Homework Assignment 1a

Due: Friday, Feb. 2, 2024, 11:59 p.m. Mountain time
Total marks: 35

Question 1. [35 MARKS]

To better visualize random variables and get some intuition for sampling, this question involves some simple simulations, which is a central theme in machine learning. You will also get some experience using `julia` and `pluto notebooks`, which you will also need to use in later assignments. Please see this document linked here with instructions on how to get started with Julia. Complete the attached notebook `A1.jl`.

In (b) and (c), the goal is to understand how much estimators themselves can vary: how different our estimate would have been under a different randomly sampled dataset. In the real world, we do not get to obtain different estimators, we will only have one; in this controlled setting, though, we can actually simulate how different the estimators could be.

In (d) and (e), the goal is to understand how we to obtain confidence intervals for our single sample average estimator.

(a) [5 MARKS] Fill in the code to calculate the sample mean, variance, and standard deviation of a vector of numbers. Do not use any packages not already loaded! Note that for the remainder of this question you will actually only use the sample mean outputted by your code, and will reason about the variability in this sample mean estimator. However, we get you to implement all three, for a bit of a practice.

(b) [7 MARKS] **WRITTEN:** Use your Julia implementation to generate 10 samples with $\mu = 0$ and $\sigma^2 = 1.0$, and compute the sample average (sample mean). Write down the sample average that you obtain. Now do this another 4 times, giving you 5 estimates of the sample average M_1, M_2, M_3, M_4 and M_5. What is the sample variance of these 5 estimates? Use the unbiased sample variance formula, $\bar{V} = \frac{1}{n-1}\sum_{i=1}^{n}(M_i - \bar{M})^2$. Note that here we want to understand the variability of the mean estimator itself, if it had been run on different datasets. Beautifully we can actually simulate this using synthetic data.

(c) [7 MARKS] **WRITTEN:** Now run the same experiment, but use **100 samples** for each sample average estimate. What is the sample variance of these 5 estimates? How is it different from the variance when you used 10 samples to compute the estimates?

(d) [8 MARKS] **WRITTEN:** Now let us consider a higher variance situation, where $\sigma^2 = \mathbf{10.0}$. Imagine the data comes from a zero-mean Gaussian with this variance, but pretend you do not know the mean. Run the code to get **30 samples**, and compute one sample average M. What is the 95% confidence interval around this M? Give actual numbers.

(e) [8 MARKS] **WRITTEN:** Now assume you know less: you **do not know** the data is Gaussian, though you still know the variance is $\sigma^2 = 10.0$. Use the same 30 samples from (d) and resulting sample average M. Give a 95% confidence interval around M, now without assuming the samples are Gaussian.

Homework policies:

Your assignment should be submitted as two pdf documents and a .jl notebook, on eClass. **Do not** submit a zip file with all three. One pdf is for the written work, the other pdf is generated from the .jl notebook. The first pdf should contain your answers for questions starting with "**WRITTEN:**". Your answers must be written legibly and scanned or must be typed (e.g., Latex). This .pdf should be named Firstname_LastName_Sol.pdf, For your code, we want you to submit it both as .pdf and .jl. To generate the .pdf format of a Pluto notebook, you can easily click on the circle-triangle icon on the right top corner of the screen, called Export, and then generate the .pdf file of your notebook. The .pdf of your Pluto notebook as Firstname_LastName_Code.pdf while the .jl of your Pluto notebook as Firstname_LastName.jl. All code should be turned in when you submit your assignment.

Because assignments are more for learning, and less for evaluation, grading will be based on coarse bins. **The grading is atypical**. For grades between (1) 81-100, we round-up to 100; (2) 61-80, we round-up to 80; (3) 41-60, we round-up to 60; and (4) **0-40, we round down to 0**. The last bin is to discourage quickly throwing together some answers to get some marks. The goal for the assignments is to help you learn the material, and completing less than 50% of the assignment is ineffective for learning.

We will not accept late assignments. There is no late penalty policy. The assignments must be submitted electronically via eClass on time, by 11:59 pm Mountain time on the due date. There is a grace period of 48 hours when assignments will be accepted. No submissions will be accepted after 48 hours after the deadline, and the assignment will be considered as incomplete if not submitted.

All assignments are individual. All the sources used for the problem solution must be acknowledged, e.g. web sites, books, research papers, personal communication with people, etc. Academic honesty is taken seriously; for detailed information see the University of Alberta Code of Student Behaviour.

Good luck!