R&D Report
Research – Methods – Development
2003:2

Statistiska centralbyrån Statistics Sweden

Estimation vid förekomst av bortfall och rambrister i undersökningen *Gymnasieungdomars studieintresse*

Henrik Gustafsson
Sixten Lundström

INLEDNING

TILL

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

Föregångare:

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

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

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

Efterföljare:

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

R & D Report 2003:2. Estimation vid förekomst av bortfall och rambrister i undersökningen Gymnasieungdomars studieintresse / Henrik Gustafsson, Sixten Lundström.
Digitalt skapad fil, anpassad efter de digitaliserade delarna i serien. Statistiska centralbyrån (SCB) 2016.

urn:nbn:se:scb-2003-X101OP0302

R&D Report **2003:2**
Research – Methods – Development

Estimation vid förekomst av bortfall och rambrister i undersökningen *Gymnasieungdomars studieintresse*

Henrik Gustafsson
Sixten Lundström

R&D Report 2003:2

Research - Methods - Development

Estimation in Presence of Nonresponse and Frame Imperfections in the survey

Transition from Upper Secondary School to Higher Education

Från trycket Oktober 2003
Producent Statistiska centralbyrån, *Statistics Sweden,* metodenheten
Box 24300, *SE*-104 51 STOCKHOLM

Förfrågningar Henrik Gustafsson
henrik.gustafsson@scb.se
Telefon 019 - 17 65 36

Sixten Lundström
sixten.lundstrom@scb.se
Telefon 019 - 17 64 96

© 2003, Statistiska centralbyrån
ISSN 0283-8680

Printed in Sweden
SCB-Tryck, Örebro 2003

Abstract

This report describes a work carried out within the Nonresponse project (*Bortfallsprojektet*). The aim of the work is to demonstrate how errors, caused by sample, nonresponse and frame imperfections, can be handled in the estimation stage.

In order to convince staff at Statistics Sweden that effective methods are available we wanted to show a concrete example. For that reason we carried out the work within a "real" survey, namely "The transition from upper secondary school to higher education" (*Övergång gymnasieskola – högskola*).

Both over- and undercoverage were present, since the target population consisted of pupils in the third (final) year in the upper secondary school and the sampling frame consisted of pupils in the second year. Moreover, about 25.5 percent of the pupils did not respond in the data collection stage.

The work is based on Statistics Sweden's handbook Estimation in the Presence of Nonresponse and Frame Imperfections. The handbook describes two main methods, namely weighting and imputation. In this case we used the former method. The weights were derived by use of the general technique calibration of weights.

This report shows that all three types of errors are reduced. The main explanation for that is that we used strong auxiliary information in the calibration stage. When the estimation procedure started the register on pupils in the third (final) year was (almost) finished, so the coverage errors could be almost eliminated. Moreover, to this register we also added many auxiliary variables from other registers in order to reduce the sampling error and the nonresponse error. Some of the variables were characteristics of their parents.

Innehållsförteckning

1. Syfte

Problemet med bortfall är fortsatt stort i statistiska undersökningar. Särskilt gäller det individ- och hushållsundersökningar, där bortfallsandelen t.o.m. uppvisar en stigande trend. I ett försök att förbättra situationen har Bortfallsprojektet startats.

Projektet syftar till att utveckla och implementera effektiva metoder för att *reducera bortfallsandelarna* i SCB:s undersökningar. Likaså ska projektet utveckla och implementera effektiva metoder för att *reducera bortfallskevheten* när väl bortfall uppkommit. Föreliggande studie ingår i den andra delen.

Två CBM har skrivits inom området, där det första, *Minska bortfallet*, är en handledning i hur man minskar bortfallsandelarna och det andra, *Estimation in the Presence of Nonresponse and Frame Imperfections*, behandlar hur man reducerar bortfallsbiasen. Det senare CBMet behandlar de två huvudåtgärderna viktning och imputering. Viktningen föreslås utföras med den generella kalibreringstekniken.

Bortfallsprojektet menar att det är viktigt att demonstrera hur goda metoder kan användas. Arbetet bakom denna rapport baserar sig (nästan) enbart på beskrivningen i nämnda CBM och beräkningarna är utförda med CLAN97. Arbetet avser en särskild undersökning, men ”verktygen” passar de flesta undersökningarna.

Vi beskriver hur kalibreringstekniken kan användas i undersökningen ”Gymnasieungdomars studieintresse - läsåret 2002/2003”. Vi vill med den försöka reducera felen som beror på urvalet, bortfallet och täckningsbrister. Dessutom vi vill vi minska problemet med bristande konsistens.

Vi ville att arbetet skulle utföras i sådan takt att den nya punktestimatorn och motsvarande variansestimator kunde användas i årets undersökning. Så har också skett. Arbetet, beskrivet i avsnitten 4-5, gjordes under stark tidspress, medan resten av arbetet kunde utföras under lugnare omständigheter.

Innan vi går in på arbetet med denna specifika undersökning introducerar vi de beteckningar som behövs för att beskriva en generell urvalsundersökning som, likt ”vår” undersökning, lider av såväl bortfalls- som täckningsproblem.

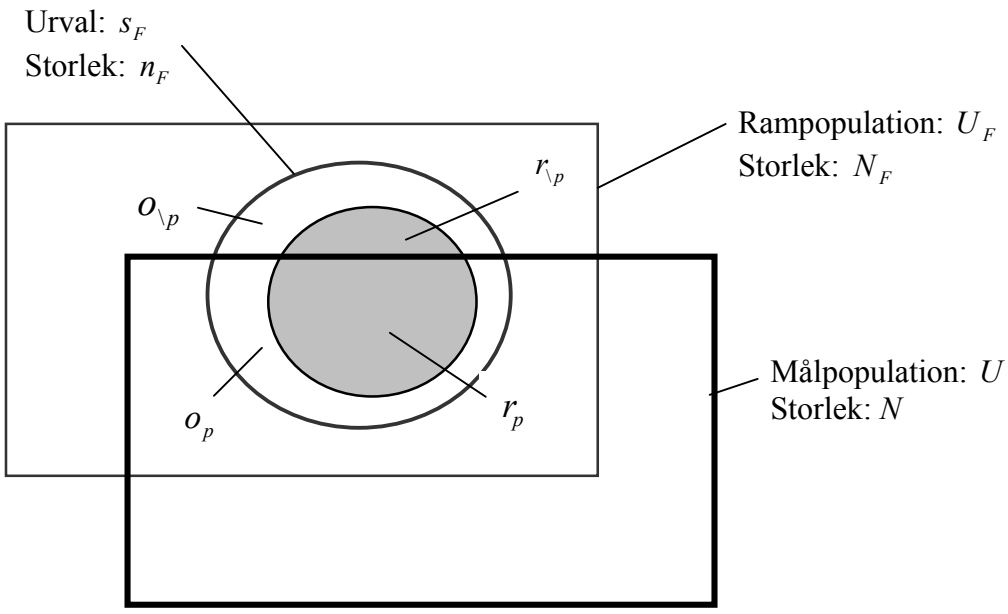
2. Beteckningar

Antag att vi vill skatta totalen

$$Y_U = \sum_U y_k \qquad (2.1)$$

där y_k är värdet på målvariabeln, y, för objekt k i målpopulationen $U = \{1,...,k,...,N\}$. Vi antar också att vi har både över- och undertäckning, d.v.s. att målpopulationen inte helt överensstämmer med rampopulationen. Situationen illustreras i figur 2.1.

Låt s_F beteckna urvalet av storleken n_F draget från rampopulationen U_F (av storleken N_F) med sannolikheten $p(s_F)$. Inklusionssannolikheterna, kända för alla $k \in U_F$, är då $\pi_k = \sum_{s_F \ni k} p(s_F)$ och *designvikten* för objekt k är $d_k = 1/\pi_k$.



Figur 2.1

Figur 2.1 illustrerar en situation där både undertäckning och övertäckning förekommer. Med *övertäckningsmängd* menas $U_F - (U \cap U_F)$ och med *undertäckningsmängd* $U - (U \cap U_F)$. Mängden av objekt som svarar *och*

tillhör målpopulationen benämns r_p och dess storlek m_p. Det gäller att $r_p \subseteq s_F$. Indexet p används här för att indikera objekt som är ”kvar” i målpopulationen och indexet $\setminus p$ avser objekt som tillhör övertäckningen. Vi använder notationen $r_{\setminus p}$ för mängden objekt som svarar *och* tillhör övertäckningen. Storleken på $r_{\setminus p}$ benämner vi $m_{\setminus p}$.

Bortfallsmängden är $o_p \cup o_{\setminus p}$, där o_p är den del som tillhör målpopulationen och $o_{\setminus p}$ den del som tillhör övertäckningen. Urvalet s_F är unionen av de fyra disjunkta mängderna r_p, $r_{\setminus p}$, o_p and $o_{\setminus p}$.

