



Statistics Sweden

Statistiska centralbyrån

Jämförelse av röjanderiskmått för tabeller

Helen Lindkvist

Ann-Marie Flygare

Rapportserien **Research and Development – Methodology reports from Statistics Sweden** publicerar rapporter med resultat från SCBs forsknings- och utvecklingsverksamhet. Fokus i rapportserien ligger på metod-utveckling för offentlig statistikproduktion. Rapportserien publicerar bidrag från alla avdelningar inom SCB och är öppen för bidrag som behandlar en vid mängd av olika metodologiska problem.

Utgivna publikationer i serien:

2006:1 **Quantifying the quality of macroeconomic variables**

2006:2 **Stochastic population projections for Sweden**

.

Jämförelse av röjanderiskmått för tabeller

Helen Lindkvist

Ann-Marie Flygare

A comparison of disclosure risk measures for tables

Statistics Sweden
2007

Producent
Producer

SCB, utvecklingsavdelningen
Statistics Sweden, Research and Development Department
SE 701 89 ÖREBRO
+ 46 19 17 60 00

Förfrågningar
Inquiries

Helen Lindkvist, +46 19 17 62 63
helen.lindkvist@scb.se

Ann-Marie Flygare, + 46 19 17 64 16
ann-marie.flygare@scb.se

Det är tillåtet att kopiera och på annat sätt mångfaldiga innehållet i denna publikation.
Om du citerar, var god uppge källan på följande sätt:

Källa: SCB, Utveckling och forskning – Metodrapporter från SCB, *Jämförelse av röganderiskmått för tabeller*.

It is permitted to copy and reproduce the contents in this publication.
When quoting, please state the source as follows:

Source: Statistics Sweden, Research and Development – Methodology Reports from Statistics Sweden,
A comparison of disclosure risk measures for tables.

Förord

Rökjanderisk avser problemet där enskilda objekts värden kan härledas ur publicerad statistik. Risk för ett rökjande av en uppgift kan uppstå vid publicerandet av tabeller. Risken är speciellt stor då antalet objekt bakom en summa i en tabellcell är litet. Rapporten beskriver detta problem och jämför olika ansatser till kvantifiering av rökjanderisken. Sådana kvantifieringsregler används som underlag för beslut om publicering av enskilda tabellvärden. I ett exempel visas även hur kunskap om använda regler för riskbedömning kan utnyttjas för rökjande av uppgift. Rapporten behandlar ett komplicerat och viktigt problem. En fundamental förutsättning för officiell statistikproduktion är skyddandet av den enskilde uppgiftslämnaren.

Statistiska centralbyrån i februari 2007

Folke Carlsson

Friskrivningsklausul

I serien Utveckling och forskning – Metodrapporter från SCB, publicerar SCB resultat från utvecklingsarbete rörande metoder och tekniker för statistikproduktion. Författaren/författarna svarar själva för innehåll och slutsatser.

Innehåll

A separate text in English is provided at the end of the publication, on page 33.

Förord	3
Sammanfattning	7
Inledning	9
Förutsättningar och avgränsningar	11
Säkerhetsregler	13
Tröskelvärdesregeln	13
Dominansregeln	13
p %-regeln	15
Jämförelse av säkerhetsregler	17
Jämförelse av dominans- och p %-reglerna	17
Jämförelse av reglerna med hjälp av trigonometriska figurer	18
Tröskelvärdesregeln	20
Dominansregeln	20
p %-regeln	22
Ett exempel från undersökningen Industrins varuproduktion	25
Rekommenderade säkerhetsregler på andra statistikbyråer	29
Konsekvensen av att säkerhetsreglernas parameter- värden är kända för användarna	30
Slutsatser	31
Referenser	32
In English	33
Summary	33

Sammanfattning

I litteraturen förekommer ett antal olika förslag på sätt att bedöma rökanderisken i tabeller som visar totaler av en kvantitativ variabel (t.ex. total omsättning uppdelat på region och bransch). Vid Statistiska centralbyrån (SCB) används i dag främst dominansregeln för tabeller av denna typ. Det har dock på senare tid framförts kritik i litteraturen mot dominansregeln och istället har p %-regeln förordats. Detta dels p.g.a. att dominansregeln är onödigt restriktiv och att den stämmer dåligt överens med tänkbara riskscenarier. Syftet med denna rapport är därför att få en, utifrån SCB:s perspektiv, fördjupad förståelse för dessa mått. En grundlig beskrivning och jämförelse av tröskelvärdes-, dominans- och p %-reglerna görs. Därefter jämförs reglerna med hjälp av ett datamaterial från undersökningen Industrins varuproduktion.

Inledning

Statistisk röjandekontroll handlar om att ha kontroll på risken att information om en enskild (t.ex. en individ eller ett företag) kan på ett statistiskt sätt röjas i ett datamaterial. Det avser således risken att information kan röjas från ett datamaterial (t.ex. en tabell) med hjälp av matematiska eller statistiska slutledningar.

I aggregerade datamaterial (t.ex. tabeller) görs röjanderiskbedömningen vanligtvis på cellnivå, dvs. röjanderisken bedöms för en cell i taget och inte för hela tabellen på en gång. Ordet röjanderisk implicerar att det är någon form av sannolikhet för röjande som beräknas. Vid tabeller som visar totaler av en kvantitativ variabel (t.ex. total omsättning uppdelat på region och bransch) innebär det dock att en bedömning av huruvida cellen är säker eller osäker görs.

Ett säkerhetsmått i kombination med en gräns används för att definiera en säkerhetsregel, där ett värde under gränsen indikerar att cellen är osäker medan ett värde över gränsen indikerar att den är säker. Det är för att underlätta arbetet med riskbedömning som det binära utfallet säker respektive osäker används trots att säkerhet egentligen kan mätas på en kontinuerlig skala.

I litteraturen förekommer synonymer till orden *osäker* och *(o)säkerhetsregel*, såsom *känslig* och *känslighetsregel*, se t.ex. Willenborg och de Waal (2001). Innebörden är dock densamma.

Vid Statistiska centralbyrån (SCB) används idag i huvudsak en kombination av två säkerhetsregler vid riskbedömning av tabeller för summor av kvantitativa variabler; *tröskelvärdesregeln* och *dominansregeln*. I litteraturen och på några andra statistikbyråer förekommer dock andra varianter, där *p* %-regeln är den regel som utöver dessa två är mest frekvent använd och som fler och fler statistikbyråer går över till att rekommendera. Det har i internationella sammanhang framförts att *p* %-regeln är mer lämplig att använda för detta ändamål, se t.ex. Robertson och Ethier (2002), eftersom det måttet i högre utsträckning mäter det som är intressant att mäta och det faktum att det för flera situationer visar sig att *p* %-regeln är något mindre restriktiv än dominansregeln med bibehållen kontroll på röjanderisken.

