Sommerskole i videregående statistik (2200-B5-5F18)
Juli 2018

Opgave mandag den 16. juli
Af cand. stat. Hans Bay

1) Regn de tre første opgaver i E 1998 (stikprøveteori). Bemærk disse tre opgaver bruger den teoretiske S^2. Du kender altså variationen i universet. Du får altså at vide universets antal er N=931, at stikprøvens størrelse er n=49, at univers gennemsnittet ($\bar{Y}$)=1266,1 og at univers spredningen S=2248,6. **Dette er lommeregneropgave.**
   De to første spørgsmål i E 1998 skal alene besvares ud fra at
   Univers-gennemsnit er 1.266,1
   Univers antal er N=931
   Univers spredning S=2.248,6
   Stikprøve n=49
   Du må meget gerne kommentere størrelsen af S.

Spg. 1) Beregn variansen for denne estimator og udregn et tilsvarende 95% konfindensinterval.

Formlen for varians af en stikprøve anvendes

$$V(\bar{y}) = \frac{N-n}{N}\frac{1}{n}S^2 = \frac{931-49}{931}\frac{1}{49}2248{,}6^2 = 97756{,}85852$$

Dette giver en varians af estimatoren 97756,86. Dernæst udregner vi konfidensintervallet.

$$\text{Konfidens} - \text{interval} = [\bar{x} + 1{,}96 \cdot \sigma; \bar{x} - 1{,}96 \cdot \sigma]$$

$$= \left[1266{,}1 + 1{,}96 \cdot \sqrt{97756{,}85}; 1266{,}1 - 1{,}96 \cdot \sqrt{97756{,}85}\right] = [1878{,}915\ ;\ 653{,}28]$$

Dermed kan man sige med 95% sikkerhed at middelværdien af stikprøven ligger mellem 1878,9 og 653,28.

Spg. 2) Angiv hvor stor stikprøven skal være, hvis længden af intervallet skal være på højst 10%.

Vi starter med at finde 10% af middelværdien

$$L_0 = 0{,}1 \cdot 1266{,}1 = 126{,}1$$

Dermed må den maksimalt gå $\frac{126{,}1}{2} = 63{,}05$ over eller under.

Først findes n_0 via følgende formel

$$n_0 = \frac{S^2}{\left(\frac{L_0}{(2 \cdot 1{,}96)}\right)^2 + \frac{S^2}{N}} = \frac{2248{,}6^2}{\left(\frac{126{,}1}{(2 \cdot 1{,}96)}\right)^2 + \frac{2248{,}6^2}{931}} = 782$$

Dermed skal stikprøven være på 782 før at stikprøvens middelværdi maksimalt svinger med 10 % omkring middelværdien.

Spg. 3) Betragt tabel 1, angiv den proportionale og den optimale allokering af en stikprøve af samlet størrelse på 49 samt estimatorernes standardafvigelse. Kommenter forskellen mellem de to allokeringer og sammenlign med simpel tilfældig udvælgelse.

stratum	N_k	S_k	W	W*S^2	W*S
	Antallet i det enkelte stratum	Spredningen i det enkelte stratum			
HR	91	4710,60	0,10	2168923,16	460,43
Øvr. Sjælland	214	1732,70	0,23	690098,12	398,28
Fyn	105	940,30	0,11	99717,75	106,05
Jylland	521	1832,30	0,56	1878802,83	1025,38
i alt	931	312,66	1,00	97757,00	312,66

Nu udregnes den spredning for den proportionale allokering

$$V(\bar{y}_p) = \frac{931-49}{931} \cdot \frac{1}{49} \cdot 4837541{,}9 = 0{,}019 \cdot 4837541{,}9 = 93.529{,}27$$

$$S(\bar{y}_p) = \sqrt{93529{,}27} = 305{,}83$$

Nu udregnes den spredning for den optimale allokering

$$V(\bar{y}_p) = \frac{931-49}{931} \cdot \frac{1}{49} \cdot 1901^2 - \frac{1}{931} \cdot 4837541{,}9 = 80829 - 5196 = 75633$$

$$S(\bar{y}_p) = \sqrt{75633} = 275{,}01$$

Den optimale allokering har en lavere spredning end den propertionale. Dette skyldes helt mekanisk den formel som vælger at vægte spredning i de forskellige grupper.

