Homework 8

Due November 21 at 11 pm

1. (Short questions)

   (a) Give an example of a nonnegative random variable $\tilde{a}$ and a constant $c > 0$ for which $P\{\tilde{a} \geq c\} = \tilde{a}/c$. What does this say about Markov's inequality?

   For our example, lets use a non-negative bernoulli random variable, let's call it $\tilde{a}$, who's pdf is defined in the following way:

   $$\begin{cases} \tilde{a} = 1 & with\ P(\theta) \\ \tilde{a} = 0 & with\ P(1-\theta) \end{cases}$$

   With $c = 1$. Then what we would have is the following inequality:

   $$P(\tilde{a} \geq c) \leq \frac{E(\tilde{a})}{c} \longrightarrow 1 \leq 1$$

   Since we have equality in the example we have used, there is no better bound than Markov's inequality.

   (b) In the notes, we have defined the sample variance of a dataset $X := x_1, x_2, \ldots, x_n$ as

   $$\sigma_X^2 := \frac{\sum_{i=1}^n (x_i - \mu_X)^2}{n}, \qquad (1)$$

   where μ_X is the sample mean. Show that this is not an unbiased estimator of the true variance if the data are i.i.d. samples from a distribution with zero mean and variance σ^2. Explain how to fix the estimator so that it is unbiased.

Intuitively, we understand that the samples we derive from a distribution of i.i.d. variables will always be subject to randomness and thus noise. Therefore, no matter what, the sample variance will be a biased estimator of the true population variance. We can show this by taking the expected value and expanding our initial equation, and doing a few math tricks to illustrate how it does not equal the true population variance of σ^2.

$$\begin{aligned}
E(\sigma_X^2) &= E(\frac{\sum_{i=1}^n (x_i - \mu_X)^2}{n}) \\
E(\sigma_X^2) &= \frac{1}{n} \times E(\sum_{i=1}^n (x_i^2 - 2\mu_X x_i + \mu_x^2)) \\
E(\sigma_X^2) &= \frac{1}{n}(E(x_i^2 + n\mu_X^2 - 2n\mu_X^2)) \\
E(\sigma_X^2) &= \frac{1}{n}((^2_X + \mu^2) - n(\frac{\sigma_X^2}{n} + \mu_X^2)) \\
E(\sigma_X^2) &= \frac{1}{n}(n\sigma_X^2 - \sigma_X^2) \\
\sigma_X^2 &\neq \sigma_X^2 - \frac{^2_X}{n} \quad \square
\end{aligned} \qquad (2)$$

We could divide our initial equation by $n-1$ instead of n and it would yield:

$$\frac{1}{n-1} \times (n-1) \times \sigma_X^2 = \sigma_X^2$$

and thus would unbias our estimator.

2. (Poll) In an online poll before an election, 60 participants intend to vote for the Democratic candidate, and the remaining 40 intend to vote for the Republican candidate.

   (a) The number of young people (between 18 and 35 years old) in the poll is 70. 50 intend to vote for the Democratic candidate. The fraction of young people among voters in general is 25%. Provide an estimate of the proportion of voters that will vote for the Democratic candidate.

   For the whole population we will have:

   $$(\frac{1}{4} \times \frac{5}{7}) + (\frac{3}{4} \times \frac{1}{3}) = \frac{3}{7}$$

   Who will vote Democrat.

   (b) Under what assumptions is your estimate unbiased? Justify your answer mathematically.

   The estimate will only be unbiased if within the two groups we have defined our population: young and not young, each behave and are sampled in i.i.d. fashion, that is to say they will have uniform likelihood and are sampled with replacement. Since we understand the proportions of young and not young voters in our population, we need to ensure that we get an unbiased estimate of the proportion of each group that votes democrat.

   (c) Let α be the proportion of young people in the population, θ_1 the proportion of young people who vote for the Democratic candidate, and θ_2 the proportion of old people who vote for the Democratic candidate. If the number of young people in a poll is n_1 and the number of old people is n_2, what is the variance of your estimator if the assumptions from the previous question hold?

