

Om HINK-undersökningens estimationsförfarande

Bengt Rosén



R & D Report
Statistics Sweden
Research - Methods - Development
U/STM - 30

Från trycket Februari 1987
Producent Statistiska centralbyrån, Enheten för statistiska metoder
Ansvarig utgivare Staffan Wahlström
Förfrågningar Bengt Rosén, tel. 08 7834490

© 1987, Statistiska centralbyrån
ISSN 0283-8680
Printed in Sweden
Garnisonstryckeriet, Stockholm 1987

INLEDNING

TILL

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

Föregångare:

Promemorior från U/STM / Statistiska centralbyrån. – Stockholm : Statistiska centralbyrån, 1986. – Nr 25-28.

Efterföljare:

R & D report : research, methods, development / Statistics Sweden. – Stockholm : Statistiska centralbyrån, 1988-2004. – Nr. 1988:1-2004:2.

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

U/STM

Bengt Rosén/Elisabeth Sandström

OM HINK-UNDERSÖKNINGENS ESTIMATIONSFÖRFARANDE

AV

BENGT ROSÉN

under medverkan av

Marianne Andersson
Tor Bengtsson
Håkan Björk
Leif Johansson
Håkan Lindström
Ingvar Lyselius
Lennart Nordberg
Göran Råbäck
Håkan Sellerfors
Peter Simon

SUMMARY

ON THE ESTIMATION PROCEDURE FOR THE SURVEY ON FAMILY INCOMES (HINK)

SCB conducts yearly the sample survey HINK with the objective to describe income conditions, primarily for different groups of families but also for different groups of individuals.

The aim of this report is that it should serve as a "manual" for how to compute estimates of group means and group totals, as well as estimates of the variances of the point estimators. The idea is that the manual should be possible to use on two different levels of insight into the HINK survey.

For the "lower" insight level, the person who wants to carry out the estimation has available a final register from the HINK survey, including appropriate weighting values. The "parameters" at disposal are

- . the variable of interest,
- . the group of interest.

Sections 1-4 aim at giving a fairly short route to a presentation of how the programme packages TAB68 and SMED83 should be set into operation on a (final) HINK register in order to deliver desired estimates.

On the "higher" insight level, the report also presents how to compute the weighting values, thereby paying regard to complications as non-response, overcoverage etc. For a proper description (and understanding) of this part of the estimation problem, one needs a more detailed account of various practical as well as technical aspects of the entire HINK survey process. These matters are dealt with in the Appendices 1 and 2.

FÖRETAL

Denna skrift är ett utflöde från den genomlysning av HINK-undersökningen som U/STM genomförde under 1986. Som framgår av titeln behandlar den HINKs estimation.

I genomlysningsarbetet gjordes tillsammans med HINK-sektionen bl a en genomgång av estimationsförfarandet, vilken gav insikt i hittillsvarande metodik och förfaranden samt även ledde till vissa nyheter. Ett huvudsyfte med föreliggande skrift är att samla ihop "gammalt och nytt" angående HINKs estimationsförfarande samt att dokumentera det hela, företrädesvis för framtida bruk. Vid utformningen av denna skrift har en vägledande tanke varit att den skall kunna fungera som "HINK-estimations-manual", och detta på två olika insiktsnivåer.

På den "lägre" insiktsnivån betraktas uppräkningsvikterna som redan beräknade och införda i de aktuella HINK-registren. De storheter som står till "estimatörens" förfogande är, dels aktuell redovisningsgrupp, dels aktuell variabel. Beskrivningen av estimationsförfarandet på denna lägre ambitionsnivå ges i Avsnitten 1-4. Härvid har ett mål varit att någorlunda snabbt nå fram till innehållet i Avsnitt 4, vilket anger hur man applicerar programvarorna TAB68 och SMED83 på HINKs register (inkluderande uppräkningsvikter) för att erhålla önskade skattningar, inklusive skattningar av estimatorernas medelfel. För att uppnå snabbhetsmålet ges i Avsnitten 1-4 en förenklad (och faktiskt något oegentlig) beskrivning av HINK-undersökningen, vilken dock förhoppningsvis ger tillräcklig bakgrund för att man skall kunna genomföra adekvata TAB68- och SMED83-körningar.

På den "högre" insiktsnivån handlar det om att också beräkna uppräkningsvikterna. För att kunna göra detta med förståelse fordras en mer ingående och detaljerad kännedom om hela HINK-processen; definitioner och klassificeringar, urvalsförfarandet, datainsamling och registerföring, bortfallshantering m m. Beskrivning av dessa saker ges i Appendixen 1 och 2, och där beskrivs också förfarandet för beräkning av uppräkningsvikterna.

Som framgår av titelbladet har denna skrift åsatts epitetet preliminär version. Skälet till detta är följande. HINK är en förhållandevis komplicerad undersökning, som innehåller ett antal olika moment i vars genomförande ett flertal olika personer är involverade. Som konsekvens härav har, vilket också framgår av titelbladet, ett flertal olika personer deltagit i den "remissomgång" som föregått den här föreliggande rapporten. Därvid har en mängd tidigare felaktigheter, luckor o dyl undanröjts, men "sammanfasningen" är troligen ännu inte fullständig. Det mest effektiva sättet att eliminera kvarstående brister torde vara att låta denna "försöksutgåva" konfronteras med den verklighet som HINK-produk-

tionen, och då särskilt framttagandet av SMen utgör. Efter en SM-omgång, säg i medio 1987, är det nog lämpligt att samla ihop erfarenheter och kritik som kan ge underlag för en förhoppningsvis väl "avslipad" estimationsmanual.

Jag emottar tacksamt synpunkter (kritiska eller ej) på föreliggande försöksutgåva.

I december 1986

Bengt Rosén

INNEHÅLLSFÖRTECKNING

1	NÅGRA GRUNDLÄGGANDE BEGREPP OCH BETECKNINGAR I ANSLUTNING TILL HINK-UNDERSÖKNINGEN
2	URVAL OCH ORGANISATION AV URVALSINFORMATIONEN
2.1	Urval och register
2.2	Ett årsurvals probabilistiska egenskaper
3	ESTIMATORER OCH ESTIMATORVARIANSER
3.1	Estimation för grupper av familjeenheter
3.2	Estimation för grupper av individer
4	DATORBERÄKNINGAR MED HJÄLP AV TAB68 OCH SMED83
4.1	Några allmänna saker
4.2	Skattningar för grupper av familjeenheter
4.3	Skattningar för grupper av individer
APPENDIX 1	YTTERLIGARE BESKRIVNING AV HINKs KLASSIFICERINGAR, URVAL, DATAINSAMLING OCH ESTIMATION
A1.1	Ytterligare klassificering av individer och familjeenheter
A1.2	Om HINKs urvalsförfarande
A1.2.1	Det för år t nya urvalet
A1.2.2	Det för år t gamla urvalet
A1.2.3	En tänkbar modifiering av förfarandet för dragning av nytt och gammalt (del)urval
A1.3	Insamling och registerföring av urvalsinformation
A1.4	Estimation
A1.4.1	Allmänt om skattning av grupptotaler på basis av stratifierade urval
A1.4.1.1	Estimation under antagande att inget bortfall föreligger
A1.4.1.2	Estimation när bortfall föreligger
A1.4.2	Sammanfattning av beteckningar, formler och beräkningsalgoritmer i Avsnitten 1-4 med den mer fullständiga beskrivning av HINK-undersökningen som givits i detta Appendix

APPENDIX 2	OM VAL AV ESTIMATORVIKTNING OCH INFORMATIONSBENÄGENHETS- KVOT
A2.1	Om estimatorviktningen
A2.2	Om åsättande av värden på informationsbenägenhetskvoterna
A2.2.1	Några teoretiska överväganden
A2.2.2	Fortsatt diskussion mot bakgrund av empiriska värden

1 NÅGRA GRUNDLÄGGANDE BEGREPP OCH BETECKNINGAR I ANSLUTNING TILL HINK-UNDERSÖKNINGEN

HINK-undersökningens syfte är att redovisa inkomstförhållanden och inkomstutveckling för grupper av familjeenheter och grupper av individer, företrädesvis på riksnivå. HINK är en urvalsundersökning som genomförs årligen.

De i HINK aktuella populationerna och redovisningsgrupperna består som sagt av individer och familjeenheter och sådana populationer undergår förändringar allteftersom tiden går (individer föds och dör, sammanboendeförhållanden ingås och upplöses, för att nämna några exempel). Bl a sådana förändringar i tiden utgör en komplicerande faktor när man skall räkna upp urvalsobservationerna i HINK. Dock, för att göra denna introducerande beskrivning någorlunda lättassimilerad skall vi i Avsnitten 1-4 arbeta under nedanstående förenklande förutsättningar;

dels att "allt av intresse" för ett visst undersökningsår (populationsfastställande, urvalsdragning, observationsinsamling) sker vid en och samma tidpunkt,

dels att målpopulationerna avseende familjeenheter och individer sammanfaller med de nedan införda populationerna U^F respektive U^I , dvs vi antar att varken över- eller undertäckning föreligger,

dels att datainsamlingen sker utan bortfall.

I Appendixen 1 och 2 kommer vi att ge en utvidgad diskussion av begrepp och beteckningar för det fall som föreligger i verkligheten, dvs utan de ovan nämnda förenklande antagandena. Där visar vi också att den i praktiken föreliggande situationen kan återföras på den förenklade situationen.

I det fortsatta resonemanget betraktar vi alltså förhållandena vid en och samma (fixerade) tidpunkt, vilken också uppfattas som "aktuellt HINK-år". Låt

$$U^I = \text{populationen av alla individer (i landet vid den aktuella tidpunkten).} \quad (1.1)$$

Individerna är antingen vuxna (≥ 18 år) eller barn (< 18 år), och denna klassificering ger uppdelningen

$$U^I = V^I \cup B, \quad (1.2)$$

där

$$V^I = \text{mängden av vuxna individer} \quad (1.3)$$

$$B = \text{mängden av barn} \quad (1.4)$$

Allmänna element i V^I och B betecknas fortsättningsvis med i respektive b.

I första omgången lämnar vi barnen utanför betraktelserna.

För vissa par av vuxna individer föreligger sammanboende och ett sådant par säges utgöra vuxendelen av en tvåvuxenfamiljeenhet. En vuxen individ ingår i högst ett sammanboendeförhållande, men behöver inte ingå i något. Om ej, säges individen vara ensamstående och utgör då vuxendelen av en envuxenfamiljeenhet.

Anmärkning 1.1 (a). För våra syften här behöver vi inte gå in på den precisa definitionen av begreppet sammanboende, bara förutsätta att en väldefinierad "hoppningsregel" föreligger. I grova drag gäller att två vuxna säges vara sammanboende om de är av motsatt kön och bor tillsammans i äktenskap eller under äktenskapsliknande förhållanden. (Individernas civilståndsförhållanden bortses ifrån.)

(b). Något oegentligt i förhållande till SCBs terminologiska riktlinjer, kommer vi emellanåt att tillåta oss att använda ordet famili som synonym till familjeenhet. Oegentligheten ligger då i att vi räknar en ensamstående vuxen utan barn som en familj medan de nämnda riktlinjerna inte räknar denna typ av familjeenhet som familj, men dock alla andra typer.

(c). Vad gäller just termen familjeenhet så gjordes från och med SMen för HINK 1984 en terminologisk omläggning. Tidigare användes ordet hushåll, ibland även HINK-hushåll, för det som här kallas familjeenhet. ●

Sammanboenderelationen ger en uppdelning av vuxenindividpopulationen,

$$V^I = D^I \cup E^I, \quad (1.5)$$

där

$$D^I = \{i: i \text{ är sammanboende}\}, \quad (1.6)$$

$$E^I = \{i: i \text{ är ensamstående}\}. \quad (1.7)$$

Sammanboenderelationen inducerar också en ny population, V^F , vars objekt är vuxendelarna av familjeenheterna. Denna kallas för populationen av vuxenfamiljeenheter. I enlighet med vad som sägs i (b) i anmärkningen ovan kommer vi att tillåta oss att ibland säga bara vuxenfamiljer.

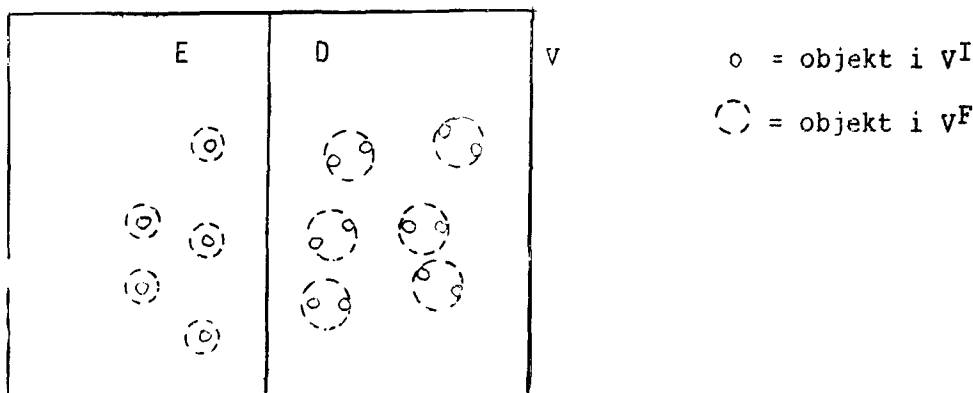
Fortsättningsvis använder vi bokstaven d för att beteckna ett allmänt element i V^F .

Till uppdelningen (1.5) av V^I svarar en analog uppdelning av V^F ,

$$V^F = D^F \cup E^F \quad (1.8)$$

i en- respektive tvåvuxenfamiljeenheter.

Figur 1



Med de små ringarna som objekt beskriver Figur 1 uppdelningen (1.5) av V^I , medan den med de stora ringarna som objekt beskriver uppdelningen (1.8) av V^F .

Härnäst inför vi beteckningar för några naturliga (med tanke på den aktuella tolkningen) avbildningar mellan de införda mängderna.

$$m: D^I \rightarrow D^I; m(i) = \text{sammanboendepartnern (ibland också kallad } \underline{\text{moatjén}} \text{) till individ } i, \quad (1.9)$$

$$d: V^I \rightarrow V^F; d(i) = \text{den vuxenfamiljeenhet som individ } i \text{ ingår i,} \quad (1.10)$$

$$\{i_1, i_2\}: D^F \rightarrow D^I \otimes D^I; \{i_1(d), i_2(d)\} = \text{det sammanboendepar som utgör vuxenhushåll } d. \text{ (Härvid betraktas } \{i_1, i_2\} \text{ som ett i matematisk mening } \underline{\text{oordnat par}}, \text{ och } D^I \otimes D^I \text{ avses beteckna "oordnad Cartesiansk produkt").} \quad (1.11)$$

Följande relationer inses vid litet eftertanke

$$\begin{aligned} d(i) &= d(m(i)), \quad i \in D^I, \quad d(i_1(d)) = d(i_2(d)) = d, \quad d \in V^F, \\ i_2(d) &= m(i_1(d)), \quad d \in V^F \end{aligned} \quad (1.12)$$

I nästa steg för vi in barnen i bilden. Vi förutsätter att varje barn är relaterat till en unik vuxenfamiljeenhet, nämligen till den vuxenfamiljeenhet som har det ekonomiska "vardagsansvaret" för barnet. Låt

$$B(d) = \text{barnaskaran} (= \text{mängden av barn}) \text{ som hör till} \\ \text{vuxenfamilj } d, d \in V^F. \quad (1.13)$$

Notera att mängden $B(d)$ mycket väl kan vara tom.

Populationen av familjeenheter, U^F , utgörs av de objekt vi får när vi med varje vuxenfamilj förenar den barnaskara som hör till vuxenfamiljen ifråga, dvs

$$U^F = \{d \cup B(d); d \in V^F\}. \quad (1.14)$$

Definitionen (1.14) ger omgående en naturlig ett-ett-korrespondens mellan familjeenhetspopulationen U^F och vuxenfamiljeenhetspopulationen V^F , nämligen den som erhålles genom att man identifierar en familjeenhet med sin vuxendel. Vi kommer därför ofta att tillåta oss att inte göra någon åtskillnad mellan U^F och V^F . Bland annat tillåter vi oss att använda bokstaven d även för att beteckna ett allmänt element i U^F .

Med en redovisningsgrupp på individnivå menar vi en delmängd av individpopulationen U^I och med en redovisningsgrupp på familjeenhetsnivå en delmängd av endera V^H eller U^H (vilken man nu tycker att det är lämpligt att arbeta med).

Med en individvariabel y menar vi en funktion på individpopulationen U^I , $y = \{y_i; i \in V^I, y_p; p \in B\}$, med en vuxenindividvariabel y , en funktion på V^I , $y = \{y_i; i \in V^I\}$ och med en familje(enhets)variabel y en funktion på endera av familje(enhets)populationerna V^F eller U^F , $y = \{y_d; d \in V^F\}$ eller $y = \{y_d; d \in U^F\}$.

