Laboration 1: Övningar på skattningar vid OSU med utnyttjande av standardprogram för surveyer.

Allmänt

I version 8 (med efterföljare) av *SAS* finns ett par procedurer som vi kan utnyttja för skattningar vid ändliga populationer, *SURVEYMEANS* och *SURVEYSELECT*. Den första proceduren beräknar skattningar från ett givet urval, medan den andra fixar urval ur en population.

SURVEYMEANS klarar skattningar av väntevärden och totaler för OSU, stratifierat urval och klusterurval samt hanterar urval med olika urvalssannolikheter. Dock klarar proceduren inte av kvot- och regressionsskattningar.

SURVEYSELECT klarar direkt av alla möjliga urvalsmetoder: OSU, systematiskt urval, stratifierat urval, och urval med olika sannolikheter (olika former av PPS). Här finns direkt tillgång till alla möjliga urvalsmetoder.

Datamaterial

I denna laboration så ska en jordbrukssurvey från USA analyseras, en närmare beskrivning av surveyen finns i exempel 2.5 på sid. 34-35. Det finns data över hela populationen (*agpop.dat* som den kallas i kursboken) och dels ett OSU ur populationen (*agsrs.dat*), dessa filer finns på kurshemsidan. För att läsa in filerna i *SAS* finns två färdiga program, *agpop.SAS* resp. *agsrs.SAS*. Variablerna är namnade som i kursboken (observera att icke-numeriska variabler måste följas av dollartecken). Data är separerade med kommatecken och detta måste man tala om. Vidare så finns variablernas namn i första raden i filen och därför påbörjas läsningen först på rad nummer 2.

Uppgift 1

Kör *agpop.SAS* och titta på utskriften att materialet bestående av 3078 observationer verkar vara korrekt inläst. Det finns ju en del bortfall kodat med -99 som vi vill ändra till *SAS* bortfallskod som är en punkt. Detta kan göras i ett nytt datasteg i *SAS* t.ex. genom att skriva (där vi kodar om variabeln acres92)

```
data agrarp;
            set agrarp;
            if acres92=-99 then acres92=.;
run;
```

Printa och se att omkodningen har blivit OK.

Ta också fram enkel beskrivande statistik genom att köra *PROC MEANS*. Det räcker med medelvärde och standardavvikelse, så koden blir:

```
proc means data=agrarp mean std;
var acres92;
run;
```

Multiplicera medelvärdet med 3078 (på miniräknare) så har vi populationens total.

(**Svar:** Populationens väntevärde är 308 582.41, totalen blir då 949 816 658)

Uppgift 2

Kör *agsrs.SAS* och kolla upp att inläsningen blivit korrekt. Inget bortfall finns i detta urval.

Vi skall nu använda *SURVEYMEANS* för att skatta populationens väntevärde och total avseende variabeln ***acres92***. Proceduren är ganska enkel och har en hel del valmöjligheter, men viktigt att specificera är förstås datafilen som skall används, hur stor populationen är (för att få med ändlighetskorrektion) och vilka variabler som skall analyseras. Man kan skriva ut en hel del olika statistik från proceduren och vill man ha ut allt används kommandot all. Följande enkla program räcker till en början:

```
proc surveymeans data=agrars total=3078 all;
var acres92;
run;
```

När man skriver ***all*** kommer det ut väldigt mycket beskrivande statistik, och allt detta kanske inte är av intresse. Om man inte skriver något alls får man ut antal observationer (***nobs***), väntevärdesskattningen (***mean***), standardavvikelsen för väntevärdesskattningen (***stderr***) och konfidensintervallgränserna (***clm***).

Ofta vill man också ha en skattning av totalen med konfidensintervall, men detta kommer i nästa laboration, då *SURVEYMEANS* kommer användas djupare. En vanlig sak att skriva i första raden förutom ***data=*** och ***total=*** kan vara ***alpha=*** som används om man inte vill ha 95 % konfidensintervall. Vill man ha t.ex. 99 % intervall får man skriva ***alpha***=0.01.

Beskrivning SURVEYMEANS

Detta är inte en övning utan en beskrivning av syntaxen i proceduren SURVEYMEANS, mer utförlig beskrivning finns så klart att finna i *SAS*-hjälpen.

```
proc surveymeans <options> <statisticskommandon>;
by variables;
class variables;
cluster variables;
strata variables;
var variables;
weight variables;
```

Huvudkommandot och raden `var` har redan använts.

`by` används om man vill göra skattningar för flera separata grupper (t.ex. om man genererat flera urval) och grupptillhörigheten finns angiven som en variabel.

