Statistical Methods for Data Science (Spring 2018)

Mini Project 1

Instructions:

- Due date: Jan 25, 2018.
- Total points = 20.
- Submit a typed report.
- You can work on the project either individually or in a group of no more than two students. In case of the latter, submit only one report for the group, and include a description of the contribution of each member.
- It is OK to discuss the project with other students in the class (even those who are not in your group), but each group must write its own code and answers. If the submitted report (including code and answer) is similar (either partially or fully) to someone else's, this will be considered evidence of academic dishonesty, and you will referred to appropriate university authorities.
- Do a good job.
- You must use the following template for your report:

  Mini Project #
  Name
  Names of group members (if applicable)
  Contribution of each group member
  Section 1. Answers to the specific questions asked
  Section 2: R code. Your code must be annotated. No points may be given if a brief look at the code does not tell us what it is doing.

---

1. Consider a discrete random variable X that takes 4 values — 1, 2, 3 and 4 with respective probabilities 1/2, 1/8, 1/8, and 1/4.

   (a) Compute $E(X)$, $var(X)$ and $P(X \leq 2)$ analytically, i.e., using their formulas.

   (b) Explain how you would simulate a draw from the distribution of X.

   (c) Approximate $E(X)$, $var(X)$ and $P(X \leq 2)$ using Monte Carlo simulation with 1,000 draws 5 times. Summarize the results in a table.

   (d) Repeat (c) with 5,000 and 10,000 draws.

   (e) Compare you results in (a), (c) and (d). Explain, with justification, what you observe.

2. Suppose $X_1, X_2, \ldots, X_n$ denotes a random sample from a Bernoulli (p) population, represented by the random variable X, and let $\overline{X}$ denote the sample mean. This sample mean also represents the proportion of 1s in the sample, say, $\hat{p}$. We know from Central Limit Theorem that $\hat{p}$ approximately follows a normal distribution when n is large. The goal of this exercise is investigate how large n should be for the approximation to be good. For this investigation, we will focus on $p = 0.10, 0.25, 0.50, 0.75, 0.90$, and $n = 10, 30, 50, 100$.

   (a) What is the approximate distribution of $\hat{p}$ when n is large?

(b) For a given (n, p) combination, simulate 500 values of $\hat{p}$, and make a normal $Q - Q$ plot of the values. Does the distribution look approximately normal?

(c) Repeat (b) for the remaining combinations of (n, p) values.

(d) What would you say about how large n should be for the approximation to be good? Does this answer depend on p? Justify your conclusions.