ENGR 5310 Probability and Random Process
Final Project

Dr. Semih Aslan & Dr. Xiaohua (Nemo) Luo
Due: Thursday, November 30, 2023 or
Friday, December 8, 2023

During this project, you will learn how to write a scientific paper about using Python for probability and statistics. You need to complete all these parts to get full credit.

Part I: Python Basics
Part II: Python Graphs
Part III: Basic Statistics and probability using Python
Part IV: Linear Regression

You need to submit your project as a .zip file. First, create a folder and name it "YourName_LastName_Project." Create a document file (MS Word) and put all your figures and codes here. You can perform copy/paste or screen capture. Also, you need to put all your Python files in your folder. Make sure you save all your code file names similar to one in the project description. After completing all your work, please zip your folder and submit it to CANVAS.

PS: It is fine if you prefer to use other program language, please make sure you can have the same results and submit results and the efficient code.

Part I - Python Basics:

This portion of the project you need to work on the basics of Python and standard libraries. Please make sure to assign your code names similar functionality of your code as well as section numbers.

1.1. *Data Types, arithmetic, and logical operators:* Please write data types in Python and give an example. Also, use arithmetic operation in Python for these data types.

1.2. Strings: Please write how to use strings in Python. Try to write a program that displays a useful code in another language. This is useful for automation and verification.

1.3. User Inputs and outputs: Please write an example to take user input from the terminal and print them. You may give an example of the average calculator.

1.4. List, Tuples, Dictionaries: Give information about lists, tuples, and dictionaries. Indicate differences and usage. Give some examples of list manipulation and creation.

1.5. Conditionals (if/else) and Loops (for/while): Please explain conditionals and loops. Give some examples and compare some of the operations. Is there a “case” conditional in Python? Can you create one?

1.6. Function: Describe the function and give some examples.

1.7. *Task:* Please complete the following tasks (20 points).
1.7.1 Create a 4x4 matrix A
1.7.2 Create a 1x4 matrix B
1.7.3 Create a 4x1 matrix C
1.7.4 Combine (column) A and B
1.7.5 Combine (row) A and C
1.7.6 Create an array (1x20) with all random values (an integer between 0 and 9) and name it x.
1.7.7 Write a function to calculate mean, standard deviation, max, min, and medium of a list.

Part II - Python Graphs:

This portion of the project you need to work on the basics of Python graphs using Matplotlibs.

2.1. *Basic line graph:* Using Matplotlib to work on basic line plots. Also, work on subplots that 1x2, 2x2, 4x4 plots with some examples.

2.2. *Bar Charts, Pie Charts, and Pareto Charts:* Using Matplotlib to work on basic bar and pie charts.

2.3. *Histogram:* Using Matplotlib to work on the basic histogram.

2.4. *Boxplots:* Using Matplotlib to work on basic boxplots.

2.5. *Task:* Please complete the following tasks (20 points).

2.5.1 Create a 10x10 sparse matrix A and visualize it.
2.5.2 Create a plot that shows sine(x), cos(x) and sin(x)cos(x) on same plot.
2.5.3 Create Bar chart and pie chart chart using any random data.
2.5.4 Create a histogram using random data (you may use data on Matplotlib website)
2.5.5 Create a boxplot with different quantiles.
2.5.6 Save figures with different file types (png, svg, jpeg, pdf) and compare them based on file size and scalability.
2.5.7 Create and change figure titles, x and y-axis labels, grid thickness, and etc.

Part III - Basic Statistics and Probability Using Python:

This portion of the project you need to work on basics data distribution using Python.

3.1. *Binomial Distribution:* The binomial distribution is a discrete probability distribution. It describes the outcome of n independent trials in an experiment. Each trial is assumed to have only two outcomes, either success or failure. If the probability of a successful trial is p, then the probability of having x successful outcomes in an experiment of n independent trials is as follows.

$$f(x) = \binom{n}{x} p^x (1-p)^{(n-x)} \text{ where } x = 0, 1, 2, ..., n$$

3.1.1.a *Problem:* Suppose there are twelve multiple-choice questions in an English class quiz. Each question has five possible answers, and only one of them is correct. Find the probability of having four or less correct answers if a student attempts to answer every question at random.

3.1.1.b *Answer:* The probability of four or fewer questions answered correctly by random in a twelve-question multiple choice quiz is 92.7%.