`class` är ett viktigt kommando, då detta talar om att numeriska variabler skall behandlas som kategoriska och man därigenom skattar andelar för varje värde på variablerna. Om den kategoriska variabeln är icke-numerisk (dvs. skriven som text) så behandlas den alltid som kategorisk och `class` behöver inte användas.

`cluster` används förstås bara om man har ett klusterurval och med variablerna ger man klustertillhörigheten. Motsvarande gäller strata.

`weight` förutsätter att det finns en variabel som ger vikten för observationen, och oftast är den vikten inverterade värdet till urvalssannolikheten. Vid OSU, stratifierat OSU och kluster-OSU behövs inte kommandot när man skattar ett väntevärde, men däremot när man skattar en total. Även vid olika typer av PPS-urval behövs `weight` -kommandot.

Uppgift 3

Analysen ska nu fortsätta och områden med stora värden på ***acres92*** ska undersökas. Litet mera specifikt ska andelen områden med värden större än 1 miljon skattas utifrån urvalet med 300 observationer. Detta gör man enkelt i *SAS* genom att utöka datasetet med en sådan variabel. Ett förslag finns nedan där den nya variabeln har döpts till ***acgroup***:

```
data agrars;
set agrars;
if (acres92>1000000) then acgroup='stor';
else acgroup='normal';
run;
```

Sedan är det ju enkelt att få ut skattningarna med hjälp av SURVEMEANS:

```
proc surveymeans data=agrars total=3078;
var acgroup;
run;
```

Studera utskriften och kolla så att rätt information om den kategoriska variabeln erhållits. Notera att det även här står rubriken **Mean** i skattningen men att rubriken för kategoriska variabler skall tolkas som **Proportion.**

Uppgift 4

Det är inte bara de stora områdena som är av intresse, utan även de små. Utöka därför variabeln ***acgroup*** med en till kategori, *liten*. Kategorierna ska nu vara *liten* som innebär ***acres92*** under 50 000, *stor* över 1 000 000 och *normal* däremellan. Utöka koden ovan för att få fram de tre kategorierna och ta fram punktskattning och 90 % intervall för dessa proportioner.

Uppgift 5

Nu ska OSU dras ur populationen *agpop* och för detta används proceduren SURVEYSELECT. Denna procedur har en syntax som i mångt och mycket liknar den i SURVEYMEANS:

```
proc surveyselect <options>;
strata variables;
control variables;
size variable;
id variables;
```

`strata` används vid stratifierat urval, och där anges vilken variabel som innehåller de olika stratumen.

Om man på något sätt vill ordna data så används `control` (t.ex. för systematiskt urval).

Om man gör ett PPS-urval måste `size` finnas med och variabeln innehålla storleksmåttet för PPS-urvalet.

Efter `id` ska de variabler anges som man vill ha i urvalet (alla från populationen kommer med om kommandot inte används).

I huvudkommandot finns en stor mängd options varav dessa främst används: **data=, out=, method=, sampsize=** och **rep=.**

Man ger namnet på utdatafilen i **out**=. Urvalsstorleken anges i **sampsize**= och i **rep=** anger man hur många urval man vill göra (skrivs ingenting görs ett urval). Urvalsmetoderna (**method=**) är många men främst används **srs** (OSU, vilket görs om inget annat skrivs), **sys**, **pps** (utan återläggning) och **pps_wr** (med återläggning).

Uppgiften är att göra ett nytt OSU omfattande 300 observationer ur *agpop* och ta med alla variabler. Urvalet skall sparas i filen *agosu*. Följande enkla program bör fungera:

```
proc surveyselect data=agrarp sampsize=300 out=agosu;
run;
```

Skriv ut filen för att kontrollera att det blivit rätt. Kör sedan SURVEMEANS för att skatta väntevärdet för ***acres92***. Täcker konfidensintervallen de sanna värdena?

Uppgift 6

Gör först med hjälp av SURVEYSELECT tio stycken OSU ur *agpop,* bestående av variablerna ***acres92*** och ***small92***. Använd SURVEYMEANS (med utnyttjande av `by`) för att i varje urval skatta väntevärdet för de två variablerna. Kommentera likheterna och olikheterna mellan de resultat som erhålls i vartdera urvalet. Täcker alla intervall sina sanna väntevärden (dvs. medelvärdet för variablerna i populationen)?

TIPS: Urvalen specificeras genom att en ny variabel ***replicate*** har skapats och att denna antar urvalsnummer som värde.

Inlämningsuppgifter

Vill man ha sin laboration rättad lämnas svar på uppgifterna 3-6 in senast **en vecka efter laborationstillfället**. Se som vanligt till att kommentera och tolka de siffror som erhållits.