ECE 131A - LEC 80 Class Project
Probability and Statistics Monday, September 2nd, 2024
Instructor: Lara Dolecek TA: Zilai WANG

Please upload your project to Gradescope by September 2nd, 11:59pm.
Please submit a single PDF directly on Gradescope
You may type your project report or scan your handwritten version. Make sure all the work is discernible.
100 points total

In this project we will further analyze random variables and learn about their utility in practical systems. Each part will have a combination of programming, mathematical analysis, and technical writing. You will be graded on all 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.

The preferred programming language is MATLAB but you **may use** other programming languages. Make sure to attach in the appendix of your project report **all programs/code** that you used to generate the data and plots.

1. **[25 points]** *Tossing a fair and unfair die.* Suppose you have a 5-sided die, with sides numbered 1,2,3,4,5.

   (a) Write a MATLAB program to simulate the tossing of a 5-sided fair die, for t =10, 50, 100, 500 and 1000 tosses. Based on the simulation, what is the estimated probability of obtaining an odd number?

   (b) Suppose X is a random variable denoting the outcome of a die toss. Based on the mathematical analysis, what is the probability that X has odd value?

   (c) Refer back to part (a). Does it agree with the theoretical result in (b)?

   (d) Repeat parts (a), (b), and (c) with a 5-sided die that has the following properties:
   - The probability of the die outcome being 1 is equal to the probability of the die outcome being 2.
   - The probabilities of the die outcome being a 3, 4, or 5 are all equal.
   - The probability of the die outcome being a 1 is twice the probability of the die outcome being 5.

   You may find useful the MATLAB function rand that generates a uniform random value in the (0, 1) interval.

2. **[25 points]** *Computer Methods for Generating Random Variables*

   (a) Plot the pdf and cdf of the gamma random variable for the following case: $\lambda = \frac{1}{2}$ and $\alpha = \frac{1}{2}, 1, \frac{3}{2}, \frac{5}{2}$

   (b) Specify the transformation method needed to generate the geometric random variable with parameter $p = \frac{1}{2}$. Find the average number of comparisons needed in the search to determine each outcome.

   (c) Let the number of event occurrences in a time interval be a Poisson random variable. It is known that the time between events for a Poisson random variable is an exponentially distributed random variable.

      i. Explain how one can generate Poisson random variables from a sequence of exponentially distributed random variables.

      ii. Use Matlab to implement the two methods when $\alpha = 3$, $\alpha = 25$, *and* $\alpha = 100$

3. **[25 points]** *Naive Bayes Classifier.* On April 15, 1912, the largest passenger liner ever made collided with an iceberg during her maiden voyage. When the Titanic sank it killed 1502 out of 2224 passengers and crew. This sensational tragedy shocked the international community and led to better safety regulations for ships. One of the reasons that the shipwreck resulted in such a loss of life was that there were not enough lifeboats for the passengers and crew. Although there was some element of luck involved in surviving the sinking, some groups of people were more likely to survive than others.

   Some demographic data are collected for 887 passengers and are provided in *titanic.csv*.The price class, gender (1 for Male and 0 for Female) and age are recorded for each survived (1) or killed (0) passenger. We use random variables S, C, G, and A for survival status, price class, gender and age, respectively. In this problem, we are going to build a popular Naive Bayes Classifier to predict S given C, G, and A.

   (a) Estimate the PMFs for S, C, G, and A by finding the fraction of each realization of these random variables among all data. Plot these PMFs.

   (b) We are interested in how C, G, and A affect S, in the context of the *Naive Bayes Classifier*; the first step is to estimate the conditional PMF conditioning on the outcome of interest, i.e., survival or not. Estimate and plot the conditional PMFs for C, G, and A conditioned on S.

   (c) In the *Naive Bayes Classifier*, we use the conditional independence assumption. For example, if C, G, and A are conditionally independent on S, then

   $$P(C, G, A|S = 0) = (C|S = 0)P(G|S = 0)P(A|S = 0)$$

   and

   $$P(C, G, A|S = 1) = (C|S = 1)P(G|S = 1)P(A|S = 1)$$

   Using this assumption, compute

   $$P(S = 0, C = 1, G = 0, A \leq 40) \textit{ and } P(S = 1, C = 1, G = 0, A \leq 40)$$

   based on your estimations in (a) and (b).

(d) Based on your result in (c), compute

$$P(S = 0|C = 1, G = 0, A \leq 40); and\ P(S = 1|C = 1, G = 0, A \leq 40)$$

Predict whether a female whose age is under 40 and who is in first class will survive or not.

4. **[25 points]** *Central Limit Theorem.* Let $X_1, X_2, X_3, \ldots$ be a sequence of i.i.d. random variables with finite mean μ and finite variance σ^2, and let Z_n be the sum of the first n random variables in the sequence:

$$Z_n = \frac{1}{n} \sum_{i=1}^{n} X_i$$

(a) Let X_n for $i = 1, 2, \ldots$ be a uniform continuous random variable taking values in the interval (3,7). Write a MATLAB program to plot the pdf of Z_n. Consider $n = 1, 3, 10, 30, 100$ and compare your results across different n's.

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

(c) Write a MATLAB program to generate a Gaussian random variable with the same mean and variance as Z_n. Superimpose its pdf on the plots from part (a).

(d) Repeat parts (a), (b), and (c) with X_i representing a toss of a fair 5-sided die that is described in Problem 1(d). Note that X_i and Z_n are discrete in this case. **Hint.** You can calculate the PDF analytically or empirically. For the latter method, use $t = 10^4$ samples and while plotting the histogram for discrete data, use 'BinWidth' as $\frac{1}{n+1}$.