Vi antar att varje svarsobjekt, $k \in r_p \cup r_{\setminus p}$, kan hänföras till antingen r_p eller $r_{\setminus p}$. Detta är vanligtvis enkelt. Mycket mer problematisk är det att i praktiken dela upp bortfallet i dess två delmängder, o_p and $o_{\setminus p}$.

Förutom totaler för hela målpopulationen skattas vanligtvis också totaler för redovisningsgrupper. Låt oss benämna redovisningsgrupper med $U_1,\ldots,U_d,\ldots,U_D$ och deras storlek $N_1,\ldots,N_d,\ldots,N_D$. Antag att vi vill skatta totalen för variabeln y för var och en av redovisningsgrupperna. Målet för estimationen är då de D storheterna $Y_1,\ldots,Y_d,\ldots,Y_D$, där $Y_d = \sum_U y_{dk}$, $d = 1,\ldots,D$, med

$$y_{dk} = \begin{cases} y_k & \text{för} \quad k \in U_d \\ 0 & \text{för} \quad k \notin U_d \end{cases}$$

Även andra typer av parametrar är efterfrågade, som t.ex. medelvärdet i redovisningsgrupp d, $\bar{Y}_d = \dfrac{Y_d}{N_d}$. Parametrarna utgör i regel funktioner av totaler och vid estimationen skattas varje total för sig.

3. Beskrivning av undersökningen ”Gymnasieungdomars studieintresse”

3.1. Inledning

Syftet med undersökningen är att belysa hur stort intresset för att börja läsa på högskolan är bland gymnasieelever, vilken inriktning på högskolestudierna som är mest lockande och hur intresset för högskoleutbildning förändras över tiden.

Undersökningen genomförs en gång per år. Uppgifterna samlas in fr.o.m. första veckan i oktober t.o.m. sista veckan i november via postenkäter till elever i årskurs tre i gymnasieskolan.

Datainsamlingen görs via postenkät med tre skriftliga påminnelser. Svarsandelen för undersökningen avseende läsåret 2002/2003 är 74.5 %.

I avsnitten 3.2-3.6 beskriver vi undersökningen kortfattat. Mer information finns i UF 36 SM 0301.

3.2. Population och urvalsdesign

Rampopulationen, U_F, hämtas från Skolverkets elevregister. Eftersom registret över elever i årskurs tre för aktuellt läsår inte är färdigt när urvalet till undersökningen ska dras, måste föregående läsårs register över elever i årskurs två användas som ram.

I undersökningen avseende läsåret 2002/2003 stratifierades rampopulationen efter riksområde, programgrupp (studieförberedande/yrkesförberedande) samt kön. Urvalet allokerades med syfte att erhålla bra skattningar för viktiga redovisningsgrupper.

Dessutom drogs på uppdrag tilläggsurval. Data från tilläggsurvalen används också i skattningarna, vilket totalt ger urvalet s_F av storleken $n_F = 9023$. Det slutliga urvalet behandlas i den ordinarie estimationen som stratifierat obundet slumpmässigt urval. Tilläggsurvalen allokerades på ett sådant sätt att i många strata ingår samtliga elever.

3.3. Parametertyper och redovisningsgrupper

De flesta parametrarna utgörs av procentuella andelen elever inom en redovisningsgrupp som har en viss egenskap. Egenskapen kan t.ex. vara att planera att börja läsa på universitet eller högskola inom tre år eller att läsa inom ett visst ämnesområde eller att läsa vid ett visst universitet eller en viss högskola.

Redovisningsgrupperna avgränsas antingen med hjälp av en registervariabel (från urvalsramen) eller utifrån svaret på en fråga. Vanliga registervariabler som används i detta syfte är program, kön och riksområde. En vanlig redovisningsgrupp av den andra typen är ”de som planerar att börja läsa på universitet eller högskola inom tre år” och bland dessa kan egenskapen vara t.ex. ämnesområde.

Förutom procenttal skattas också antalet elever med olika egenskaper. I SMet utgör dessa totaler vanligtvis storleken på redovisningsgrupper. I Sveriges Statistiska Databaser (SSD) skattas enbart totaler (antal).

3.4. Redovisning av osäkerhetsmått

I SMet presenteras 95-procentiga konfidensintervall för de flesta skattningarna av procenttal, men däremot inte för skattningar av totaler. I SSD finns inga konfidensintervall.

Några andra mått på osäkerheten anges inte, men däremot kommenteras andra fel t.ex. täckningsbrister.

3.5. Teknisk beskrivning av undersökningen

I undersökningen utgörs övertäckningen $U_F - \left(U \cap U_F\right)$ av de elever som avbrutit eller gjort uppehåll i sina studier under eller efter årskurs två och undertäckningen $U - \left(U \cap U_F\right)$ av de elever som inte fanns med i årskurs-två-registret, t.ex. på grund av studieuppehåll men som innevarande läsår går i årskurs tre.

Rampopulationen U_F är indelad i strata, U_{Fh}, $h = 1,...,H$, och urvalet s_{Fh} dras från U_{Fh} med obundet slumpmässigt urval. (När det är nödvändigt att identifiera ett stratum lägger vi till index h till beteckningarna angivna i figur 2.1.) Designvikten är då $d_k = N_{Fh} / n_{Fh}$ for $k \in U_{Fh}$.

I undersökningen skattas Y_U med följande estimator:

$$\hat{Y}_{HT} = \sum_{h=1}^{H} \frac{N_{Fh}}{m_{ph} + m_{\setminus ph}} \sum_{r_{ph}} y_k \qquad (3.1)$$

Anm.: Vi kallar denna estimator för HT-estimator, en förkortning av Horvitz-Thompson-estimator, även om den inte riktigt är en sådan estimator.

Låt oss diskutera olika överväganden kring svarssannolikheten och täckningsbrister för att förstå rimligheten i att estimator (3.1) används. Vi antar två olika fall, nämligen där

(i) svarssannolikheten är lika stor (inom varje stratum) för elever som tillhör målpopulationen som för de som tillhör övertäckningen

resp.

(ii) svarssannolikheten är lika med noll för de som tillhör övertäckningen.

I fall (i) utgör svarsmängden inom varje stratum ett (approximativt) obundet slumpmässigt urval från U_{Fh}. Då är det lätt att inse att vi skattar totalen i ”domänen” $U \cap U_F$ (se figur 2.1). Alltså får man i detta fall en (förväntad) underskattning av totalen för målpopulationen Y_U.

Det är troligt att svarssannolikheten är lägre bland de personer som tillhör övertäckningen än bland andra. Många av dessa elever kan vara på utlandsstudier och inte nåbara och andra kan tycka att frågorna är irrelevanta. Låt oss anta att det går så långt att ingen svarar i den gruppen (fall (ii)), vilket innebär att $m_{\setminus ph} = 0$ (jmf. estimator (4.3)). Det är lätt att se att om vi utnyttjar den estimatorn för att skatta storleken på populationen får vi $\hat{Y}_{HT} = N_F$. Om över- och undertäckningsmängderna är lika stora, d.v.s. $N = N_F$, så erhålles rätt ”dimension” på vikterna. I denna undersökning är det troligt att övertäckningen är betydligt större än undertäckningen och därmed skulle en kraftig överskattning erhållas.

Vi ser alltså att fall (i) ger en underskattning och fall (ii) en överskattning. Svarssannolikheten i övertäckningen bör alltså vara mindre än bland övriga, men inte noll för att viktningen i estimator (3.1) ska fungera.

3.6. Brister i statistiken

Skattningarna i undersökningen är behäftade med fel som består av flera komponenter. De felkomponenter vi ska studera är

(i) bortfallsfelet (bortfallsbias),

(ii) urvalsfelet (varians) samt

(iii) täckningsproblem.

Dessutom vill vi också studera möjligheten att reducera

(iv) bristande konsistens mellan skattningar och registerdata.

Bortfallsbias
Bortfall har en snedvridande effekt på skattningarna om de som svarar har andra egenskaper än de som inte svarar. Detta får vi aldrig veta med säkerhet, men däremot kan vi för registervariabler jämföra gruppen svarande med bortfallet. Om registervariablerna är (starkt) korrelerade med målvariablerna ger jämförelsen en indikation på bortfallsfelet.

I denna undersökning indikerar bortfallsanalysen (avsnitt 4.2.3) att bortfallet orsakar bias.

Varians
I en statistisk undersökning utgör stratifieringen och allokeringen av urvalet på strata, liksom urvalsstorleken (egentligen antalet svarande) viktiga bestämningsfaktorer för variansen. Skattningar som avser hela populationen har i regel liten varians, men däremot kan variansen vara stor i vissa redovisningsgrupper.

