Skolverket

Bilaga 2

Dokumentdatum: 2018-04-13
Dnr: 5.1.3-2017:251

Kalibreringsrapport för undersökningen av ett antal målgruppers deltagande i och uppfattning av Skolverkets skolutvecklingsinsatser inom de nationella skolutvecklingsprogrammen

Statistiska centralbyrån

Dokumentdatum: 2018-04-13

Diarienummer: 5.1.3-2017:251

Jens Malmros, PMU/MIS

Kalibreringsrapport för undersökningen ”Utvärdering av Skolverkets skolutvecklingsinsatser”

1 Inledning

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

För att underlätta användningen av statistiken är det värdefullt om storleken på felen kan uppskattas. Av nämnda typerna av fel 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, så kallad *kalibrering*. Detta beskrivs i avsnitt 2. En teknisk beskrivning av urval och estimation ges i avsnitt 3.

2 Hjälpinformation

För att reducera bortfallsfelet använder man sig av *hjälpinformation*, det vill säga registervariabler vars värden är kända för samtliga enheter i urvalsramen (eller åtminstone i urvalet). Viss hjälpinformation utnyttjas vanligtvis även före estimationen, t.ex. för bildande av stratifierade urvalsdesigner. I denna undersökning drar vi ett stratifierat obundet slumpmässigt urval. Stratum bildas utifrån variablerna personaltyp, huvudman och skolform. Det kan dock finnas ytterligare hjälpinformation som är effektiv i estimationen.

Det centrala kriteriet för att få god kvalitet på skattningarna, då kalibreringsestimatorn används, är att använda ”stark” hjälpinformation. Detta kan sammanfattas i tre kriterier (Särndal & Lundström, 2005):

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

(ii) 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.

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

Förutom information om personaltyp, huvudman och skolform som användes till att bilda stratum har vi tillgång till ett antal hjälpvariabler. Stratumvariabler och hjälpvariabler beskrivs i Tabell 1. Arbetet med att utvärdera vilka hjälpvariabler som skall användas vid beräkningen av uppräkningsvikter beskrivs i Avsnitt 2.1 och 2.2.

Tabell 1. Möjliga hjälpvariabler.

Variabel	**Möjliga värden**
Personaltyp	Rektor, lärare och annan pedagogisk personal
Huvudman	Enskild och kommun
Skolform	Grundskolan och gymnasieskolan
Ålder	För varje kön bildas sex ålderskategorier: 22-31 år, 32-41 år, 42-48 år, 49-56 år, 57-64 år samt 65 år och äldre.
Kön	Man och kvinna
Antal år som lärare	Mindre än 3 år, 4-8 år, 9-14 år, 15-20 år, 21-29 år samt 30 eller fler år.
Kommungrupp	Nio grupper enligt SKL 2016.
Skolstorlek	Mindre än 200 elever, mellan 200 och 499 elever samt 500 eller fler elever.
Tjänsteomfattning	Mindre än 50%, 50%-99% samt 100%.

2.1 Hjälpinformation och svarsandelar

Det skattade svarsandelen i undersökningen är 47,1%. Vid skattning av den totala svarsandelen används designvikten. Skattningarna representerar då den andel som vi tror hade svarat om vi hade undersökt alla individer i urvalsramen. Den skattade svarsandelen kallas också för viktad svarsandel. Den oviktade svarsandelen, det vill säga då man inte tar hänsyn till designvikten, är 42,8%. I fortsättningen redovisar vi skattade svarsandelar om inte annat nämns. I Tabell 2 visas svarsandelar per stratum. Vi ser i tabellen att

de lägsta svarsfrekvenserna finns i stratumen med personaltypen annan pedagogisk personal.

Tabell 2. Svarsandelar (%) per stratum.

Stratum	Bortfall	Svar
Lärare, enskild huvudman, grundskolan	60.4	39.6
Lärare, kommunal huvudman, grundskolan	52.1	47.9
Lärare, enskild huvudman, gymnasieskolan	55.5	44.5
Lärare, kommunal huvudman, gymnasieskolan	44.7	55.3
Rektor, enskild huvudman, grundskolan	55.6	44.4
Rektor, kommunal huvudman, grundskolan	51.3	48.7
Rektor, enskild huvudman, gymnasieskolan	56.7	43.3
Rektor, kommunal huvudman, gymnasieskolan	47.5	52.5
Annan ped. pers., enskild huvudman, grundskolan	74.1	25.9
Annan ped. pers., kommunal huvudman, grundskolan	66.7	33.3
Annan ped. pers., enskild huvudman, gymnasieskolan	76.0	24.0
Annan ped. pers., kommunal huvudman, gymnasiesk.	68.8	31.2

För att få en uppfattning om de möjliga hjälpvariablerna är lämpliga att inkludera i estimationen kan man se hur svarsandelarna varierar mellan möjliga värden på en hjälpvariabel. Om svarsandelarna skiljer sig mycket mellan hjälpvariabelns värden eller kategorier är detta en indikation på att hjälpvariabeln kan vara användbar i estimationen.

