cellfrekvensen är lika med en marginalfrekvens där cellen ingår. Med *marginalfrekvens* i en frekvenstabell avses summan av de redovisade frekvenserna för cellerna i en rad, kolumn eller annan uppdelning.

Den nämnda situationen innebär att en cell ensam innehåller alla objekt i raden, kolumnen eller motsvarande. Då är det möjligt att identifiera en grupp av objekt och säkert säga att alla dessa har den egenskap som svarar mot raden, kolumnen och så vidare. Detta kallas *gruppröjande*.

Ett annat exempel: Om cellfrekvensen är lika med en marginalfrekvens minus 1, så betyder det att någon annan cell innehåller ett objekt, och detta objekt kan med säkerhet röja värdet för de andra objekten. Om alla elever utom en i en klass får ett visst betyg, så kommer eleven med ett annat betyg att kunna säga vilket betyg alla andra har fått. Det är också ett röjande, även om det här i första hand bara är en enda individ, nämligen eleven med annat betyg, som kan få reda på de övriga elevernas betyg.

Olika tröskelvärdesregler

Tröskelvärdesreglerna för riskbedömning i en frekvenstabell kan i en grundform formuleras enligt följande:

1. En cell är en riskcell om $0 < \textit{cellfrekvens} < t_1$, där $t_1 \geq 3$.
2. En cell är en riskcell om $\textit{marginalfrekvens} - \textit{cellfrekvens} < t_2$, givet $\textit{marginalfrekvens} > 0$, där $t_2 \geq 1$.

Regel 2 avser risken för gruppröjande. Om värdet på t_2 är lika med 1 så identifieras cellen som en riskcell om frekvensen i cellen är lika med marginalfrekvensen. Om t_2 sätts till 2 så tillåts inte att ett ensamt objekt i en annan cell kan röja objekten i den bedömda cellen. Värdet på t_2 kan sättas högre om det exempelvis finns behov av beredskap för att flera objekt kan gå samman och tillsammans röja den bedömda cellen. (Regel 2 är för närvarande inte tillgänglig i de huvudsakliga IT-verktygen enligt kapitel 11.)

Det går också att lägga till en tredje tröskelvärdesregel enligt följande:

3. En cell är en riskcell om $0 < \textit{marginalfrekvens} < t_3$, där $t_3 \geq 3$.

Regel 3 går alltså ut på att införa en marginaltröskel (cellen är en riskcell om marginalfrekvensen som cellen bidrar till är lägre än marginaltröskeln), som komplement till celltröskeln i regel 1 och gruppröjandetröskeln i regel 2.

Det är lämpligt att kombinera de tre tröskelvärdesreglerna för att undvika röjanderisker på ett säkert sätt. Hur trösklarna ska sättas och kombineras kan baseras på materialets känslighet. Skyddsbehov kan finnas även om tabellen anger andelar och inte antal. Om det är sju svarande, inses direkt att 14 procent motsvarar en individ, 29 procent motsvarar två individer, och så vidare. För andelstabeller bedöms röjanderiskerna på motsvarande antalstabeller, varefter dessa skyddas och omvandlas till andelstabeller.

Möjligt resonemang för parameterval

Exempel: En enkät där frågor ska besvaras enligt en skala med de tre svarsalternativen missnöjd – varken/eller – nöjd. Utgå från ett enkelt röjandescenario där angriparna inte antas ha någon förhandskännedom om individerna som besvarat enkäten. Om då till exempel två av sju individer har svarat att de är missnöjda, så är risken (vid användning av en enkel slumpmodell med likformig fördelning) att någon av de fem inte missnöjda kan peka ut en missnöjd lika med 2/6 =1/3, vilket kan betraktas som en relativt hög risk. (Med god kännedom om övriga, vilket ofta är mer realistiskt, kan risken vara ännu högre.) En missnöjd kan peka ut en annan missnöjd med sannolikheten 1/6 och en inte missnöjd med sannolikheten 5/6. En som inte deltagit i enkäten kan peka ut en missnöjd med sannolikheten 2/7 och en inte missnöjd med sannolikheten 5/7. Om marginaltröskeln hade satts högre än 7 skulle dessa relativt höga sannolikheter för röjande kunnat undvikas. Marginaltröskeln t_3 får sättas beroende på sammanhanget, men vanligen inte under 10.

5.2 *p* %-regeln

Den så kallade *p* %-regeln är en vanlig metod för att göra riskbedömningar av magnitudtabeller med icke-negativa målvariabler. Regeln anger en gräns för hur nära en angripare ska kunna räkna ut ett objekts riktiga värde. En cell identifieras som en riskcell om det går att hitta en *övre gräns* för värdet på något objekt i cellen som är närmare än *p* procent av objektets sanna värde. En cell betraktas alltså som riskabel om det går att uppskatta bidraget från något objekt i cellen närmare än *p* procent av dess sanna värde.

Objekt som tillhör samma cell har störst möjlighet att göra en uppskattning av varandras värden. Det går att visa att det är objektet med det näst största värdet i cellen som har den bästa möjligheten att uppskatta ett annat objekts värde, och det objekt som har det största värdet i cellen är lättast att uppskatta av just det förstnämnda, det näst största. Vilket objekt som är det största antas vara känt av det näst största objektet. Riskbedömningen handlar alltså endast om att skydda objektens värden.

Grundform av *p* %-regeln

Det näst största objektet, med variabelvärdet x_2, har kännedom om sitt eget bidrag till cellsumman X och vet även att det finns ett objekt, med värdet x_1, som bidrar med mer till cellen än det självt gör. Därmed är det högsta potentiella värdet som det största objektet bidrar med lika med cellsumman minus värdet för x_2. Det minsta potentiella värdet, det vill säga den nedre gränsen, är något värde större än x_2. Det som det näst största objektet känner till om x_1 kan uttryckas på följande sätt:

$$x_2 < x_1 < X - x_2$$

För att avgöra om det näst största objektet har för mycket information om x_1 behövs att något slags kriterium sätts upp. *p* %-regeln är ett sådant kriterium. Den uttrycker hur nära en gissning av x_1 får vara för att cellvärdet ska anses säkert. Cellen identifieras som en riskcell om följande olikhet uppfylls:

$$X - x_2 - x_1 < \frac{p}{100} x_1$$

Värdet på parametern *p* kan väljas efter skyddsbehovet med hänsyn till dels en angripares bedömda identifieringsmöjligheter, dels möjlig grad av skada eller men. Principen är att ju större värde på *p*, desto strängare är riskbedömningen och desto flera celler i tabellen kommer att betraktas som riskceller. Parametern *p* sätts i praktiken ofta till 5, 10 eller 15 procent. Formeln innebär också att ju större andel av det totala värdet som de två största bidragsgivarna står för, desto mer riskabel anses uppgiften. För att belysa hur testet fungerar används tre exempel enligt följande.

Tabell 5.1 Exempel på tillämpning av *p* %-regeln

	X	x_1	x_2	Intervall	p	Test $X - x_2 - x_1 < \frac{p}{100} x_1$	Utfall
Ex. 1	100	41	40	40< x_1< 60	10	19 > 4,1	Inte risk
Ex. 2	100	59	40	40< x_1< 60	10	1 < 5,9	Risk
Ex. 3	100	50	49	49< x_1< 51	10	1 < 5	Risk

I exempel 1 är de två största företagen jämnstora, men det finns också andra aktörer på marknaden som ger betydande bidrag till cellen. Det leder till att de potentiella värdena för x_1 ryms inom ett bredare intervall. Utfallet på testet är att cellen inte är en riskcell. I exempel 2 är det största företaget betydligt större än det näst största. Intervallet för vilka de möjliga värdena är på x_1 är lika stort som i exempel 1. Det finns dock inga ytterligare aktörer som bidrar till cellvärdet på ett märkbart sätt, och därmed markerar *p* %-regeln cellen som en riskcell. I exempel 3 är de två största företagen jämnstora och står tillsammans för en helt dominerande del av cellsumman, och cellen blir markerad som riskutsatt av testet.

Kommentarer

Dessa resultat fungerar väl mot ett hypotetiskt resonemang. Låt säga att det på en marknad finns ett medelstort företag och ett mycket stort företag medan resterande aktörer är mycket små. När det näst största företaget då vill skatta vilket bidrag det största företaget har till en cellsumma, är det rimligt att anta att det bland de möjliga värdena kommer att välja ett som ligger nära den högre gränsen. Då är cellen också en riskcell.

Det finns dock begränsningar med *p* %-regeln. På en marknad med två jämnstora företag, men där de mindre aktörerna bidrar med en märkbar del till cellsumman, är det inte troligt att angriparen – det näst största företaget – skulle gissa på samma sätt som om det inte fanns några ytterligare viktiga konkurrenter. Det skulle då vara rimligare att anta att angriparen skulle välja ett värde någonstans i mitten av det tänkbara intervallet. Användning av *p* %-regeln med ett högre värde på *p* eller i kombination med en strängare tröskelvärdesregel kan då vara alternativ för identifiering av röjanderisker.

Om en myndighet producerar ekonomisk statistik över en marknad som domineras av ett fåtal aktörer, finns en risk att en stor del av statistiken kan anses röjanderiskabel, mätt utifrån *p* %-regeln. I stället för att publicera kraftigt begränsad statistik är ett alternativ i det läget att ta kontakt med de största aktörerna och se om det går att få samtycke för publicering trots att röjanderisk föreligger. Se vidare kapitel 8 om samtycke till att efterge sekretess.

Notera att *p* %-regeln implicit sätter ett tröskelvärde lika med 3 för antalet objekt som bidrar till cellvärdet. Det beror på att om det finns endast två bidragsgivare (informationsgivare) till en cell så behöver den ena enbart dra bort sitt värde från cellsumman för att få information om vilka uppgifter som lämnats av den andra bidragsgivaren.

p %-regeln för koalitioner

Regeln kan även modifieras för att hantera koalitioner. Ett exempel är en situation där de två objekten med de två näst största värdena i cellen försöker röja objektet med det största värdet. Tröskelvärdet för antalet objekt höjs i motsvarande steg; skydd mot koalitioner av storlek 2 ger tröskelvärdet minst 4 och så vidare. Följande formulering tar med koalitioner av storlek *m*-1 i bedömningen:

$$X - x_m - \ldots - x_2 - x_1 < \frac{p}{100} x_1$$

Här sätter *p* %-regeln på motsvarande sätt som ovan implicit ett tröskelvärde lika med *m*+1 för antalet objekt som bidrar till cellvärdet. Notera att *m* = 2 för *p* %-regeln utan koalitioner.

Konkret innebär förekomsten av koalitioner alltså att kraven skärps genom ett högre tröskelvärde. Det är dock lämpligt att använda samma värde på *p* med eller utan koalitioner.

pq-regeln

En utökad version av *p* %-regeln är den så kallade *pq*-regeln. *p* %-regeln kan ses som ett specialfall av *pq*-regeln (med *q* = 100). *pq*-regeln tar hänsyn till vilken kunskap som den näst största bidragsgivaren till en cell kan tänkas ha om de mindre konkurrenterna. Parametern *q* anger hur nära i procent det näst största objektet kan skatta det summerade värdet för samtliga mindre objekt. Tanken är att om den näst största aktören kan skatta hur mycket mindre aktörer bidrar till cellen, kan den informationen i sin tur användas för att bättre gissa hur mycket den största bidragsgivaren till cellen står för. Därför kan *pq*-regeln, som är en strängare regel för samma värde på *p*, behöva tillämpas. Informationsförlusten blir då också större. Ett exempel är att först sätta *p* = 20, sedan sätta *q* = 80, varmed *p* justeras till 100 *p*/*q* = 100 • 20/80 = 25. Tillämpning kan sedan göras med detta justerade, ”strängare” *p*-värde i *p* %-regeln. En detaljerad beskrivning av *pq*-regeln finns i EU:s handbok för röjandekontroll (Statistics Netherlands, 2010, s. 123–124).

5.3 Dominansregeln – (*n*, *k*)-regeln

Ett alternativ till *p* %-regeln för att göra riskbedömningar av magnitudtabeller med icke-negativa målvariabler är dominansregeln, även kallad (*n*, *k*)-regeln. Även denna regel utgår från hur bidragen till cellvärdet koncentreras till ett fåtal bidragsgivare (företag). En cell identifieras som en riskcell om de *n* största objektens variabelvärden utgör minst *k* % av celltotalen. Detta kan skrivas

$$x_1 + \cdots + x_n \geq \frac{k}{100} X$$

där *X* är cellsumman, $x_1,\ldots, x_n$ är värdena för de *n* objekt som bidrar mest till cellsumman och *k* är ett värde mellan 0 och 100. Riskbedömningen görs ofta med två eller flera uppsättningar av parametervärden samtidigt, till exempel (1,60) och (2,90).

Tillämpningsfrågor

Dominansregeln tar inte hänsyn till storleken på andra bidrag än de som inkluderas i regeln. På så sätt är regeln inte lika flexibel som *p* %-regeln. Det innebär att den i vissa fall kommer att skydda fler celler än *p* %-regeln, men i andra fall färre. Det beror på hur parametrarna sätts.

Ett fiktivt exempel på dominansregelns relevans är att ett nytt företag vill ta sig in på marknaden och då vill veta konkurrenternas kostnadsbild. En myndighet som redovisar statistik över företagens kostnader har att se till att inte röja uppgifter om en enskild aktörs kostnader.

I ett läge med ett dominerande företag kan det vara bra med (1,50) som (*n*, *k*)-regel. Därmed undertrycks alla celler där ett företag ensamt står för mer än 50 procent av verksamheten.

Värdet på parametern *k* kan väljas med hänsyn till dels en angripares bedömda identifieringsmöjligheter, dels möjlig grad av skada eller men. Principen är att ju lägre värde på *k*, desto strängare är riskbedömningen.

Om det är så att det finns två stora och många mindre företag i branschen, kan det vara lämpligt att komplettera den första regeln med ytterligare en, till exempel (2,70). Det betyder då att om två företag tillsammans står för mer än 70 procent av innehållet i cellen, betraktas den som en riskcell. Om båda dessa regler tillämpas är det troligt att uppgifter som röjer ett eller två företag inte kommer att publiceras.

Dominansregeln kan få kompletteras med andra regler när det finns behov att förebygga risker för att en koalition av företag försöker röja andra företag. Då kan tröskelvärdesregeln utgöra ett bra komplement, men det kan också vara värt att överväga att i stället använda *p* %-regeln. Tröskelvärdesregeln kan uttryckas som en dominansregel där värdet *k* har satts till 100. Således innebär till exempel dominansregeln med parametrarna (3,100) att alla celler till vilka det finns tre eller färre bidragsgivare (tröskelvärde 4) ska ses som riskceller.

I nedanstående tabell visas ett antal hypotetiska fall och hur olika tillämpningar av dominansregeln skulle fungera.

Tabell 5.2 Exempel på tillämpning av dominansregeln

	X	x_1	x_2	Intervall	(n, k)	Test	Utfall
Ex. 1	100	49	30	$30< x_1< 70$	(1,50)	49<50	Inte risk
Ex. 2	100	49	48	$48< x_1< 52$	(1,50)	49<50	Inte risk
Ex. 3	100	49	48	$48< x_1< 52$	(1,50) + (2,70)	49+48>70	Risk

Det framgår av tabellen att användning av endast (1,50)-regeln är otillräcklig då $x_1 = 49$ och $x_2 = 48$. Då regeln kompletteras med en (2,70)-regel fungerar däremot skyddet.

Jämförelse av *p* %-regeln och dominansregeln

Beroende på hur parametrarna sätts i *p* %-regeln respektive dominansregeln, kommer de att identifiera olika många riskceller och ibland olika celler. Således blir graden av skydd också beroende av valet av parametervärden. Från olika håll bland forskare och institutioner har framförts att *p* %-regeln allmänt sett är att föredra, eftersom den mer realistiskt fångar upp riskscenariot att ett enskilt objekts värde kan uppskattas inom ett snävt intervall av en angripare, se till exempel Statistiska centralbyrån (2007). Regeln fungerar också bättre tillsammans med samtycken till att efterge sekretess (se kapitel 8).

Några riktlinjer om vilken av reglerna som är lämpligast ges inte i den här handboken. Ett generellt råd är att tänka på hur de olika reglerna fungerar och bedöma hur de passar

materialen. Dominansregeln och *p* %-regeln fungerar på olika sätt, vilket kan relateras till materialens förutsättningar.

Det går att jämföra reglerna när (*n*, *k*)-regeln sätts till (2, *k*). Med de värdena tar dominansregeln i beaktande objektet med det näst största värdet. Om *k* sätts till 100 • (100/(100+*p*)) så kommer en cell som är säker enligt (2, *k)*-regeln också att vara säker enligt *p* %-regeln. Om *k* sätts till exempelvis 100 • (100/(100+25)) = 80, där alltså *p* satts till 25, så kommer en cell som är säker enligt (2, *k)*-regeln också att vara säker enligt *p* %-regeln.

5.4 Summa lika med noll

Magnitudtabeller med nollvärden

I magnitudtabeller som inte kan innehålla negativa värden så är icke-tomma celler med värdet noll (0) uppenbart riskabla, såvida det inte handlar om strukturella nollor (logiskt självklara nollor, såsom antal individer med ett examensår före födelseåret). Motsvarande kan delvis gälla när negativa värden är möjliga men sällsynta. För celler med summan noll är det uppenbart att alla objekt i redovisningsgruppen har värdet noll.

Sådan information kan vara förenad med skaderisk. Exempelvis kan kunder och andra tappa förtroende för företag som inte investerar i behövlig miljöteknik. En generell rekommendation är därför att betrakta nollvärden i tabeller som riskabla.

Frekvenstabeller med nollvärden

Nollvärden i frekvenstabeller kan delvis anses mindre problematiska än i magnitudtabeller. Nollvärden innebär dock även där gruppröjanden. Det står klart för dem som ser nollvärdet att ingen enda person bland de svarande eller i registret kan tillhöra kategorin i fråga. Ibland kan nollvärdena vara potentiellt menliga, som för antalet personer med ett visst betyg. Ofta är dock nollvärden i frekvenstabeller inte så menliga. Det kan exempelvis ses som mindre menligt att ingen person var arbetslös i en grupp ett visst år; men kanske inte helt harmlöst om någon eventuell fuskare ändå fått ersättning från A-kassa.

5.5 Urvalsdata

Det är i de flesta fall mindre riskabelt att publicera detaljerade statistiska uppgifter som baserar sig på skattningar från urval av objekt än att publicera skattningar från en totalundersökning eller ett register. Ett urvalsförfarande innebär i sig ett begränsat skydd, förutsatt att användarna inte känner till vilka objekt som ingår i urvalet.

En inklusionssannolikhet (ibland kallad urvalssannolikhet) är sannolikheten att ett objekt ingår i urvalet. Designvikterna utgör inverserna av inklusionssannolikheterna. För totalundersökta objekt som dragits med 1 som inklusionssannolikhet uppstår ett undantag. Då kan motsvarande skydd som i en totalundersökning behövas. Att det gjorts ett urval innebär alltså inte att uppgifterna automatiskt kan anses som säkra att publicera.

Undersökningar med inklusionssannolikheter som visserligen är mindre än 1 men ändå förhållandevis stora, kan ha relativt stora röjanderisker. I praktiken är det mindre vanligt i individbaserad statistik med designvikter så stora att individer som ingår i urvalet riskerar att röjas genom det. Det är vanligare med sådana röjanderisker i företagsundersökningar. Designvikterna kan i sådana fall användas i riskbedömningen (jämför underavsnittet ”Uppräkningstal och andra tekniska variabler” i avsnitt 4.3).

Följande exempel beskriver ett sätt att använda designvikter för riskbedömning i magnitudtabeller. Anta att ett urval innehåller 3 objekt för en cell, där objekten har värden

respektive designvikter 300 och 1, 100 och 2 samt 10 och 7. Skattningen av cellens summa blir 300 • 1 + 100 • 2 + 10 • 7 = 570. Det går inte att veta värdet för de största objekten i populationen, men vikterna kan användas till att räkna upp antalet objekt så att objektet med värdet 100 och vikten 2 ger 2 ”objekt” med värdet 100 och objektet med värdet 10 ger 7 ”objekt” med värdet 10. I en riskbedömning med designvikter görs bedömningen på det ”flerfaldigade” datamaterialet, vilket i det här exemplet består av värdena 300, 100, 100, 10, 10, 10, 10, 10, 10 och 10. Värdena jämförs mot den skattade cellsumman. Metoden kan användas i programvaran τ-ARGUS (se avsnitt 11.1), och programmet kan även hantera designvikter som inte är heltal.

Genom användning av uppräknade data utnyttjas samma information som en angripare har i en publicerad tabell. Bedömningen ger oftast färre riskceller jämfört med att göra bedömningen direkt på data som inte är uppräknade, genom att den tar hänsyn till att urvalsförfarandet ger ett visst skydd i sig.

När det gäller riskbedömning för frekvenstabeller, återstår det att utreda vilken metod som är lämplig.

5.6 Skuggvariabler

När *p* %- eller dominansregeln tillämpas för att identifiera riskceller, utgås från att de största objekten (till exempel företagen) i en cell är riskutsatta och behöver skydd. Ofta speglar magnitudvariabeln storleken på ett objekt, men inte alltid. Exempel på det senare är variabler såsom ekonomiskt resultat och investeringar; de kan även anta värdet noll och negativa värden.

I stället för att basera identifieringen av riskceller på den i tabellen redovisade magnitudvariabeln, kan då en annan variabel användas, en så kallad skuggvariabel, för riskidentifieringen. Som skuggvariabel kan till exempel en storleksindikator väljas för objekten, såsom omsättningen för företag.

Skuggvariabler kan utnyttjas för att hantera riskbedömning i tabeller med variabler som kan anta nollvärden och negativa värden. Ett annat skäl för skuggvariabler kan vara önskemålet att få samma undertryckningsmönster i en mängd relaterade tabeller.

En möjlighet med angreppssättet är också att förebygga röjande genom relationer mellan olika magnitudvariabler. Även för tabeller med samma bakgrundsvariabler men olika innehåll (målvariabler) kan skuggvariabler användas, om det är önskvärt att publicera liknande mönster av undertryckning över flera tabeller.

Metoden med skuggvariabler kan användas för både primär- och sekundärundertryckning (se avsnitt 7.2). En variant är att tillämpa den endast för primärundertryckning, varefter sekundärundertryckning baseras på aktuell magnitudvariabel.

Metoden är relativt enkel att implementera och minskar arbetet med bedömning av röjanderisker väsentligt om det finns ett stort antal liknande tabeller med olika magnitudvariabler. Före en implementering behöver utredas vilken skuggvariabel som är lämplig att välja för att motsvara den redovisade magnitudvariabeln med avseende på risk för röjande. Effekterna på undertryckningarna genom valet av skuggvariabel kan också vara önskvärt att utreda.

Ett exempel på användning av skuggvariabler finns i Statistiska centralbyråns undersökning Företagens ekonomi. Hundratals variabler ingår i undersökningen och tabellplanen är mycket omfattande. Somliga variabler kan anta nollvärden och negativa värden och korrelerar inte så starkt med företagsstorleken. Bedömningen av röjanderisker görs med dominansregeln med en kombination av (1,75)- och (2,90)-reglerna. Företagets omsättning

används som skuggvariabel i alla tabeller för både primär- och sekundärundertryckning. Tabellerna levereras till Eurostat med riskcellerna markerade. För den nationella resultatpubliceringen i Statistikdatabasen används grövre tabeller, kompletterade med undertryckning i några få fall.