Syftet med detta arbete är därför att få en, utifrån SCB:s och svenskt perspektiv, fördjupad förståelse för dessa mått och framförallt för att jämföra p %-regeln med den idag rekommenderade dominansregeln.

I nästa avsnitt ges en utökad beskrivning av de förutsättningar vad gäller t.ex. variabler och studiedesign som vi avgränsar oss till. I det avsnitt som följer därefter presenteras säkerhetsregler som är avsedda att användas för att bedöma om en cell är osäker eller ej. Därefter följer ett avsnitt där dessa säkerhetsregler jämförs, både matematiskt och med hjälp av bilder. En applicering av dessa säkerhetsregler har gjorts på ett datamaterial från undersökningen Industrins varuproduktion, vilken redovisas i det nästföljande avsnittet. Före det avslutande avsnittet diskuteras i två avsnitt vilka säkerhetsregler som används på andra statistikbyråer och konsekvensen av att säkerhetsreglernas parametrar är kända för användarna.

Förutsättningar och avgränsningar

Inom statistisk röjandekontroll görs en uppdelning mellan *direkt identifierande variabler*, *indirekt identifierande variabler* och *övriga variabler*. Exempel på direkt identifierande variabler är personnummer och organisationsnummer. Däremot är det mindre givet vilka variabler som är indirekt identifierande, dvs. vilka variabler en användare av statistiken kan känna till hos objekten och som användaren kan använda för att röja objekten i tabellen. Några exempel är dock kön, ålder, region och bransch.

Ett rimligt antagande är att de variabler som spänner upp tabellen (t.ex. bransch och region) är kända för en användare av statistiken, dvs. är indirekt identifierande. Därför antas att en eller flera användare, utan att studera den aktuella tabellen, har vetskap om t.ex. vilken bransch och vilken region ett specifikt företag tillhör och eller det omvända förhållandet, dvs. att användaren har kännedom om vilka företag som tillhör en specifik bransch och region.

Dessutom är det rimligt att tänka sig att en användare har vetskap om vilket objekt som har det största värdet på responsvariabeln inom den aktuella delpopulationen, t.ex. att ett företag som själv befinner sig i en specifik bransch känner till vilket företag som har den största omsättningen. Den information som eventuellt skulle kunna avslöjas när en sådan tabell publiceras är ett enskilt företags omsättning. Antingen att det exakta värdet röjs, eller att det kan uppskattas inom ett snävt intervall.

Endast tabeller som är baserade på totalräknade data behandlas i denna rapport. Detta kan, ur ett praktiskt perspektiv, ses som en avgränsning som i stort sett omöjliggör en applicering i verkliga situationer. Man ska då ha i åtanke att den situation som studeras här är den, ur röjanderisksynpunkt, värsta tänkbara. I tabeller som framställts med hjälp av icke totalräknade data är det för användaren i många fall okänt vilka objekt vars information som resultaten är baserade på. Om detta är okänt så kan användaren inte heller härleda någon information till något specifikt objekt och något röjande förekommer således inte.

Vi kommer dessutom endast behandla situationer där responsvariabeln antar icke-negativa värden. Anledningen till detta är dels

p.g.a. att säkerhetsreglerna är framtagna för just den typen av variabler men också för att det är svårare för en användare att röja någon information om ett specifikt objekt om det inte finns varken en övre eller undre gräns för möjliga värden.

Se t.ex. Willenborg och de Waal (2001) för ytterligare diskussion om riskbedömning vid icke totalräknade data och av variabler som kan anta negativa värden. Där ges bl.a. förslag på hur röjanderiskbedömning kan göras i tabeller som är baserade på data från urvalsundersökningar när man inte kan förutsätta att det för användarna är okänt vilka objekt som ingår i urvalet. Dessutom ges ett antal förslag på hur variabler som kan anta negativa värden kan hanteras.

Säkerhetsregler

Tröskelvärdesregeln

Tröskelvärdesregeln innebär att ett cellvärde som är baserat på uppgifter från färre än ett visst tröskelvärde anses vara osäker. För en tabell som t.ex. visar total omsättning uppdelat på region och bransch anses de celler som är baserade på omsättningsuppgifter från färre än 3 företag vara osäkra om tröskelvärdesregeln med tröskelvärdet $t = 3$ används.

Exempel 1. Låt säga att en total är 100 och är baserad på uppgifter från 3 objekt med värdena 59, 40 och 1. Vid användning av tröskelvärdesregeln med $t = 3$ anses totalen inte vara osäker eftersom den är baserad på 3 observationer.

Låt

$$S_t = N - t$$

Då gäller att en total, som är baserad på uppgifter från N objekt, är osäker enligt tröskelvärdesregeln med tröskelvärdet t då

$$S_t < 0$$

Om det skulle vara tillräckligt att skydda oss mot situationer där en användare av statistiken kan beräkna ett objekts värde på responsvariabeln exakt skulle vi kunna använda tröskelvärdesregeln. Tröskelvärdet 3 skulle då vara motiverat eftersom för celler som innehåller två objekt kan det ena objektet räkna ut det andra objektets värde med hjälp av den publicerade totalen och det egna värdet. Någon sådan exakt beräkning av de ingående objektens värden kan inte göras då cellens total är baserad på uppgifter från tre eller fler objekt. Detta förutsätter att det inte finns några koalitioner, dvs. några fall där ett eller flera objekt går samman för att genom vetskap om sina värden kunna härleda ett annat objekts värde.

Dominansregeln

Vid användning av tröskelvärdesregeln kan vi skydda oss mot att en användare av statistiken kan beräkna ett objekts värde exakt.

Däremot kan inte tröskelvärdesregeln fånga upp situationer där ett eller flera objekt är dominanta i en cell, med avseende på responsvariabeln. En sådan är beskriven i Exempel 1, där de två största objekten står för tillsammans 99 %, men när tröskelvärdesregeln med tröskelvärdet 3 användes ansågs totalen inte vara osäker. För att undvika sådana situationer finns dominansregeln.

