Metoder att få konsistens mellan olika skattningar och samtidigt öka precisionen

Sixten Lundström

SCB

R&D Report
Statistics Sweden
Research - Methods - Development
1993:6

INLEDNING

TILL

R & D report : research, methods, development / Statistics Sweden. – Stockholm : Statistiska centralbyrån, 1988-2004. – Nr. 1988:1-2004:2.
Häri ingår Abstracts : sammanfattningar av metodrapporter från SCB med egen numrering.

Föregångare:

Metodinformation : preliminär rapport från Statistiska centralbyrån. – Stockholm : Statistiska centralbyrån. – 1984-1986. – Nr 1984:1-1986:8.

U/ADB / Statistics Sweden. – Stockholm : Statistiska centralbyrån, 1986-1987. – Nr E24-E26

R & D report : research, methods, development, U/STM / Statistics Sweden. – Stockholm : Statistiska centralbyrån, 1987. – Nr 29-41.

Efterföljare:

Research and development : methodology reports from Statistics Sweden. – Stockholm : Statistiska centralbyrån. – 2006-. – Nr 2006:1-.

R & D Report 1993:6. Metoder att få konsistens mellan olika skattningar och samtidigt öka precisionen / Sixten Lundström.
Digitaliserad av Statistiska centralbyrån (SCB) 2016.

urn:nbn:se:scb-1993-X101OP9306

Metoder att få konsistens mellan olika skattningar och samtidigt öka precisionen

Sixten Lundström

SCB

R&D Report
Statistics Sweden
Research - Methods - Development
1993:6

Från trycket | September 1993
Producent | Statistiska centralbyrån, utvecklingsavdelningen
Ansvarig utgivare | Lars Lyberg
Förfrågningar | Sixten Lundström, tel 019-17 64 96

ISSN 0283-8680

Printed | September 1993
Producer | Statistics Sweden,
Department of Research and Development
S-115 83 Stockholm
Publisher | Lars Lyberg
Inquiries | Sixten Lundström, telephone +46 019 17 64 96

ISSN 0283-8680

Metoder att få konsistens mellan olika skattningar och samtidigt öka precisionen: Kalibrering av vikter, sammanjämkning av tabeller och raking-ratio-metoden.*)

1. Inledning

SCB har som central myndighet till uppgift att samordna den officiella statistiken. En sådan samordning kan bestå i att närbesläktade undersökningar samordnar genomförandet eller helt enkelt slås samman, den kan bestå i ett samutnyttjande av data eller enbart samordning av definitioner. Samordning kan ge kostnadsminskningar och ökat värde åt respektive undersökning. Det kan också öka statistikanvändningen då användarna slipper störas av misstämmelser.

Även om man genomför sådana åtgärder kan man erhålla irriterande skillnader mellan olika skattningar av samma storhet. Det kan bero på att man använt olika metoder - särskilt vanligt är det när en urvalsundersökning skattar storheter som också kan erhållas från register.

För att erhålla konsistenta skattningar krävs, förutom att SCBaren är uppmärksam på vilken statistik som SCB publicerar, att det finns bra tekniker för att uppnå detta. I föreliggande rapport ska vi presentera olika metoder att samutnyttja data i syfte att slippa nämnda störningar och samtidigt öka precisionen i skattningarna.

Det finns två huvudtyper av metoder. I den ena förändras (kalibreras) vikterna från urvalsdesignen så att de enkla viktade summorna som sedan används i estimationen ger konsistenta skattningar och i den andra metoden anpassas/justeras estimaten till kända storheter. Deville & Särndal [1992] beskriver den första metoden. Här ska vi ge en sammanfattning av den, samt även en beskrivning av datorprogrammet CALMAR. Den andra typen innehåller sammanjämkning av tabeller och marginaler (se Lütjohann [1985]) samt raking-ratio-metoden (se Deming & Stephan [1940]).

Den generella metoden i Deville & Särndal [1992] ger som specialfall t ex en generaliserad regressionsestimator eller en raking-ratio-estimator. Det kan dock finnas situationer då kalibrerade vikter inte efterfrågas specifikt och där det är praktiskt lättare att använda en metod från den andra gruppen. Även möjligheten att beräkna medelfel för skattningarna är en faktor i valet mellan metoder.

2. Exemplifiering av problemen.

Exempel 1. Vi vill skatta antalet personer i varje cell i korstabellen kön*ålder*inkomst i riket, där ålder och inkomst är kategoriserade variabler. En statistisk undersökning har gett oss värden på dessa variabler för ett urval av personer. Befolkningsregistret ger oss dessutom populationsuppgifter om antalet personer i respektive kön och åldersklass, d v s vi känner marginaler i den efterfrågade tabellen.

---

*) Jag har av Klas Lindström (ADB), fått god hjälp med tolkningen av SAS-programmet CALMAR. Daniel Thorburn (Stockholms universitet), Harry Lütjohann (AM/STM) och Carl-Eric Särndal (U/PLAN) har gett mig mycket värdefulla synpunkter på rapporten.

Exempel 2. Vi vill dessutom skatta totala inkomsten [1] för män resp kvinnor. Ett register innehåller för varje person i riket uppgift om inkomst av annan typ (fortsättningsvis kallar vi den för "registerinkomst").

Många typer av samplingstrategier (urvalsdesign + estimation) ger summor som inte överensstämmer med de kända marginalerna i exempel 1. Statistikanvändare kan upptäcka denna inkonsistens och bli störd av detta.

Om vi sätter in registerinkomst för urvalsobjekten i estimatorn i exempel 2 ger många samplingstrategier ett resultat som avviker från totalen som kan erhållas från registret. Det är dock inte troligt att användarna upptäcker detta, eftersom de vanligtvis inte har tillgång till vikterna och kanske inte heller finner det meningsfullt att det är överensstämmelse i det fallet. Vi kommer i avsnitt 4 att visa att om man i exempel 2 gör en kalibrering av vikterna, så att vi får konsistens i totalen för registerinkomst har vi i själva verket konstruerat en regressionsestimator, som har mindre medelfel än en vanlig Horvitz-Thompson-estimator. Ökning av precisionen är ytterligatre ett skäl till att göra en anpassning till kända storheter.

Vi kommer också att visa att det går att framställa vikter, som gör att kraven i exempel 1 och 2 uppfylls samtidigt.

Deville & Särndal [1992] visar att problemet kan beskrivas och hanteras inom området "utnyttjande av hjälpinformation". Vi gör på samma sätt i föreliggande rapport.

Det är inte säkert att de "kända" marginalerna baserar sig på registerinformation - även de kan vara urvalsbaserade skattningar. I Deville & Särndal [1992] antas att de är icke-stokastiska medan Lütjohann[1985] tillåter att marginalskattningarna är behäftade med urvalsfel. I den senare metoden samutnyttjas cellskattningarna och marginalskattningarna på bästa sätt, vilket även leder till att marginalskattningarna förändras vid anpassningen.

3. Statistiska begrepp.

Antag att vi har en ändlig population U = (1,...,k,...,N) från vilken vi drar ett urval s. Urvalsdesignen ger inklusionssannolikheterna $\pi_k = P(k \in s)$ och $\pi_{kl} = P(k\&l \in s)$. Vi vill skatta totalen för variabel y i populationen eller i en redovisningsgrupp U_q (q=1,...,Q), d v s

$$t_y = \sum_U y_k \quad \text{resp } t_{yq} = \sum_{U_q} y_k \text{ , där } U_q \subseteq U. \qquad (1)$$

Till objekt k är knutet en hjälpinformationsvektor $\mathbf{x}_k = (x_{k1}, ..., x_{kj},..., x_{kJ})'$. Populationstotalen $\mathbf{t}_x$ antas vara känd och i vissa fall också $\mathbf{t}_{xq}$. I exempel 2 (ovan) är $\mathbf{x}_k = (z_{1k}x_k , z_{2k}\, x_k)'$, där x_k = registerinkomsten för individ k och z_{1k} = 1 om individ k är man och = 0 om det är en kvinna (z_{2k} pekar på motsvarande sätt ut kvinnor). För att visa att det går att ha samma skrivningssätt för fallet beskrivet i exempel 1 inför vi följande beteckningar:

---

1) Oftast är man intresserad av medelinkomsten, men en effektiv skattning av täljaren i en kvot innebär vanligtvis en effektiv skattning av kvoten.

$x_k = (\delta_{11k},..., \delta_{hik},..., \delta_{rck})'$, där $\delta_{hik} = 1$ om objekt k ligger i rad (kön) h och kolumn (ålder) i; $= 0$ f.ö. (2)

Det innebär att $\mathbf{t}_x = (N_{11}, ..., N_{hi}, ..., N_{rc})$, där N_{hi} är antalet personer i cell hi enligt befolkningsregistret.

4. Kalibrering

Beteckningen "kalibrering" (eng calibration) är inte entydig inom statistisk teori. Ursprungligen (?) användes ordet inom naturvetenskapen när man justerade ett mätinstrument utifrån ett eller flera kända sanna värden. Sedan har statistisk teori kommit att användas inom området (se t ex Martens & Naes [1989]). För ett mindre antal mätningar försöker man finna (statistiska) samband mellan de sanna värdena och värdena instrumentet ger och ev andra värden som beskriver en mätmiljö. Dessa samband utnyttjas för att senare prediktera sanna värden utifrån uppmätta värden. Gemensamt för metoderna är att de baserar sig på enskilda objekt.

Martens & Naes [1989] säger: "To CALIBRATE is to use empirical data and prior knowledge for determining how to predict unknown quantitative information Y from available measurements X, via some mathematical transfer function."