5.7 Index och förändringstal

Inte bara frekvens- och summatabeller behöver kontrolleras för röjanderisk, utan även tabeller med index och förändringstal kan behöva granskas och skyddas. Enskilda företag kan vara stora i sina branscher relativt sett, så att deras förhållanden slår igenom och kan röjas i index och förändringstal, se även avsnitt 4.4.

Förändringstal kan anta negativa värden, vilket kan medföra särskilda möjligheter till röjande. Exempel på möjligt scenario i producentprisindex: Anta att ett företag är allmänt välkänt som dominerande i sin bransch. Om företaget kommer i en finansiell knipa och därför ensamt börjar prisdumpa, så kan branschens genomsnittliga prisutveckling tänkas avslöja skadligt mycket om det välkända företagets affärer.

Ett annat exempel: Anta att försäljningen av en vara har ökat med 10 procent sammantaget för alla företag. Om ett av de större företagen har ökat sin försäljning motsvarande mer än det värde som den 10-procentiga ökningen representerar, så kan företaget dra slutsatsen att minst ett annat företag måste ha minskat sin försäljning. På en marknad med få aktörer kan det innebära att känslig information röjs, genom att det förstnämnda företaget med sin erfarenhet av branschen kan ha möjlighet att göra en kanske alltför god gissning om vilket företaget med minskad försäljning är.

5.8 Kartdata

Den tekniska utvecklingen har ökat möjligheterna att enkelt skapa bra visualisering av statistik. En del av den trenden är en ökad efterfrågan på kartdata. Att utföra statistisk röjandekontroll av kartor är oftast i princip samma uppgift som för tabeller. Statistik över ett geografiskt område kan presenteras antingen som en tabell eller som en karta.

Kartor uttrycks digitalt med koordinater. Dessa bedöms generellt inte vara känsliga uppgifter i sig, men överväganden om skyddade identiteter är nödvändiga, se avsnitt 6.1. Det är annars i princip när olika attributdata läggs till koordinaterna som problem uppstår. En reservation kan också få göras för att nya tekniska möjligheter för användare att ännu lättare komma åt olika attributdata från koordinatdata kan komma att medföra att även de senare blir att se som generellt känsliga.

Samma resonemang som för andra geografiska nivåer gäller här. Ju mer detaljerad redovisningsnivå, desto större risk för att enskilda röjs. Koordinater är den lägsta möjliga redovisningsnivån avseende geografi och har därför i princip störst risker för röjande.

Ett exempel är uppgifter om brottsplats i polisanmälningar. Att publicera koordinater här kunde ge uppenbara risker för identifiering av enskilda individer. När Brottsförebyggande rådet genomförde en studie över så kallade hotspots, det vill säga specifikt brottsdrabbade områden, användes uppgifterna om koordinater, men resultaten presenterades i form av så kallade densitetskartor. På så sätt kunde den intressanta geografiska informationen redovisas utan att enskilda individer riskerade att röjas (Brottsförebyggande rådet 2011).

En utförligare beskrivning av olika karttyper och metoder för att kontrollera röjanderisker för dessa finns i Statistiska centralbyråns publikation *Statistisk röjandekontroll av tabeller, databaser och kartor* från 2001, s. 27–30 och s. 39 (Statistiska centralbyrån 2001b).

5.9 Länkade tabeller

Med länkade tabeller avses, som beskrivits i avsnitt 4.5, en uppsättning av två eller flera tabeller som genererats ur samma mikrodatafil eller ”supertabell”. Även om tabellerna var för sig bedöms sakna riskceller, kan det vara möjligt att röja information om enskilda objekt för den som har tillgång till hela uppsättningen av tabeller, genom att det finns överlappande information mellan de länkade tabellerna. De metoder som används för tabeller var för sig riskerar i sådana fall att inte ge ett tillräckligt skydd.

Även om det kan vara svårt att förutse alla tänkbara kopplingar mellan tabeller är det viktigt att beakta den problematiken. Ett minimikrav kan vara att bedöma om det finns utökade risker eller inte då tabeller kan länkas samman. Om det finns osäkerheter, kan en utväg vara att skärpa kraven i den metod för identifiering av röjanderisker som använts i tabellerna var för sig. Här kan det också uppstå situationer där celler i en tabell behöver undertryckas för att skydda celler i en annan tabell, som i sig betraktas som mindre intressant. En väg är att utgå ifrån ”supertabellen” och bedöma röjanderiskerna i denna. Andra typer av regler kan också användas, till exempel en regel om hur många tabeller som får genereras ur samma material eller en regel om hur många variabler tabellerna får innehålla.

Länkade tabeller kan ibland med fördel skyddas med andra metoder än undertryckning. I avsnitt 7.4 diskuteras skyddsmetoder som kan vara lämpliga vid samanvändning av tabeller via flexibla tabelluttag från statistiska databaser. Även röjandekontroll av mikrodata kan vara aktuellt som skydd, se vidare kapitel 10. Frågeställningar kring länkade tabeller diskuteras bland annat i EU:s handbok för röjandekontroll (Statistics Netherlands, 2010, s. 138).

5.10 Information om metod för bedömning av röjanderisk

En allmän regel, som nämndes i avsnitt 2.1, är att ansvariga för röjandekontroll inte ska vara alltför detaljerade i sin information (inte alltför ”transparenta”) till andra om de metoder som tillämpats för riskidentifiering och skydd av ett material. Ju mer information en potentiell angripare har om metoderna för röjandekontroll, desto större möjlighet har angriparen att ta sig förbi skyddet. Ett illustrativt exempel beskrivs av Statistiska centralbyrån (2007, s. 30).

Röjandekontrollerad statistik behöver åtföljas av förklaringar av dels *varför* röjandekontroll görs, dels allmänt *hur* den är gjord, men begränsat till vad som behövs för att ge en bild av statistikens kvalitet och användbarhet. Det kan vara viktigt att vara tydlig med om tabellceller har ”manipulerats” på något sätt. Risken är annars att användaren tolkar uppgifterna i tabellerna som mer exakta än de är.

6 Bedömning av risk för skada eller men

Utöver bedömningen av röjanderiskerna och röjandescenarierna, se kapitel 5, ska en bedömning göras huruvida ett röjande medför skada eller men för den enskilde. Bedömningen kan vara kvantitativ om den avser ekonomisk skada, men är annars vanligen kvalitativ. En individuell bedömning av risken för skada eller men måste göras i varje enskilt fall, men det finns faktorer som har en generell betydelse för bedömningen. I detta kapitel diskuteras faktorerna *typ av uppgift* och *typ av användare*.

Resultatet av skaderiskbedömningen och de beslut som fattas behöver dokumenteras. Det gäller dock att vara restriktiv med att sprida detaljerad information som kan underlätta röjanden, se avsnitt 5.10.

Den juridiska grunden för prövningen av skada och men finns i första hand i 24 kap. 8 § offentlighets- och sekretesslagen, den bestämmelse som informellt brukar kallas ”statistiksekretessen”. Enligt undantagen från statistiksekretessen får uppgifter lämnas ut endast under där angivna förhållanden, och då bara om det står klart att uppgifterna kan röjas utan skada eller men för den enskilde som uppgiften gäller eller för någon närstående till denne. En närmare redogörelse för detta finns i avsnitt 3.3.

Huruvida skaderisken är stor eller liten (men inte betydelselös) kan inverka på vilken röjanderisk som accepteras. När det till exempel gäller särskilt känsliga data, som har en stor skaderisk, kan det vara lämpligt att välja parametervärden som pekar ut flera celler som riskceller i bedömningen av röjanderisken (se kapitel 5). Exempelvis kan ett högre tröskelvärde användas vid tillämpningen av tröskelvärdesregeln.

I det följande förutsätts att de aktuella objekten (den enskilde) inte har lämnat något samtycke till att efterge sekretessen. För en del företagsstatistik, när ett fåtal företag riskerar att röjas, kan ett alternativ till att skydda objekten vara att försöka hämta in sådana samtycken. Lämnas ett sådant kan uppgifterna offentliggöras, se vidare kapitel 8.

6.1 Identifiering och attribuering

Om det saknas förutsättningar att i ett statistiskt material identifiera något enskilt objekt såsom en individ eller ett företag, så är röjanderisken så låg att den är betydelselös, och därmed finns det inget utrymme för någon skaderisk. Om *identifiering* (se avsnitt 4.3) av ett objekt däremot är möjlig, kan åtminstone någon skaderisk finnas. Om dessutom *attribuering* kan ske, det vill säga om en angripare kan få ny kunskap om ett identifierbart objekt, får skaderisken i princip bedömas vara större. Både vid identifiering och vid attribuering har typen av uppgift betydelse för bedömningen av skaderisken.

För personer som har beviljats skyddad identitet (skyddade personuppgifter) enligt 22 kap. 1 § offentlighets- och sekretesslagen kan skaderisken vid identifiering vara stor och även innefatta risk för fysiska men till liv och lem. Enligt bestämmelsen gäller sekretess för uppgift om en enskilds personliga förhållanden, om det av särskild anledning kan antas att den enskilde eller någon närstående till denne lider men om uppgiften röjs och uppgiften förekommer i bland annat folkbokföringen. Även när denna sekretess inte formellt överförs till ett statistikmaterial genom någon bestämmelse, så utgör den ett väsentligt förhållande för prövningen av men som avses i bestämmelsen om statistiksekretessen. Närhelst det inte kan uteslutas att sådana skyddade personuppgifter kan ingå i ett statistikmaterial, till exempel framgå av detaljerade kartor, behövs alltså skydd med hänsyn till den

eventualiteten. Skaderisken förstärks av att potentiella angripare kan tänkas villiga att ägna relativt mycket tid och möda åt sitt uppsåt.

Allmänt går det inte att överblicka vilken ytterligare information en angripare kan ha att tillgå utöver de aktuella statistikuppgifterna. Ett sätt att närma sig problemet kan vara att först betrakta extremfallet med ett ”värstascenario” där angriparen har de resurser och den tid som behövs för ett röjande, och sedan modifiera scenariot med hänsyn till hur troligt det bedöms vara.

Ett exempel kan vara statistik över antal fiskodlingar per kommun eller län. Anta att det bara finns en odling i en kommun och att denna kan identifieras med redan känd kunskap. I annan statistik anges miljöavfall från olika näringar (inklusive fiskodlingar) per kommun, varmed avfallet från den enskilda fiskodlingen kan röjas. Skaderisken kan då vara påfallande. Det finns alltså anledning att vidta försiktighetsåtgärder även vid risk för identifiering, bland annat för att bevara förtroendet för den statistikproducerande myndigheten.

När det gäller men för fysiska personer behöver hänsyn tas till det eventuella obehag som en enskild kan känna, särskilt om det rör sig om känsliga uppgifter. Detta gäller även i fallet att det enbart rör sig om *självidentifiering*. Att någon kan känna igen sig själv och sina egna variabelvärden i en statistisk tabell kan ge obehag och jämställas med ett integritetsintrång. Även om självidentifieringen i sig inte betraktas som ett röjande, så riskerar det att sänka förtroendet för statistikproducentens vilja och förmåga att skydda uppgifterna, vilket också bör tas hänsyn till.

6.2 Typ av uppgift

Känslighet

Som nämnts tidigare i avsnitt 3.3 avser begreppet *men* integritetskränkningar av olika slag. En integritetskränkning kan vara att någon blir utsatt för andras missaktning om personliga förhållanden blir kända. Det kan också röra sig om att uppgifter sprids som den enskilde upplever som obehagliga att andra känner till. Utgångspunkten vid bedömningen av risken för men är den subjektiva upplevelsen hos den som riskerar att drabbas. Denna kan dock behöva korrigeras med utgångspunkt i värderingar i samhället generellt, se även den avslutande delen av avsnitt 3.3. Med hänsyn till rådande värderingar i samhället kan också bedömningen av vad som kan utgöra men ändras över tid.

Känsliga uppgifter kan allmänt sägas syfta på sådana uppgifter som genom innehållets karaktär kan medföra särskilt betydande risker för skada eller men om de röjs. En uppgifts känslighet får ytterst bedömas specifikt utifrån innehållets karaktär och sammanhanget där uppgiften förekommer.

Som *känsliga personuppgifter* betraktas enligt artikel 9.1 i EU:s dataskyddsförordning följande: personuppgifter som avslöjar ras eller etniskt ursprung, politiska åsikter, religiös eller filosofisk övertygelse eller medlemskap i fackförening, genetiska uppgifter, biometriska uppgifter för att entydigt identifiera en fysisk person, uppgifter om hälsa eller uppgifter om en fysisk persons sexualliv eller sexuella läggning. Enligt artikel 10 i EU:s dataskyddsförordning gäller också särskilda bestämmelser för personuppgifter som rör fällande domar i brottmål och lagöverträdelser som innefattar brott eller därmed sammanhängande säkerhetsåtgärder[33].

[33] Med ”därmed sammanhängande säkerhetsåtgärder” avses straffprocessuella tvångsmedel, som häktning, husrannsakan, reseförbud och telefonavlyssning.

I lagen (2002:546) om behandling av personuppgifter i den arbetsmarknadspolitiska verksamheten betraktas även vissa andra personuppgifter som extra skyddsvärda. Det gäller personuppgifter om sociala förhållanden och omdömen, slutsatser eller andra värderande upplysningar om en enskild, samt uppgifter om att den enskilde har vårdats med stöd av socialtjänstlagen eller varit föremål för åtgärd enligt utlänningslagen.

Om uppgifter av det slag som nämnts ovan röjs kan det i regel leda till men. Vid denna typ av röjanden uppfylls därför inte kravet ”står klart att uppgiften kan röjas utan att den enskilde eller någon närstående till denne lider skada eller men”. En kategori uppgifter kan emellertid ha en varierande grad av känslighet. En uppgift om sexualbrott kan i allmänhet betecknas som känsligare än en uppgift om till exempel cykelstöld. På liknande sätt kan en uppgift om att en person har aids eller cancer vara känsligare än en uppgift om att en person har gått igenom en allmän hälsokontroll.

Ibland är variabeln som sådan inte generellt känslig men kan innehålla *känsliga kategorier*, se även stycket om variabelegenskaper i avsnitt 4.3. Exempelvis är födelselandet Sverige i allmänhet inte särskilt känsligt, vilket dock ett annat födelseland kan vara. Hushållssammansättning är inte så känslig allmänt sett, medan kategorin samkönad hushållssammansättning kan vara känsligare. På liknande sätt är variabeln civilstånd oftast oproblematisk, medan kategorin efterlevande (registrerad) partner kan vara känslig. Vid skyddandet av den känsliga kategorin gäller det dock att beakta risken för härledning av dess värde med hjälp av övriga kategoriers värden.

Förutom vilken typ av uppgift det rör sig om kan även typen av statistiskt objekt ha betydelse för känsligheten och därmed för skaderiskbedömningen. I synnerhet är skaderisken olika beroende på om objektet är å ena sidan ett företag eller en organisation eller å andra sidan en individ eller ett hushåll. Objekttyper diskuteras i avsnitt 4.1.

Detaljeringsgrad

En annan aspekt på uppgiftstyp gäller detaljeringsgrad. Skaderisken är i allmänhet större vid ett exakt värde än vid ett approximativt. En uppgift om att en persons årsinkomst var 679 523 kr kan vara mer integritetskränkande än en uppgift om att inkomsten låg i intervallet 500–800 tkr. Likaså är en uppgift om vilka kvantiteter mineralgödsel ett lantbruk har spritt känsligare än en uppgift om att lantbruket har spritt mineralgödsel (utan att ange kvantiteten).

Tidsaspekter

Tidsaspekter har också betydelse. Om det är färska data som publiceras kan skaderisken vara stor. Äldre data, som har några år på nacken, kan vara mindre känsliga och beroende på sammanhanget eventuellt sakna skaderisk. Konjunkturstatistik kan ha mindre nyhetsvärde efter något år. Uppgifter som kan styrka misstankar om eventuellt bokförings- eller skattebrott vid företag kan däremot vara känsliga även efter preskriptionstiden för sådana brott. Också hälsouppgifter om individer kan vara känsliga under lång tid.

Det kan vara skillnad på skaderisken beroende på om företagsbaserad statistik kommer före eller efter företagens kvartalsrapporter. Vidare kan vetskapen om en kommande statistikpublicering påkalla försiktighet om den nya statistiken kan kombineras med den befintliga. Även tidigare offentliggjord statistik inom samma område kan höja skaderisken. För att kunna göra adekvata bedömningar kring tidsaspekterna kan ingående ämneskunskaper behövas.

Enligt 24 kap. 8 § offentlighets- och sekretesslagen finns det tidsgränser för sekretessen. Sekretessen gäller i högst 70 år för uppgifter om en enskilds personliga förhållanden, och om ekonomiska förhållanden i högst 20 år.

Rumsaspekter

Vid publicering av kartdata eller liknande behöver rumsaspekter beaktas. När olika egenskaper (attribut) kopplas till koordinater kan känsliga kombinationer uppstå. En uppgift om ett enskilt avlopp eller en djurbesättning som är kopplad till en karta eller en koordinatpunkt, kan till exempel ha en hög skaderisk. Skaderisken kan vara ännu större för en uppgift avseende miljöfarligt utsläpp som genom en karta eller en koordinatpunkt kan hänföras till ett enskilt avlopp eller en enskild djurbesättning. Se även avsnitt 4.3 och 4.4.

6.3 Typ av användare

På vilket sätt och till vem uppgifterna lämnas har också betydelse för bedömningen. Om uppgifterna publiceras så att vem som helst kan ta del av dem kan skaderisken normalt bedömas som större än om de lämnas ut till enstaka användare. Det finns dock flera faktorer som ska vägas in i bedömningen.

Användarens avsikt

En faktor att beakta är användarens avsikt med uppgifterna.[34] Om det exempelvis är att publicera uppgifterna, kan ett utlämnande vara jämförbart med att statistikproducenten själv publicerar dem. Att förhöra sig om beställarens avsikter ska i normalfallet inte göras vid begäran om utlämnande av allmän handling enligt tryckfrihetsförordningen. Vid en begäran om utlämnande enligt något av undantagen i 24 kap. 8 § offentlighets- och sekretesslagen kan den som begär uppgifterna erbjudas att ange vem denna är och vad uppgifterna ska användas till (se vidare avsnitt 3.3, underavsnittet ”Forskningsändamål”). Det kan då leda till en annan bedömning i frågan om huruvida utlämnande kan ske[35], se även avsnitt 8.2.

Offentlighet eller sekretess vid myndighet

En viktig fråga vid utlämnande till en myndighet är om uppgifterna kommer att vara offentliga hos myndigheten, eller om de kommer att omfattas av sekretess, och i så fall av vilken styrka. Det går inte att utgå från att en uppgift som omfattas av sekretess hos en myndighet även gör det hos den mottagande myndigheten. Vilket sekretesskydd uppgifterna har hos den mottagande myndigheten har betydelse för bedömningen av skaderisken. Se avsnitt 3.3 beträffande utlämnande för forsknings- och statistikändamål.

En så kallad sekretessöverenskommelse används ibland för att erinra den mottagande myndigheten om vilka bestämmelser som ska iakttas beträffande de utlämnade uppgifterna. Skyddet för uppgifterna hos den mottagande myndigheten påverkas inte av en sådan överenskommelse, utan beror enbart av de sekretessbestämmelser som kan tillämpas för uppgifterna där.

Förbehåll mot enskild

Vid utlämnande till en enskild finns möjligheten att lämna ut uppgifter med ett förbehåll, som inskränker mottagarens rätt att lämna uppgifterna vidare eller utnyttja dem (10 kap. 14 § offentlighets- och sekretesslagen). Genom ett sådant förbehåll kan risken för skada eller men undanröjas, det vill säga den utlämnande myndigheten skrivs fri från ansvar för skada och men. Den möjligheten finns däremot inte gentemot en myndighet.

[34] Högsta förvaltningsdomstolen 2012, ref. 64, och prop. 1979/80:2, del A, s. 81.

[35] Prop. 1979/80:2, del A, s. 81.

7 Metoder för skydd av tabeller och kartor

Det finns i grunden två olika slags metoder för att modifiera data (avser både mikrodata, se vidare kapitel 10, och makrodata) i syfte att skydda mot röjande:

- *Perturbativa metoder* skyddar genom att *data ändras, ”förvanskas”*. En störning läggs in så att en angripare inte kan veta om den fått fram riktiga uppgifter om exempelvis en individ. Störningen läggs ofta in med en viss sannolikhet, så att både värden med och värden utan störning förekommer. (Benämningen kommer av engelska verbet ”perturb”, efter latin ”perturbare”, vilket står för att uppröra, förvirra eller störa.) För tabelldata är till exempel avrundning en perturbativ metod (beskrivs i avsnitt 7.3).
- *Icke-perturbativa metoder* skyddar genom att *data redovisas mindre detaljerat*. Det görs genom att cellvärden undertrycks, eller genom att redovisningsgrupper eller svarskategorier slås samman och redovisas i klump (aggregeras).

Kombinationer mellan perturbativa och icke-perturbativa metoder kan ofta vara lämpliga.

I detta kapitel behandlas metoder för att skydda tabeller: först icke-perturbativa metoder (aggregering och undertryckning) och sedan perturbativa metoder (främst avrundning). Slutligen ges några råd om skydd av kartor.

7.1 Aggregering

Skyddsmetoden aggregering innebär sammanslagning av redovisningsgrupper eller svarskategorier. Det betyder att celler i en tabell slås samman med celler i andra rader eller kolumner än de där cellen ingår. Ett liknande grepp är att begränsa omfattningen på tabeller med detaljerade nedbrytningar. De tabellerna görs då inte lika fullständigt täckande som tabeller med grövre nedbrytningar.

Metoden med aggregering kan ofta vara ett effektivt sätt att minska röjanderisken. Den kan dock naturligen medföra informationsförlust, särskilt om de sammanslagna grupperna blir alltför blandade och mindre relevanta för tolkning. I mycket glesa tabeller kan det vara alltför ont om möjligheter till sammanslagningar, så att metoden blir mindre användbar där.

Det är lämpligt att göra sammanslagningarna på ett sätt som är konsekvent mellan olika tabellredovisningar, till exempel olika publiceringsomgångar. Detta har två syften: dels att ge jämförbarhet, dels att inte öppna luckor i skyddet.

Det kan hända att en sammanslagning av två redovisningsgrupper framstår som nödvändig för röjandeskyddet men oacceptabel för användares krav på uppdelningar. Detta är en fundamental målkonflikt som inte direkt hänger på metoden, och frågan får då efter beredning lyftas till en beslutsnivå med ett övergripande ansvar.

Arbetsgång

Aggregering är en icke-perturbativ metod där tabellens sanna värden alltså behålls. Skyddet för en nyutformad tabell provas ut iterativt med upprepade skadeprövningar. När en redovisningsgrupp innehåller riskceller slås den ihop med en annan. Sedan görs riskbedömning för den aggregerade tabellen, och om då den nya redovisningsgruppen innehåller riskceller så får ytterligare aggregering provas. I löpande produktion med en stående tabellplan i periodiska omgångar kan skadeprövningen ofta göras som en översiktlig kontroll. Den iterativa proceduren kan då väntas komma in mest för nyutformade tabeller och vid oväntade problem genom nya svarsmönster eller annat.

Tillämpningsfrågor

Vilken redovisningsgrupp som den riskutsatta ska sammanslås med är något som en ämneskunnig får ta beslut om, eventuellt utifrån kundens specifika önskemål. Om två eller flera redovisningsgrupper är riskutsatta och naturligt relaterade kan det vara lämpligt att aggregera dem, i vart fall ur synvinkeln att informationsmängden ska vara så stor som möjligt. Det finns metoder för att slå samman redovisningsgrupper enligt ett formellt optimeringsförfarande, men risken finns att resultatet blir irrelevant och oanvändbart.

