CS 215: Data Analysis and Interpretation
Assignment: Distributions, Expectation

Instructor: Suyash P. Awate

Submission Instructions:

• IITB and CSE have zero tolerance to plagiarism.

• For the sake of effective learning, if you submit the solution to the assignment as a group, then each member of the group agrees to have participated fully (100%) in performing every part of every question in the assignment.

• If you submit the solution to the assignment (by yourself or as a group), then you agree that every line of code and every line in the report is your (or your group's) own, and isn't a copied/modified version of any other source online (on the internet) or offline (in electronic form or paper form or any other form).

• If you submit the solution to the assignment as a group and any member of the group is determined to have committed any (non-zero amount of) plagiarism, then the full penalty (a reduction of at least 1 letter grade, e.g., from AB to BB or less) will be applicable to every member of the group. The penalty will be applicable to givers and takers both.

• Submit your solution to each problem, i.e., (i) the code, (ii) the results, e.g., graphs or other data, and (iii) the report (in Adobe PDF format), for each question, through moodle. Put the code within the folder "code", the results within the folder "results", and the report within the folder "report".

• Submit all code that allows the TAs to regenerate your results, exactly as they appear in the report.

• Submit a single zip file that contains the solutions to all problems in the assignment.

• To get any possible partial credit for the code, ensure that the code is very well documented. To get partial credit for the derivations, include all derivation steps in their full details.

• To avoid non-deterministic results in each program run, and to make the results reproducible during test time, use rng(seed) where seed is a fixed hard-coded integer in your code.

• If the question suggests the use of some function in Matlab, then you can use a corresponding function in other coding frameworks/languages.

• Delayed submissions will be penalized 25% of the total points on each day after the deadline, i.e., submitting anytime within the first 24 hours after the deadline will incur a penalty of 25% of the total points.

• If you feel there is a typo in the question, please make suitable assumptions, consistent with those in the question, and proceed to solve the problem. Also, in that case, please let the TAs or the instructor know.

• **5 points** are reserved for submission in the proper format.

1. (15 points)

   For each of the following distributions, do:

   (i) plot the probability density function (PDF) based on the analytical expression. The PDF must appear smooth enough and without apparent signs of discretization.

   (ii) plot the cumulative distribution function (CDF) using Riemann-sum approximation. The CDF must appear smooth enough and without apparent signs of discretization.

   (iii) use Riemann-sum approximations to compute the approximate variance (if finite) within a tolerance of 0.01 of its true value known analytically.

   • [6 points: 1 + 2 + 3]

   Laplace PDF (`en.wikipedia.org/wiki/Laplace_distribution#Probability_density_function`) with location parameter $\mu := 2$ and scale parameter $b := 2$.

   • [6 points: 1 + 2 + 3]

   Gumbel PDF (`en.wikipedia.org/wiki/Gumbel_distribution`) with location parameter $\mu := 1$ and scale parameter $\beta := 2$.

   • [3 points: 1 + 2]

   Cauchy PDF (`en.wikipedia.org/wiki/Cauchy_distribution#Probability_density_function`) with location parameter $x_0 := 0$ and scale parameter $\gamma := 1$.

2. (15 points)

Consider two independent Poisson random variables X and Y, with parameters $\lambda_X := 3$ and $\lambda_Y := 4$.

• [5 points: 3 + 1 + 1]

Define a random variable $Z := X + Y$, having a probability mass function (PMF) $P(Z)$.

(i) Empirically obtain an estimate $\widehat{P}(Z)$ of the PMF $P(Z)$, by drawing $N := 10^6$ instances (sample points) of X and Y both. You may use the poissrnd(.) function in Matlab. Report the values of $\widehat{P}(Z = k)$ for $k = 0, 1, 2, \cdots, 25$.

(ii) What will the PMF $P(Z)$ be theoretically/analytically ?

(iii) Show and compare the values for $\widehat{P}(Z = k)$ and $P(Z = k)$ for $k = 0, 1, 2, \cdots, 25$.

• [10 points: 6 + 2 + 2]

Implement a Poisson thinning process (as discussed in class) on the random variable Y, where the thinning process uses probability parameter 0.8. Let the thinned random variable be Z.

(i) Empirically obtain an estimate $\widehat{P}(Z)$ of the PMF $P(Z)$, by drawing $N := 10^5$ instances (sample points) from Y. You may use the poissrnd(.) and binornd(.) functions in Matlab. Report the values of $\widehat{P}(Z = k)$ for $k = 0, 1, 2, \cdots, 25$.

(ii) What will the PMF $P(Z)$ be theoretically/analytically ?