I Deville & Särndal [1992] förändrar man visserligen enskilda värden (vikter), men det görs med syfte att erhålla vissa kända storheter, som består av summor av vikterna eller summor av produkter mellan vikter och kända variabelvärden. Några "sanna" vikter finns inte. Denna senare metod presenteras nedan.

4.1 Teori

En Horvitz-Thompson-estimator (HT) av t_y har följande utseende:

$$\hat{t}_{yHT} = \sum_s y_k / \pi_k = \sum_s d_k y_k , \quad (3)$$

där vi kallar $d_k = 1/ \pi_k$ för designvikten för enhet k.

Metoden består i att bestämma nya vikter w_k som avviker "så lite som möjligt" från vikterna d_k, men som ändå ger:

$$\sum_s w_k x_k = \mathbf{t}_x \quad (4)$$

Med "så lite som möjligt" menas att man söker vikter som minimerar en avståndsfunktion under bivillkoret (4). Deville & Särndal [1992] studerar de avståndsfunktioner, som anges i tabell 1. De anger också vissa varianter på dessa huvudfunktioner bestämda av begränsningar av utfallsrummet för vikterna.

Tabell 1
Exempel på avståndsfunktioner.

1	Generaliserad minsta kvadrat	$(w_k - d_k)^2 / 2\, d_k v_k$
2	Raking ratio	$(w_k \log (w_k / d_k) - w_k + d_k) / v_k$
3	Hellinger	$2\left(\sqrt{w_k} - \sqrt{d_k}\right)^2 / v_k$
4	Minimum entropy	$(-\, d_k \log (w_k / d_k) + w_k - d_k) / v_k$
5	Modifierad Chi–två	$(w_k - d_k)^2 / 2\, w_k v_k$

Anm. I tabellen är v_k en känd positiv vikt, som ofta sätts lika med 1, men ibland sätts till något annat (se nedan om kvotestimator).

Väl man genomfört kalibreringen kan punktskattningen beräknas med estimatorn:

$$\hat{t}_{yw} = \sum_s w_k\, y_k \qquad (5)$$

Detta motsvarar i formen en HT-estimator, men bytet av d_k till w_k ger skattningar för y-variabeln, som utnyttjar hjälpinformation och även summerar sig till kända marginaler för x-variablerna.

Deville & Särndal [1992] behandlar också frågan om medelfel för kalibrerings-estimatorerna. De rekommenderar att variansen skattas med hjälp av följande formel:

$$\hat{V}(\hat{t}_{yw}) = \sum_s \sum \left(1 - \frac{\pi_k \pi_l}{\pi_{kl}}\right) (w_k e_k)\, (w_l e_l) \qquad (6)$$

där

$e_k = y_k - \mathbf{x}_k' \hat{\mathbf{B}}_{ws}$ och där $\hat{\mathbf{B}}_{ws}$ utgör lösningen till normalekvationen

$$\left(\sum_s w_k v_k \mathbf{x}_k \mathbf{x}_k'\right) \hat{\mathbf{B}}_{ws} = \sum_s w_k\, v_k \mathbf{x}_k y_k \qquad (7)$$

Den enda skillnaden mot formeln för medelfelet i en generell regressionsestimator (se Särndal, Swensson & Wretman[1992]) är att designvikten d_k är ersatt med w_k.

Om vi drar ett OSU av storleken n från N i exempel 2 skulle vi få följande formel för variansskattningen vid estimation av redovisningsgruppstotalen t_{yq}, om vi sätter $v_k = 1/x_k$ och antar t_{xq} känd:

$$\hat{V}(\hat{t}_{yqw}) \approx (t_{xq} / \hat{t}_{xqHT})^2\, N_q^2\, \frac{1 - n/N}{n_q} \left[\sum_{s_q} (y_k - \hat{R}_q x_k)^2\right] / (n_q - 1), \text{ där} \qquad (8)$$

$$\hat{R}_q = \left(\sum_{s_q} y_k\right) / \left(\sum_{s_q} x_k\right)$$

Det är alltså den ofta använda formeln (motsvarar metoden med s k g-viktade residualer) för skattning av medelfelet i en kvotestimator.

Den teori som är utvecklad i Särndal, Swensson & Wretman [1992] är i hög grad relevant att använda när man kalibrerar vikter. Avståndsfunktion 1 leder till den generella regressionsestimatorn, som i sin tur innehåller en mängd kända specialfall, som den poststratifierade estimatorn, kvotestimatorn (separat och kombinerad), den klassiska regressionsestimatorn, etc. Dessutom ger övriga avståndsfunktioner estimatorer som asymptotiskt (med ökande stickprovsstorlek) går mot den generella regressionsestimatorn. Det är därför rimligt att ha en modellstödd ansats även när man kalibrerar vikter.

4.2 Beräkning av kalibrerade vikter.

Som har påpekats tidigare minimerar kalibrerade vikter en avståndsfunktion under bivillkoret (4). Första steget i lösningen av det minimeringsproblemet görs med Lagrange-metoden och för de flesta avståndsfunktionerna kan inte en explicit lösning erhållas (undantag utgör avståndsfunktion 1) utan en iterativ metod får tillgripas.

Vi gör ingen ytterligare beskrivning av teorin bakom beräkningarna utan hänvisar till Deville & Särndal [1992].

Den franska statistiska centralbyrån I.N.S.E.E. har sedan år 1990 kalibrerat vikterna i sina hushållsundersökningar och för den skull har man framställt SAS-macro-programmet CALMAR. I.N.S.E.E. (Oliver Sautory) har välvilligt ställt programmet till förfogande och det finns nu på diskett vid U/STM-Ö.

Programmet innehåller såväl en stordatorversion som en PC-version. Tyvärr finns det inte någon egentlig manual till programmet, det enda skrivna finns i Sautory [1992], som är en artikel av mer statistisk/teoretisk art, och därför vet vi inte till alla delar hur programmet fungerar. I bilaga 1 visas exempel på en körning samt kommentarer till den.

CALMAR klarar av att kalibrera vikter när man har flera kända marginaler och samtidigt kända totaler för kontinuerliga hjälpvariabler.

Outputen består i kalibrerade vikter och det är även möjligt att få ut vissa analyserande mått över viktsystemet.

Tyvärr är PC-versionen (Windows-applikationen)[1] mycket långsam - vid 200 observationer och med 2 kända marginaler och den enklaste avståndsfunktionen (nr 1) tar den 17 min. (Vi har även prövat att köra programmet utanför windows och då tar det 5 min.). Därför är den olämplig att använda när man vill göra många beräkningar, t ex studera effekten av olika avståndsfunktioner, hur estimatorerna fungerar för olika uppsättningar av marginaler, etc. Vi har istället framställt ett APL-program som utför nämnda beräkning på 5 sek.

Med hjälp av APL-programmet ska vi för ett konkret exempel illustrera kalibreringstekniken.

---

1) Vi har kört på IBM/PC, modell 70, 386 med matematikprocessor.
Efter att de flesta beräkningarna i denna rapport har genomförts har SAS/Win installerats. Den gör beräkningen på ca 3 min.

Vi har konstruerat en population bestående av 4000 personer och som ansluter till exempel 1 och 2. Vi önskar bl a skatta antalet personer i varje cell i korstabellen kön*ålder*inkomst, där ålder och inkomst är kategoriserade variabler (jämför med exempel 1). Dessutom vill vi skatta summan av inkomsten (y) för respektive kön (exempel 2). Ett register innehåller registerinkomst (x), som samvarierar med undersökningsvariabeln. Vi har gjort sambanden mellan y och x så att en klassisk regressionsestimator är bäst för män (kön 1) och en klassisk kvotestimator för kvinnor. Interceptet är 1000, lutningskoefficienten är 1 och modellvariansen är lika för alla män och för kvinnor har vi inget intercept, lutningskoefficienten är 1.5 och modellvariansen för y_k är proportionell mot x_k . Korrelationskoefficienten mellan y och x är 0.82 för män och 0.67 för kvinnor.

Vi gör denna konstruktion för att pröva huruvida det går att göra flexibla och optimala lösningar vid kalibrering och dessutom vill vi med de fortsatta beräkningarna konkretisera metoden. Arbetet tjänar mindre till att visa att hjälpinformation ger precisare skattningar, eftersom vi redan vet detta.

Populationen beskriven i korstabellen kön*ålder*inkomst har följande utseende:

Tabell 2.
Antalet personer i populationen fördelade på kön, ålder och inkomst.

		Inkomstklass			
	Ålder	1	2	3	Summa
Män	1	2	113	285	400
	2	3	183	414	600
	3	5	196	499	700
	4	1	108	191	300
Summa män		11	600	1389	2000
Kvinnor	1	233	101	16	350
	2	420	174	56	650
	3	531	203	66	800
	4	128	58	14	200
Summa kvinnor		1312	536	152	2000

Förutom cellvärden och marginaler i tabell 2 vill vi också skatta totalinkomsten för män respektive kvinnor:

$$t_{y1} = \sum_{U_1} y_k = 2691622 \text{ och } t_{y2} = \sum_{U_2} y_k = 1363630 \qquad (9)$$

Vi antar att befolkningsregistret ger marginalen kön*ålder och dessutom ger inkomstregistret:

$$t_{x1} = \sum_{U_1} x_k = 694024 \text{ och } t_{x2} = \sum_{U_2} x_k = 897128 \qquad (10)$$

Vi drar ett obundet slumpmässigt urval av storleken 200 från populationen. Designvikten (d_k) är därmed (4000/200 =) 20 för alla utvalda och skattningarna vi erhåller med HT-estimatorn återfinns i tabell 3.