För hierarkiska redovisningsgrupper, som SNI-klasser, kan aggregering göras globalt eller lokalt. Med global aggregering menas att till exempel samtliga tresiffernivåer slås ihop till den tvåsiffriga. Med lokal aggregering menas att flera tresiffriga slås ihop till en för redovisningen skapad klass. Exempelvis finns det under SNI 55 fyra treställiga koder, 551, 552, 553 och 559. Om 559 är riskutsatt så skulle antingen 559 slås ihop med 553 och redovisas som 55x (lokal aggregering) eller samtliga koder aggregeras och endast SNI 55 redovisas (global aggregering).

Vid beställningar är det lämpligt att producenten i sina resonemang med kunden tar upp vilka redovisningsgrupper som bör aggregeras. I återkommande tabeller ska normalt undvikas att aggregeringarna oplanerat varierar över tid och bryter tidsserierna. Det är viktigt att vara förutseende för detta i tabellplanen.

Ett möjligt riktmärke för när aggregering i frekvenstabeller ska göras kan vara att den genomsnittliga cellfrekvensen understiger ett visst tröskelvärde, exempelvis 5. Detta innebär ett krav på en minsta genomsnittlig cellfrekvens för att tabellen ska återges som den är.

Ett annat möjligt riktmärke för behov av aggregering kan vara att andelen tomma celler (nollfrekvenser) är högre än ett visst procenttal. Då betraktas tabellen som alltför gles, med otillåten risk för röjande, genom att celler som har populationsobjekt bakom sig ligger så vitt och relativt åtskiljbart spridda bland tomma celler. En sådan gles tabell kan också vanligen, om än kanske inte alltid, vara till mindre nytta för användare, genom bristande överblick och problem vid analys, se avsnitt 4.7.

7.2 Undertryckning

Att dölja (utelämna, maskera) värden i en tabell kallas *undertryckning*. Detta brukar markeras genom att en symbol placeras i tabellcellen: ett streck (-), ett kryss (x), en prick [punkt] (.) eller en dubbelprick [dubbel punkt] (..). Dubbelprick är troligen den vanligaste symbolen och har fördelen att den ofta också används för att markera alltför osäkra värden. Därmed kan en angripare inte veta varför (dubbel)prickning har gjorts, vilket stärker skyddet av tabellcellen (jämför avsnitt 4.6). *Dubbelprick* rekommenderas därför i denna handbok som symbol för undertryckning vid röjandekontroll, givet att den också används som symbol för okända eller alltför dåligt underbyggda värden. I teckenförklaringen rekommenderas texten ”Uppgift kan ej anges”.

Celler som i röjanderiskbedömningen bedömts som riskceller kan undertryckas. Detta är *primärundertryckning* (även primär undertryckning). Undertryckning är en icke-perturbativ metod.

I följande frekvenstabell ska tröskelvärdesregeln användas med 3 som gräns för cellfrekvensen. Celler som ska primärundertryckas markeras med rött.

Tabell 7.1 Originaltabell

	Åldersklass 1	Åldersklass 2	Åldersklass 3	**Totalt**

Region 1	10	25	125	160
Region 2	1	20	75	96
Region 3	2	15	10	27
Totalt	13	60	210	283

Det framgår att antalen för åldersklass 1 kombinerat med region 2 respektive region 3 understiger tröskelvärdet. Dessa celler bedöms som riskceller och ska alltså undertryckas enligt tabell 7.2. (Notera att undertryckning ofta inte är en särskilt lämplig metod för skydd av frekvenstabeller. De här angivna frekvenstabellerna ger dock enkla illustrationer av undertryckningsmetodiken.)

Tabell 7.2 Tabell med primärundertryckning

	Åldersklass 1	Åldersklass 2	Åldersklass 3	**Totalt**
Region 1	10	25	125	160
Region 2	..	20	75	96
Region 3	..	15	10	27
Totalt	13	60	210	283

Behov av sekundärundertryckning

Då tabellen redovisar summor av rader och kolumner, marginalsummor, går det att bakvägen härleda värdet för de nu undertryckta cellerna (96-20-75 = 1 och 27-15-10 = 2). För att hindra denna härledning måste flera celler undertryckas. Detta kallas *sekundärundertryckning* (även sekundär undertryckning eller konsekvensundertryckning). Sekundärundertryckning ska markeras med samma symbol som primärundertryckning.

Sekundärundertryckning kan behövas även om marginalsummorna inte redovisas, eftersom de kan finnas tillgängliga på annat håll, såsom i en annan statistiktabell. Det förutsätts i det följande att marginalceller för respektive rad och kolumn finns att tillgå i den mån de är relevanta.

Sekundärundertryckning innebär ett val: Vilka ytterligare celler ska undertryckas för att skydda riskcellen? Den bästa lösningen är den som minimerar informationsförlusten men samtidigt skyddar riskcellerna. Det kan tänkas att användarna av tabellen har angett en preferens, en uppgift de är mest intresserade av. I exemplet ovan har kanske beställaren angivit att den främst är intresserad av åldersklass 3; då ska i första hand åldersklass 2 undertryckas.

I tabell 7.3 nedan ges ett exempel i två varianter med liten respektive stor informations-förlust och inga preferenser när det gäller åldersklasserna. Sekundärundertryckning markeras med gult.

Tabell 7.3a Tabell med sämre sekundärundertryckning – liten informationsförlust

	Åldersklass 1	Åldersklass 2	Åldersklass 3	**Totalt**
Region 1	10	25	125	160
Region 2	..	..	75	96
Region 3	..	15	..	27
Totalt	13	60	210	283

Om dessa celler väljs kan riskcellerna fortfarande gå att räkna ut, då de sekundär-undertryckta cellerna går att räkna ut. För att skydda riskcellerna behövs ytterligare sekundärundertryckning.

Tabell 7.3b Tabell med sämre sekundärundertryckning – stor informationsförlust

	Åldersklass 1	Åldersklass 2	Åldersklass 3	**Totalt**
Region 1	10	25	125	160
Region 2	..	..	..	96
Region 3	..	..	..	27
Totalt	13	60	210	283

Endast 3 av 9 celler (i tabellens innanmäte) är nu redovisade. Informationsförlusten kan dock minskas genom att celler sekundärundertrycks inom samma dimension enligt nedan.

Tabell 7.4 Tabell med bättre sekundärundertryckning

	Åldersklass 1	Åldersklass 2	Åldersklass 3	**Totalt**
Region 1	10	25	125	160
Region 2	..	..	75	96
Region 3	..	..	10	27
Totalt	13	60	210	283

Ingen ytterligare sekundärundertryckning behövs. Nu redovisas 5 av 9 celler.

Andra hänsyn vid sekundärundertryckning

Så länge totalsummor redovisas går det alltid att räkna ut ett intervall för riskcellens värde. Regler kan sättas upp med hjälp av ett säkerhetsintervall (skyddsintervall), som visar hur litet detta intervall får vara. Om säkerhetsintervallet görs större får flera celler undertryckas. Det leder även till ytterligare sekundärundertryckningar.

I tabell 7.2 ovan kan med säkerhet sägas att de primärundertryckta cellerna ligger i intervallet 0–3, och om 0 är ett omöjligt värde blir intervallet 1–2. Detta kan bedömas som ett otillräckligt skydd, och flera celler kan behöva undertryckas. Att skydda riskceller från snäva intervall kallas att skydda från *inferential disclosure* (*inferensröjande* på svenska), och innebär en strängare syn med starkare skydd av data. Att inte använda ett säkerhetsintervall skyddar ändå cellens exakta värde, detta kallas att skydda från *exact disclosure* (*exakt röjande*). I den förenklade tabellen ovan skulle ett säkerhetsintervall göra att samtliga celler undertrycktes. Skyddet mot inferensröjande kan ses som ett extra skydd att använda vid särskilda behov. De huvudsakliga IT-verktygen enligt kapitel 11 har funktionalitet att ta hand om de nu nämnda problemen.

För optimal sekundärundertryckning behövs att informationsförlusten kan kvantifieras (se avsnitt 2.4). Oftast behövs särskild programvara för rationell hantering. Sekundärundertryckning fordrar vanligen optimering med linjärprogrammering, till exempel Integer Linear Programming, ILP, se vidare kapitel 11.

Om antalet primärundertryckningar är stort, som i en mycket gles frekvenstabell, fungerar dessa undertryckningar som skydd för varandra. Då kan antalet sekundärundertryckningar bli relativt litet. I tabeller som inte är fullt så glesa och kanske innehåller aggregerade (sammanslagna) celler kan det bli färre primärt undertryckta celler, och då kan relativt sett flera sekundärundertryckningar behövas. Det kan dock allmänt sett vara lämpligt att

undvika glesa tabeller; därför ska normalt planeringen av en röjandekontroll föregås av en översyn av tabellplanen med avseende på statistikens kvalitet.

Det kan ibland vara lämpligt att bedömningsmässigt välja regler för vilka celler som ska undertryckas för att få en över tiden konsekvent tabellstruktur.

Ett annat problem att beakta vid undertryckning är *singleton*-problematiken: värdet i en cell kommer från endast en uppgiftslämnare. Denna känner naturligtvis till sitt eget värde och kan därför vid exempelvis två undertryckta celler i en rad eller kolumn med hjälp av marginalen räkna ut värdet för den andra undertryckta cellen.

Celler med *strukturella nollor* (logiskt självklara nollor, såsom antal individer med ett examensår före födelseåret) kan inte användas för sekundärundertryckning, eftersom alla vet att dessa celler per definition är tomma.

Informationsförlusten kan vara stor vid undertryckning, se Hundepool (2012, s. 193). Detta gäller främst (stora) frekvenstabeller som har många små frekvenser. Om det är möjligt kan det därför vara bättre att använda andra metoder för skydd av frekvenstabeller, såsom aggregering, avrundning eller någon av metoderna som presenteras i avsnitt 7.4, eller ansatser med skydd av mikrodata (se kapitel 10). Undertryckning kan även användas som komplement till exempelvis aggregering.

Trots nackdelarna med minskad användbarhet av statistiken genom informationsförlust, så är kanske ändå en fördel med undertryckning som metod att dess resultat kan vara relativt lättfattligt för användare. Undertryckningen har som enda möjlig verkan att undanhålla värden vid behov, och den tar inte till några slags ersättningar med artificiella värden som eventuellt kan vara svårtolkade för användare.

7.3 Avrundning

En skyddsmetod är att avrunda värden i en frekvenstabell till heltalsmultipler av en vald bas *b*, där basen vanligen sätts lika med tröskelvärdet, till exempel 3, 5 eller 10. Detta är en perturbativ metod, där alltså cellinnehållet ändras. Observera att varje cell i en tabell kan ändras vid avrundning, det vill säga även marginalcellerna och inte bara riskcellerna i tabellens innanmäte. Ett alternativ till generell avrundning är att avrunda endast små frekvenser.

En relativ fördel med avrundning jämfört med undertryckning är att den kan motverka missvisande intryck att statistiken är exakt trots förekommande osäkerhetskällor.

Tabell 7.5 ger ett exempel på en frekvenstabell med flera riskceller (utifrån tröskelvärdet 3), dels i rad I där cellen i kolumn B och även summan är riskceller, dels i rad II, där cellerna i både kolumn A och B är riskceller.

Tabell 7.5 Exempel på frekvenstabell

	A	B	**Totalt**
I	0	1	1
II	2	2	4
III	13	7	20
Totalt	15	10	25

Det finns åtminstone tre metoder för avrundning i frekvenstabeller: deterministisk (konventionell), stokastisk (slumpmässig) och kontrollerad avrundning. De två första

metoderna är ”okontrollerade” i den meningen att de avrundade cellvärdena i tabellens innanmäte inte behöver summera exakt till de avrundade marginalvärdena.

Deterministisk avrundning

Vid *deterministisk* avrundning avrundas cellvärdet (cellfrekvensen) till närmaste multipel av *b*.

I tabell 7.6 visas resultatet av att skydda tabell 7.5 med deterministisk avrundning med bas 3. Det framgår att marginalsummorna inte alltid stämmer med summorna av ingående celler. I det här fallet är det dock ett värre problem att skyddet av cellerna i rad II misslyckas. Utfallet 3 efter deterministisk avrundning svarar ju mot värdena 2, 3 eller 4 före, och då finns det en unik lösning för rad II.

Tabell 7.6 Exempel på frekvenstabell, med deterministisk avrundning med basen 3

	A	B	**Totalt**
I	0	0	0
II	3	3	3
III	12	6	21
Totalt	15	9	24

En lösning för att inte kunna härleda cellfrekvenser med marginalsummor skulle kunna vara att beräkna marginalsummor utifrån avrundade cellfrekvenser (*justerade* marginaler), men det kan ge stora avrundningsfel för marginalsummorna. Dessutom får då marginalsummorna inte finnas tillgängliga på annat sätt. Se vidare underavsnittet ”Hantering av marginalfrekvenser” nedan.

Generellt ses deterministisk avrundning som en mindre effektiv metod än stokastisk eller kontrollerad avrundning för att skydda riskceller i frekvenstabeller (Hundepool m.fl. 2012, s. 195). Deterministisk avrundning används ibland för att antyda osäkerhet i värden.

Stokastisk avrundning

Vid *stokastisk* avrundning avrundas cellfrekvensen slumpmässigt, normalt till någon av de två närmaste multiplerna. Sannolikheterna för respektive multipel kan väljas så att de garanterar en väntevärdesriktig avrundning, det vill säga att väntevärdet[36] för en avrundad cell är lika med den ursprungliga cellfrekvensen.

Med väntevärdesriktig avrundning och bas *b* avrundas frekvenser till någon av de närmaste multiplarna med sannolikheter som är proportionella mot avståndet till motsatt multipel. Säg att frekvensen är f och låt $r = f \bmod b$, det vill säga resten vid division med b. Då avrundas f nedåt till $f - r$ med sannolikheten $(b - r)/b$ och uppåt till $f - r + b$ med sannolikheten r/b. Ett enkelt exempel på detta är $f = 1$ och $b = 3$. Då blir $r = 1$, och 1 avrundas nedåt till 0 med sannolikhet (3-1)/3 = 2/3 och uppåt till 3 med sannolikhet 1/3.

I tabell 7.7 redovisas ett utfall av stokastisk avrundning av tabell 7.5.

Tabell 7.7 Exempel på frekvenstabell, med stokastisk avrundning med basen 3

	A	B	**Totalt**
I	0	0	0
II	0	3	3
III	12	6	21
Totalt	15	9	24

Stokastisk avrundning kan ofta ge ett bättre skydd än konventionell, deterministisk avrundning, men samtidigt blir informationsförlusten större.

Hantering av marginalfrekvenser

När det gäller marginalfrekvenser finns allmänt två alternativ vid deterministisk eller stokastisk avrundning: antingen får marginalfrekvenserna vara oförändrade (*bibehållna* marginaler) eller så beräknas de som summor av cellfrekvenserna (*justerade* marginaler). Bibehållna marginaler rekommenderas om risken för röjande kan hållas nere. Då blir tabellen dock inte säkert additiv (summakonsistent). Justerade marginaler innebär att tabellen har ett skydd i sig, men det kan omintetgöras om marginalsummorna finns tillgängliga i andra publicerade tabeller. Dessutom kan justerade marginaler leda till bristande konsistens mellan tabeller.

Kontrollerad avrundning

Kontrollerad avrundning innebär att avrundningen görs så att de avrundade cellfrekvenserna i tabellens innanmäte summerar exakt till de avrundade marginal-frekvenserna. Denna ofta efterfrågade egenskap för en tabell kallas exakt summakonsistens eller exakt additivitet. Notera att additivitet *inom* en tabell inte innebär konsistens *mellan* tabeller, vilket också ofta efterfrågas.

Kontrollerad avrundning kan ha flera lösningar. Metoden bygger på att infor-mationsförlusten kan mätas och minimeras samtidigt som röjanderisken hanteras. Det är ett optimeringsproblem som beräkningsmässigt behandlas med programvara för linjär heltalsprogrammering.

Tabell 7.8 visar en kontrollerad avrundning med basen 3 av frekvenstabellen i tabell 7.5.

[36] Väntevärdet är här medelvärdet vid ett tänkt oändligt antal avrundningar.

Tabell 7.8 Exempel på frekvenstabell, med kontrollerad avrundning med basen 3

	A	B	**Totalt**
I	0	3	3
II	3	0	3
III	12	6	18
Totalt	15	9	24

Ibland används också *semikontrollerad* avrundning. Denna bygger också på linjär heltalsprogrammering för optimal avrundning av cellvärden. Kontroll görs enbart för tabellens total, det vill säga inte för alla marginalsummor.

En annan variant är kontrollerad avrundning med *nollrestriktion*. Den innebär att originalfrekvenser som är multipler av basen inte får ändras, till exempel får med basen 3 frekvenserna 0, 3, 6, 9 och så vidare inte ändras. Andra originalfrekvenser får avrundas endast till en av de två närliggande multiplerna av basen, till exempel får med basen 3 frekvensen 7 avrundas endast till 6 eller 9. Restriktionen kan begränsa möjligheterna att finna en lösning med kontrollerad avrundning.

Kontrollerad avrundning av *en* tabell löser inte problemet med att skydda länkade tabeller och samtidigt behålla konsistens mellan dessa.

Avrundning av endast små frekvenser

För frekvenstabeller kan förekomma att endast små cellfrekvenser avrundas, till exempel en konventionell, deterministisk avrundning av 1 till 0 och 2 till 3. Att endast avrunda små frekvenser kan vara ett alternativ som bevarar stora delar av en tabell, men ofta är skyddet inte tillräckligt. Vid deterministisk avrundning beror detta dels på det deterministiska tillvägagångssättet, dels på att avrundade värden i många fall kan härledas med ledning av de värden som inte avrundats.

Stokastisk avrundning av små frekvenser, till exempel 1 till 0 eller 3, och 2 till 0 eller 3, kan vara att föredra. Avrundningen kan även tillämpas på marginalerna. Dessa kan sedan bibehållas eller justeras. För mer information om stokastisk avrundning av små frekvenser, se Shlomo m.fl. (2013).

Om det finns möjligheter till differentiering (se avsnitt 4.5), ger varken deterministisk eller stokastisk avrundning av små värden tillräckligt skydd.

Val av avrundningsmetod

En slutsats är att deterministisk avrundning bör undvikas. Stokastisk avrundning är i stort sett lika lätt att tillämpa och kan ge ett bättre skydd, även om informationsförlusten också är större; skyddet riskerar dock att brytas vid möjligheter till upprepade tabelluttag. Ofta gäller att kontrollerad avrundning är det bästa alternativet. Hundepool m.fl. (2012, s. 199) har gjort en sammanställning av avrundningsmetodernas för- och nackdelar. En variant av sammanställningen redovisas nedan i tabell 7.9.

Tabell 7.9 Sammanställning av metoder för röjandekontroll med avrundning

	Avrundningsmetod		
	Deterministisk	Stokastisk	Kontrollerad
Innanmätets cell-summa stämmer med marginalernas	Nej	Nej	Ja
Metoden ger till-räckligt skydd mot röjande	Nej	Under vissa förut-sättningar	Ja
Metoden är snabb och enkel	Ja	Ja	Nej, det kan ta mycket datortid att finna en lösning

Avrundning av endast små värden kan vara en enklare lösning än kontrollerad avrundning och under lämpliga förhållanden ge liten informationsförlust, särskilt om det inte finns så många små cellfrekvenser i tabellerna.

7.4 Andra skyddsmetoder för tabeller

Det finns andra perturbativa metoder än avrundning som kan vara lämpliga att använda för att skydda tabeller med riskceller. Gemensamt för dessa metoder är att viss osäkerhet tillförs cellvärdet på ett kontrollerat sätt. Ett så kallat brus adderas till originalvärdet och medför osäkerhet om det riktiga värdet i tabellcellen och minskar därigenom röjanderisken. Metoderna är flexibla och kan lägga till brus för endast tabellens innanmäte eller för alla tabellceller, det vill säga även marginalerna.

Barnardisering

Metoden barnardisering (barnardisation på engelska) tillämpas endast för frekvenstabeller. Den innebär att värdet 1 adderas eller subtraheras från vissa cellfrekvenser. Perturbationen görs slumpmässigt, med sannolikheterna $p/2$, $1-p$ respektive $p/2$ för tilläggen +1, 0 respektive -1 till cellfrekvensen. Sannolikheten p sätts vanligen relativt lågt, varför majoriteten av cellerna förblir ojusterade. Nollor behålls oförändrade. Marginalfrekvenserna beräknas från de justerade cellfrekvenserna, därmed blir tabellerna additiva men inte konsistenta sinsemellan. En nackdel med metoden är att stora röjanderisker kan kvarstå för små cellfrekvenser, se vidare Hundepool m.fl. (2012, s. 199).

ABS-metoden för modifiering med slumpnycklar

Den metod som här benämns ABS-metoden (för modifiering med slumpnycklar) utvecklades av Australian Bureau of Statistics (ABS) för att skydda frekvenstabeller från folk- och bostadsräkningar som publiceras i ett on-line-baserat system där användaren har stora valmöjligheter. Metoden utarbetades särskilt med tanke på risken för differentiering och risken med att publicera geografisk tillhörighet på låg nivå, se till exempel Fraser och Wooton (2005), Marley och Leaver (2011) eller Thompson m.fl. (2013). ABS-metoden testas och implementeras på Statistiska centralbyrån under 2013–2015.

ABS-metoden skapar först en extra variabel, en så kallad permanent nyckel, till varje objekt i den population som tabellerna ska avse. Nyckeln för ett objekt skapas som ett slumptal, till exempel ett tiosiffrigt tal, men behålls sedan oförändrad vid all tabellframställning, även för senare produktionsomgångar.

En frekvens i en tabellcell beräknas genom att frekvenser för olika objekt summeras. I ABS-metoden summeras (med modulär aritmetik[37]) då även objektens nycklar till en cellnyckel. Cellnyckeln används till att konstruera ett radindex som har ett värde i intervallet 0–255. Värdet på radindexet bestämmer vid tabellframställningen en rad i en så kallad uppslagstabell, medan den oskyddade cellfrekvensen bestämmer en kolumn i uppslagstabellen. Det värde (brus) i uppslagstabellen som ges av rad och kolumn adderas sedan till den ursprungliga cellfrekvensen.

Bruset antar heltalsvärden (större än, mindre än eller lika med noll), genererade under vissa villkor. Dessa villkor kan specificeras utifrån användares och producenters önskemål så att önskat skydd och informationsbehov tillgodoses. Metoden är flexibel och kan anpassas så att till exempel alla nollor bevaras orörda eller små cellfrekvenser alltid tas bort.

En fördel med denna metod är att om samma cell ingår i flera olika tabeller beräknade på samma mikrodata, så kommer den skyddade cellfrekvensen att vara lika i alla tabeller. En annan fördel är att detta sker automatiskt genom att osäkerheten tillförs data samtidigt som tabellen skapas i ett produktionssystem. Tabellerna blir konsistenta, både vid samma tillfälle och över tid, eftersom samma cell alltid skyddas likadant och tillförs samma brus oavsett när eller i vilken tabell den dyker upp. Det bidrar till att det är en metod som passar för flexibla tabelluttagssystem.

Metoden är beroende av att uppslagstabellen har definierats bra. Troligen behövs bara ett fåtal uppslagstabeller för att skydda större delen av de tabeller som produceras. Att definiera uppslagstabellen är en engångsinsats, men sedan bör designen ses över regelbundet. Eftersom metoden är så flexibel är det lätt att anpassa uppslagstabellen efter ändrade förutsättningar.