I denna undersökning utgör riksområde, program och kön viktiga redovisningsgrupper. Stratifieringen (i grundurvalet) och allokeringen har gjorts med syfte att få så liten varians som möjligt i dessa redovisningsgrupper. Det är dock inte möjligt att ha alltför många strata när urvalet inte är större. Därför fick man begränsa sig till programgrupper i stället för program och reducerade därmed antalet strata från 272 till 32. Därför har man större varians för vissa program än vad som är önskvärt. Tilläggsurvalen syftade till att öka precisionen för vissa specifika grupper.

En minskning av variansen är naturligtvis välkommet för de skattningar som nu publiceras. En minskning kan också göra det möjligt att publicera skattningar, som man tidigare fått utelämna p.g.a. dålig kvalitet.

Täckningsproblem
I inledningen av avsnitt 3.5 förklarar vi varför både övertäckning och undertäckning förekommer, d.v.s. att båda mängderna $U_F - \left(U \cap U_F\right)$ och $U - \left(U \cap U_F\right)$ innehåller elever. Vi visar senare i rapporten att dessa brister i stort sett elimineras genom att utnyttja aktuell hjälpinformation.

Bristande konsistens mellan skattningar och registerdata
Resultaten från undersökningen publiceras en kort tid innan resultaten från årskurs-tre-registret publiceras. Många av uppgifterna ska överensstämma men gör det inte. Vid skattningar av, t.ex. antalet elever i olika program, erhålles olika resultat. Denna bristande konsistens är naturligtvis störande.

I föreliggande arbete visar vi hur dessa brister kan reduceras genom att utnyttja hjälpinformation i en kalibreringsestimator.

4. Kalibrering av vikter

4.1. Kalibreringsestimatorn

Både variansen och bortfallsbiasen kan reduceras genom att utnyttja ”stark” hjälpinformation i en kalibreringsestimator. Även täckningsbrister kan reduceras om det finns hjälpinformation som väl speglar målpopulationen (Lundström and Särndal, 2001, Ch. 11).

Viss hjälpinformation utnyttjas vanligtvis även före estimationen, t.ex. för bildande av stratifierade urvalsdesigner. I studerade undersökning används, som tidigare påpekats, stratifieringsvariablerna riksområde, kön och programgrupp. Det kan dock finnas ytterligare hjälpinformation som är effektiv i estimationen.

Innan vi går in på den specifika undersökningen ska vi beskriva kalibreringsestimatorn och visa dess flexibilitet vad gäller utnyttjande av hjälpinformation.

Hjälpinformation består av två delar, nämligen (i) en hjälpvektor för varje svarande objekt och (ii) en populationstotal för hjälpvektorn. Hjälpvektorn benämnes $\mathbf{x}$, och dess värden för objekt k med $\mathbf{x}_k = \left(x_{1k},...,x_{jk},...,x_{Jk}\right)'$, en

kolumnvektor med J komponenter, där x_{jk} är dess värde, för objekt k, för j:e hjälpvariabeln. Populationstotalen för hjälpvektorn är alltså $\sum_U \mathbf{x}_k$.

När vi har täckningsproblem kan det vara svårt att erhålla ett exakt värde på $\sum_U \mathbf{x}_k$. I det fallet har vi förhoppningsvis en god approximation, här benämnd $\widetilde{\mathbf{X}}$.

Vid skattning av $Y_U = \sum_U y_k$ har kalibreringsestimatorn följande utseende:

$$\hat{Y}_{UW} = \sum_{r_p} w_k y_k \tag{4.1}$$

där $w_k = d_k v_k$ och

$$v_k = 1 + \left(\widetilde{\mathbf{X}} - \sum_{r_p} d_k \mathbf{x}_k\right)' \left(\sum_{r_p} d_k \mathbf{x}_k \mathbf{x}_k'\right)^{-1} \mathbf{x}_k \quad \text{for } k \in r_p \tag{4.2}$$

Kalibreringsvikterna w_k har den önskade kalibreringsegenskapen att de ”återskapar” de kända totalerna, d.v.s. $\sum_{r_p} w_k \mathbf{x}_k = \widetilde{\mathbf{X}}$.

Vi kommer att använda den generella kalibreringsestimatorn (4.1) i detta arbete. Det kan dock vara intressant att veta att de flesta kända estimatorer utgör specialfall från kalibreringsestimatorn; ett exempel visas nedan.

EXEMPEL: *En vanlig estimator av* Y_U

Antag att s_F dras med stratifierat obundet slumpmässigt urval som beskrivs i avsnitt 3.5. En estimator av Y_U som är vanlig på SCB är

$$\hat{Y}_U = \sum_{h=1}^{H} \frac{N_{Fh}}{m_{ph}} \sum_{r_{ph}} y_k \tag{4.3}$$

I Lundström and Särndal (2001, Example 11.3.1) visas att estimator (4.3) är ett specialfall av (4.1). Det inträffar när $\mathbf{x}_k$ får identifiera tillhörighet till strata och $\widetilde{\mathbf{X}} = \left(N_{F1},...,N_{Fh},...,N_{FH}\right)'$.

I studerade undersökning (avseende läsåret 2002/2003) fanns en preliminär version av årskurs-tre-registret vid estimationstillfället. Detta register

bedömdes vara relativt väl överensstämmande med det slutliga registret, som i sin tur har mycket hög kvalitet. Därför hämtade vi $\tilde{\mathbf{X}}$ från den preliminära versionen.

Det vi beskriver längre fram i rapporten, särskilt i avsnitten 6.4 och 6.5, har vi kunnat göra först när årskurs-tre-registret var klart (april). Då kan vi studera i vilken mån täckningsbristerna har reducerats med hjälp av denna teknik. Dessutom kan vi se den verkliga skillnaden mellan den preliminära och den slutliga versionen av årskurs-tre-registret.

Det centrala arbetet för att erhålla en god kvalitet på skattningarna är att använda ”stark” hjälpinformation. I nästa avsnitt beskriver vi detta arbete för vår undersökning.

4.2. Hjälpvariabler

4.2.1. Inledning

Vid val av hjälpvariabler är det tre kriterier som ska beaktas.

Det första kriteriet är att variabeln samvarierar väl med svarsbenägenheten (-sannolikheten). Det är det viktigaste kriteriet eftersom det leder till en minskning av bortfallsbiasen för alla skattningar.

Det andra kriteriet är att variabeln samvarierar väl med (viktiga) målvariabler. Om så är fallet minskar bortfallsbiasen för de skattningar som byggs upp av dessa målvariabler. Även variansen minskar för dessa skattningar.

Det tredje kriteriet är att variabeln avgränsar (viktiga) redovisningsgrupper. Det leder framför allt till minskad varians i skattningar för dessa redovisningsgrupper.

Vår erfarenhet är att uppsättningen variabler som uppfyller första kriteriet är vanligtvis relativt lika i olika undersökningar. Kriterium (ii) är dock mer undersökningsspecifik. Även vilka redovisningsgrupper som studeras varierar mellan undersökningar.

4.2.2. Presumtiva hjälpvariabler

Populationen i denna undersökning är sådan att vissa registervariabler inte är användbara. Eftersom alla är (ungefär) lika gamla, alla har samma utbildning och ingen förvärvsarbetar ger inte registervariablerna ålder, utbildning och inkomst någon information. Däremot är det rimligt att utnyttja en del av

föräldrarnas (”vårdnadshavarens”) variabler. T.ex. vet vi från andra studier att om föräldrarna är högutbildade så är det troligt att barnet väljer en längre akademisk utbildning. Dessa uppgifter hämtas från registret över totalbefolkningen (RTB) och utbildningsregistret. Elevernas ”egna” variabler hämtar vi från den preliminära versionen av årskurs-tre-registret. Dessutom tror vi att elevens slutbetyg i grundskolan kan vara intressant och utnyttjar därför det registret.

Vi gör också vissa hopslagningar av kategorier baserat på kunskaper från tidigare kalibreringar. I tabell 4.1 visas de presumtiva hjälpvariablerna.

Tabell 4.1. *Presumtiva hjälpvariabler*

Variabel (benämning)	Kategorier
Kön	Man ; kvinna
Riksområde (NUTS)	8 områden
Födelseland	Födda i Sverige; Europa utom Sverige; övriga
Storstad	Boende i storstad; övriga
Program	17 kategorier
Vårdnadshavarens födelseland	Födda i Sverige; Europa utom Sverige; övriga
Vårdnadshavarens utbildningsnivå	Förgymnasial; gymnasial; eftergymnasial
Vårdnadshavarens utbildningsinriktning	(1)Allmän utbildning; (2) pedagogik, lärarutbildning, humaniora och konst; (3) samhällsvetenskap, juridik, handel, adm.; (4) naturvetenskap, matematik och data, teknik och tillverkning, lant- och skogsbruk och djursjukvård; (5) hälso- och sjukvård, social omsorg; (6) tjänster
Vårdnadshavarens civilstånd	Gift och registrerat partnerskap; övriga
Slutbetyg i årskurs 9	4 kategorier
Huvudman	Kommunal; landsting; fristående

I det följande analyserar vi variablerna i tabell 4.1 för att slutligen bestämma en hjälpvektor.