Tabell 3.
HT-estimatorn.

	Ålder	Inkomstklass 1	2	3	Summa
Män	1	0	60	360	420
	2	0	160	480	640
	3	20	160	560	740
	4	0	180	100	280
Summa män		20	560	1500	2080
Kvinnor	1	340	100	20	460
	2	340	220	60	620
	3	460	160	20	640
	4	100	60	40	200
Summa kvinnor		1240	540	140	1920

Som väntat överensstämmer inte marginalskattningarna med marginalen kön*ålder i tabell 2.

Dessutom ger HT-estimatorn följande skattningar av totalinkomsten för respektive kön:

$$\hat{t}_{yHT1} = \sum_{s_1} d_k\, y_k = 2792300 \text{ och } \hat{t}_{yHT2} = \sum_{s_2} d_k\, y_k = 1326940 \qquad (11)$$

s_1 (s_2) är den delmängd av s som är män (kvinnor).

Om vi gör motsvarande skattningar av storheterna (10) får vi

$$\hat{t}_{x1HT} = \sum_{s_1} d_k\, x_k = 692220 \text{ och } \hat{t}_{x2HT} = \sum_{s_2} d_k\, x_k = 847080 \qquad (12)$$

Dessa skattningar överensstämmer inte heller med de kända värdena.

Kalibrering 1.

Låt oss börja med att beräkna kalibrerade vikter w_k som ger skattningar som summerar sig till marginalerna i tabell 2. Vi använder avståndsfunktion 1 som enligt teorin ger en poststratifierad estimator med poststrata enligt marginalinformationen.

Vi antar att kön och åldersklass (4 klasser) bildar två olika variabler på datafilen. Först måste vi lägga samman dessa variabler till en variabel med (2x4=) 8 kategorier. I APL-varianten har vi gjort ett program för detta - i CALMAR finns ingen färdig lösning.

Grundinformationen för programmen är matrisen X, där kolumn k utgörs av x_k (k= 1,..., n). I föreliggande fall har vi 200 (= antal observationer i stickprovet) kolumner och 8 rader (= antal kategorier). Kolumn k innehåller en etta och resten nollor, där ettan pekar ut vilken kategori person k tillhör. Dessutom ska också marginalinformationen samt designvikterna anges som input till programmen . (I bilaga 1 visar vi hur detta görs i CALMAR).

I APL-programmet finns det också möjlighet att sätta in godtyckliga värden på v-vikterna - i CALMAR är den vikten genomgående lika med 1. När vi endast kalibrerar mot kända marginaler i en korstabell verkar det rimligt att sätta $v_k = 1$ för alla k. Det gör vi i detta fall.

Kalibreringen leder till följande skattningar:

Tabell 4.
Skattningar baserade på kalibrerade vikter till marginalerna kön*ålder i tabell 2 och med användande av avståndsfunktion 1.

		Inkomstklass			
	Ålder	1	2	3	Summa
Män	1	0	57	343	400
	2	0	150	450	600
	3	19	151	530	700
	4	0	193	107	300
Summa män		19	551	1430	2000
Kvinnor	1	259	76	15	350
	2	356	231	63	650
	3	575	200	25	800
	4	100	60	40	200
Summa kvinnor		1290	567	143	2000

Skattningarna summerar sig som synes till de kända marginalerna.

Skattningarna för totalinkomsterna blir:

$$\hat{t}_{y1w} = \sum_{s_1} w_k y_k = 2678715 \text{ och } \hat{t}_{y2w} = \sum_{s_2} w_k y_k = 1388090 \qquad (13)$$

Motsvarande skattningar för x-variabeln blir:

$$\hat{t}_{x1w} = \sum_{s_1} w_k x_k = 663329 \text{ och } \hat{t}_{x2w} = \sum_{s_2} w_k x_k = 882428 \qquad (14)$$

Dessa skattningar överensstämmer inte med de kända storheterna i (10), därför att vikterna inte kalibrerats till att reproducera dem.

Motsvarande skattningar skulle vi kunna erhålla genom att på "konventionellt" sätt bilda en poststratifierad estimator med poststrata bestämda av kön*ålder.

I nästa kalibrering gör vi dessutom en anpassning mot totalerna (10), vilket leder till skattningar, som inte kan erhållas från en konventionell estimator. Om vi bara anpassade mot totalerna (10) skulle vi få skattningar, som också kunde erhållas från en klassisk regressionsestimator med kända totaler (10).

Kalibrering 2.

Vi vill alltså bestämma vikter som ger skattningar som summerar sig till kända antal i kön*ålder och där skattningarna av totalerna (10) är lika med dessa totaler.

För att kunna beskriva de nya restriktionerna på vikterna måste två rader läggas till matrisen X (bestämd i kalibrering 1). Kolumn k i den första raden i tillägget består av värdet x_k om person k är man och 0 om det är en kvinna. I den andra raden är det tvärtom - männen åsätts nollor och kvinnor x-värdena.

Man skulle också kunna tro att man måste ha en rad som anger intercept eller ej i estimatorn, men en sådan rad gör matrisen singulär. Utelämnande innebär i princip att ett intercept tillåts, men kan skattas till noll.

Vi antar att vi faktiskt har insett att variansstrukturen är i stort sett konstant för män och proportionell mot x för kvinnor. Dessa vikter är relevanta när vi skattar inkomsttotalerna, men de påverkar också skattningen av cellvärden i tabellen kön*ålder*inkomst och är där inte relevanta. Därför väljer vi här $v_k = 1$ för alla k.

Följande resultat erhålles:

Tabell 5.
Skattningar baserade på kalibrerade vikter till marginalerna enligt tabell 2 och till inkomsttotalerna (10). Avståndsfunktion 1 används.

		Inkomstklass			
	Ålder	1	2	3	Summa
Män	1	0	51	349	400
	2	0	142	458	600
	3	16	140	543	699
	4	0	186	114	300
	Summa män	16	519	1464	1999
Kvinnor	1	253	81	16	350
	2	345	237	68	650
	3	568	207	26	801
	4	96	61	43	200
	Summa kvinnor	1262	586	153	2001

Vi ser att skattningarna summerar sig till kända marginaler (när man avrundar cellskattningarna till heltal och summerar dessa kan man få vissa avvikelser i marginalerna).

Skattningarna av inkomsttotalerna för respektive kön är:

$$\hat{t}_{y1w} = \sum_{s_1} w_k y_k = 2705800 \text{ och } \hat{t}_{y2w} = \sum_{s_2} w_k y_k = 1408576 \qquad (15)$$

Om vi sätter in registeruppgifterna x_k i estimatorn erhålles värdena 694024 resp 897128, d v s exakt lika med de kända värdena. Detta är som sig bör och bekräftar att programmet har utfört viktberäkningen på korrekt sätt.

Kalibrering 3.

I kalibrering 2 har vi beräknat skattningar som inte kan erhållas med en konventionell estimator. Vi ska även i denna kalibrering visa ett sådant fall.

Antag att vi vet antalet män och kvinnor i populationen och vi vet också hur hela populationen fördelar sig på åldersklasser. Däremot vet vi inte hur männen och kvinnorna fördelar sig på åldersklasser.

Tabell 6.
Skattningar baserade på vikter kalibrerade till kända köns- resp åldersmarginaler. Avståndsfunktion 1 används.

		Inkomstklass			
	Ålder	1	2	3	Summa
Män	1	0	48	290	338
	2	0	152	455	607
	3	21	167	585	773
	4	0	181	100	281
Summa män		21	548	1430	1999
Kvinnor	1	304	89	18	411
	2	353	228	62	643
	3	522	182	23	727
	4	109	66	44	219
Summa kvinnor		1288	565	147	2000

I denna tabell upptäcker man att summeringen inom varje kön summerar sig till de kända storheterna. En elementär beräkning visar att de även summerar sig till känd åldersfördelning (750,1250, 1500, 500).

Kalibrering 4.

Vi ska titta på ett knepigt fall som kan inträffa och som också visar att konventionell skattningsteknik inte är möjlig att använda eller åtminstone att resultaten kan ifrågasättas.

Antag att vi har ytterligare en variabel "region". Antag vidare att vi vill redovisa resultaten fördelade på två tabeller, där första tabellen är region*kön*inkomst och den andra region*ålder*inkomst (se tabell 7 och 8). Tabellerna är uppbyggda så att båda innehåller delsummorna region*inkomst. Vi vill då inte ha olika skattningar av dessa delsummor.

Vid en konventionell ansats skulle vi behöva ha region*kön*ålder som poststrata. Om region består av många kategorier skulle urvalet ge ostadiga skattningar i vissa celler. I

kalibreringsansatsen kan vi däremot kalibrera enbart mot marginalerna region*kön resp region*ålder. Vi ska här visa att vi får samma skattningar av delsummorna.

I den population som använts tidigare har vi lagt till en variabel som pekar ut två regioner och samtidigt slagit samman ålderskategori 1 och 3 samt 2 och 4. På populationsnivå har vi följande två tabeller:

Tabell 7
Populationen fördelad på region, kön och inkomstklasser.

		Inkomstklass			
	Kön	1	2	3	Summa
Region 1	1	11	597	541	1149
	2	737	44	0	781
	Summa	748	641	541	1930
Region 2	1	0	3	848	851
	2	575	492	152	1219
	Summa	575	495	1000	2070

Tabell 8
Populationen fördelad på region, ålder och inkomstklasser.