2) Udregn DEFF (=DEsign EFfekt) som er forholdet mellem variansen for det aktuelle design og variansen for en simpeltilfældig udvælgelse. I E 1998 sp3 er der to design, for hvert design skal du udregne DEFF. Hvordan vil du fortolke de to DEFF størrelser?

Vi udregner DEFF for begge allokeringstyper via følgende udtryk

$$DEFF = \frac{V(\bar{y}_i)}{V(\bar{y})}$$

Dernæst indsættes den fundne varians fra 1.3

$$DEFF_p = \frac{93529{,}27}{97756{,}85} = 0{,}96$$

$$DEFF_{opt} = \frac{75633}{97756{,}85} = 0{,}77$$

Jo tættere på 1 DEFF er jo tættere er allokeringens varians på den oprindelige varians. Da begge er under nul er variansen faldet og det er faldet mest ved den optimale allokering, hvilket er forventeligt.

3) I Greens opinions måling fra marts 2017 skal du eftervise at konfidensintervallet for liste (A) på 2,4 % er korrekt udregnet. Hvilke forudsætninger har du gjort? Hvorfor er dette konfidensinterval større end konfidensintervallet for liste (F)?

Formlen for varians af en stikprøve anvendes

$$V(\bar{y}) = \frac{N-n}{N}\frac{1}{n-1}P(1-P) = \frac{4100000-1253}{4100000}\frac{1}{1253-1} \cdot 0{,}263 \cdot (1-0{,}263) = 0{,}000154$$

Dette giver en varians af estimatoren 97756,86. Dernæst udregner vi konfidensintervallet.

$$\text{Konfidens} - \text{interval} = [\bar{x} + 1{,}96 \cdot \sigma; \bar{x} - 1{,}96 \cdot \sigma]$$

$$= \left[26{,}3 + 1{,}96 \cdot \sqrt{0{,}000154}; 26{,}3 - 1{,}96 \cdot \sqrt{0{,}000154}\right] = [26{,}32; 26{,}27]$$

Dermed kan man sige med 95% sikkerhed at middelværdien af stikprøven ligger mellem 26,27% og 26,32%. Dette passer med de 2,4% af 26,3 som GREEN fandt.

Jeg har forudsat at populationen er antallet af stemmeberettigede i Danmark og at sandsynligheden for at stemme på A er lig andelen der ville stemme på dem.

Liste A har et større konfidensinterval end liste F, da liste F har en mindre varians

$$\frac{4100000-1253}{4100000}\frac{1}{1253-1}\cdot 0{,}048\cdot(1-0{,}048)=0{,}0000365$$

Dermed har liste F per definition et lavere konfidensinterval.

4) Hvor stor skal stikprøven være hvis konfidensintervallet for liste (A) skal være på 1,2 %?

Vi starter med at finde L_0 som er den tilladte variation på hver side ganget 2.

$$L_0 = 0{,}012\cdot 2 = 0{,}024$$

Dernæst findes n_0 via følgende formel

$$n_0 = \frac{S^2}{\frac{L_0}{(2\cdot 1{,}96)}+\frac{S^2}{N}}$$

Vi mangler at finde spredningen som vi finder via følgende udtryk

$$\hat{S}^2 = S^2 = \frac{n}{n-1}\hat{p}(1-\hat{p}) = \frac{1253}{1253-1}0{,}263(1-0{,}263) = 0{,}19398581709265$$

Denne indsættes i formlen for L_0

$$n_0 = \frac{0{,}19398581709265}{\left(\frac{0{,}024}{(2\cdot 1{,}96)}\right)^2+\frac{0{,}19398581709265}{4100000}} = 5169$$

Dermed skal stikprøven være på 5169 før at stikprøvens middelværdi maksimalt svinger med 1,2 % omkring middelværdien.