4.2.3. Samvarierar med svarsbenägenheten

I detta avsnitt skattar vi procentuella andelen svarande i olika redovisningsgrupper i populationen $U \cap U_F$. Redovisningsgrupperna avgränsas med hjälp av de presumtiva hjälpvariablerna.

Skattningarna görs med följande estimator:

$$\hat{P} = \frac{\sum_{r_p} d_k I_k}{\sum_{r_p + o_p} d_k I_k},$$

där $I_k = \begin{cases} 1 \text{ om elev k tillhör studerade redovisningsgrupp} \\ 0 \text{ för övrigt} \end{cases}$

Tabell 4.2. Skattad procentuell andel svarande fördelat på kön.

Kön	Man	Kvinna
Svarsandel (%)	68.6	80.7

Tabell 4.3. Skattad procentuell andel svarande fördelat på riksområden (NUTS).

NUTS	1	2	3	4	5	6	7	8
Svarsandel (%)	70.2	74.0	73.6	74.6	79.5	75.2	74.6	71.7

Tabell 4.4. Skattad procentuell andel svarande fördelat på födelseland.

Födelseland	Sverige	Europa utom Sverige	Övriga
Svarsandel (%)	75.3	66.5	68.5

Tabell 4.5. Skattad procentuell andel svarande fördelat på storstad/icke storstad.

Storstad	Boende i storstad	Övriga
Svarsandel (%)	74.3	74.6

Tabell 4.6. Skattad procentuell andel svarande fördelat på program (koderna för programmen är förklarade i tabell 6.1).

Program	77	81	83	84	85	86	87	88
Svarsandel (%)	74.2	66.9	60.3	69.3	72.9	64.6	79.5	72.8

Program	89	90	91	92	93	94	95	96	97
Svarsandel (%)	74.0	69.6	73.0	78.4	74.4	73.8	80.7	76.8	76.1

Tabell 4.7. Skattad procentuell andel svarande fördelat på vårdnadshavarens födelseland.

Vårdnadshavarens födelseland	Sverige	Europa utom Sverige	Övriga
Svarsandel (%)	75.1	72.9	68.4

Tabell 4.8. Skattad procentuell andel svarande fördelat på vårdnadshavarens utbildningsnivå.

Vårdnadshavarens utbildningsnivå	Förgymnasial	Gymnasial	Eftergymnasial
Svarsandel (%)	70.9	75.3	79.1

Tabell 4.9. Skattad procentuell andel svarande fördelat på vårdnadshavarens utbildningsinriktning (koderna är förklarade i tabell 4.1).

Vårdnadshavarens utbildningsinriktning	1	2	3	4	5	6
Svarsandel (%)	72.6	78.4	76.8	74.0	74.9	66.9

Tabell 4.10. Skattad procentuell andel svarande fördelat på vårdnadshavarens civilstånd.

Vårdnadshavarens civilstånd	Gift eller registrerat partnerskap	Övriga
Svarsandel (%)	77.4	68.8

Tabell 4.11. Skattad procentuell andel svarande fördelat på slutbetyg i årskurs 9.

Slutbetyg i årskurs 9	0-160	161-200	201-240	241-320
Svarsandel (%)	63.8	68.1	76.9	84.6

Tabell 4.12. Skattad procentuell andel svarande fördelat på huvudman

Huvudman	Kommunal	Landsting	Fristående
Svarsandel (%)	74.6	70.8	76.2

Tabellerna 4.2-4.12 visar att de starka hjälpvariablerna (beträffande kriterium (i)) är framför allt kön, elevens slutbetyg i årskurs 9, vårdnadshavarens utbildningsnivå och civilstånd. Även variabeln vårdnadshavarens födelseland är relativt stark (födda i Europa skulle dock kunna bilda en grupp). Svarsbenägenheten varierar också en hel del mellan olika program. T.ex. är andelen svarande inom elprogrammet (kod 83) 60.3 % och inom vårdprogrammet (kod 95) 80.7 %. Skillnaden mellan storstad och övriga landet är däremot (ovanligt!) litet. Vårdnadshavarens utbildningsinriktning är inte särskilt stark.

Innan vi utesluter någon variabel undersöker vi i vilken mån kriterium (ii) uppfylls för olika variabler.

4.2.4. Samvarierar med målvariabler

Vi har plockat ut 7 viktiga målvariabler och från dessa bildat dikotoma variabler. Dessa konstruerade variabler är förklarade i tabell 4.13.

Tabell 4.13. Konstruerade målvariabler

Målvariabel	Förklaring (se även frågeformuläret)
Till universitet	Fr 1; 1=Ja
Fast program	Fr 3; 1=Ett fast program som ger en bestämd utb.
Studietid	Fr 4; 1= Mer än 4 år
Visst universitet	Fr 5; 1=Viktigast att komma till visst universitet
Teknik/natur	Fr 6; 1= 10+11
Studier utomlands	Fr 11; 1= Ja, det kan jag tänka mig att göra
Studenthandbok	Fr 12A; 1 = Mycket

I nedanstående tabeller skattar vi procentuella andelen inom olika redovisningsgrupper som har en viss egenskap (se tabell 4.13). Den ”population” vi begränsar oss till är den del av $U \cap U_F$ som skulle ha svarat om en totalundersökning hade genomförts med samma metod (och resurser) som i undersökningen. Denna ”population” benämns ibland ”svarsstratum”.

Estimatorn är

$$\hat{P}_y = \frac{\sum_{r_p} d_k I_k y_k}{\sum_{r_p} d_k I_k},$$

där $y_k = \begin{cases} 1 \text{ om elev k har studerade egenskap} \\ 0 \text{ för övrigt} \end{cases}$

Frågan ”Har du planer på att börja på universitet eller högskola inom de närmaste åren? ” ställs till alla i urvalet och alltså avser vår konstruerade variabel ”Till universitet” hela svarsstratum. De övriga variablerna begränsar sig till elever som har sådana planer.

I tabellerna 4.14 – 4.24 redovisas resultaten fördelade efter de presumtiva hjälpvariablernas kategorier.

Tabell 4.14. Kön

Målvariabel	Man	Kvinna
Till universitet	45.4	56.5
Fast program	52.4	61.4
Studietid	22.1	28.1
Visst universitet	18.4	12.5
Teknik/natur	45.6	14.2
Studier utomlands	42.2	55.3
Studenthandbok	3.0	11.2

Tabell 4.15. Riksområden (NUTS)

Målvariabel	NUTS							
	1	2	3	4	5	6	7	8
Till universitet	60.8	51.6	48.5	50.9	49.3	48.0	47.5	46.1
Fast program	55.2	50.9	67.8	57.4	60.0	59.2	61.5	54.6
Studietid	27.0	23.4	19.8	32.2	28.5	20.0	24.8	17.5
Visst universitet	17.9	13.2	9.1	18.5	16.4	12.6	13.5	9.7
Teknik/natur	25.9	25.0	23.1	29.3	32.9	21.8	27.4	27.1
Studier utomlands	56.6	43.5	48.9	46.6	55.1	44.8	48.7	45.8
Studenthandbok	5.8	4.9	7.7	10.8	9.5	7.4	7.8	8.8

Tabell 4.16. Födelseland

Målvariabel	Födelseland		
	Sverige	Europa utom Sverige	Övriga
Till universitet	50.2	65.0	61.4
Fast program	56.3	68.3	69.6
Studietid	24.5	29.1	37.3
Visst universitet	13.1	24.8	33.5
Teknik/natur	27.5	27.5	23.5
Studier utomlands	48.5	59.2	63.3
Studenthandbok	7.6	3.0	12.3

Tabell 4.17. Storstad/övriga

Målvariabel	Storstad/övriga	
	Boende i storstad	Övriga
Till universitet	59.3	48.1
Fast program	55.0	59.0
Studietid	31.0	22.9
Visst universitet	17.8	13.6
Teknik/natur	26.6	27.5
Studier utomlands	56.5	46.7
Studenthandbok	8.1	7.6

Tabell 4.18 Program (koderna för programmen är förklarade i tabell 6.1)