		Inkomstklass			
	Ålder	1	2	3	Summa
Region 1	1	440	335	315	1090
	2	308	306	226	840
	Summa	748	641	541	1930
Region 2	1	331	278	551	1160
	2	244	217	449	910
	Summa	575	495	1000	2070

Vi drar, som i föregående fallen, ett OSU med n=200 och genomför en kalibrering med avståndsfunktion 1 och med antagande att vi känner marginalerna för region*kön samt region*ålder. Dessutom sätter vi v_k=1 för alla k. Resultatet blir följande:

Tabell 9
Skattningar baserade på kalibrerade vikter till marginalerna region*kön.
Avståndsfunktion 1 används.

		Inkomstklass			
	Kön	1	2	3	Summa
Region 1	1	19	561	568	1148
	2	697	84	0	781
	Summa	716	645	568	1929
Region 2	1	0	0	851	851
	2	452	647	111	218
	Summa	452	647	970	2069

Tabell 10
Skattningar baserade på kalibrerade vikter till marginalerna region*ålder.
Avståndsfunktion 1 används.

		Inkomstklass			
	Ålder	1	2	3	Summa
Region 1	1	380	344	366	1090
	2	337	301	202	840
	Summa	717	645	568	1930
Region 2	1	250	384	526	1160
	2	203	263	444	910
	Summa	453	647	970	2070

Vi ser att skattningarna summerar sig till kända marginaler och dessutom överensstämmer skattningarna av cellvärdena i tabellen region*inkomstklass mellan tabell 9 och 10.

Medelfel.

Låt oss titta på medelfelen för ovanstående estimatorer. I tabell 11 anges skattningar av medelfelen av varje cellskattning baserad på samma urval som tidigare redovisade resultat. Medelfelsskattningarna utförs med ett specialgjort APL-program som följer formel (6).

Tabell 11.
Medelfel för ..... HT-estimatorn (första siffran)
estimatorn i kalibr. 1 (andra siffran)
" " 2 (tredje siffran).
" " 3 (fjärde siffran)

		Inkomstklass 1				2				3			
Ålder													
Män	1	0	0	0	0	34	30	27	26	79	30	27	50
	2	0	0	0	0	54	45	42	48	90	45	42	66
	3	19	18	16	20	54	46	42	53	96	48	43	74
	4	0	0	0	0	57	38	37	46	43	38	37	39
Kvinnor	1	77	31	32	52	43	29	30	36	19	15	16	17
	2	77	57	54	66	63	55	54	58	34	34	35	34
	3	88	62	61	75	54	60	59	57	49	24	25	22
	4	43	31	30	41	34	28	29	34	27	25	26	29

Vi ser att kalibreringen som motsvarar poststratifiering (andra siffran) ger en kraftig reducering av medelfelen jämfört med HT-estimatorn (första siffran). När vi lägger ytterligare restriktioner på vikterna (tredje siffran) får vi ytterligare någon reducering av medelfelen. När vi endast känner fördelningen på kön resp ålder (fjärde siffran) får vi sämre skattningar än då vi känner fördelningen på kön*ålder, men oftast bättre skattningar än där vi inte utnytjar någon hjälpinformation alls (första siffran).

Medelfelen för skattningar av inkomsttotalerna är:

Tabell 12.
Skattade medelfel för estimatorerna av inkomsttotaler.

	HT-estim.	Kalibrering 1	2	3
Män	190932	41112	25760	44329
Kvinnor	112969	62084	51179	63192

Vi ser en kraftig reducering av medelfelen från HT-estimatorn till kalibrering 1. Poststratifieringen har alltså en god effekt även vid skattning av dessa parametrar. En ytterligare reducering görs i kalibrering 2. Det är en kraftigare samvariation mellan y och x för män än för kvinnor och därför får vi en kraftigare reduktion av medelfelen när ytterligare hjälpinformation utnyttjas. Kalibrering 3 ger något större medelfel än kalibrering 1.

Val av avståndsfunktion.

Deville & Särndal [1992] ger vissa kommentarer till valet av avståndsfunktion. De säger t ex att avståndsfunktionen "...has only a modest impact on such essential properties as the variance of the estimator." De säger också att: "Computational convenience more than anything else may then dictate the choise...". Valet av funktion kan också påverka skattningar i redovisningsgrupper negativt: "One may want to avoid a function ... that give overly extreme weights, because applying these weights to make estimates for various subpopulations (domains) may produce unrealistic estimates for some domains..."

De ger också några kommentarer till de enskilda avståndsfunktionerna:

a) Man kan alltid finna lösningar för funktionerna 1 och 2, men för de andra kan det ibland vara svårt.

b) Funktion 1 kan ge negativa vikter, men dock inte de andra.

c) Funktion 2 kan ge mycket stora vikter.

Nu presenterar de även varianter på dessa funktioner, som innebär att man begränsar utfallsrummet.

Vi har även gjort kalibreringar med de restriktioner som finns för kalibrering 2 med utnyttjande av övriga avståndsfunktioner och endast funnit små skillnader mellan skattningarna - den största skillnaden i någon cell är 6 personer.

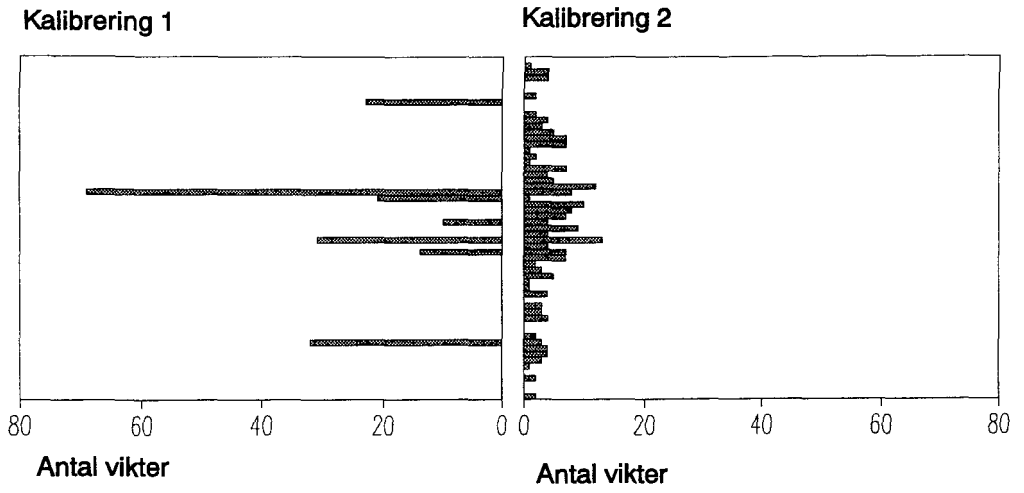
Stukel & Boyer [1992] analyserar vikterna för olika avståndsfunktioner på data från Statistics Canada's Labor Force Survey och finner att funktion 1 "...is the winner". En god egenskap (utan att motivera varför) är enligt författarna att den uppvisar vikter som är nästan normalfördelade.

Även Sautory [1992] gör en omfattande analys av viktsystemen och CALMAR ger också möjlighet att beräkna ett antal analyserande mått.

Vi har också något studerat fördelningen av olika viktsystem. För de restriktioner som gäller för kalibrering 2 ser vi endast små skillnader mellan viktsystemen för olika avståndsfunktioner och standardavvikelserna för varje system är nästan identiska.

Låt oss även titta på hur vikterna förändras när mer hjälpinformation inkluderas. I nedanstående diagram visas fördelningen av vikterna för kalibrering 1 och 2.

Diagram 1.
Fördelning av vikterna för kalibrering 1 och 2.



Den första kalibreringen ger en separat vikt för varje cell, medan den andra ger olika vikter för alla objekt.

Den slutsats vi drar kring val av avståndsfunktion är:

Den första funktionen kräver minsta beräkningsarbetet och leder också till en generell regressionsestimator, vilken är den estimator som övriga asymptotiskt går mot med ökande stickprov. Om man finner att negativa värden är besvärande kan det kan vara rimligt att begränsa utfallsrummet till enbart positiva värden (rekommenderas av Stukel & Boyer [1992]).

5. Anpassning av estimat till kända storheter.

I kalibreringen bestäms vilka marginaler och andra storheter som anpassningen ska göras till före estimationsmomentet. De metoder som presenteras i detta avsnitt används efter att den konventionella estimationen är gjord. Vi antar att man har använt en viss hjälpinformation i såväl urvalsdesignen som i estimationen, men att man har "råkat" få störande avvikelser mellan skattningar och kända storheter eller skillnader mellan skattningar av samma parameter på olika ställen i den statistiska rapporten. Det är inte säkert att de kända storheterna är i någon mening "sanna", utan de kan vara behäftade med t ex urvalsfel.

Vi kommer att presentera två metoder, sammanjämkning av tabeller med utnyttjande av minsta-kvadrat-metoden samt raking-ratio-metoden.

Båda metoderna behandlar fallet med korstabeller och därför startar vi med att definiera problemet i statistiska termer.

5.1 Statistiska begrepp.

Problematiken kan finnas i en mångdimensionell tabell, men ofta kan de återföras till ett 2-dimensionellt fall då man kan bilda en ny variabel av två eller flera grundvariabler. Det är även möjligt att beskriva metoderna med många dimensioner, men för att förenkla beskrivningen begränsar vi oss till två dimensioner.

Antag alltså att vi har en korstabell med r rader och c kolumner. Vi antar också att det är antalsparametrar som skattas.

Parametervärdena vi vill skatta är:

N_{hi} där $h = 1,..., r$ och $i = 1,..., c$

Dessa storheter kan skrivas i form av en vektor:

$$\mathbf{N} = \{ N_{11}, N_{12},..., N_{1c}, N_{21},...., N_{hi}, ...N_{rc}\}' \qquad (16)$$

Antag att vi har skattat dessa parametrar med $\hat{N}_{hi}$.