I Tabell 3 visas svarsandelar per kön korsad med ålder. Vi ser att svarsandelarna varierar stort över kategorierna och att de i allmänhet ökar med stigande ålder. Vi har även prövat att ytterligare dela in kategorierna i Tabell 3 efter skolform vilket resulterar i totalt tjugofyra kategorier. Man kan då se att det finns skillnader i svarsfrekvens mellan skolform inom kategorier av ålder x kön. Exempelvis har män i åldern 22-31 år en svarsfrekvens på 17,2% i grundskolan och en svarsfrekvens på 38,5% i gymnasieskolan. Vi återkommer till denna indelning i Avsnitt 2.2.

I Tabell 4 visas svarsandelar per antal år som lärare. Vi ser att svarsandelen varierar stort mellan grupperna och att den ökar med antal år som lärare. I Tabell 5, Tabell 6 och Tabell 7 visas svarsandelar per kommungrupp, skolstorlek respektive tjänsteomfattning. Vi ser att svarsandelarna varierar relativt lite mellan grupperna för dessa variabler.

Tabell 3. Svarsandelar (%) per kön och ålder.

Kön och ålder	Bortfall	Svar
Man, 22-31 år	77.5	22.5
Man, 32-41 år	69.1	30.9
Man, 42-48 år	54.5	45.5
Man, 49-56 år	52.9	47.1
Man, 57-64 år	47.9	52.1
Man, 65- år	44.7	55.3
Kvinna, 22-31 år	65.2	34.8
Kvinna, 32-41 år	59.9	40.1
Kvinna, 42-48 år	57.0	43.0
Kvinna, 49-56 år	47.3	52.7
Kvinna, 57-64 år	40.9	59.1
Kvinna, 65- år	43.8	56.2

Tabell 4. Svarsandelar (%) per antal år som lärare.

Antal år som lärare	Bortfall	Svar
0-3 år	68.0	32.0
4-8 år	59.3	40.7
9-14 år	53.9	46.1
15-20 år	47.1	52.9
20-29 år	42.9	57.1
30 år eller fler	39.5	60.5

Tabell 5. Svarsandelar (%) per kommungrupp.

Kommungrupp	Bortfall	Svar
Storstäder	54.1	45.9
Pendlingskommun nära storstad	57.6	42.4
Större stad	51.8	48.2
Pendlingskommun nära större stad	50.2	49.8
Lågpendlingskommun nära större stad	53.7	46.3
Mindre stad/tätort	50.6	49.4
Pendlingskommun nära mindre stad/tätort	47.4	52.6
Landsbygdskommun	52.6	47.4
Landsbygdskommun med besöksnäring	54.7	45.3

Tabell 6. Svarsandelar (%) per skolstorlek.

Skolstorlek	Bortfall	Svar
0-199 elever	57.0	43.0
200-499 elever	51.1	48.9
500- elever	50.8	49.2

Tabell 7. Svarsandelar (%) per tjänsteomfattning.

Tjänsteomfattning	Bortfall	Svar
<50 %	60.3	39.7
50-99%	50.5	49.5
100%	53.0	47.0

2.2 H_3-indikatorn och val av hjälpvektor

H_3-indikatorn ger en indikation på hur väl en hjälpvektor kan reducera bias som har uppkommit som ett resultat av bortfall (Särndal & Lundström, 2010). Med hjälpvektor avses den mängd av hjälpvariablerna som används i estimationen. H_3 avser ingen särskild undersökningsvariabel utan visar på en hjälpvektors förmåga att reducera bias för samtliga undersökningsvariabler (kriterium (i)). I Tabell 8 visas värdet på H_3 multiplicerat med 100 när hjälpvariablerna var och en för sig utgör hjälpvektor. I tabellen ses att

de högsta H_3-värdena fås för hjälpvariablerna ålder x kön, ålder x kön x skolform samt antal år som lärare. Övriga variabler har låga värden. En stegvis ansats med avseende på H_3-värde visar att hjälpvektorn bör innehålla ålder x kön x skolform samt antal år som lärare och att övriga variabler har liten påverkan på H_3 när de läggs till en hjälpvektor som innehåller de två redan nämnda hjälpvariablerna. Analysen med avseende på H_3 visar också att korsningen av ålder, kön och skolform har större potential för att reducera bortfallsbias än korsningen av ålder och kön.

Tabell 8. H_3-värden för enskilda hjälpvariabler.

