# Hjemmeopgave 5

**1.del**

Brug ESS datasættet for 2014 udtag de danske data. Dette datasæt bestående af 1.502 personer er nu dit univers. Dvs. at N=1.502.

Der konstrueres 6 strata som er defineret ved køn og alder således:

1) Sammenlign gennemsnittene for TRSTEP og TRSTPRL for de to datasæt DK7 og DK7a.

vi importerer data via følgende kode:

```
proc import datafile='/courses/d284cd65ba27fe300/Sommerskole
2018/Uge 2/ESS7e02_1' out=ud7 dbms=sav replace;
data dk7;
set ud7;
where cntry='DK';
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;
run;
proc standard data=dk7 replace out=dk7a;
var trstep trstprl;
run;
```

Vi vil gerne have gennemsnittene for pågældende variable TRSTEP og TRSTPRL for hhv. dk7 og dk7a. Dette gøres via følgende kode i SAS:

```
proc means data=dk7a;
var trstep trstprl;
run;
proc means data=dk7;
var trstep trstprl;
run;
```

Dette giver følgende output fra dk7a:

| Variable | Label | N | Mean | Std Dev | Minimum | Maximum |
|---|---|---|---|---|---|---|
| trstep | Trust in the European Parliament | 1502 | 4.7793394 | 2.3482190 | 0 | 10.0000000 |
| trstprl | Trust in country's parliament | 1502 | 5.9087860 | 2.4311430 | 0 | 10.0000000 |

Det bemærkes klart, at tilliden til det nationale parlament er større end tilliden til EU parlamentet og har ligeledes en større spredning i std. dev. Dette er ikke overraskende da folk muligvis har større tillid til personer og institutioner tættere på dem end længere væk.

Dette giver følgende output fra dk7:

| Variable | Label | N | Mean | Std Dev | Minimum | Maximum |
|---|---|---|---|---|---|---|
| trstep | Trust in the European Parliament | 1423 | 4.7793394 | 2.4125656 | 0 | 10.0000000 |
| trstprl | Trust in country's parliament | 1491 | 5.9087860 | 2.4401005 | 0 | 10.0000000 |

Sammenlignes dk7a og dk7, ses det, at de har samme middelværdi, hvilket er forventeligt.

**Brug datasættet DK7a.**

2) Udtag en simpel tilfældig stikprøve på 100 personer og beregn den gennemsnitlige tiltro med "Europa Parlamentet" samt tilhørende usikkerhed.

Vi udtrækker en tilfældig stikprøve og finder gennemsnittet via følgende kode:

```
proc surveyselect data=dk7a seed=100 n=100 out=stik1;
run;
proc surveymeans data=stik1;
var trstep trstprl;
run;
```

Dette giver følgende output:

| Statistics | | | | | | |
|---|---|---|---|---|---|---|
| Variable | Label | N | Mean | Std Error of Mean | 95% CL for Mean | |
| trstep | Trust in the European Parliament | 100 | 4.966760 | 0.233589 | 4.50326918 | 5.43025155 |
| trstprl | Trust in country's parliament | 100 | 6.119088 | 0.222609 | 5.67738424 | 6.56079148 |

Af ovenstående tabel fremgår det, at begge variable har en større middelværdi og dermed en højere gennemsnitlig tiltro til begge parlamenter. Usikkerheden er faldet da standardfejlen på middelværdien er faldet for begge middelværdier.

Den tilhørende usikkerhed er 2 gange standardafvigelsen. Først med EU parlamentet:

$$0,234 \cdot 2 = 0,468$$

Dette medfører at middelværdiens usikkerhed er $\pm 0,468$. Dernæst med det nationale parlament:

$$0,223 \cdot 2 = 0,446$$

Dette medfører at middelværdiens usikkerhed er $\pm 0,446$

3) Udtag en simpel tilfældig stikprøve på 100 personer. Beregn den gennemsnitlige tiltro til Europarlamentet i forhold til den gennemsnitlige tiltro til det nationale parlament (brøk estimation). Angiv den tilhørende usikkerhed.

