MVE155, Statistical Inference
Assignment 1

February 7, 2020

Ella Guiladi
930509-0822
guiladi@student.chalmers.se

Exercises

The code to the following solutions and data can be found in the source code files attached in the mail. Since the code is written in cells, you need to run through all the cells in order for the code to work.

a.

In the following tasks we were supposed to estimate population parameters, calculate the estimated standard error of these estimates and form 95 % confidence intervals of a simple random sample of 600 families. The source code for this subsection can be found in the attached file, by opening the script-file "ExerciseA"

(i) The parameter that one want to estimate is the proportion of husband-wife family. The estimator for the following is the sample proportion $\hat{p}$. To solve the problem, on used the following equations

$$\hat{p} = \bar{x}, \tag{1}$$

i.e the sum of the husband-wife families divided by the sample size. The sample mean turns into a sample proportion, giving an unbiased estimate of p. The estimated standard error is given by:

$$s_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n-1}}, \tag{2}$$

where n is the sample size, which is 600. The last equation is used to form the 95% confidence intervals

$$I_p \approx \hat{p} \pm z_{\frac{\alpha}{2}} \cdot s_{\hat{p}}, \tag{3}$$

where $z_{\frac{\alpha}{2}} = 1.96$ for a 95% confidence interval.

All estimates are presented in the following table,

Table 1: Table with estimates for the proportion of husband-wife family.

$\hat{p}$	$s_{\hat{p}}$	95% confidence interval
0.762	0.0174	[0.7275,0.7958]

(ii) Now we want to find out the average number of children per family as well as the population parameters, the estimated standard error and the 95% confidence interval. The following equations was used,

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{4}$$

in order to calculate the average number of children per family, where n is the sample size.

In order to calculate the estimated standard error, following equation was used for the standard deviation was used,

$$s = \sqrt{\frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2}. \tag{5}$$

The estimated standard error for the sample mean is calculated through,

$$s_{\bar{x}} = \frac{s}{\sqrt{n}}, \tag{6}$$

and,

$$I_\mu \approx \bar{x} \pm z_{\frac{\alpha}{2}} \cdot s_{\bar{x}}, \tag{7}$$

gives the 95% confidence interval. All estimates are presented in the following table,

Table 2: Table with estimates for number of children per family

$\bar{x}$	$s_{\bar{x}}$	95% confidence interval
0.980	0.0482	[0.8855,1.0745]

(iii) The average number of persons per family and as well as the population parameters, the estimated standard error and the 95% confidence interval. The equations (4), (6) and (7) was used in the same way as in task aii), in order to get the following table,

Table 3: Table with estimates for number of persons per family

$\bar{x}$	$s_{\bar{x}}$	95% confidence interval
3.220	0.0555	[3.1111, 3.3289]

b.

In the following tasks 100 samples of size 400 was used. The source code for this subsection can be found in the attached file, by opening the script-file "ExerciseB"

(i) For each sample, the average family income was found by looping a sample income of the size 100x400 (i.e samples x size of samples) over the number of samples, to get a sample of the income with this dimension. Then it is possible to use the build in function "mean" in Matlab, that returns the mean of the input, which gives the average income for each sample as a vector. The vector can be visualized in the code since it is of dimension 100x1.

(ii) The average of the 100 estimates was found as in bi), i.e by using the build in function "mean" with the average vector as input. The standard deviation was found through the built in function "std" in Matlab.The "std" function returns the standard deviations of the sample income matrix.

The average and the standard deviation of the 100 estimates are 41026.75 respectively 1567.38 and the following histogram of the estimates is presented in figure [1].The histogram in [1] and [2] have different scales due to the fact that the histogram in figure [2] is normalised, in order to get a better scale on the axis.

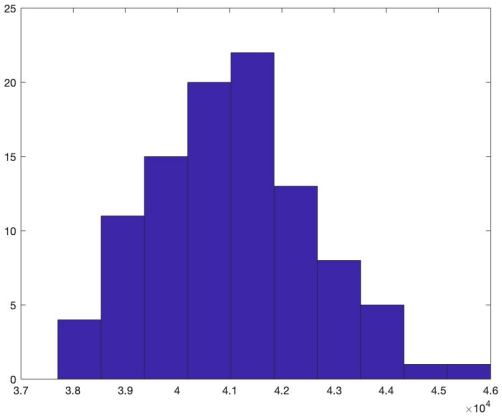

Figure 1: Histogram of the estimates

(iii) As we can see from figure [2], the histogram and the normal probability density function have similar shapes. The reason for this might be due to the properties of the central limit theorem. The central limit theorem states that regardless of the population distribution model, with an increasing sample size, the mean will approach a normal distribution around the population mean and the standard deviation becomes smaller as the sample size increases. However, since it's still a finite sample size, there will be small deviations from the normal distribution.