5) Test hypotesen at Liste (A), (O) og (V) alle er uforandret i forhold til sidste folketingsvalg.

Vi tester samlet om liste A, O og V er uforandret. Dette gøres via følgende kode

```
*Opgave 5;
Data opgave5;
input Parti $ Andel; *$ betyder at det er en karaktervariable;
datalines;
A 0.263
O 0.211
V 0.195
Andre 0.331
;
run;

proc freq data=opgave5;
table parti/chisq testp=(0.263 0.174 0.188 0.375);
weight andel;
run;
```

Dette giver følgende output:

Chi-Square Test for Specified Proportions	
Chi-Square	0.2309
DF	3
Pr > ChiSq	0.9725
WARNING: 100% of the cells have expected counts less than 5. Chi-Square may not be a valid test.	

Testes nulhypotesen om at der er ingen forskel mellem sidste folketingsvalg og GREEN's måling på et 5% signifikansniveau så finder vi, givet en p-værdi på 97,5%, at nulhypotesen ikke kan afvises og der dermed ikke er nogen forskel mellem sidste folketingsvalg og GREEN's måling.

De følgende spørgsmål skal helst løses vha. surveyselect, og (i sp. 7) surveymeans, i SAS. (Ikke alle spørgsmål kan alene løses i disse procedurer).
Endvidere skal der så vidt muligt beregnes både den teoretiske standardafvigelse og den estimerede standardafvigelse på estimaterne.

6) Hvor mange respondenter har fået værdien 5.5? Hvad mener du om denne form for erstatning af missings værdier (imputation)?

Vi importerer data og danner 6 strata via følgende kode med 1502 observationer:

```
proc import datafile='/courses/d284cd65ba27fe300/Sommerskole
2018/Uge 2/ESS7e02_1' out=ud7 dbms=sav replace;
data dk7;
set ud7;
where cntry='DK';
*** nedenstående udarbejdes 6 strata baseret på køn og alder ****;
*** missing for trsetep og trstun = midtpunkt af skala ***;
if agea < 40 then age1=1;
if 40<= agea< 70 then age1=2;
if agea =>70 then age1=3;
if age1=1 and gndr=1 then strat=1;
if age1=2 and gndr=1 then strat=2;
if age1=3 and gndr=1 then strat=3;
if age1=1 and gndr=2 then strat=4;
if age1=2 and gndr=2 then strat=5;
if age1=3 and gndr=2 then strat=6;
if trstep=. then trstep=5.5;
if trstun=. then trstun=5.5;
run;
proc freq data=dk7;
table strat;
run;
```

Dernæst kigger vi på hvor mange der har fået værdi 5,5 via følgende kode:

```
proc freq data=dk7;
table trstep trstun;
run;
```

Og der fås følgende output:

Trust in the European Parliament				
trstep	Frequency	Percent	Cumulative Frequency	Cumulative Percent
No trust at all	108	7.19	108	7.19
1	43	2.86	151	10.05
2	127	8.46	278	18.51
3	128	8.52	406	27.03
4	163	10.85	569	37.88
5	281	18.71	850	56.59
5.5	79	5.26	929	61.85
6	203	13.52	1132	75.37
7	188	12.52	1320	87.88
8	127	8.46	1447	96.34
9	37	2.46	1484	98.80
Complete trust	18	1.20	1502	100.00

Trust in the United Nations				
trstun	Frequency	Percent	Cumulative Frequency	Cumulative Percent
No trust at all	27	1.80	27	1.80
1	17	1.13	44	2.93
2	48	3.20	92	6.13
3	75	4.99	167	11.12
4	99	6.59	266	17.71
5	220	14.65	486	32.36
5.5	58	3.86	544	36.22
6	193	12.85	737	49.07
7	302	20.11	1039	69.17
8	276	18.38	1315	87.55
9	109	7.26	1424	94.81
Complete trust	78	5.19	1502	100.00

Heraf ses det at 79 har fået værdi trstep=5,5 og 58 har fået værdi trstun=5,5. Dette virker som en meget forsigtig måde af erstatte folk på, da 5,5 ligge i midten af mulighederne og dermed ikke trækker i nogen retning.