De skyddade tabellerna är inte additiva, vilket kan vara en nackdel. Det går att hantera, men till kostnaden av förlorad konsistens. Risken för gruppröjande kan kvarstå och får då hanteras separat, förslagsvis med någon annan skyddsmetod. En praktisk nackdel kan vara tillgången till färdigutvecklade it-verktyg. Funktionalitet finns dock i SuperCROSS (se avsnitt 11.6), och utveckling i SAS pågår på Statistiska centralbyrån.

Ett enkelt exempel på ett val av uppslagstabell innebär att stokastisk (väntevärdesriktig) avrundning till 0 eller 3 görs för celler med värde 1 eller 2, medan tillägget genereras från en likformig fördelning mellan -3 och 3 för celler med värde 3 eller större. I och med att alla frekvenser över 2 tilldelas likformiga slumptal mellan -3 och 3 blir bruset relativt mindre ju större cellfrekvensen är.

Controlled tabular adjustment (CTA)

Skyddsmetoden Controlled tabular adjustment (CTA) utvecklades i början på 2000-talet för magnitudtabeller. Den introducerades av Dandekar och Cox (2002). Metoden kan även användas för frekvenstabeller. På svenska kan metoden benämnas ”kontrollerad tabelljustering”. Tanken med metoden är att hitta närmaste säkra tabell till den osäkra originaltabellen genom att göra små förändringar i cellernas värden. CTA förlitar sig på optimeringsmetoder, huvudsakligen linjär (heltals)programmering (Mixed Integer Linear Programming, MILP, och Linear Programming, LP), som används för att informationsförlusten ska bli så liten som möjligt. Mer information om CTA ges i Castro och González (2011).

[37] Modulär aritmetik är ett område inom algebra. Vid summering läggs två eller flera tal samman, varefter det erhållna talet divideras med en utpekad nämnare. Den ”modulära” summan blir lika med den uppkomna resten vid divisionen.

Ett första steg i skydd med CTA är vanligen att bestämma cellernas typ: *känslig*, *får ändras* och *ska bevaras*. För alla celler bestäms gränsvärden som minimum och maximum för vad cellen får anta. För celler som ska bevaras är gränsvärdet lika med cellvärdet, och dessa celler kommer att få oförändrade värden i tabellen efter att skyddsmetoden har applicerats. Ett kritiskt moment är att bedöma hur gränsvärdena ska sättas.

För *känsliga* celler anges även så kallade skyddsgränser. Skyddsgränserna bestäms utifrån värdet i cellen, det vill säga för en frekvenstabell som originalvärdet ± *n*, där *n* är ett heltal. Skyddade celler ska få värden utanför intervallet som utgörs av originalvärdet ± *n*.

Fördelar med skyddsmetoden CTA är att tabellerna kan förbli additiva och att en stor flexibilitet erbjuds för olika tillämpningar. Vissa egenskaper kan bevaras i den skyddade tabellen, till exempel bibehållna marginaler. CTA ger ett bra skydd och informationsförlusten kan minimeras. Skydd mot differentiering åstadkoms genom att ett brus som ger en osäkerhet för alla cellvärden adderas.

En nackdel är att tabellerna normalt inte blir konsistenta med varandra. Förslag på sätt att skapa konsistenta tabeller med CTA finns men har ännu inte testats. En misstanke finns att lösningen skulle bli för prestandakrävande i vissa tillämpningar. En annan nackdel kan vara det ibland omfattande arbetet med att klassificera celler och ta fram cellgränsvärden och skyddsgränser. CTA kan ge ett visst skydd mot gruppröjande, men knappast tillräckligt för alla typer av tabeller. Det är oklart om CTA passar som metod för flexibla tabelluttagssystem.

CTA har testats i mindre utsträckning på Statistiska centralbyrån. Ett exempel som kan illustrera metoden var att alla små frekvenser, 1 och 2, avrundades slumpmässigt till 0 eller 3 med sannolikheterna 1/3 respektive 2/3, som i exemplet för ABS-metoden ovan. Dessa celler sattes sedan till att bevaras. Celler med frekvenser mellan 4 och 19 ansågs känsliga och fick skyddsgränserna ± 3 i förhållande till originalfrekvenserna och cellgränserna ± 10 i förhållande till originalfrekvenserna. Övriga icke tomma celler, även marginalerna, kunde förändras fritt inom cellgränserna ± 10. Om cellgränserna blev negativa ändrades de till 0. Marginalerna tilläts att justeras. Vikten (kostnaden för att förändra cellen) sattes till 1/cellfrekvensen, vilket innebar att celler med stora värden kostade mindre att förändra.

Skydd av tabeller med metoder som arbetar på mikrodata

Skydd av tabeller kan alternativt göras med metoder som arbetar på mikrodata, före aggregeringen till statistik i tabeller. Sådana metoder beskrivs i avsnitt 10.3.

7.5 Skydd av kartor

Riskbedömningen för kartor är likartad som för tabeller, vilket beskrevs i avsnitt 5.8. Skyddandet av kartor är även det delvis likartat men skiljer sig i en del aspekter.

Behovet av exakta värden är ofta mindre i kartor än i tabeller. I kartor redovisas ofta *intervallvärden* i olika färgskalor. I en karta som redovisar frekvenser kan det vara lämpligt att sätta det lägsta intervallet högre än tröskelvärdet. För magnitudkartor blir denna skyddsmetod mindre effektiv men utgör likväl ett skydd.

En skyddsmetod speciellt för kartor är *störning av koordinater*. I en karta återgiven som en pricktabell på koordinatnivå kan prickarna flyttas till andra koordinater än de faktiska. Prickar kan byta plats, flyttas till några givna (felaktiga) koordinater eller flyttas slumpmässigt. Slumpmässiga flyttningar kan dock ge onaturliga resultat, som att prickar för fastigheter hamnar i sjöar eller liknande, om inte modifierande villkor för detta kan användas.

De ovan beskrivna skyddsmetoderna aggregering och undertryckning (avsnitt 7.1 och 7.2) går att tillämpa även på kartor. I stället för att undertryckas (dubbelprickas) som en cell kan ett geografiskt område exempelvis vitmarkeras. Liksom för tabeller kan sekundärundertryckning behövas för att hindra härledning. För kartor är aggregering att föredra framför undertryckning. I stället för att undertryckas kan geografiska områden slås samman genom att artificiellt ges samma färgmarkering för ett värdeintervall på ett statistiskt mått, tillsammans med en tydlig markering att så har skett.

7.6 Översikt över metoder för skydd av tabeller

Följande tablå sammanfattar principer samt för- och nackdelar hos olika metoder för skydd av tabeller. Tablån är inte uttömmande, utan summarisk. De angivna egenskaperna hos metoderna gäller under förutsättning att dessa används på lämpligt sätt.

Benämning (och avsnitt)	Principer	Fördelar	Nackdelar
Aggregering (7.1)	Slår ihop celler	Kan enkelt skydda bra	En del problem i glesa tabeller
Undertryckning (7.2)	”Prickar” celler	Lättfattligt	Informationsförlusten kan bli stor
Avrundning (7.3)	Avrundar till multipel av vald bas	Begränsar informations-förlusten	Konsistent endast med särskild metod
Barnardisering (7.4)	För frekvenstabeller. Ökar/ minskar antal med 1	Lättfattligt, konsistent inom tabeller	Mindre säkert skydd, ej konsistent mellan tabeller
ABS: slump-nycklar (7.4)	För frekvenstabeller. Lägger på brus	Konsistent mellan tabeller och mellan uttagstillfällen	Nyckeln behöver passa alla tabeller; ej konsi-stent inom tabeller
CTA: skydds-gränser (7.4)	För magnitudtabeller, även frekvenstabeller. Utgår från känslighets-klassning av celler	Flexibelt, minimerar informationsförlusten, konsistent inom tabeller	Beror av klassningen; ej konsistent mellan tabeller
PRAM, data-växling (10.3)	Ändrar i mikrodata	Välkontrollerad informa-tionsförlust; konsistent inom och mellan tabeller, och mellan uttagstillfällen	Mindre lättfattligt; ändringen i mikrodata behöver anpassas till tabellplanen

8 Samtycke till att efterge sekretess

Samtycke till att efterge sekretess kan ibland vara ett alternativ till röjandeskydd av tabeller. Detta har sin juridiska grund i följande bestämmelse i 12 kap. offentlighets- och sekretesslagen:

Sekretess i förhållande till den enskilde själv

1 § Sekretess till skydd för en enskild gäller inte i förhållande till den enskilde själv, om inte annat anges i denna lag.

2 § En enskild kan helt eller delvis häva sekretess som gäller till skydd för honom eller henne, om inte annat anges i denna lag.

Om en enskild samtycker till att en uppgift som är sekretessbelagd till skydd för honom eller henne lämnas till en annan enskild endast under förutsättning att myndigheten gör ett förbehåll som inskränker den enskilde mottagarens rätt att lämna uppgiften vidare eller utnyttja den, ska myndigheten göra ett sådant förbehåll när uppgiften lämnas ut.

Att den tystnadsplikt som uppkommer genom ett sådant förbehåll som anges i andra stycket inskränker den rätt att meddela och offentliggöra uppgifter som följer av 1 kap. 1 § tryckfrihetsförordningen och 1 kap. 1 och 2 §§ yttrandefrihetsgrundlagen följer av 13 kap. 5 § andra stycket.

Sekretess gäller normalt inte mot den enskilde själv, det vill säga företaget eller individen, och den kan helt eller delvis efterges av denne. I de fall sekretessen skyddar den enskilde, kan inte den bestämmelsen om sekretess åberopas som skäl för att inte lämna ut uppgiften till den enskilde som uppgiften avser. I likhet med detta kan personen lämna sitt godkännande till att uppgiften ges till någon som annars inte skulle ha fått uppgiften på grund av sekretessen, till exempel via publicering av detaljerad statistik. Bedömer uppgiftslämnaren själv att uppgiften inte kan leda till skada eller men, får det också sägas stå klart att uppgiften inte leder till skada eller men och att den därmed kan röjas. Det är en tydlig viljeyttring från den enskilde som behövs. Detta kallas *samtycke* (eller *medgivande*) till att efterge sekretess.

Nedan behandlas samtycke till offentliggörande av statistik (avsnitt 8.1) respektive samtycke till utlämnande av mikrodata (avsnitt 8.2).

8.1 Samtycke till offentliggörande av statistik

Lagstiftningen och dess förarbeten ger inga riktlinjer för när eller hur ofta samtycke till att efterge sekretess kan användas för att kunna offentliggöra statistik på en detaljnivå som gör att enskilda eventuellt kan röjas. Inte heller finns något stöd för hur en begäran om samtycke ska utformas.

Exempelvis har Statistiska centralbyrån av tradition varit förhållandevis restriktiv med att efterfråga samtycke till att efterge sekretess. Ett argument för restriktivitet kan vara att myndigheten inte ska sätta statistiksekretessen ur spel genom att slentrianmässigt se olika situationer som undantag. Ett annat skäl är att en enskild som kontaktas av en myndighet kan uppleva ett tryck på sig att medverka eller en rädsla för att avslöjas som den som inte ställde upp (Samuelson 2004, s. 72–73). Ett alltför vanligt användande av samtycke bedöms eventuellt kunna skada förtroendet för myndigheten samt uppgiftslämnares och andras samarbetsvilja.

Typ av statistik

Samtycke till att efterge sekretess kan bli aktuellt främst för statistik som baseras på uppgifter från företag och organisationer. Ett exempel på en situation där samtycke kan användas är när det finns viktiga användarönskemål om redovisning av statistikvärden för en liten bransch som domineras av något enstaka företag. Hur känsliga de aktuella uppgifterna är måste också bedömas innan samtycke väljs som lösning.

För individ- och hushållsundersökningar är behovet av samtycke för att kunna publicera statistik inte så vanligt. Det hänger samman med att mycket detaljerad statistik normalt inte ska offentliggöras av bland annat relevans- och kvalitetsskäl, se även avsnitt 4.7. Användarönskemål om publicering kan normalt inte bedömas räcka för att kompensera de nämnda nackdelarna med samtycken. För akademisk forskning finns möjligheten att efter prövning få tillgång till mikrodata.

Utformning av begäran om samtycke

När det gäller hur begäran om samtycke ska utformas, har exempelvis Statistiska centralbyrån (utkast till) interna riktlinjer för detta. Enligt dessa ska samtyckesförklaringen undertecknas av firmatecknaren för företaget. Det ska finnas ett tydligt slutdatum för publicering, vilket inte ska vara senare än två år efter datumet för brevet med begäran om samtycke. Det ska tydligt framgå vilka uppgifter och uppdelningar som samtycket avser. Vidare ska det stå att samtycket är frivilligt och att företaget har rätt att återkalla det när som helst.

I en bilaga till denna handbok finns en mall för begäran om samtycke, som bygger på Statistiska centralbyråns ansats. Mallen kompletteras lämpligen med sidhuvud med myndighetens och eventuellt uppdragsgivarens logotyp och med sidfot med kontaktuppgifter till myndigheten. Mallen innehåller färgkoder enligt följande:

- *Svart text* är fast och ska inte tas bort annat än om den inte är relevant.
- *Röd text* ska anpassas till den aktuella situationen och antingen ändras och göras svart eller tas bort helt.
- *Blå kursiv text* innehåller instruktioner och ska alltid tas bort.

Tid för begäran om samtycke

Mallen i bilagan avser en begäran om samtycke som görs *efter* datainsamlingen, vilket normalt kan vara lämpligt för att inte störa svarsprocessen. Det är dock lämpligt att redan vid utformandet av en undersökning ta ställning till vilka uppgifters sekretess som behöver efterges för att den avsedda tabellplanen ska kunna realiseras. Utifrån tidigare produktionsomgångar kan ofta bedömas vilka företags uppgifter som ”står i vägen” för en redovisning enligt tabellplanen. Arbetet med att ta in svaren med eventuella samtycken från företagen kan bli relativt omfattande.

Mer undantagsvis kan det vara tänkbart att begära samtycke redan *under* datainsamlingen. Företaget tillfrågas då om samtycke i den utskickade blanketten. Enligt Statistiska centralbyråns rutin räcker det i detta sammanhang med att samtycket ges genom att sätta ett kryss i den ruta (i stället för namnteckning) som finns i anslutning till en text som tydligt anger att krysset innebär ett samtycke till att efterge sekretessen. Samtycket ska också uppfylla de övriga krav som framgått ovan.

Påverkan på undertryckning

I samband med röjandeskydd av en tabell genom primär- och sekundärundertryckning ska ett samtycke beaktas på ett statistiskt riktigt sätt. Annars kan ”oväntade” röjanden uppstå. Anta exempelvis att endast ett företag bidrar till ett cellvärde, varmed cellen behöver primärundertryckas. Efter förfrågan har dock företaget gett sitt samtycke. Om då cellen

sekundärundertrycks kan detta företag avslöja primärundertryckta cellers värden och därmed röja andra företag. Om det är ett företag som är dominerande (enligt valt kriterium) i en cell och primärundertryckning behövs men samtycke har getts, och den cellen sekundärundertrycks, kan detta företag också avslöja primärundertryckta cellers värden tillräckligt nära och därmed röja andra företag. Slutsatsen av detta är att celler som baseras på företag som lämnat samtycke till att efterge sekretess inte utan vidare ska få sekundärundertryckas. För mer information om röjandekontroll anpassad till samtycken, se s. 148 i Hundepool m.fl. (2012).

8.2 Samtycke till utlämnande av mikrodata

Samtycke till att efterge sekretess kan även bli aktuellt vid utlämnande av mikrodata. Denna handbok fokuserar inte på utlämnandefrågor, men ger i kapitel 10 råd för röjandekontroll av mikrodata, bland annat vid utlämnande av mikrodata. I det följande ges därför en del kommentarer kring samtycke för att efterge sekretess i det sammanhanget.

För att en myndighet ska kunna lämna ut en svarsfil med identitet, till exempel med företagsnamn eller personnummer, behövs stöd i det första eller andra undantaget från den absoluta sekretessen i 24 kap. 8 § offentlighets- och sekretesslagen eller i någon annan sekretessbrytande bestämmelse i samma lag. Om några sådana bestämmelser inte kan tillämpas, behövs samtycke från alla företag eller individer som kan identifieras av uppgifterna för att ett utlämnande ska kunna ske.

Underförstått samtycke

Statistiska centralbyrån har bedömt att det på vissa villkor kan räcka med ett så kallat underförstått (presumtivt) samtycke som begärs *under* datainsamlingen. Ett underförstått samtycke kan användas om utlämnandet endast avser uppgifter från själva enkäten och om uppgifterna endast rör uppgiftslämnaren själv. Det räcker då att uppgiftslämnaren svarar på enkäten för att det ska ses som ett samtycke, givet att tillräcklig information om utlämnandet har lämnats. Samtycket baseras således på att uppgiftslämnaren, i samband med att uppgifterna lämnas, får full information om hur uppgifterna kommer att användas, och därefter själv beslutar om uppgifterna ska lämnas eller inte. Ett uppgiftslämnande betraktas då som ett samtycke genom konkludent handlande[38]. Detta ställer höga krav på att uppgiftslämnaren tydligt informeras om vilka konsekvenser uppgiftslämnandet för med sig. När det gäller uppgifter från företag ska det kunna visas att en behörig person, det vill säga en firmatecknare, har lämnat uppgifterna för att det underförstådda samtycket ska vara giltigt.

Uttryckligt samtycke

I alla andra fall bedömer Statistiska centralbyrån att det behövs ett sedvanligt uttryckligt samtycke. Detta innebär att det måste finnas en tydlig viljeyttring i texten genom att uppgiftslämnaren lämnar sin underskrift eller kryssar i en ruta i anslutning till en text som tydligt anger att underskriften/krysset innebär ett samtycke till att efterge sekretessen. Ett exempel på ett sådant fall är att enkäten fylls i av någon annan än den fysiska eller juridiska personen som uppgifterna avser, till exempel av en anhörig till en äldre person. Andra exempel är då uppgifter lämnas om andra än uppgiftslämnaren, till exempel om personer i uppgiftslämnarens hushåll, att enkätuppgifterna kopplas ihop med registeruppgifter, eller att utlämnandet gäller redan insamlade uppgifter.

[38] Avtal genom konkludent handlande sluts när inget formellt avtal träffats mellan parterna, men när dessa agerar *som om* ett avtal existerar. Ett exempel från vardagslivet är att någon går på en buss och därigenom är skyldig att köpa biljett.

När det gäller fysiska personer är huvudregeln att endast den som uppgifterna avser kan lämna uttryckligt samtycke. Undantag görs bland annat när vårdnadshavare lämnar samtycke för sina minderåriga barn. För juridiska personer kan endast firmatecknare lämna uttryckligt samtycke.

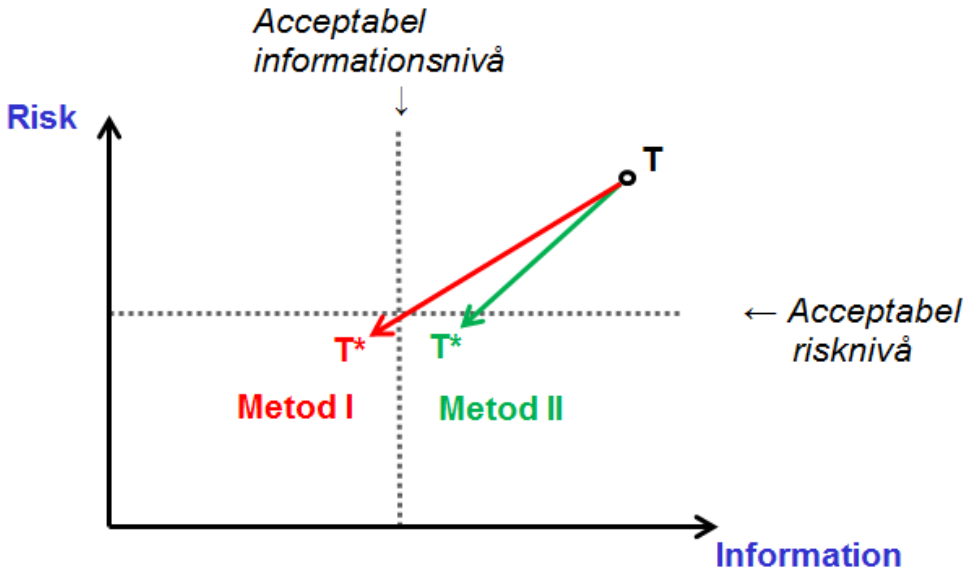
Samtycket kan lämnas direkt på svarsblanketten eller på annat sätt för exempelvis uppgifter som redan är insamlade.

För mikrodata liksom för statistik (enligt avsnitt 8.1) bedömer Statistiska centralbyrån att ett samtycke måste innehålla information om vilka uppgifter som ska lämnas ut. För mikrodata gäller också att det måste framgå till vem uppgifterna ska lämnas och hur sekretesskyddet ser ut hos den som får uppgifterna. En myndighet har rätt att fråga vad begärda uppgifter ska användas till och vem som begär dem, i den utsträckning som behövs för att kunna göra en sekretessprövning.

9 Metoder för bedömning av informationsförlust

Åtgärder för att minska röjanderisken medför i princip att informationen minskar, genom att cellvärden ändras eller tas bort. Med minskad information finns en ökad risk för begränsningar av de slutsatser som användaren kan dra utifrån tabellerna. För att minimera informationsförlusten kan en form av förlustfunktion skapas. Denna beräknas likt en kostnadsfunktion i en optimeringsrutin där målet är att hitta det sätt att skydda tabellen för röjande som ger minst informationsförlust. Valet av förlustfunktion är ett bedömningsmässigt inslag i kontrollen, och det kan därför finnas skäl att titta på flera sådana funktioner för att få en tydligare bild av effekten av skyddsmetoderna.

Figur 9.1 En förenklad bild av relationen mellan information och risk



Det är bra om det på något sätt går att mäta informationsinnehållets nytta (eller användbarhet) för statistikanvändarna, för att ställa i relation till risken. Kvantifieringen kan vara svår och handlar ofta mer om subjektiva önskemål om hur tabellerna bör se ut. Risken kan avse antingen röjandesannolikheten eller den sammantagna skaderisken, som speglar både röjandesannolikheten och storleken på skada och men. Figur 9.1 är en förenklad illustration av förhållandet mellan informationen, eller rättare sagt informationens nytta, och risken. Punkten T motsvarar en situation där informationsinnehållet i tabellen är stort, men samtidigt är risken för röjande också stor. Att jämföra hur olika metoder påverkar informationsinnehållet och risken ger en grund för att uppnå ett så stort informations-innehåll som möjligt givet önskad riskreduktion.

Vid sammanslagning av redovisningsgrupper eller dylikt får tabellen en ny utformning. En nackdel kan vara att tidsserier bryts om information går förlorad på grund av den nya utformningen av tabellen. Det är bra att försöka bedöma om de slutsatser som kan dras från den aggregerade tabellen i någon högre grad skiljer sig från slutsatserna från originaltabellen.

I det följande beskrivs mått på informationsförlust vid sekundärundertryckning (avsnitt 9.1) och avrundning (avsnitt 9.2).

9.1 Informationsförlust vid sekundärundertryckning

Informationsförlusten vid sekundärundertryckning kan uttryckas som en kostnadsfunktion som indikerar den relativa betydelsen eller vikten hos varje cell. Målet är att undertrycka de celler som tillsammans ger minsta möjliga kostnad i form av informationsförlust. I princip kan olika mått på informationsförlusten tänkas, men några har befunnits enkla och därmed användbara (Statistics
Netherlands 2010, s. 140–142).