Skrivet på analogt sätt i vektorform erhåller vi $\hat{\mathbf{N}}$.

Antag vidare att vi känner

$$\text{radsummorna } N_{h.} = \sum_{i}^{c} N_{hi} \text{ och / eller kolumnsummorna } N_{.i} = \sum_{h}^{r} N_{hi} \qquad (17)$$

Uttryckt i vektorform har vi

$$\mathbf{N}_r = \{ N_{1.}, ..., N_{h.}, ..., N_{r.}\}' \text{ och } \mathbf{N}_c = \{ N_{.1}, ..., N_{.i}, ..., N_{.c}\}'$$

Problemet är att om vi summerar cellskattningarna överensstämmer inte summan med de kända storheterna.

Nu kan det vara så att även marginalerna är skattade (väntevärdesriktigt!) i någon annan undersökning och behäftade med urvalsfel. Valet av dessa till rättesnören torde i regel bero på att de baserar sig på större urval än i föreliggande undersökning.

5.2 Sammanjämkning av tabeller och marginaler med användande av minsta-kvadrat-metoden.

5.2.1 Teori

Denna metod har sina rötter i Deming & Stephan [1940]. Det är dock framför allt från Lütjohann [1985] vi hämtar informationen. Här gör vi endast en översiktlig beskrivning av metoden och hänvisar till nämnda referenser för mer detaljer.

Metoden baserar sig på följande antaganden:

a) N_{hi} skattas väntevärdesriktigt med $\hat{N}_{hi}$.

b) Cellskattningarna är ömsesidigt okorrelerade.

Om dessa antaganden gäller kan teorin kring vägd regressionsansats med linjära restriktioner användas (se t ex Johnston [1972]).

Låt **z** vara en (stående) vektor som anger de kända marginalerna. Om vi enbart känner radsummorna är $\mathbf{z} = \mathbf{N}_r$ och om vi bara känner kolumnsummorna är $\mathbf{z} = \mathbf{N}_c$, och om vi känner både radsummor och kolumnsummor är **z** en (r + c, 1) vektor där vi först återfinner $\mathbf{N}_r$ och sedan $\mathbf{N}_c$.

Låt vidare $\mathbf{N}^{sam}$ vara de reconcilierade skattningarna.

Vi vill ju att skattningarna ska uppfylla följande villkor:

$\mathbf{Z}\,\mathbf{N}^{sam} = \mathbf{z}$, där **Z** är en design–matris som pekar ut vilka element i $\mathbf{N}^{sam}$ som ska summeras.

A. Icke-stokastiska marginaler.

Om vi antar att **z** är icke-stokastisk då ger följande skattningar en minsta-kvadrat-lösning under ovanstående bivillkor:

$\mathbf{N}^{sam} = \hat{\mathbf{N}} - \mathbf{W}\,\mathbf{Z}'\,[\mathbf{Z}\,\mathbf{W}\,\mathbf{Z}']^{-1}\,(\mathbf{Z}\,\hat{\mathbf{N}} - \mathbf{z})$, där **W** är en diagonalmatris (18) med $V(\hat{\mathbf{N}})$ "uträtad" i diagonalen.

$V(\hat{\mathbf{N}})$ är designvariansen, d v s den innehåller inte någon varianskomponent som beror av modellen. Om vi kände $V(\hat{\mathbf{N}})$ och därmed den rätta **W** skulle den sammanjämkade estimatorn ge skattningar med minsta varians bland gruppen linjära estimatorer. Detta gäller även för små stickprov. Nu måste vi skatta $V(\hat{\mathbf{N}})$ och därmed måste vi föra in ett "stor-sampel"-resonemang.

Deming & Stephan [1940] antog att $V(\hat{N}_{hi})$ är proportionell mot $\hat{N}_{hi}$. Lütjohann [1985] säger bl a "If the weights are not known exactly, they should be estimated or even guessed..."

Theil [1971] visar att variansen för $\mathbf{N}^{sam}$ kan beräknas med följande formel:

$$V(\mathbf{N}^{sam}) = \mathbf{W} - \mathbf{W}\,\mathbf{Z}'\,[\mathbf{Z}\,\mathbf{W}\,\mathbf{Z}']^{-1}\,\mathbf{Z}\,\mathbf{W} \qquad (19)$$

B. Stokastiska marginaler.

Om däremot även marginalinformationen är stokastisk (består av urvalsbaserade skattningar) blir skattningsförfarandet annorlunda. Antag att vi har bra skattningar av både rad- och kolumnsummorna $\hat{\mathbf{N}}_r$ resp $\hat{\mathbf{N}}_c$ och variansen för dessa $V(\hat{\mathbf{N}}_r)=\mathbf{W}_r$ resp resp $V(\hat{\mathbf{N}}_c)=\mathbf{W}_c$. Här skattas cellvärdena med

$$\mathbf{N}^{sam} = (\mathbf{X}'\mathbf{W}_t^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}_t^{-1}\mathbf{Y}, \text{ där} \qquad (20)$$

X är en design-matris, som kan skrivas:

$$\mathbf{X} = \begin{bmatrix} \mathbf{I} \\ \mathbf{Z}_r \\ \mathbf{Z}_c \end{bmatrix} \qquad (21)$$

där **I** är en diagonalmatris ((r x c), (r x c)) med ettor i marginalen, $\mathbf{Z}_r$ ($\mathbf{Z}_c$) är **Z** - varianten när rad (kolumn) - summorna är kända.

$$\mathbf{Y} = \begin{bmatrix} \hat{\mathbf{N}} \\ \hat{\mathbf{N}}_r \\ \hat{\mathbf{N}}_c \end{bmatrix} \qquad (22)$$

Dessutom är $\mathbf{W}_t$ en diagonalmatris där $V(\hat{\mathbf{N}})$ (eller motsvarande vikter) , $\mathbf{W}_r$ och $\mathbf{W}_c$ ligger uträtade och efter varandra i diagonalen.

Detta ger skattningar av cellparametrarna och de nya marginalskattningarna ges av summeringar av cellskattningarna.

Theil [1971] visar att variansen för $\mathbf{N}^{sam}$ när vi har marginaluppgifter, som är behäftade med urvalsfel, kan skrivas som:

$$V(\mathbf{N}^{sam}) = (\mathbf{X}'\mathbf{W}^{-1}\mathbf{X})^{-1} \qquad (23)$$

5.2.2 Beräkning av sammanjämkade tabeller.

Lütjohann har framställt två SAS - program för stordator - det första används när man har icke-stokastiska marginaler och det andra vid stokastiska marginaler. Programmen ger punktskattningar, men däremot inga medelfelsskattningar. Programmet har en del begränsningar och verkar (vi har inte provat!) vara något krävande vid inmatning av data.

Därför har vi gjort motsvarande två APL-program, som inte har de begränsningarna. Dessutom har vi också framställt program för medelfelsberäkningar.

Med dessa APL-program vill vi visa hur denna metod fungerar och även göra jämförelser med övriga metoder i föreliggande rapport.

Vi fortsätter med det urval vi använt tidigare och genomför en sammanjämkning enligt de förutsättningar som kalibrering 1 har genomförts. Det innebär bl a att marginalerna är icke-stokastiska. I den första beräkningen låter vi **W** bestämmas av de skattade (design)-varianserna för HT-estimatorn.

Tabell 13.
Skattningar sammanjämkade till den kända marginalen kön*ålder. Viktmatrisen **W** är bestämd av de skattade varianserna i HT-estimatorn.

		Inkomstklass			
	Ålder	1	2	3	Summa
Män	1	0	57	343	400
	2	0	149	451	600
	3	19	151	530	700
	4	0	193	107	300
Summa män		19	550	1431	2000
Kvinnor	1	260	75	15	350
	2	356	231	63	650
	3	572	202	25	799
	4	100	60	40	200
Summa kvinnor		1288	568	143	1999

Om vi jämför med kalibreringen som utnyttjar avståndsfunktion 1 (se tabell 4) ser vi att skillnaden är minimal.

Låt oss jämföra de skattade medelfelen för den reconcilierade estimatorn med motsvarande skattade medelfel för estimatorn i kalibrering 1. Dessutom anger vi också skattade medelfel för HT-estimatorn. (Tabell 7 innehåller delvis samma information, men för jämförbarhetens skull gör vi denna dubblering). De skattade medelfelen med användande av fomel (19), med insatta skattningar för **W**, är följande:

Tabell 14.
Medelfel för ..... HT-estimatorn (första siffran)
estimatorn i kalibr. 1 (andra siffran)
sammanjämkade estim. (tredje siffran).

	Ålder	Inkomstklass 1	2	3
Män	1	0 0 1	34 30 31	79 30 31
	2	0 0 1	54 45 46	90 45 46
	3	19 18 19	54 46 47	96 48 49
	4	0 0 1	57 38 34	43 38 34
Kvinnor	1	77 31 40	43 29 38	19 15 19
	2	77 57 52	63 55 50	34 34 32
	3	88 62 48	54 60 46	49 24 19
	4	43 31 31	34 28 28	27 25 24

Viktningen har stor betydelse, vilket framgår av nästa tabell. Där har vi bestämt W så att samma vikt ges åt varje cellskattning (ävem om det i praktiken förefaller osannolikt att varje cellskattning har samma varians).

Tabell 15
Skattningar sammanjämkade till den kända marginalen kön*ålder. Samma vikt för varje cellskattning.

