Hjemmeopgave 6

```
proc import datafile='/courses/d284cd65ba27fe300/Sommerskole
2018/Uge 2/ESS7e02_1' out=ud7 dbms=sav replace;
data dk7;
set ud7;
where cntry='DK';
if 1<=PRTVTcDK <=9;
if 1<= PRTCLcDK<=9;
stemte_A=0;
if prtvtcdk=1 then stemte_A =1;
vil_A=0;
if PRTCLcDK=1 then vil_A=1;
proc freq data=dk7;
table prtvtcdk;
table stemte_A*vil_A/norow nocol nopercent;
run;
```

1) **Forklar hvad der sker i programstumpen, se også tabellerne i bilaget**

Dnne opgave importerer data fra ESS7e02_1 og genererer datasættet dk7 på baggrund af ESS7e02_1. Dernæst vælges kun data fra Danmark. Dernæst genereres en dummy for om man stemte på Socialdemokratiet eller ej og en for at om man føler sig tættest på socialdemokratiet eller ej. Dernæst printer den data for hvad folk stemte ii en frekvenstabel. Der laves også en tabel der viser frekvenserne for de to dummyer.

2) **Der er 188 personer, der har stemt på (A) svarende til en andel på 21,8%. Ved folketingsvalget i 2011 fik (A) 24,8% af stemmerne. Er datasættet repræsentativt mht. (A)?**

I stikprøven ses det, at Socialdemokratiet er underrepræsenteret da der er en forskel på 24,8-21,8=3%-point mellem de to. Umiddelbart er datasættet ikke repræsentativt netop pga. af socialdemokraterne er underrepræsenteret.

Tabel1: Folketingsvalget i 2011 inddelt efter tilslutning til (A).

stratum	antal	Univers-vægte (W_k)	Stikprøve	Stikprøve-vægte (w_k)
stemte på A i 2011	879.615	0,248	188	0,218
stemte på andet i 2011	2.665.753	0,752	675	0,782
I alt	3.545.368	1	863	1

3) Udstyr datasættet på 863 med politiske vægte, der gør datasættet "repræsentativt" mht. stemmeafgivningen på A ved folketingsvalget i 2011.

vi udstyrer datasættet med vægtet efter formlen

$$Vægte = \frac{Univers\_vægt}{Stikprøve\_vægt}$$

Dette gøres via følgende kode:

```
data dkw7;
set dk7;
if stemte_A=1 then vgt=0.248/0.218;
if stemte_A=0 then vgt=0.752/0.782;
run;
```

Vægtene er altså forholdet mellem hvor mange pct. der stemte på A ved sidste folketingsvalg og hvor mange der siger de stemte på dem i stikprøven.

4) Brug variablen prtclcdk som udtryk for hvad respondenter vil stemme på ved det kommende valg. Lav en prognose for A vil det kommende valg.

Vi bruger følgende kode, hvor vi bruger vægtene fra 3), for at tage højde for den skæve stikprøve:

```
proc freq data=dkw7;
table PRTCLcDK;
weight vgt;
run;
```

Dette giver følgende output:

Which party feel closer to, Denmark

prtclcdk	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Socialdemokraterne - the Danish social democrats	191.7982	22.23	191.7982	22.23
Det Radikale Venstre - Danish Social-Liberal Party	70.55144	8.18	262.3497	30.40
Det Konservative Folkeparti - Conservative	45.37291	5.26	307.7226	35.66
SF Socialistisk Folkeparti - Socialist People's Party	76.76745	8.90	384.49	44.55
Dansk Folkeparti - Danish peoples party	142.1588	16.47	526.6488	61.03
Kristendemokraterne - Christian democrats	9.792346	1.13	536.4411	62.16
Venstre, Danmarks Liberale Parti - Venstre	217.5059	25.20	753.947	87.37
Liberal Alliance - Liberal Alliance	37.67981	4.37	791.6268	91.73
Enhedslisten - Unity List - The Red-Green Alliance	71.34959	8.27	862.9764	100.00

Det ses, at givet vægtene og data fra hvilket parti de føler sig tættest på at Socialdemokratiet står til 22,23% af stemmerne ifølge denne prognose.

5) Brug vægtvariablen for variablene TRSTEP, TRSTPRL, PRTVTcDK og TVTOT. Beregn deres vægtede gennemsnit og sammenlign med de tilsvarende uvægtede resultater.

