Bonus Assignment 1 – Survey Sampling

Qixun Qu
901001-5551
qixun@student.chalmers.se

Problem A

In this problem, three parameters need to be estimated, which are the proportion of male-headed families(i), the average number of persons per family(ii) and the proportion of heads of households who received at least a Bachelor's degree(iii). To do the estimation, 5 different simple random samples are extracted, and the size of each sample is 500. Thus, three parameters of each sample can be computed, the results are shown as table 1. Since the data in sample is extracted randomly, three parameters have many different values with varying data. Table 1 just shows five possible results of five samples.

Paras	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5
i	0.060000	0.054000	0.064000	0.046000	0.054000
ii	3.224000	3.156000	3.072000	3.208000	3.228000
iii	0.210000	0.206000	0.220000	0.216000	0.218000

Table 1. Possible Results of Three Parameters in Each Sample

Three parameters of each sample can be regarded as independent with the other samples' parameters, since data is extracted randomly in each sample. Thus, the sample mean and the estimated standard error for sample mean of each parameter can be calculated as Equation 1 and Equation 2.

$$\bar{X} = \frac{X_1 + \cdots + X_n}{n} \qquad \text{Eq. 1}$$

$$S_{\bar{X}} = \frac{s}{\sqrt{n}} \qquad \text{Eq. 2}$$

So, the 95% confident interval for each parameter can be computed as $\bar{X} \pm 1.96 \times S_{\bar{X}}$.
In this case, *n* equals to 5 here. The results of each parameter's sample mean and estimated standard error are shown in Table 2.

Paras	Sample Mean	Estimated Standard Error	95% Confident Interval
i	0.055600	0.003059	[0.049604, 0.061596]
ii	3.177600	0.029356	[3.120063, 3.235137]
iii	0.214000	0.002608	[0.208889, 0.219111]

Table 2. Estimation of Each Parameter

Problem B

1. Sample Size is 400

In this section, 100 samples of size 400 are extracted at first. In each sample, the average education level of head of household is computed. The average and standard deviation of these 100 estimates for all samples are also calculated, which are shown in Table 3. The histogram of these 100 estimates and the normal density of estimates are plotted as Figure 1. Again, the results shown in the table is just one possible result. Every time the program generates different result in a reasonable range.

Average of 100 Sample Estimates	Standard Deviation of 100 Sample Estimates
39.432100	0.136857

Table 3. Average and Standard Deviation of 100 Sample Means

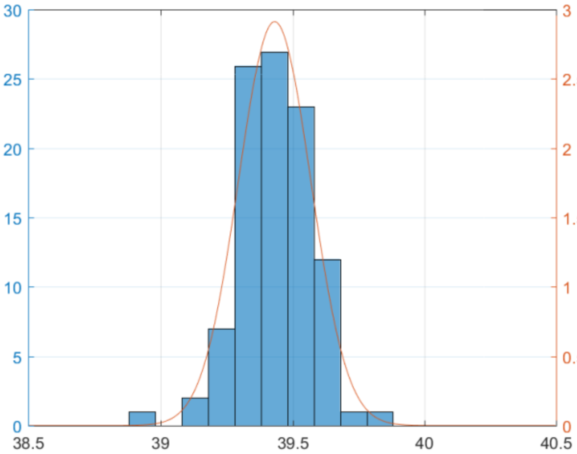

Figure 1. Histogram and Normal Density of 100 Estimates

It can be seen from the plot that the histogram is pretty similar with the normal density.

In this case, each sample is a simple random sample, the sample mean has been obtained as the average education level by Equation 1. However, the estimated standard error should be the unbiased one that is presented as Equation 3. Here, the *n* is the sample size 400, *N* is the population size. After which, 95% confident interval for each sample can be computed a $\bar{X} \pm 1.96 \times S_{\bar{X}}$.

$$S_{\bar{X}} = \frac{s}{\sqrt{n}} \times \sqrt{1 - \frac{n}{N}} \qquad \text{Eq. 3}$$

The confident intervals of first three samples are displayed as follows:
[39.264271, 39.835729], [39.071432, 39.678568], [38.878584, 39.456416].

2. Sample Size is 100

When the size of each sample in decreased to 100, the same process will be carried out to obtain plot and confident intervals. The average and standard deviation of these 100 estimates for all samples are also calculated, which are shown in Table 4. The histogram of these 100 estimates and the normal density of estimates are plotted as Figure 2.

Average of 100 Sample Estimates	Standard Deviation of 100 Sample Estimates
39.467500	0.310052

Table 4. Average and Standard Deviation of 100 Sample Means

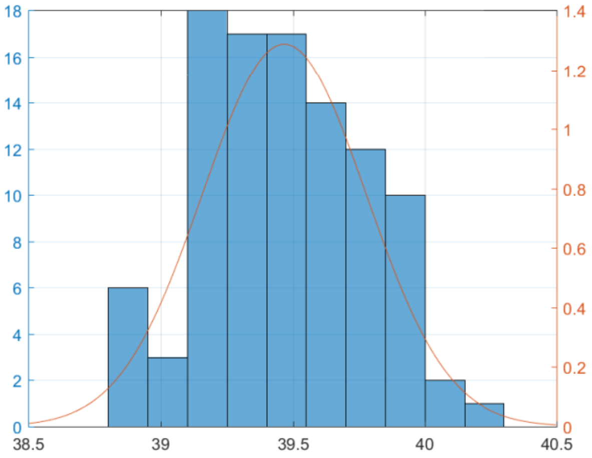

Figure 2. Histogram and Normal Density of 100 Estimates

The confident intervals of first three samples are displayed as follows:
[38.752995, 39.987005], [38.503619, 39.776381], [39.082126, 40.177874].