		Ålder	Inkomstklass 1	2	3	Summa
Män		1	-7	53	353	399
		2	-13	147	467	601
		3	7	147	547	701
		4	7	187	107	301
	Summa män		-6	534	1474	2002
Kvinnor		1	303	63	-17	349
		2	350	230	70	650
		3	513	213	73	799
		4	100	60	40	200
	Summa kvinnor		1266	566	166	1998

Vi kan konstatera att viktningen har stor betydelse och den kan t o m leda till negativa skattningar.

Vi antar att marginalerna är skattade i någon annan undersökning och med varianser som endast utgör 10% av den varians HT-estimatorn (baserad på föreliggande urval) uppvisar. Marginalskattningarna antas vara desamma som de vi tidigare använt som fixa. Vi får följande resultat:

Tabell 16.
Skattningar sammanjämkade till den skattade marginalen kön*ålder. Viktningen görs med de skattade varianserna i HT-estimatorn för cellerna och motsvarande 10% av variansen för marginalskattningarna.

	Ålder	Inkomstklass 1	2	3	Summa
Män	1	0	57	345	402
	2	0	150	453	603
	3	19	151	533	703
	4	0	192	107	299
Summa män		19	550	1438	2007
Kvinnor	1	267	77	16	360
	2	355	230	63	648
	3	563	199	25	787
	4	100	60	40	200
Summa kvinnor		1285	566	144	1995

Vi ser att marginalskattningarna förändras något från tabell 13, liksom även cellskattningarna.

ANM: De program som antar stokastiska marginaler ger samma resultat, som de med icke-stokastiska marginaler, om vi i programmet, som antar stokastiska marginaler, sätter varianserna för marginalskattningarna till små tal (det går inte att sätta storheterna till 0).

5.3 Raking-ratio teknik.

5.3.1 Teori

Låt oss börja med att studera en 2-dimensionell tabell, där vi känner båda marginalerna. Våra skattningar av cellantalen är $\hat{N}_{hi}$ som dock inte summerar sig till de kända

marginalerna N_r och N_c. Därför vill vi framställa Raking-Ratio-skattningar (RR) som har den egenskapen. RR-skattningen faller ut som lösning till följande iterativa process (eng. Iterative Proportional Fitting [IPF]):

Startvärde i den iterativa processen är
$\hat{N}_{hi}^{(0)} = \hat{N}_{hi}$

Steg k i IPF-proceduren:

$$ {}_1\hat{N}_{hi}^{(k)} = \frac{\hat{N}_{hi}^{(k-1)}}{\hat{N}_{h.}^{(k-1)}} N_{h.} \qquad (24) $$

$$ \hat{N}_{hi}^{(k)} = \frac{{}_1\hat{N}_{hi}^{(k)}}{{}_1\hat{N}_{.i}^{(k)}} N_{.i} $$

Efter ett antal steg, säg k_0 , är förändringen av skattningen från steg k_0-1 så liten att vi avbryter iterationen. Då är

$$ N_{hi}^{RR} = \hat{N}_{hi}^{(k_0)} $$

Om vi enbart känner en marginal t ex $\mathbf{N_r}$ då behövs ingen iteration utan skattningen erhålles ur en explicit formel:

$$ N_{hi}^{RR} = \frac{\hat{N}_{hi}}{\hat{N}_{.h}} N_{h.} \text{ , d v s en kvotskattning} \qquad (25) $$

Det är lätt att utvidga IPF-proceduren till fler dimensioner. Vi skriver inte ner uttrycken här, men däremot har vi framställt APL-program som beräknar RR-skattningar för en 3-dimensionell tabell, där vi känner alla tre marginalerna eller bara två marginaler.

Om man känner antalen i korstabellen variabel1*variabel2 kan en variabel bildas av dessa två och därmed kan de flesta fallen återföras till en 3-dimensionell tabell.

Man kan fråga sig vilka statistiska egenskaper dessa RR-estimatorer har? Redan i avsnittet om kalibrering har vi sett att avståndsfunktion 2 leder till en RR-estimator och därmed vet vi en del om den. Historiskt sett har den utvecklats från Deming & Stephan [1940] via bl a Ireland & Kullback [1968], Fienberg [1978]. Den har gått från beräkningar i enklare tabeller till teorier om log-linjära modeller.

RR-estimatorn är optimal i den meningen att den maximerar likelihood-funktionen under de givna marginalvillkoren. Likelihoodfunktionen baserar sig ju på en sannolikhetsfördelning för cellvärdena och frågan är då hur den ser ut. För ett OSU är en rimlig samplingmodell den multinomiala fördelningen där n är fix. Om vi låter marginalkategorierna utgöra strata kan en produkt-nomial fördelning vara rimlig. I båda dessa fall är RR-estimatorn en maximum-likelihood estimator.

5.3.2 Beräkning av RR-skattningar.

Låt oss studera ett 3-dimensionellt fall där vi känner två marginaler. Vi antar att vi känner köns- resp åldersfördelningen i det fall vi analyserat tidigare.

Tabell 17.
RR-skattningar med köns- resp åldersmarginalen kända.

	Ålder	Inkomstklass 1	2	3	Summa
Män	1	0	49	293	342
	2	0	152	455	607
	3	21	167	583	771
	4	0	180	100	280
Summa män		21	548	1431	2000
Kvinnor	1	302	89	18	409
	2	353	228	62	643
	3	524	182	23	729
	4	110	66	44	220
Summa kvinnor		1289	565	147	2001

Om man gör en summering av skattningarna till åldersmarginalen erhåller vi (bortsett från avrundningsfel) just den kända marginalen (750, 1250, 1500, 500).

Vi har även gjort en kalibrering med utnyttjande av avståndsfunktion 2 och med de två kända marginalerna. Teorin stämmer - resultaten blir identiska med resultaten i ovanstående tabell!

6. Kommentarer till de tre metoderna.

A. Allmänna kommentarer.

En effektiv urvalsundersökning kräver en god urvalsdesign och goda estimatorer. Det innebär i sin tur att den utnyttjar mesta möjliga hjälpinformation, som man fördelar på ett "optimalt" sätt mellan de två momenten.

Det konventionella förfarandet innebär att inklusionssannolikheterna från urvalsdesignen läggs in på en urvalsfil och efter datainsamlingen eventuellt justeras p g a bortfall och övertäckning. Därefter brukar estimationsmomentet vidta. Det kräver tillgång till information från register, speciella datorprogram och det kräver goda statistiska kunskaper.

I framtiden kan det bli aktuellt att skicka (avidentifierade) filer med objektsinformation till t ex de statistikansvariga myndigheterna. Vid urvalsundersökningar kan det bli svårt att behålla en effektiv strategi om estimationsmomentet läggs ut på användarna. Exempelvis har SCB som central statistikmyndighet tillgång till mycket registerinformation, vilket inte användarna har.

Mot bakgrund av detta framtidsscenarium blir kalibreringstekniken intressant. Den avviker mot det konventionella förfarandet genom att all hjälpinformation läggs in i vikterna. Användarna kan då använda enkla tabuleringsprogram och ändå utnyttja effektiva estimationsmetoder.

I bilaga 2 åskådliggör vi med hjälp av en figur skillnaderna mellan ett konventionellt förfarande och ett förfarande som innehåller kalibrering av vikter.

De två andra metoderna har inte denna goda egenskap. Det finns dock även problem förknippade med kalibreringstekniken, vilket framgår av fortsättningen av detta avsnitt.

Kalibreringstekniken kräver, om man vill undvika alla inkonsistenser, att man har en fullständig överblick över vilken statistik, som ska framställas från undersökningen innan framställningen startar. Det är inte enbart en nackdel att tänka genom resultatredovisningen i förväg, men i en undersökning, som får många uppdrag i efterhand, kan det vara svårt att göra detta. Detta tillsammans med att framtidsscenariet inte gäller alla undersökningar är det relevant att även hålla övriga metoder "levande".

Kalibreringstekniken kräver att man har samma v-vikter för alla enheter och därför kanske inte lika bra, som när man i det konventionella förfarandet "specialsyr" estimator för olika parametrar. Denna restriktion betyder dock troligen lite.

Sammanjämkningstekniken har en fördel framför de andra metoderna, nämligen att den även kan beakta att marginalinformationen är behäftad med urvalsfel.

Skattningar vill vi ska vara unbiased och behäftade med små urvalsfel. Dessutom vill vi kunna beräkna och redovisa urvalsfelet. Skattningar avser inte bara hela populationen utan även redovisningsgrupper. I fortsättningen ska vi diskutera metoderna utifrån dessa aspekter.

Som vi har sett innehåller kalibreringstekniken en mängd alternativ. I denna rapport har vi redovisat fem olika avståndsfunktioner och ytterligare varianter finns beskrivna i Deville & Särndal [1992]. Den individuella vikten v_k ska också väljas. I bakgrunden finns också hela teorin om modellstödda metoder (se Särndal, Swensson & Wretman [1992]), som ger kunskap om hur man bör förfara vid val av urvalsdesign och generell regressionsestimator, vilket i sin tur påverkar val av poststrata, v-vikter, etc.

B. Väntevärdesriktighet.

Den kalibrerade estimatorn är asymptotiskt design unbiased, d v s när den baserar sig på många observationer är biasen försumbar. Därför torde esimatorn vara unbiased vid skattning av populationsparametrar, men frågan är vad som händer vid (små) redovisningsgrupper. Avståndsfunktion 1 leder till regressionsestimatorer och för dessa vet vi rätt väl vad som händer i redovisningsgrupper (se t ex Lundström [1991]). Biasen är inget stort problem ens för relativt små redovisningsgrupper. Om man väljer avståndsfunktion 2 kan man få enstaka stora vikter, vilket "...may produce unrealistic estimates for some domains.."(Deville & Särndal [1992]). I sådana fall blir frågan om väntevärdesriktigheten ointressant.