Mått på informationsförlust

Följande mått på informationsförlust bildas som en summa av vikter för cellerna:

1. Antalet undertryckta celler. Varje cell ges lika stor vikt.
2. Antalet undertryckta objekt. Varje cell ges en vikt lika med antalet objekt i cellen.
3. Summan av undertryckta cellvärden. Cellvikten är lika med summan i cellen.
4. Summan av undertryckta cellvärden för någon annan variabel än den för vilken värden undertrycks. Cellvikten är lika med motsvarande cellvärde för någon annan variabel.
5. Transformationer av måtten i punkt 1–4, till exempel genom att cellvikten ersätts med dess logaritm, kvadrat eller kvadratrot. (Kommentar: Valet av transformation, eller att inte transformera, ska göras med kvalificerat omdöme. Logaritm eller kvadratrot kan allmänt vara lämpliga val om cellvikterna varierar starkt över tabellen, kvadrat om celler med små vikter är närmast ointressanta.)
6. Cellerna får olika vikt värderade utifrån hur det kan bedömas att användarna värderar innehållet i de olika cellerna.

Antalsuppgifterna för undertryckningarna kan redovisas med avseende på primär-undertryckning, sekundärundertryckning respektive samtliga. I optimeringsproceduren varieras endast sekundärundertryckningens bidrag till informationsförlusten. För glesa frekvenstabeller kan andel undertryckta celler redovisas *med* respektive *utan* tomma celler (nollfrekvenser) i nämnaren. Andelen kommer att se liten ut när den ställs i förhållande till alla celler i tabellen, jämfört med i relation till de celler som har värden.

Exempel med frekvenstabell

I de två tabellerna nedan har samma information förberetts för sekundärundertryckning på två olika sätt. Det som ska primärundertryckas markeras med rött, och det som ska sekundärundertryckas markeras med gult. Antalet sekundärundertryckta *celler* (punkt 1 ovan) är 3 för båda tabellerna. Däremot skiljer sig antalet sekundärundertryckta *objekt* (punkt 2 ovan), som blir 27 (12+8+7) respektive 16 (4+5+7). Informationsförlusten blir med det måttet alltså lägre i tabell 9.2.

Tabell 9.1 Frekvenstabell med primär- och sekundärundertryckningar, alternativ 1

Variabel X	Variabel Y				**Totalt**
	A	B	C	D	
A	20	2	2	1	25
B	15	***12***	***8***	15	50
C	2	4	5	1	12
D	***7***	10	16	2	35
Totalt	44	28	31	19	122

Tabell 9.2 Frekvenstabell med primär- och sekundärundertryckningar, alternativ 2

Variabel X	Variabel Y				**Totalt**
	A	B	C	D	
A	20	2	2	1	25
B	15	12	8	15	50
C	2	***4***	***5***	1	12
D	***7***	10	16	2	35
Totalt	44	28	31	19	122

Exempel med frekvens- och magnitudtabell

Det följande exemplet illustrerar punkt 3 och 4 ovan. Önskemålet är att redovisa tabell 9.3 nedan. Samma siffror som finns i tabell 9.1 och 9.2 ovan avser här antal företag inom fyra olika branscher (A–D) och ligger på vänstra halvan av tabellen. Den högra halvan visar total omsättning i miljoner kronor för dessa företag.

Här antas att det bedöms att ett tröskelvärde på minst tre företag per cell ger tillräckligt gott skydd. Det syns direkt att några celler inte lever upp till det tröskelvärdet (fetmarkerade siffror).

Tabell 9.3 Antal företag och total omsättning i branscherna A–D indelat efter antal anställda

Antal anställda	Antal företag				Total omsättning			
	Bransch				Bransch			
	A	B	C	D	A	B	C	D
0–9	20	**2**	**2**	**1**	320	27	15	3
10–49	15	12	8	15	227	212	45	32
50–249	**2**	4	5	**1**	17	68	93	2
250–	7	10	16	**2**	53	150	41	8
Totalt	44	28	31	19	617	457	194	45

Det förra exemplet visade att informationsförlusten skilde sig mellan tabell 9.1 och tabell 9.2 när cellerna vägdes enligt punkt 2, medan punkt 1 gav samma informationsförlust. Resultatet blev sekundärundertryckning enligt tabell 9.2.

Låt nu cellvärdena för en annan variabel, total omsättning, enligt punkt 4 ovan, bestämma hur antalet företag ska sekundärundertryckas. Detta resulterar i tabell 9.4. Bakom det ligger att 53+68+41 = 162, som motsvarar tabell 9.4, är lägre än 53+68+93 = 214, som motsvarar sekundärundertryckning enligt tabell 9.2.

En fördel med att använda punkt 4 kan i andra sammanhang vara den praktiska förenklingen i att utnyttja samma variabel för många olika, besläktade tabeller.

Det går då även att undertrycka omsättningssiffrorna för motsvarande celler. Det skulle också bli resultatet om omsättningssiffrorna sågs som en fristående tabell och där informationsförlusten beräknades med cellerna vägda enligt punkt 3.

Tabell 9.4 Antal företag och total omsättning i branscherna A–D indelat efter antal anställda, med primär- och sekundärundertryckningar

Antal anställda	Antal företag				Total omsättning			
	Bransch				Bransch			
	A	B	C	D	A	B	C	D
0–9	20	**2**	**2**	**1**	320	27	15	3
10–49	15	12	8	15	227	212	45	32
50–249	**2**	4	5	**1**	17	68	93	2
250–	7	10	16	**2**	53	150	41	8
Totalt	44	28	31	19	617	457	194	45

Måttens egenskaper

De olika måtten har olika för- och nackdelar beroende på typ av tabell.

Om antalet undertryckta celler används som mått är risken större att marginaler undertrycks. För att minska den risken är antalet undertryckta objekt eller summan av de undertryckta cellvärdena ett bättre mått på informationsförlust. Genom att summan av undertryckta cellvärden används bevaras så många stora celler som möjligt, men detta mått fungerar inte för negativa magnituder.

Vid bedömningen av informationsförlust är det också viktigt att ha användarnas perspektiv och utgå från hur uppgifterna ska användas. Begränsningar i att dra slutsatser från röjandekontrollerade data diskuteras mer i avsnitt 10.4.

9.2 Informationsförlust vid avrundning

Olika metoder för skyddande av tabeller genom avrundning tas upp i avsnitt 7.3. Med ett avståndsmått på skillnaden mellan de avrundade och de ursprungliga värdena (i samtliga celler, även marginaler), går det att beräkna den totala informationsförlusten vid avrundning. Avståndsmåttet kan exempelvis vara den absoluta differensen mellan det avrundade värdet X_{ij}' och det ursprungliga värdet X_{ij}, där i = rad och j = kolumn. Med absolut differens menas förenklat att ingen hänsyn tas till om skillnaden mellan X_{ij}' och X_{ij} blir ett positivt eller negativt tal, utan alla differenser behandlas som positiva och summeras.

Följande formel, som anger den summerade differensen, kan då användas vid beräkning av informationsförlusten Y

$$Y = \sum_i \sum_j \left| X_{ij} - X_{ij}' \right|$$

Ett alternativt sätt att beräkna informationsförlusten är att multiplicera varje differens ovan med en cellvikt och sedan summera. På så vis kan olika celler ges olika betydelse.

Det har visat sig att avrundning av endast små frekvenser tenderar att ge större informationsförlust än kontrollerad avrundning för glesa tabeller (med många små frekvenser). För tabeller som inte är glesa och alltså inte har många små frekvenser, kan avrundning av endast små frekvenser ge en mindre informationsförlust.

Vid avrundning väljs en bas att avrunda utifrån. Allmänt gäller att en större bas tenderar att ge större justeringar från originalfrekvenserna, och därmed större informationsförlust.

Det finns ett stort antal olika sätt att välja kostnadsfunktion på, se vidare s. 183 i Statistics Netherlands (2010).

10 Introduktion till röjandekontroll av mikrodata

Detta kapitel behandlar översiktligt statistisk röjandekontroll av *mikrodata*, alltså filer, register, databaser eller motsvarande med poster för enskilda individer, företag eller motsvarande. Tidigare kapitel har ägnats åt röjandekontroll av *makrodata* (aggregat av typ tabeller, diagram och kartor), vilket är den röjandekontroll som vanligtvis har genomförts av svenska statistikproducenter. Internationellt är mikrodatakontroll vanligt förekommande. Syftet med röjandekontroll av mikrodata respektive makrodata är detsamma: att skydda enskilda från att deras uppgifter kan röjas.

Behovet av mikrodata och hur mikrodata kan tillgängliggöras beskrivs i avsnitt 10.1. Riskerna med utlämnande av mikrodata gås igenom i 10.2. Avsnitt 10.3 beskriver tillgängliga metoder för skyddande av mikrodata, och 10.4 beskriver hur skattning och analys kan påverkas av de insatta skyddsåtgärderna.

10.1 Behov av och tillgång till mikrodata

Forskare och utredare behöver mikrodata för sitt arbete med sådant som sambandsanalyser och belysning av speciellt avgränsade grupper. De kan enligt avsnitt 3.3 under specifika villkor få tillgång till uppgifter som behövs för forsknings- eller statistikändamål. Samma möjlighet gäller inte för till exempel journalister eller företag. Från forskare och utredare uttrycks ett stort och troligen växande behov av mikrodata.

Enligt kommittédirektivet (dir. 2013:8) *Förutsättningar för registerbaserad forskning* fick en utredare i uppdrag att bland annat lämna förslag i syfte att ”registeransvariga myndigheter i större utsträckning ska kunna lämna ut uppgifter för forskningsändamål med hänsyn tagen till skyddet för den enskildes integritet”. Utredningen antog namnet Registerforskningsutredningen. I juni 2014 överlämnade Registerforskningsutredningen sitt betänkande *Unik kunskap genom registerforskning* (SOU 2014:45). De förslag som lämnades är nu under beredning.

Efterfrågan på mikrodata

Det finns flera anledningar till att mikrodata efterfrågas. Mikrodata *behövs ofta i studiet av komplexa samhällsfrågor*. Mikrodata fordras vid komplettering från register med mera, vid avgränsning av särskilda delpopulationer och i sambandsanalyser. Tekniskt lättanvända statistikprogram ger forskare ett brett spektrum av möjligheter att genomföra olika statistiska analyser.

Forskning och utredning *ökar samhällsvärdet* av statistiken. Mikrodata ökar i omfattning på grund av nya register, och forskare kan bidra till att ta till vara mikrodatas potential.

Statistikproducenten kan *få återkoppling* från forskare och utredare på kvaliteten i data. Detta även om de primära syftena i officiell statistik kan vara mer begränsade än att fullt ut svara mot en mikrodatakvalitet som kunde vara önskvärd för andra användningar, genom högre krav på granskning med mera.

Syften med röjandekontroll av mikrodata

Röjandekontroll av mikrodata kan även tjäna till att skydda makrodata. Lämpligt upplagd kan röjandekontroll på mikrodata ”slå två flugor i en smäll” och skydda både mikro- och tabelldata:

Mikrodata behöver röjandeskyddas för att kunna lämnas ut. Det räcker normalt inte med anonymisering eller avidentifiering (se begreppsförklaringar i kapitel 13), utan mer ingående bedömningar av röjanderisk och skaderisk behövs, av motsvarande skäl som för statistiktabeller (se avsnitt 2.1–2.2). Det juridiska, administrativa och tekniska skydd som tillämpas kan eventuellt kompletteras med statistiska skyddsmetoder, som beskrivs i avsnitt 10.3. Frågan kan bli aktuell särskilt om mikrodata ska lämnas till en bredare krets än vad som traditionellt varit gängse i Sverige, något som är vanligare internationellt.

Tabelldata som ska publiceras eller levereras kan röjandeskyddas på det sättet att källdata på mikronivå kontrolleras och skyddas före aggregeringen och tabelleringen. En fördel framför att skydda makrodata efter aggregeringen, som enligt metoderna i kapitel 7, är att konsistensproblem inom och mellan tabeller undviks vid skydd av mikrodata. Vidare behövs inte ett eventuellt mer arbetskrävande skydd av makrodata. Ansatsen kan vara lämplig när användare ges tillgång till flexibla uttag från statistiska databaser, även om ABS-metoden enligt avsnitt 7.4 kan vara ett alternativ för detta. Möjliga nackdelar innefattar att informationsförlusten kan bli större än vid skydd av tabelldata vid uttag, se Shlomo (2007).

Ansatser för skydd av mikrodata

Givet att mikrodata ska kunna lämnas ut och skyddas mot röjande, finns det i grunden två ansatser att tillgå:

Den första ansatsen är att mikrodata skyddas genom så kallad *accesskontroll*, det vill säga kontroll av tillgången till mikrodata. Detta kan till exempel ske i den formen att endast en väldefinierad krets dataanvändare, såsom angivna medarbetare i ett forskarteam, får tillgång till data. Skyddet bygger här på en standardiserad administrativ process med juridiskt stödd prövning av förutsättningarna och villkoren för de avsedda användarnas rätt att komma åt data. I prövningen regleras frågor som anonymisering alternativt avidentifiering, eventuellt sekretessförbehåll och begränsningar i var data lagras eller kan kommas åt.

Accesskontrollen innehåller vidare en teknisk del med anordningar för att tillhandahålla data på ett säkert avgränsat och användarvänligt sätt. Detta kan till exempel ske med hjälp av en särskilt säkrad internetförbindelse eller i särskilda lokaler, som hos statistikproducenten. Tillgången till data omgärdas av stränga regler, och kan bedömas inte behöva något skydd med sådana statistiska metoder som beskrivs i avsnitt 10.3. Denna ansats används på Statistiska centralbyrån: mikrodata tillgängliggörs över internet via systemet MONA (Microdata On-Line Access) efter godkänd utlämnandeprövning. MONA[39] beskrivs inte vidare i denna handbok.

Den andra ansatsen är att skydda mikrodata genom *statistiska metoder* för röjandekontroll. I praktiken innebär det att en del data tas bort eller ändras, se vidare avsnitt 10.3. Det ska sedan stå klart att uppgifterna kan lämnas ut utan att berörda enskilda lider skada eller men. Denna ansats har traditionellt använts i endast ringa omfattning i Sverige, men kan komma att bli mer aktuell i framtiden. Internationellt är det inte ovanligt med röjandekontroll av mikrodata utöver anonymisering eller avidentifiering, särskilt för mikrodata som sprids offentligt, främst med censusdata eller data från sociala undersökningar (se vidare nästa underavsnitt).

Ofta kombineras de juridiska, administrativa och tekniska ansatserna med statistiska ansatser för att ge ett tillräckligt skydd.

[39] Information om MONA finns på http://scb.se/mona/.

Olika typer av mikrodatafiler

Ett internationellt inte ovanligt sätt att förse forskare och utredare med mikrodata är att släppa avidentifierade mikrodatafiler för allmänt bruk utanför statistikmyndigheten. Dessa mikrodatamaterial kallas på engelska Public Use Files (PUFs), vilket kan översättas till *offentliga mikrodatafiler*. De är även användbara för undervisningsändamål på universiteten. Eftersom datamaterialet kan användas brett utan administrativa eller tekniska hinder är behovet av skydd extra stort. Bland annat ska i princip inte röjande genom identifiering kunna ske ens om mikrodatafilerna matchas med andra datafiler. Därför används statistisk röjandekontroll av mikrodata för att skydda dessa filer. För länder med liten folkmängd kan det vara svårt att skydda data utan att det uppstår en stor informationsförlust. Nackdelen med förfarandet är att störningarna som lagts in i mikrodata kan minska användbarheten av data för forskning. USA har offentliggjort mikrodatafiler sedan folkräkningen 1970. Företagsdata har dock inte lämnats ut.

Licensierade mikrodatafiler är en annan typ av anonymiserade eller avidentifierade datamaterial, där accesskontroll enligt ovan sker med hjälp av juridiska och administrativa begränsningar. Därmed är det endast forskare och utredare som myndigheten har godkänt ett utlämnande till som får tillgång till materialet. Skyddet behöver inte vara lika starkt som för offentliga mikrodatafiler. Ansatsen har använts på Statistiska centralbyrån, på senare år tillsammans med tekniska begränsningar genom utlämning av mikrodata via MONA-systemet. Det statistiska skyddet har då varit litet eller inte förekommit alls. Australian Bureau of Statistics lämnar ut licensierade mikrodatafiler (Confidentialised Unit Record Files), som dock är statistiskt skyddade. Ett liknande förfarande har även den nederländska centralbyrån.

10.2 Risker med utlämnande av mikrodata

Det finns olika riskscenarier för röjande av uppgifter om enskilda. Det tänkta röjandet görs av en angripare (se avsnitt 1.2), som med hjälp av statistiskt material, egen bakgrundskunskap och logiska slutledningar skaffar sig ny kunskap om känsliga egenskaper hos enskilda objekt. Angriparen kan exempelvis vara ett konkurrerande företag, en journalist, en datahacker eller en forskare. För att skydda mikrodata mot röjande är det lämpligt att föreställa sig olika typer av angripare i olika röjandescenarier.

Vid ett scenario med en forskare som har fått tillgång till en mikrodatafil under stränga regler, kan spontana igenkännanden uppstå då forskaren hittar ovanliga kombinationer av värden på nyckelvariabler och därmed kan identifiera en person, för vilken forskaren sedan frestas att titta på övriga uppgifter. Lämplig åtgärd kan vara att göra det omöjligt att ta fram tabeller med mycket små frekvenser i vissa celler, till exempel genom att ta bort en geografisk variabel ur filen. Ett annat scenario är att en vetenskapsjournalist tittar på forskarens publicerade tabeller som inte är tillräckligt skyddade, varpå journalisten kan röja sekretessbelagda uppgifter om enskilda. Det är väsentligt att forskaren har tillräcklig kunskap och verktyg att skydda publicerade tabeller på ett adekvat sätt.

Den utlämnande myndigheten har inte något formellt ansvar för att exempelvis en forskare på ett universitet som fått tillgång till mikrodata av misstag eller okunskap tar fram osäkra tabeller. Ansvaret för sekretessen har ju överförts via sekretessförbehåll eller genom att universitetets verksamhet omfattas av statistiksekretess. Det är ändå rimligt att den utlämnande myndigheten ser till att skyddet av mikrodata är tillräckligt, eftersom det i praktiken inte finns någon garanti för ett fullgott skydd av producerade tabeller från utlämnade mikrodata. Även om myndigheten inte har ansvaret lär den förlora anseende och förtroende vid ett röjande.

Risker med mikrodata

Det finns ett antal risker förknippade med mikrodata som kan motivera en statistisk röjandekontroll utöver anonymisering eller avidentifiering:

- Det finns en mängd *databaser*, till exempel hos företag avseende kunder eller för marknadsföring, eller via internet, som kan innehålla identiteter som kan användas för länkning.
- *Longitudinella data* eller paneldata kan öka risken för identifiering.
- Data för *små geografiska områden* kan vara alltför lätta att röja.
- *Outliers* (avvikande värden) kan uppstå genom ovanliga kombinationer av variabelvärden, vilket kan hota sekretessen.
- Om *många variabler* ingår i mikrodatafilen är risken stor att någon individ kan identifieras genom sambearbetning med en annan databas.
- *Detaljerade, precisa variabelvärden* kan leda till ökad risk för länkning och identifiering.
- *Händelsedata* kan vara riskabla. Exempel på händelsedata är övergångar mellan olika tillstånd, till exempel migration, flyttning inom Sverige eller att en person går in i och ut ur sjukskrivning.

Bedömning av röjanderisk

Vid röjandekontroll av mikrodata behöver risken för så kallad återidentifiering beaktas. Med återidentifiering menas att en anonymiserad eller avidentifierad individ kan identifieras på nytt. Det innebär alltså ett fastställande av identitet för en specifik individ, eller annat målobjekt, i den utlämnade mikrodatafilen (där ju alla poster/individer är anonymiserade eller avidentifierade).

Risken för återidentifiering per individ, i den mån den är möjlig att beräkna, kan utnyttjas för att peka ut riskindivider och därmed anvisa vilka individer som behöver skyddas. Risken kan beräknas som inversen av antalet individer med samma kombination av värden på nyckelvariabler. Om det exempelvis finns tre kvinnor i åldern 60–64 år i en viss kommun som har yrket tandläkare, så är risken 1 på 3, det vill säga 0,33. Om det bara finns en individ med den specifika kombinationen av variabelvärden, sägs individen vara populationsunik.

En indikation på risker i en mikrodatafil är andelen populationsunika individer. För urvalsdata gäller att en individ i urvalet med en unik kombination av egenskaper inte behöver vara unik i populationen med denna kombination, men det går inte att utesluta att så är fallet.

För ett exempel på riskbedömning, se avsnitt 10.3, underavsnittet ”Exempel på skydd av tabeller”.

10.3 Metoder för skydd genom röjandekontroll av mikrodata

Det finns olika typer av variabler som leder till mikrodata med olika egenskaper, vilket påverkar hur mikrodata kan skyddas. Vanligen skyddas nyckelvariabler (se avsnitt 4.3) i syfte att försvåra identifiering, men det är även möjligt att skydda målvariabler. It-verktyg för skyddande av mikrodata beskrivs kortfattat i kapitel 11.

I ingressen till kapitel 7 delades metoderna för att modifiera data i syfte att skydda mot röjande in i två typer: perturbativa och icke-perturbativa. Denna indelning är som nämnts relevant dels för metoder som arbetar på makrodata, dels för metoder som arbetar på mikrodata. Båda typerna av metoder modifierar data för att skydda mot röjande, utan att onödigt störa datas kvalitet och användbarhet. Skillnaden mellan perturbativa och icke-

perturbativa metoder är att perturbativa metoder ändrar, ”förvanskar” data, medan icke-perturbativa metoder i stället tar bort data så att informationen förgrovas. I det följande beskrivs ett antal metoder av de två typerna för att skydda mikrodata.

Omkodning

En klass icke-perturbativa skyddsmetoder går ut på omkodning för aggregering. Förfarandet innebär att en grövre klassindelning införs, det vill säga att variabelkategorier slås samman där det är lämpligt.

Lokal omkodning betyder att sammanslagningen av kategorier görs postvis, så att exempelvis olika individer kan få olika kategorier, vilket naturligtvis kan komplicera efterföljande analyser avsevärt. Metoden är därför i regel mindre lämplig.

Global omkodning innebär sammanslagning av kategorier, till exempel två eller flera åldersgrupper, på samma sätt för alla individer. Därmed kan informationsinnehållet och således röjanderisken begränsas. Några specialfall av global omkodning är följande:

- *Fri omkodning* används för att slå samman kategorier enligt exempelvis ämnesmässiga behov och bedömda röjanderisker.
- *Fast omkodning* tillämpar en förbestämd sammanslagning av kategorier och kan därmed passa bra för automatiskt skyddande av omfattande mikrodata. Det kan till exempel gälla geografisk indelning.
- *Hierarkisk omkodning* innebär att en siffernivå i en hierarkisk klassifikation tas bort, det vill säga genom att gå upp en nivå i en näringsgrens- eller yrkesindelning eller en regional indelning. Metodiken kan göras mer flexibel genom att ge tillåtelse att gå upp olika antal nivåer för olika grenar av klassifikationen. Ett problem då är att handskas med olika långa strängar för värdemängden.
- *Variabelundertryckning* innebär att en variabel tas bort helt och hållet för alla objekt. Detta kan formellt ses som omkodning av en variabels alla värden till ett och samma värde, vilket gör variabeln informationslös. Exempel på detta är anonymisering genom att ta bort personnummer eller namn och adress.
- *Topp- och bottenkodning* (öppna kategorier) tillämpas på ordnade variabler. Vid toppkodning skapas en ”toppkategori” som innefattar variabelvärden över en viss nivå. (Ett alternativ är att tilldela medel- eller medianvärdet för alla i den översta kategorin.) Exempelvis skapas kategorin ”förvärvsinkomst över 2000 tkr”. Vid bottenkodning skapas en ”bottenkategori”, till exempel ”förvärvsinkomst under 100 tkr”. Toppkodning används ofta för variabeln ålder, exempelvis tillämpas kategorin ”100 år och äldre” för individer. En svårighet med metodiken är att hitta ett lämpligt gränsvärde för avgränsning av den öppna kategorin.

