Bonus assignment 1

Ludwig Tranheden
ludtra@student.chalmers.se
9406238791
MVE155

2016/01/28

Contents

1 Introduction **2**

2 A **2**
2.1 i - The proportion of female-headed families 3
2.2 ii - The average number of children per family 4
2.3 iii - The proportion of head of households who did not receive a high school diploma, iv - The average family income 5
2.3.1 (iii) 5
2.3.2 (iv) 6

3 B **6**
3.1 i - For each sample, the average income 6
3.2 ii to v - Average and standard deviation of the 100 estimators in part (i) and some plots 7
3.3 vi - 95% CI 9
3.4 vii - Comparison with 100 size samples 10

4 C **12**

5 D **14**
5.1 i - Compare incomes 15
5.2 ii - Family sizes 15
5.3 iii - Education level 16

6 Final comment **16**

1 Introduction

The assignment consists of working with a data set called families. It contains information about 43886 families. The first thing i did was to write some initiating code wich sorted the families into structs who contained the information given in the task. See below.

```
%Init-Code
NR = 43886;
fam = readtable('families.txt');
field1 = 'FamilyType';
field2 = 'NoP';
field3 = 'NoC';
field4 = 'Inc';
field5 = 'Region';
field6 = 'Educ';
s = struct(field1,0,field2,0,field3,0,field4,0,field5,0,field6,0);


Families(1,43886) = struct(s);

for i=1:43886
    Families(i) = struct(field1,fam.x_TYPE_(i),field2,fam.x_PERSONS_(i),
                         field3,fam.x_CHILDREN_(i),field4,fam.x_INCOME_(i),
                         field5,fam.x_REGION_(i),field6,fam.x_EDUCATION_(i));
end
```

2 A

In this task a you are supposed to estimate some different population parameters aswell as calculate the standard error of the estimate. Then form a 95% CI for each parameter. This is to be done 5 times with a sample of 500 families for each estimate.

To acquire the n = 500 samples i use the following code, then just repeating it in the following subsections five times for each parameter

```
%Sample 500 families
load Families
sampleA(1,500) = struct(s);
taken = zeros(1,500);
for i=1:length(sampleA)
   x = floor(rand()*43886);
   sampleA(i) = Families(i);
end
```

2.1 i - The proportion of female-headed families

The parameter to be estimated is the proportion of female-headed families, the estimator is denoted $\hat{p}$. The following equations will be used.

$$\hat{p} = \text{The proportion that have trait} \tag{1}$$

$$S_{\hat{p}}^2 = \frac{\hat{p}(1-\hat{p})}{n-1} \tag{2}$$

$$CI = \hat{p} \pm S_{\hat{p}} * 1.96 \tag{3}$$

Where 1.96 is due to the fact that we want a 95% Confidence intervall. And equation (2) is missing the finite population correction due to it's negligible value, the population in total is a lot bigger then the sample size. All estimates

Test	$\hat{p}$	$S_{\hat{p}}$	CI
1	0.1980	0.0178	[0.1630,0.2330]
2	0.2000	0.0179	[0.1649,0.2351]
3	0.1920	0.0176	[0.1574,0.2266]
4	0.2160	0.0184	[0.1799,0.2521]
5	0.1780	0.0171	[0.1444,0.2116]

Table 1: The proportion of female-headed families

are relativly close to eachother except maybe the last one. The confidence intervalls will contain the true value in 95% of the cases. So if we did this more than 5 times, say 100. We would expect that 95 of these CI:s would contain the true value and 5 would not.

The code used is the following:

```
%Code for part i
nOfemaleheads = 0;
for i=1:length(sampleA)
    if sampleA(i).FamilyType == 3
        nOfemaleheads = nOfemaleheads +1;
    end
```

```
end
Prop1 = nOfemaleheads/length(sampleA)
SE1 = sqrt(Prop1*(1-Prop1)/(length(sampleA) - 1))
CI95_1 = [Prop1-1.96*SE1 Prop1+1.96*SE1]
```

2.2 ii - The average number of children per family

The parameter to be estimated is the the average number of children per family, the estimator is $\bar{X}$. The following equations will be used.

