


Analysis of Household Yearly Income and Expenditure on Entertainment Using Resampling Techniques

Group 5: Xiyi (Britney) Wang, Terry Yuan, Yi Lu, Yuhao Zhuang

Introduction

Washington is the 13th-most population state, with more than 7.7 million people. As headquarters of many big companies like Microsoft, Amazon, Starbucks, Washington state has a relatively strong economy, with a total gross state product of \$612,996.5 million in 2019. With such high GDP, Washington State was able to provide the second highest minimum wage of the US^[4]. Moreover, Washington is a leading agricultural state. Up to 2018, the total value of Washington's agriculture has reached about \$10.6 billion. Washington also ranked first in many agricultural products such as raspberries, hops. However, since 2019 the economy of WA did not continue to grow due to the appearance of the COVID-19.

Coronavirus disease (COVID-19) is an infectious disease that can cause people to experience mild to moderate respiratory illness. The virus can spread from an infected person's mouth or nose in small liquid particles. The stay-at-home "lockdown" policy made by the WA government forced people to stay inside to prevent the spread of the virus. Without the labor force, the company cannot generate enough profit to sustain their business. The combined effect of unemployment and remote work inevitably caused a decrease in household income.

In this project, we want to know how the household income in urban, suburban and rural areas is distributed during the pandemic. Urban area is usually defined as a human settlement with a high population density and infrastructure of built environment. Suburban area then is mainly a residential area outside of a principal city of a metropolitan area. Different from the above two areas, the rural area is a geographic area that is located outside towns and cities. Therefore, the job categories will be different in these three areas due to their geographic difference. To study the effect brought by the virus to household income in these areas, we will first find their individual distribution. Then we would like to see whether there is a difference among these distributions. Also, in the project, we will employ the Parametric and Empirical Bootstrap to estimate some key numbers in the data set like mean, median and IQR. We want to compare the performance of these two resampling methods on estimating these key numbers of the household income data. 

Not only the household was harmed by the COVID-19 but also the entertainment industries in WA were affected. Therefore, in the second part of the project, we will study the relationship between the household entertainment cost and their income. Here, we are only interested in outdoor entertainment such as the cost of fairs, museums, movies, and restaurants. We will also assume that

they have a positive linear relationship because it makes sense that the lower the income the household gets, the lower money they will put on entertainment. Therefore, we want to test what degree does the household income relate to their entertainment expenditure during the pandemic.



Data Description

We generate data sets of household incomes and household expenditures on entertainment in units of one thousand dollars in urban, suburban and rural areas in WA. To bring the simulated data closer to reality, we use the following simulation methods. We first decide the sample size for each subpopulation in proportion to the real residential population in each area which is urban= 0.49 , suburban= 0.32 and rural= 0.19 ^[3] and have the total number of samples to be $10,000$. In reality we will get our data from financial surveys which investigate the living address, household income and expenditure on entertainment and will pick out 10,000 i.i.d samples based on the address and pick up the corresponding proportional number of surveys for each area. Then we will simulate all the samples from t-distribution and only take the non-negative part into account because both the income and the expenditure cannot be negative.

For the income data which is the independent variable, we plan to set parameters as the following Table 1 because we know that the average household income in WA is $\$77,006$ ^[1] and is $\$95,009$ ^[2] in King County which is an urban area.

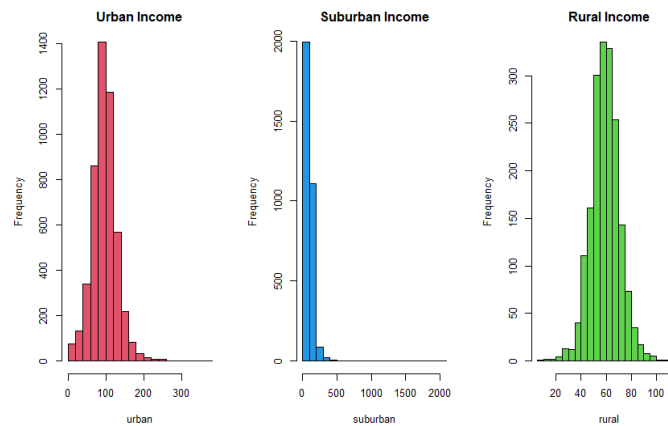


	Mean	SD	Degree of Freedom
Urban	95	25	4
Suburban	80	30	3
Rural	59	10	7

Table 1: Parameter of Distribution of Income(independent variable)