Målvariabel	Program							
	77	81	83	84	85	86	87	88
Till universitet	65.6	32.8	3.9	21.9	6.8	4.4	47.0	29.2
Fast program	53.2	77.8	96.9	56.0	96.4	73.1	47.0	53.9
Studietid	23.9	12.2	3.1	11.9	7.9	15.7	14.3	4.1
Visst universitet	17.5	13.5	0.0	18.2	81.2	4.8	5.7	29.4
Teknik/natur	76.8	0.2	50.5	80.1	100.0	89.5	2.8	5.5
Studier utomlands	33.1	28.1	46.4	22.5	7.3	50.2	61.8	60.7
Studenthandbok	5.8	1.3	0.0	0.4	0.0	0.0	7.2	9.5

Målvariabel	Program (forts.)								
	89	90	91	92	93	94	95	96	97
Till universitet	17.0	11.2	9.3	34.4	35.9	22.2	84.8	51.4	64.8
Fast program	54.7	52.4	37.9	61.1	52.6	66.6	57.0	88.0	56.8
Studietid	0.4	4.6	0.0	0.0	13.0	11.3	42.3	28.3	19.0
Visst universitet	15.3	12.1	4.8	2.8	10.9	0.5	12.9	26.5	16.0
Teknik/natur	0.0	0.0	34.3	0.0	1.2	36.4	49.5	2.4	3.0
Studier utomlands	54.9	64.5	62.1	58.3	49.8	37.7	52.9	32.9	53.9
Studenthandbok	3.4	0.0	0.0	0.0	4.3	6.4	6.6	12.2	10.6

Tabell 4.19 Vårdnadshavarens födelseland

Målvariabel	Vårdnadshavarens födelseland		
	Sverige	Europa utom Sverige	Övriga
Till universitet	50.1	53.3	67.2
Fast program	56.0	68.0	68.4
Studietid	24.0	33.0	37.9
Visst universitet	12.6	23.4	34.0
Teknik/natur	27.8	27.2	21.5
Studier utomlands	47.8	61.0	64.0
Studenthandbok	7.6	4.2	12.0

Tabell 4.20. Vårdnadshavarens utbildningsnivå

Målvariabel	Vårdnadshavarens utbildningsnivå		
	Förgymnasial	Gymnasial	Eftergymnasial
Till universitet	38.0	52.5	67.2
Fast program	58.7	59.6	55.2
Studietid	17.5	23.7	33.3
Visst universitet	18.4	16.0	11.4
Teknik/natur	23.8	27.3	29.8
Studier utomlands	44.0	49.3	54.8
Studenthandbok	7.6	7.6	8.1

Tabell 4.21. Vårdnadshavarens utbildningsinriktning (koderna är förklarade i tabell 4.1).

Målvariabel	Vårdnadshavarens utbildningsinriktning					
	1	2	3	4	5	6
Till universitet	44.2	59.2	54.2	52.4	52.9	34.7
Fast program	62.6	60.6	53.4	54.3	58.5	56.3
Studietid	26.2	27.4	23.2	27.8	24.6	22.7
Visst universitet	22.2	11.2	15.0	14.1	13.4	17.3
Teknik/natur	27.4	26.4	23.8	32.6	27.1	22.9
Studier utomlands	44.4	55.2	52.6	47.0	49.4	50.5
Studenthandbok	7.9	9.9	5.3	6.8	8.6	8.6

Tabell 4.22. Vårdnadshavarens civilstånd

Målvariabel	Vårdnadshavarens civilstånd	
	Gift eller registrerat partnerskap	Övriga
Till universitet	53.5	46.1
Fast program	58.0	56.9
Studietid	26.1	24.2
Visst universitet	15.1	14.5
Teknik/natur	26.9	28.1
Studier utomlands	49.6	50.6
Studenthandbok	7.8	7.6

Tabell 4.23. Slutbetyg i årskurs 9

Målvariabel	Slutbetyg i årskurs 9			
	0-160	161-200	201-240	241-320
Till universitet	23.7	27.9	55.9	78.2
Fast program	58.0	55.2	55.8	59.6
Studietid	18.4	12.8	16.8	36.0
Visst universitet	24.1	20.9	13.4	13.0
Teknik/natur	20.1	23.0	25.3	30.6
Studier utomlands	53.7	48.3	39.9	56.2
Studenthandbok	5.9	5.6	7.1	9.1

Tabell 4.24. Huvudman

Målvariabel	Huvudman		
	Kommunal	Landsting	Fristående
Till universitet	51.2	23.1	68.5
Fast program	58.0	71.1	49.6
Studietid	25.5	5.0	30.8
Visst universitet	15.0	0.4	16.4
Teknik/natur	27.2	33.5	26.0
Studier utomlands	48.9	38.7	69.5
Studenthandbok	7.8	0.5	7.8

De variabler som särskilt förklarar svarsbenägenheten, kön, vårdnadshavarens utbildningsnivå, vårdnadshavarens civilstånd och elevens slutbetyg i årskurs 9 är också starka variabler för kriterium (ii). Även

vårdnadshavarens födelseland är relativt stark. Variabeln storstad stärker här sin ”ställning”.

4.2.5. Avgränsar redovisningsgrupper

Det är viktigt att använda variablerna riksområde, kön och program eftersom de avgränsar redovisningsgrupper.

4.2.6. Slutligt val av hjälpvektor

Det är värdefullt om hjälpvektorn är ”stabil” över tiden och således kan användas i de kommande årens undersökningar. Erfarenhetsmässigt fungerar hjälpvariabler valda utifrån de tre kriterierna bra över tiden. Det man bör kontrollera är att inte någon grupp får alltför få observationer. Av de enskilda variablerna är program den mest känsliga. I årets urval finns det endast 18 elever i ett program. Trots detta anser vi, att program bör ingå i hjälpvektorn eftersom den avgränsar viktiga redovisningsgrupper. Vi undviker också att korstabulera hjälpvariabler för att inte få alltför små celler.

Efter en sammanvägning av analysen kring de tre kriterierna samt efter kontroll av vikternas fördelning föreslår vi följande hjälpvektor:

Kön+NUTS+program+vårdnadshavarens födelseland+ vårdnadshavarens utbildningsnivå+vårdnadshavarens civilstånd+slutbetyg i årskurs 9

Den genomsnittliga storleken på v_k-vikten (4.2) är 1.45 och det minsta värdet är 0.71 och det största är 2.35. Dessa värden håller sig inom de i litteraturen redovisade rekommendationerna över variationsvidden för vikterna.

5. Variansskattningar

Produkten redovisar konfidensintervall i anslutning till de flesta punktskattningarna i SMet (se avsnitt 3.4). Vid beräkningen av variansskattningarna används CLAN97. Det är naturligtvis viktigt att konfidensintervall kan beräknas även när KAL-estimatorn används.

I många av SCB:s undersökningar är täckningsproblemen obetydliga och därför kan mer ”konventionella” tekniker användas. De behandlas i Lundström och Särndal (2001), kapitel 1-10. I föreliggande undersökning har vi stora täckningsproblem och därmed befinner vi oss i den situation som beskrivs i kapitel 11 i samma CBM. Vi har en mycket god approximation av hjälptotalerna i målpopulationen och dessutom kan vi identifiera mängden

o_p vilket gör att variansestimator (11.3.6) med vikten (11.2.11) i nämnda CBM kan användas.

Även i detta ”okonventionella” fall beräknar CLAN97 variansskattningar.

Variansestimator utgör summan av samplingsvariansen och bortfallsvariansen. CLAN97 beräknar summan utan att särredovisa de två termerna. Vi vill gärna få en uppfattning om hur stora termerna är och därför har vi gjort ett eget program för detta. Det behandlar dock bara variansskattningar för totalskattningar. I bilaga A redovisas utseendet på de två termerna.

6. Resultat

6.1. Inledning

Skattningar baserade på kalibreringsestimatorn har använts i årets (2002/2003) undersökning. Estimatorn utnyttjar den hjälpinformation som anges i avsnitt 4.2.6. Kan vi säkert säga att kalibreringsestimatorn har gett bättre skattningar än den gamla estimatorn som inte utnyttjar hjälpinformation? Och vad menas i så fall med ”bättre”?

Som vi påpekat i avsnitt 3.4 menar vi att bytet av estimator kan reducera fyra olika brister i statistiken, nämligen bortfallsbias, varians, täckningsproblem samt bristande konsistens. I det följande försöker vi bedöma om så är fallet.

6.2. Reduktion av bortfallsbiasen

Det är inte möjligt att mäta bortfallsbiasen annat än i konstruerade fall. Det krävs nämligen att man känner svarssannolikheten, vilket man aldrig gör i verkligheten. Ett vanligt sätt att studera effekten av bortfall är att genomföra simuleringar inom olika populationer och med olika svarsmodeller. I Lundström (1997) redovisas en mängd simuleringsstudier och en slutsats är att bortfallsbiasen kan reduceras kraftigt om stark hjälpinformationen används i en kalibreringsestimator.