Vi kører følgende kode:

```
proc freq data=dkw7;
table TRSTEP TRSTPRL PRTVTcDK TVTOT;
weight vgt;
run;
proc freq data=dkw7;
table TRSTEP TRSTPRL PRTVTcDK TVTOT;
```

```
run;
```

Dette giver følgende output, hvor vi har uvægtet til nederst og vægtet øverst:

Variable	Label	N	Mean	Std Dev	Minimum	Maximum
trstep	Trust in the European Parliament	829	4.8508722	2.3469011	0	10.0000000
trstprl	Trust in country's parliament	861	6.1548495	2.3682909	0	10.0000000
prtvtcdk	Party voted for in last national election, Denmark	863	4.4243230	2.6965653	1.0000000	9.0000000
tvtot	TV watching, total time on average weekday	862	4.1730129	1.9477797	0	7.0000000

Variable	Label	N	Mean	Std Dev	Minimum	Maximum
trstep	Trust in the European Parliament	829	4.8419783	2.3437395	0	10.0000000
trstprl	Trust in country's parliament	861	6.1416957	2.3744554	0	10.0000000
prtvtcdk	Party voted for in last national election, Denmark	863	4.5608343	2.6598668	1.0000000	9.0000000
tvtot	TV watching, total time on average weekday	862	4.1531323	1.9463735	0	7.0000000

Det ses at tilliden til EU-parlamentet falder marginalt, hvilket ligeledes gælder tilliden til det nationale parlament. Det giver ikke mening at kommentere på ændringer i party voted, da dette er indeks variabel. Det ses at antallet af timer tv set er faldet.

6) Lav en tabel for Liberal alliance (LA) som er tilsvarende den i tabel1.

stratum	antal	Univers-vægte (W_k)	Stikprøve	Stikprøve-vægte (w_k)
stemte på LA i 2011	176.585	0,052	33	0,038
stemte på andet i 2011	3.368.783	0,948	830	0,962
I alt	3.545.368	1	863	1

7) Er LA repræsentativt repræsenteret i datasættet?

Dette er ligesom ved socialdemokraterne ikke tilfældet. Det ses at 3,8% i stikprøven stemte på Liberale Alliance mens i populationen stemte 5,2% på Liberale Alliance. Dette er en afvigelse 1,4% fra population til stikprøve

8) Udfør en regressions analyse hvor ESCS forklarer score_n. Forklar betydningen af hældningskoefficienten

Vi kører kode som importet PISA dataen og kører en simpel lineære regression:

```
proc import
datafile="/home/nielseriksen0/my_courses/anders.milhoj/Sommerskole
2018/Uge 2/min_PISA_renset.sav" out=ud15 dbms = sav replace;
  data dk2015;
  set ud15;
  where cnt='DNK';
national=0;
if 2<=immig<=3 then national=1;
```

```
run;
proc reg data=dk2015;
model score_n=ESCS;
run;
```

Det giver følgende parameterestimater:

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	472.04564	1.11659	422.76	<.0001
ESCS	Index of economic, social and cultural status (WLE)	1	35.39603	1.06410	33.26	<.0001

Det ses, at estimatet af socioøkonomiske indeks øger scoren med 35 når denne øges med 1. Det er ligeledes meget signifikant.

9) Brug PROC MI til at "impute" 5 værdier pr manglende elev.

Vi kører følgende kode til til at impute værdi 5:

```
proc mi data=dk2015 nimpute=5 seed=100 out=dk2015a;
em itprint outem=dk2015b;
var ESCS score_n;
run;
```

Det giver følgende output:

Missing Data Patterns						
					Group Means	
Group	ESCS	score_N	Freq	Percent	ESCS	score_N
1	X	X	6985	97.54	0.446890	487.863789
2	.	X	176	2.46	.	426.116238

Vi ser, at 176 observationer mangler ESCS information.

10) Brug PROC MIANALYZE til at estimere hældningskoefficienten

Vi bruger MIANALYSE til at estimere et nyt parameterestimat, som gøres via følgende kode:

```
proc reg data=dk2015a outest=dk2015c covout noprint;
model score_n=ESCS;
by _Imputation_;
run;
proc mianalyze data=dk2015c edf=35804;
modeleffects ESCS;
run;
```

Hvor antallet af frihedsgrad fås via følgende kode og n-1:

```
proc means data=dk2015a n;
var ESCS;
run;
```

Dette giver følgende output:

Parameter Estimates (5 Imputations)										
Parameter	Estimate	Std Error	95% Confidence Limits		DF	Minimum	Maximum	Theta0	t for H0: Parameter=Theta0	Pr > |t|
ESCS	35.510182	1.058165	33.43609	37.58427	20419	35.349512	35.576742	0	33.56	<.0001

Det ses, at parameterestimatet eller hældningskoefficienten 35,51, som siger, at scoren stiger med 35,5 når det socioøkonomiske stiger med en. En lille stigning ift. før. Det ses samtidigt, at dette resultat er signifikant. Det ses ligeledes, at estimatet er faldet. Proceduren MI har "udfyldt" de manglende svar og proc mianalyze har kørt en regression. Dette antager dog, at der ikke er en grund til at de ikke har svaret.