We assume the variance of suburban household income is large and degree of freedom is small because some people living in the suburban areas might choose the suburban as a livable place while having their well-paid job in urban areas. Because of that, there will be some outliers which represent the richest people in the suburban data set. In contrast to it, for rural areas, the standard deviation is relatively small and the degree of freedom is large because we think the income for rural areas is quite similar from place to place without large fluctuations. And in the urban areas, there will be some outliers due to the rich. Plot 1 below shows the histogram of simulated income data.





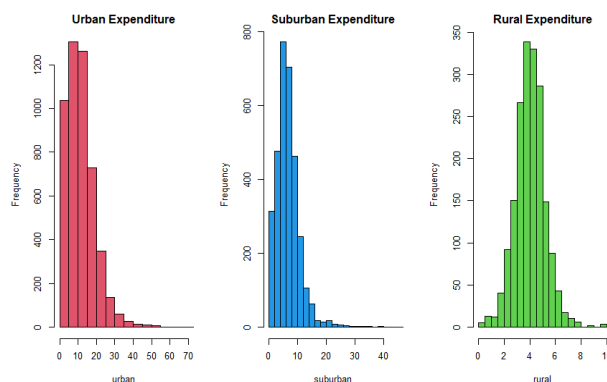
Plot 1: Histogram of Distribution of Income Data(independent variable)

For the expenditure part (dependent variable), we plan to set parameters as Table 2.

	Mean	SD	Degree of Freedom
Urban	10	7	4
Suburban	6	3	3
Rural	4	1	7

Table 2: Parameter of Distribution of Expenditure(dependent variable)

We think in reality, the entertainment venues and recreational facilities are mostly gathered in the urban areas, leading to more chances for urban people to spend money on them. Even though the fast living pace in urban areas might lead to a decreased enthusiasm for entertainment, there still exists a lot of high-cost entertainment which will lead to an overall right-skewed distribution. As for the suburban and rural areas, instead of having a high expenditure on entertainment, we think people there will spend a large proportion of money on something other than entertainment like transportation and food due to the inconvenience and the relatively low income in those areas. Plot 2 shows the histogram of simulated expenditure data.



Plot 2: Histogram of Distribution of Expenditure Data(dependent variable)




Methodology

In this section, we mainly discuss how we apply different methods to explore the difference between the distributions of each subpopulation, the performance of estimation of population, and the relationship between the household income and household expenditure on entertainment.

ECDF

To test the differences between the distribution of three subpopulations, the CDF curves drawn together on a same graph would give us the direct visual difference. Therefore, we drew the empirical CDF curves for each of our three subpopulations to see if they overlap with each other or not. To further test whether the ECDF curves come from different distributions, we will use T-test, KS-test and Permutation Test.

T-test

T-test is the first test method we chose to test whether the distributions of the three regions' household income come from the same distribution. We did the test between each of the three groups. This method tests on the mean difference between the samples. The null hypothesis (H_0) is that they have the same mean, which means that the difference between their mean would be around zero. The alternative hypothesis (H_1) is that their means are different. However, even if their mean is the same  that does not prove that they are from the same distribution, because other parameters of these distributions, such as median, IQR, sd, may be different. Therefore, we will further use other methods to prove the difference between the three sets of data. We used a significance level of 0.05 for the test.

KS-test

Another test method we used to test the difference between the three groups is the KS-test. We applied it to each two of the three groups. Unlike t-test testing on the mean difference between each of the subpopulations, KS-test is based on the maximum (vertical) difference between the two ECDF curves. The null hypothesis (H_0) is the two samples come from the same distribution. The alternative hypothesis (H_1) is that they come from different distributions. We still used a significance level of 0.05 for the test.