I föreliggande arbete har vi ett enda urval för vilket vi kan beräkna två värden för varje tabellcell, nämligen skattning med (i) HT-estimatorn (tidigare utnyttjad metod) och med (ii) KAL-estimatorn. Som vi tidigare påpekat så har kalibreringsestimatorn den goda egenskapen att den reducerar bortfallsbiasen. Men vi kan inte utifrån ett urval skatta bortfallsbiasen. Däremot tror vi att skillnaden mellan värde (i) och (ii) antyder storleken på

förbättringen eller, uttryckt på annat sätt, reduceringen av bortfallsfelet. Vi försöker också samtidigt reducera täckningsfelen och effekten av detta går inte att särskilja från bortfallsfelet. Vi vet inte heller hur stort det resterande bortfallsfelet är.

Vi väljer att begränsa jämförelsen till den första tabellen i SM:et, som är en av de viktigaste. Den innehåller två typer av parametrar. I den första delen skattas procentuella andelen som har olika planer på att börja läsa på universitet/högskola och i den andra delen skattas antalet årskurs-tre-elever i de olika programmen.

Låt oss först jämföra skattningar av procenttalen. HT-estimatorn för totaler är beskriven i uttryck (3.1) och motsvarande KAL-estimator i (4.1). Vid skattning av procenttal används respektive estimator i både täljare och nämnare.

Tabell 6.1. Jämförelse mellan skattningar av procenttal baserade på HT-estimatorn och KAL-estimatorn

Program (Inom parentes anges koder för programmen)	Planer på att börja läsa på universitet/högskola?								
	HT			KAL			Differens HT-KAL		
	Ja	Inte best.	Nej	Ja	Inte best.	Nej	Ja	Inte best.	Nej
Samtl pgm	49.1	25.6	25.3	49.0	25.5	25.6	0.1	0.1	-0.3
Kvinnor	54.7	28.8	16.5	54.9	28.9	16.2	-0.2	-0.1	0.3
Män	43.8	22.5	33.7	43.3	22.2	34.6	0.5	0.3	-0.9
Barn- och fritid (77)	31.5	44.9	23.6	31.3	43.9	24.8	0.2	1.0	-1.2
Bygg (81)	4.2	11.9	83.9	3.4	12.1	84.5	0.8	-0.2	-0.6
El (83)	22.6	27.8	49.6	21.5	28.5	50.0	1.1	-0.7	-0.4
Energi (84)	6.3	30.5	63.2	7.5	29.6	62.8	-1.2	0.9	0.4
Estetiska (85)	44.7	41.5	13.8	46.4	39.9	13.7	-1.7	1.6	0.1
Fordon (86)	5.3	8.7	86.0	4.9	8.3	86.8	0.4	0.4	-0.8
Handel (87)	27.9	38.4	33.7	29.1	37.9	32.9	-1.2	0.5	0.8
Hantverk (88)	18.3	27.3	54.3	16.9	27.7	55.4	1.4	-0.4	-1.1
Hotell (89)	12.1	41.0	46.9	10.9	41.6	47.4	1.2	-0.6	-0.5
Industri (90)	8.0	23.3	68.7	8.7	21.6	69.7	-0.7	1.7	-1.0
Livsmedel (91)	28.3	23.9	47.8	33.4	19.3	47.4	-5.1	4.6	0.4
Medie (92)	35.1	41.7	23.2	34.4	44.1	21.5	0.7	-2.4	1.7
Naturbruk (93)	21.3	25.9	52.8	21.3	24.5	54.2	0.0	1.4	-1.4
Naturvetensk (94)	84.4	11.2	4.4	84.8	11.1	4.1	-0.4	0.1	0.3
Kvinnor	84.6	12.2	3.3	85.0	12.3	2.7	-0.4	-0.1	0.6
Män	84.3	10.5	5.3	84.7	10.2	5.1	-0.4	0.3	0.2
Omvårdnad (95)	50.9	34.3	14.8	48.1	35.9	15.9	2.8	-1.6	-1.1
Samh.vetensk (96)	62.7	24.8	12.4	63.6	24.1	12.3	-0.9	0.7	0.1
Kvinnor	66.6	23.0	10.3	67.7	22.3	10.0	-1.1	0.7	0.3
Män	55.8	28.0	16.2	56.2	27.5	16.4	-0.4	0.5	-0.2
Teknik (97)	64.1	18.6	17.4	63.3	18.8	17.8	0.8	-0.2	-0.4

Förändringen beräknad i procentenheter är inte särskilt stor för de flesta skattningarna. För elever som exempelvis går på programmet för livsmedel (liten grupp) ser vi dock en kraftig förändring av skattningarna (”ja” resp. ”inte best”). För omvårdnadsprogrammet ser vi en relativt stor överskattning av de som har planer på att börja läsa på universitet/högskola enligt HT-estimatorn.

Är möjligen biasreduktionen stor relativt medelfelet? För att få en uppfattning om biasreduktionen (bortfallsfelet) är stor eller liten jämför vi

den med medelfelet. Antag att KAL-estimatorn är unbiased och att differenserna HT-KAL visar biasen för HT-estimatorn. Vi följer Cochran (1963) och bildar kvoten mellan biasen och medelfelet där han visar hur konfidensgraden påverkas av en ökande kvot. Om den är mindre än 0.1 har den ingen betydelse för konfidensgraden, men har den ett värde på 1.0 har vi en verklig konfidensgrad som är en helt annan än den angivna (t.ex. 95 %). Vi har beräknat kvoterna för varje cell i studerade tabell och funnit att 80 % av kvoterna ligger under 0.4, att ingen kvot är över 1.0 och det högsta värdet är 0.7. Man kan alltså inte säga att biasreduktionen vid skattning av procenttalen är särskilt betydande.

Låt oss därefter jämföra skattningarna av totaler (antal). Resultatet presenteras i tabell 6.2.

KAL-skattningarna i tabellerna överensstämmer helt med skattningarna baserade på registret (se även avsnitt 6.4, utom skattningarna i de celler som avser kvinnor och män inom naturvetenskap resp. samhällsvetenskap. Det beror på att vi använt hjälptotaler från årskurs-tre-registret för samtliga celler med undantag för de nämnda.

HT-estimatorn är stratifierad efter riksområde, programgrupp och kön och därför är de tre första skattningarna, ”Samtliga”, ”Kvinnor”, ”Män”, hämtade från årskurs-två-registret. Skillnaden mellan skattningarna för dessa beskriver alltså enbart täckningsfelet i HT-estimatorn.

Övriga skillnader, utom skattningarna i de celler som avser kvinnor och män inom naturvetenskap resp. samhällsvetenskap, visar totalfelet, d.v.s. den sammanlagda effekten av urvalsfel, bortfallsfel och täckningsfel i HT-estimatet.

Tabell 6.2. Jämförelse mellan skattningar baserade på HT-estimatorn och KAL-estimatorn vid skattning av antal elever i olika program.

Program	HT	KAL	Differens HT-KAL
Samtl pgm	84854	80494	4360
Kvinnor	41330	39359	1971
Män	43525	41135	2390
Barn- och fritid	3032	3507	-475
Bygg	2597	2553	44
El	4805	4019	786
Energi	622	555	67
Estetiska	5093	4475	618
Fordon	2653	3074	-421
Handel	4295	3816	479
Hantverk	1896	1397	499
Hotell	4245	4180	65
Industri	1388	1283	105
Livsmedel	368	425	-57
Medie	3922	3787	135
Naturbruk	2488	2071	417
Naturvetenskap	15643	13768	1875
Kvinnor	6956	5963	993
Män	8687	7805	882
Omvårdnad	3154	2913	241
Samhällsvetenskap	22252	22663	-411
Kvinnor	14270	14619	-349
Män	7983	8044	-61
Teknik	6401	6008	393

Differensen mellan skattningarna i de celler som avser kvinnor och män inom naturvetenskap resp. samhällsvetenskap beror på totalfelet i HT-estimatet, men kan delvis bero på kvarvarande fel i KAL-estimatet.

6.3. Reduktion av variansen

Det är inte möjligt att beräkna variansen för olika skattningar utan endast *skattningar* av variansen. Variansestimatorn för HT-estimatorn utnyttjar (implicit) en mycket enkel modell för svarsbenägenheten medan variansestimatorn för KAL-estimatorn baserar sig på en betydligt mer avancerad svarsmodell. Därför är det rimligt att tro att den senare

variansestimatorn ger säkrare skattningar än den förra. Det är därför svårt att dra säkra slutsatser av jämförelsen mellan skattningar presenterade i tabell 3 och 4.