7) Udtag en simpel tilfældig stikprøve på 120 personer (dvs. n=120) og beregn den gennemsnitlige tiltro med "Europa Parlamentet" samt tilhørende usikkerhed. Gentag denne procedure 10 gange (dvs. vælg et nyt seed 10 gange). Hvad skal gennemsnittet af gennemsnittene blive teoretisk. Hvad skal gennemsnittet af stikprøvernes varians teoretisk blive.

Vi starter med at kigge på gennemsnittet af hele populationen først. Dette gøres via følgende kode:

```
proc means data=ud61 n mean std maxdec=2 ;
var trstep;
run;
```

Som giver følgende output

The MEANS Procedure

Analysis Variable : trstep Trust in the European Parliament		
N	Mean	Std Dev
36878	4.19	2.53

Dernæst udtrækkes 10 stikprøver via følende kode:

```
*1.;
proc surveyselect data=ud61 seed=101 n=120 out=ud1;
run;
proc surveymeans data=ud1;
var trstep;
run;
*2.;
proc surveyselect data=ud61 seed=102 n=120 out=ud2;
run;
proc surveymeans data=ud2;
var trstep;
run;
*3.;
proc surveyselect data=ud61 seed=103 n=120 out=ud3;
run;
proc surveymeans data=ud3;
```

```
var trstep;
run;
*4.;
proc surveyselect data=ud61 seed=104 n=120 out=ud4;
run;
proc surveymeans data=ud4;
var trstep;
run;
*5.;
proc surveyselect data=ud61 seed=105 n=120 out=ud5;
run;
proc surveymeans data=ud5;
var trstep;
run;
*6.;
proc surveyselect data=ud61 seed=106 n=120 out=ud6;
run;
proc surveymeans data=ud6;
var trstep;
run;
*7.;
proc surveyselect data=ud61 seed=107 n=120 out=ud7;
run;
proc surveymeans data=ud7;
var trstep;
run;
*8.;
proc surveyselect data=ud61 seed=108 n=120 out=ud8;
run;
proc surveymeans data=ud8;
var trstep;
run;
*9;
proc surveyselect data=ud61 seed=109 n=120 out=ud9;
run;
proc surveymeans data=ud9;
var trstep;
run;
*10.;
proc surveyselect data=ud61 seed=110 n=120 out=ud10;
run;
proc surveymeans data=ud10;
var trstep;
run;
```

Dette giver følgende output:

Statistics						
Variable	Label	N	Mean	Std Error of Mean	95% CL for Mean	
trstep	Trust in the European Parliament	116	4.163793	0.236461	3.69541007	4.63217614

Statistics						
Variable	Label	N	Mean	Std Error of Mean	95% CL for Mean	
trstep	Trust in the European Parliament	102	4.127451	0.266076	3.59962841	4.65527355

Statistics						
Variable	Label	N	Mean	Std Error of Mean	95% CL for Mean	
trstep	Trust in the European Parliament	110	3.872727	0.247459	3.38227189	4.36318265

Statistics						
Variable	Label	N	Mean	Std Error of Mean	95% CL for Mean	
trstep	Trust in the European Parliament	112	4.776786	0.225009	4.33091561	5.22265582

Statistics						
Variable	Label	N	Mean	Std Error of Mean	95% CL for Mean	
trstep	Trust in the European Parliament	108	3.981481	0.229746	3.52603682	4.43692614

Statistics						
Variable	Label	N	Mean	Std Error of Mean	95% CL for Mean	
trstep	Trust in the European Parliament	111	4.027027	0.266460	3.49896660	4.55508746