Permutation Test

Lastly, we used permutation tests to give a final shot on the difference between the subpopulations. Like KS-test, the Permutation Test tells us whether our three groups come from the

same distribution. The null hypothesis (H_0) is the two samples come from the same distribution. The alternative hypothesis (H_1) is that they come from different distributions. Permutation Test use statistic $(x_1, \dots, x_n, y_1, \dots, y_n)$ to measure the difference between observation of type-X and type-Y. We set the method resampling with each of the original subpopulation simulated data for 10,000 times. For here, we planned to first use the medians of the three income distributions as our test parameter. If their differences are small enough to reject H_0 , we would use standard deviations to test again. As the same as the last two test methods, we used a significance level of 0.05 for the test.

Bootstrap Sampling

In order to see the difference of household income distribution between each subpopulation, we also did analysis on both parametric and empirical bootstrap sampling to estimate the statistics of our simulated data, such as mean, median, and IQR. Here each sample statistic we get is an estimate of the statistic of population. Because they are estimators, we can compute the bias, variance, and mean square error (MSE) of sample statistics, thus leading to a better understanding of how good our estimated distributions are. In addition to these basic descriptive statistics, we also conducted bootstrap sampling to see the 99% quantile of the data since we want to explore whether our original distributions have high skewness, or heavy tails, and we are interested in those extreme values or outliers. For each replication process to create the bootstrap samples, we set $B = 10,000$. We also compared the performance of these two sampling methods.

Linear regression

We used the linear regression model to see the relationship between the dependent variables and independent variables. We used the least squares to find values of $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k$ that define a line that minimizes the squared distance between the points and line. The vector **beta hat** that minimizes the sum of squared distance is:

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

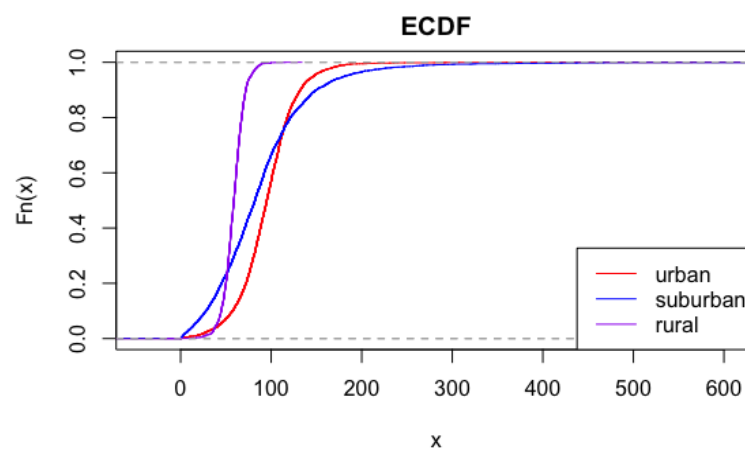
After we computed the estimated intercept and slope, we calculated their p-value using the t statistic to see whether our estimated value is statistically significant. In the next part, we checked the R-squared and adjusted R-squared to see how much of the variation of the response variable is explained by the explanatory variable. Finally, we can calculate the MSE of the linear regression method using the function below:

$$MSE = \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$$

Results

ECDF

Below is the figure of the empirical CDF curves for the household income in three different areas (urban, suburban, and rural). We can see that the three curves do not overlap with each other. The suburban area has the widest range of the increasing probability from 0 to 1. The rural area has the narrowest range, and the urban area is in the middle. This corresponds to how we set the data. We set our three subpopulations to t-distribution, but give them different parameters. The larger the standard deviation, the larger the range of ECDF. Therefore, based on the visual demonstration, we infer that the three subpopulations come from different distributions. This conclusion would be further supported by the tests in the next section.



