



Statistiska centralbyrån Statistics Sweden

Kalibreringsrapport

Utländska doktorander

SCB, Stockholm
08-506 940 00

SCB, Örebro
019-17 60 00

www.scb.se



Inledning

I en urvalsundersökning är alltid skattningarna behäftade med *urvalsfel* beroende på att endast en delmängd (urval) av populationen studeras. Ett annat fel uppkommer om vi inte lyckas få svar från alla personer (bortfall) och om de avviker från de svarande med avseende på undersökningsvariablerna. Detta fel kallas för *bortfallsfel*.

I en totalundersökning uppkommer fel om vi inte lyckas få svar från alla personer (bortfall) och om de avviker från de svarande med avseende på undersökningsvariablerna (*bortfallsfel*).

För att underlätta användningen av statistiken är det värdefullt om storleken på felen kan uppskattas. Av nämnda feltyper är det endast storleken på urvalsfelet som kan skattas med hjälp av urvalsinformation. Kunskap om bortfallsfelet kan i regel bara fås på ett indirekt och approximativt sätt genom att utnyttja registervariabler.

Både urvalsfel och bortfallsfel kan reduceras genom att använda ett effektivt uppräkningsförfarande. I följande avsnitt redovisas hur det görs i denna undersökning.

Parametrar

För samtliga frågor kommer den procentuella andelen personer med viss egenskap, t ex procentuella andelen med examen från högskola i Sverige, samt uppräknat antal personer att redovisas. Resultaten kommer att redovisas per kön, ålder, Antal terminer, forskarämnesområde, lärosäte, medborgarskapsland, födelseland samt totalen.

Hjälpinformation

Viss hjälpinformation utnyttjas vanligtvis även före estimationen, t.ex. för bildande av stratifierade urvalsdesigner. Denna undersökning är både en totalundersökning och en urvalsundersökning, därför har stratifiering gjorts. Normalt görs ingen stratifiering vid totalundersökningar. I studerade undersökning används stratifieringsvariablerna antal terminer, kön och forskarämnesområde. Det kan dock finnas ytterligare hjälpinformation som är effektiv i estimationen.

Det centrala arbetet för att få god kvalitet på skattningarna, då kalibreringsestimatorn används, är att använda "stark" hjälpinformation. I följande avsnitt beskrivs detta arbete för denna undersökning.

Tänkbara hjälpvariabler

Vid val av hjälpvariabler är det tre kriterier som ska beaktas (se Lundström och Särndal 2001):

- Det första kriteriet är att variabeln samvarierar väl med svarsbenägenheten (-sannolikheten). Det är det viktigaste kriteriet eftersom det leder till en minskning av bortfallskevheten 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 framförallt till minskad varians i skattningar för dessa redovisningsgrupper.

Tänkbara hjälpvariabler, det vill säga variabler som tros uppfylla de ovan uppsatta kriterierna, hämtades ifrån RTB (Registret över totalbefolkningen) och Universitets- och högskolesregistret. En genomgång av variablerna i dessa två register resulterade i att 8 variabler valdes ut.

Exempelvis vet vi från tidigare undersökningar att svarsfrekvensen brukar skilja sig åt mellan män och kvinnor, dessutom brukar svarsfrekvensen bero av civilstånd och bostadsregion. Både kön och ålder är viktiga redovisningsgrupper och tas därför med i analysen. Andra viktiga redovisningsvariabler är antal terminer, forskningsämnesområde, lärosäte och födelseland så även dessa tas med.

De sammanslagningar av kategorier som gjorts baseras på kunskaper från tidigare kalibreringar. Dessutom har hänsyn tagits till hur de redovisningsgrupper som sedan ska användas är avgränsade. Hjälpvariablerna är definierade enligt tabell 1.

Tabell 1 Tänkbara hjälpvariabler

Variabel (benämning)	Kategorier (koder)
KÖN	1 = Man 2 = Kvinna

ÅLDER	(Ålder: år) 1 = – 25 2 = 26 – 30 3 = 31 – 35 4 = 36 – 40 5 = 41 –
CIVILSTÅND	1 = Gift+registrerat partnerskap 2 = Övriga
FÖDELSELAND	1 = Födda i Asien 2 = Födda i Ocianien, Nordamerika och Europa utom Sverige 3 = Födda i Sverige 4 = Övriga
STORSTAD	1 = Boende i Storstockholm 2 = Boende i Storgöteborg 3 = Boende i Stormalmö 4 = Övriga
ANTAL TERMINER	1 = 1 – 2 2 = 3 – 4 3 = 5 – 6 4 = 7 – 8 5 = 9 – 10 6 = 11 – 20
FORKNINGSÄMNESOMRÅDE	1 = Naturvetenskap 2 = Teknik 3 = Medicin och hälsovård 4 = Övriga
LÄROSÄTE	1 = Universitet med läkarutbildning + KI - ORU 2 = Övriga universitet + KTH 3 = Övriga högskolor

I följande avsnitt analyserar vi variablerna i tabell 1 för att slutligen bestämma en hjälpvektor.