Lokal undertryckning

Lokal undertryckning betyder att ett variabelvärde tas bort för ett specifikt objekt, det vill säga att ett partiellt bortfall införs, vilket är mindre genomgripande än att eliminera variabeln helt (variabelundertryckning). Lokal undertryckning är inte en perturbativ metod. Ansatsen är postvis och leder inte till några inkonsistenser i data. Däremot kan skevheter uppstå om de mest extrema värdena undertrycks. Lokal undertryckning rekommenderas därför mest som ett komplement till global omkodning, med syfte att eliminera ett fåtal kvarvarande osäkra poster med ovanliga kombinationer av nyckelvärden. Genom den kombinerade ansatsen minskas informationsförlusten jämfört med global undertryckning.

En variant av lokal undertryckning är att ersätta det borttagna värdet med ett *imputerat värde*, i stället för att se det som ett partiellt bortfall. Lokal undertryckning med imputering är en perturbativ metod. Samma imputeringsmetod som används för det ordinarie partiella bortfallet i statistikprodukten kan tillämpas.

Urval

Urval som skyddsmetod innebär att endast ett urval av poster, till exempel individer, lämnas ut i en mikrodatafil. Detta är en icke-perturbativ metod. För en totalundersökning eller registerbaserad undersökning kan ett urval på någon eller några procent lämpligen dras. Metodiken används bland annat i USA och Kanada för skydd av censusdata som ska lämnas ut. Den kombineras där med omkodning till större geografiska områden. För en urvalsundersökning kan ett underurval av lämplig storlek dras. Formellt kan urvalsskydd ses som lokal undertryckning av alla variabelvärden för individer utanför urvalet. Uppgifter för dessa individer kan uppenbarligen inte röjas. Men även identifieringen av en person i urvalet försvåras, i och med att det inte är säkert att en urvalsunik individ är populations-unik.

Addition av brus

Att addera brus (slumpvärden) till en kvantitativ variabels värden utgör en perturbativ skyddsmetod. Bruset kan exempelvis följa en normalfördelning och genereras oberoende för olika objekt. Genom att sätta väntevärdet för slumpfördelningen till 0, undviks systematiska fel i skattningar från materialet. Ju högre variansen väljs, desto starkare skyddas mikrodata. Störningen av data kan påverka dataanalysen på ett för forskaren ovälkommet sätt och behöver därför utformas med dataanvändningarna i åtanke.

Avrundning

Avrundning är också en perturbativ teknik som kan tillämpas på kvantitativa variabler. Avrundning görs vanligen till en bas, såsom 5-tal, 10-tal eller 100-tal. Antingen görs detta deterministiskt till närmaste bas eller slumpmässigt enligt en vald fördelning. Avrundning kan eventuellt vara mer transparent för användaren än addition av brus.

Mikroaggregering

Mikroaggregering går ut på att ersätta individuella kvantitativa variabelvärden med medelvärden (eller medianvärden eller typvärden) inom avgränsade grupper. Metoden är perturbativ och kan ge effekter liknande dem för avrundning. En fördel är att totalsummor bevaras vid mikroaggregering med medelvärden, vilket ofta inte är fallet vid addition av brus eller avrundning. En betydande nackdel är att variansen reduceras onaturligt. Tekniken går ut på att sortera objekten efter variabelvärdet, till exempel förvärvsinkomst, avgränsa storleksgrupper och slutligen ersätta respektive variabelvärde med medelvärdet inom gruppen. Storleksgrupperna kan väljas likstora enligt en tröskel för antal objekt eller så att variationsområdet inom grupperna är lika stora. Ju större gruppen väljs, desto starkare skydd och större informationsförlust. (Se Domingo-Ferrer och Mateo-Sanz, 2002, samt Domingo-Ferrer och Torra, 2001.)

En variant av mikroaggregering är att gruppera på annat sätt än i storleksgrupper. Även vid denna osorterade mikroaggregering bevaras totalsummorna, men informationsförlusten kan bli större. Givet gruppstorlek gäller det att bilda så homogena grupper som möjligt för att hålla ned informationsförlusten.

PRAM

PRAM står för post-randomization method och innebär att variabelvärden (oordnade kategorier) ändras efter en förbestämd sannolikhetsfördelning. Tekniken är alltså perturbativ. Om en svensk individ har Bhutan som födelseland, så slumpas ett nytt land inom Asien (exempelvis) fram, med högre sannolikhet för länder som många svenskar har som födelseländer. (Se Willenborg och de Waal, 2000, avsnitt 1.8.10.)

Dataväxling

Den metod som benämns data swapping på engelska kan på svenska kallas dataväxling eller databyte. Dataväxling kan formellt ses som ett specialfall av PRAM och innebär att variabelvärden *byts* mellan objekt. I ett första steg slumpväljs utan återläggning ett urval par av objekt (individposter) ur materialet, under restriktionen att de två objekten i varje valt par ska matcha varandra genom att ha lika värden på specificerade kategorivariabler. I ett andra steg byts sedan värdena på andra specificerade variabler, ”bytesvariablerna”, mellan de två objekten i respektive par. Alla övriga variabelvärden i materialet lämnas kvar utan ändring. (Se Willenborg och de Waal, 2000, avsnitt 1.8.11.)

För givna grupper bevaras bytesvariablernas fördelningar och därmed de statistiska resultat som bygger på dem såsom medelvärden. Även sambandsmått kan bevaras. Skyddets störande inverkan på datamaterialets användbarhet för analyser kan på så sätt hållas under kontroll. För att detta ska fungera optimalt förutsätts en effektiv kommunikation med gemensam metodinsikt mellan dem som röjandeskyddar respektive analyserar datamaterialet.

Syntetiska data

En ansats för röjandekontroll av mikrodata är att inte släppa några verkliga data utan i stället anpassa en statistisk modell till verkliga data och sedan generera (imputera) en syntetisk datamängd med motsvarande fördelning. Det gäller då att modellen avspeglar de variabelsamband som användarna ska kunna analysera, vilket kan vara svårt att åstadkomma. Metoden är perturbativ genom att skapa nya data.

En mellanform är att framställa syntetiska data endast för en eller ett fåtal riskvariabler, medan mikrodata behålls oförändrade för övriga variabler.

En annan variant är att generera *flera* syntetiska datamängder med multipel imputering. Därmed kan vid rätt användning mer rättvisande slutsatser dras från skattning och analys.

Syntetiska data som ges ut från U.S. Census Bureau avseende *Survey of Income and Program Participation (SIPP)* är ett exempel. Byrån erbjuder även gratis validering på originaldata av forskarnas analysresultat från dessa syntetiska data.

Exempel på skydd av tabeller

Ett exempel på skydd av tabeller genom att applicera skydd på mikrodata kan tas från Statistiska centralbyråns metod för att skydda tabeller i Census 2011. Syftet var att tillföra osäkerhet i tabellerna för att uppnå ett tillräckligt skydd samtidigt som informationsförlusten blev liten. En förenklande omständighet för censusen var att tabellplanen var fullständigt bestämd från början (av Eurostat), med 60 fördefinierade mångdimensionella huvudtabeller och ett stort antal deltabeller. Riskbedömningen utgick helt ifrån vad en användare skulle kunna få för information i de publicerade tabellerna, och de skyddsmetoder som valdes skyddade därför enbart tabellerna och gjorde inte automatiskt att mikrodata skulle kunna lämnas ut utan risk. Exemplet handlar alltså inte om skydd för utlämnande av mikrodata.

Individer definierades som utsatta för risk om de tillhörde någon tabellcell med frekvens 1 i någon variabelkombination med *utpekande* variabler, eller om de tillhörde någon tabellcell med frekvens 1 eller 2 i någon variabelkombination med *känsliga* variabler. Vilka variabler eller kategorier som var speciellt utpekande (allmänt kända, lätta att hitta i publicerad statistik och som i kombination enkelt identifierar en enskild individ, men inte i sig känsliga) respektive känsliga (kan ge upphov till men vid röjande) fastställdes av metodstatistiker i samråd med ämnesstatistiker och jurist. Hushåll definierades som utsatta för risk om de innehöll minst en individ utsatt för risk.

Som skyddsmetod användes en kombination av dataväxling och PRAM. Dataväxling (som metoden tillämpades i censusen) innebär att geografisk tillhörighet byttes mellan riskutsatta hushåll på NUTS 2-nivå[40] genom matchning av hushåll på variablerna hushållsstorlek, könsfördelning, åldersfördelning och upplåtelseform för bostad. På så sätt kunde viktiga marginaler bibehållas. Geografisk tillhörighet är lämplig som bytesvariabel, eftersom den inte ger upphov till lika många ologiska variabelkombinationer i data som till exempel byte av variabeln ålder på individnivå skulle kunna göra. De kombinationer som ändå kunde uppkomma togs om hand i efterkontroller.

För att skydda individer som är unika i Sverige avseende födelseland i kombination med kön och åldersklass, det vill säga saknar en matchande individ, tillämpades PRAM. I censusen innebar det att värde på födelseland eller medborgarskap för dessa individer byttes mot ett ersättande värde för ett land inom samma världsdel och region enligt en sannolikhetsmodell.

Alternativ till skydd av mikrodata

Det ska också sägas att röjandekontroll av makrodata såsom tabeller kan tillämpas i stället för röjandekontroll av mikrodata i samband med utlämnande av resultat från mikrodatafiler till forskare och utredare (via en säkrad miljö över internet eller i särskilda lokaler). Detta kallas ”output checking”, som kontrast till ”input checking” (då mikrodata kontrolleras och skyddas). Kontrollen görs innan tabeller och liknande får lämna den säkrade miljön. Ansatsen tillämpas i exempelvis Storbritannien och Italien. Röjandekontrollen av tabeller med mera blir här ofta komplicerad och kontextberoende, eftersom forskarna prövar sig fram och fritt kan framställa resultat (”output”) av många olika slag, att jämföra med officiell statistik där tabellplanerna vanligen är fasta. Se även Brandt m.fl. (2010).

10.4 Påverkan på skattning och analys

I kapitel 9 beskrevs problematiken med informationsförlust. Åtgärder som skyddar mikro- eller makrodata, genom att ta bort eller ändra vissa uppgifter, försämrar förutsättningarna för statistisk analys, det vill säga för att skatta statistiska storheter för en population. Här avses framställning av beskrivande statistik, med andra ord skattning av deskriptiva parametrar såsom totalsummor, medelvärden och andelar. Här avses även statistisk analys i mer begränsad mening: skattning av analytiska parametrar såsom koefficienter i regressionsanalys. Det är väsentligt att användarnas analysperspektiv beaktas vid utformning av röjandekontrollen. I det följande beskrivs några konsekvenser för den statistiska analysen som kan uppstå vid skyddande av mikrodata.

Skyddandet av mikrodata kan leda till att vissa statistiska storheter *inte kan skattas*. Ett enkelt exempel på detta uppstår vid undertryckning av målvariabler.

Systematiska fel (bias) kan uppkomma exempelvis vid lokal undertryckning, särskilt om det är de extrema värdena som bedöms som riskfyllda och därmed undertrycks.

Vid skydd genom urval *ökas variansen* för skattningarna, vid totalundersökningar från noll. Den reducerade precisionen kan medföra att viss analys inte blir användbar. Även lokal omkodning, addition av brus, slumpmässig avrundning och PRAM leder till ett ökat slumpmässigt fel.

[40] NUTS står för Nomenklatur för statistiska territoriella enheter (på engelska Nomenclature of Units for Territorial Statistics) och utgör EU:s regionala indelning.

Åtgärderna i röjandeskyddet kan komplicera den statistiska analysen. För att tillförlitlig inferens[41] ska kunna göras kan analysmetodiken behöva modifieras, vilket kan vara kostsamt eller ogörligt för mikrodataanvändarna. Ett exempel är att addition av brus gör att den vanliga skattningen av regressionskoefficienter inte är väntevärdesriktig.

[41] Slutledning om populationsegenskaper utifrån observerat material, till exempel skattning av parametrar utifrån en urvalsundersökning eller ett registermaterial.

11 It-verktyg

Detta kapitel introducerar tillgängliga it-verktyg för röjandekontroll. Endast översiktliga beskrivningar av nuläget ges, eftersom förutsättningarna inom it-området är snabbt föränderliga. För mer detaljerade beskrivningar hänvisas till webbsidor med manualer och dylikt eller till kontakter med ägarna till de olika verktygen.

11.1 τ-ARGUS

It-verktyget τ-ARGUS används för att aggregera mikrodata till tabelldata och sedan skydda dessa eller för att skydda redan tabellerade data. τ-ARGUS har till största delen utvecklats av den nederländska statistikbyrån Centraal Bureau voor de Statistiek (CBS)/Statistics Netherlands, bland annat inom flera projekt som har finansierats av EU. Programmet är gratis och finns att ladda ned tillsammans med en användarmanual på CBS webbplats.[42] τ-ARGUS har ett grafiskt gränssnitt vilket gör det förhållandevis användarvänligt. Programmet är gjort enbart för Microsoft Windows 2000 eller senare, så användare av andra operativsystem får använda en annan programvara. En öppen-källkod-variant av τ-ARGUS har dock nyligen tagits fram i en första version. Den varianten av τ-ARGUS fungerar för både Windows-baserade och Unix-/Linux-baserade plattformar.

τ-ARGUS kan bland annat hantera riskbedömning med tröskelvärdesreglerna 1 och 3 (se avsnitt 5.1, cell- eller marginalfrekvens kontrolleras mot tröskelvärde), aggregering av rader och kolumner, primär- och sekundärundertryckning av celler och olika typer av avrundning. Programmet kan ta hänsyn till koalitioner vid undertryckning, till exempel så kallade holdings (se avsnitt 4.1). Designvikter kan beaktas vid urvalsundersökningar. Olika kostnadsfunktioner för bedömning och minimering av informationsförlust kan specificeras.

Moduler för sekundärundertryckning

I det följande beskrivs kort de olika modulerna i τ-ARGUS som behövs för sekundärundertryckning.

Hypercube (algoritmen GHMITER)

Modulen Hypercube har en snabb metod som till skillnad från övriga metoder inte ställer krav på ett externt beräkningsprogram för minimering av informationsförlusten. Metoden har ofta en tendens att undertrycka flera celler än nödvändigt och överbeskydda tabellen. Algoritmerna fungerar dock för tabeller med ända upp till sju dimensioner.

Modular (algoritmen HiTaS)

Modulen Modular kräver en fristående kommersiell LP-solver för att fungera (fram till 2014 är Cplex och Xpress de LP-solvers som fungerat), vilket är förenat med en viss kostnad. En LP-solver är ett beräkningsprogram som använder linjärprogrammering, som tillämpas för optimering utifrån en matematisk modell som representeras via linjära relationer. Modular använder vid behov Integer Linear Programming (ILP) på ”undertabeller” för att skydda hela tabellen. Med hierarkiska tabeller optimerar Modular informationsförlusten för varje tabell på lägsta hierarkiska nivå. Algoritmerna fungerar för tabeller med maximalt tre dimensioner.

Optimal

Modulen Optimal kräver en fristående kommersiell LP-solver. Optimal använder metoden ILP. Algoritmen minimerar informationsförlusten och kontrollerar skyddsmönstret på alla

[42] http://neon.vb.cbs.nl/casc/.

nivåer för hierarkiska tabeller. Optimal kan ta väldigt lång tid, men det finns en möjlighet att avgränsa tidsåtgången och acceptera en lösning som nästan är optimal med avseende på informationsförlusten.

Network

Modulen Network hanterar nätverksflöden och kan endast användas med tvådimensionella tabeller med en hierarkisk variabel. Singleton-problematiken hanteras inte. Metoden genererar en nära optimal lösning och använder sig av LP-solvers som är gratis (PPRN eller Dijkstra).

Jämförelser mellan olika moduler för sekundärundertryckning

En begränsning med modulen Network är att den endast kan användas med tvådimensionella tabeller med en hierarkisk variabel.

Hypercube har nackdelen att ge en större informationsförlust än Modular och Optimal, vilka minimerar informationsförlusten eftersom de undertrycker färre celler än Hypercube. Optimal kan ta mycket lång tid, men garanterar en optimal eller nästan optimal lösning och ett fullgott skydd.

Både Hypercube och Modular bryter ned och utför röjandekontroll på den lägsta hierarkiska nivån för hierarkiska tabeller. Det betyder att vissa primärundertryckta celler kanske inte skyddas helt (se Statistics Netherlands 2011). Efter utförd sekundär-undertryckning kan en övervakningsrutin utföras av τ-ARGUS som visar på vilka celler som inte har fullgott skydd.

Hypercube och Modular kan hantera celler med negativa värden. Dessa moduler har även implementerade lösningar för singleton-problematiken, vilket Network saknar. Hypercube och Modular kan också användas interaktivt i τ-ARGUS för att undertrycka länkade tabeller.

Sammantaget kan Modular bedömas vara den mest användbara modulen i τ-ARGUS för sekundärundertryckning, givet att det finns tillgång till en lämplig LP-solver. För en detaljerad jämförelse mellan olika algoritmer, se Giessing (2013).

Modul för kontrollerad avrundning

En annan metod som kan användas för skydd av tabeller är kontrollerad avrundning, se avsnitt 7.3. Denna metod lämpar sig för frekvenstabeller och kräver en LP-solver till τ-ARGUS. För en given tabell kan det existera mer än en lösning med kontrollerad avrundning. Vilken som helst av dessa är godtagbar, men det finns en optimal lösning som ger minst informationsförlust. Denna kan dock ta lång tid att räkna fram för större tabeller, då den beräknas iterativt. Av den anledningen kan programmet avbryta beräkningarna efter vald tidsgräns eller när den första godtagbara lösningen har hittats, eftersom denna oftast är nära den optimala. Om ingen godtagbar lösning hittas, kan programmet med hjälp av en metod som kallas RAPID skapa en approximativ lösning.

Funktionalitet som saknas

τ-ARGUS har en bred funktionalitet, men saknar också en del funktioner. Bland annat finns ingen funktion som låter programmet avrunda endast små frekvenser. Det finns inte heller någon metodik implementerad för hantering av gruppröjanden (tröskelvärdesregel 2, se avsnitt 5.1). Vidare saknas funktionalitet för att hantera länkade tabeller som inte kan reduceras till en ”supertabell”. Detta kan dock göras via ett SAS-tillägg skapat av den tyska centralbyrån (Destatis), se Schmidt och Giessing (2011).

11.2 µ-ARGUS

It-verktyget µ-ARGUS är τ-ARGUS motsvarighet för mikrodata och har vissa algoritmer gemensamma med τ-ARGUS. µ-ARGUS är framtaget enbart för Microsoft Windows 2000 eller senare. Programmet hanterar röjandekontroll av mikrodata. Exempelvis kan det förhindra att en individs identitet röjs i en mikrodatafil. Programmet är gratis och finns liksom användarmanual tillgängligt för nedladdning på CBS webbplats.[43]

µ-ARGUS innefattar funktionalitet för röjanderiskbedömning och för skydd genom bland annat global omkodning, lokal undertryckning, avrundning, mikroaggregering och PRAM (se en beskrivning av dessa skyddsmetoder i avsnitt 10.3). Programmet beaktar storleken på informationsförlusten vid skydd av mikrodata. Utdata från µ-ARGUS består av en skyddad mikrodatafil och tillhörande rapportfil.

11.3 R-program

R är en programvara och programmeringsmiljö som används till statistiska beräkningar och grafisk presentation av dessa. R-program (liksom källkod) är gratis och finns att ladda ned på R-projektets webbplats.[44] R finns för Windows-, Mac- och Unix-baserade plattformar. Till R finns olika paket som användare ideellt ställt till allmänt förfogande.

sdcTable

Röjandekontrollsprogrammet sdcTable har utvecklats på den österrikiska statistikbyrån och bygger på lösningar från τ-ARGUS, men saknar en del av dess funktionalitet i nuläget. Paketet kan användas för både primär- och sekundärundertryckning utifrån röjanderiskbedömning med tröskelvärdesregeln, *p* %-regeln och dominansregeln. Skyddet gäller endast mot exakta röjanden, inte mot inferensröjanden. En annan brist är att singleton-problematiken inte hanteras. En gratis LP-solver, GLPK, finns att ladda ned. Metoden ILP tillämpas. Algoritmerna Hypercube, HiTaS och Optimal kan användas för sekundärundertryckning inom sdcTable.

De främsta fördelarna med sdcTable jämfört med τ-ARGUS har varit att mjukvaran inte begränsas till Windows-plattformen, att gratis LP-solvers finns att tillgå och att källkoden kan studeras för ökad förståelse av algoritmerna. En betydande nackdel är att prestandan ännu inte är så hög, varför beräkningstiderna inte tillåter röjandekontroll av lika stora tabeller som för τ-ARGUS. Paketet har vidare inget grafiskt användargränssnitt, utan kräver arbete i batch-läge och därmed förtrogenhet med R.

sdcMicro

För röjandekontroll av mikrodata finns R-paketet sdcMicro att tillgå. Det har utvecklats på Wiens tekniska universitet och den österrikiska statistikbyrån och bygger på funktionalitet från ARGUS-program (µ-ARGUS). Paketet ger funktionalitet för bedömning av röjanderisker och för skydd av mikrodata genom bland annat omkodning, lokal undertryckning, addition av brus, mikroaggregering och PRAM (se avsnitt 10.3). Ett grafiskt användargränssnitt är tillgängligt via sdcMicroGUI.

11.4 Tilläggsprogram för beräkning

Vissa moduler i ARGUS-programmen kräver en extern LP-solver (se förklaring i avsnitt 11.1) för att fungera. I τ-ARGUS krävs en extern LP-solver för alla moduler för

[43] http://neon.vb.cbs.nl/casc/.

[44] www.r-project.org.

sekundärundertryckning utom en. Även för kontrollerad avrundning fordras en LP-solver. De kommersiella programmen Xpress och Cplex har den beräkningskapacitet som krävs.

En öppen-källkod-variant av τ-ARGUS har dock nyligen tagits fram i en första version. Den varianten av τ-ARGUS fungerar för både Windows-baserade och Unix-/Linux-baserade plattformar. Det ska inte heller vara nödvändigt att använda en kommersiell LP-solver. Exempel på gratis tillgängliga LP-solvers är GLPK, lp_solve och Coin-or linear programming (CLP).

11.5 Bifrost

Statistiska centralbyrån har utvecklat Bifrost som ett standardverktyg för röjandekontroll. Tanken är att så många produkter som möjligt ska använda samma system för röjandekontroll i syfte att kvalitetssäkra röjandekontrollen och underlätta framtida underhåll av it-stödet. Bifrost kan lämnas ut till andra statistikansvariga myndigheter efter beslut i varje specifikt fall.

Bifrost består av SAS2ARGUS, τ-ARGUS och Xpress (ett program som kräver egen licens). Programmet SAS2ARGUS utgörs av en samling makron i programsystemet SAS som möjliggör användandet av τ-ARGUS i en SAS-miljö. Programmet skapar de filer som behövs för röjandekontroll med τ-ARGUS och anropar sedan τ-ARGUS i batch-läge. Det innebär att τ-ARGUS inte behöver öppnas interaktivt. Programmet SAS2ARGUS kan importera resultat från τ-ARGUS till arbetskatalogen i SAS och kan även skriva ut loggen från τ-ARGUS i SAS-loggen. Med SAS2ARGUS kan de flesta funktionaliteter i τ-ARGUS användas, se vidare Kraftling (2011). Sekundärundertryckning av länkade tabeller kräver dock interaktivt arbete i τ-ARGUS.