Plot 3: Empirical CDF of Income Data(independent variable)

Hypothesis Tests

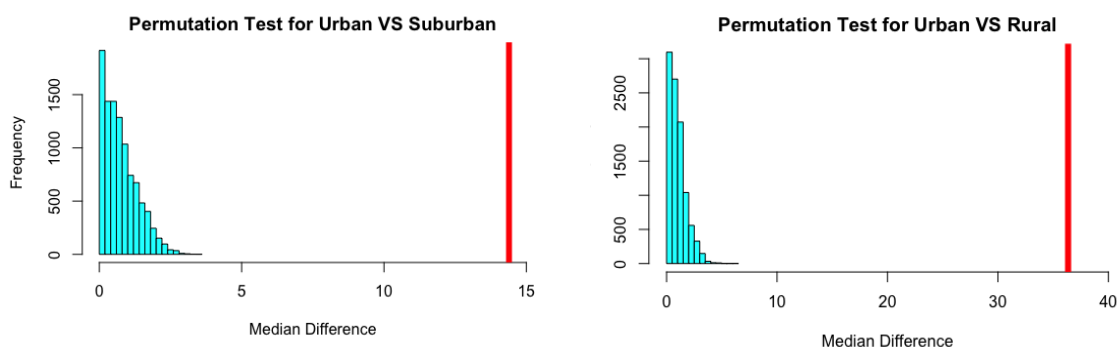
We concluded the results (p-values) of T-test, KS-test, and Permutation Test in Table 3 below. From the table, we can see that all of the p-values for T-tests and KS-tests and Permutation tests are below the significant level. The negative result for T-tests means that all of the three groups have different means. This corresponds to the parameters we set in our original data (urban: mean=95, suburban: mean=80, rural: mean=59). The **insignificant** results for KS-tests and Permutation tests show that all of the three groups come from different distributions. This is consistent with the fact that we set our three groups of data to different t-distributions, and also that the ECDF curves for the three subpopulations do not overlap with each other.

	t_Test <dbl>	KS_Test <dbl>	Permutation_Test <dbl>
Urban	3.238444e-24	0	9.999e-05
Suburban	0.000000e+00	0	9.999e-05
Rural	5.517033e-74	0	9.999e-05

Table 3 Test Results for Income(independent variable) Distribution

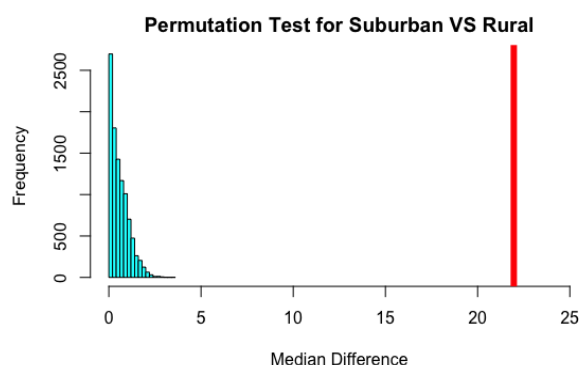
For the Permutation Test, we also got three result figures between each two of the three subpopulations, which are listed below. We can see that for all of the graphs, the red bar (true value) is far from the frequency histograms. They agree with small p-values calculated above: three data groups come from different distributions. Also, since the permutation tests on the medians have all rejected the hypothesis, we do not need to use another parameter to do the tests again.

In summary, based on the ECDF curves, the results of T-test, KS-test, and Permutation Test, all of our three income groups come from different distributions.



Plot 4: Permutation Test for Median Difference Between Urban and Suburban

Plot 5: Permutation Test for Median Difference Between Urban and Rural



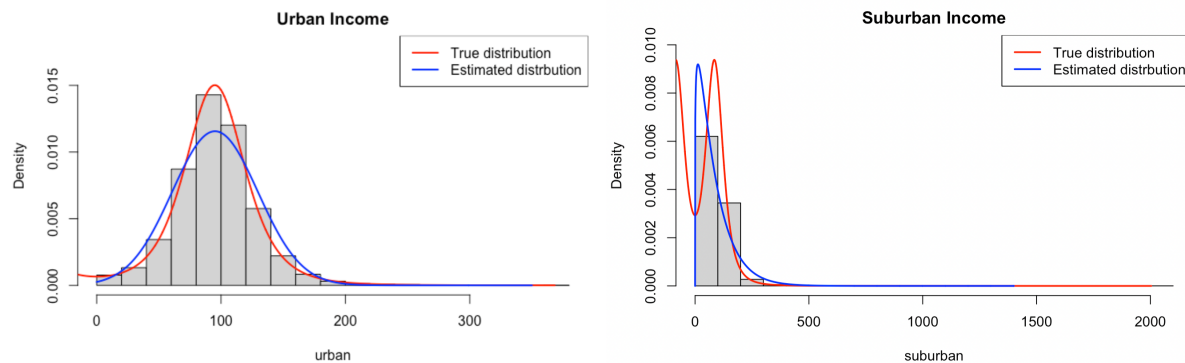
Plot 6: Permutation Test for Median Difference Between Suburban and Rural