Analys av hjälpinformation

Kriterium 1: Variabeln samvarierar med svarsbenägenheten

För att se huruvida hjälpvariablerna uppfyller det första kriteriet, studeras sambandet mellan den dikotoma variabeln svarande/bortfall och hjälpvariablerna. Det görs genom att beräkna skattad andel svarande i olika grupper, bestämda av respektive hjälpvariabel. Vid skattningen används designvikten (vid stratifierat OSU: N_h/n_h).

Vid stora skillnader mellan svarsandelarna utgör variabeln en stark kandidat till hjälpvariabel.

Tabell 2 Skattad procentuell andel svarande fördelat på kön

Kön	Man	Kvinna
Svarsandel (%)		
Inresande	58.1	57.7
Kontrollgrupp	65.1	72.3

Tabell 3 Skattad procentuell andel svarande fördelat på ålder

Ålder	25 år eller yngre	26 till 30 år	31 till 35 år	36 till 40 år	Äldre än 40 år
Svarsandel (%)					
Inresande	61.1	60.1	56.0	54.1	48.9
Kontrollgrupp	60.2	74.1	65.5	62.9	69.3

Tabell 4 Skattad procentuell andel svarande fördelat på kön

Civilstånd	Gifta, registrerat partnerskap	Övriga
Svarsandel (%)		
Inresande	57.9	57.9
Kontrollgrupp	69.4	67.7

Tabell 5 Skattad procentuell andel svarande fördelat på födelseland

Födelseland	Asien	Ocianien, Nordamerika Europa	Sverige	Övriga
Svarsandel (%)				
Inresande	55.7	60.9		56.3
Kontrollgrupp	47.9	59.8	71.0	50.7

Tabell 6 Skattad procentuell andel svarande fördelat på bostadsregion

Bostadsregion	Storstockholm	Storgöteborg	Stormalmö	Övriga
Svarsandel (%)				
Inresande	55.4	59.4	56.8	60.8
Kontrollgrupp	67.8	65.0	67.0	70.5

Tabell 7 Skattad procentuell andel svarande fördelat på antal terminer

Antal terminer	1 till 2	3 till 4	5 till 6	7 till 8	9 till 10	11 till 20
Svarsandel (%)						
Inresande	63.2	62.6	56.7	55.7	49.3	34.5
Kontrollgrupp	68.7	75.9	71.2	69.3	67.4	55.0

**Tabell 8 Skattad procentuell andel svarande fördelat på forsknings-
ämnesområde**

Forskningsämnesområde	Naturvetenskap	Teknik	Medicin och hälso- vård	Övriga
Svarsandel (%)				
Inresande	58.1	59.0	53.1	62.6
Kontrollgrupp	68.3	69.7	65.4	71.6

Tabell 9 Skattad procentuell andel svarande fördelat på lärosäte

Lärosäte	Universitet med läkarutb.	Övriga universitet	Övriga högskolor
Svarsandel (%)			
Inresande	57.3	61.5	54.9
Kontrollgrupp	68.2	70.2	65.7

Tabellerna 3,5 och 7 visar att de starka hjälpvariablerna (beträffande kriterium 1) det är ålder, födelseland och antal terminer. Även variabeln kön, bostadsregion, forskningsämnesområde och lärosäte är relativt stark (för inresande är det små variationer i svarsfrekvens för variabeln kön).

Skillnaden mellan gifta eller registrerade parter och övriga är däremot liten.

Innan någon variabel utesluts undersöks i vilken mån det andra kriteriet uppfylls för de tänkbara hjälpvariablerna.

Kriterium 3: Variabeln avgränsar (viktiga) redovisningsgrupper

Om hjälpvariabeln avgränsar viktiga redovisningsgrupper kan kvaliteten bli bättre i dessa grupper. Framförallt blir skattningarna säkrare om hjälpvariabeln väl avgränsar redovisningsgruppen.

Kalibreringsestimaten ger konsistenta skattningar i den meningen att estimaten ger exakta skattningar för utnyttjade registertotaler. Det är därför viktigt att vid val av hjälpvariabler fundera på vilka skattningar man vill få konsistenta.

Variablerna kön, ålder, antal terminer, forskningsämnesområde och lärosäte avgränsar redovisningsgrupper och bör därför vara med i hjälpvektorn. Dessutom används variablerna födelseland som redovisningsvariabel för inresande.