Den regel som idag rekommenderas vid SCB, enligt avsnitt 8 i Statistiska centralbyrån (2001), är *dominansregeln*, ibland även kallad *(n,k)-regeln*. Dominansregeln innebär att om de n största bidragen tillsammans står för mer än k procent av det totala cellvärdet är cellen osäker.

Exempel 2. I en specifik cell presenteras totalen 100 som är baserad på uppgifter från 3 objekt med värdena 59, 40 och 1. Vid användning av dominansregeln med t.ex. $n = 1$ och $k = 60$ anses cellen inte vara osäker eftersom det största värdet är mindre än 60.

Regeln kan även uttryckas matematiskt. Låt, för en specifik cell, $x_{(1)}$ beteckna värdet på responsvariabeln x för den observation med störst värde och $x_{(2)}$ det nästa största värdet, osv. Låt N beteckna antalet observationer i cellen. Vi rangordnar således de N värdena på följande sätt: $x_{(1)} \geq x_{(2)} \geq \dots \geq x_{(N)} \geq 0$ med restriktionen $x_{(1)} > 0$. Försättningsvis förenklar vi framställningen något genom att ta bort parentesen runt indexet, trots att indexets värde avser observationens ordningstal.

Låt

$$S_{(n,k)} = \frac{k}{100} \sum_{i=1}^N x_i - \sum_{i=1}^n x_i$$

Då gäller att en total är osäker enligt dominansregeln med parametrarna n och k då

$$S_{(n,k)} < 0$$

Notera att här har beteckningen n en annan betydelse än inom traditionell urvalsteori. Eftersom begreppet (n,k) -regeln är väl vedertagen inom detta område använder vi oss av den beteckningen.

p %-regeln

Om syftet är att identifiera celler där användare av statistiken har möjlighet att beräkna ett enskilt objekts värde exakt eller approximativt, dvs. inom ett snävt intervall, finns regler som bättre mäter detta än dominansregeln. Anledningen är att dominansregeln inte är avsedd att mäta hur exakt ett enskilt objekts värde kan uppskattas av en användare av statistiken.

Ett mått som bättre mäter detta är p %-regeln. Enligt den är en cell osäker om en övre eller en undre gräns av något av de ingående värdena kan hittas närmare än p procent av aktuell observations sanna värde, dvs. om en användare kan uppskatta något objekts värde, säg objekt t :s värde, inom intervallet

$$x_t \pm \frac{p}{100} x_t$$

där x_t är objekt t :s värde i den aktuella cellen.

Det är de objekt som själva tillhör cellen som har störst möjlighet att uppskatta ett annat objekts värde eftersom objekten har vetskap om sina egna värden och kan använda den informationen för att beräkna de andra objektens värden. För att hitta en övre gräns för något av de andra objektens värden kan det egna värdet subtraheras från celltotalen.

Objekt nr s kan då uppskatta en övre gräns av objekt t :s värde närmare än p procent av objekt t :s sanna värde om

$$\sum_{i=1}^N x_i - x_s < x_t + \frac{p}{100} x_t \quad \Leftrightarrow \quad \sum_{\substack{i=1 \\ i \neq s, t}}^N x_i - \frac{p}{100} x_t < 0$$

Eftersom objekten inte förutsätts ha någon förhandsinformation om de andra objektens värden (förutom att de förväntas vara positiva) räcker det med att fokusera på den övre gränsen.

Vi ser genom att studera ovanstående uttryck att fallet med störst röjanderisk är då vänsterledet är som minst, vilket inträffar då $s = 2$ och $t = 1$. Det betyder att det objekt som har den bästa möjligheten (med det snävaste intervallet) att uppskatta ett annat objekts värde är det objekt som har det näst största värdet. Det objektet har bäst möjlighet att uppskatta det största värdet.

Exempel 3. I en specifik cell presenteras totalen 100 som är baserad på uppgifter från 3 objekt med värdena 59, 40 och 1. Vid använd-

ning av p %-regeln med $p = 10$ anses cellen vara osäker eftersom det objekt med värdet 40 vet att $x_1 \leq 100 - 40 = 60$, vilket ligger närmare än 10 % från det största värdet 59.

Låt

$$S_{p\%} = \sum_{i=3}^N x_i - \frac{p}{100} x_1$$

Då gäller att en celltotal är osäker enligt p %-regeln med parametern p då

$$S_{p\%} < 0$$

(Notera att den första termen i $S_{p\%}$ är lika med 0 då $N \leq 2$)

I litteraturen förekommer en generalisering av p %-regeln vid namn *pq-regeln*, se t.ex. Loeve (2001). Den regeln är uppbyggd på liknande sätt som p %-regeln men innehåller ytterligare en parameter, q . Parametern q talar om hur nära objekten kan uppskatta övriga objekts värden innan de känner till totalen. Vid $q = 100$ förenklas pq -regeln till p %-regeln. Generaliseringen av pq -regeln har främst teoretiska syften och är ovanlig i praktiken (Loeve, 2001). Därför behandlas inte pq -regeln här.

Den beskrivna varianten av p %-regeln förutsätter att det inte finns några koalitioner. Vill man ta i beaktande att sådana finns, t.ex. koalitioner om två objekt, så ska summeringen i högerledets första term inte starta från objekt nummer 3 utan från $m + 2$, där m är antalet objekt i koalitionen.

Jämförelse av säkerhetsregler

Jämförelse av dominans- och p %-reglerna

Ett visst samband finns mellan dominansregeln och p %-regeln. Framförallt gäller att (n,k)-regeln är mer restriktiv än p %-regeln då $p = 100 - k$ vid $n \geq 2$. Med mer restriktiv avses att det är fler eller lika många osäkra celler enligt (n,k)-regeln som enligt p %-regeln. Denna slutsats kan vi dra genom att för $n \geq 2$ studera för vilka värden på p (som en funktion av k) som $S_{(n,k)}$ är mindre än $S_{p\%}$:

$$\begin{aligned}
 S_{(n,k)} - S_{p\%} &< 0 \\
 \frac{k}{100} \sum_{i=1}^N x_i - \sum_{i=1}^n x_i - \sum_{i=3}^N x_i + \frac{p}{100} x_1 &< 0 \\
 \frac{p}{100} x_1 &< \sum_{i=1}^n x_i + \sum_{i=3}^N x_i - \frac{k}{100} \sum_{i=1}^N x_i \\
 \frac{p}{100} x_1 &< \sum_{i=3}^n x_i + \left(1 - \frac{k}{100}\right) \sum_{i=1}^N x_i \\
 p &< \frac{100}{x_1} \sum_{i=3}^n x_i + (100 - k) + \frac{100 - k}{x_1} \sum_{i=2}^N x_i
 \end{aligned}$$

Detta betyder att för $n \geq 2$ gäller att

$$S_{(n,k)} - S_{p\%} < 0 \Leftrightarrow p < \frac{100}{x_1} \sum_{i=3}^n x_i + (100 - k) + \frac{100 - k}{x_1} \sum_{i=2}^N x_i$$

vilket innebär att för $n \geq 2$ är dominansregeln mer restriktiv än p %-regeln om och endast om ovanstående relation mellan p och k är uppfylld. (Notera att den första termen i det andra uttryckets högerled är lika med 0 då $n = 2$)

Detta gäller eftersom vi förutsätter att $0 < k < 100$ och $x_i > 0$ för alla i . I dessa fall är första och sista termen positiva. Av detta följer att (n,k)-regeln är mer restriktiv än p %-regeln då $p = 100 - k$.