För en familjevariabel y och en redovisningsgrupp G på familje(enhets)nivå definierar vi grupptotalen $\tau(y; G)$, gruppstorleken $g(G)$ och gruppmedelvärdet $\mu(y; G)$ enligt nedanstående formler (1.16)-(1.18). Vi ger några alternativa framställningar av storheterna i fråga och 1 betecknar familjevariabeln $1 = \{y_d = 1; d \in V^F(U^F)\}$ medan 1_G betecknar gruppindikatorvariabeln.

$$1_G(d) = \begin{cases} 1 & d \in G \\ 0 & d \in U^F \setminus G. \end{cases}$$

Vidare, låt

$$\#(\cdot) = \text{antal element i } (\cdot), \quad (1.15)$$

$$\tau(\underline{y}; G) = \sum_{d \in G} y_d = \sum_{d \in U^F} y_d 1_G(d), \quad (1.16)$$

$$g(G) = \#G = \sum_{d \in G} 1 = \sum_{d \in U^F} 1_G(d) = \tau(1; G), \quad (1.17)$$

$$\mu(\underline{y}; G) = \frac{\tau(\underline{y}; G)}{g(G)} = \frac{\tau(\underline{y}; G)}{\tau(1; G)}. \quad (1.18)$$

Ovanstående begrepp och formlerna (1.14)-(1.18) har uppenbara motsvarigheter för grupper av individer och grupper av vuxenindivider. Vi nöjer oss med att skriva ned följande formel för individergrupp totalen (vid individvariabel $\underline{y}$ och individgrupp G).

$$\begin{aligned} \tau(\underline{y}; G) &= \sum_{j \in G} y_j = \sum_{j \in U^I} y_j 1_G(j) = \sum_{i \in V^I \cap G} y_i + \sum_{b \in B \cap G} y_b = \\ &= \sum_{i \in V^I} y_i 1_G(i) + \sum_{b \in B} y_b 1_G(b). \end{aligned} \quad (1.19)$$

2 URVAL OCH ORGANISATION AV URVALSINFORMATIONEN

2.1 URVAL OCH REGISTER

Vi fortsätter att betrakta situationen med ett fixerat undersökningsår och vi bibehåller de i föregående avsnitt införda förenklande antagandena bl a att det inte förekommer bortfall i datainsamlingsprocessen. I praktiken förekommer naturligtvis bortfall och i Appendixen 1 och 2 ger vi en förhållandevis utförlig diskussion av hur bortfallet skall beaktas i utvärderingsprocessen.

HINKs årliga urval kan uppfattas som sammansatt av två åtskilda delurval. Skälet är att HINKs design inrymmer två urvalspaneler, där varje urvalspanel observeras under två på varandra följande år och där urvalspanelerna är "fasförskjutna" ett år i förhållande till varandra. Vid de estimationsförfaranden som vi skall behandla framdeles kommer panelteknikens specifika möjligheter dock ej att utnyttjas (t ex möjligheten att för objekten i "andraårspanelen" mäta förändringen från året innan). För enkelhetens skull föredrar vi därför att undertrycka begreppet "panel" och i stället tala om två delurval per år. I Appendix 1 ger vi en ingående beskrivning av hur de två delurvalen (urvalspanelerna) genereras.

Vid beskrivningen av urvalsförfarandet och insamlandet av urvalsinformationen börjar vi i den "hanteringsmässiga" ändan och tar därefter upp urvalets probabilistiska egenskaper.

De två delurvalen är i det väsentliga ekvivalenta och vi börjar med att beskriva gången i urvals- och datainsamlingsprocessen för ett (vilket som helst) av dem.

Först dras från vuxenindividpopulationen V^I ett primärt individurval, av storlek n . Vi betecknar de urvalda idividerna med $U_1, U_2, \dots, U_n$. Detta primära individurval genererar sedan det för delurvalet totala individurvalet genom att man till urvalet för alla individer (vuxna och barn) som ingår i samma familjeenhet som de primärt urvalda individerna. De nytillkommande individerna säger vi vara sekundärt urvalda.

För delurvalet ifråga upprättas delurvalets familjeenhetsregister, R_F ,

$$R_F = \left\{ \begin{array}{ll} \text{Post 1} & U_1:s \text{ familjeenhet} \\ & 2 \quad U_2:s \text{ familjeenhet} \\ & \cdot \\ & \cdot \\ & \cdot \\ & n \quad U_n:s \text{ familjeenhet.} \end{array} \right. \quad (2.1)$$

Notera att familjeenhetsregistret också kan uppfattas som ett register över de primära urvalsindividerna.

För urvalet upprättas också delurvalets individregister R_I , vilket ställs upp enligt nedan, där a_i betecknar antalet medlemmar i U_i :s hushåll, $i=1,2,\dots,n$.

$$R_I = \left\{ \begin{array}{l} \text{Post 1 } U_1 \text{ (= Medlem 1 i } U_1\text{:s familjeenhet)} \\ \quad 2 \text{ Medlem 2 i } U_1\text{:s familjeenhet} \\ \quad 3 \text{ Medlem 3 i } U_1\text{:s familjeenhet} \\ \quad \cdot \\ \quad \cdot \\ \quad \cdot \\ \quad \text{Medlem } a_1 \text{ i } U_1\text{:s familjeenhet} \\ \\ U_2 \text{ (= Medlem 1 i } U_2\text{:s familjeenhet)} \\ \quad \text{Medlem 2 i } U_2\text{:s familjeenhet} \\ \quad \cdot \\ \quad \cdot \\ \quad \cdot \\ \quad \text{Medlem } a_2 \text{ i } U_2\text{:s familjeenhet} \\ \\ \cdot \\ \cdot \\ \cdot \\ U_n \text{ (= Medlem 1 i } U_n\text{:s familjeenhet)} \\ \quad \text{Medlem 2 i } U_n\text{:s familjeenhet} \\ \quad \cdot \\ \quad \cdot \\ \quad \cdot \\ \quad \text{Medlem } a_n \text{ i } U_n\text{:s familjeenhet.} \end{array} \right. \quad (2.2)$$

Antalet poster i individregistret R_I är alltså $a_1 + a_2 + \dots + a_n$.

Kommentar 2.1: I individregistret står urvalspersonen alltid som Medlem 1. Oftast är det också så, att om en familjeenhet omfattar två vuxna så står den andre vuxne som Medlem 2 i familjeenheten ifråga. Denna regel gäller dock inte utan undantag. Om familjens andre vuxne är "tillskriven" (se Appendix 1) så kan denne få ett annat medlemsnummer än 2.

I de så upprättade familjeenhetsregistrets och individregistrets poster införs sedan information avseende familjeenheterna och individerna i registren. Denna information hämtas från ett flertal olika källor; en del information avser själva urvalsdragningen, en del insamlas med hjälp av enkät, "kärninformationen" hämtas från individernas deklarationsblanketter för det aktuella året, en del kommer från myndigheter som Riksförsäkringsverket, Värnpliktsverket, Centrala Studiestödsnämnden m fl.

Till stor del erhålles uppgifter avseende en familj genom att aggregera uppgifter över individerna i familjen ifråga. (Så erhålles t ex familjens faktorinkomst som summan av familjemedlemmarnas faktorinkomster.) Vissa intäkts- och utgiftsposter hör dock så uttalat till familjen ifråga (t ex bostadsbidrag och socialbidrag) att de förs in endast i familjeregistret (och dessa uppgifter kan således ej erhållas via individregistret).

Därmed är tillräckligt sagt om hur man förfar med ett delurval och vi övergår till det för det aktuella året totala urvalet. Som tidigare sagts, så utgörs detta av två delurval, vilka vi kallar för Delurval 1 respektive Delurval 2. Vi använder oss fortsättningsvis av följande be-teckningskonvention:

Superindex ⁽¹⁾ markerar att storheten ifråga hänför sig till Delurval 1 och ⁽²⁾ att den hänför sig till Delurval 2. (2.3)

Vart och ett av de två delurvalen dras enligt det ovan beskrivna mönstret; först ett primärt vuxenindividurval vilket sedan kompletteras med "övriga familjemedlemmar". För de två delurvalen upprättas individregister $\mathcal{R}_I^{(1)}$ och $\mathcal{R}_I^{(2)}$ samt familjeenhetsregister $\mathcal{R}_F^{(1)}$ och $\mathcal{R}_F^{(2)}$.

Det för det aktuella året totala individregistret $\mathcal{R}_I$ respektive det totala familjeenhetsregistret uppställs enligt nedan,

$$\mathcal{R}_I = \begin{cases} \mathcal{R}_I^{(1)} \\ \mathcal{R}_I^{(2)} \end{cases} \quad (2.4)$$

$$\mathcal{R}_F = \begin{cases} \mathcal{R}_F^{(1)} \\ \mathcal{R}_F^{(2)} \end{cases} \quad (2.5)$$

Kommentar 2.2: Den ovan införda struktureringen med "åtskilda" register $\mathcal{R}^{(1)}$ och $\mathcal{R}^{(2)}$ kommer att underlätta den fortsatta diskussionen, men den återspeglar kanske inte riktigt vad som händer i verkligheten. Där ligger, åtminstone hittills, registren $\mathcal{R}^{(1)}$ och $\mathcal{R}^{(2)}$ "hopblandade" i det aktuella totala registret.

I verkligheten är också datahanteringsprocessen för att nå fram till de för en HINK-årgång slutliga, totala individ- och familjeregistren avsevärt mer komplicerad än vad ovanstående beskrivning ger vid handen. ●

Därmed har vi redogjort för den "hanteringsorganisatoriska" strukturen för insamlandet och lagrandet av urvalsinformationen, och vi övergår till att beskriva urvalets probabilistiska egenskaper. Vi ger dock först en avslutande kommentar för den som är bekant med HINKs panelsystem.

Med i HINK-sammanhang gängse språkbruk kan de två delurvalen för år t beskrivas på följande sätt;

Delurval 1 är urvalet i HINK t , Panel p (inklusive tilläggssurval för 18-åringar) ("gamla panelen"), (2.6)

Delurval 2 är urvalet i HINK t , Panel($p+1$) ("nya panelen").

2.2 ETT ÅRSURVALS PROBABILISTISKA EGENSKAPER

De två delurvalen i en HINK-årgång är sannolikhetsurval. I (i) och (ii) nedan specificeras den probabilistiska struktur som fortsättningsvis antages föreligga för det totala urvalet i en HINK-årgång.

- (i) Såväl Delurval 1 som Delurval 2 är stratifierade urval (med oberoende OSU:n från de olika stratana) från vuxenindividpopulationen. Stratifieringarna betecknas

$$V^I = A_1^{(\gamma)} \cup A_2^{(\gamma)} \cup \dots \cup A_{k(\gamma)}^{(\gamma)}, \quad \gamma=1,2, \quad (2.7)$$

och stratumstorlekar och stickprovsstorlekar betecknas med

$$N_1^{(\gamma)}, N_2^{(\gamma)}, \dots, N_{k(\gamma)}^{(\gamma)} \quad \text{respektive} \quad n_1^{(\gamma)}, n_2^{(\gamma)}, \dots, n_{k(\gamma)}^{(\gamma)}, \quad \gamma=1,2. \quad (2.8)$$

- (ii) Delurvalen 1 och 2 drages oberoende av varandra. (2.9)

Kommentar 2.3: I Appendix 1 ges en utförligare beskrivning av hur urvalen genereras sett ur litet mer praktisk synvinkel. Denna beskrivning möjliggör också en bedömning av realismen av antagandena (i) och (ii).

Om dessa två antaganden kan sägas att de säkert inte är exakt uppfyllda (men denna typ av förhållande är på intet sätt specifikt för HINK). För att exemplifiera, HINK använder s k negativ urvalssamordning (det gäller t ex att en individ som urväljs till Panel p inte kan urväljas till Panel ($p+1$)). Bl a härigenom kan inte antagande (ii) vara uppfyllt i helt strikt mening.

Vad vi menar med att antaga att (i) och (ii) är uppfyllda är följande. Överväganden och beräkningar (t ex om väntevärdesriktighet eller skattningar av estimatorvarianser) som görs under förutsättningarna (i) och (ii) är kanske inte till 100 procent tillämpliga på den aktuella verkligheten (dvs HINK), men de gäller med så god approximation att man kan bortse från eventuella diskrepanser (förhoppningsvis i alla fall). ●

Kommentar 2.4: Notera att i (i) så görs inga förutsättningar om att de två stratifieringarna $\{A_1^{(1)}, A_2^{(1)}, \dots, A_{k(1)}^{(1)}\}$ och $\{A_1^{(2)}, A_2^{(2)}, \dots, A_{k(2)}^{(2)}\}$ skall förhålla sig till varandra på något särskilt sätt. Bl a kan antalen strata i de två uppdelningarna, dvs $k^{(1)}$ och $k^{(2)}$, mycket väl skilja sig åt. ●

Vi passar på att här införa ett par beteckningar för framtida bruk.

$$h^{(\gamma)}(i) = \text{det stratum som individ } i \text{ tillhör vid stratifieringen (2.7), } i \in V^I, \gamma=1,2. \quad (2.10)$$

För $d \in R_F$, dvs för en familjeenhet i urvalet, sätt

$$u(d) = \text{den vuxne individ } i \text{ som urvaldes i det primära urvalet} \\ (= \text{urvalsindividen i familjen } d). \quad (2.11)$$

3 ESTIMATORER OCH ESTIMATORVARIANSER

3.1 ESTIMATION FÖR GRUPPER AV FAMILJEEHETER

Låt

$$\underline{x} = \{x_d; d \in U^F\} \text{ vara en familjevariabel,} \quad (3.1)$$

och

$$G \text{ vara en (fixerad) grupp av familjeenheter.} \quad (3.2)$$

Vårt huvudintresse här gäller estimation av $\tau(\underline{x}; G)$, $g(G)$ och $\mu(\underline{x}; G)$ enligt (1.16)-(1.18). Vi börjar med en kortfattad rekapitulation av några i detta estimationssammanhang grundläggande begrepp och formler, vilka finns utförligare behandlade i Rosén (1987).

I fortsättningen betecknar $\underline{\alpha}$ en (familje)balanserad estimatorviktning, dvs en vuxenindividvariabel $\underline{\alpha} = \{\alpha_i; i \in V^I\}$ som satisfierar (kom ihåg (1.9))

$$\alpha_i = 1, i \in E^I, \quad \alpha_i + \alpha_{m(i)} = 1, i \in D^I. \quad (3.3)$$

Vidare betecknar I_i (primär)urvalsindikatorer, dvs I_i är den slumpvariabel som indikerar om individ $i \in V^I$ kommer med i primärurvalet eller ej.

Först betraktar vi fallet när endast ett av de två delurvalen föreligger. Nedanstående estimatorer $\hat{\tau}$ och $\hat{g}$ ger då väntevärdesriktig estimation av $\tau(\underline{x}; G)$ och $g(G)$;

$$\hat{\tau}(\underline{x}; G; \underline{\alpha}) = \sum_{h=1}^k \frac{N_h}{n_h} \sum_{i \in A_h} \alpha_i \cdot x_{d(i)} \cdot 1_G(d(i)) \cdot I_i, \quad (3.4)$$

$$\hat{g}(G; \underline{\alpha}) = \sum_{h=1}^k \frac{N_h}{n_h} \sum_{i \in A_h} \alpha_i \cdot 1_G(d(i)) \cdot I_i. \quad (3.5)$$

Estimatorn $\hat{g}$ är naturligtvis bara ett specialfall av $\hat{\tau}$, nämligen det specialfall som svarar mot $\underline{x} \equiv 1$. Vi kommer därför i fortsättningen att i stort nöja oss med att betrakta $\hat{\tau}$.

Kommentar 3.1: Balanseringsvillkoret (3.3) är tillräckligt (och även nödvändigt) för att estimatorerna i (3.4) och (3.5) skall vara väntevärdesriktiga. Villkoret medför dock inte entydig fastläggning av α -vikterna för sammanboende, utan åtskillig frihet kvarstår. I det konkreta fallet måste man naturligtvis bestämma sig för vilka α -vikter man vill använda. Denna fråga diskuteras ytterligare i Appendix 2. ●

Formeln (3.4) kan ges nedanstående alternativa form, där summationen löper över det till det aktuella delurvalet svarande familjeenhetsregistret $\mathcal{R}_F$ (kom ihåg beteckningarna i (2.10) och (2.11));

$$\hat{\tau}(\underline{x}; G; \underline{\alpha}) = \sum_{d \in \mathcal{R}_F} \frac{N_h(u(d))}{n_h(u(d))} \cdot \alpha_{u(d)} \cdot x_d \cdot 1_G(d). \quad (3.6)$$

Vi går nu över till situationen där (det totala) urvalet sammansätts av två delurval enligt vad som sades i föregående avsnitt och vi fortsätter att använda beteckningskonventionen (2.3).

För vart och ett av delurvalen kan man skatta τ och g enligt ovan med $\hat{\tau}^{(1)}$ och $\hat{\tau}^{(2)}$ respektive $\hat{g}^{(1)}$ och $\hat{g}^{(2)}$. Totalskattningar fås sedan genom lämplig hopvägning av delurvalsskattningarna. Eftersom vi antar att de två delurvalen är i allt väsentligt lika (till struktur och storlek) samt oberoende, så ges den optimala hopvägningen av estimatorernas aritmetiska medelvärde. Vi använder alltså följande årsskattningar:

$$\hat{\tau}(\underline{x}; G) = \frac{1}{2} \cdot [\hat{\tau}^{(1)}(\underline{x}; G; \underline{\alpha}^{(1)}) + \hat{\tau}^{(2)}(\underline{x}; G; \underline{\alpha}^{(2)})], \quad (3.7)$$

$$\hat{g}(G) = \frac{1}{2} \cdot [\hat{g}^{(1)}(G; \underline{\alpha}^{(1)}) + \hat{g}^{(2)}(G; \underline{\alpha}^{(2)})]. \quad (3.8)$$

Vidare använder vi följande (naturliga) årsskattning för gruppmedelvärde $\hat{\mu}(\underline{x}; G)$;

$$\hat{\mu}(\underline{x}; G) = \frac{\hat{\tau}(\underline{x}; G)}{\hat{g}(G)} = \frac{\hat{\tau}(\underline{x}; G)}{\hat{\tau}(\underline{1}; G)}. \quad (3.9)$$

Ovanstående formler (3.7)-(3.9) kommer att spela en fundamental roll i fortsättningen och vi gör oss därför besväret att ge dem några alternativa framställningar. Återigen nöjer vi oss med att betrakta $\hat{\tau}$. Explicit utskrift av (3.7), leder med beaktande av (3.4) till

$$\hat{\tau}(\underline{x}; G) = \sum_{\gamma=1}^2 \sum_{h=1}^{k(\gamma)} \frac{N_h^{(\gamma)}}{n_h^{(\gamma)}} \sum_{i \in A_h^{(\gamma)}} \alpha_i^{(\gamma)} \cdot \frac{1}{2} \cdot x_{d(i)}^{(\gamma)} \cdot 1_G(d(i)) \cdot I_i^{(\gamma)}. \quad (3.10)$$

Vi skall härnäst göra en modifikation av (3.10), vilken egentligen bara har "kosmetiskt" innehåll, men som hjälper en att se hur man kan applicera existerande standarddatorprogram.