Slutligt val av hjälpvektor

Från kriterium 1 är det framförallt variablerna ålder, födelseland och antal terminer som bör ingå och alternativa variabler är kön, bostadsregion, forskningsämnesområde och lärosäte. Från kriterium 3 är det framförallt variablerna kön, ålder, antal terminer, forskningsämnesområde och lärosäte som bör ingå. Variablerna ålder och antal terminer visar till viss del samma sak. När ett antal terminer är avklarade har även minst en viss ålder uppnåtts. Ålder har på grund av att det genererar stora vikter exkluderats från hjälpvektorn, men det kompenseras till en del av att antal terminer tas med. Några forskningsämnesområden är knutna till olika lärosäten, å

andra sidan finns det några lärosäten som har fler av forskningsämnesområdena. Så även om det finns en viss samvariation mellan lärosäte och forskningsämnesområde är den inte så stor. Pågrund av att forskningsämnesområde ger stora onormalt stora vikter har variabeln inte kunnats tagits med för att skapa hjälpvektorn.

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

Kön+födelseland+ antal terminer+lärosäte+bostadsregion

Teknisk beskrivning av urval och estimation

Vi har en population U bestående av N personer. De parametrar vi är intresserade av är vanligtvis funktioner av två totaler $Y = \sum_U y_k$ och $Z = \sum_U z_k$, där y_k är värdet på variabel y för person k och z_k värdet på en annan variabel för samma person. Vanligtvis är y (och även z) en dikotom variabel, d.v.s.

$$y_k = \begin{cases} 1 & \text{om person } k \text{ har studerade egenskap} \\ 0 & \text{för övrigt} \end{cases} \quad (4.1)$$

Vanligtvis är vi också intresserade av parametrar för redovisningsgrupper. Låt oss benämna dessa $U_1, \dots, U_d, \dots, U_D$, där $U = \bigcup_{d=1}^D U_d$. Totalen för redovisningsgrupp d kan skrivas

$$Y_d = \sum_U y_{dk} \quad (4.2)$$

$$\text{där } y_{dk} = \begin{cases} y_k & \text{för } k \in U_d \\ 0 & \text{för övrigt.} \end{cases}$$

Z_d bildas på likartat sätt.

En generell parameter för redovisningsgrupp d (d kan också avse hela populationen) kan skrivas $\theta_d = C \frac{Y_d}{Z_d}$, där C är en konstant.

Den vanligaste parametern är en procentuell andel, som erhålles när $C = 100$ och $z_k = 1$ för alla k , och y är definierad enligt (4.1). Om vi låter N_d vara antalet personer i redovisningsgrupp d , då kan parametern skrivas

$$P_d = 100 \frac{\sum_U y_{dk}}{N_d} \quad (4.3)$$

Vi drar ett obundet slumpmässigt urval s_h av storleken n_h från stratum h ($h = 1, \dots, H$), men p.g.a. övertäckning och bortfall har vi endast svarsmängden r_h av storleken m_h att utföra beräkningarna på. Storleken på stratum h ger vi beteckningen N_h .

Den "konventionella" estimatorn (för Y_d), har följande form:

$$\hat{Y}_d = \sum_{h=1}^H \frac{N_h}{m_h} \sum_{r_h} y_{dk} \quad (4.4)$$

I estimator (4.4) används ingen ytterligare hjälpinformation än stratifieringsinformationen.

I syfte att erhålla en estimator med mindre urvalsfel och bortfallsskevhet än estimator (4.4) utnyttjar vi hjälpinformation också i estimationen. Vi bildar en hjälpvektor $\mathbf{x}_k$, som anger till vilka kategorier av

Kön+ födelseland+ antal terminer+lärosäte+bostadsregion

som person k hör. Från RTB och Universitets- och högskolesregistret framställer vi hjälptotalerna $\sum_{U_d} \mathbf{x}_k$. Vi utnyttjar denna hjälpinformation i en kalibreringsestimator.

Kalibreringsestimatorn för totalen Y_d har följande utseende:

$$\hat{Y}_{wd} = \sum_r d_k v_k y_{dk} \quad (4.5)$$

där

$$d_k = N_h / n_h \text{ för } k \in r_h$$

och

$$v_k = 1 + (\sum_U \mathbf{x}_k - \sum_r d_k \mathbf{x}_k)' (\sum_r d_k \mathbf{x}_k \mathbf{x}_k')^{-1} \mathbf{x}_k \quad (4.6)$$

Vid skattning av en parameter av typen $\theta_d = C \frac{Y_d}{Z_d}$ skattas respektive total med hjälp av kalibreringsvikterna $d_k v_k$.

Denna kalibreringsvikt uppfyller kalibreringsvillkoret: $\sum_r w_k \mathbf{x}_k = \sum_U \mathbf{x}_k$, vilket innebär att om vikterna läggs på variabler som ingår i hjälpvektorn summeras dessa upp till de hjälptotaler vi hämtat från registren.

Referenser:

Lundström S. och Särndal C.-E. (2001). *Estimation in the Presence of Nonresponse and Frame Imperfection*. Stockholm: Statistics Sweden