Bootstrap Sampling

Given the conclusions from the previous hypothesis tests, we know that the household incomes in three areas come from different distributions. Thus when we did parametric bootstrap, we generated the samples from different approximated distributions which follow the shapes of the true distributions.

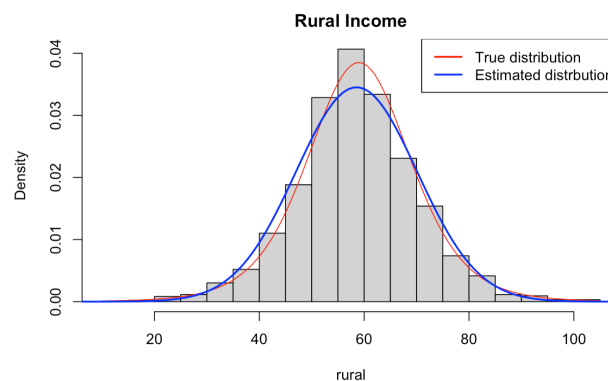
From the below plots of histograms and the density curves of original data, we discovered that the samples of both urban and rural areas are roughly normally distributed, while for suburban areas, since the household incomes there vary a lot, its distribution is more right-skewed. Therefore we used normal distribution to estimate urban incomes and rural incomes, where the input parameters are the

true mean and true standard deviation from the original data. For suburban income, we then used gamma distribution, where we calculate the alpha and beta based on the true mean and true variance($\alpha = \text{mean}^2/\text{variance}$, while $\beta = \text{mean}/\text{variance}$) to be the parameters. As for the sample sizes, since the density of population is largest in urban areas, next in suburban, and the smallest in rural areas, we should expect that the sizes of data we are approximating should also follow this relationship.



Plot 7: Histogram and density curves for urban income

Plot 8: Histogram and density curves for suburban income




Plot 9: Histogram and density curves for rural income

Then, we conducted both parametric and empirical bootstraps to estimate the statistics. Table 1 in the Appendix shows the results of parametric bootstrap sampling on the mean, the median, and the IQR. Table 2 shows the results of empirical bootstrap sampling on the median and the IQR.

Based on Table 1, for the subpopulations in urban areas and rural areas, we can discover that parametric bootstrap methods did well in estimating the mean and median of the original distributions, and the true values of both mean and median fall into each estimated confidence intervals. However, for the suburban samples, the estimated results fall apart from the true values, and the true values are not even in the confidence interval we get. That's mainly because the gamma distribution did not do well in capturing the pattern of the true suburban distribution. As for the IQR, which represents the range of the middle 50 percent of the data values, we have our estimated IQR of each subpopulation

greater than the IQR of each original sample, indicating our approximated distributions have heavier tails than the true distributions.

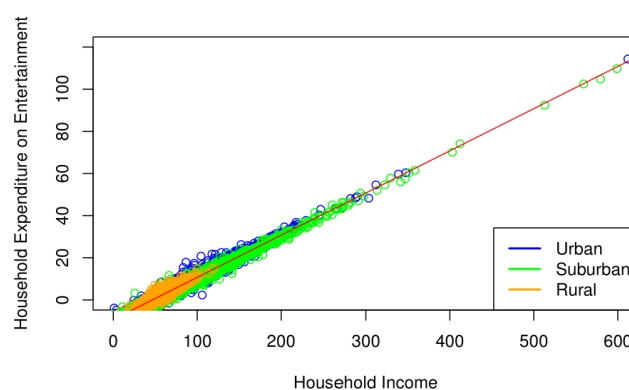
Based on Table2, since empirical bootstrap is resampling with replacement from the original dataset, it did well in estimating the original median and IQR of each subpopulation. And the variances of estimated values also agree with the MSE, implying that the estimators are unbiased.

