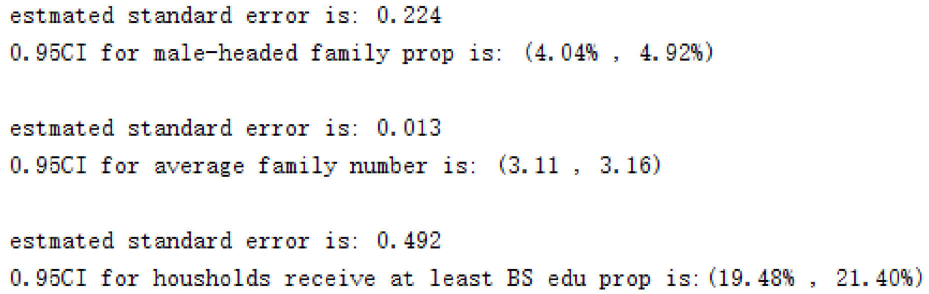
Assignment1
MVE155 Statistical inference

Yankun Xu (940630-0237)
yankun@student.chalmers.se

1. Question A

Take five samples with size 500, we should firstly get five sample values from per 500 data. Now we can get the sample mean according to $E(X_i) = \mu$, then what we want is estimated standard errors. We can estimate the variance from these five samples, then use $s^2 = \frac{n}{n-1}\left(\overline{X^2} - \overline{X}^2\right)$ and $s_{\overline{X}} = \frac{s}{\sqrt{n}}$ to get estimated standard errors, by the way, in this case n=4 and samples have the property of I.I.D. Although the number of sample is not big, higher sample size may offset this disadvantage. The 95%CI (Confidence Interval) can be calculated by formula $\mu = \overline{X} \pm 1.96 s_{\overline{X}}$. The answers of three questions are shown as following: (I got screenshot from Matlab)



```
estmated standard error is: 0.224
0.95CI for male-headed family prop is: (4.04% , 4.92%)

estmated standard error is: 0.013
0.95CI for average family number is: (3.11 , 3.16)

estmated standard error is: 0.492
0.95CI for housholds receive at least BS edu prop is:(19.48% , 21.40%)
```

Figure 1. Answers of three questions in question a

2. Question B

2.1 The first question is easy to implementation. The result of 100 means is shown in the data sample_aedu (in Matlab).

We can compare the two datasets in question ii and v. Under the size 400, the average of 100 samples is 39.42 and standard deviation is 0.14. And, under size 100, the average is 39.45 and standard deviation is 0.30. Two histograms are shown as figure 2. We can notice that, the average is nearly same and histograms are both approximately satisfying normal distribution, difference is size 400 has normal distribution with less variance. But standard deviation of size 100 is half of size 400, which means size 400 perform better. The reason is obvious when we know the relationship between standard deviation and sample size :