The average of 100 estimates of samples whose size is 100 has the same level with the average in section 1. But, the standard deviation is obvious larger than the previous one. Since in this section, the size of each sample is much smaller. The histogram and the normal density line in Figure 2 are also wider than in Figure 1. According to the Equation 3, the smaller the sample size is, the larger the estimated standard error is, resulting in a wider confident interval in this case.

Problem C

1. Income of Four Regions

Take simple random samples of size 400 from each region of the city, the parallel boxplots for income of each region are shown in Figure 3. From the plot, the median incomes of four regions are almost at the same level. Also, the range of interquartile and inner fence are almost same. Thus, there is no clearly difference in incomes among these four regions.

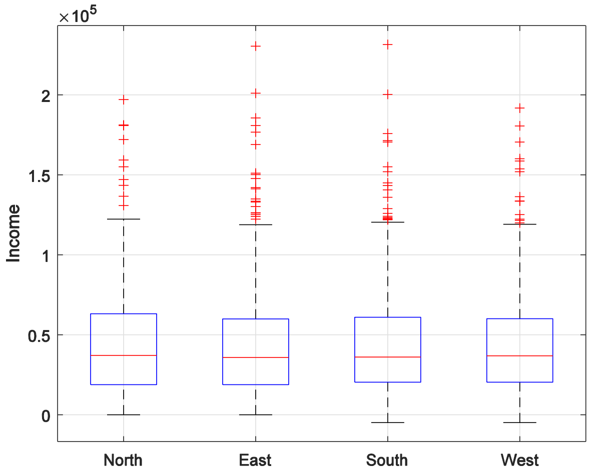

Figure 3. Parallel Boxplots for Incomes of Four Regions

2. Stratified Sample

2.1 Simple Random Sample

At first, a simple random sample of size 300 is extracted, the sample mean and the estimated standard error can be calculated as Equation 1 and 2, which is able to generate the 95% confident interval. The results are shown in Table 5.

2.2 Proportional Allocation

The proportion of each strata should be computed to obtain the size of each strata, which is shown in Equation 4. Here, n_i is the number of families in each type, *N* is the population size. Then, the proportion of each strata is obtained as $W = [W_1\ W_2\ W_3]$.

$$W_i = \frac{n_i}{N}, \qquad i = 1, 2, 3 \qquad \text{Eq. 4}$$

The mean of this strata sample is shown in Equation 5, in which $\bar{X}_i$ is the sample mean of each strata. The estimated variance is shown in Equation 6. Then the 95% confident interval can be generated. The results of this sample is shown in Table 5.

$$\bar{X}_{sp} = \sum_{i=1}^{3} W_i \bar{X}_i \qquad \text{Eq. 5}$$

$$s^2_{\bar{X}_{sp}} = \sum_{i=1}^{3} W_i^2\, s^2_{\bar{X}_i} = \sum_{i=1}^{3} \frac{W_i^2 s_i^2}{n_i} \qquad \text{Eq. 6}$$

In this step, the standard deviation of each strata σ_i and the average of standard deviation $\bar{\sigma}$ need to be estimated as Equation 7 and 8 to prepare for the optimal allocation.

$$\sigma_i:\ s_i \qquad \text{Eq. 7}$$

$$\bar{\sigma}:\ \bar{s} = \sum_{i=1}^{3} W_i s_i \qquad \text{Eq. 8}$$

2.3 Optimal Allocation

Then, calculate the size of each strata with optimal allocation as Equation 9, in which *n* is the size of the sample, and n_l' is the size of each strata. Next, the new proportion of each strata in sample should be calculated as Equation 10. Now, the sample mean and estimated standard error can be computed to obtain the 95% confident interval for this stratified sample. All results are shown in Table 5.

$$n_l' = n\frac{W_l \sigma_l}{\bar{\sigma}}, \qquad l = 1, 2, 3 \qquad \text{Eq. 9}$$

$$W_l' = \frac{n_l'}{n}, \qquad l = 1, 2, 3 \qquad \text{Eq. 10}$$

2.4 Results

The results of three samples are shown below. From the table, it can be seen that the sample mean and standard error of each sample is fairly close. When the size of the sample is increased, although the confident interval becomes narrower, results of three sample are also similar. A stratified sample has no help in estimating the average family income. The reason of this case may be that the incomes of three family types are pretty closely.

Sample	Strata Size	Sample Mean	Standard Error	95% Confident Interval
Simple Random	300	39499.18	1758.56	[36052.40, 42945.96]
Proportional Allocation	228, 14, 58	41051.01	1851.36	[37422.35, 44679.67]
Optimal Allocation	239, 10, 51	40967.59	1793.90	[37451.55, 44483.63]

Table 5. All Results of Three Samples

2017/02/03