Therefore, based on the results of these two tables, we found that empirical parametric does better in estimating the statistics of population with lower error. That's mainly because the parametric bootstrap generates its samples from the assumed distribution of the data, using the estimated parameter values. That means the results of estimators will be highly affected if the approximated distribution cannot capture the pattern of true distribution well. Also, since our sample sizes are large enough, we may expect that the non-parametric method could work well in estimating the true populations. 

As for the 99% quantile of the data in each subpopulation, based on the Plot 3 and Plot 4 in the appendix, we can discover that empirical bootstrap does better than parametric bootstrap in estimating the true 99% quantile value. We believe that's mainly because in parametric bootstrap, although the assumed distributions have the similar shape of the original distributions, they still cannot capture the whole pattern of the dataset, including some extreme values. Thus Plot 3 indicates that compared to our estimated distributions, the original distributions have heavier tails, and thus higher skewness.

Linear Regression

We ran the linear regression function in R using the simulated household income and household expenditure data. We then plot the result below:



Plot 10: Linear regression model on household income and entertainment expenditure

From the graph, we can see that we have seven outliers from the urban and suburban household income data. It is not surprising because we expect that there are some super rich people in these two



areas. From the plot we can see that the household income and expenditure follows a strong positive linear relationship, and both the R-squared and adjusted R-squared is about 0.9783 which means that 97.82% of variation of household entertainment expenditure is explained by their income. The fitted value of intercept and slope we calculated is -9.2596628 and 0.1999000, and their t test statistic is -327.9 and 671.9. Both t statistics are quite large, so we have enough evidence to support that the values of intercept and slope are significantly away from 0. Therefore, the linear regression model we get is:

$$\widehat{householdexp} = -9.2597 + 0.1999householdinc$$

where '*householdexp*' = household entertainment expenditure and '*householdinc*' = household income. We can see that the intercept has no real-life meaning when the household income is 0. The slope shows that one unit increase of the household will increase the household expenditure by about 0.2 unit. It means that during the pandemic, the income has only a small effect on the household expenditure on entertainment. The mean squared error is approximately 1.991 which is in our expectation since we use the least squares method.



Conclusion

In summary, by comparing the plots and mean square error generated by ECDF, T-test, KS-test, permutation test and bootstrap sampling, we find the distribution of household income from every subpopulation is definitely different. And we find that there is a **weak** positive linear relationship between household yearly income and expenditure on entertainment which means in reality, the higher income people get, slightly more expenditure they might have on entertainment. The results we get in the simulation and test process are very realistic because the differences in economic structure and economic development pace in different kinds of areas have led to differences in income and expenditure.

Reference

1. https://datacommons.org/place/geoId/53?utm_medium=explore&mprop=income&popt=Person&cpv=age%2CYears15Onwards&hl=en
2. <https://kingcounty.gov/independent/forecasting/King%20County%20Economic%20Indicators/Household%20Income.aspx>
3. [https://www.ruralhealthinfo.org/states/washington#:~:text=Washington%20covers%2066%2C544%20square%20miles,Washington%20\(USDA%20DERS\).](https://www.ruralhealthinfo.org/states/washington#:~:text=Washington%20covers%2066%2C544%20square%20miles,Washington%20(USDA%20DERS).)
4. [https://en.wikipedia.org/wiki/Washington_\(state\)#Western_Washington](https://en.wikipedia.org/wiki/Washington_(state)#Western_Washington)
5. https://www.who.int/health-topics/coronavirus#tab=tab_1