Der udtages en ny tilfældig stikprøve og brøk estimation udføres via følgende kode:

```
proc surveyselect data=dk7a seed=101 n=100 out=stik2;
run;
proc surveymeans data=stik2;
var trstep trstprl;
ratio trstep/trstprl;
run;
```

Dette giver følgende output:

| Statistics | | | | | | |
|---|---|---|---|---|---|---|
| Variable | Label | N | Mean | Std Error of Mean | 95% CL for Mean | |
| trstep | Trust in the European Parliament | 100 | 4.980141 | 0.210435 | 4.56259272 | 5.39768837 |
| trstprl | Trust in country's parliament | 100 | 6.200000 | 0.233117 | 5.73744573 | 6.66255427 |

| Ratio Analysis | | | | | | |
|---|---|---|---|---|---|---|
| Numerator | Denominator | N | Ratio | Std Error | 95% CL for Ratio | |
| trstep | trstprl | 100 | 0.803248 | 0.032988 | 0.73779216 | 0.86870479 |

Det ses af brøk estimationen giver et estimat på 0,803, som betyder at tilliden til det EU-parlamentet er 80% af hvad den er til det nationale parlament. Den tilhørende usikkerhed er:

$$0,329 \cdot 2 = 0,658$$

Dette medfører at middelværdiens usikkerhed er $\pm 0,658$

og hvis det af ratio-estimation så bliver usikkerheden

$$0,033 \cdot 2 = 0,066$$

Dette medfører at middelværdiens usikkerhed er $\pm 0,066$

4) Gentag ovenstående 10 gange.

Vi kører en loop, som kører koden fra opgave 3. 10 gange:

```
proc surveyselect data=dk7a seed=102 n=100 out=stik3 reps=10;
run;
proc surveymeans data=stik3;
var trstep trstprl;
ratio trstep/trstprl;
run;
```

Dette giver følgende output

| Statistics | | | | | | |
|---|---|---|---|---|---|---|
| Variable | Label | N | Mean | Std Error of Mean | 95% CL for Mean | |
| trstep | Trust in the European Parliament | 1000 | 4.701629 | 0.076204 | 4.55209042 | 4.85116749 |
| trstprl | Trust in country's parliament | 1000 | 5.847362 | 0.076175 | 5.69788102 | 5.99684198 |

| Ratio Analysis | | | | | | |
|---|---|---|---|---|---|---|
| Numerator | Denominator | N | Ratio | Std Error | 95% CL for Ratio | |
| trstep | trstprl | 1000 | 0.804060 | 0.011210 | 0.78206147 | 0.82605834 |

Det bemærkes at ratio-estimatet er steget minimalt, men at standardfejlen er faldet markant, som skyldes den større stikprøven. Den tilhørende usikkerhed af ratio-estimation bliver

$$0,011 \cdot 2 = 0,022$$

Dette medfører at middelværdiens usikkerhed er $\pm 0,022$

6) Betragt det store datasæt ud7. I dette datasæt skal du teste om ”trusten til EP” er den samme i de tre lande: Danmark, Belgien og England [cntry-værdierne er ’DK’, ’BE’ og ’GB’].

Vi udfører et test af om tilliden til EP er den samme på tværs af de tre lande. Dette gøres via følgende kode:

```
Data opgave6;
set ud;
where cntry='DK' or cntry='BE' or cntry='GB';
run;
Proc anova data=opgave6;
class cntry;
model trstep=cntry;
run;
```

Dette giver følgende output

| Source | DF | Anova SS | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| cntry | 2 | 3768.875661 | 1884.437831 | 318.98 | <.0001 |

Vi tester nulhypotesen om at der ikke er nogen forskel i tillid mellem de 3 lande. Denne hypotese testes på et 5% signifikansniveau og det ses, at en p-værdi på under 0,1% at denne nulhypotese kan forkastes og der er dermed forskellig tillid til EU-parlamentet i de 3 lande.