(iv) For each of the 100 samples, we want to find a 95% confidence interval for the population average income. This confidence interval is calculated in the same way as in equation (7). In order to know how many of these intervals that contains the population target, it is necessary to calculate the total population target, i.e the average income of the total 43886 number of families. The degree of confidence (95% in this case) describes the probability that the confidence interval captures the true population target. This means that since we have 100 samples, approximately 95 of the confidence intervals should

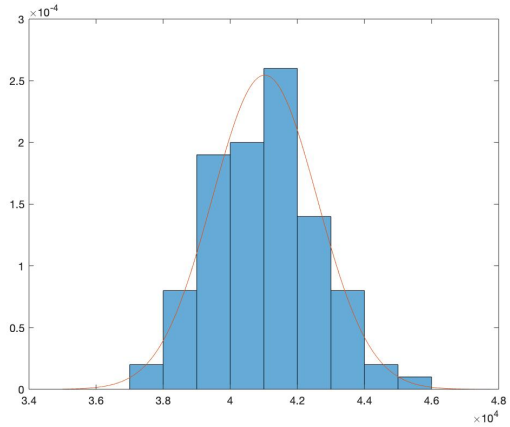

Figure 2: Superimposed plot of a normal density with that mean and standard deviation of the histogram

contain the true population target. From figure [3], one can see that this is in fact true since approximately 5 of the confidence intervals aren't containing the true value. In figure [3], the total population target is visualized as an horizontal line and the confidence intervals are visualized as vertical lines.

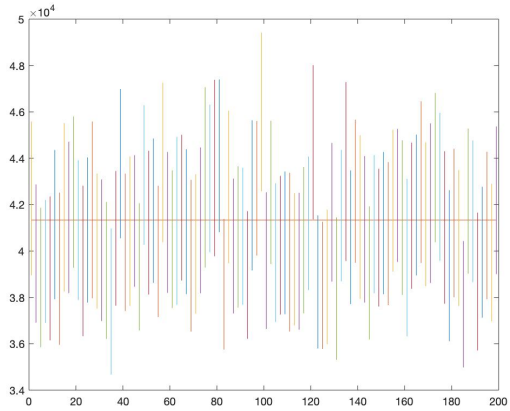

Figure 3: Confidence interval on the vertical axis and the true population target on the horizontal axis.

(v) Now a 100 samples of size 100 was chosen and then the averages, standard deviations and histograms was compared to the 100 samples of size 400. The average and the standard deviation of the 100 estimates of size 400 are approximately 41027 respectively 1567 and the average and the standard deviation of the 100 estimates of size 100 are approximately 41697 respectively 2923. The averages and standard deviation for the 100 samples of size 100 was calculated in the same way as in bii).

Simple random sampling is meant to be an unbiased representation of a population, but is sensitive to sampling error since the randomness of the selection can result in that the sample doesn't accurately represent the population that it's supposed to represent. Thereby, a larger sample size would result in a smaller sampling error.

The averages have a small difference, with the average of the larger sample size being slightly closer to the true value, since an increasing sample size leads to the sample mean

clustering more and more around the true population mean. The true value is approximately 41300, which can be seen from figure (3)). The standard deviation for the smaller sample size is on the other hand twice as big as the standard deviation for the larger sample size. The standard error increases when the standard deviation increases due to an increase in the variance of the population, i.e observed values are further from the true value, which agrees with the theory of simple random sampling.

As presented in figure (4) with the overlapping histograms, one can see that the histogram with the smaller size is more spread out over each bin then the larger sampling size, which is inline with above discussion.