Vi säger att de A-stratan som ingår i (3.10) är löpnumrerade om de numreras efter att ha ställts i följande ordningsföljd; $A_1^{(1)}, A_2^{(1)}, \dots, A_k^{(1)}, A_1^{(2)}, A_2^{(2)}, \dots, A_k^{(2)}$.

Även vid löpnumrering använder vi h som stratumindex vilket då rör sig över $h=1, 2, \dots, k^{(1)}+k^{(2)}$, samt de tidigare bokstäverna A_h, N_h och n_h med sina naturliga tolkningar.

På löpnumrerad form kan då estimatorn i (3.10) skrivas på följande sätt

$$\hat{\tau}(\underline{x}; G) = \sum_{h=1}^{k^{(1)}+k^{(2)}} \frac{N_h}{n_h} \sum_{i \in A_h} \alpha_i \cdot \frac{1}{2} \cdot x_{d(i)} \cdot 1_G(d(i)) \cdot I_i. \quad (3.11)$$

Ytterligare en variation på temat fås genom att uttrycka (3.11) med hjälp av en summation över totala familjeenhetsregistret $\mathcal{R}_F$ (jämför med (3.6)),

$$\hat{\tau}(\underline{x}; G) = \sum_{d \in \mathcal{R}_F} \frac{N_h(u(d))}{n_h(u(d))} \cdot \alpha_{u(d)} \cdot \frac{1}{2} \cdot x_d \cdot 1_G(d). \quad (3.12)$$

Från (3.11) och antagandet (2.9) att de två delurvalen är oberoende av varandra får vi följande variansformel för τ ;

$$V(\hat{\tau}(\underline{x}; G)) = \sum_{h=1}^{k(1)+k(2)} \left(\frac{N_h}{n_h} \right)^2 V \left(\sum_{i \in A_h} \alpha_i \cdot \frac{1}{2} \cdot x_{d(i)} \cdot 1_G(d(i)) \cdot I_i \right). \quad (3.13)$$

Den för fortsättningen mest intressanta aspekten på formlerna (3.11), (3.13) och (3.9) formuleras nedan. Kom ihåg den tidigare påpekade ekvivalensen mellan individerna i det primära urvalet och posterna i familjeenhetsregistret.

LEMMA 3.1: Betrakta objekten i $\mathcal{R}^F$ som ett "oegentligt", stratifierat urval från V^I med stratan (i löpnumrering) $A_1, A_2, \dots, A_{k(1)+k(2)}$. (Det oegentliga ligger i att de nämnda A :na inte ger en ordinär stratifiering av V^I utan snarare en "dubbelstratifiering".) Beteckna motsvarande stratum- och urvalsstorlekar med $N_1, N_2, \dots, N_{k(1)+k(2)}$ respektive $n_1, n_2, \dots, n_{k(1)+k(2)}$.

- a) Grupptotalestimatet $\hat{\tau}(\underline{x}; G)$ och ett estimat av grupptotalestimators varians $V(\hat{\tau}(\underline{x}; G))$ erhålles genom att i ovan nämnda stratifierat-urval-situation applicera gängse formler för skattning och estimatorvariansskattning avseende populationstotalen svarande till vuxen-individvariabeln.

$$x_i = \frac{1}{2} \cdot \alpha_i \cdot x_{d(i)} \cdot 1_G(d(i)), \quad i \in V^I. \quad (3.14)$$

- b) Gruppstorleksestimatet $\hat{g}(G)$ och ett estimat av gruppstorleksestimators varians $V(\hat{g}(G))$ erhålles genom att i ovan nämnda stratifierat-urval-situation applicera gängse formler för skattning och estimatorvariansskattning avseende populationstotalen svarande till vuxen-individvariabeln

$$y_i = \frac{1}{2} \cdot \alpha_i \cdot 1_G(d(i)), \quad i \in V^I. \quad (3.15)$$

c) Medelvärdesestimatet $\hat{\mu}(\underline{x}; G)$ erhålles som kvotestimatet

$$\hat{\mu}(\underline{x}; G) = \frac{\hat{\tau}(\underline{x})}{\hat{\tau}(\underline{y})}, \quad (3.16)$$

och ett estimat av medelvärdesestimatorns varians $V(\hat{\mu}(\underline{x}; G))$ erhålles genom att i ovan nämnda stratifierat-urval-situation applicera gängse variansskattningsformel för kvotestimatorn i (3.16).

3.2 ESTIMATION FÖR GRUPPER AV INDIVIDER

Här låter vi

$$\underline{x} = \{x_j; j \in U^I = V^I \cup B\}, \quad (3.17)$$

beteckna en individvariabel (vilken som synes förutsätts vara definierad för såväl vuxna som barn). Vidare, i detta avsnitt betecknar

$$G \text{ en (fixerad) grupp av individer (dvs en delmängd av } U^I). \quad (3.18)$$

Helt i analogi med det tidigare intresserar vi oss för nedanstående grupptotal, gruppstorlek och gruppmedelvärde (vilka nu ligger på "individnivå"),

$$\tau(\underline{x}; G) = \sum_{j \in G} x_j, \quad g(G) = \#G, \quad \mu(\underline{x}; G) = \frac{\tau(\underline{x}; G)}{g(G)}. \quad (3.19)$$

Det är storheterna i (3.19) vi vill skatta. Den typ av urval (och urvalsinformation) som står till buds förutsättes vara enligt Avsnitt 2.1. Liksom tidigare, så betecknar $\underline{x}$ även här en familjebalanserad estimator-viktning (dvs (3.3) är uppfyllt).

Som förut så inskränker vi oss i första omgången till fallet när endast ett av de två delurvalen föreligger. Enligt Rosén (1987) erhålles en väntevärdesriktig estimator $\hat{\tau}$ av $\tau(\underline{x}; G)$ på följande sätt. Inför nedanstående hjälpvariabel $\underline{z}$ (som har kvaliteten familjevariabel),

$$z_d(\underline{x}; G) = \sum_{j \in d} x_j \cdot 1_G(j), \quad d \in U^F. \quad (3.20)$$

I ord innebär $z_d(\underline{x}; G)$ summan av x -värdena för de individer som dels ingår i familjeenheten d , dels tillhör redovisningsgruppen G .

En väntevärdesriktig estimator av $\tau(\underline{x}; G)$ ges nu av

$$\hat{\tau}(\underline{x}; G; \underline{\alpha}) = \sum_{h=1}^k \frac{N_h}{n_h} \sum_{i \in A_h} \alpha_i \cdot z_d(i)(\underline{x}; G) \cdot I_i. \quad (3.21)$$

Från (3.21) och (3.20) framgår att $\hat{\tau}$ också kan uttryckas medelst en summation över (delurvalets) individregister R_I på följande sätt,

$$\hat{\tau}(\underline{x}; G; \underline{\alpha}) = \sum_{j \in R_I} \frac{N_h(u(d(j)))}{n_h(u(d(j)))} \cdot \alpha_{u(d(j))} \cdot x_j \cdot 1_G(j). \quad (3.22)$$

Vi går nu över till situationen när (det totala) urvalet sammansätts av två delurval. Även här väger vi ihop bidragen från de två delurvalen enligt (3.7) och (3.8) och ger årsskattningen för gruppmedelvärdet enligt (3.9). Den så konstruerade årsskattningen för grupptotalen för $\underline{x}$ blir

$$\hat{\tau}(\underline{x}; G) = \sum_{\gamma=1}^2 \sum_{h=1}^{k^{(\gamma)}} \frac{N_h^{(\gamma)}}{n_h^{(\gamma)}} \sum_{i \in A_h^{(\gamma)}} \alpha_i^{(\gamma)} \cdot \frac{1}{2} \cdot z_d(i)(\underline{x}; G) \cdot I_i, \quad (3.23)$$

eller skriven på löpnumrerad form,

$$\hat{\tau}(\underline{x}; G) = \sum_{h=1}^{k^{(1)}+k^{(2)}} \frac{N_h}{n_h} \sum_{i \in A_h} \alpha_i \cdot \frac{1}{2} \cdot z_d(i)(\underline{x}; G) \cdot I_i. \quad (3.24)$$

Som specialfall har vi följande gruppstorleksestimator

$$\hat{g}(G) = \sum_{h=1}^{k(1)+k(2)} \frac{N_h}{n_h} \sum_{i \in A_h} \alpha_i \cdot \frac{1}{2} \cdot z_d(i)(1;G) \cdot I_i. \quad (3.25)$$

Som nämnts årsskattar vi gruppmedelvärdet med

$$\hat{\mu}(\underline{x};G) = \frac{\hat{\tau}(\underline{x};G)}{\hat{g}(G)}. \quad (3.26)$$

Vi kan också uttrycka estimatorerna $\hat{\tau}$ och $\hat{g}$ genom summationer över (det totala) individregistret $\mathcal{R}_I$ på följande sätt (jämför (3.22)),

$$\hat{\tau}(\underline{x};G) = \sum_{j \in \mathcal{R}_I} \frac{N_h(u(d(j)))}{n_h(u(d(j)))} \cdot \alpha_{u(d(j))} \cdot \frac{1}{2} \cdot x_j \cdot 1_G(j). \quad (3.27)$$

Genom att sätta $x_j = 1$ i (3.27) fås motsvarande formel för $\hat{g}(G)$.

Nedanstående resultat följer nu tämligen omgående från det som sades i Avsnitt 3.1 samt formlerna (3.24)-(3.26).

LEMMA 3.2: Det som sägs i Lemma 3.1 äger tillämplighet i fallet när $\underline{x}$ står för en individvariabel och G för en grupp av individer under förutsättning att (vuxenindivid)variablerna $\underline{x}$ och $\underline{y}$ definieras på följande sätt,

$$x_i = \frac{1}{2} \cdot \alpha_i \cdot \left\{ \sum_{j \in d(i)} x_j 1_G(j) \right\} = \sum_{j \in d(i)} \frac{\alpha_i x_j}{2} \cdot 1_G(j), \quad i \in V^I, \quad (3.28)$$

$$y_i = \frac{1}{2} \cdot \alpha_i \cdot \left\{ \sum_{j \in d(i)} 1_G(j) \right\} = \sum_{j \in d(i)} \frac{\alpha_i}{2} \cdot 1_G(j), \quad i \in V^I. \quad (3.29)$$

Kommentar 3.1: Man kan även här, i analogi med (3.12), framställa $\tau(\underline{x}; G)$ medelst en summation över familjeenhetsregistret $\mathcal{R}_F$ på följande sätt

$$\tau(\underline{x}; G) = \sum_{d \in \mathcal{R}_F} \frac{N_h(u(d))}{n_h(u(d))} \cdot \frac{\alpha_{u(d)}}{2} \cdot z_d(\underline{x}; G). \quad (3.30)$$

Notera dock, att i det allmänna fallet kan man inte använda (3.30) som en beräkningsformel som man låter operera på familjeenhetsregistret. Skälet till detta är att observationsvärdena på familjevariabeln

$$z_d(\underline{x}; G) = \sum_{j \in d} x_j \cdot 1_G(j) \quad (3.31)$$

normalt inte finns tillgängliga i familjeenhetsregistret, utan de måste genereras med hjälp av informationen i individregistret $\mathcal{R}_I$.

4 DATORBERÄKNINGAR MED HJÄLP AV TAB68 OCH SMED83

4.1 NÅGRA ALLMÄNNA SAKER

Vi skall här beskriva hur de i föregående avsnitt behandlade skattningarna av gruppstorlekar, grupptotaler och gruppmedelvärden kan beräknas numeriskt med hjälp av programvarorna TAB68 och SMED83. Det senare programmet ger också möjlighet att numeriskt beräkna skattningarna av estimatorernas varianser (och även andra variationsmått). Med tanke på att SMED83 är dyrare att köra än TAB68, så är det väl dock först när man förutom av punktskattningar också är intresserad av att få osäkerhetsmått som man tillgriper SMED83.

För att uppnå våra beskrivningssyften här, räcker det att betrakta framtagning av ett par mycket enkla typer av tabeller, nämligen nedanstående.

Tabell 1

$\hat{g}(G)$
$\hat{t}(\underline{x}; G)$
$\hat{\mu}(\underline{x}; G)$

Tabell 2

$\hat{g}(g)$	$\sqrt{\hat{V}(\hat{g}(G))}$
$\hat{t}(\underline{x}; G)$	$\sqrt{\hat{V}(\hat{t}(\underline{x}; G))}$
$\hat{\mu}(\underline{x}; G)$	$\sqrt{\hat{V}(\hat{\mu}(\underline{x}; G))}$

där G , som vanligt, står för en redovisningsgrupp (antingen av familjenheter eller av individer).

Tabell 2 innehåller standardavvikelser och kan endast framtagas av SMED83, medan Tabell 1 kan fås både av TAB68 och SMED83. (Tabell 1 ingår ju också som del av Tabell 2.) TAB68- och SMED83-körningar skall naturligtvis resultera i så gott som identiska Tabell 1-delar. Aningar av differenser kan dock uppstå på grund av något olika beräkningsprecisioner med TAB68 och SMED83.

SMED83 är en vidareutveckling av TAB68 och det man kan göra med hjälp av TAB68 kan man också göra med hjälp av SMED83, i de allra flesta situationer "på precis samma sätt". Vissa "inkonsistenser" finns dock mellan TAB68 och SMED83 och en av dessa blir aktuell i det här sammanhanget, nämligen följande.

Vid skattning av en grupptotal på basis av ett stratifierat urval så måste man vid en TAB68-körning "själv" multiplicera på "stratumuppräkningsfaktorn" N_h/n_h innan man begär summation över posterna.

Vid en SMED83-körning görs dessa stratumuppräkningsfaktorer automatiskt, men i gengäld måste varje post förses med uppgift om dess stratumtillhörighet, h , samt värdena på N_h och n_h .

(4.1)

I Kommentar 3.1 pekade vi på att α -vikterna måste specificeras för att de aktuella estimatorerna skall vara entydigt fastlagda, vilket naturligtvis är en förutsättning för att de skall kunna beräknas numeriskt. Härav följer att man vid numerisk beräkning av estimatorer och estimatorvarianser, någonstans i beräkningsgången måste föra in information om vilka α -vikter som skall användas och införandet av α -informationen diskuteras härnäst.

Den "inkonsistens" som vi pekade på i (4.1) föranleder att man lämpligen lägger in α -informationen i två olika former såväl i det (slutliga) familjeenhetsregistret som i det (slutliga) individregistret. Fortsättningsvis kommer vi att förutsätta att varje post i såväl familjeenhetsregistret som i individregistret innehåller, dels en storhet VIKT, dels en storhet SMEDV(ikt). Storheten VIKT är den mest ändamålsenliga att använda vid TAB68-körningar, medan storheten SMEDV är den lämpliga vid SMED83-körningar.

I familjeenhetsregistret är VIKT och SMEDV beräknade enligt nedan

$$VIKT(d) = \frac{N_{h(u(d))}}{n_{h(u(d))}} \cdot \alpha_{u(d)} \cdot \frac{1}{2}, \quad d \in R_F, \quad (4.2)$$

$$SMEDV(d) = \alpha_{u(d)} \cdot \frac{1}{2}, \quad d \in R_F. \quad (4.3)$$

I individregistret är de beräknade på följande sätt,

$$VIKT(j) = \frac{N_{h(u(d(j)))}}{n_{h(u(d(j)))}} \cdot \alpha_{u(d(j))} \cdot \frac{1}{2}, \quad j \in R_I, \quad (4.4)$$

$$SMEDV(j) = \alpha_{u(d(j))} \cdot \frac{1}{2}, \quad j \in R_I. \quad (4.5)$$

Kommentar 4.1:

- a) I ord kan sambanden mellan formlerna (4.2)-(4.5) formuleras så här. I individregistret har alla medlemmar i en familjeenhet samma värde på VIKT respektive på SMEDV, nämligen den primärt urvalda individens värden, vilka också återfinnes i familjeenhetsregistret.
- b) I äldre HINK-årgångar finns endast storheten VIKT lagrad i registren. I sådana fall kan man naturligtvis, vid en SMED83-körning enkelt beräkna SMEDV enligt formeln

$$SMEDV = \frac{n_h}{N_h} \cdot VIKT. \quad (4.6)$$

Skälet till att man numera lägger in båda vikterna i registren är beräkningsekonomiskt. I och med ökad mängd SMED83-körningar är det fördelaktigt att beräkna SMEDV en gång för alla i stället för att göra det vid varje ny SMED83-körning. ●

4.2 SKATTNINGAR FÖR GRUPPER AV FAMILJEENHETER

I detta avsnitt står $\underline{x}$ för en familjevariabel och G för en grupp av familjeenheter.