Den sammanjämkade estimatorn är modell-unbiased givet känd variansstruktur. Om man inte känner variansstrukturen och gissar "fel" är estimatorn design-biased.

C. Medelfel

Gemensamt för presenterade estimatorer är att de utnyttjar hjälpinformation. Det leder till att vi kan förvänta oss mindre medelfel.

Man kan förvänta sig att medelfelen för kalibrerade estimatorer för små redovisningsgrupper underskattas vid användande av formel (6), precis som fallet är med de generella regressionsestimatorerna.

Det finns ingen formel för beräkning av medelfel för den reconcilierade estimatorn. Det har inte heller funnits för RR-estimatorer, men Deville & Särndal [1992] ger via kalibreringstekniken en lösning.

D. Bortfall

En vanlig metod att reducera bortfallsbiasen är att poststratifiera på ett sådant sätt att strata blir homogena avseende undersökningsvariabeln och/eller avseende svarssannolikheterna. Om man hittar starka samband mellan undersökningsvariabeln och en hjälpvariabel och bildar en regressionsestimator kan man uppnå en betydande reduktion av bortfallsbiasen (se Bethlehem [1988]).

Den generella kalibreringsestimatorn har som specialfall nämnda estimatorer och har därför förutsättningen att reducera bortfallsbiasen. RR-estimatorn är också ett specialfall av kalibreringsestimatorn och torde ha denna biasreducerande effekt. I sammanjämkningsalternativet ger bortfallet biased skattningar av cellvärdena och därmed är inte grundvillkoren för estimatorn uppfyllda. Huruvida detta har någon praktiskt betydelse vet vi inte.

E. Beräkningsaspekter.

När vi enbart har kategoriska variabler (marginaler) som vi anpassar skattningarna mot, behöver man inte genomföra en kalibrering enligt det sätt vi visat i tidigare avsnitt.

Vi har sett att med raking-ratio-tekniken kan man bilda skattningar utgående från de konventionella skattningarna av cellvärdena. Om vi vill ha kalibrerade vikter, t ex för att beräkna medelfel för RR-estimatorn, kan man sätta $\hat{N}_{hi}^{w} = N_{hi}^{sam}$ i nedanstående formel.

$$w_k = d_k \frac{\hat{N}_{hi}^{\,w}}{\hat{N}_{hi}} \qquad (26)$$

Detta är de vikter som man erhåller med avståndsfunktion 2 i kalibreringsmetoden.

Man kan även utföra en likartad beräkning av kalibrerade vikter för avståndsfunktion 1 genom att utgå från de konventionella skattningarna. Följande iterativa förfarande finns beskrivet i Särndal & Deville [1990] och i Andersson [1991]. Lösningen avser fallet med två dimensioner.

Startvärdet sätts till:

$N_{hi}^{(0)} = \hat{N}_{hi}$, d v s vi startar med de konventionella estimaten samt $b_i^{(0)} = 0$.

Vid steg k beräknas:

$$a_h^{(k)} = \frac{N_{h.} - \sum_i^c (N_{hi}^{(k-1)} + N_{hi}^{(k-1)} b_i^{(k-1)})}{\sum_i^c N_{hi}^{(k-1)}} \qquad (27)$$

och

$$b_i^{(k)} = \frac{N_{.i} - \sum_h^r (N_{hi}^{(k-1)} + N_{hi}^{(k-1)} a_h^{(k)})}{\sum_h^r N_{hi}^{(k-1)}} \text{ och sätt } b_c = 0. \qquad (28)$$

$$N_{hi}^{(k)} = N_{hi}^{(k-1)} (1 + a_h^{(k)} + b_i^{(k)}) \qquad (29)$$

Iterationen avbryts då maximala skillnaden mellan två stegs cellskattningar är mindre än ett förutbestämt värde, säg vid k=k_0. (Ofta måste man göra fler steg i denna iteration än i IPF-proceduren.) Vi har då

$$N_{hi}^{lin} = N_{hi}^{(k_0)} .$$

Vi har framställt ett APL-program som utför denna beräkning.

Vi har därvid en punktskattning av cellvärdena som överensstämmer med den skattning vi får genom att utnyttja vikter kalibrerade med avståndsfunktion 1 ($v_k = 1$ för alla k). Om vi går via nyss presenterade metod kan vi bilda kalibrerade vikter genom att utnyttja formel (26), där vi sätter $\hat{N}_{hi}^{w} = N_{hi}^{lin}$

Frågan är dock vilken beräkningshjälp som kan fås? Vi har visat att CALMAR är en möjlighet vid kalibrering. I studiens slutskede har SAS/Win installerats på nätet och det visar sig att den versionen har betydligt större kapacitet än det vi tidigare använt. Vi har provat att köra CALMAR på ett urval av 4000 personer. Vi antog att vi känner marginalerna kön*ålder och totalerna för registerinkomst för män och kvinnor. Beräkningarna är gjorda med avståndsfunktion 1. Om vi ska göra en generell kalibrering med både kategoriska och kontinuerliga variabler och med många observationer räcker inte PC-varianten till (i varje fall inte med den utrustning som använts här). Vi har inte prövat stordatoralternativet och kan därför inte uttala oss om dess möjlighet. I.N.S.E.E. visar dock att, med ca 7000 observationer och upp till 7 marginaler, är det möjligt att utnyttja CALMAR.

REFERENSER

C. Andersson [1991]. Bortfallet i FoB90 II - förslag till kompensationsvägning. SCB, I/MET.

J.G. Bethlehem [1988]. Reduction of Nonresponse Bias Through Regression Estimation. Journal of Official Statistics, Vol 4, No. 3, 1988. Statistics Sweden.

J-C Deville & C-E Särndal [1990]. Calibration and Generalized Raking Techniques in Survey Sampling. R&D Report 1990:1. Statistics Sweden.

J-C Deville & C-E Särndal [1992]. Calibration Estimators in Survey Sampling. JASA, June 1992, Vol. 87, No. 418.

W.E. Deming & F.F. Stephan (1940). On a Least Squares Adjustment of a Sampled Frequency Table When the Expected Marginal Totals are Known. Annals of Mathematical Statistics, 11.

S.E. Fienberg [1978]. The Analysis of Cross-Classified Categorical Data. The MIT Press.

C.T. Ireland & S. Kullback (1968). Contingency tables with given marginals. Biometrica 55.

J. Johnston [1972]. Econometric Methods. McGraw-Hill, Inc.

H. Jönrup [1977]. Korstabulering med utnyttjande av supplementär information. Statistisk Tidskrift 1977:5. SCB.

S. Lundström [1991]. Egenskaper hos modellbaserade estimatorer för redovisningsgruppstotaler. R&D Report 1991:19. Statistics Sweden.

H. Lütjohann [1985]. Reconciling Tables and Margins Using Least-Squares Promemorior från P/STM. Statistics Sweden.

H. Martens T. Naes [1989]. Mulivariate Calibration. Wiley &Sons Ltd.

O. Sautory [1992]. Calibration on Known Marginal Counts for Sample Surveys: Practical Experiencies at I.N.S.E.E. Proceedings från Workshop on uses of auxiliary information in surveys. Statistics Sweden och University of Örebro, Sweden.

D.M. Stukel & R. Boyer [1992]. Calibration Estimation: An Application to the Canadian Labour Force Survey. SSMD #009E. Statistics Canada.

C-E Särndal, B. Swensson & J. Wretman [1992]. Model Assisted Survey Sampling. Springer-Verlag New York, Inc.

H. Theil [1971]. Principles of Econometrics. North-Holland and Wiley.

Bilaga 1

Exempel på körning av CALMAR - programmet.

CALMAR är skrivet i SAS-macro och finns i en stordator- och en PC-version. I föreliggande exempel har vi använt den senare versionen.

Det exempel vi visar här motsvarar APL-bearbetningen till kalibrering 2, d v s där vi söker skattningar av antalet personer i populationen fördelade på kön, ålder och inkomst och där populationstotalerna i kön*ålder är kända samt skattningar av totalinkomsterna för män resp kvinnor där vi känner motsvarande populationstotaler för registerinkomst. Kalibreringen utförs med avståndsfunktion 1.

Vi har en fil över urvalet av 200 personer med variablerna kön, ålder, klassindelad inkomst, inkomst, registerinkomst och inkusionssannolikheten.

Först måste man skapa en SAS-fil, som innehåller kön, ålder och klassindelad inkomst (vi benämner dem för KON, ALDER och INK) och så bildar vi KONALD, som består av 8 kategorier, där 1-4 står för män i de fyra åldersklasserna och 5-8 är motsvarande kategorier för kvinnor. Dessutom bildar vi XMA , som är lika med registerinkomst för män, men lika med noll för kvinnor. Vi bildar också XKV, vilken är motsvarande variabel för kvinnor. Dessutom bildar vi VIKT, som utgör inverterade inklusionssannolikheten. Filen kallar vi för S.URV.

CALMAR måste först kompileras (laddas in och köras) innan det kan användas.