$$S^2 = \frac{1}{n-1} * \sum_{i=1}^{n} (X_i - \bar{X})^2 \tag{4}$$

$$\bar{X} = \frac{1}{n} * \sum_{i=1}^{n} X_i \tag{5}$$

$$S^2_{\bar{X}} = \frac{S^2}{n} \tag{6}$$

$$CI = \bar{X} \pm S_{\bar{X}} * 1.96 \tag{7}$$

The motivation for neglecting the finite population correction is the same as that of the previous subsection. The estimated values are close to eachother

Test	$\bar{X}$	$S_{\bar{X}}$	CI
1	0.2300	0.0188	[0.1931,0.2669]
2	0.1980	0.0178	[0.1630,0.2330]
3	0.2180	0.0185	[0.1818,0.2542]
4	0.2340	0.0190	[0.1969,0.2711]
5	0.2160	0.0184	[0.1799,0.2521]

Table 2: The average number of children per family

as is the standard errors. The random 95% confidence intervalls as in the previous subsections will contain the true value in 95% or 19/20 of the cases. So if we do 20 tests instead of 5 we could expect one CI won't contain the true value.

The code used is the following:

```
%Code for part ii
meanch = mean([sampleA.NoC])
SE2 = std([sampleA.NoC])/sqrt(length(sampleA))
CI95_2 = [meanch-1.96*SE2 meanch+1.96*SE2]
```

2.3 iii - The proportion of head of households who did not receive a high school diploma, iv - The average family income

The solution for part (iii) is the same as in part (i). And the solution for part (iv) is the same as in part (ii). With some minor obvious modifications to the code. Therefore i will just show the results and discuss them without declaring any formulas more than one time.

2.3.1 (iii)

My interpretation of not receiving a high school diploma is that you have some form of lower degree of education. I.e the condition for not receiving a high school diploma is x $<$ 39, where x is the education level of the head of the family. The estimator is denoted $\hat{p}$.

Test	$\hat{p}$	$S_{\hat{p}}$	CI
1	0.1980	0.0178	[0.1630,0.2330]
2	0.2000	0.0179	[0.1649,0.2351]
3	0.1920	0.0176	[0.1574,0.2266]
4	0.2160	0.0184	[0.1799,0.2521]
5	0.1740	0.0170	[0.1407,0.2073]

Table 3: The proportion of head of households who did not receive a high school diploma

The estimated partion of head of households who did not receive a high school diploma is around 19% and the conclusion of the CI:s is simular of part (i). The modification done to the code:

```
%Code for part iii
highscool= 0;
for i=1:length(sampleA)
    if sampleA(i).Educ < 39
        highscool = highscool +1;
    end
end
Prop3 = highscool/length(sampleA)
SE3 = sqrt(Prop3*(1-Prop3)/(length(sampleA) - 1))
CI95_3 = [Prop3-1.96*SE3 Prop3+1.96*SE3]
```

2.3.2 (iv)

The estimate of the average family income is denoted $\bar{X}$. The average family

Test	$\bar{X}$ (10^4)	$S_{\bar{X}}(10^3)$	CI (10^4)
1	4.0201	1.3585	[3.7538,4.2863]
2	4.4020	1.5733	[4.0936,4.7104]
3	3.9916	1.3806	[3.7210,4.2622]
4	4.2134	1.4166	[3.9357,4.4911]
5	4.0292	1.3612	[3.7624,4.2960]

Table 4: The average family income

income lies around $4 * 10^4$ (note the scaling in the header of Table 4). The conclusion of the CI:s is the same of part (ii). The standard errors seems to be propertional to the other estimates, one decimal less then the estimated value. The modified code:

```
%Code for part iv
meaninc = mean([sampleA.Inc])
SE4 = std([sampleA.Inc])/sqrt(length(sampleA))
CI95_4 = [meaninc-1.96*SE4 meaninc+1.96*SE4]
```

3 B

Now we are going to take 100 samples of size 400 using to following code. Each row of sampleB contains 400 families sampled.

```
%Sample 400 families 100 times.
load Families
sampleB(100,400) = struct(s);
for k=1:100
    for i=1:400
       x = floor(rand()*43886);
       sampleB(k,i) = Families(x);
    end
end
```