11) Brug PROC SURVEYIMPUTE til at erstatte de manglende værdier for ESCS

VI anvender en følgende kode:

```
proc surveyimpute data=dk2015 method=hotdeck(selection=srswor);
var ESCS;
output out=dk2015d;
run;
```

Dette giver følgende output:

Imputation Summary		
Observation Status	Number of Observations	Sum of Weights
Nonmissing	6985	6985
Missing	176	176
Missing, Imputed	176	176
Missing, Not Imputed	0	0

12) Gentag regressionsanalysen

Vi kører samme regression som i opgave 9 med følgende kode:

```
proc reg data=dk2015d;
model score_n=ESCS;
run;
```

Dette giver følgende output:

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	470.81786	1.11227	423.29	<.0001
ESCS	Index of economic, social and cultural status (WLE)	1	34.57270	1.06089	32.59	<.0001

Parameteren for ESCS estimeres til 34,57, altså lidt lavere men meget tæt på de andre estimater fra 12) og 10).

13) Sammenlign resultaterne

Her indsættes resultat fra opgave 8 til sammenligning:

Variable	Opgave 8	Opgave 10	Opgave 12
ESCS	35,39	35,51	34,57

Standardfejl	1,06410	1,05817	1,06089

Det ses, at estimatet fra opgave 12 har det laveste estimat for effekten af ESCS på pisascoren, hvor de manglende værdier har fået tildelt en værdi. Estimatet fra opgave 10 er det højeste hvor de manglende værdier har fået tildelt en tilfældig værdi 5 gange og gennemsnittet af denne er taget.
Det ses ift. standardfejlene at opgave 10 giver det mest efficiente estimat da denne har den laveste standardfejl, dernæst kommer opgave 12 estimat, dog er forskellen generelt meget lille og det virker til ikke at have haft den store indflydelse på estimatet.

14) Udarbejd vægte til det danske PISA datasæt

Vi udregner vægten via nedenstående formel og med data fra tabellen med strata

$$Vægte = \frac{Univers\_vægt}{Stikprøve\_vægt}$$

Vi får følgende resultater:
For DNK - stratum 1 fås:

$$Vægt_1 = \frac{0,065}{0,249} = 0,2610441767$$

For DNK - stratum 2 fås:

$$Vægt_2 = \frac{0,225}{0,262} = 0,858778626$$

For DNK - stratum 3 fås:

$$Vægt_3 = \frac{0,521}{0,370} = 1,4081081081$$

For DNK - stratum 4 fås:

$$Vægt_4 = \frac{0,189}{0,119} = 1,5882352941$$

15) Udregn Danmarks gennemsnitlige score for naturfag.

Vi anvender følgende kode

```
proc means data=dk2015 n mean std max min;
var score_n;
run;
```

Dette giver følgende output:

Analysis Variable : score_N score_N				
N	Mean	Std Dev	Maximum	Minimum
7161	486.3461845	91.1326897	810.1034000	207.5794000

Det ses at det simple uvejede gennemsnit giver en gennemsnitlig score i naturfag på 486,34 med en std. afv. på 91,132 på hver side af middelværdien. Givet måden stikprøven er udtaget på er den meget ikke-repræsentativ og giver dermed et skævt resultat, som tages højde for i opgave 16.

16) Sammenlign med beregninger, hvor du bruger vægtvariablen W_fstuwt (final student weight) som er i datamaterialet.

Vi anvender følgende kode

```
proc means data=dk2015 n mean std max min;
var score_n;
weight W_fstuwt;
run;
```

Dette giver følgende output:

Analysis Variable : score_N score_N				
N	Mean	Std Dev	Maximum	Minimum
7161	501.7617893	252.5068200	810.1034000	207.5794000

Når vi vægter resultatet efter vægten, som er givet i datamaterialet ser vi, at den gennemsnitlige score i naturfag stiger fra 486,35 til 501,76. Dette skyldes, at antallet af individer med en anden etnicitet end dansk er meget højere i stikprøven end i populationen. Disse klarer sig typisk ringere end etniske danskere og dermed giver stikprøven et skævt billede hvilket der tages højde for med en vægtning ift til populationens, som giver et mere retvisende billede af danske folkeskole elevers evner.