Nedan beskrivs, dels ett TAB68-program för framtagning av Tabell 1, dels ett SMED83-program för framtagning av Tabell 2. Att dessa program verkligen leder till de påstådda tabellerna följer från dels Lemma 1, dels funktionssätten för programvarorna TAB68 och SMED83 (se lämpliga manualer). För att underlätta korresponderandet med programmanualerna vill vi peka på att formeln (3.12) med hjälp av de ovan införda beteckningarna kan skrivas på nedanstående former

$$\hat{\tau}(\underline{x}; G; \underline{\alpha}) = \sum_{d \in R_F} VIKT(d) \cdot x_d \cdot 1_G(d), \quad (4.7)$$

eller

$$\hat{\tau}(\underline{x}; G; \underline{\alpha}) = \sum_{d \in R_F} \frac{N_h(d)}{n_h(d)} \cdot SMEDV(d) \cdot x_d \cdot 1_G(d). \quad (4.8)$$

TAB68-program för beräkning av Tabell 1 för en grupp av familjeenheter.

Filen (totala) familjeenhetsregistret R_F för aktuellt år förutsättes tillgänglig.

```

FIL    VAR  = (VIKT,      ,      ),
              (EX,        ,      ), (specifikation av  $\underline{x}$ )
              (BESKR,      ,      ); (för specifikation av G)

TRANSGEN    DUMMY=1;
              VEX=VIKT*EX;
              HELP=VIKT*100;

GRUPP        GRU(BESKR)=1( ); (där villkor på BESKR specificerar G)

TABELL        KOL(1)=DUMMY(1);
              RAD(1)=GRU(1), S(VIKT);
              RAD(2)=GRU(1), S(VEX);
              RAD(3)=GRU(1), S(VEX), P(4);
              RAD(4)=GRU(1), S(HELP);
              S=3;

SLUT          KÖR;

```

Som nämndes i (4.1); för att SMED83 skall kunna operera måste varje post i familjeenhetsregistret vara försedd med information av följande innebörd,

STR = den urvalda familjemedlemmens stratumnummer h , (4.9)

STORAN = storleken N_h på den urvalda familjemedlemmens stratum, (4.10)

LILLAN = storleken n_h på urvalet från det stratum som den urvalda familjemedlemmen kommer ifrån. (4.11)

SMED83-program för beräkning av Tabell 2 för en grupp av familjeenheter.

Filen (totala) familjeenhetsregistret R_F för aktuellt år förutsättes tillgänglig och förberedd för körning med SMED83, STRATEGI=1, dvs posterna är sorterade efter stratum (STR).

```

FIL    VAR  = (STR,      ,      ),
              (STORAN,   ,      ),
              (LILLAN,   ,      ),
              (SMEDV,    ,      ),
              (EX,       ,      ), (specifikation av  $\underline{x}$ )
              (BESKR,   ,      ); (för specifikation av G)

VARIANS    STRATEGI=1;
           STRATUM=STR;
           ANTRAM=STORAN;
           ANTSVAR=LILLAN;
           REDOVIS=[enligt önskemål; för Tabell 2=S]

TRANSGEN   DUMMY=1;
           SVEX=SMEDV*EX;
           HELP=SMEDV*100;

GRUPP      GRU(BESKR)=1( ); (där villkor på BESKR specificerar G)

TABELL     KOL(1)=DUMMY(1);
           KOL(2)=DUMMY(1),V;
           RAD(1)=GRU(1),S(SMEDV);
           RAD(2)=GRU(1),S(SVEX);
           RAD(3)=GRU(1),S(SVEX),P(4);
           RAD(4)=GRU(1),S(HELP);
           S=3;

SLUT       KÖR;

```

4.3 SKATTNINGAR FÖR GRUPPER AV INDIVIDER

I detta avsnitt står $\underline{x}$ för en individvariabel och G för en grupp av individer.

Från Avsnitt 3.2 framgår att det här är (totala) individregistret R_I som är den aktuella informationsfilen (se särskilt (3.27) och Kommentrar 3.1). Den grundläggande skattningsformeln är (3.27), vilken med hjälp av beteckningarna i (4.4) och (4.5) också kan skrivas på nedanstående former,

$$\hat{\tau}(\underline{x}; G; \underline{\alpha}) = \sum_{j \in R_I} \text{VIKT}(j) \cdot x_j \cdot 1_G(j), \quad (4.12)$$

eller

$$\hat{\tau}(\underline{x}; G; \underline{\alpha}) = \sum_{j \in R_I} \frac{N_h(u(d(j)))}{n_h(u(d(j)))} \cdot \text{SMEDV}(j) \cdot x_j \cdot 1_G(j). \quad (4.13)$$

Nedan beskrivs, dels ett TAB68-program för framtagning av Tabell 1, dels ett SMED83-program för framtagning av Tabell 2. Att dessa program verkligen leder till de påstådda tabellerna följer från dels Lemma 3.2, dels funktionssätten för programvarorna TAB68 och SMED83. (Se lämpliga programmanualer, vilken ifråga om SMED83 måste vara "up to date" eftersom den 1986 införda STRATEGI=3 kommer till användning här.) Formlerna (4.7) och (4.8) är behjälpliga vid korresponderandet med programmanualerna.

TAB68-program för beräkning av Tabell 2 för en grupp av individer.

Filen (totala) individregistret R_I för aktuellt år förutsättes tillgänglig.

FIL VAR = (VIKT, ,),
(EX, ,), (specifikation av $\underline{x}$)
(BESKR, ,); (för specifikation av G)

TRANSGEN DUMMY=1;
VEX=VIKT*EX;
HELP=VIKT*100;

GRUPP GRU(BESKR)=1(); (där villkor på BESKR specificerar G)

TABELL KOL(1)=DUMMY(1);
RAD(1)=GRU(1),S(VIKT);
RAD(2)=GRU(1),S(VEX);
RAD(3)=GRU(1),S(VEX),P(4);
RAD(4)=GRU(1),S(HELP);
S=3;

SLUT KÖR;

Härnäst går vi in på beräkning av Tabell 2 med hjälp av SMED83. Det som sades i (4.1) angående olikheter mellan TAB68 och SMED83 är tillämpligt även här. En konsekvens av detta är att vid SMED-beräkning så måste varje post i individregistret R_I vara försedd med den typ av information som beskrevs i (4.9)-(4.11). Härvid skall samtliga medlemmar i en familjeenhet påföras de värden på STR, STORAN och LILLAN som gäller för den primärt urvalda familjemedlemmen.

Vidare, även om variansskattningarna, här till sin logiska struktur är väsentligen desamma som i föregående avsnitt (beräkningarna utföres på "familjevärden"), så finns en ur beräkningssynvinkel sett väsentlig skillnad. I föregående avsnitt förutsattes all önskad familjeinformation finns tillgänglig på "familjenivå" (dvs i R_F). Här måste den önskade familjeinformationen först genereras genom familjevissa summationer i individregistret (se (3.28), (3.29) och Kommentar 3.1). Med sin nya STRATEGI=3 kan SMED83 utföra sådana summationer under förutsättning att individposterna i R_I är försedda med information om individens familjetillhörighet. Sådan information finns i HINKs individposter och den brukar kallas familjeenhetens UB-nummer, fortsättningsvis betecknat UBNR (ub emanerar från urvalsblankett).

SMED83-program för beräkning av Tabell 2 för en grupp av individer.

Filen (totala) individregistret R_I för aktuellt år förutsättes tillgänglig och förberedd för körning med SMED83, STRATEGI=3, dvs posterna är sorterade efter stratum (STR) och inom stratum efter familjeenhet (UBNR).

```
FIL   VAR   = (UBNR,      ,      ),
              (STR,       ,      ),
              (STORAN,    ,      ),
              (LILLAN,    ,      ),
              (SMEDV,     ,      ),
              (EX,        ,      ), (specifikation av  $\underline{x}$ )
              (BESKR,     ,      ); (för specifikation av G)
```

```
VARIANS      STRATEGI=3;
              ID=UBNR;
              STRATUM=STR;
              ANTRAM=STORAN;
              ANTSVAR=LILLAN;
              REDOVIS=[enligt önskemål; för Tabell 2=S1;
```

```
TRANSGEN      DUMMY=1;
              SVEX=SMEDV*EX;
              HELP=SMEDV*100;
```

```
GRUPP         GRU(BESKR)=1( ); (där villkor på BESKR specificerar G)
```

TABELL	KOL(1)=DUMMY(1); KOL(2)=DUMMY(1),V; RAD(1)=GRU(1),S(SMEDV); RAD(2)=GRU(1),S(SVEX); RAD(3)=GRU(1),S(SVEX),P(4); RAD(4)=GRU(1),S(HELP); S=3;
SLUT	KÖR;

APPENDIX 1 YTTERLIGARE BESKRIVNING AV HINKS KLASSIFICERINGAR, URVAL, DATAINSAMLING OCH ESTIMATION

I Avsnitten 1-4 arbetade vi under vissa förenklande antaganden och framställningen där är inte tillräckligt fullständig för att ge en beskrivning av de moment som i praktiken måste genomgå innan beräkningsalgoritmerna i Avsnitt 4 kan appliceras på lämpliga HINK-register. Med syfte att undanröja, eller åtminstone mildra, dessa ofullständigheter skall vi här ge vissa kompletterande beskrivningar, framför allt avseende

- klassificering av individer och familjeenheter samt avgränsning av undersökningspopulationerna,
- urvalsförfarandet,
- datainsamling och registerföring,
- estimation under förekomst av övertäckning och bortfall.

Vi avslutar sedan med att beskriva hur man i den "praktiska HINK-situationen" når fram till önskade skattningar genom att på lämpligt sätt korrespondera storheterna i den följande mer detaljerade beskrivningen med de (i viss mening förenklade) begreppen, formlerna och beräkningsrutinerna i Avsnitten 1-4.

A1.1 YTTERLIGARE KLASSIFICERING AV INDIVIDER OCH FAMILJEENHETER

Innan vi går in på de särskilda klassificeringsproblemen vill vi peka på att följande årsklassificeringskonvention används i HINK, liksom i de flesta andra SCB-undersökningar. Klassificeringar avseende individer och familjeenheter är normalt formulerade så att de avser förhållandena vid en bestämd tidpunkt. HINKs redovisningar avser kalenderår och under loppet av ett år kan en individ/familjeenhet undergå förändringar både en och flera gånger. Vid årsklassificering hänförs en individ/familjeenhet till det tillstånd som den tillbragte längst tid i under året ifråga.

Ytterligare en kommentar av allmän natur. Syftet nedan är inte att presentera en "fullt operationell" klassificeringskatalog, utan endast att indikera huvuddragen i de mest aktuella begreppen. Dokumentation över precisa gränsdragningar får sökas annorstädes.

Vi går nu in på de mer speciella sakerna och vi börjar med att ge (verbala) definitioner av HINKs undersökningspopulationer av individer respektive familjeenheter. De är

alla, under det aktuella året, i riket bosatta individer som
ej bor på institution, (A1.1)

respektive

alla, under det aktuella året, i riket bosatta familjeenheter
som ej är institutionsfamiljeenheter. (A1.2)

Vidare,

HINKs redovisningsgrupper förutsättes, förutom att ha den egen-
skap som specificerar gruppen ifråga, också (och utan att det
särskilt utsäges) vara delmängder av de aktuella undersöknings-
populationerna (dvs inga utrikes- eller institutionsindivider/
familjeenheter ingår i grupperna). Om vi vill trycka på att en
viss (redovisnings)grupp verkligen är en delmängd av (A1.1)/
(A1.2), så gör vi detta genom att säga att det handlar om en
intressant (redovisnings)grupp. (A1.3)

Den följande mer ingående klassificering av individer och familjeenhe-
ter, kommer bl a att precisera definitionerna (A1.1) och (A1.2). Dessa
preciseringar är dock inte huvudsyftet med begreppen ifråga, vilka i
vissa avseenden faktiskt är tämligen abstrakta och sofistikerade. Huvud-
syftet är att införa precisa instrument för den tekniska diskussionen
av HINKs urvals- och estimationsförfaranden.

I de fortsatta betraktelserna behöver vi ha tiden explicit med i bil-
den. Vi vill laborera med såväl "kontinuerligt flytande" tid, vilken vi
betecknar med τ , som "årsvis stegande" tid, vilken vi betecknar med t . I
anslutning till dessa bokstäver inför vi följande beteckningskonven-
tion. Bokstaven τ indikerar att en klassificering görs enligt de vid
(den momentana) tidpunkten τ föreliggande förhållanden medan bokstaven
 t indikerar att en klassificering görs relativt år t , enligt den tidigare
införda årsklassificeringskonventionen.

Låt

$U^I(\tau)$, $V^I(\tau)$, $V^F(\tau)$ och $U^F(\tau)$ beteckna (rikets) individpopu-
lation, vuxenindividpopulation, vuxenfamiljeenhetspopulation
respektive familjeenhetspopulation vid tidpunkten τ . I enlig-
het med den nyss införda beteckningskonventionen betecknar
 $U^I(t)$, $V^I(t)$, $V^F(t)$ och $U^F(t)$ motsvarande (årsklassificerade)
populationer under år t . (A1.4)

Låt τ_0 stå för en fixerad tidpunkt samt låt

R beteckna en (specificerad) delmängd av $U^I(\tau_0)$. (A1.5)

Vår fortsättningsvisa konkretisering av en sådan mängd R kommer att vara
att den står för en vid tidpunkt τ_0 fastlagd rampopulation.

Vi skall intressera oss för att följa utvecklingarna i tiden av de populationer av individer, vuxenindivider och familjeenheter som genereras av individer i R. Vid beskrivningen vill vi uppnå att varje individ som initialt (dvs vid tidpunkt τ_0) är med i R,

dels ligger kvar i "evärdelig" tid och för varje år t från och med " τ_0 -året" ger upphov till ett objekt i var och en av generaliserade versioner av $U^I(t)$, $V^I(t)$ och $V^F(t)$, (A1.6)

dels har en, med hänsyn till relevanta förändringar för individen ifråga, adekvat tillståndsstatus i var och en av de nämnda (generaliserade) populationerna. (A1.7)

De förändringar på individnivå som föranleder behov av nya termer är de som inträffar när en individ

- intas (långvarigt) på institution,
 - flyttar utomlands,
 - dör,
 - blir vuxen.
- (A1.8)

Innebörden i stort av de i (A1.8) nämnda förändringarna framgår väl av själva benämningarna. Med institution menas (i grova drag) sjukvårdsinrättning, kriminalvårdsanstalt, o dyl.

För att kunna tillmötesgå önskemålen i (A1.6) och (A1.7) måste vi tillåta vissa generaliserade objekt (i förhållande till populationen av i riket bosatta individer/familjeenheter) i våra populationer. Vi börjar med individer och vi noterar först att möjligheterna i (A1.8) inte är varandra uteslutande. För enkelhets skull vill vi dock att varje individ skall ha unik status i den generaliserade individpopulationen och vi låter förändringarna i (A1.8) ha företräde i ordningen, dödsfall, utomlandsflyttning, institutionsintagning. En individ som genomgått en av dessa förändringar kallar vi dödsindivid, utomlandsindivid respektive institutionsindivid. Övriga individer säges vara ordinära.

Sätt

$U^I(R;t)$ = den population av (en del generaliserade) individer som erhålles när individerna i R år- t -klassificeras med avseende på ordinär, på institution, död respektive utomlands. (A1.9)

Även på "vuxenindividnivå" använder vi de ovan nämnda typerna av generaliserade individer. Här behöver vi också ytterligare en typ av generaliserad vuxenindivid, beroende på att R tillåts omfatta barn (vilka ju inte är vuxna). Som objekt i den generaliserade vuxenindividpopulationen kallar vi ett barn för en fiktiv vuxenindivid. För entydighetens skull bestämmer vi att fiktiv har företräde framför död, utomlands- och institutions-.

Sätt

$\overset{*}{V}I(R;t)$ = den population av (en del generaliserade) vuxen-individer som erhålles när individerna i R år-t-klassificeras efter vuxenindividsklassificeringen; ordinär, på institution, död, utomlands, fiktiv. (A1.10)

Härnäst går vi till familjeenheterna och vi vill ge varje individ i

$\overset{*}{V}I(R;t)$ status som ensamstående eller sammanboende. Härvid använder vi följande konvention. Alla de "konstiga" typerna av vuxenindivider i

$\overset{*}{V}I(R;t)$, dvs döds-, utomlands-, institutions- och fiktiva vuxenindivider (= barn), räknas som ensamstående (utan barn). Motsvarande familjeenheter (vilka alltså är envuxenfamiljeenheter) kallas dödsfamiljeenheter, utomlandsfamiljeenheter, institutionsfamiljeenheter respektive fiktiva familjeenheter. (Anm: Denna konvention kan ge en i sak oegentlig beskrivning av sammanboendeförhållandena för en individ som flyttat utomlands, men konventionen fungerar för HINKs syften.) Övriga individer i

$\overset{*}{V}I(R;t)$ är ordinära vuxenindivider och för dem appliceras de sedvanliga sammanboendedefinitionerna. De vuxenfamiljeenheter som genereras av dessa övriga vuxenindivider (dvs de vuxenfamiljeenheter där åtminstone den ene vuxne kommer från R) säges vara ordinära vuxenfamiljeenheter. Notera att en ordinär vuxenfamiljeenhet omfattar endast ordinära vuxenindivider (en eller två).

Sätt

$\overset{*}{V}F(R;t)$ = den population av (en del generaliserade) vuxenfamiljeenheter som genereras av R, år-t-klassificerade enligt ovan nämnda indelning. (A1.11)

Populationen $\overset{*}{V}F(R;t)$ kan delas upp i sin ordinära, institutions-, utomlands-, döds- och fiktivdel,

$$\overset{*}{V}F(R;t) = \overset{*}{V}F(R;t)_{ORD} \cup \overset{*}{V}F(R;t)_{INST} \cup \overset{*}{V}F(R;t)_{UTL} \cup \overset{*}{V}F(R;t)_{DÖD} \cup \overset{*}{V}F(R;t)_{FIKT}. \quad (A1.12)$$

Vid jämförelse av uppdelningen (A1.12) och definitionen av undersökningspopulationen i (A1.2) framgår att det är precis den ordinära delen av

$\overset{*}{V}F(R;t)$ som är intressant (se (A1.3) i den meningen att den ingår i undersökningspopulationen, medan övriga delar är ointressanta och utgör övertäckning i förhållande till undersökningspopulationen. Mot den bakgrunden gör vi också följande uppdelning av $\overset{*}{V}F(R;t)$,

$$\check{V}^F(R;t) = V^F(R;t)_{INTR} \cup \check{V}^F(R;t)_{\check{O}VERT}, \quad (A1.13)$$

där $V^F(R;t)_{INTR} = \check{V}^F(R;t)_{ORD}$ och $\check{V}^F(R;t)_{\check{O}VERT}$ är resten.