3.1 i - For each sample, the average income

This task is pretty straightforward, i will just show the simple code i used and not make any conclusions. To show the data (100 averages) will take to much space. But the result is in the vector AvIncome.

```
%% i
AvIncome = zeros(1,100);
for i=1:100
    AvIncome(i) = mean([sampleB(i,:).Inc]);
end
```

3.2 ii to v - Average and standard deviation of the 100 estimators in part (i) and some plots

So we are going to find the standard deviation and average of the 100 estimators from the previous subsection and plot a histogram. Using equation (5) for the average and the root of equation (4) for the standard devation we get the following values.

$$\textit{Average } \mu = 41237$$

$$\textit{Standard deviation } \sigma = 1604.8$$

Next we are to plot the histogram togheter with the normal density with the information about the average and standard deviation we already have (see figure (1)).

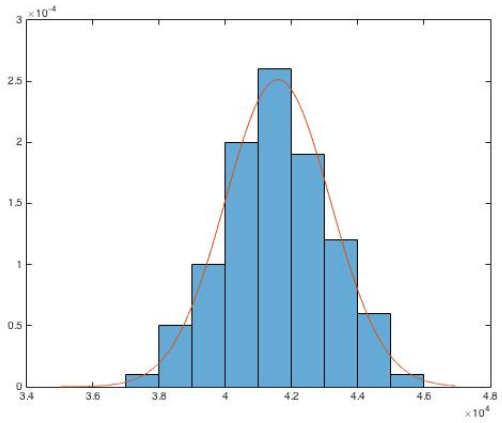

Figure 1: Histogram togheter the normal density function

As we see the histogram and the normal pdf looks very simular. The reason for this is probably the CLT.

The next task is to plot the empirical cumulative distribution function togheter with the normal the normal cumulative distribution function. The result is show in figure (2) where we can se that the normal cumulative distribution function fits our empirical cumulative distribution function very well.

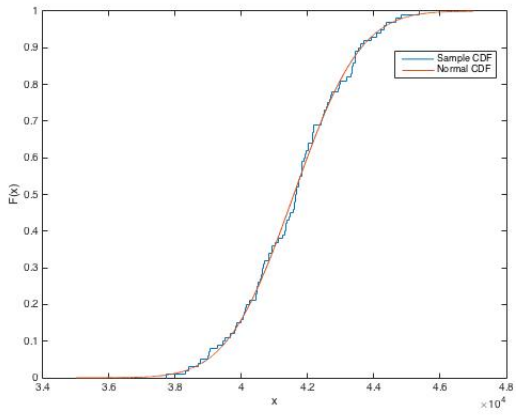

Figure 2: Empirical cdf and the Normal cdf

The last plot to be made for examining a normal approximation is the normal probability plot. The purpose of the plot is to detect departures from normality in a illustrative graphical way. The result is shown in figure (3) As we see our

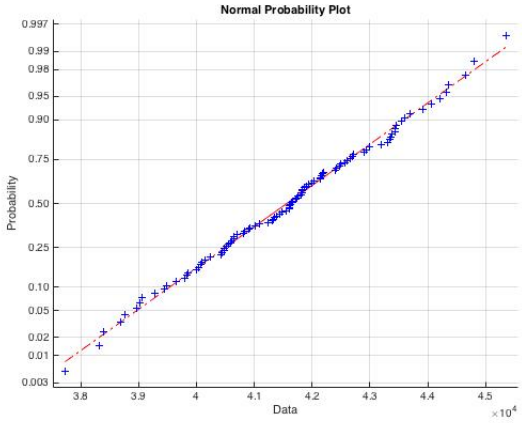

Figure 3: Normal probability plot

data is a little bit of the red line on several places, especially at the beginning and end. This indicates that we have something called "Long Tails" i.e that we have more variance than we would expect in a normal distribution. The code for these tasks is shown below.

```
%% ii - v
load AvIncome
M = mean(AvIncome);
std = std(AvIncome);
figure(1)
histogram(AvIncome,'Normalization','pdf');
hold on
X = 35000:1:47000;
Y = normpdf(X,M,std);
figure(1)
plot(X,Y);

figure(2)
ecdf(AvIncome);
hold on
CdN = cdf('Normal',X,M,std);
figure(2)
plot(X,CdN)
legend('Sample CDF','Normal CDF')
hold on

figure(3)
normplot(AvIncome);
```