   Assuming that we sample with i.i.d. fashion, let our estimator be the following: $Estimator = \alpha\theta_1 + (1-\alpha)\theta_2$. Then we can calculate the variance as such:

   $$\begin{aligned} Var(Estimator) &= E(Estimator^2) - E(Estimator)^2 \\ Var(Estimator) &= (\alpha\theta_1)^2 + 2\theta_1\alpha\theta_2(1-\alpha) + (\theta_2(1-\alpha))^2 - (\theta_1\alpha + \theta_2 - \alpha\theta_2)^2 \quad (3) \\ Var(Estimator) &= 0 \end{aligned}$$

   Thus, since the population as a whole only has one proportion of Democratic voters, if we are estimating it and we know the population of young and not young and the proportion of democratic voters in these groups, our estimate will not vary from the population proportion, as we are essentially using ground truth population statistics in our estimates.

3. (Length of confidence interval) We are interested in estimating the mean height in a population from a finite set of random samples. We would like to have a 95% confidence interval for our estimate of width equal to 5 cm.

   (a) Use Chebyshev's inequality to determine how many samples we need to take. Explain any assumptions you make.

   Lets assume that the sample standard deviation would have to be bounded at 75. Then we would have:

$$\begin{aligned} \frac{\sigma}{\sqrt{\alpha n}} &= \frac{width}{2} \\ \frac{75}{\sqrt{.05n}} &= 2.5 \\ 2.5 \times \sqrt{.05n} &= 75 \\ .05n &= 30^2 \\ n &= 18,000 \end{aligned} \tag{4}$$

   (b) Use the central limit theorem to determine how many samples we need to take, assuming that the sample standard deviation of the data equals 10 cm.

   We are assuming $\sigma = 10$ therefore:

$$\begin{aligned} \frac{\sigma}{\sqrt{n}}\theta^{-1}(1 - \frac{\alpha}{2}) &= \frac{width}{2} \\ \frac{10}{\sqrt{n}}\theta^{-1} \times .975 &= 2.5 \\ n &= \text{ 61.46 round up} \\ n &= 62 \quad \square \end{aligned} \tag{5}$$

   Therefore, we would need to take at least 62 samples.

4. (Radioactive sample) Consider the following experiment. We have a radioactive sample situated at unit distance from a line of sensors. Each time a sensor detects a particle emitted from the sample we obtain a reading of the position of the sensor in the x axis (we assume that we have so many sensors that you can model this position as a continuous random variable). We model the measurements as an i.i.d. sequence distributed as a random variable $= c + \tilde{x}$ where the pdf of $\tilde{x}$ is symmetric around the origin, that is $f_{\tilde{x}}(x) = f_{\tilde{x}}(-x)$ for all real numbers x. Your task is to estimate the position of the sample c from these data.

Figure 1: Diagram of the experiment.

   (a) The file *radioactive_sample_1.txt* contains a vector of measurements $m_1, m_2, \ldots$. Plot a moving average of the measurements $\frac{1}{n}\sum_{i=1}^{n} m_i$ for $n = 1, 2, 3, \ldots$ (and submit the plot). Use the plot to give an estimate for the value of c. (Hint: What is the expected value of ?)

Under what assumptions on $\tilde{x}$ can you prove that the estimation method you propose work?

If we assume $\tilde{x}$ is an iid. random variable and we treat each observation as an iid. experiment on a random variable, then we know that the law of large numbers should come into effect and the sample mean should converge to the population mean. This also assumes that there will be no anomalies with the data due to the sensor position.

(b) The file *radioactive_sample_2.txt* contains a vector of measurements corresponding to a different radioactive sample. Does the estimation method described above work in this case? Submit the plot of the new moving average.

The assumptions and approach with the first samples do not work necessary with the separate samples as the means do not tend to approach any given number.

(c) A colleague suggests that the angle α between the trajectory of the particles emitted by the new sample and the vertical axis (illustrated in Figure 1) might be well modeled by a random variable $\tilde{a}$ that is uniformly distributed between $-\pi/2$ and $\pi/2$. Compute the pdf and mean of $\tilde{x}$ under this assumption.
(Hint: remember the trigonometric function tan and its inverse arctan.)
Would such model explain your observations in (b)?

Lets calculate the CDF of $\tilde{x}$:

$$\begin{aligned} CDF\ of\ \tilde{x} &= P(\tilde{x} \leq x) \\ &= P(tan(\tilde{a}) \leq x) \\ &= P(\tilde{a} \leq arctan(x)) \\ &= \int_{\frac{-\pi}{2}}^{arctan(x)} \frac{\tilde{a}}{\pi} da \\ &= \frac{arctan(x)}{\pi} + \frac{p\imath}{2\pi} \end{aligned} \tag{6}$$

Now lets calculate the PDF of $\tilde{x}$:

$$PDF\ of\ \tilde{x} = \frac{arctan(x)}{\pi} \frac{d}{dx} = \frac{1}{\pi(1+x^2)}$$

Which we can identify as the Cauchy PDF. Therefore, $E(\tilde{x})$ doesn't exist which is why our estimate does not converge to any given value.

