Homework Assignment # 1

Due: Friday, Jan. 27, 2023, 11:59 p.m. Mountain time
Total marks: 100

Question 1. [15 MARKS]

Let X be a random variable with outcome space $\Omega = \{a, b, c\}$ and $p(a) = 0.1, p(b) = 0.2$, and $p(c) = 0.7$. Let

$$f(x) = \begin{cases} 10 & \text{if } x = a \\ 5 & \text{if } x = b \\ 10/7 & \text{if } x = c \end{cases}$$

(a) [4 MARKS] What is $E[f(X)]$?

(b) [3 MARKS] What is $E[1/p(X)]$? (Note: This is an expectation about the probability)

(c) [4 MARKS] For an arbitrary pmf p, what is $E[1/p(X)]$? (Note: This is an expectation about the probability)

(d) [4 MARKS] What is $E[f(X)^2]$ and $E[f(X)]^2$?

Question 2. [15 MARKS]

Suppose you have three coins. Coin A has a probability of heads of 0.75, Coin B has a probability of heads of 0.5, and Coin C has a probability of heads of 0.25.

(a) [5 MARKS] Suppose you flip all three coins at once, and let X be the number of heads you see (which will be between 0 and 3). What is the expected value of X, $E[X]$?

(b) [10 MARKS] Suppose instead you put all three coins in your pocket, select one at random, and then flip that coin 5 times. You notice that 3 of the 5 flips result in heads while the other 2 are tails. What is the probability that you chose Coin C? Hint: Define random variable D as the observed data, and notice that you can compute $p(D = 3 \text{ heads and 2 tails} | \text{Coin} = C)$.

Question 3. [10 MARKS]

Alberta Hospital occasionally has electrical problems. It can take some time to find the problem, though it is always found in no more than 10 hours. The amount of time is variable; for example, one time it might take 0.3 hours, and another time it might take 5.7 hours. The time (in hours) necessary to find and fix an electrical problem at Alberta Hospital is a random variable, say X, whose density is given by the following uniform distribution

$$p(x) = \begin{cases} \frac{1}{10} & \text{if } 0 \leq x \leq 10 \\ 0 & \text{otherwise} \end{cases}$$

Such electrical problems can be costly for the Hospital, more so the longer it takes to fix it. The cost of an electrical breakdown of duration x is x^3. What is the expected cost of an electrical breakdown? Show your work.

Question 4. [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. Complete the attached notebook `A1.jl` and follow the `instructions.md` to get setup.

For the first two questions, 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.

For the second two questions, 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] 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] 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] Now let us consider a higher variance situation, where $\sigma^2 = \mathbf{10.0}$. Imagine you know this variance, and that the data comes from a Gaussian, but that you do not know the true 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] 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.

Question 5. [25 MARKS]

We have talked about the fact that the sample mean estimator $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is an unbiased estimator of the mean μ for identically distributed $X_1, X_2, \ldots, X_n$: $\mathbb{E}[\bar{X}] = \mu$. The straightforward variance estimator, on the other hand, is not an unbiased estimate of the true variance σ^2: for $\bar{V}_b = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$, we get that $\mathbb{E}[\bar{V}_b] = (1 - \frac{1}{n})\sigma^2$. Instead, the following bias-corrected sample variance estimator is unbiased: $\bar{V} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$. This unbiased estimator is typically what is called the *sample variance*.

(a) [15 MARKS] Use the fact that $\mathbb{E}[\bar{V}] = \sigma^2$ to show that $\mathbb{E}[\bar{V}_b] = (1 - \frac{1}{n})\sigma^2$. **Hint:** The proof is short, it can be done in a few lines. To start, consider if you can rewrite $\bar{V}_b$ in terms of $\bar{V}$ and then use your expectation rules.

(b) [10 MARKS] We also discussed the variance of the sample mean estimator, and concluded that $\text{Var}[\bar{X}] = \frac{1}{n}\sigma^2$, for iid variables with variance σ^2. We can similarly ask what the variance is of the sample variance estimator. Deriving the formula is a bit more complex for general random variables, so let's assume the X_i are zero-mean Gaussian. For zero-mean Gaussian X_i, we can use $\bar{V} = \frac{1}{n}\sum_{i=1}^{n} X_i^2$, which is unbiased, i.e., $\mathbb{E}[\bar{V}] = \sigma^2$. (Note that this $\bar{V}$ is different from 5a. Here the estimator subtracts the **true mean**, which we know is zero, from each sample. The previous estimator subtracted the **sample mean**. This is why this estimator is unbiased although it divides the sum by n while the previous estimator had to divide the sum by $n-1$ to be unbiased). Then we know that the following is true (though we omit the derivation): $\text{Var}[\bar{V}] = \frac{2(n-1)}{n^2}\sigma^4$.

This variance enables us to use Chebyshev's inequality, to get a confidence interval. Recall that Chebyshev's inequality states that for a random variable Y with known variance v, we know that

$$\Pr(|Y - \mathbb{E}[Y]| < \epsilon) > 1 - \frac{v}{\epsilon^2}$$

For some confidence level δ, this means we can set $\delta = v/\epsilon^2$ and solve for ϵ in terms of the two knowns: δ and v. Derive the confidence interval for $\bar{V}$, using Chebyshev's inequality.

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 .jl notebook. The first pdf containing your answers of the write-up questions 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) 80-100, we round-up to 100; (2) 60-80, we round-up to 80; (3) 40-60, 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. Plan for this and aim to submit at least a day early. If you know you will have a problem submitting by the deadline, due to a personal issue that arises, please contact the instructor as early as possible to make a plan. If you have an emergency that prevents submission near the deadline, please contact the instructor right away. Retroactive reasons for delays are much harder to deal with in a fair way. If you submit a few minutes late (even up to an hour late), then this counts as being on time, to account for any small issues with uploading in eClass.

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!