3.2. *Poisson Distribution:* The Poisson distribution is the probability distribution of independent event occurrences in an interval. If λ is the mean occurrence per interval, then the probability of having x occurrences within a given interval is:

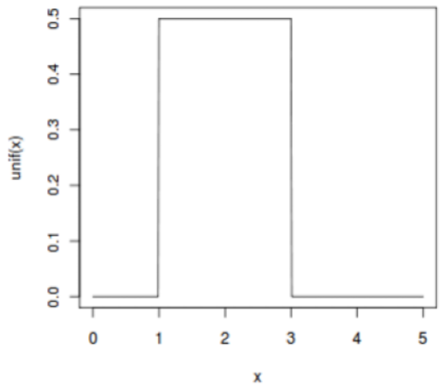
$$f(x) = \frac{\lambda^x e^{-\lambda}}{x!} \text{ where } x = 0, 1, 2, ...$$

3.2.1.a *Problem:* If there are twelve cars crossing a bridge per minute on average, find the probability of having seventeen or more cars crossing the bridge in a particular minute.

3.2.1.b *Answer:* If there are twelve cars crossing a bridge per minute on average, the probability of having seventeen or more cars crossing the bridge in a particular minute is 10.1%.

3.3. *Continuous Uniform Distribution:* Here is a graph of the continuous uniform distribution with a = 1, b = 3.

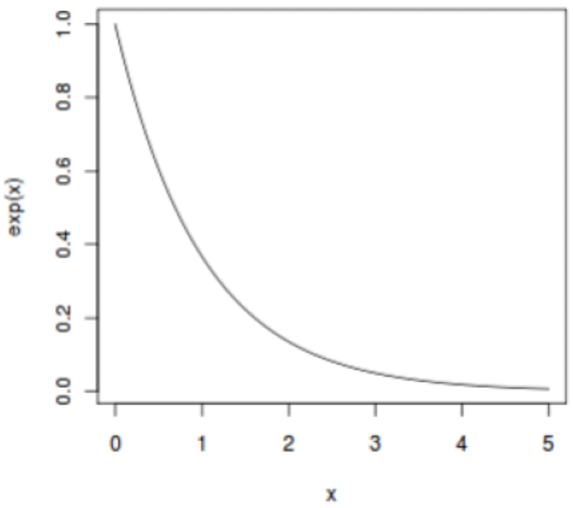
$$f(x) = \begin{cases} \frac{1}{b-a}, & a \le x \le b. \\ 0, & \text{otherwise.} \end{cases}$$



3.4. *Exponential Distribution:* The exponential distribution describes the arrival time of a randomly recurring independent event sequence. If μ is the mean waiting time for the next event recurrence, its probability density function is:

$$f(x) = \begin{cases} \frac{1}{\mu} e^{-x/\mu}, & x \geq 0. \\ 0, & \text{otherwise.} \end{cases}$$

Here is a graph of the exponential distribution with μ = 1.



3.4.1.a *Problem:* Suppose the mean checkout time of a supermarket cashier is three minutes. Find the probability of a customer checkout being completed by the cashier in less than two minutes.

3.4.1.b *Answer:* The probability of finishing a checkout in under two minutes by the cashier is 48.7%

3.5. *Normal Distribution:* The exponential distribution describes the arrival time of a randomly recurring independent event sequence. If μ is the mean waiting time for the next event recurrence, its probability density function is:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

If a random variable X follows the normal distribution, then we write X ~ N(μ, σ^2). In particular, the normal distribution with μ = 0 and σ^2 = 1 is called the standard normal distribution and is denoted as N(0,1). It can be graphed as follows.

The normal distribution is important because of the Central Limit Theorem, which states that the population of all possible samples of size n from a population with mean μ and variance σ^2 approaches a normal distribution with mean μ and σ^2 / n when n approaches infinity.