Statistics						
Variable	Label	N	Mean	Std Error of Mean	95% CL for Mean	
trstep	Trust in the European Parliament	109	3.880734	0.233793	3.41731500	4.34415289

Statistics						
Variable	Label	N	Mean	Std Error of Mean	95% CL for Mean	
trstep	Trust in the European Parliament	111	4.198198	0.239717	3.72313494	4.67326146

Statistics						
Variable	Label	N	Mean	Std Error of Mean	95% CL for Mean	
trstep	Trust in the European Parliament	111	4.243243	0.231513	3.78443967	4.70204682

Statistics						
Variable	Label	N	Mean	Std Error of Mean	95% CL for Mean	
trstep	Trust in the European Parliament	113	4.212389	0.241954	3.73298906	4.69178970

Teoretisk set skal gennemsnittet af de enkelte gennemsnit blive 4,19 som er gennemsnittet for populationen. Hvis stikprøverne er trykket tilfældigt og uafhænigigt burde variansen bliver den samme som populationen.

8) Forklar hvad der sker i ovenstående program. Hvordan vil du estimere den gennemsnitlige tillid til EU parlamentet (TRSTEP) ud fra data i udtrak1? Forklar hvorfor dette gennemsnit ikke afviger meget fra gennemsnittet i forrige opgave.

Derudføres en tilfældig stikprøve på 120 personer, som gentages 10 gange via følgende kode:

```
proc sort data=ud61;
by strat;
run;
proc surveyselect data=yd61 n=20 out=udtrak1 seed=1997;
strata strat;
run;
```

Dette giver følgende output:

Selection Method	Simple Random Sampling
Strata Variable	strat

Input Data Set	DK7
Random Number Seed	1997
Stratum Sample Size	20
Number of Strata	6
Total Sample Size	120
Output Data Set	UDTRAK1

Dette program sorteret det importede data efter strata og udtrækker dernæst en stikprøve opdelt på strata som hver har 20 observationer med seednummer 1997 og gemmer dataen i udtrak1.

Gennemsnittet af TRSTEP estimeres ved via en proc means:

```
proc surveymeans data=udtrak1;
var trstep;
run;
```

Dette giver følgende output:

Statistics						
Variable	Label	N	Mean	Std Error of Mean	95% CL for Mean	
trstep	Trust in the European Parliament	120	4.704167	0.221126	4.26631418	5.14201915

Da dette er en tilfældigt udtrukket stikprøve af tillid til EU, så er det forventeligt at denne tilnærmer sig populations gennemsnit af tilliden.

9) Skriv et program der udtrækker en stikprøve på n=120 der er proportionalt allokeret. Hvordan kan estimatet udregnes?

Skriver et program der udtrækker en stikprøve med n=120:

```
proc surveyselect data=dk7 method=srs seed=110 n=120 out=udtrak2;
```

```
strata strat;
run;
proc surveymeans data=udtrak2;
var trstep;
run;
```

Estimatet udregnes med surveymeans, som tidligere. Dog her kan estimatet for monenterne i det samlede dataudregnes som momentet og giver følgende output:

Statistics

Variable	Label	N	Mean	Std Error of Mean	95% CL for Mean	
trstep	Trust in the European Parliament	480	4.985417	0.105290	4.77852896	5.19230437

10) Forklar hvad der sker i ovennævnte program

```
proc sort data=dk7;
by strat;
proc means data=dk7;
by strat;
var trstep;
output out=antal n=_TOTAL_;
data antal;
set antal;
drop _FREQ_ _TYPE_;
proc print data=antal;
sum _TOTAL_;
run;
```

Dette program sorteret data i datasættet ”dk7” med hensyn til strata. Deres finder den samtlige statistiske momenter og sorteret dem efter strata og ift. Variable trstep og gemmer outputet som ”antal” n=_TOTAL_ angiver antallet og gemmer det som totalt. Dernæst dannes et nyt datasæt antal hvor den slettet freq_type. Dernæst printer den data fra sættet ”antal” med summen ift. ”_TOTAL”.