Appendix

Parametric Bootstrap

Subpopulation	True value (mean)	Estimated value (mean)	BT- Variance	BT - MSE	BT- 95%CI
Urban	95.65745	95.6549011	0.2570614	0.2570422	(95.6407552, 95.6690470)
Suburban	84.76037	86.657124	0.9457863	4.5433593	(86.6228668, 86.6913813)
Rural	58.94391	58.94468634	0.07124882	0.0712423	(58.93289611, 58.95647656)

Subpopulation	True value (median)	Estimated value (median)	BT- Variance	BT - MSE	BT- 95%CI
Urban	95.52546	95.6532074	0.4068579	0.4231365	(95.6354109, 95.6710039)
Suburban	85.81369	75.840778	1.253455	100.712364	(75.801341, 75.880216)
Rural	58.91194	58.9464271	0.1131109	0.1142887	(58.9315717, 58.9612825)

Subpopulation	True value (IQR)	Estimated value (IQR)	BT- Variance	BT - MSE	BT- 95%CI
Urban	37.9403	48.5183416	0.6530089	112.5479044	(48.4957955, 48.5408877)
Suburban	61.84028	67.341371	2.362153	32.623956	(67.287232, 67.395510)
Rural	13.72982	16.1686815	0.1825625	6.1305675	(16.1498086, 16.1875544)

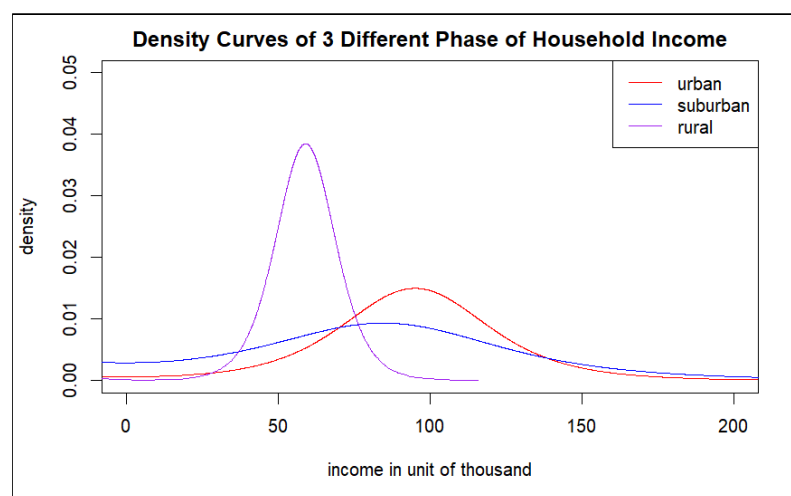
Table 1: Parametric Bootstrap Estimation Results

Empirical Bootstrap

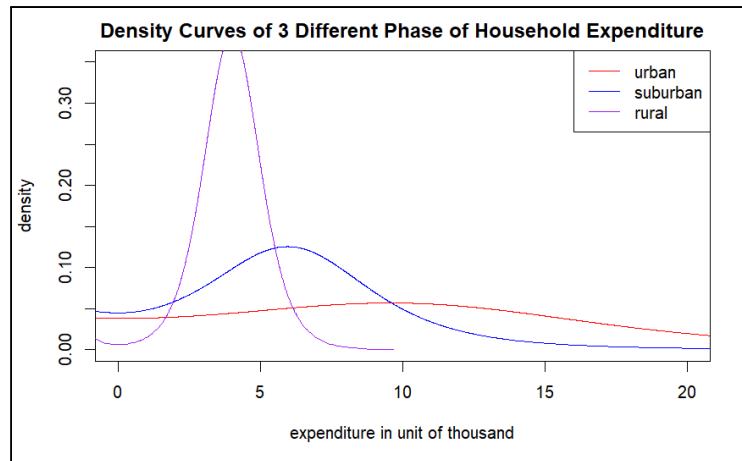
Subpopulation	True value (median)	Estimated value (median)	BT- Variance	BT - MSE	BT- 95%CI
Urban	95.52546	95.5057238	0.2459115	0.2462765	(95.4918881, 95.5195595)
Suburban	85.81369	85.7984276	0.9958404	0.9959739	(85.7705852, 85.8262701)
Rural	58.91194	58.90464583	0.06939035	0.06943669	(58.89729626, 58.91199540)

