



SOCIAL STATISTICS I
(SOCI 328/SECTION 202)
Professor Guy Stecklov

Exercise 1

Due Date: 16 March 2019

This is the first of two exercises due over the course of the semester. Each exercise is worth 20 percent of your grade. Each problem set can be completed on one's own or in a group of up to 3 persons in total. The names of each individual should be clearly marked on the problem set that is handed at the time of submission.

You will use the gapminder dataset that is publicly available and that we have repeatedly used throughout the semester. The gapminder dataset is obtained using the following two steps in the Rstudio Cloud website,

```
install.packages("gapminder")  
library(gapminder)  
gapminder <- as.data.frame(gapminder)
```

You will only be using the 2007 data from gapminder and we want to make sure the data is saved as a data frame. For these two reasons, you can use the following command to drop all other years that are not relevant and force the new object, gapminder, to be a dataframe:

```
gapminder <- as.data.frame(gapminder[gapminder$year==2007,])
```

All of your work needs to be CLEARLY documented and runnable as R code. Thus, you are required to submit your code for each question and sub-question. If the relevant code does not run then you do not get points for the answer.

1. Open the dataset.
 - a. Briefly describe the level of measurement of the following variables: (3 points)
 - i. Continent
 - ii. Year
 - iii. lifeExp
 - b. The variable lifeExp uses a particular naming convention in terms of caps and underscores that we discussed in class. What is it called? (1 point)
2. Create a boxplot for the lifeExp variable with separate boxes drawn for each continent on the same graph. Please point out the key differences that emerge when comparing continents.

Focus on both measures of central tendency and measures of variability that are visible in the boxplot. (3 points)

3. The World Bank employs a classification of countries in the world according to the Gross National Income. This is quite similar to our *gdpPercap* measure. There are four categories in this classification: high, upper-middle, lower-middle, and low.
- Your first task is to use a series of commands in R to create a new variable, *wb_inc_cat*, which is based on GNI and assigns each country in the data into one of the four categories: (2 points)

Threshold	July 2019/\$
Low income	<1,026
Lower-middle income	1,026 - 3,995
Upper-middle income	3,996 - 12,375
High income	> 12,375

- After you have created this new variable, *wb_inc_cat*, create a frequency distribution of the variable to show how many countries fall in each category and present this result. (1 points)
4. Create a histogram for the *gdpPercap* variable. Briefly explain the histogram and key features evident from its shape. Is the histogram skewed and if so in what way. Present your writing analysis with R output and commands. (3 points)
5. A national survey that studies the social statistics knowledge of Canadian adults with 3,500 respondents includes a question on the normal distribution. The responses to this question show that 40% of the respondents believe that the normal distribution is an empirical distribution, while the rest believe that it is a theoretical distribution.
- According to this sample, calculate the proportion of Canadians holding the correct view on the normal distribution and identify the 95% confidence level for this proportion. (Provide the response with precision to two decimal places.) (2 points).
 - Provide an interpretation of the confidence interval that has sufficient detail to comprehend the meaning of the confidence interval and its relationship to the population parameter. (1 point)
6. Last month, 35 students took a quiz. The mean for their marks is 63 with a standard deviation of 22. Answer the following questions assuming the distribution of the marks is normal:
- Anna receives 85 on the quiz, what is the Z score that corresponds to her score? (1 point)
 - Anna wants to find out how well she did compared to other students in the class on the quiz, please use z-table.com to find out the what proportion of students who took the quiz scores below than Anna. Explain your answer! (2 points)
 - The z score corresponding to Jay's score on the quiz is 74, what share of scores in the class are expected to fall between Jay and Anna's scores? (1 point)

7. Optional Bonus Credit Question (2 points): The variable for gdpPercap is measured in USD. The Canadian government has asked for analysis of these data and would like the data to be converted to CAD. The exchange rate for CAD is 1 USD = 1.34 CAD.

- a. Create a new variable, gdpPercapCad, as the Canadian dollar equivalent, and then proceed to calculate the standard deviation for this variable. Compare the standard deviation for gdpPercap and gdpPercapCad. Are they different if yes in what way? Also create a new variable, gdpPercapBonus, which is simply $\text{gdpPercap} + 100$. It turns out this new variable is the result of a new program by the World Bank to add 100 USD to every person in every country raising the average national income in every location by 100 USD. Calculate the standard deviation for gdpPercapBonus and compare it to gdpPercap. Are they the same or different.
- b. Can you draw any general insight to compare what happens when calculate the standard deviation for a variable after multiplying it by a constant factor versus adding a constant factor?