Tabell 6.3. Jämförelse mellan halva 95%-iga konfidensintervallets bredd för skattningar av procenttal baserade på HT-estimatorn och KAL-estimatorn

Program	Planer på att börja läsa på universitet/högskola?								
	HT			KAL			Differens HT-KAL		
	Ja	Inte best.	Nej	Ja	Inte best.	Nej	Ja	Inte best.	Nej
Samtl pgm	1.7	1.6	1.5	1.6	1.6	1.4	0.1	0.0	0.1
Kvinnor	2.4	2.3	1.9	2.4	2.3	1.9	0.0	0.0	0.0
Män	2.4	2.3	2.4	2.3	2.3	2.3	0.1	0.0	0.1
Barn- och fritid	9.4	10.4	8.8	8.6	9.7	8.5	0.8	0.7	0.3
Bygg	5.3	8.0	9.2	5.0	9.0	10.0	0.3	-1.0	-0.8
El	7.7	7.6	8.9	8.2	8.7	9.9	-0.5	-1.1	-1.0
Energi	9.5	20.0	20.9	11.5	19.7	21.3	-2.0	0.3	-0.4
Estetiska	7.5	7.6	4.9	7.3	7.4	4.7	0.2	0.2	0.2
Fordon	4.5	5.5	6.9	3.6	4.5	5.6	0.9	1.0	1.3
Handel	8.1	8.3	8.5	8.7	8.8	8.9	-0.6	-0.5	-0.4
Hantverk	10.1	11.2	12.7	12.7	14.8	16.8	-2.6	-3.6	-4.1
Hotell	5.4	8.8	8.8	5.2	9.0	9.0	0.2	-0.2	-0.2
Industri	7.5	13.5	14.6	7.3	12.2	13.5	0.2	1.3	1.1
Livsmedel	28.1	20.5	27.9	22.1	13.8	21.4	6.0	6.7	6.5
Medie	8.0	8.7	7.4	7.3	8.1	6.9	0.7	0.6	0.5
Naturbruk	9.3	9.9	11.7	10.7	10.8	13.2	-1.4	-0.9	-1.5
Naturvetenskap	3.1	2.7	1.8	3.2	2.8	1.8	-0.1	-0.1	0.0
Kvinnor	4.6	4.2	2.1	4.8	4.5	2.1	-0.2	-0.3	0.0
Män	4.3	3.5	2.8	4.3	3.4	2.8	0.0	0.1	0.0
Omvårdnad	10.1	9.8	6.8	10.3	10.2	7.3	-0.2	-0.4	-0.5
Samhällsvetenskap	3.5	3.1	2.4	3.3	3.0	2.4	0.2	0.1	0.0
Kvinnor	4.1	3.6	2.7	4.0	3.4	2.7	0.1	0.2	0.0
Män	6.4	5.9	4.7	6.1	5.6	4.6	0.3	0.3	0.1
Teknik	6.6	5.2	5.2	6.4	5.1	5.1	0.2	0.1	0.1

För de flesta cellskattningarna ser vi en minskning av det skattade medelfelet även om storleken på minskningen är obetydlig.

Tabell 6.4. Jämförelse mellan halva 95%-iga konfidensintervallets bredd för skattningar baserade på HT-estimatorn och KAL-estimatorn vid skattning av antal elever i olika program.

Program	HT	KAL	Differens HT-KAL
Samtl pgm	281	0	281
Kvinnor	219	0	219
Män	176	0	176
Barn- och fritid	595	0	595
Bygg	593	0	593
El	738	0	738
Energi	278	0	278
Estetiska	726	0	726
Fordon	555	0	555
Handel	717	0	717
Hantverk	466	0	466
Hotell	705	0	705
Industri	441	0	441
Livsmedel	205	0	205
Medie	646	0	646
Naturbruk	568	0	568
Naturvetenskap	1109	0	1109
Kvinnor	730	517	213
Män	835	517	318
Omvårdnad	598	0	598
Samhällsvetenskap	1115	0	1115
Kvinnor	750	582	168
Män	825	582	243
Teknik	788	0	788

Den stora reduktionen av medelfel visas alltså i skattningen av totala antalet elever på olika program. I själva verket har vi inget urvalsfel för de skattningar som också används som hjälptotaler. Däremot är medelfelen större än noll vid skattning av antalet män resp. kvinnor inom ett specifikt program. Det beror på att vi inte använt hjälptotaler från korstabellen kön*program utan endast marginalsummorna. Skälet till att vi inte använt den informationen är att antalet observationer blev för litet i vissa celler. (Det bör dock vara möjligt att inför nästa undersökning göra en mer ingående analys av tabellplanen. Det skulle kunna leda till att t.ex. endast programmen naturvetenskap och samhällsvetenskap delas upp på kön i hjälpinformationen.)

I de fall medelfelen är större än noll för KAL-estimatorn har vi beräknat komponenterna i variansen. För naturvetenskap, kvinnor (lika för män) utgör bortfallsvariansen 22.7 % av den totala variansen och för samhällsvetenskap, kvinnor (lika för män) är motsvarande siffra 32.2 %.

6.4. Reduktion av täckningsfelen

Som vi har påpekat tidigare så bestod urvalsramen av årskurs-två-registret och vi har i estimationen använt en preliminär version av årskurs-tre-registret. Nu finns det slutliga registret som vi kan göra jämförelser med. Det slutliga registret utgör en mycket god beskrivning av målpopulationen.

Vi antog, efter kontakter med den produktansvarige för registret, att den preliminära versionen av årskurs-tre-registret var tillräckligt bra för att kunna användas i estimationen (i slutet av januari). Låt oss titta på i vilken mån vårt antagande var riktigt.

Den slutliga versionen innehåller 80379 elever och 40 av dessa finns inte i den preliminära versionen. Dessutom innehöll den preliminära versionen 155 elever som inte ska vara med. Det är alltså en mycket god överensstämmelse. Särskilt tydligt blir detta när vi jämför med urvalsramen. Den gemensamma delen mellan urvalsramen och den slutliga versionen är 77 814 elever. Antalet elever som tillhör övertäckningen är 7784 och antalet i undertäckningen är 2565.

Låt oss också titta på hur skillnaden mellan den preliminära versionen och den slutliga fördelar sig på program.

Tabell 6.5. Skillnaden mellan den preliminära och den slutliga versionen fördelad på program.

Program	Preliminär	Slutlig	Differens
Barn- och fritid	3507	3499	8
Bygg	2553	2554	-1
El	4019	4027	-8
Energi	555	555	0
Estetiska	4475	4477	-2
Fordon	3074	3075	-1
Handel	3816	3807	9
Hantverk	4180	4170	10
Hotell	1397	1382	15
Industri	1283	1287	-4
Livsmedel	425	425	0
Medie	3787	3787	0
Naturbruk	2071	2058	13
Naturvetenskap	13768	13751	17
Omvårdnad	2913	2910	3
Samhällsvetenskap	22663	22612	51
Teknik	6008	6003	5

Även tabell 6.5 visar att kalibrering av vikter med anpassning mot totaler i den preliminära versionen ger nästan perfekta skattningar av totalerna.

Vid bedömning av täckningsfelens storlek vid totalskattningar ger alltså tabell 6.2 tillförlitlig information. Undantag utgör uppgiften om antalet män och kvinnor som ingår i programmen naturvetenskap och samhällsvetenskap. De uppgifterna är också behäftade med urvalsfel.

Tabell 6.2 visar att HT-estimatorn ger en överskattning på 4360 elever. Om det inte vore någon skillnad i svarsbenägenhet (inom varje stratum) mellan elever tillhörande målpopulationen och övertäckningen skulle HT-estimatorn (3.1) ge en underskattning motsvarande storleken på undertäckningen (2565). Nu är inte svarsbenägenheten lika stor; av de som tillhör målpopulationen svarar 82.8 % (oviktat) och av de som tillhör övertäckningen 44.6 % (oviktat). En annan förklaring till överskattningen är att övertäckningen är betydligt större än undertäckningen. Den kraftigaste överskattningen erhålls när ingen svarar i övertäckningen, d.v.s. $m_{\setminus p} = 0$. Då kommer antalet elever att skattas till 85598.

6.5. Förbättrad konsistens mellan skattningar och registerdata

Tidigare år har HT-skattningar presenterats i mars och uppgifter från årskurs-tre-registret ett par månader senare. HT-skattningarna av antal elever i t.ex. varje program har då inte överensstämt med uppgifterna från registret. Användningen av KAL-estimatorn eliminerar denna inkonsistens för de flesta jämförbara storheter. En viss liten skillnad erhålls visserligen p.g.a. att vi endast har tillgång till en preliminär version av årskurs-tre-registret. Förhoppningsvis är det möjligt att kommande år ytterligare minska skillnaden mellan den version som finns vid estimationstillfället och den slutliga versionen.