Helt analoga uppdelningar kan göras av $\check{U}^I(R;t)$ (visavi undersökningspopulationen i (A1.1)).

Avslutningsvis inför vi ytterligare en population som genereras av R.

$$U^I(R;t) = \text{alla } \check{a}\text{-}t\text{-ordinära individer (vuxna och barn) som ingår i de familjeeenheter vars vuxendel tillhör } \check{V}^F(R;t). \quad (A1.14)$$

Notera att mängden $U^I(R;t)$ alltid är intressant i den meningen att den ingår helt i $\check{a}\text{-}t\text{-undersökningspopulationen}$ enligt (A1.1).

A1.2 OM HINKs URVALSFÖRFARANDE

Som har framhållits tidigare, så består HINKs årliga urval av två delurval vilka vi här skall kalla för det nya respektive det gamla delurvalet. I Avsnitten A1.2.1 respektive A1.2.2 beskriver vi hur dessa urval framtages enligt hittillsvarande förfarande, medan A1.2.3 behandlar en tänkbar variation av det hittillsvarande förfarandet.

A1.2.1 Det för år t nya urvalet

Nedan beskrivs huvudmomenten vid framtagandet av det för år t nya (del)urvalet.

(i) Följande urvalsram, $URAM(t)$, upprättas,

$$URAM(t) = \{RTB \text{ med aktualitet } \sim \text{ mitten av juli } \check{a}\text{r } t\} \setminus \{\text{individer födda } \check{a}\text{r } t-18 \text{ eller senare}\} \setminus \{\text{de primära urvalsindividerna i } \check{a}\text{r } (t-1):s \text{ nya urval}\}. \quad (A1.15)$$

Kommentar A1.1: Förra årets nya urval borttages ur ramen för att individer i detta urval ej skall utsättas för excessiv uppgiftslämnarbyrå. ●

(ii) Urvalsramen stratifieras i på förhand bestämda strata,

$$URAM(t) = A_1(t) \cup A_2(t) \cup \dots \cup A_k(t) \quad (A1.16)$$

För stratifieringen används information från RTBS, SCBTB samt lantbruksregistret LBR. Ett av stratana (tidigare en grupp av stratan) utgörs av de individer som är upptagna som "brukare" i LBR och detta stratum (tidigare grupp av stratan) kallas fortsättningsvis för LBR-stratumet.

I detta sammanhang görs också en populationsräkning (mer adekvat kanske, en stratumräkning) varmed avses beräkning av $N_h(t) = \#A_h(t)$, $h = 1, 2, \dots, k(t)$.

(iii) Från den så stratifierade urvalsramen dras ett stratifierat urval av (primära) urvalsindivider med på förhand fastlagda åsyftade urvalsstorlekar från de olika stratana.

För närvarande dras urvalen från de olika stratana som systematiska urval. På grund av urvalssamordning mellan HINK och DU (= deklarationsundersökningen för jordbrukare) dras urvalet från LBR-stratumet som ett suburval ur DU-urvalet. Mer detaljerad beskrivning av LBR-urvalet ges anorstädes.

(iv) I anslutning till urvalsdragningen inhämtas också registerinformation angående

- sammansättningen, enligt RTBS, av de familjeenheter som de (primärt) urvalda individerna ingår i. (Notera dock, att RTBS endast registrerar civilståndsförhållanden, ej samboförhållanden.)
- stratumtillhörighet för urvalsindividernas make/maka (enligt RTBS).

(v) Urvalsförfarandet innehåller rutiner som tillser att en RTB-familj urväljs högst en gång.

Därmed har man genererat

$$N(t) = \text{brutto(primär)urvalet för år } t: \text{s nya (del)urval (även kallat nya urvalspanelen),} \quad (A1.17)$$

varvid också framkommer bruttourvalsstorlekarna

$$n_1(t), n_2(t), \dots, n_k(t)(t) \quad (A1.18)$$

från de olika stratana. Dessa bruttourvalsstorlekar faller mycket nära de tidigare nämnda åsyftade urvalsstorlekarna. Vissa smärre diskrepanser kan dock uppstå, dels på grund av "stegningsmissar" vid de systematiska urvalen, dels på grund av eliminationen av ev "dubblettfamiljer".

I de fortsatta resonemangen arbetar vi under följande antagande,

$N(t)$ är ett stratifierat urval (med oberoende OSU-urval från de olika stratana) med urvalsstorlekar (A1.18) från den stratifierade urvalsramen (A1.16). (A1.19)

A1.2.2 Det för år t gamla urvalet

Undersökningsåret är fortfarande t. Huvudtanken bakom det urval vi skall kalla för det gamla urvalet är att det skall utgöras av förra årets nya urval, dvs av $N(t-1)$. Detta urval har dock olägenheten att år t:s 18-åringar blir oreprresenterade i urvalet. För att komma runt den komplikationen dras vid urvalstillfället för år t:s nya urval ett kompletteringsurval (till $N(t-1)$), betecknat $K(t)$, av den föreskrivna storleken $n_{k(t-1)+1}(t-1)$ ur nedanstående stratum, vilket är en delmängd av U^I vid tidpunkten för dragning av år t:s nya urval,

$$A_{k(t-1)+1}(t-1) = \{\text{individer födda år } t-18\}, \quad (A1.20)$$

vars storlek betecknas $N_{k(t-1)+1}(t-1)$.

Härefter sättes

$$G(t) = N(t-1) \quad K(t) = \text{brutto(primär)urvalet för år t:s gamla urval (även kallat gamla urvalspanelen)}. \quad (A1.21)$$

I fortsättningen arbetar vi under följande antagande.

$G(t)$ är draget från nedanstående stratifierade urvalsram, vilken är en delmängd av U^I vid tillfället för dragningen av år $(t-1)$:s nya urval,

$$\widetilde{URAM}(t-1) = A_1(t-1) \cup \dots \cup A_{k(t-1)}(t-1) \cup A_{k(t-1)+1}(t-1), \quad (A1.22)$$

där $A_{k(t-1)+1}(t-1)$ tolkas som "17-åringssstratumet i U^I vid urvalstillfället" $G(t)$ är draget med oberoende OSUn från de olika stratana och med urvalsstorlekar $n_1(t-1), \dots, n_{k(t-1)}(t-1), n_{k(t-1)+1}(t-1)$. (A1.23)

Ovanstående antagande ger en något oegentlig behandling av tilläggsstratumet $A_{k(t-1)+1}(t-1)$ såtillvida att det betraktas som ett stratum i U^I vid "fel" år, men vi gör antagandet för enkelhets skull. Vi vill dock påpeka att de beräkningsformler vi baserar på antagandet faktiskt kan ges ett stringent rättfärdigande.

A1.2.3 En tänkbar modifiering av förfarandet för dragning av nytt och gammalt (del)urval

En annan möjlighet att komma runt svårigheten med att år t:s 18-åringar är oreprresenterade i $N(t-1)$ är att göra på följande sätt.

(i) Drag det nya urvalet $N'(t)$ enligt vad som sades i A2.2.1, med den modifikation att i definitionen (A1.15) av $URAM(t)$ byts t-18 mot t-17. (A1.24)

(ii) Modifiera (A1.21) till $G(t) = N'(t-1)$. (A1.25)

Härigenom blir år t :s 18-åringar representerade i $N(t-1)$, men i gengäld kommer $N(t)$ att innehålla ett antal 17-åringar. Dessa räknas då som övertäckning i $N(t)$ (se nästa avsnitt).

A1.3 INSAMLING OCH REGISTERFÖRING AV URVALSINFORMATION

Vi skall här beskriva de stora dragen i processen för insamling och registerföring av den urvalsinformation som avser undersökningsåret t . I det sammanhanget behöver vi inte göra någon åtskillnad på primära urvalsindivider från det nya respektive det gamla urvalet och vi skall inte heller göra det. (Dock när vi kommer till estimation i nästa avsnitt vill vi återigen kunna skilja urvalen åt.) Vi väljer därför en "neutral" bokstav för att beteckna ett urval. Låt

U vara ett (primärt) urval av individer, vilket dragits år t eller tidigare. (A1.26)

För urvalet U upprättas ett register med en post för varje individ i U , vilket vi betecknar och kallar,

$R_F^{Br}(U;t) = \text{bruttofamilje(enhets)registret för urvalet } U, \text{ avseende}$
förhållandena år } t. (A1.27)

Först några ord angående behandling av övertäckningsfall. I olika skeden i informationsinsamlingsprocessen kan uppdagas att en urvalsindivid, och därmed dennes familjeenhet är ointressant (se Avsnitt A1.1) vid år- t -klassificering, dvs att urvalsindividuens familjeenhet tillhör övertäckningen relativt undersökningspopulationen (A1.2). Detta inträffar om urvalsindividens är institutionsindivid, utomlandsindivid, död eller barn. (Det sista kan ej inträffa vid den hittillsvarande typen av gamla och nya urval, men väl vid den modifierade typ som diskuterades i Avsnitt A1.2.3.) Så snart ointressanthenhet uppdagas avbryts informationsinsamlandet för den aktuella urvalsindividens och dess familjeenhet, och anledningen till ointressanthenheten noteras i registret $R_F^{Br}(U;t)$.

Nu till gången i informationsinsamlandet. Det första steget gäller information som hör ihop med/framkommer vid urvalsdragningen (diverse uppgifter ur RTB och RTBS, stratumtillhörighet för urvalsindivid och (RTB)-make/maka, m m). I fall då barn kan komma med i urvalet U uppdagas denna ointressanthenhetsorsak redan i detta första steg.

Därefter vidtar den "egentliga" datainsamlingen vilken börjar med att de (primära) urvalsindividerna tillställs en enkät, vilken bl a innehåller de vid urvalsdragningen framtagna RTB-uppgifterna om sammansättningen av urvalsindividuens familjeenhet. (Rent praktiskt sker enkätbesvarandet

sedan 1984 med hjälp av en telefonintervju.) Ett viktigt moment vid enkätbesvarandet är att granska och komplettera RTB-uppgifterna om familjens sammansättning. Denna granskning och komplettering kan medföra strykningar i urvalsindividen "RTB-lista" (därför att individer på listan inte ingår i urvalsindividen familj) samt tillskrivningar av familjemedlemmar som inte står på listan.

I samband med enkätbesvarandet, och även redan vid den ID-körning som föregår enkätomgången kan framkomma uppgifter som medför att urvalsindividen, och därmed dennes familjeenhet, är ointressant. I enlighet med den tidigare nämnda regeln avbryts då datainsamlingen och orsaken till ointressanheten införs i registret $\mathcal{R}_F^{Br}(\mathcal{U};t)$.

Vid enkätbesvarandet kan den komplikationen uppstå att urvalsindividen förklarar sig vägra deltaga i undersökningen. I sådana fall avbryts vidare datainsamling och orsaken vägran införs i $\mathcal{R}_F^{Br}(\mathcal{U};t)$. En annan komplikation som kan inträffa i enkätomgången är att urvalsindividen är oanträffbar för enkätintervjun. I sådana fall avbryts dock inte datainsamlingen. Urvalsindividen får ligga kvar i undersökningen med den familjesammansättning han/hon har enligt RTB.

På basis av resultaten i enkätomgången upprättas nedanstående register, vilket innehåller en post för varje familjemedlem i de familjeenheter som i enkätomgången inte vägrat att medverka eller klassats som ointressanta. (För ej anträffade primära urvalsindivider ingår familjemedlemmar enligt RTB-listan.)

$$\mathcal{R}_I^{Br}(\mathcal{U};t) = \text{bruttoindividregistret för urvalet } \mathcal{U}, \text{ avseende för-} \\ \text{hållandena är } t. \quad (\text{A1.28})$$

För individerna i $\mathcal{R}_I^{Br}(\mathcal{U};t)$ och familjeenheter i $\mathcal{R}_F^{Br}(\mathcal{U};t)$ fortsätter nu datainsamlandet från diverse externa källor (se Avsnitt 2). Ett huvudmoment i denna informationsinsamling är att man från skattemyndigheten inhämtar självdeklarationer (avseende beskattningsåret t), och detta görs för varje individ i $\mathcal{R}_I^{Br}(\mathcal{U};t)$. Som tidigare nämnts är även andra informationskällor aktuella, olika statliga myndigheter, kommuner m m. Införskaffade uppgifter införs i $\mathcal{R}_I^{Br}$ eller $\mathcal{R}_F^{Br}$ beroende på om de avser individ eller familjeenhet.

Härnäst några ord om klassificeringen svarande, övertäckning och bortfall.

En familjeenhet i $\mathcal{R}_F^{\text{Br}}(\mathcal{U}; t)$ betraktas som svarande om

- urvalsindividen, och därmed familjeenheten klassats som intressant,
- alla önskade uppgifter har insamlats för samtliga medlemmar i familjeenheten ifråga.

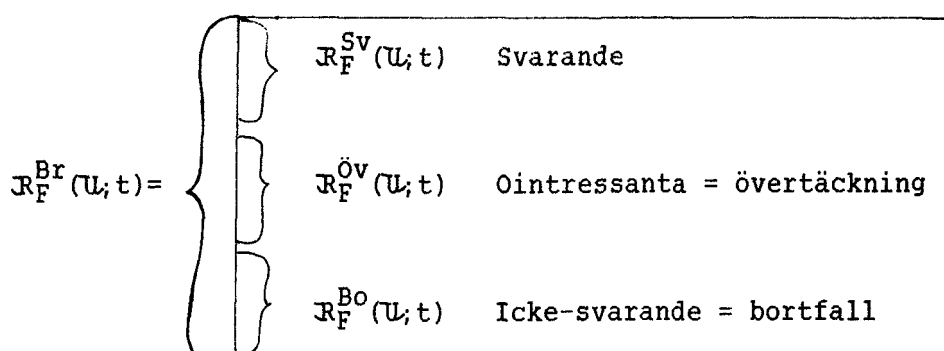
Som tidigare sagts så hänförs en familjeenhet till övertäckningen om urvalsindividen, och därmed familjeenheten har klassats som ointressant. Notera att alla familjeenheter i övertäckningen utgörs av ensamstående (utan barn).

Som icke-svarande (eller bortfallande) betraktas en familjeenhet om

- urvalsindividen vägrat att medverka i undersökningen,
- alla önskade uppgifter ej har insamlats för samtliga medlemmar i familjeenheten ifråga. (Den vanligaste anledningen är att deklara-tionsuppgifter saknas för en eller flera familjemedlemmar.)
- man av andra skäl funnit lämpligt att betrakta familjeenheten som icke-svarande.

Vid varje tidpunkt kan bruttofamiljeregistret $\mathcal{R}_F^{\text{Br}}(\mathcal{U}; t)$ delas upp i de tre disjunkta delarna; svarande, ointressanta (=övertäckning) samt icke-svarande (=bortfall) enligt nedanstående figur.

Figur 2



Bruttoindividentregistret $\mathcal{R}_I^{\text{Br}}(\mathcal{U}; t)$ kan delas upp på liknande sätt. Denna uppdelning kommer att spela roll framför allt vid estimation för grupper av individer. Som påpekades på i Avsnitt 3 så framtas även individskattningarna genom uppräknings av "familjeenhetsförhållanden". Detta är förklaringen till nedanstående definition av svarandedelen av bruttoindividentregistret, och vi nöjer oss med att specificera denna del av $\mathcal{R}_I^{\text{Br}}(\mathcal{U}; t)$,

$$\mathcal{R}_I^{\text{Sv}}(\mathcal{U}; t) = \text{de individer i } \mathcal{R}_I^{\text{Br}}(\mathcal{U}; t) \text{ som ingår i svarande familjeenheter i } \mathcal{R}_F^{\text{Br}}(\mathcal{U}; t). \quad (\text{A1.29})$$

Vid någon tidpunkt förklaras "slut på år t :s datainsamling". De då föreliggande familjeenhetsregistret och individregistret "frysas" och benämnes år t :s slutliga register.

A1.4 ESTIMATION

Vårt huvudsyfte i detta avsnitt är att nå fram till en beskrivning av hur formlerna och beräkningsalgoritmerna i Avsnitten 3 och 4 skall appliceras på observationerna i år t :s (gamla och nya) urval för att de skall leda fram till önskade skattningar för år t . Dock, innan vi ger oss in på HINKs specifika urval skall vi behandla estimationsproblematiken i en mer allmän tappning.

A1.4.1 Allmänt om skattning av grupptotaler på basis av stratifierade individurval

Som vanligt står t för aktuellt undersökningsår. Låt R beteckna en stratifierad urvalsram, vilken är en delmängd av individpopulationen $U^I(\tau_0)$ där τ_0 är en tidpunkt som vi antar ligga före utgången av år t ,

$$U^I(\tau_0) \supseteq R = A_1 \cup A_2 \cup \dots \cup A_k. \quad (\text{A1.30})$$

Storlekarna av stratana $A_1, A_2, \dots, A_k$ betecknas som vanligt med $N_1, N_2, \dots, N_k$.