7) I Tyskland er der 81 millioner personer (i Danmark er der 5 millioner personer). Hvor stor skal den tyske stikprøve være, hvis usikkerheden på tilliden til EP skal være den samme som i Danmark [cntry=’DE’]?

Vi finder først spredningen for begge via følgende kode:

```
proc  surveymeans data=dk7a total=5000000;
var trstep trstprl;
run;
proc means data=ud n mean std;
var trstep trstprl;
where cntry='DE';
run;
```

Dette giver følgende output:

| Statistics | | | | | | |
|---|---|---|---|---|---|---|
| Variable | Label | N | Mean | Std Error of Mean | 95% CL for Mean | |
| trstep | Trust in the European Parliament | 1000 | 4.701629 | 0.076197 | 4.55210537 | 4.85115253 |
| trstprl | Trust in country's parliament | 1000 | 5.847362 | 0.076167 | 5.69789597 | 5.99682703 |

| Variable | Label | N | Mean | Std Dev |
|---|---|---|---|---|
| trstep | Trust in the European Parliament | 2921 | 3.9856214 | 2.4348635 |
| trstprl | Trust in country's parliament | 3012 | 4.9960159 | 2.4039125 |

Dernæst udregner vi variansen for middelværdien til EU for begge lande

$\sigma^2_{DK} = 0,076167^2 = 0,0058$
$\sigma^2_{DE} = 2,43486^2 = 5.92855$
Vi anvender dernæst følgende formel

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

Vi mangler $L_0$ som udregnes som længden af konfidensintervallet, som er den længde som DK har, som ønskes for tyskland

$$L_0 = 4,8511 - 4,5521 = 0,299$$

Dernæst kan alle data indsættes i formlen

$$n_0 = \frac{5.92855}{(\frac{0,299}{(2 \cdot 1,96)})^2 + \frac{5.92855}{81.000.000}} = 1018,997$$

Det ses at for at få samme konfidensintervalslængde skal stikprøvestørrelsen for Tyskland være 1.019 individer. Dette virker forkert givet at den burde være omkring 1500. Dette skyldes at længden af L0 er for stor.

8) Fra datasættet ESS7 og ESS8 udtages England [cntry='GB'] for hvert datasæt udregnes forholdet mellem TRSTEP og TRSTPRL. Har der været en signifikant udvikling?

Først udregnes de to ratioer (TRSTEP/TRSTPRL). Den ene for datasættet ESS7 og den anden for ESS8.

```
 *Opgave 8;
*Importerer data;
proc import datafile='/courses/d284cd65ba27fe300/Sommerskole
2018/Uge 2/ESS8e01_1' out=ud8 dbms=sav replace;
proc import datafile='/courses/d284cd65ba27fe300/Sommerskole
2018/Uge 2/ESS7e02_1' out=ud7 dbms=sav replace;
run;
proc surveymeans data=ud7;
where cntry='GB';
var trstep trstprl;
ratio trstep/trstprl;
run;
proc surveymeans data=ud8;
where cntry='GB';
var trstep trstprl;
ratio trstep/trstprl;
run;
```

Det giver følgende output:
For ESS7:

| Ratio Analysis | | | | | | |
|---|---|---|---|---|---|---|
| Numerator | Denominator | N | Ratio | Std Error | 95% CL for Ratio | |
| trstep | trstprl | 2078 | 0.711334 | 0.011105 | 0.68955631 | 0.73311259 |

For ESS8:

| Ratio Analysis | | | | | | |
|---|---|---|---|---|---|---|
| Numerator | Denominator | N | Ratio | Std Error | 95% CL for Ratio | |
| trstep | trstprl | 1854 | 0.767110 | 0.010958 | 0.74561884 | 0.78860214 |

Konfidensintervallet for ratioen i ESS7 (0,6895;0,73311) overlapper ikke konfidensintervallet for ratioen i ESS8 (0,74562;0,78860). Det vil sige, at der har været en signifikant udvikling i forholdet mellem tilliden til EP og det nationale parlament fra ESS7 til ESS8. Tilliden til EP er styrket i forhold til tilliden til det nationale parlament.