Marginalinformationen måste läggas in som särskild sats i programmet. I nedanstående utdrag visas både hur marginalinformationen läggs in och hur parametervärdena sätts in:

```
LIBNAME BASE 'C:\SASNET\SASDLIB';
%global marges;
/*                                                          */
/*  Hr lgges vrdena fr de knda marginalerna in     */
/*                                                          */
/*                                                          */
%let marges={400,600,700,300,350,650,800,200,694024,897128};
/*                                                          */
/*  Hr sker anropet till proceduren                        */
/*                                                          */
/*                                                          */
%calmar(data=s.urv,
        m=1,
        vqual=konald,
        vquan=xma xkv,
        nbmod=8,
        poids=vikt,
        poidc=viktf);
```

I nedanstående utdrag visar vi hela uppsättningen av parametrar och kommenterar (där vi har något att säga).

```
%MACRO CALMAR   (DATA      =          , /* TABLE SAS EN ENTREE
                 M         =          , /* FONCTION DISTANCE
                 VQUAL     =          , /* VARIABLES QUALITATIVES
                 VQUAN     =          , /* VARIABLES QUANTITATIVES
                 NBMOD     =          , /* NOMBRE DE MODALITES DES
                                         /*  VARIABLES QUALITATIVES
                 POIDS     =          , /* PONDERATION INITIALE
                 POIDC     =          , /* PONDERATION FINALE (CALAGE)
                 POIDL     =          , /* LABEL DE LA PONDERATION
                 LO        =  0.00    , /* BORNE INF (LOGIT OU LINEAIRE
                                         /*  TRONQUEE
                 UP        =  4.00    , /* BORNE SUP (LOGIT OU LINEAIRE
                                         /*  TRONQUEE
                 REGION    =          , /* LES VARIABLES REGION ET LISTE
                 LISTE     =          , /*  PRISES ENSEMBLE PERMETTENT
                                         /*  D'IDENTIFIER LES MENAGES
                 TRANSF    =          , /* TRANSFORMATION EN VUE
                                         /*  D'ASSIGNER UN MEME POIDS
                                         /*  A TOUS LES MEMBRES D'UN
                 EDIT      =          , /* EDITION DES POIDS PAR CELLULE
                 STAT      =          , /* STATISTIQUES SUR LES POIDS
                 SEUIL     = 0.0005); /* SEUIL POUR LE TEST D'ARRET
```

- DATA anger i vanlig SAS-ordning namnet på INPUT-filen.

- M anger vilken avståndsfunktion man önskar. I exemplet har vi satt M=1, vilket innebär att vi har utnyttjar den första avståndsfunktionen. (Methode: Lineaire).

- VQUAL anger namnen på de kategoriska variablerna. I exemplet KONALD.

- VQUAN anger namnen på de kontinuerliga (kvantitativa) variablerna. I exemplet har XMA och XKV. Om man inte har någon utelämnas parametern.

- NBMOD anger antalet kategorier för de variabler som är uppräknade under VQUAL (se dock nedan). I exemplet har KONALD 8 kategorier (= 2 x 4).

- POIDS anger de inverterade inklusionssannolikheterna. I exemplet är det VIKT.

- POIDC anger det namn vi vi sätta på de kalibrerade vikterna. I exemplet sätter vi POIDC=VIKTF.

- LO resp UP anger nedre och övre gränsen för de begränsade avståndsfunktionerna 1 och 3 (se Deville & Särndal [1992]).

Övriga parametrar har vi endast vaga uppfattningar om: LISTE och TRANF har att göra med hushållsundersökningar, där metoden föreslagen av Lemaître and Dufour [1987] används. STAT ger viss statistik över de kalibrerade vikterna. När skillnaden mellan två iterationssteg understiger SEUIL-värdet avbryts processen.

När man har flera kända marginaler anges den första marginalen och därefter följer den andra marginalen med undantag av totalen för den sista kategorin. Övriga marginaler läggs därefter, men med samma reducering som den andra variabeln. Reduceringen görs för att undvika singulära matriser. Under VBMOD anges det antal kategorier som finns med under "%let marges".

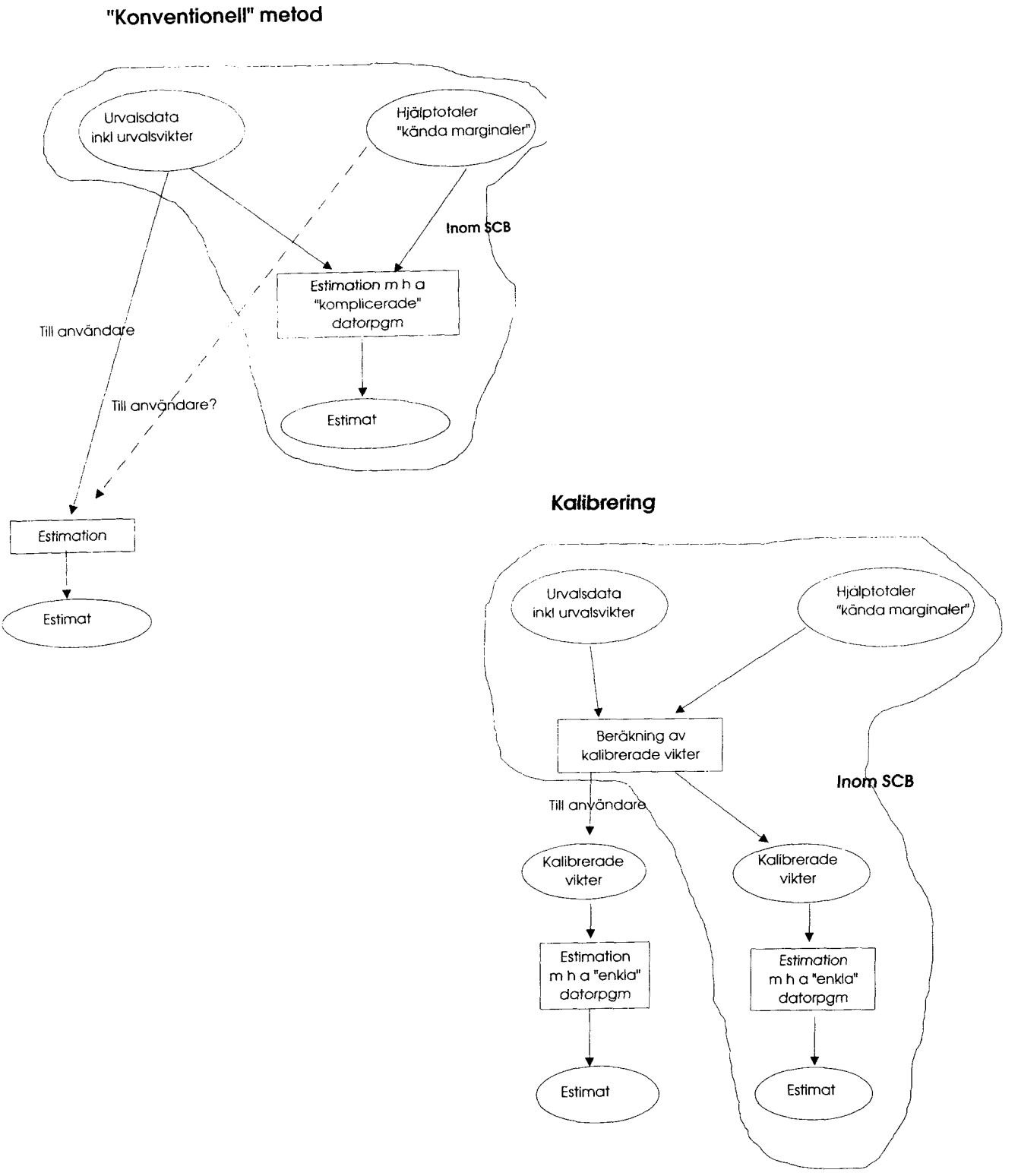
OUTPUTen består av data över körningen, t ex skattningen av marginalerna med ursprungsvikterna och motsvarande skattning med kalibrerade vikter. Dessutom erhålles en SAS-fil med de kalibrerade vikterna. I exemplet heter den S.POIDS och variabeln heter, som tidigare påpekats, VIKTF. Dessutom kan man erhålla en del statistik över de kalibrerade vikterna via parametern STAT.

REFERENSER.

Lemaître, G. and Dufour, J [1987]. An Integrated Method for Weighting Persons and Families. Survey Methodology, vol 87.

Bilaga 2

Figur över konventionell metod resp kalibrering.



R & D Reports är en för U/ADB och U/STM gemensam publikationsserie, som fr o m 1988-01-01 ersätter de tidigare "gula" och "gröna" serierna.

R & D Reports Statistics Sweden are published by the Department of Research & Development within Statistics Sweden. Reports dealing with statistical methods have green (grön) covers. Reports dealing with EDP methods have yellow (gul) covers.

Reports published 1993:

1993:1 (grön)	Kvalitetsfonden 1991/92 - Projekt, handläggning och utvärdering (Roland Friberg)
1993:2 (grön)	Översyn av Undersökningen av lastbilstransporter i Sverige (UVAV) (Bengt Rosén, Mahnaz Zamani)
1993:3 (grön)	Bortfallsbarometern nr 8 (Mats Bergdahl, Pär Brundell, Jan Hörngren, Håkan Lindén, Peter Lundquist, Monica Rennermalm)
1993:4 (gul)	Guidelines on the Design and Implementation of Statistical Meta-information Systems (Bo Sundgren)
1993:5 (grön)	Kvalitetsrapporten 1995 - Utveckling av kvaliteten för SCBs statistikproduktion (Jan Eklöf, Per Nilsson)

Tidigare utgivna **R&D Reports** kan beställas genom Ingvar Andersson, SCB, U/LEDN, 115 81 STOCKHOLM (telefon 08-783 41 47, telefax 08-783 45 99).

R&D Reports listed above as well as issues from 1988-92 can - in case they are still in stock - be ordered from Statistics Sweden, att. Ingvar Andersson U/LEDN, S-115 81 STOCKHOLM, SWEDEN (telephone nr 46 08 783 41 47, telefax nr +46 08 783 45 99).