EE 131A | Matlab project
Probability | Monday, November 26, 2018
Instructor: Professor Roychowdhury | Due: Friday, December 7, 2018

In this project we will further analyze random variables and their various properties. We will also investigate how are random variables used to model practical systems. Each part will have a combination of MATLAB programming, mathematical analysis and technical writing. You will be graded on all three components.

When producing your plots **clearly indicate** the x-axis, the y-axis and what is being plotted (using legends, title etc.). You may need to rescale x-axis to ensure that your plot is showing the right quantity.

Make sure to attach in the appendix of your project report **all MATLAB programs** that you used to generate the data.

1. *Simulation of Game of cards.* Recall the bonus problem in the midterm:

   Shuffle a deck of 52 cards randomly and lay them down one after the other. What is the probability that the first King is immediately followed by the first Ace?

   In this problem, we run simulations in MATLAB to estimate this probability p.

   Consider the following experiment:

EXPERIMENT 1
A shuffle generates a random permutation of 52 cards. Keep shuffling cards until the desired pattern occurs, then record the number of shuffles it takes. The number of shuffles until the desired pattern is a random variable, denoted as X.

   We denote the number of repetition of this experiment as m.

   The values of X observed are denoted as $\{x_i : i \in \{1, 2, \ldots, m\}\}$.

   (a) Verification of randomness: 1st order and 2nd order

   Repeat the shuffling for $100,000$ times. Record the shuffling results in a 100000×52 sized matrix.

   i. Check whether at each position, each card is equally likely to show up. Specifically, **within each column, count the number of occurrences of each of the 52 cards, and see whether the numbers are evenly distributed. This should result in 52 sets of 52 count numbers**.

ii. Similarly, check whether at the following pairs of positions, each of 52×51 card pairs is equally likely to show up.
$(5, 20)$ $(45, 51)$ $(2, 32)$
In this case, for each pair of positions, you will have $52 \times 51 = 2652$ count numbers. **Please plot one histogram with 50 bins of the count numbers for each pair of positions.**

(b) Let $m = 100,000$; that is, repeat the experiment 1 for $100,000$ times. Record the x_i's.

i. Plot the distribution of the x_i's you get. Specifically, calculate the vector N, so that N_n is the number of x_i's that are equal to n. $n \in \{1, 2, \ldots, \max(\{x_i\}_{i=1}^m)\}$.

For example, if we have 5 x_i's, namely,

$$x_1 = 2, x_2 = 1, x_3 = 4, x_4 = 1, x_5 = 4$$

then

$$N = [N_1, N_2, N_3, N_4] \quad \text{and} \quad N_1 = 2, N_2 = 1, N_3 = 0, N_4 = 2$$

Scatter plot N_n vs n.

ii. Show that X follows a geometric distribution. What is the relation between its parameter and p?

Hint: Probably useful functions: `randperm`, `histcounts`, `plot`

(c) Assume that X follows a geometric distribution with parameter p.

i. What is the probability $P(X = k \mid p)$?

ii. We can also see $\{x_i\}_{i=1}^m$ as samples of m random variables $\{X_i\}_{i=1}^m$ respectively, where $X_i \overset{\text{iid}}{\sim}$ **Geometric**(p).
What is the probability that one takes m independent samples of X and gets the x_i's you observed? (i.e. what is $P(X_1 = x_1, \ldots, X_m = m_m \mid p)$?)

iii. What is the value of p that can maximize $P(X_1 = x_1, \ldots, X_m = m_m \mid p)$? Show that this value is

$$\hat{p} = \frac{1}{\bar{x}} = \frac{m}{\sum_{i=1}^m x_i}, \text{ where } \bar{x} \equiv \frac{\sum_{i=1}^m x_i}{m}$$

Hint: first take the logarithm of $P\big(\{x_i\}_{i=1}^m \mid p\big)$, then find the p that makes the derivative zero.