Betragt nu hele datasættet (det danske) på i alt 1.502 observationer. Reducer datasættet til dem der har oplyst om stemmeafgivningen ved folketingsvalget i 2011. (Vi har lige haft et folketingsvalg i 2015, dette er selvfølgelig ikke med her).
Brug variablen *prtvtcdk*. Dette skulle give et datasæt på ca. 1.179 personer. Dette datasæt betragtes som en simpel tilfældig stikprøve ud af 3.500.000 vælgere (dem der vælger at stemme)[1]

9) I dette datasæt skal beregnes Radikale Venstres (prtvtcdk=2) andel med tilhørende usikkerhed.

Først laves et nyt datasæt dk8 med de konkrete rensninger ovenfor med følgende kode

```
data dk8;
set ud;
where cntry='DK';
if prtvtcdk=. then delete;
run;
```

Dernæst findes Radikale Venstres andel af de samlede stemmer med følgende kode:

```
proc freq data=dk8;
table prtvtcdk;
run;
```

Dette giver følgende output:

| Party voted for in last national election, Denmark | | | | |
|---|---|---|---|---|
| prtvtcdk | Frequency | Percent | Cumulative Frequency | Cumulative Percent |
| Socialdemokraterne - the Danish social democrats | 268 | 22.73 | 268 | 22.73 |
| Det Radikale Venstre - Danish Social-Liberal Party | 134 | 11.37 | 402 | 34.10 |
| Det Konservative Folkeparti - Conservative | 65 | 5.51 | 467 | 39.61 |
| SF Socialistisk Folkeparti - Socialist People's Party | 108 | 9.16 | 575 | 48.77 |
| Dansk Folkeparti - Danish peoples party | 143 | 12.13 | 718 | 60.90 |
| Kristendemokraterne - Christian democrats | 8 | 0.68 | 726 | 61.58 |
| Venstre, Danmarks Liberale Parti - Venstre | 311 | 26.38 | 1037 | 87.96 |
| Liberal Alliance - Liberal Alliance | 48 | 4.07 | 1085 | 92.03 |
| Enhedslisten - Unity List - The Red-Green Alliance | 77 | 6.53 | 1162 | 98.56 |
| Andet - other | 17 | 1.44 | 1179 | 100.00 |

Det ses her, at det radikale venstre fik 11,37 procent af stemmerne.

Dernæst findes variansen, hvor det antages at andelen der har stemt på et parti kan bruges som sandsynligheden for at stemme på det parti.

$$V(\bar{y}) = \frac{N-n}{N}\frac{1}{n-1}P(1-P) = \frac{3500000-1179}{3500000}\frac{1}{1179-1}\cdot 0,1137\cdot(1-0,1137) = 0,000085516438$$

Dernæst dernæst finder vi konfidensintervallet

$$1,96\cdot\sqrt{0,000085516438} = \pm 0,018125119$$

Dermed vil andelen svinge 1,8%-point omkring middelværdien med 95% sikkerhed.

10) Ved folketingsvalget i 2011 fik Det Radikale Venstre 9,5 % af de afgivne stemmer. Test om Det Radikale Venstres andel i stikprøven på de 1.179 personer kan antages at være på 9,5 %.

Det testes om det i stikprøven kan antages at 9,5% af stikpørven stemmer B. Dette gøres via følgende kode

```
Data opgave10; input Parti $ Andel;
datalines;
B 0.1127
Andre 0.8873
;
run;
proc freq data=opgave10;
table parti/chisq testp=(0.095 0.905);
weight andel;
run;
```

Dette giver følgende output:

| Chi-Square Test for Specified Proportions | |
|---|---|
| Chi-Square | 7.3014 |
| DF | 1 |
| Pr > ChiSq | 0.0069 |
| WARNING: 100% of the cells have expected counts less than 5. Chi-Square may not be a valid test. | |

Vi tester nulhypotesen, som siger at man godt kan antages at 9,5% stemmer B. Denne hypotese testes på et 5% signifikansniveau. Givet et p-værdi på 0,7% kan vi afvise nulhypotesen og dermed kan vi ikke antage at 9,5% af stikprøven stemmer radikalt.