11.6 Andra it-verktyg

G-Confid

It-verktyget G-Confid har tagits fram av den kanadensiska statistikbyrån, Statistics Canada, se Wright (2013). Utvecklingen startade i början av 1980-talet. Föregångarna till G-Confid hette CONFID och CONFID2. G-Confid består av komponenter (procedurer och makron) i SAS programmiljö. Verktyget passar därför bäst i en produktionsmiljö med SAS. Möjliga plattformar är Windows, Unix och Lunix. Ett grafiskt användargränssnitt är tillgängligt via SAS Enterprise Guide. Som LP-solver används SAS/OR. G-Confid kan inköpas från Statistics Canada för 30 000 kanadensiska dollar; support kan köpas till.

Verktyget används för tabelldata och har funktionalitet för riskbedömning och undertryckning i magnitudtabeller. Sekundärundertryckning görs så att informations-förlusten minimeras. För optimering används algoritmen Linear Programming (LP), vilken är effektivare beräkningsmässigt än ILP (som används i τ-ARGUS). För att uppnå fullvärdigt skydd för stora hierarkiska tabeller är därför G-Confid ett bättre alternativ. Undertryckningsmönstret kan också utvärderas via ett särskilt makro.

G-Confid är utformat för Statistics Canadas behov och från början egentligen inte avsett för spridning. En brist med verktyget är att stöd för avrundning inte ingår. Från ett svenskt perspektiv, där frekvenstabeller baserade på registerdata är vanliga, är detta problematiskt. Inte heller cellsummor lika med noll (se avsnitt 5.4) hanteras. Länkade tabeller kräver en hel del manuell hantering. Indata till programmet måste vara mikrodata. Sammantaget erbjuder G-Confid mindre flexibilitet och färre metoder än τ-ARGUS.

SuperCROSS

SuperCROSS är ett tabellerings- och analysverktyg för Windows-plattformar som saluförs av det australiensiska företaget Space-Time Research. Det är användarvänligt med ett grafiskt användargränssnitt. Programmet tillhandahåller vissa funktioner för riskbedömning. Det finns också möjlighet att anropa vissa τ-ARGUS-funktioner från SuperCROSS, till exempel sekundärundertryckning och kontrollerad avrundning, men detta kräver SuperCROSS-licens. Funktionaliteten är begränsad, men en fördel med verktyget är att ABS-metoden för modifiering med slumpnycklar (se avsnitt 7.4) finns implementerad. För mer information om SuperCROSS, se webbplatsen för Space-Time Research.

12 Exempel med råd om hantering

I detta kapitel ges några större exempel på röjandekontroll för viktiga typer av data. De fyra avsnitten följer en uppdelning av tabeller med avseende på två aspekter:

- frekvenstabell eller magnitudtabell
- totalräknade data eller urvalsdata.

Syftet med detta är att dessa aspekter påverkar vilka metoder som ska användas i röjandekontrollsprocessen. Varje aspekt har två möjliga alternativ, vilket ger totalt fyra olika möjliga kategorier av tabeller. Det är en grov uppdelning och givetvis finns undantag; ibland måste användaren konsultera flera avsnitt nedan för att utföra röjandekontrollen. Exemplen visar dock de olika frågor som behöver utredas i en given situation. Underavsnitten om riskscenario, riskbedömning, skyddsmetoder och implementering pekar ut vad som behöver beaktas för varje tabell.

Inom respektive kategori kan specifika indelningar förekomma. För frekvenstabeller görs en indelning efter förekomst av variabler som kan betraktas som potentiella målvariabler, för magnitudtabeller görs en indelning efter om målvariablerna kan anta negativa värden. För varje statistikmaterial behöver ställning tas till vilket eller vilka parametervärden som ska användas vid riskbedömning och skydd samt om det behövs kontroll för koalitioner (se kapitel 5).

12.1 Frekvenstabeller med totalräknade data

Endast nyckelvariabler

Riskscenario

Det här fallet rör tabeller som enbart spänns upp med hjälp av nyckelvariabler, det vill säga indirekt identifierande variabler. Med hjälp av en sådan tabell framkommer inte någon ny information om objekten, utan all information är redan uppenbart känd. Det som då återstår är risken för att någon individ kan identifieras och känna sig utpekad och därmed lida men. Det är i tabeller med små frekvenser, till exempel 1:or och 2:or, som det finns risk att de individer som dessa låga frekvenser avser ska känna sig utpekade. Då ska dock beaktas att för att någon ska veta vilket objekt som en 1:a avser så behöver den också känna till det aktuella objektets värde på de variabler som definierar tabellen.

Metoder för riskbedömning

Det är celler med låga frekvenser som primärt kan vara riskceller, och därför används tröskelvärdesregeln. Enligt tröskelvärdesregeln är en cell en riskcell om frekvensen understiger ett visst tröskelvärde. Tröskelvärdet får vara lägst 3. Ett högre tröskelvärde kan vara motiverat om risken för koalitioner är stor.

Skyddsmetoder

Om tabellen vid en första tabellering visar sig innehålla många låga frekvenser och om den uppspännande variabelns kategorier på ett naturligt sätt kan omdefinieras så bör detta göras (aggregering).

Ett alternativ kan vara stokastisk väntevärdesriktig avrundning av små frekvenser med bibehållna marginaler. Marginalfrekvenser lägre än tröskelvärdet avrundas på samma sätt. Undertryckning är sällan lämplig i fall som detta, eftersom informationsförlusten vanligen blir för stor (se Hundepool, 2012, s. 193).

Bibehållna marginaler innebär att för tabeller som innehåller marginalsummor större än tröskelvärdet så ska dessa utgöra de ursprungliga marginalerna och inte summan av de avrundade värdena. Fördelen med att behålla de ursprungliga marginalerna är att det inte kommer att uppstå diskrepanser mellan tabeller som har gemensamma variabler, utan konsistensen bevaras. Nackdelen är att tabellen ändå inte är helt skyddad och att det kan uppstå situationer där det går att bakvägen räkna ut det ursprungliga värdet i en cell via marginalerna. Det ska dock vägas mot det faktum att det både för statistikproducenten och för användarna av statistiken kan innebära svårigheter att hantera tabeller med olika marginaler. Risken för att någon via marginalerna ska kunna räkna ut de ursprungliga värdena är dessutom mindre än vid deterministisk avrundning.

Implementering och it-verktyg

Tröskelvärdesregeln finns implementerad i τ-ARGUS. Stokastisk väntevärdesriktig avrundning av små frekvenser ingår dock inte i τ-ARGUS, utan skulle få programmeras separat. Det ska helst vara möjligt att reproducera exakt samma tabell trots att avrundningen görs slumpmässigt.

Exempel

Tabell 12.1 är ett exempel på en frekvenstabell som endast innehåller indirekt identifierande nyckelvariabler. Detta är också ett exempel på att det inte alltid är entydigt om en variabel är trolig som indirekt identifierande eller som målvariabel.

Det finns till exempel endast en kvinna som är 16 år och är gift eller har registrerat partnerskap. Ingen kan lära sig någon ny information om denna kvinna, eftersom det skulle behövas uppgift om hennes ålder och civilstånd för att visa vem denna 1:a avser.

Tabell 12.1 Utdrag ur Folkmängd efter kön, civilstånd och ålder (15–19-åringar) i ettårsklasser den 31 december 2006

		Ogifta	Gifta	Skilda	Änkor/änklingar	**Totalt**
15 år	Män	66 672	0	0	0	66 672
	Kvinnor	62 779	0	0	0	62 779
16 år	Män	67 172	0	0	0	67 172
	Kvinnor	63 744	1	0	0	63 745
17 år	Män	63 729	0	0	0	63 729
	Kvinnor	60 229	6	1	0	60 236
18 år	Män	62 325	38	1	0	62 364
	Kvinnor	58 361	346	0	0	58 707
19 år	Män	58 088	152	0	0	58 240
	Kvinnor	55 081	979	17	1	56 078

Väntevärdesriktig slumpmässig avrundning av alla 1:or resulterar i tabell 12.2, där det i den publicerade tabellen skulle stå 0 *eller* 3 (0 med sannolikheten 2/3 och 3 med sannolikheten 1/3):

Tabell 12.2 Utdrag ur Folkmängd efter kön, civilstånd och ålder (15–19-åringar) i ettårsklasser den 31 december 2006, med stokastisk avrundning

		Ogifta	Gifta	Skilda	Änkor/Änklingar	**Totalt**
15 år	Män	66 672	0	0	0	66 672

	Kvinnor	62 779	0	0	0	62 779
16 år	Män	67 172	0	0	0	67 172
	Kvinnor	63 744	0 eller 3	0	0	63 745
17 år	Män	63 729	0	0	0	63 729
	Kvinnor	60 229	6	0 eller 3	0	60 236
18 år	Män	62 325	38	0 eller 3	0	62 364
	Kvinnor	58 361	346	0	0	58 707
19 år	Män	58 088	152	0	0	58 240
	Kvinnor	55 081	979	17	0 eller 3	56 078

För 19-åriga kvinnor kommer det efter avrundningen fortfarande vara möjligt att med hjälp av marginalsumman enkelt härleda att det finns en änka, eftersom 56 078 - 55 081 - 979 - 17 = 1. Motsvarande gäller dock inte för de 16- och 17-åriga kvinnorna eller de 18-åriga männen, eftersom det finns en eller flera sanna 0:or på samma rad. Det går i den publicerade tabellen inte att veta om dessa 0:or var en 0:a, 1:a eller 2:a i den ursprungliga tabellen. För dessa rader i tabellen gäller att om frekvenserna under tröskelvärdet avrundas till en 0:a så går det i efterhand inte att veta vilken 0:a (på aktuell rad) som var en 1:a från början. Avrundas frekvenserna under tröskelvärdet till en 3:a så framgår att de ursprungligen var 1:or, eftersom summan av de avrundade värdena är 2 enheter större än marginalsumman.

Trots de ovan påtalade bristerna med att låta marginalerna vara oförändrade så överväger fördelarna. Syftet är att få bort de iögonfallande låga frekvenserna och därmed till viss del försvåra bakvägsidentifiering.

Både nyckelvariabler och potentiella målvariabler

Riskscenario

Även i detta fall finns en risk att objekt som tillhör tabellceller med små frekvenser kan känna sig utpekade och därmed lida men. Dessutom tillkommer risken att användarna av tabellen kan få ny information om de objekt som ligger till grund för statistiken. Denna situation kan inträffa om minst en av variablerna som spänner upp tabellen utgör ny information som inte redan är känd av den som försöker utföra röjandet, samtidigt som de övriga variablerna är tillräckliga för att identifiera ett objekt. Detta är ett allvarligare riskscenario än scenariot med endast nyckelvariabler. I detta scenario finns det en risk att grupper av objekt kan röjas (se exempel nedan).

Metoder för riskbedömning

Tabellceller som innehåller små frekvenser utgör riskceller, därför används tröskelvärdesregeln. Dessutom bör celler där cellens frekvens är lika med *marginalen* eller *marginalen -1* definieras som riskceller, på grund av risken för gruppröjande.

Skyddsmetoder

Ett första steg är att om möjligt göra om klassindelningen av nyckelvariabler.

Som komplement eller alternativ till aggregering föreslås stokastisk väntevärdesriktig avrundning av små frekvenser med *justerade* marginaler, det vill säga de avrundade cellvärdena summeras till ett nytt marginalvärde. Justeringen av marginalerna motiveras av att riskscenariot är att betrakta som allvarligare jämfört med fallet med endast nyckelvariabler, men det gäller att väga detta mot eventuella olägenheter med att publicera flera tabeller med delvis gemensamma variabler där marginalerna inte överensstämmer.

Avrundning av små frekvenser ger inget skydd för celler med frekvens lika med *marginalen* eller *marginalen -1*. Det är möjligt att undertrycka sådana celler, men sekundärundertryckningen kan orsaka problem och kräver en noggrann analys. Det finns i dagsläget inte någon standardlösning för avvägningen som behövs vid skydd av dessa celler.

Implementering och it-verktyg

Samma it-verktyg krävs som i fallet med endast nyckelvariabler. Tröskelvärdesregeln finns implementerad i τ-ARGUS.

Exempel

I detta konstruerade exempel används kommun, kön och ålder som nyckel och utbildningsnivå som målvariabel. Med hjälp av tabell 12.3 kan vem som helst röja att samtliga män i kommunen i åldersgruppen 25–29 år har låg utbildningsnivå. Vidare kan den man som har utbildningsnivå 2 i åldersgruppen 30–34 år dra slutsatsen att alla andra män i hans åldersgrupp i kommunen har utbildningsnivå 1. Riskabla celler i denna typ av riskscenario definieras alltså inte enbart genom låga cellfrekvenser.

Tabell 12.3 Antal män efter ålder och utbildningsnivå

Kommun A		Utbildningsnivå				**Totalt**
Kön	Ålder	1	2	3	4	
Män	25–29	90	0	0	0	90
	30–34	75	1	0	0	76
	35–39	80	40	10	15	145

Jämfört med individstatistik ska för företag beaktas att den geografiska informationen kan få stort genomslag, till exempel kan stora företag som dominerar på små orter eller i glesbygdsregioner lätt kännas igen. Populationen kan vara sned med avseende på nyckelvariablerna, vilket kan göra vissa företag lättare att identifiera än andra.

12.2 Frekvenstabeller med urvalsdata

Endast nyckelvariabler

Riskscenario

Riskscenariot motsvarar scenariot vid totalräknade data, men det faktum att det är ett urval gör att risken för identifiering är mindre, speciellt vid små urvalsfraktioner och symmetriska fördelningar.

I individundersökningar är det mycket svårt att säkert identifiera en individ i en tabell utan kännedom om att just den individen ingår i urvalet. Med hjälp av den information om urvalsförfarandet som normalt tillhandahålls av statistikproducenten går det knappast att säkert säga att någon individ tillhör ett urval; det måste till annan information som inte brukar publiceras. Den största risken borde därför vara att någon som ingår i undersökningen känner igen sig själv, se text om självidentifiering i avsnitt 6.1.

Ett möjligt scenario är också att någon tycker att en kategori som beskrivs i en tabell är mycket lik någon specifik individ, men det kan inte betraktas som ett utpekande. Eftersom det är ett urval finns alltid en möjlighet att det är någon annan.

Ett urval av företag är inte direkt jämförbart med ett urval av individer. I företags-undersökningar förekommer ofta att stora företag ingår i undersökningen med sannolikhet

1. För dessa företag är inte urvalsförfarandet en skyddsåtgärd i sig och situationen blir därför jämförbar med motsvarande för totalräknade företag, och i stort sett samma metoder för riskbedömning och skydd är lämpliga.

Metoder för riskbedömning

Riskbedömningen görs på skattade (uppräknade) frekvenser, med tröskelvärdesregeln. Skillnaden mot totalundersökta data som redovisas enbart på nyckelvariabler är att ett lägre tröskelvärde kan användas, men minst värdet 3. Det är dock av kvalitetsskäl tämligen osannolikt i praktiken att celler där uppräknade frekvenser har mycket låga värden är aktuella för publicering.

Skyddsmetoder

Eftersom risken med denna typ av tabell är liten, är det ganska troligt att skadeprövningen kommer fram till att inget skydd behövs. Om skydd ska appliceras, eller om det finns objekt som ingår i data med sannolikheten 1, så används samma metoder som i avsnittet om totalräknade data, det vill säga först slås om möjligt redovisningsgrupper ihop och sedan avrundas små värden med bibehållna marginaler.

Implementering och it-verktyg

It-lösningen är densamma som för totalräknade data.

Exempel

Se exemplet i avsnitt 12.1 ovan om totalräknade data.

Både nyckelvariabler och potentiella målvariabler

Riskscenario

Risken för identifiering i en urvalsundersökning av objekt är troligen liten, men risken för skada eller men om ett objekt skulle kunna identifieras eller känna igen sig själv är större när potentiella målvariabler redovisas i tabellen.

Liksom i motsvarande scenario för totalundersökningar finns även problemet att grupper av objekt kan röjas när ett cellvärde är lika med *marginalen* eller *marginalen -1*. När det är frågan om urval förutsätter detta att någon (eller några i koalition) har kännedom om att en hel grupp ingår i urvalet. Den risken borde dock vara betydelselös.

I undersökningar med totalundersökta strata, typiskt företagsundersökningar, där objekten ingår i undersökningen med sannolikhet 1, är riskscenariot detsamma som för totalräknade data. Riskbedömning och skydd ska hanteras på samma sätt.

Metoder för riskbedömning

Riskbedömningen görs på skattade frekvenser. Ett lägre tröskelvärde kan användas än vid motsvarande situation för totalundersökta data, men minst värdet 3. Även här avgör sannolikt kvalitetsskäl i första hand om så låga cellvärden är aktuella för publicering. Celler med frekvens lika med *marginalen* eller *marginalen -1* bör identifieras som riskceller även vid urvalsdata.

Skyddsmetoder

Även här kan skadeprövningen komma fram till att urvalsmekanismen ger ett tillräckligt skydd så att inget ytterligare skydd behövs. Om skydd ska appliceras eller om det finns objekt som ingår i data med sannolikheten 1, så används samma metoder som i avsnittet om totalräknade data, det vill säga först sammanslagning av celler och sedan avrundning av små värden med justerade marginaler. Om skyddsbehovet bedöms som litet kan bibehållna marginaler övervägas.

Implementering och it-verktyg

It-lösningen är densamma som i avsnitt 12.1 om totalräknade data.

Exempel

Exemplet visar en urvalsundersökning vars syfte är att belysa användningen av informationsteknik i svenska företag. Populationen är svenska företag med tio eller flera anställda. Populationen är stratifierad efter bransch och storleksklass, där obundet slumpmässigt urval inom strata används för strata med färre än 200 anställda, medan strata med 200 eller flera anställda totalundersöks. En av undersökningsvariablerna är huruvida företaget använder dator eller inte. Resultatet från 2006 års undersökning ges i nedanstående tabell.

Tabell 12.4 Datoranvändning i företag (med tio anställda eller flera) efter företagsstorlek, år 2006

Antal anställda	Andel företag, procent	Felmarginal, procentenheter
10–19	94	2
20–49	97	2
50–99	100	0
100–199	100	0
200–499	100	0
500–	100	0

Det föreligger inte något röjande i de två minsta storleksklasserna. Andelen är inte 100 procent och dessutom är det rimligen inte känt för användarna vilka företag som ingår i urvalet. För storleksklasserna 50–99 och 100–199 anställda är andelen 100 procent, men inte heller där föreligger något röjande om det kan förutsättas att det inte är känt vilka företag som ingår i urvalet. I de två största storleksklasserna skulle det kunna föreligga röjanderisker (men knappast skaderisker), eftersom användaren troligen genom dokumentation och dylikt vet att samtliga företag i dessa klasser ingår i urvalet. Däremot vet inte användaren vilka företag som utgör bortfall.

12.3 Magnitudtabeller med totalräknade data

Icke-negativa målvariabler

Riskscenario

Typiskt är det celler där endast ett litet antal objekt bidrar till cellsumman som är riskfyllda. Det enklaste exemplet är en cell med endast två objekt som bidrar till den redovisade summan; i detta fall kan vartdera objektet genom att dra bort sina respektive värden från cellsumman röja det andra objektets värde.

Celler med fler än två objekt är inte garanterat säkra. Beroende på hur bidragen inom en cell är fördelade storleksmässigt kan en användare ändå dra mer eller mindre säkra slutsatser om objekt i cellen. I celler med många objekt men där ett eller två objekt bidrar med extremt stora värden relativt sett, kan det vara lätt att dra slutsatser om dessa. Den största risken är att objektet med det näst största värdet i cellen subtraherar sitt värde från

cellens summa och försöker dra slutsatser om objektet med det största värdet (se avsnitt 5.2).

Ett annat riskscenario är koalitioner av två eller flera objekt i en cell. Koalitionens medlemmar subtraherar sina värden från cellsumman för att därigenom försöka röja information om någon utanför koalitionen.

Riskscenariot är detsamma för företagsdata och individdata, men med vissa typiska skillnader. Risken för koalitioner bör tas på större allvar när det gäller företag än när det gäller individer. Risken för sneda fördelningar är större för företag, till exempel när ett litet antal stora företag dominerar inom sin bransch eller region. Fördelningen av värden är ofta mindre sned i individdata, vilket medför dels att det är färre individer som har avvikande värden, dels att storleken på avvikande värden vanligen inte är lika extrema. Vidare är det lättare att identifiera företag, eftersom de är mer offentliga och viss information kan förväntas vara känd om dem, som vilket eller vilka företag som dominerar i sin bransch.

Metoder för riskbedömning

Som primär metod för riskbedömning kan *p* %-regeln användas. Om risken för koalitioner bedöms tillräckligt stor kan *p* %-regeln för koalitioner användas.

Icke-tomma celler där variabeln summerar till noll markeras som riskfyllda.

Typiskt är utgångsläget att summor redovisas. I fall med tabeller som redovisar enklare funktioner av summan, till exempel medelvärden, ska utgås från att antalen i cellerna finns tillgängliga för angriparen. Riskbedömningen görs på den underliggande magnitudtabellen med summor.

Skyddsmetoder

Om möjligt kan övervägas att omdefiniera tabellen för att undvika celler med för hög risk, det vill säga sammanslagning av redovisningsgrupper (aggregering). I annat fall (eller i kombination med aggregering) används undertryckning. För att förhindra att de undertryckta värdena ska gå att räkna ut, undertrycks även ett antal celler som inte i sig är riskceller. I företagssammanhang finns också ibland möjligheten att få enskilda företags tillåtelse att publicera uppgiften genom samtycke (medgivande), enligt kapitel 8.

Implementering och it-verktyg

τ-ARGUS har de nödvändiga funktionerna för identifiering av riskceller: *p* %-regeln med hänsyn till koalitioner samt identifiering av ”nollceller”. Sekundärundertryckning med optimeringsrutiner som tar hänsyn till informationsförlusten finns också i *τ*-ARGUS.

Negativa målvariabler

Riskscenario

I fallet med endast icke-negativa värden som summerar till en celltotal finns det en nedre och en övre gräns för de enskilda bidragens möjliga värden. Detta gör det möjligt för en användare av tabellen att subtrahera kända bidrag från cellsumman för att se vad som återstår och därefter justera dessa gränsvärden. Om det däremot är en ”negativ variabel”, som alltså även kan anta negativa värden, är det inte längre möjligt att dra samma slutsatser när antalet objekt som bidrar till cellsumman är stort.

Om cellen endast innehåller två objekt kan dessa fortfarande röja varandras värden, men är antalet större blir det genast svårare. Som ett enkelt exempel kan tas att tre objekt bidrar till en summa enligt 7-5+3 = 5. Om ett av dessa drar bort sitt värde från summan återstår -2, 10 eller 2, beroende på vem det är som drar bort sitt värde. Eftersom gränserna för de olika bidragen (i teorin) tillåts variera mellan $\pm \infty$ är inte detta en användbar metod, om det inte

går att lägga troliga övre och undre gränser för bidragen med hjälp av någon annan information.

Metoder för riskbedömning

Riskbedömningen görs inte på målvariabeln utan på cellernas frekvenser, med tröskelvärdesregeln och tröskelvärde minst lika med 3. Om risken för koalitioner kan antas vara stor sätts värdet högre. Andra metoder för att hantera negativa variabelvärden har föreslagits i litteraturen, se s. 128 och s. 154 i Statistics Netherlands (2010).