3.3 vi - 95% CI

In this task we are supposed to find 95% confidence intervalls on the average income for these 100 samples. And then check how many of these intervalls contain the actual population target. In theory 95% of the 100 CI:s would contain the population target, i.e 95 of them. The 100 CI:s are calculated with equation (7). And the actual population target, i assume, is the actual average income of the total 43886 families. The result is show below (figure(4)) where the horizontal line is the true value and the vertical lines are the 100 different confidence intervals. As we expected approximatly 5 ended up not containing

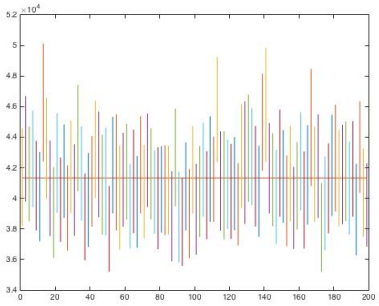

Figure 4: Confidence intervalls and true value

the true value, actually 6 in this case. The code used is displayed below.

```
%% ii - v
load SampleB
load Families
truevalue = mean([Families.Inc])
t = 1:2:200;
plot(t,ones(100)*truevalue);
hold on

for i=1:100
    av = mean([sampleB(i,:).Inc]);
    SE = std([sampleB(i,:).Inc])/sqrt(400);
    plot([t(i) t(i)],[(av - SE*1.96),(av+SE*1.96)])
    hold on
end
```

3.4 vii - Comparison with 100 size samples

Now we are to take instead of 100 samples of size 400, 100 samples of size 100. We will compare the averages, standard devations and histogram of the two different cases. I illustrate this using two scatter plots, one for the averages and one for the standard deviations (figure (5) and (6)).

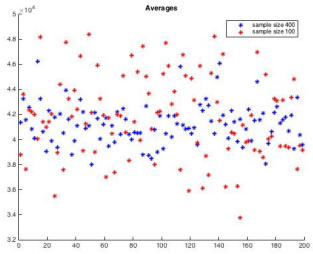

Figure 5: The averages of 100 and 400 sample size

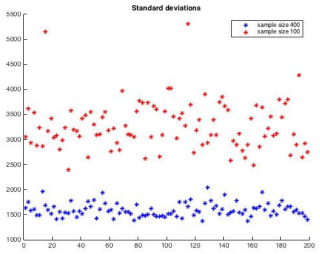

Figure 6: The standard deviations of 100 and 400 sample size

As we see the averages of the larger sample is more concentrated and compact around the true value (≈ 41300). While the smaller sample size yields a much larger spread both above and below the true value. The figure of the standard deviations talk for itself, the standard deviations of the larger sample are smaller and more concentrated. Since the larger sample is about 4 times bigger then the smaller a reliable hypothesis would be that the standard deviations of the larger sample should be $\approx \sqrt{4} = 2$ times smaller than the standard deviation of the smaller sample. Which absolutely seems to be the case, thus relating the result to the theory of simple random sampling.

Moving on the the histograms. As we see in figure (7) the histogram of the smaller sample size the values are more spread out "over the bins" than the larger sample size. Which justifies the result from the scatterplots.