**2. del (lommeregner)**

Regn de tre første opgaver i Eksamen i stikprøveteori 2003.
I opgave 3 skal der stå: Hvor stor skal stikprøvestørrelsen være, når der allokeres optimalt, hvis længden af det tilsvarende 95% konfidensinterval skal være på 0,05.

1) Antag at der er foretaget en stratifikation efter de tre landsdele. Kommenter stratifikationen. Estimer andelen i befolkningen (blandt de 4.200.000), der ikke kan deltage i undersøgelsen pga. sprogvanskeligheder samt beregn den tilsvarende usikkerhed.

| | |
|---|---|
| Hovedstadsregionen | 3/1327=0,00226 |

| | |
|---|---|
| Øvr. Sjælland + øer | 18/770=0,02337 |
| Jylland | 2/1903=0,00105 |
| Hele landet | =0,02668 |

Stratifikationen virker nogenlunde rimelig. Landsdelene minder tilnærmelsesvis om hinanden.
Dernæst udregner vi de 3 usikkerheder som i opgave 8:
Vi starter med hovedstadsregionen:

$$V(\bar{y}) = \frac{N-n}{N}\frac{1}{n-1}P(1-P) = \frac{4200000-4000}{4200000}\frac{1}{4000-1}\cdot 0,02668\cdot(1-0,02668) = 0,0000064874834$$

Dernæst dernæst finder vi konfidensintervallet

$$1,96\cdot\sqrt{0,0000064874834} = \pm\ 0,0049922255678$$

Dermed vil andelen svinge 0,5%-point omkring middelværdien med 95% sikkerhed.
2) Med udgangspunkt i Tabel 1, angiv da den optimalt allokerede stikprøve (stikprøvestørrelsen er stadig 4.000) og udregn den tilsvarende spredning.
Vi starter med at finde den optimale allokering (ses længst til højre)

| | $W_k$ | $n_k$ | Antal med angivelse af sprogvanskeligheder | s2 | s | s2w | sw | yOptimal |
|---|---|---|---|---|---|---|---|---|
| Hovedstadsregionen | 33 | 1.327 | 3 | 0,0022515 | 0,0474497 | 0,0007430 | 1,5658399 | 1.051 |
| Øvr. Sjælland + øer | 19 | 770 | 18 | 0,0223087 | 0,1493608 | 0,0042386 | 2,8378560 | 1.905 |
| Jylland | 48 | 1.903 | 2 | 0,0010489 | 0,0323867 | 0,0005035 | 1,5545608 | 1.044 |
| I alt | 100 | 4.000 | 23 | | | | | |

Dernæst finder vi den tilsvarende spredning

$$V(y_{opt}) = \frac{1}{4000}\cdot\sum sW - \frac{1}{4200000}\cdot\sum s^2W = 0,0088752$$

$$S(y_{opt}) = \sqrt{0,0088752} = 0,0942$$

3) Hvor stor skal stikprøvestørrelsen være, når der allokeres optimalt, hvis længden af det tilsvarende 95% konfidensinterval skal være på 0,001.

Dette betyder at længden af konfidensintervallet er 0,001=L0. Dette kan indsættes i formlen

$$n_0 = \frac{S^2}{(\frac{L_0}{(2\cdot 1,96)})^2 + \frac{S^2}{N}} = \frac{0,0088752}{(\frac{0,001}{(2\cdot 1,96)})^2 + \frac{0,0088752}{4200000}} = 132090$$

Dermed skal stikprøven være på 132090 før at 95% konfidensintervallet har længden 0,001.