Vi kan exempelvis studera vilket värde p ska anta för att (2,k)-regeln ska vara mer restriktiv än p %-regeln genom att göra en omskrivning av relationen mellan p , k och x -värdena:

$$p < \frac{100}{x_1} \sum_{i=3}^2 x_i + \frac{100-k}{x_1} \sum_{i=1}^N x_i = \frac{100-k}{x_1} \sum_{i=1}^N x_i$$

$$\frac{\sum_{i=1}^N x_i}{x_1} > \frac{p}{100-k}$$

Om t.ex. $p = 15$ och $k = 90$ så måste $\sum_{i=1}^N x_i / x_1 > 1,5$ för att dominansregeln fortfarande ska vara mer restriktiv än p %-regeln.

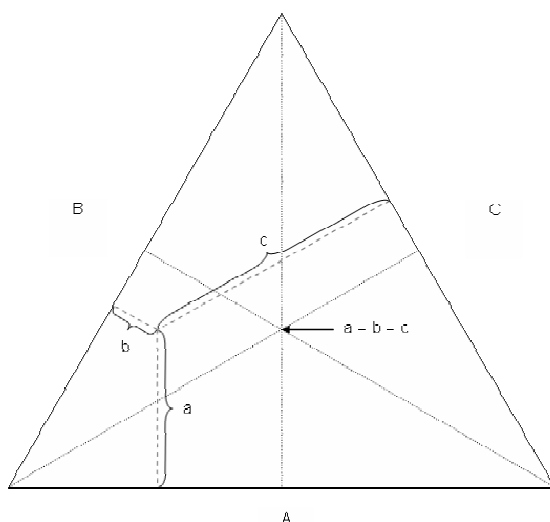
Jämförelse av reglerna med hjälp av trigonometriska figurer

Genom att begränsa oss till totaler som är baserade på uppgifter från 3 objekt kan vi få en god uppfattning om hur dominans- och p %-reglerna fungerar och dess relation till varandra.

Låt dessa tre objekt betecknas med A , B och C och respektive objekts värden med a , b och c . För enkelhets skull tänker vi oss att $a + b + c = 100$. Alla tänkbara fördelningar för dessa tre objekts värden kan illustreras med hjälp av en liksidig triangel, vilket är ett sätt som används i Robertson och Ethier (2002).

Låt alla tänkbara punkter på den liksidiga triangelns yta representera alla möjliga relativa fördelningar för de tre objekts värden. T.ex. är a det vinkelräta avståndet till sida A , som representerar objekt A och motsvarande för b och c . I Figur 1 illustreras två tänkbara fördelningar.

Figur 1. Liksidig triangel som illustrerar två tänkbara fördelningar för de tre objekten A , B och C .



I Figur 1 har den liksidiga triangeln delats in i sex lika stora rätvinkliga trianglar på ett sådant sätt att i den punkt där samtliga sex trianglar gränsar mot varandra gäller att $a = b = c$. Här har således objekten A , B och C samma värde. I de sex rätvinkliga triangelarna gäller olika förhållanden mellan a , b och c . Exempelvis representerar den övre högra triangeln sådana fördelningar där $a > b > c$. I alla tänkbara punkter i den liksidiga triangeln gäller att $a + b + c$ är konstant, dvs. att summan av de tre vinkelräta avståenden till respektive sida i triangeln är konstant, t.ex. 100.

Notera att det egentligen räcker med att använda en av de sex rätvinkliga triangelarna för att illustrera tänkbara fördelningar. Detta om det på förhand bestäms att t.ex. objekt A utgör det objekt med störst värde, objekt B näst störst värde och objekt C minst värde.

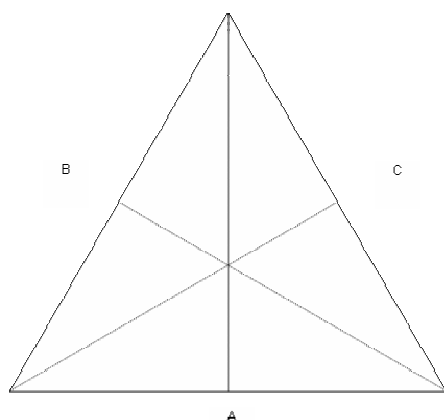
Vi använder denna liksidiga triangel för att visa på vilka områden där tröskelvärdesregeln (S_t), dominansregeln ($S_{(n,k)}$) respektive p %-regeln ($S_{p\%}$) är under eller över 0, dvs. där enligt säkerhetsmåttens totalen är osäker alternativt säker. Fortsättningsvis är de vita fälten i triangeln de områden som representerar fördelningar där säkerhetsmättet har ett värde som är större än 0 (säker) och de gråa fälten där säkerhetsmättet är mindre än 0 (osäker).

Tröskelvärdesregeln

Vid användning av tröskelvärdesregeln med $t = 3$ är alla tänkbara celler (cellers fördelningar) säkra. Den liksidiga triangelns utseende visas i Figur 2, där alla tänkbara fördelningar är säkra.

Eftersom tröskelvärdesregeln endast sätter en restriktion vad gäller antalet objekt spelar det ingen roll hur fördelningen för dessa tre objekts värden ser ut.

