**ORCA 4500: Foundations of Data Science** Fall 2019

# Assignment 1

*Deadline: 09/20/2019 11:59pm*

**Instruction:** Please hand in two files on courseworks. One contains answers to the written questions. The second is a jupyter notebook that answers question 4.

**Exercise 1**

A study at UCLA investigated factors that might result in greater attention to the health and education of girls in rural India. One such factor is information about job opportunities for women. The idea is that if people know that educated women can get good jobs, they might take more care of the health and education of girls in their families, as an investment in the girls future potential as earners. The study focused on 160 villages outside the capital of India, all with little access to information about call centers and similar organizations that offer job opportunities to women. In 80 of the villages chosen at random, recruiters visited the village, described the opportunities, recruited women who had some English language proficiency and experience with computers, and provided ongoing support free of charge for three years. In the other 80 villages, no recruiters visited and no other intervention was made. At the end of the study period, the researchers recorded data about the school attendance and health of the children in the villages.

1. Did this analysis have a treatment group and a control group? If so, describe the two groups.
2. Was this an observational study or a randomized controlled experiment?
3. The study reported (Jensen, Department of Public Policy at UCLA, 2010), Girls aged 5-15 in villages that received the recruiting services were 3 to 5 percentage points more likely to be in school and experienced an increase in Body Mass Index, reflecting greater nutrition and/or medical care. However, there was no net gain in height. For boys, there was no change in any of these measures. Why do you think the author points out the lack of change in the boys?

**Exercise 2**

An ad for ADT Security Systems says, When you go on vacation, burglars go to work ... According to FBI statistics, over 25% of home burglaries occur between Memorial Day and Labor Day[1]. Do the data in the ad support the claim that burglars are more likely to go to work during the summer vacation period than at other times?

**Exercise 3**

Myopia, or nearsightedness, results from a number of genetic and environmental factors. In 1999, Quinn et al studied the relation between myopia and ambient lighting at night (for example, from night lights or room lights) during childhood.

1. The data were gathered by the following procedure, reported in the study. Between January and June 1998, parents of children aged 2-16 years ... that were seen as outpatients in a university pediatric ophthalmology clinic completed a questionnaire on the childs light exposure both at present and before the age of 2 years. Was this study observational, or was it a controlled experiment?

[1] According to Wikipedia, memorial Day is considered the unofficial start of summer in the United States, while Labor Day marks the unofficial start of Autumn on the first Monday of September.

2. The study found that of the children who slept with a room light on before the age of 2, 55% were myopic. Of the children who slept with a night light on before the age of 2, 34% were myopic. Of the children who slept in the dark before the age of 2, 10% were myopic. The study concluded that, The prevalence of myopia ... during childhood was strongly associated with ambient light exposure during sleep at night in the first two years after birth. Do the data support this statement? Why?

3. On May 13, 1999, CNN reported the results of this study under the headline, "Night light may lead to nearsightedness". Does the conclusion of the study claim that night light causes nearsightedness?

4. The final paragraph of the CNN report said that several eye specialists had pointed out that the study should have accounted for heredity. Myopia is passed down from parents to children. In what way do you think this fact might have affected the data?

**Exercise 4**
Create a jupyter notebook that does the following:

- Creates an array with the following daily temperature readings in Fahrenheit:

$$(50, 54, 66, 81, 80, 84, 85, 67, 45, 40, 42, 38, 50, 40, 30, 20, 25, 35, 40)$$

- Convert the array to Centigrade (or Celsius).
- Print the average temperature in Celsius.
- Print the maximum and minimum temperatures in Celsius.
- Print the number of days in which the temperature is more than the day before
- Print the biggest percentage increase and the biggest percentage decrease.

The jupyter notebook should intersperse text (the questions answered) and code to compute the quantities listed above.