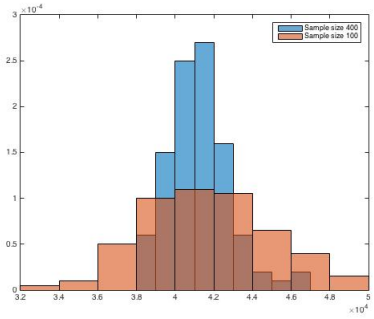

Figure 7: Overlapping histograms of the two different sample sizes

The code for the section is displayed below.

```
%vii
load families
load sampleB
sample2(100,100) = struct(s);
for k=1:100
    for i=1:100
       x = floor(rand()*43886);
       sample2(k,i) = Families(x);
    end
end

t = 1:2:200;
for i=1:100
    av1 = mean([sampleB(i,:).Inc]);
    av2 = mean([sample2(i,:).Inc]);
    figure (1)
    scatter(t(i),av1,'*','b');
    hold on
    scatter(t(i),av2,'*','r');
    legend('sample size 400','sample size 100');
    title('Averages');

    SE1 = std([sampleB(i,:).Inc])/sqrt(400);
    SE2 = std([sample2(i,:).Inc])/sqrt(100);
    figure (2)
    scatter(t(i),SE1,'*','b');
    hold on
    scatter(t(i),SE2,'*','r');
    legend('sample size 400','sample size 100');
    title('Standard deviations')
end

%Histograms
AvIncome_1 = zeros(1,100);
AvIncome_2 = zeros(1,100);
for i=1:100
    AvIncome_1(i) = mean([sampleB(i,:).Inc]);
```

```
    AvIncome_2(i) = mean([sample2(i,:).Inc]);
end
figure(1)
histogram(AvIncome_1,'Normalization','pdf');
hold on
histogram(AvIncome_2,'Normalization','pdf');
hold on
legend('Sample size 400','Sample size 100');
```

4 C

Now we are to take a sample of 500 families and and compare the income of the three different family types via histograms and boxplots. First i use some initiation code for sampling and sorting the different family types into separat arrays.

```
%% C
clear all
close all
clc
clf

load families
sample(1,500) = struct(s);
FT1 = 0;
FT2 = 0;
FT3 = 0;

%Sample and keep count of the different families.
for i=1:length(sample)
   x = floor(rand()*43886);
   sample(i) = Families(x);
   if sample(i).FamilyType == 1
       FT1 = FT1 + 1;
   elseif sample(i).FamilyType == 2
       FT2 = FT2 + 1;
   else
       FT3 = FT3 +1;
   end
end

F1(1,FT1) = struct(s);
F2(1,FT2) = struct(s);
F3(1,FT3) = struct(s);
k=1;
j=1;
l=1;
%sort the families
for i=1:length(sample)
   if sample(i).FamilyType == 1
       F1(k) = sample(i);
       k = k+1;
   elseif sample(i).FamilyType == 2
```

```
        F2(j) = sample(i);
        j = j+1;
    else
        F3(l) = sample(i);
        l = l+1;
    end
end
```

Using the data from F1,F2 and F3 i plot histograms and boxplots shown below in figure (8) and figure (9).

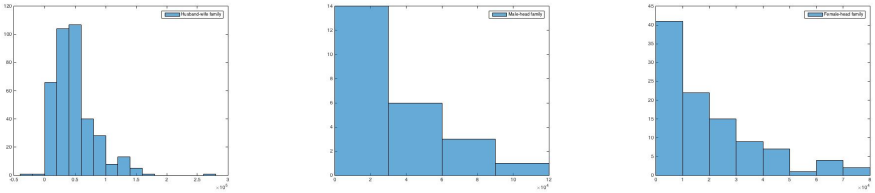

Figure 8: Histograms

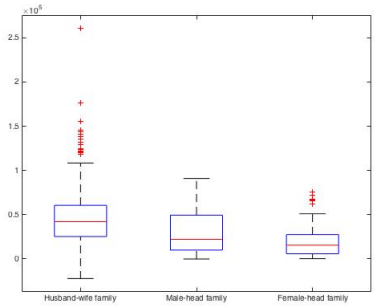

Figure 9: Boxplots regarding income of family types

As we notice Husband-wife families have a in general higher income, which is logical because there are two providers instead of one. It seems like in general male-head familiys have a higher income than female-head familys. There are a few "extreme values" in both the first and third family type. which is also quite logical, "the richest 1% have more wealth then the remaining 99%"[1].
The code is as usual below.

```
%% ii - v
figure(1)
histogram([F1.Inc]);
legend('Husband-wife family');
```

[1] http://www.bbc.com/news/business-30875633

```
hold on

figure(2)
histogram([F2.Inc]);
legend('Male-head family');
hold on

figure(3)
histogram([F3.Inc]);
hold on
legend('Female-head family');


C = [[F1.Inc],[F2.Inc],[F3.Inc]];
grp = [zeros(1,length([F1.Inc])),ones(1,length([F2.Inc])),2*ones(1,length([F3.Inc]))];
figure(4)
boxplot(C,grp,'labels',{'Husband-wife family','Male-head family','Female-head family'});
hold on
```