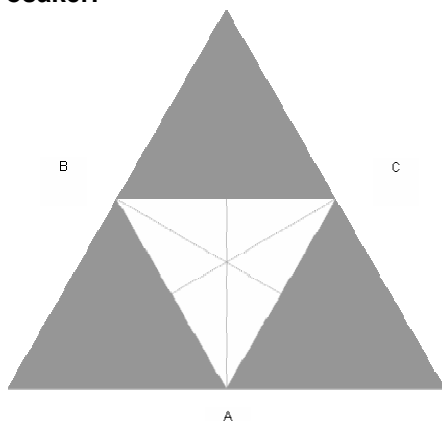
Figur 2. Tröskelvärdesregeln med $t = 3$, där vitt fält illustrerar fördelningar där totalen är säker.



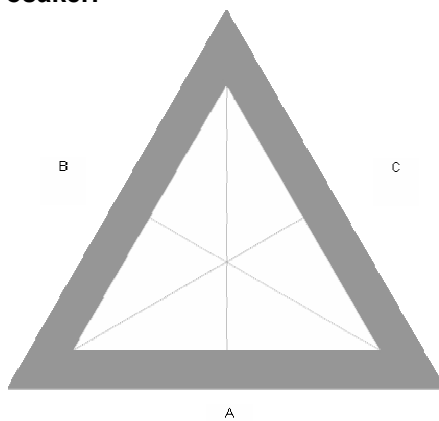
Dominansregeln

Vid användning av dominansregeln (1,50) visas triangelns utseende i Figur 3.

Figur 3. Dominansregeln med $n = 1$ och $k = 50$, där vitt fält illustrerar fördelningar där totalen är säker och grått fält osäker.



Figur 4. Dominansregeln med $n = 2$ och $k = 90$, där vitt fält illustrerar fördelningar där totalen är säker och grått fält osäker.



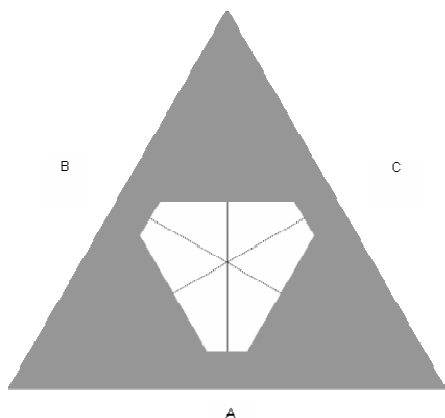
För alla varianter där a , b eller c står för mer än 50 % av totalen anses cellen vara osäkert enligt dominansregeln (1,50). Exempelvis så representerar den övre grå triangeln fördelningar där A står för mer än 50 % av totalen. Den grå triangelns höjd är därmed hälften av höjden på den stora liksidiga triangeln. Här ser vi att en stor andel av den stora triangelns yta är grå.

För dominansregeln (2,90) visas triangeln i Figur 4. Där ser vi att de fall då a , b eller c bidrar med mindre än 10 % (vilket i sin tur betyder att de övriga två står för mer än 90 %) är osäker enligt dominansregeln (2,90).

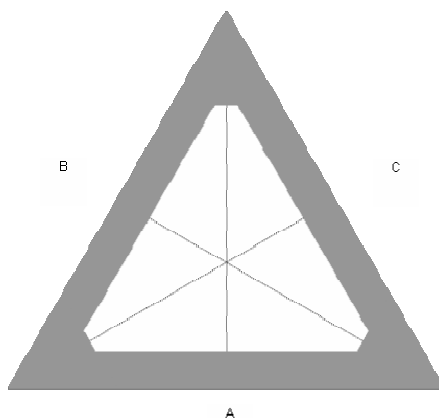
Figur 3 och 4 kan läggas samman, för att illustrera fallet då en kombination av dominansreglerna (1,50) och (2,90) används. En cell är osäker om endera (1,50) eller (2,90) är uppfylld. Triangeln får då det utseende som visas i Figur 5. Den största delen av den grå ytan är grå p.g.a. dominansregeln (1,50) medan en liten del är grå p.g.a. (2,90)-regeln. Naturligtvis bör det inte bara fokuseras på hur stor del av triangeln som är grå för att bedöma hur stort inflytande det får på antalet osäkra celler i en stor tabell eftersom det inte är rimligt att tänka sig att, för ett stort antal totaler, de är jämnt fördelade över triangelns yta.

Ett alternativ till (1,50) kan vara (1,75). Figur 6 visar (1,75) och (2,90). Valet av n och k bestäms utav hur stora rökanderisker som kan accepteras. Det är skillnad mellan (1,50) och (1,75) i hur stor del av ytan som är grå. Som nämnts tidigare betyder detta inte nödvändigtvis att det är stor skillnad i antal säkra totaler eftersom vi inte vet något om fördelningen i den liksidiga triangeln.

Figur 5. Dominansregeln med $n = 1$ och $k = 50$ och $n = 2$ och $k = 90$, där vitt fält illustrerar fördelningar där totalen är säker och grått fält osäker.



Figur 6. Dominansregeln med $n = 1$ och $k = 75$ och $n = 2$ och $k = 90$, där vitt fält utgör fördelningar där totalen är säker och grått fält osäker.

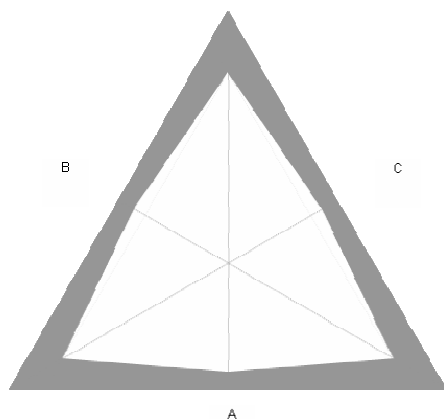


p %-regeln

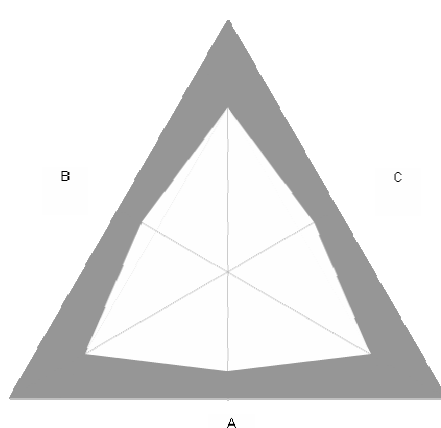
Figur 7 och 8 visar hur den liksidiga triangeln ser ut när $p = 10$ respektive $p = 15$ används.

