Homework Assignment 1

Data Analysis 2

CEU Economics 2017/8

Submission deadline: Sunday 1 October 9.00 AM.

Late submissions won't be considered at all.

In this problem set you are going to have hands on exercise about some of the material covered in the first two classes. While you are allowed to cooperate among groups. If two groups submit identical or suspiciously similar solutions, both groups receive automatically zero point in the assignment.

Use the software that you are signed to and carefully document your steps.

Submit a .R or .do file with all your codes with comments and submit a separate pdf file with your graphs, tables, descriptives commenting your own results. Include in both files the group number and the names of the persons involved in each group.

1. Use the amazon_compare.csv file and load it into Stata or R.
2. Drop observations where price or price online or both are missing. Drop observations that are larger than the 99-percentile. How do you call these observations? Make an argument on why should we keep them in the sample. Make another argument on why should we drop them.
3. Create a random number generator from a distribution of your own choice setting the seed equal to your group number. Call this variable id_rand. Sort the variable and select the first 1000 observations in your dataset (**Non mandatory! Extra point**).
4. Make summary statistics (mean, standard deviations, quartiles, percentiles) of the variables price and online price. Are there any differences between these statistics?
5. Create variables two variables in your dataset: i) variance of the price variable, and ii) it's standard deviation (using the formula included in your Handout and Slides). How do they compare to the variance and standard deviations you got using Built-in commands (either in R or Stata).
6. Create a dummy (indicator) variable in case the price is greater than the online price. (**Non mandatory! Extra point**).
7. Create a variable that is the difference between the price and online price. Call this variable diff_price. Create a dummy variable taking value one if the diff_price is positive. How does this variable compare to the one above? (**Non mandatory! Extra point)**.
8. How often do you observe a positive value for the diff_price? How often do you observe no price difference between online and offline prices?

9. What is the probability of observing a positive price difference if the good category is Electronics? What is the probability of observing a zero price difference for the category Home and Appliances?
10. Create a scatterplot with price and online price. Interpret it. Create a scatterplot with price and amazon price. Interpret it. Do you see any difference between these two scatterplots? What are they telling us about the correlation between these variables? (**Non mandatory for PART TIME STUDENTS!)**.
11. Make box plots and histograms for the price difference over the categories Electronics and Pharmacy and Health. What can you tell us about variation in this variable? In which plot, histogram or boxplot, is easier to see the quartiles of the distribution?