7. Slutsatser

Objektsbortfallet i undersökningen är ca 26 % och bortfallsanalysen indikerar att bortfallet är snedvridande. I analysen har vi tagit med både hjälpvariabler som anger en direkt egenskap hos gymnasieungdomar samt några av vårdnadshavarens egenskaper. Till den första gruppen variabler hör kön, bostadsort, programtillhörighet och slutbetyg i årskurs 9. Till den senare gruppen hör födelseland, utbildningsnivå och –inriktning samt civilstånd.

Vi konstaterar att framför allt kön, vårdnadshavarens utbildningsnivå, vårdnadshavarens civilstånd och elevens slutbetyg i årskurs 9 är variabler som är viktiga för att beskriva variationen i svarsbenägenheten. Det är särskilt viktigt att hitta starka sådana variabler eftersom det reducerar bortfallsbiasen för alla skattningar.

I våra försök att hitta variabler som förklarar variationen i viktiga målvariabler konstaterar vi att de variabler som förklarar svarsbenägenheten också är viktiga här. Dessutom stärker variabeln vårdnadshavarens födelseland sin position att ingå i den slutliga hjälpvektorn.

Variablerna riksområde, kön och program ska användas för uppdelning av redovisningsgrupper och bör därför vara med i hjälpvektorn. Detta tredje kriterium ledde till att också ta med riksområde och program.

I undersökningen presenteras skattningar av både procenttal och antal. I SMet utgör huvuddelen procenttal och i SSD är det enbart antal. Den tidigare använda estimatorn överskattar antalet elever i årskurs tre med 5.4 %, vilket naturligtvis också i varierande grad gäller övriga antalsskattningar. Överskattningen beror på att övertäckningen är mycket större än undertäckningen och att svarsbenägenheten är mycket lägre bland elever som tillhör övertäckningen än de som tillhör målpopulationen.

Kalibreringsestimatorn ger exakta skattningar för många antalsparametrar. För andra antalsparametrar kan man räkna med att urvalsfelet, bortfallsfelet och täckningsfelet har reducerats.

Kalibreringsestimatorn reducerar också problemet med att skattningarna i undersökningen inte överensstämmer med statistik baserad på årskurs-tre-registret.

Skattningar av kvoter, t.ex. procenttal, påverkas inte i lika hög grad av nämnda svaghet, eftersom felen finns både i täljare och nämnare och det faktum att felen vanligtvis går ”åt samma håll”. Därför blir det ofta mindre skillnad mellan metoder vid skattning av kvoter än vid skattning av totaler (antal). Det syns också i denna undersökning.

Kalibreringsestimatorn reducerar också variansen mer för antalsskattningar är för kvotskattningar. I denna undersökning är det inte ens troligt att det blir någon minskning av variansen för den senare parametertypen.

Referenser

Cochran (1963). Sampling Techniques. Wiley & Sons.

UF 36 SM 0301. Övergång gymnasieskola-högskola.

Lundström, S. (1997). Calibration as a standard method for treatment of nonresponse. Stockholm University.

Lundström, S. and Särndal, C.E. (1999). Calibration as a standard method for treatment of nonresponse. Journal of Official Statistics, 15, 305-327.

Lundström, S. and Särndal, C.E. (2001). Estimation in the Presence of Nonresponse and Frame Imperfections. Statistics Sweden.

Bilaga A: Variansestimator för kalibreringsestimatorn

Vi anpassar variansestimator (11.3.6) i Lundström och Särndal (2001) till designen STOSU och med uppdelning på termerna $\hat{V}_{SAM}$ och $\hat{V}_{NR}$:

$$\hat{V}(\hat{Y}_{UW}) = \hat{V}_{SAM} + \hat{V}_{NR} \tag{A.1}$$

där

$$\hat{V}_{SAM} = \sum_{h=1}^{H} N_h^2 (1 - \frac{n_h}{N_h}) \frac{1}{n_h} \frac{1}{n_h - 1} \left[\sum\nolimits_{r_{ph}} (v_{sk} e_k)^2 - \frac{1}{n_h} (\sum\nolimits_{r_{ph}} v_{sk} e_k)^2 \right] -$$

$$- \sum_{h=1}^{H} \frac{N_h}{n_h} \left(\frac{N_h}{n_h} - 1 \right) \sum\nolimits_{r_{ph}} v_{sk} (v_{sk} - 1) e_k^2 \tag{A.2}$$

och

$$\hat{V}_{NR} = \sum_{h=1}^{H} \left(\frac{N_h}{n_h} \right)^2 \sum\nolimits_{r_{ph}} v_{sk} (v_{sk} - 1) e_k^2 \tag{A.3}$$

där

$$e_k = y_k - \mathbf{x}_k' \hat{\mathbf{B}}_v \tag{A.4}$$

och

$$\hat{\mathbf{B}}_v = (\sum_{h=1}^{H} \frac{N_h}{n_h} \sum\nolimits_{r_{ph}} v_{sk} \mathbf{x}_k \mathbf{x}_k')^{-1} \sum_{h=1}^{H} \frac{N_h}{n_h} \sum\nolimits_{r_{ph}} v_{sk} \mathbf{x}_k y_k \tag{A.5}$$

samt

$$v_{sk} = 1 + (\sum_{h=1}^{H} \frac{N_h}{n_h} \sum\nolimits_{r_{ph} \cup o_{ph}} \mathbf{x}_k - \sum_{h=1}^{H} \frac{N_h}{n_h} \sum\nolimits_{r_{ph}} \mathbf{x}_k)' (\sum_{h=1}^{H} \frac{N_h}{n_h} \sum\nolimits_{r_{ph}} \mathbf{x}_k \mathbf{x}_k')^{-1} \mathbf{x}_k \tag{A.6}$$

för $k \in r_p$.

Förteckning över utkomna R&D Reports

R&D Reports är en för IT-enheten och Metodenheten gemensam publikationsserie, som 1988-01-01 ersatte de tidigare ”gula” och ”gröna” serierna.

Reports published during 2000 and onwards:

2000:1 Kalibrering av vikter – beskrivning av tekniken och de SCB-fall den prövats i *(Sixten Lundström et al)*

2000:2 On Inclusion Probabilities and Estimator Bias for Pareto πps Sampling *(Nibia Aires and Bengt Rosén)*

2000:3 Bortfallsbarometer nr 15 *(Per Nilsson, Ann-Louise Engstrand, Sara Tångdahl, Stefan Berg, Tomas Garås och Arne Holmqvist)*

2000:4 Bortfallsanalys av SCB-undersökningarna HINK och ULF *(Jan Qvist)*

2000:5 Generalized Regression Estimation and Pareto πps *(Bengt Rosén)*

2000:6 A User's Guide to Pareto πps Sampling *(Bengt Rosén)*

2001:1 Det statistiska registersystemet. Utvecklingsmöjligheter och förslag *(SCB, Registerprojektet)*

2001:2 Order πps Inclusion Probabilities Are Asymptotically correct (*Bengt Rosén)*

2002:1 On the Choice of Sampling Design under GREG Estimation in Multiparameter Surveys *(Anders Holmberg)*

2002:2 Model-based calibration for survey estimation, with an example from expenditure analysis Surveys *(Claes Cassel, Peter Lundquist and Jan Selén)*

2002:3 On the Choice of Sampling Design in Business Surveys with Several Important Study Variables *(Anders Holmberg, Patrik Flisberg and Mikael Rönnqvist)*

2002:4 The Sampling- and the Estimation Procedure in the Swedish Labour Force Survey *(Hassan Mirza and Jan Hörngren)*

2003:1 Översyn av undersökningen Inrikes och utrikes trafik med svenska lastbilar *(Johan Eriksson, Per Anders Paulson och Bengt Rosén)*

2003:2 Estimation vid förekomst av bortfall och rambrister i undersökningen Gymnasieungdomars studieintresse *(Henrik Gustafsson och Sixten Lundström)*

ISSN 0283-8680

Tidigare utgivna **R&D Reports** kan beställas genom Katarina Klingberg eller Anna-Lena Carlström, SCB, U/MET, Box 24 300, 104 51 STOCKHOLM (telefon 08-506 940 00, fax 08-506 945 99, e-post katarina.klingberg@scb.se eller anna-lena.carlstrom@scb.se).
R&D Reports from 1988-1999 can - in case they are still in stock - be ordered from Statistics Sweden, attn. Katarina Klingberg or Anna-Lena Carlström, U/MET, Box 24 300, SE-104 51 STOCKHOLM (telephone +46 8 506 940 00, fax +46 8 506 945 99, e-mail katarina.klingberg@scb.se or anna-lena.carlstrom@scb.se).