Figur 7 och 8 liknar i viss utsträckning Figur 4 på ett sådant sätt att det är de fall där ett eller två objekt utgör en liten del av totalen som anses vara osäkra, dvs. att det gråa fältet ligger som en ram runt respektive triangel. Skillnaden mellan figurerna för p %-regeln (Figur 7 och 8) och den för (2,90)-regeln (Figur 4) är att det går att för ett fixt men lågt värde på ett objekt, t.ex. a (säg $a = 10$ i Figur 8), befinna sig i antingen det grå eller det vita fältet beroende på fördelningarna för b och c . Detta betyder att a kan vara lägre om b och c är lika stora men ändå klassas som säker än om t.ex. b är större än c .

Figur 7. p %-regeln med $p = 10$, där vitt fält illustrerar fördelningar där totalen är säker och grått fält osäker.



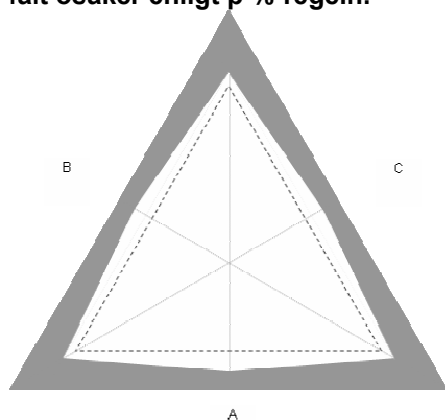
Figur 8. p %-regeln med $p = 15$, där vitt fält illustrerar fördelningar där totalen är säker och grått fält osäker.



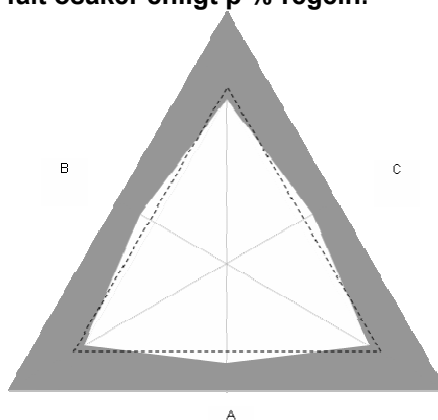
I Figur 9 görs en jämförelse mellan p %-regeln med $p = 10$ och dominansregeln med (2,90). Den grå ytan visar vilka områden som är osäkra enligt p %-regeln medan ytan utanför den streckade liksidiga triangeln visar vilka områden som är osäkra enligt dominansregeln. Här kan det, liksom tidigare konstateras att dominansregeln (2,90) är mer restriktiv (fler osäkra celler) än p %-regeln med $p = 10$.

I Figur 10 jämförs p %-regeln med $p = 15$ och dominansregeln med (2,90). Där kan det konstateras att, för specialfallet med tre objekt, dessa två reglerna är i stort sett identiska.

Figur 9. p %-regeln med $p = 10$ och dominansregeln med (2,90), där vitt fält illustrerar fördelningar där totalen är säker enligt dominansregeln och grått fält osäker enligt p %-regeln.



Figur 10. p %-regeln med $p = 15$ och dominansregeln med (2,90), där vitt fält illustrerar fördelningar där totalen är säker enligt dominansregeln och grått fält osäker enligt p %-regeln.



Ett exempel från undersökningen Industrins varuproduktion

Industrins varuproduktion är en undersökning vars syfte är att årligen belysa den svenska industriproduktionens varufördelning på en detaljerad nivå. Både produktion av varor och industriella tjänster ingår i statistiken. En varuklassificering i enlighet med EU:s kombinerade nomenklatur utnyttjas för rapportering av produktionsuppgifterna. Den innehåller c:a 10 000 varor där det för c:a 4 000 av dessa redovisas produktion. Den kombinerade nomenklaturen är hierarkiskt uppbyggd vilket möjliggör redovisning på varunivå (8- och 9-siffernivå i nomenklaturen) och på olika varugrupsnivåer (t.ex. 4- och 6-siffernivå).

Undersökningsvariabler är leveransvärde, produktions- och leveranskvantitet. Parameter av intresse är totalen. Här studeras variabeln leveransvärde.

I Tabell 1 visas hur de säkerhetsregler som diskuterats tidigare med några olika parametervärden klassificerar de c:a 4 000 varorna (på 8- och 9-siffernivå) för årgångarna 2003 och 2004. För tröskelvärdesregeln visas resultat vid tröskelvärdet 3 eftersom det är ett på SCB (och även internationellt) vanligt förekommande parametervärde. Vad gäller dominansregeln så redovisas (1,50), (1,75) och (2,90) eftersom det främst är kombinationer av dessa som används vid SCB. Vad gäller p %-regeln så redovisas $p = 5, 10, 15$ och 20 . Valet av p beror på hur nära det sanna värdet som SCB kan acceptera att ett objekts värde kan uppskatta. Att använda ett lägre värde än 5% bör inte vara aktuellt. Däremot är det möjligt att 20% i vissa sammanhang inte är tillräckligt.

Här bör det noteras att varor där $\sum_{i=1}^N x_i = 0$ bör bedömas vara osäkra

eftersom samtliga företag som (vanligtvis) producerar den varan har värdet 0 (eftersom x inte kan anta negativa värden). Sådana situationer hanteras dock inte av de regler som tidigare beskrivits. Genom att studera fördelningen för 2003 och 2004 kan det konstateras att fördelningarna för de två årgångarna i stort sett sammanfaller. En stor del, knappt 60% , av de c:a 4 000 varorna har färre än 3 företag.

Tabell 1. Antal osäkra varor enligt olika säkerhetsregler

Regel och parametervärde	2003 (%)	2004 (%)
Tröskelvärdesregeln, $t = 3$	2 316 (58)	2 240 (57)
Dominansregeln, (1,50)	3 385 (85)	3 290 (84)
Dominansregeln, (1,75)	2 641 (66)	2 586 (66)
Dominansregeln, (2,90)	2 985 (75)	2 905 (75)
p %-regeln, $p = 5$	2 732 (69)	2 625 (67)
p %-regeln, $p = 10$	2 883 (72)	2 799 (72)
p %-regeln, $p = 15$	2 998 (75)	2 928 (75)
p %-regeln, $p = 20$	3 097 (78)	3 010 (77)
$\sum x = 0$	17 (0,4)	20 (0,5)
Totalt antal varor	3 988	3 897

Det är av särskilt intresse att studera ett antal korstabeller utifrån Tabell 1 för att se hur reglerna samvarierar med avseende på andelen celler som är osäkra respektive säkra. Detta eftersom reglerna till stor del går omlott. Exempelvis är per definition alla varor som är osäkra enligt tröskelvärdesregeln med tröskelvärdet 3 även osäkra enligt dominansregeln (2,90). Ett undantag är de varor där antalet företag är 1 eller 2 och deras värden är lika med 0.