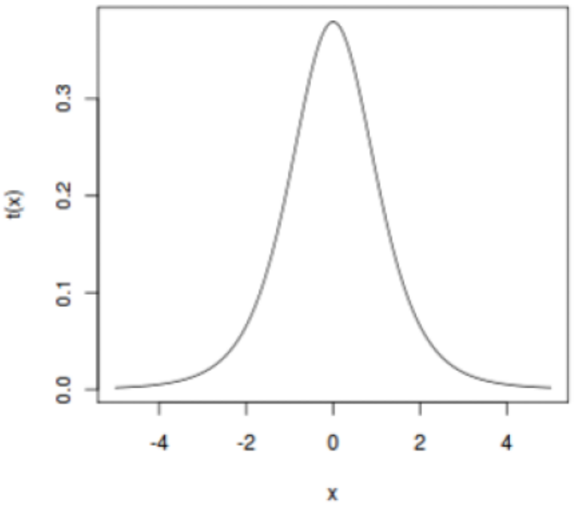
3.5.1.a *Problem:* Assume that the test scores of a college entrance exam fit a normal distribution. Furthermore, the mean test score is 72, and the standard deviation is 15.2. What is the percentage of students scoring 84 or more in the exam?

3.5.1.b *Answer:* The percentage of students scoring 84 or more in the college entrance exam is 21.5%.

3.6. *Student t Distribution:* Assume that a random variable Z has the standard normal distribution, and another random variable V has the Chi-Squared distribution with m degrees of freedom. Assume further that Z and V are independent, then the following quantity follows a Student t distribution with m degrees of freedom.

$$t = \frac{Z}{\sqrt{V/m}} \sim t(m)$$

Here is a graph of the Student t distribution with 5 degrees of freedom.



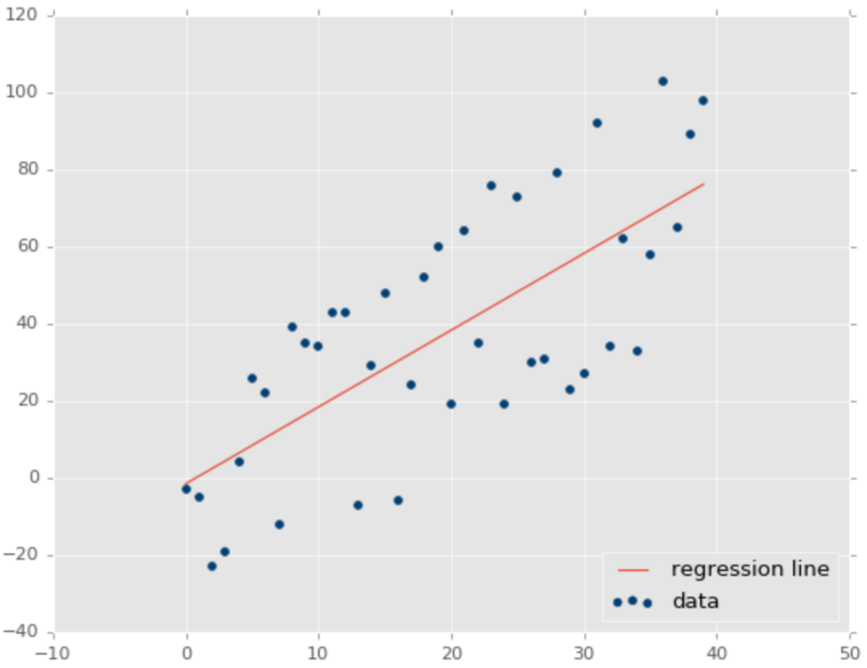
3.7. Task: Please complete the following tasks (30 points).

3.7.1 Solve two of the textbook examples for the above distribution using Python.
3.7.2 Plot these distributions using Python and Matplotlib.

Part IV – Linear Regression Using Python:

This portion of the project you need to work on basics one variable linear regression. There are two types of supervised machine learning algorithms: Regression and classification. The former predicts continuous value outputs while the latter predicts discrete outputs. For instance, predicting the price of a house in dollars is a regression problem, whereas predicting whether a tumor is malignant or benign is a classification problem.

The term "linearity" in algebra refers to a linear relationship between two or more variables. If we draw this relationship in a two-dimensional space (between two variables), we get a straight line. Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y(output). Hence, the name is Linear Regression. If we plot the independent variable (x) on the x-axis and dependent variable (y) on the y-axis, linear regression gives us a straight line that best fits the data points, as shown in the figure below.



https://towardsdatascience.com/a-beginners-guide-to-linear-regression-in-python-with-scikit-learn-83a8f7ae2b4f

Task: Using the website above or your own website gives a simple example (30 points).

Deadlines:

Please complete all the tasks above before **Friday 11 PM, December 8, 2023**. Please include all your codes and MS Word file. This will be graded based on completion of all tasks. If there are extra efforts, you may get extra credit.