(iii) Show and compare the values for $\widehat{P}(Z = k)$ and $P(Z = k)$ for $k = 0, 1, 2, \cdots, 25$.

3. (30 points)

Simulate $N := 10^4$ independent random walkers (as discussed in class) along the real line, each walker starting at the origin, and each walker taking 10^3 steps each of length 10^{-3}.

• [5 points]

Plot a histogram of the final locations of all the random walkers. You may use the hist(.) function in Matlab.

• [5 points]

For the first 10^3 walkers, plot space-time curves that show the path taken by each walker (as depicted in the class slides). On the graph, draw each path in a different randomly-chosen color for better clarity.

• [10 points]

Submit the well-documented code for all of the above.

• [5 points: 1 + 4]

Consider a random variable X. Consider a dataset that comprises N independent draws (e.g., modeled by $X_1, \cdots, X_N$) from the distribution of X.

Use the law of large numbers to show that the random variable $\widehat{M} := (X_1 + \cdots + X_N)/N$ converges to the true mean $M := E[X]$ as $N \to \infty$.

Prove that the expected value of the random variable $\widehat{V} := \sum_{i=1}^{N} (X_i - \widehat{M})^2/N$ tends to the true variance $V := \text{Var}(X)$ as $N \to \infty$. It can also be shown that the variance of the random variable $\widehat{V}$ tends to zero as $N \to \infty$; the proof is a bit tedious though (see `https://mathworld.wolfram.com/SampleVarianceDistribution.html`).

• [1 points]

Report the empirically-computed mean $\widehat{M}$ and the empirically-computed variance $\widehat{V}$ of the final locations of the random walkers.

• [3 points]

What should the values of the true mean and the true variance be for the random variable that models the final location of the random walker, as function of the step length and the number of steps ?

• [1 points]

Report the error between the empirically-computed mean and the true mean.

Report the error between the empirically-computed variance and the true variance.

4. (25 points)

Consider a continuous random variable X that has an M-shaped probability density function (PDF) $P_X(\cdot)$ as follows:

$P_X(x) := 0$ for $|x| > 1$, and

$P_X(x) := |x|$ for $x \in [-1, 1]$.

Consider independent continuous random variables $\{X_i : i = 1, 2, \cdots, \infty\}$ with PDFs identical to that of X.

Define random variables $Y_N := (1/N) \sum_{i=1}^{N} X_i$, for $N = 1, 2, \cdots, \infty$, which have associated distributions $P_{Y_N}(\cdot)$.

• [5 points]

Write code to generate independent draws from $P_X(\cdot)$. Your code can use only the uniform random-number generator rand() (no other generator). Submit this code.

• [5 points: 2 + 3]

Show plots of (i) the histogram (with 200 bins) and (ii) cumulative distribution function (CDF), both using $M := 10^5$ draws from the PDF $P_X(\cdot)$.

• [8 points]

Use the code written in the previous sub-question to write code to generate independent draws from $P_{Y_N}(\cdot)$. Submit this code.

• [7 points: 3 + 4]

Show plots (separately) of histograms using 10^4 draws from each of the PDFs $P_{Y_N}(\cdot)$ for $N = 2, 4, 8, 16, 32, 64$.

Show plots, on the same graph, of all the CDFs associated with Y_N for $N = 1, 2, 4, 8, 16, 32, 64$, computed using 10^4 draws from each $P_{Y_N}(\cdot)$. Plot each CDF curve using a different color. You may use the cdfplot(.) function in Matlab.

5. (25 points)

Generate a dataset comprising a set of N real numbers drawn from the uniform distribution on $[0, 1]$.

Consider various dataset sizes $N = 5, 10, 20, 40, 60, 80, 100, 500, 10^3, 10^4$.

For each N, repeat the following experiment $M := 100$ times:

(i) first, generate the data,

(ii) then, compute the average $\widehat{\mu}$, and

(iii) finally, measure the error between the computed average $\widehat{\mu}$ and the true mean μ as $|\widehat{\mu} - \mu_{\text{true}}|$.

• [5 points] For the uniform distribution, plot a single graph that shows the distribution of errors (across M repeats) for all values of N using a box-and-whisker plot. You may use the boxplot(.) function in Matlab.

• [5 points] Repeat the above question by replacing the uniform distribution by a Gaussian distribution with $\mu := 0$ and $\sigma^2 := 1$.

• [5 points] Interpret what you see in the graphs. What happens to the distribution of error as N increases ?

• [10 points: 5 + 5] Submit the well-documented code for both uniform and Gaussian cases.