5 D

Now we are to take a random sample of 400 from each of the four regions. It is done using the following code.

```
%Sample from different regions
load Families
SampleNorth(1,400) = struct(s);
SampleEast(1,400) = struct(s);
SampleSouth(1,400) = struct(s);
SampleWest(1,400) = struct(s);
j=1;
l=1;
k=1;
i=1;
while l<=400 || j<=400 || k<=400 || i<=400
       x = floor(rand()*43886);
       temp = Families(x).Region;
       if temp == 1 && j<=400
           SampleNorth(j) = Families(x);
           j = j+1;
       elseif temp == 2 && l<=400
           SampleEast(l) = Families(x);
           l = l+1;
       elseif temp == 3 && k<=400
           SampleSouth(k) = Families(x);
           k = k+1;
       elseif temp == 4 && i <=400
           SampleWest(i) = Families(x);
           i = i+1;
       end
end
```

Now the different Sample+"Region" contains 400 samples from respectively region.

5.1 i - Compare incomes

We are now to compare the different regions with respect to the income by making boxplots.

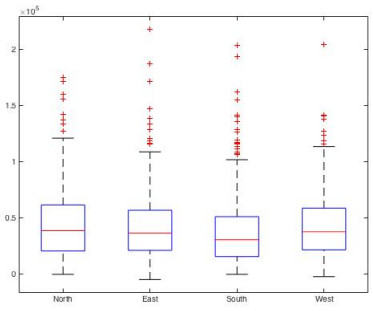

Figure 10: Boxplots regarding income in different regions

The medians look basicly the same, with some small differences. The south-region income is in general less then the other. But the south show more extreme values than the others. The code for generating the boxplot is in essence the same as in section 4.

5.2 ii - Family sizes

To compare the family sizes of the different regions we again make parallell boxplot as in the previous comparison of income. The code is as then simular to section 4.

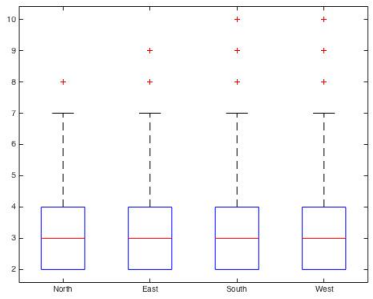

Figure 11: Boxplots regarding number of persons in families in different regions

The medians are almost identical and so are the means when calculating them ($\approx 3,1$). However the extreme values are fewer in the north and east region.

5.3 iii - Education level

We now turn to the question about difference in the educational level between the regions. The educational level is ranked in integers ranging from 31 to 46. Where a lower value indicates a lower form of education and a higher indicates a higher form of education. So we should be able to do the boxplots in the same fashion as in the cases of income and persons in families because the educational level is measured in numbers. Hence the same code is reused and the boxplots produced is displayed in figure (12).

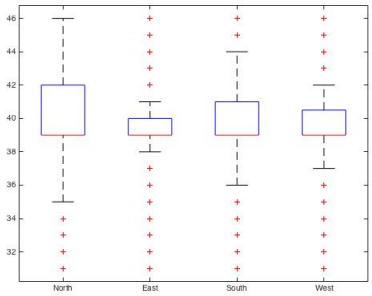

Figure 12: Boxplots regarding the education in different regions

The medians seems to be close to eachother, but they are all on the border of their respective region. This indicates a very high degree of difference among the educationel levels seen in the regions for themselves. They all have alot of extreme values. However in the case of the north we se a underrepresentation in educational level higher then a college degree.

6 Final comment

In this bonus assignment i displayed the code that belongs to a specific problem in the section where i discussed the problem. I think this way it becomes more clear how i have thought solving the problem. In my oppinion this was the best way to represent it (there is alot of code) instead of displaying all of it in the last pages. But i would appreciate, if you think it is better for me to put it in an appendix, to let me know.

I tried to keep the pagenumbers down because it is quite a long report for being one of four bonus assignments. With that in mind i tried not to repeat

myself when simular tasks appeared. I also performed the simple random sample with replecement, the problem did not specify which type to use. But it should produce essentially the same result because we are doing rather small samples in a big population[2].

[2] https://en.wikipedia.org/wiki/Simple_random_sample