Subpopulation	True value (IQR)	Estimated value (IQR)	BT- Variance	BT - MSE	BT- 95%CI
Urban	37.9403	37.8809721	0.4841263	0.4875977	(37.8615592, 37.9003851)
Suburban	61.84028	61.643125	2.781418	2.820009	(61.584378, 61.701873)
Rural	13.72982	13.7728259	0.1603288	0.1621619	(13.7551395, 13.7905122)

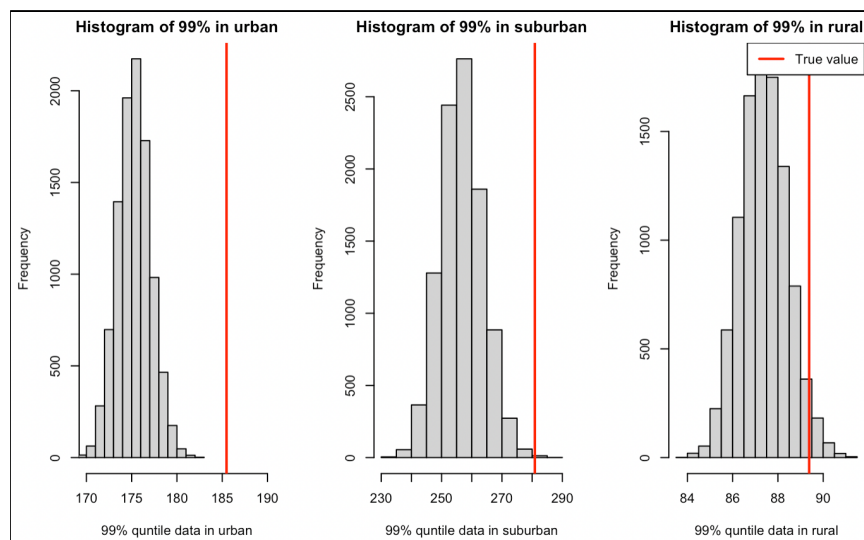
Table 2: Empirical Bootstrap Estimation Results



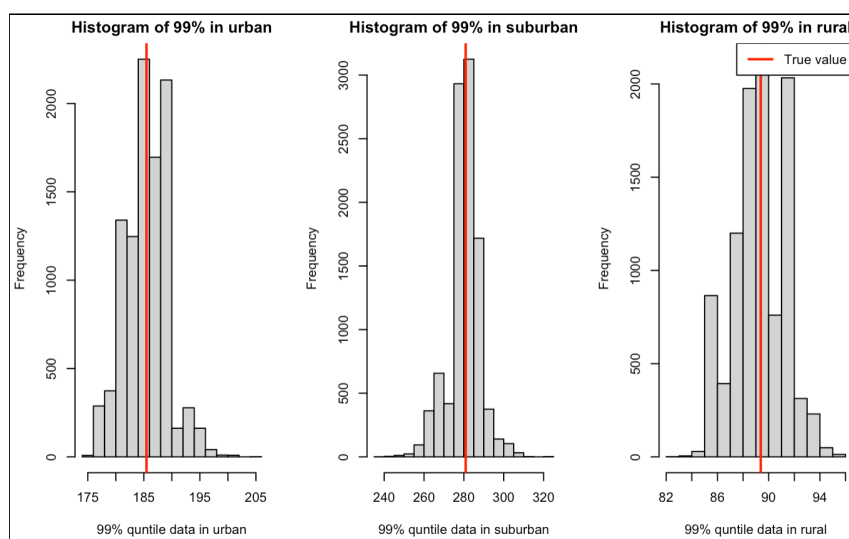
Plot 1: Density Curve of Distribution of Income Data(independent variable)



Plot 2 Density Curve of Distribution of Expenditure Data(dependent variable)



Plot 3: Histograms of 99% quantile data of Parametric bootstrap samples



Plot 4: Histograms of 99% quantile data of Empirical bootstrap samples