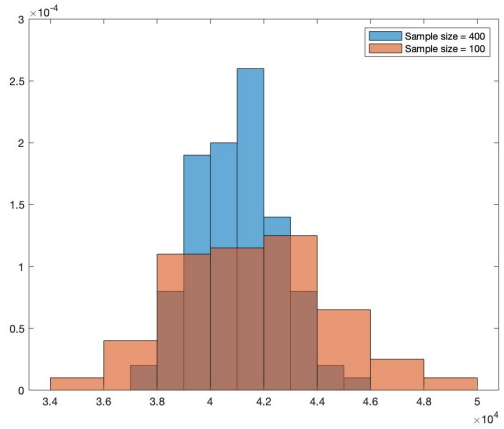

Figure 4: Compared histograms of samples with size 400 respectively 100

c.

In this task, the families was stratified into four strata by region (North, East, South and West). Where North contained all the 1 in the region column, East contained all of the 2, South contained all of the 3 and West contained the part of the column with all of the 4. The source code for this subsection can be found in the attached file, by opening the script-file "ExerciseC"

(i) The sampling fraction is given by the equation for proportional allocation respectively optimal allocation, divided by the entire population, i.e 43886. The equation for proportional allocation is given by,

$$n_j = n \cdot w_j, \tag{8}$$

and the equation for optimal allocation is given by,

$$n_j = n \cdot \frac{w_j \cdot \sigma_j}{\bar{\sigma}}. \tag{9}$$

The proportion $_j$ is given by,

$$w_j = \frac{n_j}{n}. \tag{10}$$

The sampling fraction for the proportional allocation is thereby given by the equation,

$$Samplefraction = w_j. \tag{11}$$

The results are presented in the table below

Table 4: Table with sample fraction of the proportional allocation

North	East	South	West
0.2313	0.2367	0.3066	0.2254

The sampling fraction for the optimal allocation is given by,

$$Samplefraction = \frac{w_j \cdot \sigma_j}{\bar{\sigma}}, \tag{12}$$

where σ_j is calculated by creating an income matrix for each region, containing only the income for that region. So for example, the north region (with only the number 1 in the region column), has one column with one and one more column with the corresponding income for each row. All income matrices can be found in the code. To find each σ_j, the built in function "std" is used with the income matrix for each region as an input. In order to calculate $\bar{\sigma}$, the following equation was used,

$$\bar{\sigma} = \sum_{i=1}^{k} w_i \cdot \sigma_i. \tag{13}$$

The results are presented in the table below,

Table 5: Table with sample fraction of the optimal allocation

North	East	South	West
0.2529	0.2220	0.2892	0.2359

(ii) In the last sub-exercise, 500 observations was allocated proportionally to the four regions by choosing $500 \cdot w_j$ observations sampled uniformly at random from the data in each income matrix. From this sample, the mean and the standard deviation are calculated with the built in function "mean" and "std" as before.

The stratified sample mean is calculated by,

$$\bar{x_s} = \sum_{i=1}^{k} w_i \cdot \bar{x_i}, \tag{14}$$

and the estimated error is calculated by,

$$\sqrt{s^2_{\bar{x}_s}} = \sqrt{w_1^2 s^2_{\bar{x}_1} + \ldots + w_k^2 s^2_{\bar{x}_k}} = \sqrt{\frac{w_1^2 s_1^2}{n_1} + \ldots + \frac{w_k^2 s_k^2}{n_k}}. \tag{15}$$

The equation for the 95% confidence interval is,

$$I_\mu = \bar{x_s} \pm z_{\frac{\alpha}{2}} \cdot s_{\bar{x_s}}, \tag{16}$$

Table 6: Table with estimates for the stratified sample

$\bar{x_s}$	$s_{\bar{x_s}}$	95% confidence interval
39350.66	147.39	[39781, 40359]

The average and the standard deviation of the 100 estimates of size 400 are approximately 41027 respectively 1567 and the average and the standard deviation of the 100 estimates of size 100 are approximately 41697 respectively 2923. The average and the standard deviation for the random sample (both sample sizes) are larger then the stratified sample, meaning that the estimated error is smaller for the stratified sample. One can also see from Table 6 that the estimated standard error is much smaller for the stratified sample, then for the samples from the simple random sampling. This might be due to the fact that when we have some information about the population (such as region and income) it may be favourable to use the stratified sampling technique in order to account for the differences within the population, which the random simple sampling doesn't take into account.