Fortsättningsvis studerar vi endast 2003 års årgång eftersom årgångarna är i stort sett identiska med avseende på fördelningen av antal osäkra och säkra varor. Eftersom en stor del av varorna endast har ett eller två företag faller en stor del för tröskelvärdesregeln, varför en uppdelning mellan säkra och osäkra varor enligt tröskelvärdesregeln görs.

I Tabell 2 visas en korstabellering mellan tröskelvärdesregeln, dominansregeln (2,90) och dominansregeln (1,50), respektive (1,75). Notera att det är de två första, respektive de två sista, raderna som summerar till 100 %. Som konstaterats tidigare är andelen varor som är osäkra enligt tröskelvärdesregeln stor (58 %). Jämför vi Tabell 2 med Figur 5 kan vi notera att det är en liten andel (0,5 %) som är osäker enligt (2,90) men säker enligt (1,50) och säker enligt tröskelvärdesregeln, vilket i figuren illustreras av den lilla grå yta som beror på (2,90)-regeln men inte (1,50)-regeln.

Tabell 2. Andel osäkra och säkra celler enligt tröskelvärdes- och dominansreglerna

Andel			$N \leq 2$, osäker		$N \geq 3$, säker	
			Dominansregel (2,90)			
			osäker	säker	osäker	säker
Dominans-regel	(1,50)	osäker	58 %	-	17 %	11 %
		säker	-	0,4 %	0,5 %	14 %
	(1,75)	osäker	53 %	-	12 %	2 %
		säker	5 %	0,4 %	5 %	23 %

I Tabell 3 visas fördelningen för 2003 års varor som är säkra enligt tröskelvärdesregeln över fyra varianter av p %-regeln och (2,90)-regeln mot två varianter av (1, k)-regeln. Därmed består Tabell 3 av åtta stycken 2x2-tabeller.

Tabell 3. Andelen varor som är osäkra respektive säkra enligt dominansreglerna (2,90), (1,50) eller (1,75) och p %-regeln för 2003 och $N \geq 3$

Andel			p %-regeln							
			$p = 5 \%$		$p = 10 \%$		$p = 15 \%$		$p = 20 \%$	
			osäker	säker	osäker	säker	osäker	säker	osäker	säker
Domi-nans-regler (2,90)	(1,50)	osäker	26 %	40 %	35 %	31 %	42 %	24 %	48 %	18 %
		säker	-	34 %	-	34 %	-	34 %	-	34 %
	(1,75)	osäker	26 %	19 %	35 %	10 %	41 %	4 %	45 %	0,6 %
		säker	-	55 %	-	55 %	0,3 %	55 %	3 %	52 %

För $p = 5 \%$ och $p = 10 \%$ får vi som förväntat resultatet (se tidigare resonemang vid jämförelse av dominans- och p %-regelerna) att samtliga celler som kategoriseras som osäkra enligt p %-regeln även är kategoriserade som osäkra enligt dominansreglerna. Vi ser att när värdet på p ökar så ökar andelen osäkra varor. Dessutom gäller att de som är osäkra enligt $p = 5 \%$ även är det enligt övriga värden på p .

Det är de fall då p %-regeln och dominansreglerna kategoriserar olika som är av intresse. Detta för att studera hur en eventuell förändring av säkerhetsreglerna skulle påverka antalet osäkra och

säkra varor. Vi ser att för (2,90) och (1,50) så innebär oavsett värdet på p att p %-regeln klassificerar en högre andel säkra celler än vad (2,90) och (1,50) gör. Skillnaden är stor även vid $p = 20$, där det är 18 % som är säkra enligt p %-regeln men osäkra enligt dominans-reglerna.

Vi ser i de två nedre högra deltabellerna att det förefaller som om en kombination av (2,90)- och (1,75)-reglerna i stort sett sammanfaller med $p = 15$ %, eller ännu bättre med $p = 20$ %, med avseende på andelen celler som klassificeras lika vid användning av dominans- eller p %-reglerna. Detta stämmer väl överens med Figur 10 som indikerar att p %-regeln med $p = 15$ och (2,90)-regeln är näst intill identiska.

Rekommenderade säkerhetsregler på andra statistikbyråer

Någon kartläggning har inte gjorts specifikt för detta syfte för att ta reda på vilket mått som andra statistikbyråer rekommenderar att använda för att bedöma röjanderisken i tabeller för summor av kvantitativa variabler.

I samband med skrivandet av Statistiska centralbyrån (2001), skickades ett antal frågor om statistisk röjandekontroll till statistikbyråer (eller motsvarande) i USA, Kanada, Australien, Nederländerna, Storbritannien och Finland. Resultat av detta finns att läsa i Statistiska centralbyrån (2001). Där framgår att på US Bureau of the Census rekommenderas p %-regeln framför dominansregeln. Dessutom framgår att Statistics Canada använder en kombination av dominansregler men att de ansåg att p %-regeln och pq-regeln är bättre. Dessutom framhöll de att de parametrar som används inte publiceras. Australian Bureau of Statistics svarade att de använder en kombination av två dominansregler, nämligen att inget objekt får bidra med mer än 85 % till totala cellvärdet och att två objekt tillsammans inte får bidra med mer än 90 % av totala cellvärdet.

I Felsö, Theeuwes och Wagner (2001) rapporteras resultatet av en enkätundersökning som har skickats ut till alla statistikbyråer i Nordamerika, Nya Zeeland och EU-länderna. Där ges även en sammanfattning av de tidigare undersökningar som genomförts för att studera vilka röjandekontrollmetoder som används. Det var totalt 13 statistikbyråer som svarade på frågorna som gällde tabeller som redovisar totaler av kvantitativa variabler. En av frågorna gällde vilken metod de använder för att identifiera osäkra celler. Av dessa 13 angav 8 att de använder dominansregeln, 3 att de använder p %-regeln, 1 pq-regeln, 1 tröskelvärdesregeln och 1 annan metod. (Det var möjligt att ange mer än en metod).

Domingo-Ferrer och Torra (2004) diskuterar röjanderisk och konstaterar bl.a. att US Census Bureau ändrade 1992 från dominansregeln till p %-regeln och även tyska statistikbyrån gjorde det under 2001.