När vi i fortsättningen talar om ett stratifierat urval så underförstår vi att urvalet består av oberoende OSU-urval från de olika stratana. Låt

$$U \text{ beteckna ett stratifierat urval från urvalsramen } R \text{ i (A1.30),} \\ \text{med urvalsstorlekar } n_1, n_2, \dots, n_k. \quad (\text{A1.31})$$

Notera att vi här betraktar en något allmännare situation än i Avsnitt 2.2, där rampopulationen var V^I medan den här tillåts gå utöver V^I (och den kan i princip vara hela U^I).

Som vanligt börjar vi med att betrakta skattning av en grupptotal för en grupp av familjeenheter.

Om urvalstidpunkten τ_0 ligger före år t så kan urvalet U innehålla individer som år t -år av generaliserad typ (se Avsnitt A1.1). Vi skall därför till att börja med behandla estimationsproblematiken i en generaliserad tappning. Vi använder då de beteckningar som infördes i Avsnitt A1.1, bl a betecknar $\check{V}^F(R;t)$ den i (A1.11) införda populationen. Vidare låt,

$\underline{x}^*$ beteckna en generaliserad familjevariabel, dvs en variabel som är definierad på (åtminstone) $\check{V}^F(R;t)$, (A1.32)

G^* beteckna en generaliserad redovisningsgrupp, med vilket avses en delmängd av $\check{V}^F(R;t)$. (A1.33)

Motsvarande grupptotal betecknas

$$\tau(\underline{x}^*;G^*) = \sum_{d \in \check{V}^F(R;t)} x_d^* \quad (A1.34)$$

I det fortsatta resonemanget arbetar vi (för enkelhets skull) under följande antagande;

Om den ene i ett år- t -sammanboendeförhållande ingår $\check{V}^I(R;t)$ så gör också den andre det. (A1.35)

Låt

α^* vara en generaliserad familjebalanserad estimatorviktning, varmed menas en variabel som är definierad för åtminstone alla (ev generaliserade) vuxenindivider i $\check{V}^I(R;t)$, och som satisfierar

$$\alpha_i^* = 1 \text{ om } i \in (\check{V}^I(R;t)) \text{ är ensamstående,} \quad (A1.36)$$

$$\alpha_i^* + \alpha_{m(i)}^* = 1 \text{ om } i \in \epsilon \tilde{V}^I \text{ är sammanboende.} \quad (A1.37)$$

Kommentar A1.1: Antagandet (A1.34) innebär att om i tillhör $\tilde{V}^I(R;t)$ så gör också sammanboendepartnern $m(i)$ det. Därmed är $\alpha_{m(i)}^*$ definierad, och (A1.37) är en väldefinition.

Vidare, eftersom alla icke-ordinära typer av vuxenindivider räknas som ensamstående, så är relationen (A1.37) faktiskt bara aktuell på den ordinarie delen av $\tilde{V}^I(R;t)$. Varje "ordinär" familjebalanseerad estimatorviktning $\underline{\alpha}$ kan därför också uppfattas som en generaliserad, familjebalanseerad estimatorviktning. ●

Vi antar nu att man för urvalet $\mathcal{U}$ har genomfört datainsamling och registerföring enligt det mönster som beskrevs i Avsnitt A1.3, bl a har man upprättat registret $R_F^{Br}(\mathcal{U};t)$ som vid estimationstillfället uppdelas i delregistren $R_F^{Sv}(\mathcal{U};t)$, $R_F^{Öv}(\mathcal{U};t)$ och $R_F^{Bo}(\mathcal{U};t)$.

A1.4.1.1 Estimation under antagande att inget bortfall föreligger

I första omgången skall vi arbeta under följande förenklande antagande;

Datainsamlandet för urvalet $\mathcal{U}$ har genomförts utan bortfall. (A1.38)

Ett mer tekniskt sätt att uttrycka antagandet (A1.38) på är följande,

$$R_F^{Br}(\mathcal{U};t) = R_F^{Sv}(\mathcal{U};t) \cup R_F^{Öv}(\mathcal{U};t). \quad (A1.39)$$

I detta sammanhang antar vi att den generaliserade variabeln $\underline{x}^*$ verkligen har observerats för samtliga familjeenheter i $R_F^{Br}(\mathcal{U};t)$ dvs även för familjeenheter som tillhör övertäckningen.

Vid litet eftertanke inses att formlerna (3.1)-(3.6) är tillämpliga på den här aktuella situationen under förutsättning att följande identifiering görs,

$$R_F = R_F^{Br}(\mathcal{U};t), \quad G = G^*, \quad \underline{x} = \underline{x}^*, \quad \underline{\alpha} = \underline{\alpha}^*. \quad (A1.40)$$

Formeln (3.6) ger då följande väntevärdesriktiga estimator för $\tau(\underline{x}^*; G^*)$,

$$\hat{\tau}(\underline{x}^*; G^*; \underline{\alpha}^*) = \sum_{d \in R_F^{Br}(U; t)} \frac{N_h(u(d))}{n_h(u(d))} \cdot \alpha^*_{u(d)} \cdot x^*_d \cdot 1_{G^*}(d). \quad (A1.41)$$

Låt nu

$$G \text{ vara en intressant delmängd av } V^F(t) \text{ (se Avsnitt A1.1)}. \quad (A1.42)$$

Det vi härnäst vill göra, åtminstone i princip är att applicera formeln (A1.41) med $G^*=G$. Det kan vi dock inte göra utan vidare, ty (A1.41) förutsätter att G^* är en delmängd av $\check{V}^F(R; t)$, vilket mängden G i (A1.42) inte behöver vara. Däremot är $G \cap \check{V}^F(R; t)$ en delmängd som vi kan, och skall stoppa in i (A1.41). Notera att om G är intressant och om (A1.39) är uppfylld, så gäller följande relationer

$$G \cap \check{V}^F(R; t) = G \cap V^F(R; t)_{INTR}, \quad (A1.43)$$

$$1_{R_F^{Br}}(U; t) \cdot 1_{G \cap \check{V}^F(R; t)} = 1_G \cdot 1_{R_F^{Sv}}(U; t). \quad (A1.44)$$

Genom att sätta $G^*=G \cap \check{V}^F(R; t)$ i (A1.41) och beakta (A1.43) och (A1.44) fås

$$\hat{\tau}(\underline{x}^*; G \cap \check{V}^F(R; t)_{INTR}; \underline{\alpha}^*) = \sum_{d \in R_F^{Sv}(U; t)} \frac{N_h(u(d))}{n_h(u(d))} \cdot \alpha^*_{u(d)} \cdot x^*_d \cdot 1_G(d). \quad (A1.45)$$

Vi kan göra ytterligare förenklingar i skattningsformeln. Låt $\underline{x}$ och $\underline{\alpha}$ beteckna restriktionerna av $\underline{x}^*$ respektive $\underline{\alpha}^*$ till $V^F(t)_{INTR}$, dvs vi nöjer oss med att betrakta $\underline{x}$ och $\underline{\alpha}$ för intressanta familjeenheter och vuxen-individer. Vid litet eftertanke inses att (A1.45) då kan skrivas på följande form,

$$\hat{\tau}(\underline{x}; G \cap V^F(R; t)_{INTR}; \underline{\alpha}) = \sum_{d \in R_F^{SV}(U; t)} \frac{N_h(u(d))}{n_h(u(d))} \cdot \alpha_u(d) \cdot x_d \cdot 1_G(d). \quad (A1.46)$$

Från (A1.46) framgår följande, vilket är tämligen självklart (bortsett från alla beteckningar och begrepp vi har infört). När G är en intressant redovisningsgrupp av familjeenheter, så behöver man bara ha observerat $\underline{x}$ för de intressanta familjeenheter i urvalet för att kunna skatta totalen $\tau(\underline{x}; G \cap V^F(R; t))$.

Därmed har vi uppnått följande resultat

LEMMA A1.1: Låt förutsättningar och beteckningar vara enligt ovan i Avsnitt A1.4. Då ger

$$\sum_{d \in R_F^{SV}(U; t)} \frac{N_h(u(d))}{n_h(u(d))} \cdot \alpha_u(d) \cdot x_d \cdot 1_G(d). \quad (A1.47)$$

väntevärdesriktig estimation av

$$x\text{-totalen över } G \cap V^F(R; t)_{INTR}. \quad (A1.48)$$

Vi övergår nu till skattning av totaler för grupper av individer. Med i allt väsentligt samma typ av resonemang som det som ledde fram till Lemma A1.1 kommer man fram till följande resultat. Vi utelämnar detaljer.

LEMMA A1.2: Låt $\underline{x}$ vara en individvariabel och G en intressant redovisningsgrupp av individer. Låt i övrigt förutsättningar och beteckningar vara som hittills i detta avsnitt. Bl a förutsättes (A1.35) och (A1.38) vara i kraft. Då ger

$$\sum_{j \in R_I^{Sv}(U;t)} \frac{N_h(u(d(j)))}{n_h(u(d(j)))} \cdot \alpha_{u(d(j))} \cdot x_j \cdot 1_G(j) \quad (A1.49)$$

väntevärdesriktig estimation av (för beteckning se (A1.14))

$$\underline{x}\text{-totalen över } G \cap U^I(R;t). \quad (A1.50)$$

A1.4.1.2 Estimation när bortfall föreligger

Vid framtagandet av resultaten i de två ovanstående lemmorna arbetade vi under antagandet (A1.38) att inget bortfall uppstår i datainsamlingsprocessen. I verkligheten förekommer naturligtvis bortfall, och vi skall betrakta den situationen här. Bortsett från (A1.38) så använder vi förutsättningar och beteckningar enligt det tidigare.

Här tillåter vi alltså att alla de tre komponenterna i bruttoregistret $R_F^{Br}(U;t)$, $R_F^{Sv}(U;t)$, $R_F^{\ddot{O}v}(U;t)$ och $R_F^{Bo}(U;t)$ (se Figur 2) kan vara icke-tomma. Sätt (vid aktuell estimationstidpunkt)

$$m_{hSv} = \text{antal svarande familjeenheter med urvalsindivider från stratum } h, \quad (A1.51)$$

$$m_{h\ddot{O}} = \text{antal familjeenheter med urvalsindivider från stratum } h, \text{ vilka tillhör övertäckningen (= klassats som ointressanta)}, \quad (A1.52)$$

$$m_{hBo} = \text{antal familjeenheter med urvalsindivider från stratum } h, \text{ vilka klassats som icke-svarande (= förts till bortfallet)}. \quad (A1.53)$$

Vi har alltså

$$n_h = m_{hSv} + m_{h\ddot{O}} + m_{hBo}. \quad (A1.54)$$

När man vill skatta på basis av ett observationsmaterial som utsatts för bortfall, så kommer man inte ifrån att göra antaganden om bortfallsprocessen. Vi skall arbeta under följande antagande.

BORTFALLSANTAGANDE: Huruvida en urvald familjeenhet, vid den aktuella estimationstidpunkten, är informationsgivande (=svarande om familjeenheten är intressant, identifierad som tillhörande övertäckningen om den gör det) eller ej beror av slumpen.

Olika familjeenheter är informationsgivande oberoende av varandra och stratumvis gäller att intressanta individer leder till en informationsgivande familjeenhet med en och samma sannolikhet, β_{hI} för stratum h , samt att individer i övertäckningen är informationsgivande med en och samma sannolikhet, β_{hO} för stratum h . Vi kallar β_{hI} och β_{hO} för informationsbenägenheterna.

Följande storheter, kallade informationsbenägenhetskvoterna, betraktas som kända

$$\kappa_h = \frac{\beta_{hI}}{\beta_{hO}}, \quad h=1, 2, \dots, k, \quad (A1.55)$$

medan β_{hI} , β_{hO} , N_{hi} (=antalet intressanta individer i urvalsstratum h) och $N_{hO} = N_h - N_{hI}$ betraktas som okända.

Följande resultat rättfärdigas annorstädes.

LEMMA A1.3: Låt förutsättningar och beteckningar vara enligt ovan. Om bortfall föreligger sätt

$$\tilde{n}_h = m_{hSv} + \kappa_h \cdot m_{hO}, \quad h=1, 2, \dots, k. \quad (A1.56)$$

a) Då gäller att estimatorn

$$\sum_{d \in R_F^{Sv}(U, t)} \frac{N_h(u(d))}{\tilde{n}_h(u(d))} \cdot \alpha_u(d) \cdot x_d \cdot 1_G(d) \quad (A1.57)$$

har (så gott som) samma väntevärde som estimatorn (A1.47) har under förutsättningarna i Lemma A1, dvs (A1.57) ger (så gott som) väntevärdesriktig estimation av

$$\underline{x}\text{-totalen över } G \cap V^F(R; t)_{INTR}. \quad (A1.58)$$

b) Vidare har estimatoren

$$\sum_{j \in R_I^{Sv}(\mathcal{U}; t)} \frac{N_h(u(d(j)))}{\tilde{n}_h(u(d(j)))} \cdot \alpha_{u(d(j))} \cdot x_j \cdot 1_G(j) \quad (A1.59)$$

(så gott som) samma väntevärde som estimatoren (A1.49) har under förutsättningarna i Lemma A1.2, dvs

$$\underline{x}\text{-totalen över } G \cap U^I(R; t) \quad (A1.60)$$

Kommentar A1.2: Om inget bortfall föreligger tolkar vi (A1.56) så att $\kappa=1$ (jämför med Lemmana A1.1 och A1.2). ●

Kommentar A1.3: I det konkreta estimationsfallet måste man bestämma sig för κ_h -värden som man tror vara ändamålsenliga. Vi återkommer till denna problematik i Appendix 2. ●

Kommentar A1.4: I estimatorerna (A1.57) och (A1.59) är, enligt bortfallsantagandet, nämnarna n slumpstorheter. Man kan därför diskutera om de tidigare anförda variansskattningsformlerna kan appliceras på (A1.57) och (A1.59) eller om dessa bör modifieras med hänsyn till att n är slumpmässig.

Eftersom de skattade varianserna framför allt är avsedda för inferens på basis av det föreliggande urvalet (att bilda konfidensintervall, eller bara helt allmänt få en uppfattning om estimatorns slumposäkerhet, o dyl) så bör man, enligt min mening, "resonera betingat" givet erhållna värden på m_{hSv} och $m_{h\bar{O}}$. Detta leder till att man vid variansskattning bör betrakta n som fixt och applicera tidigare variansskattningsformler. ●

A1.4.2 Sammanknnytning av beteckningar, formler och beräkningsalgoritmer i Avsnitten 1-4 med den mer fullständiga beskrivning av HINK-undersökningen som givits i detta Appendix 1

I Avsnitten 1-4 arbetade vi under förenklade antaganden. De formler och beräkningsalgoritmer som presenterades där leder dock rätt även i den "fullständiga" HINK-situationen om man korresponderar storheter på lämpligt sätt. Nedan formuleras hur korresponderandet skall, eller åtminstone kan, ske och vi ger efteråt några rättfärdiganden.

LEMMA A1.4: För såväl grupper av familjeenheter som grupper av individer gäller att resultaten i Avsnitt 3, och därmed beräkningsrutinerna i Avsnitt 4, är tillämpliga på HINKs observationsmaterial om storheterna i Avsnitten 3 och 4 väljs på det speciella sätt som anges i nedanstående Tabell 3.

Tabell 3.

Korresponderenstablå mellan "konkret" HINK-situation och storheter enligt Avsnitten 1-4 för beräkning av år-t-skattningar. (En huvudingrediens är att de år t gamla respektive nya urvalen spelar rollerna som ⁽¹⁾-urval och ⁽²⁾-urval.)

Storhet enligt Avsnitt 2-4	Väljes i den konkreta HINK-situationen som nedanstående speciella storhet (betecknad enligt detta Appendix).
$\underline{x}$	Aktuell variabel.
G	Intressant redovisningsgrupp.
$k^{(1)}$	Antalet stratan i det gamla urvalet. I det "hittillsvarande fallet" = $k(t-1)+1$.
$N_1^{(1)}, N_2^{(1)}, \dots, N_k^{(1)}$	Stratumstorlekar för det gamla urvalet (vid det för urvalet aktuella urvallstillfället).
$n_1^{(1)}, n_2^{(1)}, \dots, n_k^{(1)}$	$n_h^{(1)} = m_{hSv}^{(1)} + \kappa_h^{(1)} \cdot m_{h\bar{o}}^{(1)}, \quad h=1, 2, \dots, k^{(1)},$ där $m_{hSv}^{(1)}$, $m_{h\bar{o}}^{(1)}$ och $\kappa_h^{(1)}$ ges av (A1.51), (A1.52) och (A1.55) relativt gamla urvalet. Rekommenderat val, $\kappa_h=1$. För ytterligare diskussion se Appendix 2.2.
$R_F^{(1)}$	$R_F^{Sv}(g(t); t)$
$R_I^{(1)}$	$R_I^{Sv}(g(t); t)$

Tabell 3 (forts)

$\underline{\alpha}^{(1)}$	$\alpha_i^{(1)} = \begin{cases} 1 & \text{om } i \in E^I, \\ \frac{\frac{n_{h(i)}^{(1)}}{N_{h(i)}^{(1)}}}{\frac{n_{h(i)}^{(1)}}{N_{h(i)}^{(1)}} + \frac{n_{h(m(i))}^{(1)}}{N_{h(m(i))}^{(1)}}} & \text{om } i \in D^I. \end{cases}$ För ytterligare diskussion av valet av $\underline{\alpha}$ se Appendix 2.1.
$\begin{matrix} k^{(2)} \\ N_1^{(2)}, N_2^{(2)}, \dots, N_k^{(2)} \\ n_1^{(2)}, n_2^{(2)}, \dots, n_k^{(2)} \\ \mathcal{R}_F^{(2)}, \mathcal{R}_I^{(2)} \text{ och } \underline{\alpha}^{(2)} \end{matrix}$	Väljes helt i analogi med det ovanstående, med den modifikation att gamla urvalet byts mot nya urvalet.
$\mathcal{R}_F$	$\mathcal{R}_F^{Sv}(g(t); t) \sim \mathcal{R}_F^{Sv}(N(t); t)$
$\mathcal{R}_I$	$\mathcal{R}_I^{Sv}(g(t); t) \sim \mathcal{R}_I^{Sv}(N(t); t)$

Vi anser att påståendena i Lemma A1.4 i det väsentliga är rättfärdigade genom de tidigare överläggningarna. Vi skall dock ge en sista kommentar, varvid vi nöjer oss med att betrakta skattning av $\underline{x}$ -totalen för en grupp av familjeenheter.