Variabel (benämning)	$\boldsymbol{H_3 \times 100}$
Ålder x kön	21
Ålder x kön x skolform	24
Antal år som lärare	19
Kommungrupp	6
Skolform	6
Tjänsteomfattning	6

Utifrån vår analys ser vi att flera av de möjliga hjälpvariablerna kan ha goda möjligheter att förklara svarsbenägenheten och därmed är intressanta att inkludera i hjälpvektorn. Det är också av intresse att inkludera ett flertal av de möjliga hjälpvariablerna i hjälpvektorn då de kan utgöra redovisningsgrupper för undersökningen. När samtliga hjälpvariabler (med ålder x kön x skolform men inte ålder x kön) inkluderas i hjälpvektorn fås g-vikter mellan 0,35 och 2,55. Det betyder att de uppräknade designvikterna (se Avsnitt 3) har modifierats av en faktor i detta intervall vilken har beräknats i kalibreringsproceduren utifrån värdena på hjälpvariablerna. Utifrån detta ser vi inga hinder för att inkludera samtliga hjälpvariabler i hjälpvektorn. Den slutgiltiga hjälpvektorn ges alltså av ålder x kön x skolform + antal år som lärare + kommungrupp + skolstorlek + tjänsteomfattning. Hjälpvektorn inkluderar även variablerna personaltyp och huvudman som användes i konstruktionen av strata. Variabeln skolform finns i ålder x kön x skolform och inkluderas därför inte separat.

3 Teknisk beskrivning av urval och estimation

Antag att vår population av intresse ges av U och att den består av N personer. De parametrar som är av intresse är främst funktioner av två totaler $Y = \sum_U y_k$ och $Z = \sum_U z_k$, där y_k är värdet på variabeln y för person k och z_k värdet på en annan variabel för samma person. Vi kan definiera y (och även z) som en dikotom variabel, dvs.

$$y_k = \begin{cases} 1 \text{ om person } k \text{ har studerad egenskap;} \\ 0 \text{ annars.} \end{cases} \quad (1)$$

Det finns givetvis också intresse av parametrar för olika redovisningsgrupper. Låt oss benämna dessa $U_1,\ldots,U_d,\ldots,U_D$, där $U = \bigcup_{d=1}^{D} U_d$. Totalen för redovisningsgrupp d kan skrivas

$$Y_d = \sum_U y_{dk}$$

där

$$y_{dk} = \begin{cases} y_k \text{ för } k \in U_d; \\ 0 \text{ annars.} \end{cases}$$

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 P_d, som erhålles när $C = 100$, $z_k = 1$ för alla k och y är definierad enligt (1). Om vi låter N_d vara antalet personer i redovisningsgrupp d så kan parametern skrivas

$$P_d = 100\frac{\sum_U y_{dk}}{N_d}.$$

Vi drar ett obundet slumpmässigt urval s_h av storleken n_h från stratum $h, h = 1, \ldots, H$, 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. Designvikten ges av

$$d_k = \frac{N_h}{n_h}.$$

Den ”konventionella” estimatorn för Y_d har följande form:

$$\hat{Y}_d = \sum_r d_k^* y_k \quad (2),$$

där r är svarsmängden och

$$d_k^* = \frac{N_h}{m_h}$$

är den uppräknade designvikten, vilket kommer av att d_k räknas upp med termen n_h/m_h. I estimator (2) används ingen ytterligare hjälpinformation än stratifieringsinformationen. Denna estimationsmetod brukar kallas ”rak uppräkning inom strata” eftersom man kompenserar för bortfallet genom att använda m_h istället för n_h i designvikten.

I syfte att erhålla en estimator med mindre urvalsfel och bortfallsskevhet än estimator (2) utnyttjar vi hjälpinformation också i estimationen. Vi bildar en hjälpvektor $\mathbf{x}_k$, som anger till vilka kategorier av *Ålder x kön x skolform* + *Antal år som lärare* + *Kommungrupp* + *Skolstorlek* + *Tjänsteomfattning* + *Personaltyp* + *Huvudman* + *Skolform* som individ *k* tillhör. Från register framställer vi sedan hjälptotalerna $\sum_{U_d} \mathbf{x}_k$, där hjälpvektorn $\mathbf{x}_k$ summeras över dess värden i hela populationen. Vi utnyttjar denna hjälpinformation i en kalibreringsestimator.

Kalibreringsestimatorn för totalen Y_d har följande utseende:

$$\hat{Y}_{wd} = \sum_r g_k d_k^* y_{dk,}$$

där

$$g_k = 1 + \left(\sum_U \boldsymbol{X}_k - \sum_r d_k^* \boldsymbol{X}_k\right)' \left(\sum_r d_k^* \boldsymbol{X}_k \boldsymbol{X}_k'\right)^{-1} \boldsymbol{X}_k.$$

Vid skattning av en parameter av typen $\theta_d = C\frac{Y_d}{Z_d}$ skattas respektive total med hjälp av kalibreringsvikterna $\boldsymbol{w_k} = \boldsymbol{g_k d_k^*}$. Denna kalibreringsvikt uppfyller kalibreringsvillkoret $\sum_r w_{\mathbf{k}} \mathbf{x}_k = \sum_U \mathbf{x}_k$, vilket innebär att om vikterna läggs på variabler som ingår i hjälpvektorn summeras dessa upp till de hjälptotaler vi hämtat från registren.

4 Referenser

Särndal, C.-E. & Lundström, S., 2005. *Estimation in Surveys with Nonresponse.* Chichester: John Wiley & Sons.

Särndal, C.-E. & Lundström, S., 2010. Design for estimation: Identifying auxiliary vectors to reduce nonresponse bias. *Survey Methodology*, pp. 131-144.