Skyddsmetoder

Risken i tabeller där variabeln kan anta negativa värden är mindre än i tabeller med variabler som bara kan anta icke-negativa värden. Skadeprövningen kan komma fram till att skyddsbehovet är litet. Om skydd ska appliceras används samma metoder som för icke-negativa magnituder, det vill säga aggregering eller undertryckning.

Implementering och it-verktyg

Tröskelvärdesregeln och sekundärundertryckning kan tillämpas med τ-ARGUS.

12.4 Magnitudtabeller med urvalsdata

Icke-negativa målvariabler

Riskscenario

Riskscenariot motsvarar scenariot vid totalräknade data. Det faktum att det är ett urval gör att risken för identifiering är mindre, speciellt vid små urvalsfraktioner och symmetriska fördelningar, och om det kan antas att det inte är känt vilka objekt som ingår i urvalet.

Vid företagsundersökningar är dock designvikterna att beakta då de ofta är framtagna med ledning av en storleksvariabel som kan vara snedfördelad. I företagsundersökningar förekommer ofta att stora företag ingår i undersökningen med sannolikhet 1. För dessa företag är inte urvalsförfarandet en skyddsåtgärd i sig och situationen blir därför jämförbar med motsvarande för totalräknade företag. Samma metoder för riskbedömning och skydd är då lämpliga.

Metoder för riskbedömning

Som primär metod för riskbedömning kan *p* %-regeln användas. Designvikter kan användas i riskbedömningen enligt metoden som beskrivs i avsnitt 5.5. Om risken för koalitioner bedöms stor kan *p* %-regeln för koalitioner användas.

Icke-tomma celler där variabeln summerar till noll markeras som riskfyllda.

Skyddsmetoder

Skadeprövningen kan komma fram till att urvalsmekanismen ger ett tillräckligt skydd, så att inget ytterligare skydd behövs. Om skydd ska appliceras, eller om det finns objekt som ingår i urvalet med sannolikhet 1, så används samma metoder som för totalräknade data, det vill säga aggregering eller undertryckning.

Implementering och it-verktyg

Funktioner för identifiering av riskceller enligt *p* %-regeln med användning av designvikter och med hänsyn till koalitioner samt för identifiering av ”nollceller” finns i τ-ARGUS. Programmet har även utvecklade optimeringsfunktioner för sekundärundertryckning.

Exempel

Exemplet visar en tabell med testdata från företag där målvariabeln är en summavariabel som inte kan anta negativa värden, till exempel omsättning, och bakgrundsvariablerna är

region och storlek. τ-ARGUS används för riskbedömning och skydd av tabellen. Riskbedömningsmetod är *p* %-regeln med *p* = 20 och med hänsyn till att koalitioner av storlek 2 kan förekomma. Designvikterna används i riskbedömningen. Som skyddsmetod används undertryckning. Den rödmarkerade cellen identifierades som en riskcell och ska därför undertryckas. De gulmarkerade cellerna är valda för sekundärundertryckning för att det inte ska gå att räkna ut värdet i den primärt undertryckta cellen. Valet av celler för sekundärundertryckning gjordes med hyperkub-metoden och med tillämpning av säkerhetsintervall, se avsnitt 7.2. Informationsförlusten mättes som summan av de undertryckta cellvärdena, med minskad risk för att marginaler undertrycks.

Tabell 12.5 Omsättning efter storlek och region

		Region				**Totalt**
		Nord	Öst	Väst	Syd	
Storlek	1	1 538	8 935	-	2 479	12 952
	5	9 441 804	7 914 392	7 197 277	6 373 271	30 926 744
	6	8 867 808	6 530 004	6 426 570	5 675 564	27 499 946
	7	9 379 567	6 629 907	7 526 487	6 014 609	29 550 571
	8	10 913 175	8 395 858	9 169 322	6 443 118	34 921 474
	9	17 877 116	16 453 648	22 961 918	15 397 085	72 689 767
	Totalt	56 481 855	45 932 745	53 281 575	39 906 126	195 602 301

Negativa målvariabler

Riskscenario

Riskscenariot motsvarar scenariot vid totalräknade data och negativa målvariabler, med reservation för det eventuella skydd som urvalsförfarandet kan ge. Liksom i tidigare beskrivna scenarier ska objekt med urvalssannolikhet 1 ses som tillhörande ett totalundersökt datamaterial. Dock försvårar förekomsten av negativa värden ett röjande.

Metoder för riskbedömning

Tröskelvärdesregeln ska användas med tröskelvärde minst lika med 3. Om risken för koalitioner kan antas vara stor sätts värdet högre.

Skyddsmetoder

Skadeprövningen kan resultera i att skyddsbehovet bedöms som litet. Om skydd ska appliceras används aggregering eller undertryckning, enligt avsnitt 12.3 om totalräknade data.

Implementering och it-verktyg

Tröskelvärdesregeln och sekundärundertryckning kan tillämpas med τ-ARGUS.

13 Förklaring av några begrepp

Additivitet

Se *Konsistent*.

Angripare

Någon som med eller utan avsikt försöker komma åt skyddad information genom att utnyttja möjligheter till röjande.

Anonymisering

Att ur datamaterial ta bort de uppgifter (*identifierare*) som kan hänföra uppgifter till enskilda objekt i en population av personer, företag eller motsvarande. När en anonymisering gjorts är dataskyddsregleringen ändå tillämplig, eftersom materialet fortfarande kan innehålla personuppgifter. Kan även betyda avidentifiering (jämför nedan), varför förtydligande kan behövas.

Attribuering

Att på något sätt utläsa en egenskap (ett värde på en målvariabel) för ett enskilt objekt i en population. Jämför med *Identifiering*.

Attributröjande

Se *Attribuering*.

Avidentifiering

Att ur datamaterial ta bort alla personuppgifter och uppgifter som kan hänföra uppgifter till enskilda objekt i en population av personer, företag eller motsvarande. När en avidentifiering gjorts är dataskyddsregleringen inte tillämplig, eftersom materialet inte innehåller personuppgifter. Kan även betyda anonymisering (jämför ovan), varför förtydligande kan behövas.

Celler

Betyder i denna handbok de olika aggregat som värdena i en tabell avser, även marginalsummor och liknande.

Direkt identifiering

Identifiering med hjälp av *identifierare*.

Holdings

Ett slags hierarkiska objekt, där varje objekt kan föras till en grupp av flera objekt i en hierarkisk struktur med två eller flera nivåer. Ett exempel är koncerner med dotterföretag i olika län. Risken vid förekomst av holdings är att olika objekt som tillhör samma grupp kan finnas utspridda över flera olika celler i tabellen, vilket i sig kan utgöra en ökad risk för röjande via koalitioner. Ett sätt att skydda mot röjande är att utgå från koncernerna som objekt då röjanderisken ska bedömas, även om det kan bli arbetskrävande att uppdaterat hänföra rätt företag till rätt koncern.

Identifierare

Uppgifter som är väsentligt särskiljande, inte nödvändigtvis unikt, för objekt i en population, såsom person- eller organisationsnummer, namn, adress och telefonnummer.

Identifiering

Att på något sätt urskilja vilket enskilt objekt i en population som ligger bakom en uppgift i ett statistikmaterial. Kallas även *Identitetsröjande.* Jämför med *Attribuering*.

Identitetsröjande

Se *Identifiering*.

Indirekt identifiering

Identifiering på något annat sätt än *direkt identifiering*.

Innanmäte (i en statistiktabell)

Avser i denna handbok mängden av cellerna i tabellens inre, det vill säga tabellens celler utom cellerna i marginalerna.

Koalition

En koalition uppstår om någon kommit över icke offentliga data om två eller flera objekt och då kan dra slutsatser om ytterligare objekt. Läs mer om koalitioner i avsnitt 4.2.1 i stycket vid *Minimum frequency rule* på s. 119 i Statistics Netherlands (2010).

Konsistent

Egenskap hos en skyddsmetod att värden stämmer med varandra som förväntat. Kan avse överensstämmelse mellan olika angivelser av samma värde (till exempel *konsistens mellan tabeller*), eller att summor över till exempel rader och kolumner inom en tabell stämmer med de värden på dem som anges i tabellen (kallas *Summakonsistens* eller *Additivitet*).

Makrodata

Data som är i vid mening sammanräknade över en population av exempelvis individer eller företag, alltså sådana data som normalt redovisas i statistiktabeller, diagram eller kartor.

Marginalfrekvens

I en frekvenstabell: Den redovisade summan av de redovisade frekvenserna för cellerna i en rad, kolumn eller annan motsvarande uppdelning.

Medgivande

Se *Samtycke* (till att efterge sekretess).

Mikrodata

Data som inte är sammanräknade över en population utan avser enskilda observationer för exempelvis individer eller företag.

Parametrar (styrparametrar)

För röjandekontroll: Talvärden eller koder för valbara inställningar av funktionaliteten i metoder och programvara för riskidentifiering och skydd.

Personuppgifter

Varje upplysning som avser en identifierad eller identifierbar fysisk person, varvid en identifierbar fysisk person är en person som direkt eller indirekt kan identifieras särskilt med hänvisning till en identifierare som ett namn, ett identifikationsnummer, en lokaliseringsuppgift eller onlineidentifikatorer eller en eller flera faktorer som är specifika för den fysiska personens fysiska, fysiologiska, genetiska, psykiska, ekonomiska, kulturella eller sociala identitet. (definition från artikel 4.1 i EU:s dataskyddsförordning)). Med personuppgifter avses enligt EU:s dataskyddsförordning endast uppgifter om levande personer (skäl 27 i EU:s dataskyddsförordning).

Publicera

Betyder i denna handbok att *göra offentlig och/eller tillgänglig* utanför statistikproducenten. Det behöver inte innebära spridning på webben eller i tryck och gäller både makro- och mikrodata.

Riskabel (riskfylld)

Egenskapen hos cell, värde, värdeintervall, variabel, redovisningsgrupp eller annat att redovisning där potentiellt kan medföra en inte betydelselös risk för röjande, vilket behöver uppmärksammas i röjandekontrollen.

Riskcell

Cell vars redovisning skulle medföra en inte betydelselös risk för röjande, så att någon förebyggande åtgärd normalt behövs för cellen före publicering.

Röjandescenario

Ett möjligt tillvägagångssätt för en angripare. Exempel: Att subtrahera ett känt variabelvärde från en summa i en cell för att dra slutsatser om övriga objekt i cellen.

Samtycke

Samtycke definieras i artikel 4.11 i EU:s dataskyddsförordning som varje slag av frivillig, specifik, informerad och otvetydig viljeyttring, genom vilken den registrerade, antingen genom ett uttalande eller genom en entydig bekräftande handling, godtar behandling av personuppgifter som rör honom eller henne. Denna beskrivning av ett samtycke kan tjäna som utgångspunkt även när en fysisk eller juridisk person lämnar ett samtycke (medgivande) till att efterge sekretess för en given uppgift. Allmänt gäller att samtycket ska lämnas av den enskilde själv eller, när det gäller företag, av en legal företrädare (firmatecknare). Vidare gäller att samtycket ska avse statistikresultat som offentliggörs under en avgränsad tid. Samtycket ska för att gälla inhämtas på ett sätt som är tydligt för den som lämnar det.

Sekretess

Ett förbud [med stöd i lag] att röja en uppgift, vare sig det sker muntligen, genom utlämnande av en allmän handling eller på något annat sätt (enligt definition från 3 kap. 1 § offentlighets- och sekretesslagen (2009:400)).

Singleton

Med singleton avses att värdet i en cell kommer från endast en uppgiftslämnare. Denna känner naturligtvis till sitt eget värde och kan därför vid exempelvis två undertryckta celler i en rad eller kolumn med hjälp av marginalen räkna ut värdet för den andra undertryckta cellen. Singleton-problematiken behöver därför beaktas vid undertryckning av cellvärden.

Summakonsistens

Se *Konsistent*.

14 Svensk-engelsk ordlista

I det följande ges en ordlista från svenska till engelska inom röjandekontrollsområdet. Några av de svenska orden har knappast använts tidigare. Orden kan dock rekommenderas, mot bakgrund av att svenska myndigheter enligt språklagen (2009:600) har ett särskilt ansvar för att svenska språket används och utvecklas. Syftet med ordlistan är bland annat att underlätta sökning i den internationella litteraturen om röjandekontroll.

Svenska	Engelska
Addition av brus	Noise addition, adding noise
Additivitet, summakonsistens	Additivity, summation consistency
Aggregering	Table redesign, table restructuring, table recoding
Angripare	Intruder, attacker, snooper
Anonymisering	Anonymization
Attribuering	Attribution
Avidentifiering	Deidentification
Avrundning	Rounding
Barnardisering	Barnardisation
Dataväxling, databyte	Data swapping
Deterministisk avrundning, konventionell avrundning	Deterministic rounding, conventional rounding, ordinary rounding
Differentiering	Differencing
Dominansregel, (n, k)-regel	Dominance rule, (n,k) rule, concentration rule
Exakt röjande	Exact disclosure, complete disclosure
Frekvenstabell	Frequency table, table of frequency (count) data
Gruppröjande	Group (attribute) disclosure
Icke-perturbativ	Nonperturbative
Identifierare	Identifier
Identifiering	Identification
Inferensröjande	Inferential disclosure
Informationsförlust	Information loss
Koalition (av uppgiftslämnare)	Coalition (of respondents)
Kontrollerad avrundning	Controlled rounding
Kontrollerad tabelljustering	Controlled tabular adjustment (CTA)
Känsliga uppgifter	Sensitive data
Länkade tabeller	Linked tables
Magnitudtabell	Magnitude table, quantitative table, table of magnitude data
Makrodata	Macrodata
Marginalsumma	Marginal total
Mikroaggregering	Microaggregation
Mikrodata	Microdata
Målvariabel	Target variable
Nyckel	Key, identification key

Nyckelvariabel	Key variable
Offentliga mikrodatafiler	Public use files (PUFs)
Omkodning	Recoding
Personuppgifter	Personal data
Perturbativ	Perturbative
p %-regel	*p*% rule
pq-regel	*pq* rule, prior-posterior (ambiguity) rule
Primärundertryckning	Primary suppression
Publicera, offentliggöra	Publish, disseminate
Riskcell, riskutsatt cell	Risky cell, disclosive cell, sensitive cell
Röjande	Disclosure
Röjandekontroll	Disclosure control, disclosure limitation
Röjanderisk	Disclosure risk, probability of disclosure
Röjandescenario	Disclosure scenario
Samtycke (till att efterge sekretess), medgivande	Consent, waiver
Sekretess	Secrecy, confidentiality
Sekundärundertryckning, konsekvensundertryckning	Secondary suppression, complementary suppression, residual suppression
Skadeprövning	Risk assessment
Skaderisk	Risk of damage, risk of harm
Skuggvariabel	Shadow variable
Skydd av data	Data protection
Stokastisk avrundning, slumpmässig avrundning	Stochastic rounding, random rounding
Tabelldata	Tabular data
Transparens, spridning av information om metod för röjandekontroll	Transparency
Tröskelvärdesregel	Threshold rule, (minimum) frequency rule, *n* rule
Undertryckning	Suppression, cell suppression

15 Referenser

Brandt, M., Franconi, L., Guerke, C., Hundepool, A., Lucarelli, M., Mol, J., Ritchie, F., Seri, G., och Welpton, R. (2010). *Guidelines for the Checking of Output Based on Microdata Research.* Final report of ESSnet Sub-Group on Output SDC.

Brottsförebyggande rådet (2011). *Hot spots för brott i sex svenska städer – en studie av förutsättningarna för platsbaserat polisiärt arbete i Sverige.* Rapport 2011:17. Stockholm: Brottsförebyggande rådet.

Castro, J., och González, J.A. (2011). *Present and Future Research on Controlled Tabular Adjustment.* Paper presented at the Joint UNECE/Eurostat work session on statistical data confidentiality, Tarragona [Spanien], October 2011.

Dandekar, R.A., och Cox, L.H. (2002). *Synthetic Tabular Data: An Alternative to Complementary Cell Suppression.* [Opublicerat manuskript.]

Domingo-Ferrer, J., och Mateo-Sanz, J.M. (2002). *Practical data-oriented microaggregation for statistical disclosure control*. IEEE Transactions on Knowledge and Data Engineering, vol. 14, s. 189–201.

Domingo-Ferrer, J., och Torra, V. (2001). *A quantitative comparison of disclosure control methods for microdata*. Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies, Amsterdam: North-Holland, s. 111–133.

von Essen, U. (2003). *Biobanksforskning – forskares möjligheter att få tillgång till vävnadsmaterial och personuppgifter.* Förvaltningsrättslig tidskrift 2/2003.

European Commission, Eurostat (2014). *Recommendations on the Treatment of Statistical Confidentiality in Tabulated Business Data in Eurostat.* Internal document, draft. Luxembourg, July 2014.

Fraser, B., och Wooton, J. (2005). *A Proposed Method for Confidentialising Tabular Output to Protect Against Differencing.* Paper presented at the Joint UNECE/Eurostat work session on statistical data confidentiality, Geneva [Schweiz], November 2005.

Giessing, S. (2013). *Software Tools for Assessing Disclosure Risk and Producing Lower Risk Tabular Data.* Data Without Boundaries, Deliverable 11.1 – Part B, March 2013. [Delrapport från EU-projekt inom Seventh Framework Programme.]

Groves, R. (2004). *Survey Errors and Survey Costs.* Hoboken, New Jersey: John Wiley & Sons Inc.

Hellners, T., och Malmqvist, B. (2010). *Förvaltningslagen med kommentarer*. Tredje upplagan. Norstedts Juridik AB.

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Nordholt, E.S., Spicer, K., och de Wolf P.-P. (2012). *Statistical Disclosure Control*. Chichester, UK: John Wiley & Sons Ltd.

Kraftling, A. (2011). *SAS2ARGUS User Manual*. Statistiska centralbyrån.

Lenberg, E., Geijer, U., och Tansjö A. (2010). *Offentlighets- och sekretesslagen: en kommentar*. Stockholm: Norstedts Juridik AB.

Marley, J.K., och Leaver, V.L. (2011). *A Method for Confidentialising User-Defined Tables: Statistical Properties and a Risk-Utility Analysis*. International Statistical Institute, Proceedings 58th World Statistical Congress, 2011, Dublin [Irland].

Rådet för den officiella statistiken (2013). *Riktlinjer för europeisk statistik (Code of Practice) – en handledning och exempelsamling.* April 2013.

Samuelson, P. (2004). *Statistikrätt: Om officiell statistik.* Stockholm: Norstedts Juridik AB.

Schmidt, K., och Giessing, S. (2011). *A SAS-Tool for Managing Secondary Cell Suppression on Sets of Linked Tables by τ-ARGUS Modular*. Paper and poster presented at NTTS (Conferences on New Techniques and Technologies for Statistics), Brussels [Belgien], February 2011.

Shlomo, N. (2007). *Statistical Disclosure Control Methods for Census Frequency Tables.* International Statistical Review 75, 2, 199–217.

Shlomo, N., Antal, L., och Elliot, M. (2013). *Disclosure Risk and Data Utility in Flexible Table Generators.* Paper presented at NTTS (Conferences on New Techniques and Technologies for Statistics), Brussels [Belgien], March 2013.

Statistics Netherlands (2010). *Handbook on Statistical Disclosure Control, Version 1.2*. ESSnet SDC. Tillgänglig som http://neon.vb.cbs.nl/casc/handbook.htm.

Statistics Netherlands (2011). *τ-ARGUS User's Manual, Version 3.5*. ESSnet SDC. Tillgänglig som http://neon.vb.cbs.nl/casc/tau.htm.

Statistiska centralbyrån (2001a). *Kvalitetsbegrepp och riktlinjer för kvalitetsdeklaration av officiell statistik.* Meddelanden i samordningsfrågor för Sveriges officiella statistik, MIS 2001:1.

Statistiska centralbyrån (2001b). *Statistisk röjandekontroll av tabeller, databaser och kartor*. Current Best Methods [CBM], september 2001.

Statistiska centralbyrån (2007). *Jämförelse av röjanderiskmått för tabeller*. Research and Development – Methodology reports from Statistics Sweden, 2007:1.

Thompson, G., Broadfoot, S., och Elazar, D. (2013). *Methodology for the Automatic Confidentialisation of Statistical Outputs from Remote Servers at the Australian Bureau of Statistics.* Paper presented at the Joint UNECE/Eurostat work session on statistical data confidentiality, Ottawa [Kanada], October 2013.

Willenborg, L., och de Waal, T. (2000). *Elements of Statistical Disclosure Control.* Series: Lecture Notes in Statistics. New York: Springer-Verlag.

Wright, P. (2013). *G-Confid: Turning the tables on disclosure risk*. Paper presented at the Joint UNECE/Eurostat work session on statistical data confidentiality, Ottawa [Kanada], October 2013.

Bilaga: Begäran om samtycke till att efterge sekretess

FÖRETAGETS NAMN *(motsv.)*
Ev. precisering, en eller två rader
ADRESS
POSTORT

Om färgmarkeringarna:
Svart: *Fast text (tas bort endast då den inte är relevant).*
Röd*: Text som ska anpassas till den aktuella situationen och ändras eller helt tas bort.*
Blå*: Instruktioner (tas alltid bort).*

Månad år

TILL FIRMATECKNARE

Undersökningens namn *anges alltid*

Offentliggörande av statistikresultat

Ni har inom ramen för Undersökningens namn lämnat uppgifter till statistik om ämnesområdet, som Myndigheten framställer med Uppdragsgivarens namn som uppdragsgivare. *Ange ämnesområdet på ett för företagen lättförståeligt sätt.* Uppgifterna skyddas av statistiksekretessen i 24 kap. 8 § offentlighets- och sekretesslagen (2009:400), vilket innebär att Myndigheten inte får föra vidare uppgifter om enskilda uppgiftslämnare.

För att statistiken ska bli till bästa nytta önskar Uppdragsgivarens namn och Myndigheten ert samtycke att få offentliggöra statistikresultat uppdelat på kommuner/branscher/… . *Ange benämning på uppdelning(ar) som avses.* Se information på baksidan. Samtycket är frivilligt och omfattar statistikresultat som offentliggörs t.o.m. månad år. *Ange månad och år, inte senare än två år efter brevets datum.*

Så här gör ni

Skicka in samtyckesförklaringen (se baksidan) till oss i det bifogade svarskuvertet. Kontakta oss gärna om ni önskar ytterligare information.

Tack för er medverkan!

Namnteckning

Namnförtydligande
Produktansvarig/Enhetschef

Var god vänd!

Lämnade uppgifter

Uppgifterna ni lämnat avser Företagets namn med organisationsnummer xxxxxx-xxxx.

Samtyckets ändamål

Offentliggörandet innebär att statistikresultat i form av sammanräknade sifferuppgifter blir tillgängliga offentligt genom internet eller på annat sätt. Med uppdelningen på kommuner/branscher/… kan förhållanden för enskilda uppgiftslämnare komma att framgå. Sekretesskyddet för de uppgifter som ert företag har lämnat innebär att Myndigheten inte får offentliggöra uppgifterna uppdelade på detta sätt utan samtycke från er som uppgiftslämnare. Detta gäller även om motsvarande uppgifter redan har offentliggjorts av er eller någon annan. Samtycket är frivilligt.

Samtycke till att efterge sekretessen i 24 kap. 8 § offentlighets- och sekretesslagen

Jag har tagit del av informationen ovan och intygar att Företagets namn samtycker till att Myndigheten offentliggör statistikresultat på det sätt som anges där. Företagets namn har rätt att ta tillbaka samtycket när som helst och utan att behöva ge någon motivering.

……

Ort och datum

……

Underskrift (behörig firmatecknare)

……..…..

Namnförtydligande