(d) $\hat{p}$ is called the max-likelihood estimator of p.
Note that $\hat{p}$ is also the ratio of the "valid" permutations among the entire m permutations.

i. Calculate $\hat{p}$ for $m = 100,000$

ii. Compare the variance of $\hat{p}$ for different m:

Calculate $\hat{p}$ for 100 times for $m = 30, 100, 300, 1000, 3000$, respectively.
(1) Report the sample variance of the $\hat{p}$'s for each m. (2) Use the `boxplot` function to plot a boxplot where each m value corresponds to a box.

(e) One can also make use of the empirical distribution N to estimate p.
Let $m = 100,000$, use `polyfit` to fit a line for $\log(N)$ vs n. Use the slope to estimate p.

(f) One may find that it is quite intuitive to first estimate $E[X]$ using sample mean $\overline{X} = \frac{\sum_{i=1}^{m} X_i}{m}$ then use the relation $E[X] = \frac{1}{p}$ to estimate p. Is it also valid if we use $\overline{\left(\frac{1}{X}\right)} = \frac{\sum_{i=1}^{m} \frac{1}{X_i}}{m}$ to estimate p?

i. Show that $E\left[\overline{X}\right] = E[X]$ and $E\left[\overline{\frac{1}{X}}\right] = E\left[\frac{1}{X}\right]$

ii. Use Markov inequality to show $E\left[\frac{1}{X}\right] \geq p$.

iii. *Calculate $E[\frac{1}{X}]$. Is it equal to p?

2. *Simulation of binary transmission systems* A binary transmission system sends a "0" bit using a -2 voltage signal and a "1" bit by transmitting a $+2$. The received signal is corrupted by a noise N that has a Laplacian distribution with parameter α. Assume that "0" and "1" bits are equiprobable.

(a) Write a MATLAB program to generate Laplacian distribution with parameter $\alpha = 2, 0.5$ respectively. Use the transformation method (refer to Section 4.9 in the textbook) based on $t = 10,000$ samples of the unit-interval uniform random variable. Plot the resulting empirical (Laplacian) cdf and pdf in each case.

(b) Assume the received signal Y is given by $Y = X + N$. Suppose that the receiver decides a "0" was sent if $Y < 0$, and a "1" was sent if $Y \geq 0$. Write a MATLAB program to simulate the transmission of 10,000 bits under the SNR = 0dB for this channel and compute the empirical error probability.

(c) Derive analytically the expression for the error probability under SNR = 0dB. How does the analysis compare to your simulation from part (b)?

For the above, SNR denotes the Signal-to-Noise Ratio and SNR(dB)=$10\log_{10}(\frac{1}{\sigma^2})$. You need to express the standard deviation of Laplacian distribution, i.e. σ, in terms of its parameter α.

3. *Central Limit Theorem* Let X_1,X_2,... be a sequence of iid random variables with finite mean μ and finite variance σ^2, and let S_n be the sum of the first n random variables in the sequence:

$$S_n = X_1 + X_2 + ... + X_n.$$

Use $t = 10,000$ samples in the questions below.

(a) Let X_i be a continuous uniform random variable taking values in the interval $(1, 6)$. Write a MATLAB program to plot the empirical pdf and cdf of S_n. Consider $n = 1, 5, 15, 50$ and compare your results.

(b) Calculate analytically the mean and the variance of X_i and of S_n in part (a).

(c) Write a MATLAB program to draw the pdf and cdf curves of the Gaussian distribution with the same mean and variance as S_n. Superimpose this plot with the plots from part (a).

(d) Similarly, let $Y_1, Y_2 \ldots$ be another sequence of iid random variables following the exponential distribution with $\lambda = 1$, and let $S_n^{'}$ be the sum of the first n random variables. Based on your observation and understanding, pick a suitable n, write a MATLAB program to plot the empirical pdf and cdf of $S_n^{'}$, and superimpose the corresponding Gaussian distributions.