Även Storbritanniens ONS använder p %-regeln och då i kombination med tröskelvärdesregeln. De lämnar, under inga omständigheter, ut de parametervärden som de använder.

Konsekvensen av att säkerhetsreglernas parametervärden är kända för användarna

I Statistiska centralbyrån (2001) rekommenderas specifika parametervärden vid användning av dominansregeln. Dock är fallet att om en användare känner till vilka parametervärden som har använts så har användaren möjlighet att få mer information om materialet än vad som är önskvärt. Därför rekommenderar bl.a. Willenborg och de Waal (2001) att de parametervärden som används i säkerhetsreglerna inte publiceras. Nedan visas med ett exempel varför parametervärdena inte bör göras kända.

Exempel 4. Antag att p %-regeln används och att det till en specifik cell bidrar $N = 9$ objekt, t.ex. 9 företag, och att totalen 120 har publicerats då den anses vara säker enligt p %-regeln med aktuell parameter. Antag att $x_1 = 65$ och att det företaget som värdet gäller (företag A) känner till att $p = 10$ används liksom sitt eget värde. Antag vidare att $x_2 = 48$, men att det inte är känt för företag A. Företag A känner således följande:

$$x_1 = 65$$

$$x_1 \geq x_2 \geq \dots \geq x_9 \geq 0$$

$$x_1 + x_2 + \dots + x_9 = 120$$

$$S_{10\%} = \sum_{i=3}^9 x_i - \frac{10}{100} \cdot 65 \geq 0 \quad \Rightarrow \quad x_3 + x_4 + \dots + x_9 \geq 6,5$$

Utifrån denna information kan företag A dra slutsatsen att $x_2 \leq 48,5$ och på så sätt finna en övre gräns som ligger närmare än 10 % från det sanna värdet.

Slutsatser

I denna rapport har säkerhetsregler med syfte att bedöma om en total av en kvantitativ variabel är osäker eller ej studerats. De säkerhetsregler som har jämförts är tröskelvärdesregeln, dominansregeln och p %-regeln. Syftet med detta arbete har varit att få en grundligare förståelse för syftet med att använda dessa regler och reglernas innebörd och relation till varandra. Den främsta anledningen till detta är att det på flera håll i litteraturen, i t.ex. Robertson och Ethier (2002), konstateras att dominansregeln egentligen inte fångar upp det riskscenario som vi vill skydda oss emot. Det riskscenario som vi vill skydda oss emot är att när ett enskilt objekts värde kan uppskattas inom ett snävt intervall av någon användare eller något objekt som tillhör samma delpopulation.

Vi har bekräftat med hjälp av både analytiska och empiriska resultat att dominansregeln är mer restriktiv än p %-regeln. Vid användning av dominansregeln med (1,50) och (2,90) så är det (1,50) som står för den största andelen osäkra celler. Effekten på andelen osäkra varor visade sig i materialet från undersökningen Industrins varuproduktion i stort sett vara densamma som vid användning av $p = 15\%$ eller $p = 20\%$ som vid användning av dominansreglerna (2,90) och (1,75).

Referenser

Domingo-Ferrer, J. och Torra, V. (2004). Disclosure risk assessment in statistical data protection. *Journal of Computational and Applied Mathematics*, vol. 164-165, s. 285-293.

Felsö, F., Theeuwes, J. och Wagner, G.G. (2001). Disclosure Limitation Methods in Use: Results of a Survey, i Doyle, P, Lane, J.I., Theeuwes, J.J.M och Zayatz, L.V. (red), *Confidentiality, Disclosure, and Data Access*, Amsterdam: Elsevier.

Loeve, J.A. (2001) Notes on sensitivity measures and protection levels. Research paper, Voorburg: Statistics Netherlands.

Robertson, D. A. och Ethier, R (2002) Cell suppression: Experience and theory, i Domingo-Ferrer, J. (red), *Inference control in statistical databases*, Berlin: Springer-Verlag.

Statiska centralbyrån (2001). Statistisk röjandekontroll av tabeller, databaser och kartor, Örebro: Statistiska centralbyrån.

Willenborg, L. och de Waal, T. (2001) *Elements of Statistical Disclosure Control; Series: Lecture Notes in Statistics*, vol. 155, New York: Springer-Verlag.

In English

Summary

This report includes a number of proposals about how to assess disclosure risks in tables that show totals of a quantitative variable (for instance, total turnover by region and industry). At Statistics Sweden today, the dominance rule is mainly used for such tables. In recent years however, the dominance rule has been criticised in the literature and the p-percent rule has been recommended instead. This is partly because the dominance rule is very restrictive and is not in agreement with probable risk scenarios. The aim of this report is to obtain a better understanding of these measures, from Statistics Sweden's perspective. Thorough descriptions of the threshold, dominance, and p-percent rules are given and comparisons of these rules are made. Afterwards, the rules are compared by means of a data set from the survey Production of Commodities and Industrial Services.

Jämförelse av rökanderiskmått för tabeller

I litteraturen förekommer ett antal olika förslag på sätt att bedöma rökanderisken i tabeller som visar totaler av en kvantitativ variabel (t.ex. total omsättning uppdelat på region och bransch). Vid Statistiska centralbyrån (SCB) används i dag främst dominansregeln för tabeller av denna typ. Det har dock på senare tid framförts kritik i litteraturen mot dominansregeln och istället har p %-regeln förordats. Detta dels p.g.a. att dominansregeln är onödigt restriktiv och att den stämmer dåligt överens med tänkbara riskscenarier.

Syftet med denna rapport är därför att få en, utifrån SCB:s perspektiv, fördjupad förståelse för dessa mått. En grundlig beskrivning och jämförelse av tröskelvärdes-, dominans- och datamaterial från undersökningen Industrins varuproduktion.

ISSN 1653-7149 (online)

Publikationstjänsten:

E-post: publ@scb.se, tfn: 019-17 68 00, fax: 019-17 64 44. Postadress: 701 89 Örebro.

Information och bibliotek: E-post: information@scb.se, tfn: 08-506 948 01, fax: 08-506 948 99.

Försäljning över disk, besöksadress: Biblioteket, Karlavägen 100, Stockholm.

Publication services:

E-mail: publ@scb.se, phone: +46 19 17 68 00, fax: +46 19 17 64 44. Address: SE-701 89 Örebro.

Information and Library: E-mail: information@scb.se, phone: +46 8 506 948 01, fax: +46 8 506 948 99.

Over-the-counter sales: Statistics Sweden, Library, Karlavägen 100, Stockholm, Sweden.