Management Sciences, University of Waterloo
Statistical Methods for Data Analytics
Winter 2022
Instructor: Jangho Yang
TA: Yekta Amirkhalili & Muhammad Saadi Azi

Assignment 1

Due Feb 3 11:59 PM (Thursday)

*Note: The assignment submission should include i) a write-up for all your answers in a pdf form and ii) a separate R script. If you use Rmarkdown and combine i) and ii), you can submit the Rmd file only. The questions with *R coding only* do not require a written response. We will check your R coding only. Your R script needs to be fully commented, meaning that you need to explain your code line by line. Further, your R script needs to be self-sufficient so that graders can execute your code without any bug.

1. Probability calculation: As of Jan 18, 332,210,522 people in the world got COVID, and 5,494,965 died from it. The world population is 7,794,795,127. [10 point]

   (a) What is the probability of getting COVID? [2 point]

   (b) What is the probability of dying from COVID? [2 point]

   (c) What is the probability of dying conditional that one has COVID? [3 point]

   (d) See the chart on the infection fatality rate by age group in slide 6 in Topic 1. What is the probability of dying conditional that you have COVID and you're a 21 year old man? Give a rough number based on the chart. [3 point]

2. Simple urn simulation with R: There is an urn with 2 black balls (B), 3 white balls (W), and 3 red balls (R). We draw all of these 8 balls sequentially. [25 point]

   (a) What is the probability of observing B first and W last when we draw without replacement. The sequence between the first B and the last W does not matter. Calculate the probability by simulation. [15 point]

   (b) What is the probability of observing B first and W last when we draw with replacement. The sequence between the first B and the last W does not matter. Calculate the probability by simulation. [10 point]

3. Generalized urn simulation with R: The initial composition of colours in the urn is 20 black, 10 white, 7 red, and 7 green. When a ball of a certain colour is picked, we remove another ball of the same colour from the urn. For example, when a black ball is picked for the first draw, we remove another black ball before we start the second draw, so that there are 2 black balls removed in total. [35 point]

   (a) Simulate the composition of colours in each draw for 20 trials. *R coding only* [15 point]

   (b) Repeat this experiment 3 more times (with different set.seed). *R coding only* [10 point]

(c) Visualize the results of the above 4 experiments in a 2 by 2 plot. In each pot, visualize the time series of the number of balls of each colour. For example, see slide 33 in Topic 1. [10 point]

4. MLE for normal distribution [30 point]

   (a) Generate 50 random numbers with set.seed(200) from the normal distribution with $\mu = 0$ and $\sigma = 2$.*R coding only* [5 point]

   (b) Given this synthetic data in (a), estimate μ and σ using the MLE. Use **optim()** or **nlm()** function. Check whether the estimated parameters are close to the actual ones. [15 point]

   (c) Repeat this exercise with 5000 generated numbers instead of 50 numbers with set.seed(200). What difference can you tell? [10 point]

Bonus Questions

1 Read slides 51-52 in Topic 1. Calculate the probability of Sally Clark being innocent based on Bayes' theorem. Do you think Sally Clark is innocent? For simplicity, assume that there are only two possible causes of children's death: SIDS and murder. [10 point]

2 Cross-validation for Normal and Gamma distribution. [20 point]

   (a) Generate 100 random numbers from the normal distribution with $\mu = 10$ and $\sigma = 2$ using set.seed(200). This is your synthetic data.

   (b) Suppose you have two different models: the normal distribution and the Gamma distribution. Use a 10 fold cross-validation to see which model has a higher predictive power. [10 point]

   (c) Generate 100 random numbers from the Gamma distribution with $r = 2$ and $\lambda = 0.5$ using set.seed(200). Given this new synthetic data, repeat (b). Which one has a higher predictive power? [10 point]