Från Lemmorna A1.3 och A1.2 framgår att estimatorn $\hat{\tau}(\underline{x}; G)$ (vid korresponderande enligt Tabell 3) ger (så gott som) väntevärdesriktig estimation av storheten

$$\frac{1}{2} \cdot [\tau(\underline{x}; G \cap \widetilde{V^F(\widetilde{URAM}(t-1), t)}) + \tau(\underline{x}; G \cap V^F(URAM(t); t))] \approx \tau(\underline{x}; G). \quad (A1.61)$$

Betrakta följande relationer

$$V^F(\widetilde{URAM}(t-1); t) \approx V^F(t), \quad (A1.62)$$

$$V^F(URAM(t); t) \approx V^F(t). \quad (A1.63)$$

Om nu relationerna (A1.62) och (A1.63) vore helt exakta så vore också (A1.61) exakt. Dock (A1.62) och (A1.63) gäller endast approximativt, men approximationerna är så pass goda (särskilt den i (A1.63)) att man i praktiken kan bortse från de diskrepenser som föreligger. Den diskrepens som finns i (A1.62) bottnar huvudsakligen i att vuxenindivider som immigrerar i landet under slutet av år $t-1$ och början av år t saknas i

$\widetilde{URAM}(t-1)$ men ingår i $V^F(t)$.

APPENDIX 2 OM VAL AV ESTIMATORVIKTNING OCH INFORMATIONSBENÄGENHETSKVOTER

A2.1 OM VAL AV ESTIMATORVIKTNING

Det enda villkor som lagts på estimatorvikterna $\alpha = \{\alpha_i; i \in V^I\}$ är att de skall vara (familje)balanserade, dvs satisfiera villkoret (3.3). Så snart detta villkor är uppfyllt blir total- och medelvärdesestimatorerna väntevärdesriktiga (om vi bortser från kvotestimatorers obetydliga bias).

För ensamstående är villkoret (3.3) kategoriskt, α skall sättas till 1. För sammanboende däremot ger villkoret

$$\alpha_i + \alpha_{m(i)} = 1, \quad i \in D^I \quad (A2.1)$$

spelutrymme och olika viktuppsättningar är möjliga, och olika viktsuppsättningar har såväl diskuterats som använts i HINK.

De mest diskuterade möjligheterna utgörs av dels 1/2-viktningen,

$$\alpha_i = \alpha_{m(i)} = 1/2, \quad i \in D^I, \quad (A2.2)$$

dels (urvalssannolikhets)proportionalviktningen,

$$\alpha_i = \frac{\frac{n_{h(i)}}{N_{h(i)}}}{\frac{n_{h(i)}}{N_{h(i)}} + \frac{n_{h(m(i))}}{N_{h(m(i))}}}, \quad i \in D^I. \quad (A1.3)$$

Valet av estimatorvikter påverkar som sagt inte estimatorernas väntevärdesriktighet, men däremot deras precision. Detta val behandlas utförligt i Rosén (1987), och där visas att proportionalviktningen ger (ur praktisk synvinkel sett) optimal estimatorviktning, åtminstone vid skattning för grupper av familjeenheter.

I HINK har man också under senare år använt ett viktsystem som siktat mot just proportionalviktningen, men som av "praktiska" skäl ligger litet vid sidan av denna. Den praktiska komplikation som föranlett avvikelsen är att man har saknat uppgift om sammanboendepartners stratumtillhörighet i de fall när urvalsindivid och dennes partner ej varit gifta, eller ekvivalent i de fall då partnern tillskrivits på enkäten. Det man härvid har gjort är att man har tilldelat partners stratumtillhörighet enligt följande "oegentliga" regel.

$$\tilde{h}(m(i)) = \begin{cases} h(m(i)) & \text{när } i \text{ och } m(i) \text{ är gifta,} \\ h(i) & \text{när } i \text{ och } m(i) \text{ ej är gifta.} \end{cases} \quad \begin{matrix} i \in \text{primärurvalet,} \\ i \in D^I, \end{matrix} \quad (A2.4)$$

Därefter applicerar man följande modifierade version av proportionalviktningen (A2.3).

$$\alpha_i^{(\gamma)} = \frac{\frac{n^{(\gamma)}_{\tilde{h}^{(\gamma)}(i)}}{N^{(\gamma)}_{\tilde{h}^{(\gamma)}(i)}}}{\frac{n^{(\gamma)}_{\tilde{h}^{(\gamma)}(i)}}{N^{(\gamma)}_{\tilde{h}^{(\gamma)}(i)}} + \frac{n^{(\gamma)}_{\tilde{h}^{(\gamma)}(m(i))}}{N^{(\gamma)}_{\tilde{h}^{(\gamma)}(m(i))}}}, \quad i \in \text{primärurvalet, } i \in D^I, \gamma=1,2. \quad (A2.5)$$

Ett alternativt sätt att skriva (A2.5) på, och från vilket det framgår att α är väldefinierad på hela vuxenindividpopulationen, är följande

$$\alpha_i = \begin{cases} \frac{\frac{n_{h(i)}}{N_{h(i)}}}{\frac{n_{h(i)}}{N_{h(i)}} + \frac{n_{h(m(i))}}{N_{h(m(i))}}}, & \text{om } i \in D^I \text{ och är gift (med } m(i)), \\ \frac{1}{2} & \text{om } i \in D^I \text{ men ej är gift (med } m(i)). \end{cases} \quad (A2.6)$$

Vi avslutar med några kommentarer angående den sist betraktade viktningen.

Kommentar A2.1: För att ha något namn som skiljer viktningen i (A2.6) (eller ekvivalent i (A2.5)) från den i (A2.3) kallar vi den för gifta-proportionalviktningen. ●

Kommentar A2.2: Man kontrollerar utan svårighet att gifta-proportionalviktningen enligt (A2.6) verkligen är familjebalanserad. ●

Kommentar A2.3: Det finns inget principiellt skäl för att den brist på information om stratumtillhörighet, som är den "praktiska" anledningen till valet av gifta-proportionalviktning, verkligen måste föreligga. Den kan undanröjas genom att man kör det färdigupprättade individregistret mot den tidigare upprättade stratifierade urvalsramen. En sådan körning medför dock naturligtvis en viss kostnad (vilken dock torde vara rätt liten för en "fyr-stratifierad" HINK). ●

Kommentar A2.4: Notera att vid såväl proportionalviktning som vid gifta-proportionalviktning gäller följande. För en given urvalspanel ändrar sig värdet på n_h något från det ena estimationstillfället till det andra. Därför, även om en (sammanboende) urvalsindivid inte ändrar sitt sammanboende från det gamla till det nya urvalet, så kommer dess α -vikt att ändras något, oftast dock bara litet. Estimatorvikterna i en urvalspanel måste därför beräknas för varje estimationstillfälle. Dock, när ett register görs slutgiltigt kan också estimatorvikterna "låsas" slutgiltigt. ●

A2.2 OM ÅSÄTTANDE AV VÄRDEN PÅ INFORMATIONSBENÄGENHETSKVOTERNA

För att kunna genomföra HINK-estimation enligt vad som sägs i Appendix 1.4.2, inklusive beräkning av uppräkningsvikter, måste man åsätta värden på informationsbenägenhetskvoterna $\kappa_h^{(\gamma)}$ i relationerna (jämför (A1.56)),

$$\tilde{n}_h^{(\gamma)} = m_{hSv}^{(\gamma)} + \kappa_h^{(\gamma)} \cdot m_{hO}, \quad h=1, 2, \dots, k^{(\gamma)}, \quad \gamma=1, 2. \quad (A2.7)$$

Om inget bortfall föreligger så är, se Kommentar A1.2, de adekvata κ -valen,

$$\kappa_h^{(\gamma)} = 1 \quad \text{för alla } h, \quad \gamma=1, 2. \quad (A2.8)$$

När bortfall föreligger brukar estimation enligt (A1.57), (A2.7) och (A2.8) åberopas som (kompensationsvägning) med rak uppräknings. Inom den ram som ges av bortfallsantagandet i Appendix A1.4.1.2 svarar den raka uppräknings mot $\beta_{HI} = \beta_{H0}$, dvs att man har samma informationsbenägenhet för intressanta objekt som för objekt i övertäckningen.

Jag vill rekommendera, åtminstone intill dess att det finns empiriskt underlag för att göra annorledes, att

HINK-estimation genomförs med rak uppräknings. (A2.9)

Huvudskälet för rekommendationen (A2.9) är dock inte att jag tror att antagandet $\beta_{HI} = \beta_{H0}$ nödvändigtvis är uppfyllt, utan nedanstående två förhållanden (A2.10) och (A2.11), och att dessa två förhållanden kombinerar sig med varandra på ett sådant sätt att den raka uppräknings är en väl så god approximation som någon annan möjlighet. (Det är ju ofrånkomligen så att varje förfarande för bortfallskompensation måste betraktas som en approximation, genom att det stöder sig på något grundläggande antagande vilket, dels oftast är överifierbart, dels normalt är mer eller mindre orealistiskt särskilt vad gäller förutsättningar angående sambandet mellan benägenheten att falla bort och värdet på variabeln under intresse.)

Övertäckningsgraden i HINK är förhållandevis låg. (A2.10)

Bortfallsgraden i HINK är förhållandevis låg. (A2.11)

Vi skall härnäst ge en utförligare diskussion av rekommendationen (A2.9) där vi börjar med att ge teoretiska argument för att (A2.10) och (A2.11) "drar mot" (A2.9) och därefter ger empiriskt stöd för att (A2.10) och (A2.11) är tillräckligt väl uppfyllda för att rekommendationen (A2.9) skall vara berättigad. Låt oss dock framhålla att det för närvarande tyvärr inte finns så fylligt empiriskt underlag som man skulle önska. Det torde dock vara förhållandevis lätt att ta fram fördjupat material i avslutning till nu pågående och framtida HINK-undersökningar.

A2.2.1 Några teoretiska överväganden

I första omgången antar vi att bara ett stratum föreligger och vi utelämnar index h . Vidare antar vi att

κ är inte markant större än 1, dvs att β_I inte är markant större än β_0 . (A2.12)

Med övertäckningsgraden avser vi N_0/N , där N är antalet individer i populationen och N_0 antalet individer i övertäckningen.

LEMMA A2.1: Vi betraktar skattning av grupptotalen $\tau(y;G)$ för en variabel y över en grupp G av intressanta objekt. För enkelhets skull antar vi att $y_i \geq 0$. Vidare arbetar vi under bortfallsantagandet i Appendix A1.4.1.2 samt förutsättningen (A2.12). Notera att Lemma A1.3 anger estimatorn

$$\hat{\tau}(y; G; \kappa) = \frac{N}{m_{SV} + \kappa \cdot m_{\bar{O}}} \sum y_i \cdot \frac{1}{G}(i) \cdot I_i \quad (A2.13)$$

såsom varande en (så gott som) väntevärdesriktig estimator för $\tau(y; G)$.

Om man använder rak uppräknings, dvs estimatoren i (A2.13) med $\kappa=1$, när det sanna κ -värdet är

$$\kappa_0 = 1 - \varrho_0, \quad (A2.14)$$

så blir estimatorns bortfallsskevhets, åtminstone med god approximation för (säg) $|\varrho_0| \leq 0.2$, enligt nedan.

$$\left. \begin{array}{l} \text{Bortfallsskevhets} \\ \text{relativa storlek} \end{array} \right\} \approx |\varrho_0(1+\varrho_0)| \times \text{övertäckningsgraden}, \quad (A2.15)$$

vilket i sin tur domineras av

$$\leq \text{övertäckningsgraden} \times \text{bortfallsgraden}. \quad (A2.16)$$

Motiveringar för påståendena i lemmat: Från (A2.13) och (A2.14) får vi

$$\begin{aligned} \text{Relativa felet} &= 1 - \frac{m_{SV} + (1 - \varrho_0)m_{\bar{O}}}{m_{SV} + m_{\bar{O}}} = \varrho_0 \cdot \frac{m_{\bar{O}}}{m_{SV} + m_{\bar{O}}} \approx \\ &\approx \varrho_0 \cdot \frac{N_{\bar{O}}\beta_{\bar{O}}}{(N - N_{\bar{O}})\beta_I + N_{\bar{O}}\beta_{\bar{O}}} \approx \varrho_0 \frac{N_{\bar{O}}}{N(1 - \varrho_0) + N_{\bar{O}}} = \\ &= \varrho_0 \cdot \frac{N_{\bar{O}}/N}{(1 - \varrho_0) + \varrho_0(N_{\bar{O}}/N)} \approx \varrho_0(1 + \varrho_0) \frac{N_{\bar{O}}}{N}. \end{aligned} \quad (A2.17)$$

Därmed är (A2.15) motiverat. Vi "härleder" (A2.16) genom att ge en empirisk övre gräns för ϱ_0 och använda denna i (A2.15).

Låt n , n_I och $n_{\bar{O}}$ beteckna den totala stickprovsstorleken samt (de okända) stickprovsstorlekarna från den intressanta delen respektive övertäckningsdelen av populationen. Vidare låt, i enlighet med (A1.51)–(A1.53), m_{SV} , $m_{\bar{O}}$ och $m_{\bar{P}O}$ beteckna antalen svarande, identifierade övertäckningsobjekt respektive icke-svarande (= bortfallande). Vi har då följande relationer,

$$n = n_I + n_O = m_{SV} + m_O + m_{BO}, \quad (A2.18)$$

$$m_{SV} \leq n_I, \quad m_O \leq n_O. \quad (A2.19)$$

Medelst "naturliga" skattningar får vi

$$\hat{p}_O = 1 - \hat{\beta}_I / \hat{\beta}_O = 1 - \frac{m_{SV}/n_I}{m_O/n_O} = 1 - \frac{m_{SV}}{m_O} \cdot \frac{n_O}{n_I}. \quad (A2.20)$$

Under (A2.18) och (A2.19), då särskilt villkoret $m_O \leq n_O$, gäller att $n_O/n_I \geq m_O/(m_{SV} + m_{BO})$ som ger

$$\hat{p}_O \leq 1 - \frac{m_{SV}}{m_{SV} + m_{BO}} = \frac{m_{BO}}{m_{SV} + m_{BO}} = \frac{m_{BO}}{n} \cdot \frac{1}{1 - \frac{m_O}{n}}$$

$$\approx \frac{m_{BO}}{n} \left(1 - \frac{n_O \beta_O}{n} \right)^{-1} \approx \frac{m_{BO}}{n} \left(1 - \frac{N_O}{N} \right)^{-1}$$

$$\approx (\text{vid någorlunda liten övertäckningsgrad}) \approx \text{bortfallsgraden}. \quad (A2.21)$$

Ovanstående empiriska övre extremalgräns för p_O svarar mot antagandet att hela bortfallet kommer från den intressanta delen av populationen. Helt analogt kan man få en empirisk undre extremalgräns för p_O vilken svarar mot antagandet att hela bortfallet kommer från övertäckningen. Använd att, under (A2.18) gäller att $n_O/n_I \leq (m_O + m_{BO})/m_{SV}$. Denna typ av uppskattning leder till begränsningen att $p_O >$ -bortfallsgrad/övertäckningsgrad, vilken vi dock avför från intresselistan såsom varande inkompatibel med antagandet (A2.12). Detta antagande leder till aprioribegränsningen $p_O \leq 0$ vilken tillsammans med (A2.21) leder till $|p_O(1+p_O)| <$ bortfalls-

graden, som insatt i (A2.15) ger (A2.16). Därmed är påståendena i lemmat rättfärdigade (någorlunda i alla fall). ●

Kommentar A2.5: I ovanstående lemma nöjde vi oss med att betrakta fallet ett OSU-urval från hela den aktuella populationen. Resultaten i Lemma A2.1 kan dock på tämligen uppenbart sätt utsträckas till fallet med ett stratifierat urval, enligt nedan. Vi lämnar detaljerna åt läsaren.

Om man på basis av ett stratifierat urval skattar en grupptotal genom stratumvis rak uppräknings, när de sanna κ_h -värdena är $\kappa_{h0} = 1 - p_{h0}$ så gäller följande.

$$\left. \begin{array}{l} \text{Bortfallsskevhets} \\ \text{relativa storlek} \end{array} \right\} \approx \sum_{h=1}^k v_h \cdot e_{ho} \cdot (\text{övertäckningsgrad})_h, \quad (\text{A2.22})$$

som domineras av

$$\leq \sum_{h=1}^k v_h \cdot (\text{övertäckningsgrad})_h \cdot (\text{bortfallsgrad})_h, \quad (\text{A2.23})$$

där

$$v_h = \text{den andel av den aktuella grupptotalen som faller inom stratum } h. \quad (\text{A2.24})$$

Kommentar A2.6: Det resonemang som ledde fram till uppskattningen (A2.21) hänger nära ihop med den välkända kompensationsvägningsmetod som nedanstående specialfall av det tidigare bortfallsantagandet leder till.

SPECIALFALL AV BORTFALLSANTAGANDET: Alla urvalda objekt från övertäckningen kan identifieras (= inget bortfall i övertäckningen). Litet mer tekniskt kan antagandet formuleras så här,

$$\beta_0 = 1 \quad \text{medan } \beta_1 \text{ är okänt.} \quad (\text{A2.25})$$

Under (A2.25) har vi $\kappa = \beta_1$, och κ kan då skattas

$$\hat{\kappa} = \hat{\beta}_1 = \frac{m_{Sv}}{n - m_0}. \quad (\text{A2.26})$$

I hittillsvarande HINKar har estimationen av grupptotaler utförts enligt Lemma A1.3 med κ_h stratumvis beräknat enligt (A2.26).

