

Laboration 3 – skattningar vid stratifierade urval

Allmänt

På förra datorövningen utnyttjades SAS för att beräkna skattningar vid OSU med hjälp av proceduren SURVEYMEANS. Vidare användes SURVEYSELECT för att göra OSU ur en ändlig population. Under denna laboration kommer dessa procedurer att användas lite djupare.

Datamaterialet

Vi ska fortsätta analysera den stora jordbrukssurveyen från USA, så börja först med att läsa in de två datamaterialet (*agpop* och *agsrs*) på samma sätt som i laboration 2. Glöm inte bort att koda om bortfallet för variablerna *acres92* och *acres87* i *agpop*.

Uppgift 1

Först ska totalen för *acres92* skattas utifrån urvalet med 300 observationer. I proceduren SURVEYMEANS finns två kommandon, *sum* och *clsum*. Så för att få ut punktskattning samt intervall för totalen borde man kunna skriva på detta sätt:

```
proc surveymeans data=agrars total=3078 mean clm sum clsum;  
var acres92;  
run;
```

Pröva att skriva och köra koden ovan. Får ni det förväntade resultatet i utskriften? Svaret är nej. Om ni jämför väntevärdesintervall med totalintervallet så ser ni att totalintervallet bara är 300 gånger större, och inte 3078 som förväntat. Detta beror på att SAS använder sig av urvalsvikter, och om dessa inte anges sätts dessa till 1. Vid OSU så är urvalsvikterna **N/n**, så en sådan variabel måste skapas.

Skapa med hjälp av ett **data**-steg en variabel som består av urvalsvikterna för varje observation (3078/300 för alla observationer) och kalla den för **samplingweight**. Använd sedan proceduren SURVEYMEANS igen, men lägg till en rad **weight** efter **var**-raden, där ni anger urvalsvikterna.

Hur blir totalintervallet nu, innehåller intervallet den sanna totalen?

Uppgift 2

När man använder sig av proceduren SURVEYSELECT erhålls urvalsvikter automatiskt, dock inte för OSU. Genom att skriva **stats** i första raden så erhålls urvalsvikter även för OSU.

Gör nu ett OSU om 300 observationer ur *agpop* som i förra laborationen men lägg till **stats** och ta bara med variablerna *county*, *state* och *acres92*. Detta skrivs på följande sätt:

```
proc surveyselect data=agrapr method=srs sampsize=300 stats output=agosu;  
id county state acres92;  
run;
```

Öppna *agosu* och se vad innehållet är. Förutom de beställda variablerna har även **selectionprob** och **samplingweight** tillkommit. Med hjälp av **samplingweight** kan totalen skattas med hjälp av SURVEYMEANS på samma sätt som i uppgift 1. Gör detta!

Uppgift 3

Nu ska vi börja använda oss av stratifierade urval. I detta exempel låter vi de fyra regionerna (NC, NE, S och W) vara strata, och ur varje stratum görs oberoende OSU av ett visst antal områden. Det finns en färdig fil, *agstrat.dat*, som innehåller ett sådant urval med totalt 300 områden. Förutom de vanliga 15 variablerna finns även *rn* och *weight*, vilka är slumptalet som användes vid urvalet samt urvalsvikten (N_h/n_h). Läs in datamaterialet med hjälp av programmet *agstrat.sas*.

Nu skall det beräknas en skattning av väntevärdet med tillhörande konfidensintervall på *acres92*, se formler 3.2 och 3.5 i kursboken. Använd proceduren MEANS för att få medelvärde och standardavvikelse i varje stratum på det stratifierade urvalet. Men, för att kunna beräkna dessa skattningar behöver ni veta stratumstorlekarna N_h . Använd en lämplig procedur på datamaterialet över populationen för att ta reda på dessa. Glöm inte bort att använda en variabel som ej har bortfall, t.ex. *small92*.

Nu när ni vet medelvärdet, standardavvikelsen och stratumstorleken för de fyra strata kan ni beräkna väntevärdet och dess konfidensintervall på miniräknare. Spara resultatet!

Uppgift 4

Så klart går det att lösa uppgift 3 med hjälp av proceduren SURVEYMEANS i SAS. Dock måste stratumstorlekarna anges, så för detta ändamål skapas en fil som innehåller dessa storlekar:

```
data strtot;
input region $ _total_;
cards;
NC 1054
NE 220
S 1382
W 422
;
run;
```

Istället för att skriva *total* = N skrivs:

```
proc surveymeans data=agrarstr total=strtot;
strata region;
var acres92;
run;
```

Genomför körningen och titta på punktskattningen. Den blev väl fel jämfört med den som beräknades i uppgift 3? Det beror på samma sak som i uppgift 1, att urvalsvikterna måste anges. Filen med stratumstorlekarna används bara för att få till ändlighetskorrektionerna.

Datamaterialet med det stratifierade urvalet innehåller ju urvalsvikterna i variabeln *weight*. Komplettera programmet ovan med dessa urvalsvikter. Be även SURVEYMEANS att ge intervall för både medelvärdet och totalen för *acres92*. Jämför dessa skattningar med de från uppgift 2.

Uppgift 5

Nu ska nya stratifierade urval med OSU dras ur population med hjälp av SURVEYSELECT. Efter beräkningar med Neyman-allokering blir de optimala urvalsstorlekarna (se uppgift 3.12) ur de olika strata enligt *data*-steget nedan:

```
data urvst;
input region $ _nsize_;
cards;
NC 69
NE 7
S 122
W 102
;
run;
```

Skapa detta datamaterial, och detta ska sedan anges i SURVEYSELECT enligt nedanstående program:

```
proc surveyselect data=agrar method=srs samsize=urvst out=agstr;
strata region;
id county state acres92;
run;
```

OBS: Populationen måste vara sorterad efter *region* för att detta ska fungera.

Öppna datamaterialet och titta så att det har fungerat. Notera att variablerna ***selectionprob*** och ***samplingweight*** har tillkommit.

Skriv nu ett program för att skatta väntevärdet och totalen för *acres92* med konfidensintervall till båda skattningarna. Glöm inte bort att ange *total* och *weight*.

Inlämningsuppgifter

Om man vill ha laboration rättad ska den lämnas in senast en vecka efter labbtillfället. Labbrapporten ska innehålla svar, **kommentarer, tolkningar** och eventuell SAS-kod.