(d) The sample mean can be affected by extreme values and outliers, whereas the sample median is more robust. The sample median converges to the median of an iid sequence of random variables even when the mean is not well defined. Use the sample median from *radioactive_sample_2.txt* to estimate c.

The sample median of the second sample text document is about 3.5, which makes sense as the sample median of the position of the sensors should correspond to the median of c because the best estimator of c for any value of m is equal to the value at m.

In [19]:

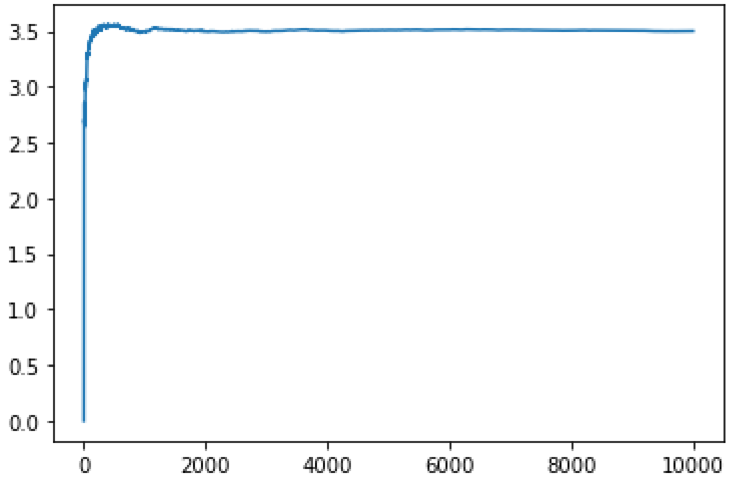
```
#Import libraries
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
```

In [20]:

```
#Import data
data1 = pd.read_csv("radioactive_sample_1.txt",sep=" ",header=None)
data2 = pd.read_csv('radioactive_sample_2.txt',sep=" ",header = None)

#Transpose the data accordignly
data1=data1.T
data2=data2.T

#Drop the na values
data1 = data1.dropna()
data2 = data2.dropna()

#Convert to list
data1list= data1[0].tolist()
data2list= data2[0].tolist()
```

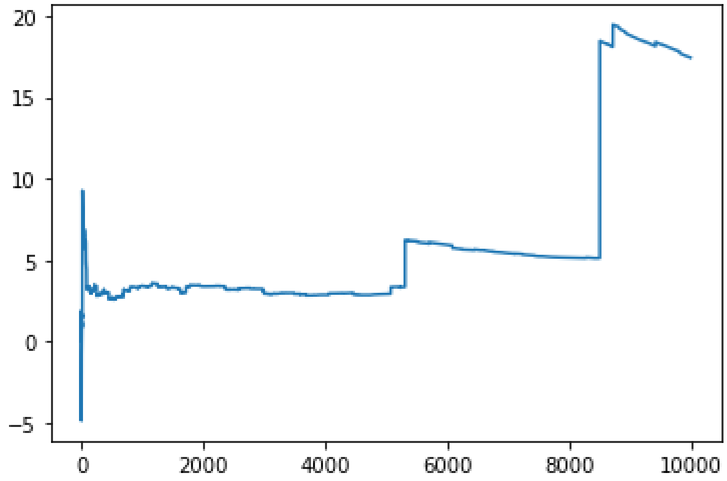
In [21]:

```
#Initialize list variables
data1mov = list()
data2mov = list()

#Loop through the values in data1list
for i in range(len(data1list)):
    data1mov.append(sum(data1list[:i])/(i+1))

#Plot the law of large numbers demonstration
plt.plot(data1mov)
plt.show()
```



In [22]:

```
#Loop through the numbers in data2list
for i in range(len(data2list)):
    data2mov.append(sum(data2list[:i])/(i+1))

plt.plot(data2mov)
plt.show()
```



C)

In [23]:

```
#Print the sample median, which appears to approach 3.5
print(np.round(np.median(data2list),decimals=6))
```

```
3.494066
```

We know that the best estimator of C for any value of m is equal to that value of m, thus the sample median of the position of the sensors should correspoond the the median of c.