Laboration 2: CGS och stratifierat urval (allokering)

Allmänt

Denna laboration syftar till att öka förståelsen gällande centrala gränsvärdessatsen (CGS) samt hur stratifierat urval och dess olika allokeringar kan genomföras med hjälp av *SAS.* Det kan vara bra att under denna laboration även ha tillgång till instruktionen för laboration 1, då vi kommer använda en del saker som togs upp där.

Datamaterial

Även i denna laboration kommer jordbrukssurveyen från USA att undersökas, datamaterialen och koder för att läsa in dessa finns på kurshemsidan.

KOM IHÅG ATT KODA OM BORTFALLET VID IMPORTERING AV DATA.

Uppgift 1

I denna uppgift ska proceduren SURVEYSELECT utnyttjas för att göra ett stort antal oberoende urval ur populationen. Vi kommer att öka stickprovsstorleken *n* succesivt för att se hur det påverkar fördelningen för stickprovsstatistikan. Läs igenom hela texten nedan, uppgiften sammanfattas i slutet.

Den variabel som kommer undersökas är ***acgroup***, som ni skapade på förra laborationen. Dock kommer definitionerna på den att ändras lite, så koden ni ska använda denna gång är:

```
data agpop;
set agpop;
if (acres92>1000000) then acgroup='stor';
else if (acres92<100000) then acgroup='liten';
else acgroup='normal';
run;
```

Kör SURVEYMEANS för att ta reda på populationsandelarna.

För att kunna genomföra ett stort antal oberoende urval används åter igen ***replicate*** (***rep***) i proceduren SURVEYSELECT. Vi kommer för varje stickprovsstorlek genomföra 1000 stycken oberoende urval. Spara de 1000 urvalen i ett dataset (kalla det lämpligtvis för *agosu*), och beräkna därefter enbart andelarna för varje enskilt urval med hjälp av SURVEYMEANS (standardavvikelsen kommer med automatiskt). Nyttja åter igen att variabeln ***replicate*** i datasetet *agosu* anger vilket urval observationerna tillhör och kom ihåg att ange antal observationsenheter i populationen.

För att kunna undersöka hur fördelningen för stickprovsandelarna ser ut måste dessa sparas i ett dataset. Detta görs genom att lägga till raden:

```
ods output statistics=prop;
```

i proceduren SURVEYMEANS. Då skapas ett dataset med namnet *prop* som innehåller de beräknade andelarna, och det finns sparat i mappen *Work.* För att skapa ett histogram över andelarna kan följande kod användas:

```
proc univariate data=prop noprint;
   histogram Mean;
   class varlevel;
run;
```

`class` anger att SAS ska skapa ett histogram för de tre olika andelar vi skattar. Ta därefter hjälp av proceduren MEANS (använd `class` på samma sätt som för histogram) för att beräkna genomsnittliga andelar och standardavvikelser för de 1000 andelar som finns beräknade i datasetet *prop*. Jämför de genomsnittliga andelarna med populationens sanna andelar och jämför standardavvikelserna med de teoretiska standardavvikelserna (för stickprovsandelarna) som går att beräkna från populationsvärdena.

Efter all denna information så kan det ni ska göra sammanfattas så här:

a) Genomför 1000 oberoende urval med stickprovsstorlekarna 10, 50 och 300 (ni ska alltså göra 1000 oberoende urval tre gånger). Undersök hur fördelningarna för stickprovsandelarna förändras och koppla detta till CGS. Är det någon skillnad på små andelar och lite större andelar? Klistra in histogrammen i er laborationsrapport.
b) Beräkna genomsnittliga andelar och standardavvikelser för de 1000 andelarna för respektive stickprovsstorlek. Undersök om dessa överensstämmer med de faktiska och teoretiska värdena, samt diskutera skattningarnas väntevärdesriktighet.

Uppgift 2

I denna uppgift ska vi istället för OSU börja använda stratifierat urval. På denna laboration kommer vi enbart ta upp hur de olika urvalsstorlekarna bestäms med hjälp av de olika allokeringarna (proportionell, Neyman och optimal). Hur intervall och liknande beräknas kommer att tas upp på laboration 3. Den stratifieringsvariabel som kommer användas är ***region*** (består av fyra stycken regioner) och vi kommer utgå från datamaterialet för hela populationen.

Det första som måste göras är att datamaterialet måste sorteras efter stratifieringsvariabeln, vilket kan göras med följande kod:

```
proc sort data=agpop;
by region;
run;
```

Stratifierat urval görs med SURVEYSELECT och nedan visas hur ett stratifierat urval med proportionell allokering görs.

```
proc surveyselect data=agpop sampsize=300;
strata region / alloc=prop nosample;
run;
```

Som synes ska det totalt väljas ut 300 observationer. Raden `strata` anger vilken variabel som är stratifieringsvariabel, och efter / anges olika villkor för denna stratifiering. `alloc` säger vilken allokering som gäller, och `nosample` säger åt SAS att inte dra något urval utan att bara beräkna urvalsstorlekarna. Ni kan läsa mer om allokeringarna på denna länk:

http://support.sas.com/documentation/cdl/en/statug/65328/HTML/default/viewer.htm#statug_surveyselect_syntax07.htm

För optimal allokering ska koden ha följande utseende:

```
proc surveyselect data=agpop sampsize=300;
strata region / alloc=optimal var=(variansstrata1 variansstrata2 osv)
costs=(kostnadstrata1 kostnadstrata2 osv) nosample;
run;
```

Där det såklart ska stå siffror istället för bokstäver inom parenteserna och de ska komma i samma ordning som de olika stratumen (***region***) gör i datasetet. För Neymanallokering tas helt enkelt kostnadsdelen bort.

Man kan också specificera vilken felmarginal man vill ha på intervallet för populationsmedelvärdet genom att lägga till `margin`=siffra efter /. Dock kan man inte i samband med detta ange en önskad total urvalsstorlek.

I kommande uppgifter ska ni tänka er att det ska genomföras en ny undersökning gällande antalet tunnland (acres) och ni ska då bestämma hur många observationer som ska göras i de olika regionerna, totalt ska det göras 300 observationer. När varianserna ska beräknas använder ni information från hela populationen för variabeln ***acres92***.

a) Hur ska de 300 observationerna fördelas om proportionell allokering används?
b) Hur ska de 300 observationerna fördelas om Neymanallokering används?
c) Vi antar att kostnaderna för att undersöka en observation i de olika regionerna är: NC = 2, NE = 3, S = 5 och W = 6. Hur ska de 300 observationerna fördelas?
d) Antag att man i samband med allokeringen i c) önskar att felmarginalen ska vara 30 000. Hur ska urvalet fördelas i detta fall?
e) Varför kan man inte ange en önskad total urvalsstorlek när felmarginalen specificieras?

Inlämning

Laborationerna är som sagt ej obligatoriska, men om ni lämnar in denna laboration **inom en vecka** efter labbtillfället kommer den att rättas och kommenteras.