Jessica Smith
11602418
STATS 419
Monte J. Shaffer
8/30/2020

1. Where did the VLSS data come from? Do some research and provide a URL for a link to the official page with the data. Describe how you found it. How much does it cost to purchase? [Please don't buy it.] If you can find an online copy of the VLSS data, please also provide a link.

> To find an online source for the data, I first tried Googling the names of the data files: VLSSage.dat and VLSSperCapita.txt. This was unproductive, so next I tried looking in the appendix of the reading material, which it turns out was not provided. I then searched for "Vietnam Living Standards Survey (VLSS)" and found the following link: https://microdata.worldbank.org/index.php/catalog/2694
>
> The cost of the data is $500 (US) for citizens from a developed country, and is available via written request the General Statistical Office in Hanoi.

2. How were the 3 research questions derived? Are they constrained by the data? If so, how should you derive research questions?

> The first question, "What is the age distribution . . ." is supported by the data, which contains the ages of 28,633 individuals. The second question, "Are there differences in the annual household per capita expenditures . . ." is also supported by the available data, which contains household per capita expenditures for 5999 households along with demographic information. The third question "Are there differences in the annual household per capita expenditures . . ." requires a graphical summary of the data. The data contains a factor for whether each tuple is classified as rural or urban, as well as identifiers for 61 provinces that encompass seven regions of the country. Research questions should be derived with the limitations of the data in mind, so that the data can support finding answers to the questions that are being asked. These research questions are answerable with the given data.

3. Review the different graphs and the R code to generate them. From Figure 1.6, is there evidence to conclude that Urban homes have higher expenditures than Rural homes? How would you logically defend your conclusion?

> The density of rural expenditures is much narrower, whereas the urban distribution has a wider range. Both are positively skewed. The mean rural expenditure falls lower on the x axis than the mean urban expenditure. It would appear that many urban homes have higher expenditures per capita than most rural households.

4. How was Figure 1.7 plotted? What was the R code to do this?

> The text does not explicitly provide the R code that was used to generate Figure 1.7. I found several resources online for how to create maps similar to this in R. Here is an example video tutorial that could be used to reproduce Figure 1.7 or something similar: https://youtu.be/GMi1ThlGFMo.

5. From Figure 1.8 and Figure 1.9, can we conclude that the South East region has higher expenditures than the other regions? Would it be possible to graph similar plots of the data by both region (7 choices) and by Rural/Urban (2 choices)?

The boxplots in Figures 1.8 and 1.9 compare the sample means for each group along with a measure of the variation. Figure 1.9 shows the South East region has higher expenditures compared to the other six regions. Figure 1.8 also shows this, but it really stands out in Figure 1.9. We can conclude from these plots that the region has higher expenditures. It would be possible to make similar plots using the urban/ rural classifications, and it would be interesting to see this comparison as well. Based on earlier plots, I would expect the graph to show higher mean urban expenditures and greater sample variation.