$$s_{\overline{X}} = \frac{s}{\sqrt{n}}$$

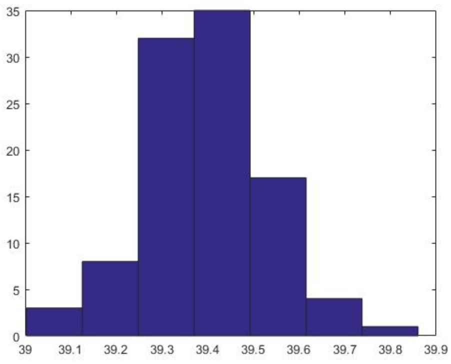


Figure 2.1. Histogram of size 400

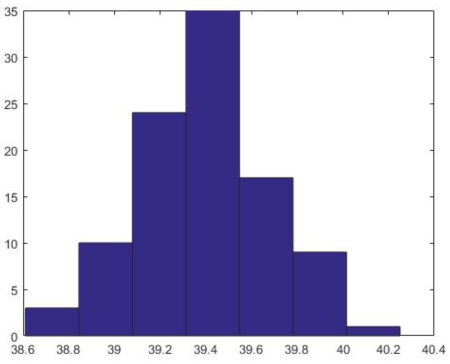


Figure 2.2. Histogram of size 100

2.2 Question iii is to superimpose normal density plot and histogram. Figure3 is shown as following.

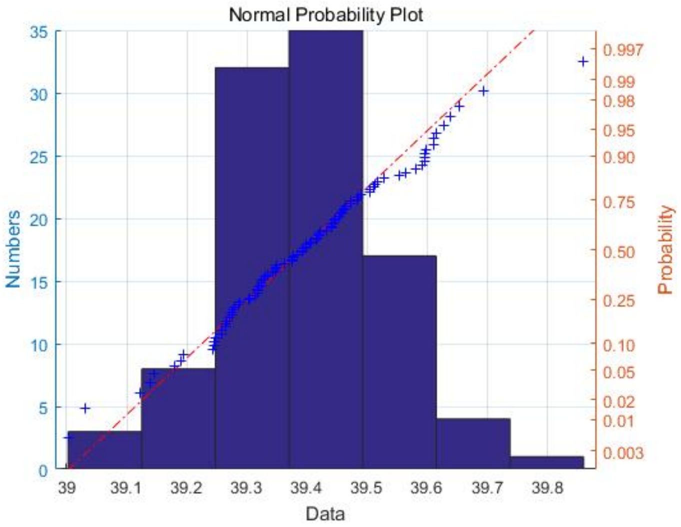


Figure 3. Histogram and normal density plot

We could notice that the appearance of data fitting normal distribution is good because most of data is close to the straight line $y = ax + b$.

2.3 Question iv is to estimate the CI of population average education level based on each sample. We need use $s^2 = \frac{n}{n-1}\left(\overline{X^2} - \overline{X}^2\right)$ again, but in this case, we should be very careful because this sample is not I.I.D. So we use formula $s_{\overline{X}} = \frac{s}{\sqrt{n}} \cdot \sqrt{1 - \frac{n}{N}}$. Then, we could use $CI = \overline{X} \pm 1.96 s_{\overline{X}}$ to calculate 95%CI. Answer is stored in data CI_pop.

3. Question C

3.1 Question i is to compare family incomes from 4 regions, boxplot is shown in figure4. Although the average incomes are similar , so many data is outside the box.

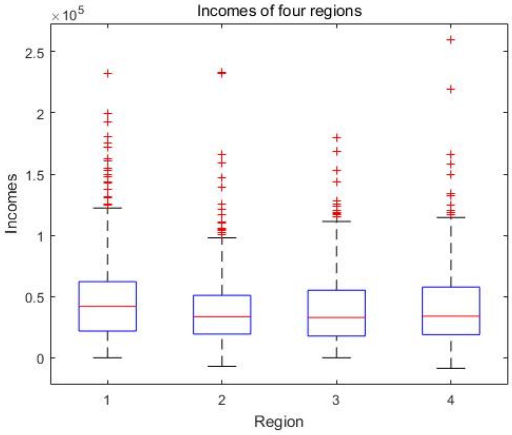
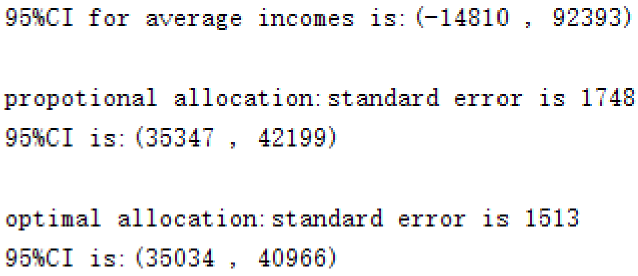


Figure 4. Boxplot of four regions

3.2 Next, we need figure out whether or not stratification method is useful to estimate the average family incomes. Firstly, we calculate the 95%CI under common method, due to its huge standard deviation, the left threshold is even negative, it is not what we want. Then, we implement proportional and optimal allocation method. In proportional allocation, strata n is based on the proportion of family type in population parameters ($n_i = nw_i$). And in optimal case, $n_i = n\frac{w_i\sigma_i}{\bar{\sigma}}$ ($\bar{\sigma} = \sum_i w_i\sigma_i$). Then we need stratified sample mean and variance. Three formulas will be used : $\bar{X}_s = \sum_i \bar{X}_i w_i$, $\mathrm{Var}(\bar{X}_{so}) = \frac{\bar{\sigma}^2}{n}$, $\mathrm{Var}(\bar{X}_{Sp}) = \frac{\overline{\sigma^2}}{n}$. In this case, we use sample variance including all data. Finally, 95%CI can be achieved. The whole result is shown as figure5. We can notice that stratification method is much better than common method, and optimal allocation has smaller standard error and better performance.



```
95%CI for average incomes is:(-14810 , 92393)

propotional allocation:standard error is 1748
95%CI is:(35347 , 42199)

optimal allocation:standard error is 1513
95%CI is:(35034 , 40966)
```

Figure 5. Results of common and stratification method

(Some data stated in the report will be achieved after running the Matlab code)