Vid litet eftertanke inses att gränserna i (A2.16) och (A2.23) faktiskt anger de precisa värdena (åtminstone med god approximation) för den relativa differens som erhålles mellan skattningarna med rak uppräkningsmetod respektive med "övertäckningsidentifierad" uppräkningsmetod vid skattning av en grupptotal.

Kommentar A2.7: Den hittills förda diskussionen av bortfallsskevheter har gällt skattning av grupptotaler. Vid skattning av gruppmedelvärden har uppräkningsmetoden avsevärt mindre inflytande eftersom gruppmedelvärden skattas som kvoter av skattningar av grupptotaler (se (3.9)). Eventuella skevheter i täljare och nämnare kommer i långa stycken att ta ut varandra.

A2.2.2 Fortsatt diskussion mot bakgrund av empiriska värden

Vi skall här fortsätta diskussionen av rekommendationen (A2.9) med hjälp av empiriska data i anslutning till HINK-undersökningen. Som nämnts så finns dock inte empiriskt underlag av riktigt det slag och med den tillförlitlighet man skulle önska. I avvaktan på ett fördjupat empiriskt material angående övertäckning och bortfall blir därför de följande resonemangen något spekulativa, men de leder förhoppningsvis ändå till någorlunda korrekta storleksordningar.

Vi disponerar diskussionen på följande sätt. Först uppskattar vi kvantitativt storleksordningen av den (relativa) bortfalls-skevhets som maximalt kan föranledas av den raka uppräknings, dvs $\kappa=1$, när det sanna (men tyvärr okända) κ -värdet är någonting annat. Vi avslutar sedan med en mer kvalitativ diskussion av vad det adekvata κ -värdet skulle kunna vara i några olika faser i datainsamlingsprocessen.

I det sista avseendet finns faktiskt olikartade situationer av praktiskt intresse, vilka skiljer sig påtagligt ifråga om bortfallsgrad. I den "typiska" HINK-estimationssituationen vill man skatta på basis av HINKs slutregister. Härvid är bortfallsgraden låg, cirka 2 å 3 procent, och bortfallet består framför allt av deklara-tionsbortfall (cirka 2.5 procent) och vägrare (cirka 0.3 procent). En helt annan typ av estimerings-situation, fortsättningsvis kallad snabbstatistiksituationen, beskrivs i följande exempel.

Exempel. 1986 års "snabbstatistik" avseende inkomstförhållandena under 1985: Som ett led i den då pågående välfärdsutredningen ville man i mitten av 1986 ta fram statistik avseende inkomstförhållandena under 1985 för ett antal redovisningsgrupper (av individer), varvid socioekonomisk indelning var av intresse.

Den "reguljära" statistik som fanns tillgänglig, via HINK, IoF, m m, vid tidpunkten ifråga gick inte längre än t o m 1984. Vidare, ifråga om deklara-tionsbaserade inkomstuppgifter för 1985 var såväl HINKs som IoFs register utan material. De inkomstuppgifter avseende 1985 som fanns tillgängliga var

- Riksskatteverkets kontrolluppgiftsregister över löner och andra ersättningar som utbetalats under 1985,
- Riksförsäkringsverkets register över sjukpenningar, pensioner m m som utbetalats under 1985.

För utredningens syften var dock dessa register bristfälliga såtillvida att de inte innehåller information om variabler som t ex tillhörighet till socioekonomisk grupp. Vid den aktuella tidpunkten var dock HINKs enkätomgång avseende 1985 genomförd, och insamlad information fanns tillgänglig i HINK85:s register. Denna enkätinformation omfattar just variabler som t ex socioekonomisk grupp.

Det man gjorde för att komma vidare i syftet att belysa inkomstförhållanden under 1985 var att i HINK85:s register påföra information från de ovan nämnda kontrolluppgiftsregistret och riksförsäkringsregistret. (De senare uppgifterna skulle ju påföras HINKs register i vilket fall som helst.) Det så upprättade registret, vilket vi kallar snabbstatistikregistret, används sedan för att skatta inkomstförhållanden för 1985.

I denna typ av estimationssituation föreligger andra bortfallsförhållanden än vid estimation på basis av ett slutregister. Här räknas som bortfallande alla familjeenheter som ej besvarat enkäten, inklusive vägrarna, (cirka 10 procent). Bortfallet här är alltså betydligt högre än i slutskedet. Detta kan i förstone låta som en motsägelse, men förklaringen är följande. För urvalsfamiljer som ej besvarar enkäten (eller som besvarar den "slarvigt") hämtas diverse "enkätinformation" (t ex socioekonomisk grupp) från de senare inkommande deklarationsblanketterna. ●

Därmed är exemplet givet och vi går tillbaka till den mer allmänna diskussionen av valet av κ och vi går då in på att ge empiriska uppskattningar av den bortfallsskevhets som (maximalt) kan föranledas av den raka uppräkningsstrategin. Vårt huvudhjälpmedel skall vara begränsningen (A2.16) i dess stratifierade version enligt (A2.33). För detta behöver vi säga några ord om HINKs stratifiering. Denna lades om förhållandevis radikalt från och med HINK86, Panel 13, varvid man övergick till att använda endast 4 strata mot tidigare 16. Eftersom det näppeligen kommer att vara av intresse att djuploda i beräkningen av uppräkningsvikterna i "gamla" HINKar, nöjer vi oss med att betrakta den "framtida" situationen med 4 stratan, vilka är (för en utförligare beskrivning se t ex Rosén 1986-10-06);

LBR-stratumet
Företagar-/jordbrukarstratumet
Alderspensionärsstratumet
Löntagarsstratumet.

I nedanstående Tabell 4 redovisas stratumvis troliga (någorlunda i alla fall) värden för övertäckningsgrad, bortfallsgrad och motsvarande begränsning av den maximala relativa bortfallsskevhets vid rak uppräkningsstrategi. Vi ger längre fram en litet mer utförlig beskrivning av hur värdena framtagits.

Tabell 4

Stratumvisa uppskattningar av övertäckningsgrad, bortfallsgrad och maximal relativ bortfallsskevheter vid rak uppräknings vid estimation av en grupp total (jämför Kommentar A2.5).

	Övertäckningsgrad		Bortfallsgrad		Gräns för bortfallsskevheter = övertäckningsgrad $\times$ bortfallsgrad	
	Gammalt urval	Nytt urval	Gammalt urval	Nytt urval	Gammalt urval	Nytt urval
Estimation på basis av <u>slutregistret</u>						
LBR	1.1 %	0.6 %	5.0 %	5.0 %	$5.5 \cdot 10^{-4}$	$3.0 \cdot 10^{-4}$
Företagare	1.7 %	1.0 %	4.7 %	4.3 %	$8.0 \cdot 10^{-4}$	$4.3 \cdot 10^{-4}$
Alderspensionärer	9.9 %	4.1 %	0.6 %	0.6 %	$5.9 \cdot 10^{-4}$	$2.5 \cdot 10^{-4}$
Löntagare	1.2 %	0.4 %	1.8 %	2.4 %	$2.2 \cdot 10^{-4}$	$1.0 \cdot 10^{-4}$
Estimation på basis av <u>snabbstatistikregistret</u>						
LBR	1.1 %	0.6 %	6.3 %	4.9 %	$0.7 \cdot 10^{-3}$	$0.3 \cdot 10^{-3}$
Företagare	1.7 %	1.0 %	11.4 %	13.3 %	$1.9 \cdot 10^{-3}$	$1.3 \cdot 10^{-3}$
Alderspensionärer	9.9 %	4.1 %	5.2 %	5.6 %	$5.1 \cdot 10^{-3}$	$2.3 \cdot 10^{-3}$
Löntagare	1.2 %	0.4 %	11.9 %	12.1 %	$1.4 \cdot 10^{-3}$	$0.5 \cdot 10^{-3}$

De maximala relativa bortfallsskevheter som redovisas i Tabell 4 är små, storleksordningen är promille eller bråkdelar därav. Litenheten understyks när man relaterar skevheter till den slumpmässiga osäkerhet som estimaten är behäftade med. I nedanstående Tabell 5 redovisas det slumpmässiga felets (ungefärliga) relativa storlek vid estimation av den speciella grupptotalen = gruppstorleken.

Tabell 5

"Sann" gruppstorlek	Relativ storlek av $\pm 2 \times$ estimatorns standardavvikelse
50 000	$\pm 19.5 \%$
100 000	$\pm 13.7 \%$
200 000	$\pm 9.6 \%$
300 000	$\pm 7.7 \%$
500 000	$\pm 5.8 \%$

Vi påminner också om det som sades i Kommentar A2.7, nämligen att vid skattning av gruppmedelvärden så blir eventuell bortfallsskevhets avsevärt lägre storleksordning än vid skattning av grupptotaler.

Den slutsats vi drar av det ovanstående, åtminstone i avvaktan på ytterligare empiriskt underlag, är att det inte spelar någon egentlig roll vilken metod för bortfallskompensation man använder sig av. Att jag då rekommenderar den raka uppräkningsmetoden beror dels på det som längre fram sägs om förmodade κ -värden i snabbstatistiksituationen, dels på ett anammande av den s k Occams rakkniv (av likvärdiga möjligheter föredrar man en enklare framför en krångligare).

Härnäst beskriver vi hur värdena på övertäcknings- och bortfallsgrader i Tabell 4 framtagits, och vi börjar med övertäckningsgraderna.

Till övertäckningen förs individer som befinner sig på institution. Nedan ges grova skattningar av antalen individer som befinner sig minst halva året på de olika huvudtyperna av institution;

Tabell 6

Typ av institution	Antal individer som tillbringar mer än halva året på institutionen ifråga
Långvård	50 000
Mentalvård, utvecklingsstörda	25 000
Fängelse	4 000

I nedanstående Tabell 7 fördelas, så gott det nu går, de ovan nämnda individerna på de aktuella stratana.

För det gamla urvalet uppträder som övertäckning, dels de individer som dött mellan urvalsaktualitetstillfället (~juli år t-1) och halvårsskiftet år t, dels de individer som utvandrat under samma tidsperiod. Antalen ifråga uppskattas till

döda 90 000, utvandrade 24 000.

I Tabell 7 fördelas de ovan nämnda individerna, återigen så gott det nu går på de aktuella stratana.

Tabell 7

Övertäckningen fördelad på stratan

Stratum	Övertäckningsorsak och antal				Stratum-storlek	Övertäckningsgrad
	Inst	Döda	Utv	Totalt		
<u>För gammalt urval</u>						
LBR	500	500		1 000	88 000	1.1 %
Företagare	3 000	1 000	1 000	5 000	293 000	1.7 %
Alderspensionärer	55 000	75 000	1 500	131 500	1 334 000	9.9 %
Löntagare	20 500	13 500	21 500	55 500	4 785 000	1.2 %
<u>För nytt urval</u>						
LBR	500	0	0	500	88 000	0.6 %
Företagare	3 000	0	0	3 000	293 000	1.0 %
Alderspensionärer	55 000	0	0	55 000	1 334 000	4.1 %
Löntagare	20 500	0	0	20 500	4 785 000	0.4 %

Därmed har vi redovisat bakgrunden för de värden på övertäckningsgrad som angavs i Tabell 4. Vad gäller bortfallsgraderna har vi stött oss på resultat från de sk HIPO-beräkningarna, dels för HINK84, dels för det ovan nämnda snabbstatistikregistret. Det bör dock sägas ifrån att de nämnda materialen innehåller en del konstigheter, men i brist på annat material angående bortfall och övertäckning har de fått ligga till grund för uppskattningarna av bortfallsgraderna i Tabell 4.

Som tidigare nämnts så gäller gränserna för bortfallsskevheten i Tabell 4 oavsett vilka de sanna κ -värdena är (åtminstone under förutsättningen (A2.12)). De möjliga skevhetera är som sagt små, och därför behöver man egentligen inte bekymra sig särskilt mycket om vad de "korrekta" κ -värdena är. Nedan görs i alla fall några spekulationer.

Omständigheter som drar åt hållet $\kappa < 1$ (dvs $\beta_I < \beta_O$)

- Döda kan inte vägra att delta i undersökningen.
- I slutregisterfallet. Här beror stordelen av bortfallet på deklara-
tionsbortfall. Notera dock att även individer i övertäckningen kan
hamna i bortfallet genom att de utgör deklara-
tionsbortfall, om de
först har utgjort enkätbortfall.
 - a) Samtliga intressanta individer "ligger under risk" att hamna i dek-
larationsbortfallet. I övertäckningen ligger endast de ej "övertäck-
ningsidentifierade" under sådan risk.
 - b) Ju färre individer i en familjeenhet, ju mindre risk för deklara-
tionsbortfall. Individer i övertäckningen är huvudsakligen ensamstå-
ende.
 - c) En anledning till deklara-
tionsbortfall är att taxeringsnämndens be-
slut har överklagats. Denna bortfallsanledning torde vara vanligare
för intressanta individer än för individer i övertäckningen (?).

Omständigheter som drar åt hållet $\kappa > 1$ (dvs $\beta_I > \beta_O$)

- I slutregisterfallet. En annan anledning till deklara-
tionsbortfall är
att individer sköntaxerats. Denna bortfallsorsak torde vara vanligare i
den oidentifierade delen av övertäckningen än för övriga individer (?).
- I snabbstatistiksituationen. Här är bortfall väsentligen ekvivalent med
enkätbortfall. Individer i övertäckningen torde ha överbenägenhet att
bli enkätbortfall (?).



Skulle jag våga mig på en gissning så blir den att i slutregisterfallet
är κ -värdena < 1 medan de i snabbstatistikfallet är > 1 .

Tidigare nummer av Promemorior från U/STM:

NR

- 1 Bayesianska idéer vid planeringen av sample surveys. Lars Lyberg (1978-11-01)
- 2 Litteraturförteckning över artiklar om kontingenstabeller. Anders Andersson (1978-11-07)
- 3 En presentation av Box-Jenkins metod för analys och prognos av tidsserier. Åke Holmén (1979-12-20)
- 4 Handledning i AID-analys. Anders Norberg (1980-10-22)
- 5 Utredning angående statistisk analysverksamhet vid SCB: Slutrapport. P/STM, Analysprojektet (1980-10-31)
- 6 Metoder för evalvering av noggrannheten i SCBs statistik. En översikt. Jörgen Dalén (1981-03-02)
- 7 Effektiva strategier för estimation av förändringar och nivåer vid föränderlig population. Gösta Forsman och Tomas Garås (1982-11-01)
- 8 How large must the sample size be? Nominal confidence levels versus actual coverage probabilities in simple random sampling. Jörgen Dalén (1983-02-14)
- 9 Regression analysis and ratio analysis for domains. A randomization theory approach. Eva Elvers, Carl Erik Särndal, Jan Wretman och Göran Örnberg (1983-06-20)
- 10 Current survey research at Statistics Sweden. Lars Lyberg, Bengt Swensson och Jan Håkan Wretman (1983-09-01)
- 11 Utjämningsmetoder vid nivåkorrigering av tidsserier med tillämpning på nationalräkenskapsdata. Lars-Otto Sjöberg (1984-01-11)
- 12 Regressionsanalys för f d statistikstuderande. Harry Lütjohann (1984-02-01)
- 13 Estimating Gini and Entropy inequality parameters. Fredrik Nygård och Arne Sandström (1985-01-09)
- 14 Income inequality measures based on sample surveys. Fredrik Nygård och Arne Sandström (1985-05-20)
- 15 Granskning och evalvering av surveymodeller, tiden före 1960. Gösta Forsman (1985-05-30)
- 16 Variance estimators of the Gini coefficient - simple random sampling. Arne Sandström, Jan Wretman och Bertil Waldén (Memo, Februari 1985)
- 17 Variance estimators of the Gini coefficient - probability sampling. Arne Sandström, Jan Wretman och Bertil Waldén (1985-07-05)
- 18 Reconciling tables and margins using least-squares. Harry Lütjohann (1985-08-01)

- 19 Ersättnings och uppgiftslämnarbördans betydelse för kvaliteten i undersökningarna om hushållens utgifter. Håkan L. Lindström (1985-11-29)
- 20 A general view of estimation for two phases of selection. Carl-Erik Särndal och Bengt Swensson (1985-12-05)
- 21 On the use of automated coding at Statistics Sweden. Lars Lyberg (1986-01-16)
- 22 Quality Control of Coding Operations at Statistics Sweden. Lars Lyberg (1986-03-20)
- 23 A General View of Nonresponse Bias in Some Sample Surveys of the Swedish Population. Håkan L Lindström (1986-05-16)
- 24 Nonresponse rates in 1970 - 1985 in surveys of Individuals and Households. Håkan L. Lindström och Pat Dean (1986-06-06)
- 25 Two Evaluation Studies of Small Area Estimation Methods: The Case of Estimating Population Characteristics in Swedish Municipalities for the Intercensal Period. Sixten Lundström (1986-10-14)
- 26 A Survey Practitioner's Notion of Nonresponse. Richard Platek (1986-10-20)
- 27 Factors to be Considered in Developing a Reinterview Program and Interviewer Debriefings at SCB. Dawn D. Nelson (1986-10-24)
- 28 Utredning kring bortfallet i samband med den s k Metropolit-debatten. Sixten Lundström (1986-11-10)
- 29 Om HINK-undersökningens design- och allokeringsproblematik. Bengt Rosén (1986-10-06)

Kvarvarande exemplar av ovanstående promemorior kan rekvireras från
Elseliv Lindfors, U/STM, SCB, 115 81 Stockholm, eller per telefon
08 7834178