CS 3753 & 5163 Data Science Summer 2020

Homework 3 (110 points)

Submission:

1. submit a single python script (abc123_hw3.ipynb or abc123_hw3.py) through blackboard. All the results are outputted from your Python code.
2. You should submit a readMe.txt file. Give the instruction to the TA. So, he knows how to run your code. It should run successfully either in the basic Python3 environment or in Jupyter Notebook.
3. A figure should be generated and displayed after running your code.
4. Do not compress your files
5. The late submission will lose 15% points. Your code should run successfully. There is a limit of half points max if the code cannot run.
6. You can submit your homework 3 times before the deadline.

Questions

1. Matrix multiplication (10 pts)

   A= $\begin{pmatrix} 2 & 8 & 4 \\ 5 & 4 & 2 \end{pmatrix}$; B= $\begin{pmatrix} 4 & 1 \\ 6 & 4 \\ 5 & 3 \end{pmatrix}$; C= $\begin{pmatrix} 4 & 1 & 2 \\ 6 & 4 & 3 \\ 5 & 3 & 4 \end{pmatrix}$; D= $\begin{pmatrix} 4 & 1 & 2 \\ 6 & 4 & 3 \end{pmatrix}$

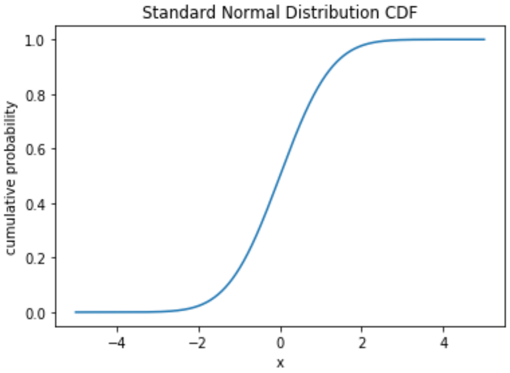
   Write a python code to calculate A·B, A·C, and A·D. If any of them cannot be calculated, please explain why it cannot be calculated.

2. x = np.array([50, 68, 74, 70, 65, 61, 63, 74, 62])
   y = np.array([170, 173, 209, 130, 215, 127, 108, 152, 183]) (10 pts)
   a. define the function of zscore() (3 pts)
   b. scatter plot of the zscore of x and y. (3 pts)
   c. Calculate the Pearson correlation coefficient using your own methods and compare your result with the result from the corrcoef function in numpy. Are they the same (use the print function to display your answer)? (4 pts)

3. x = np.array([50, 68, 74, 70, 65, 61, 63, 74, 62, 20])
   y = np.array([170, 173, 209, 130, 215, 127, 108, 152, 183, 800]) (12 pts)
   a. Calculate the Pearson correlation coefficient using the corrcoef function. (2 pts)
   b. Calculate the Spearman rank correlation coefficient using the corrcoef function. (4 pts)
   c. Are the two coefficients the same? Which one is better? Why? (3 pts)
   d. What do the values (-1, 0, and 1) of the coefficient indicate? (3 pts)

4. Complete the Python code to calculate the following probability in the standard normal distribution. (8 pts)

$$P(-0.5 \leq X < 0.5) = \mathrm{CDF}_X(0.5) - \mathrm{CDF}_X(-0.5)$$



   a. norm.cdf(0.5) - ______________________

   b. 2*(______________________________)

5. Properties of normal distribution (12 pts)
   If the distribution of a random variable X is a normal distribution, $X \sim N(\mu, \sigma^2)$.
   a. If we have X' = $aX + b$, then what is the distribution if X'?
   b. If the distribution of a random variable X is a normal distribution, $X \sim N(\mu, \sigma^2)$. If we convert it into a standard normal distribution, $Z \sim N(0, 1)$. What is the relationship between X and Z? In other words, how do you represent Z using X?
   c. Write a Python code to calculate the probability P(2 <= X <= 7) where $X \sim N(5, 9)$ is in the normal distribution.
   d. Calculate the probability P(-1.5 <= X <= 1.5) where $X \sim N(0, 1)$ is in the standard normal distribution.

6. Complete the following Probabilistic Calculus. (8 pts)
   a. $P(A \cup B) = P(A) +$______________

   b. $P(A|B) =$ _________________$\mathrm{P(B)}$

   c. $P(A \cap B) = \mathrm{P(B)}$________________

   d. If A and B are independent, $P(A \cap B) =$______________________

7. Assume d is the tossed number of an 8-face die. Will the probabilities P(d = even) and P(d < 5) be independent? Write down your derivation and intermediate steps. (8 pts)

8. Use the theorem of total probability and Bayes theorem to solve the following problem. (8 pts)

   A box of dices: 95% fair, 5% loaded (50% at six). If we get 4 six in a row, what's the chance that the die is loaded?

9. Suppose that one person in 10,000 people has a rare genetic disease. There is an excellent test for the disease; 99.9% of people with the disease test positive and only 0.02% who do not have the disease test positive. (10 points)

   a. What is the probability that someone who tests positive has the genetic disease? (5 pts)

   b. What is the probability that someone who tests negative does not have the disease? (5 pts)

10. To complete this question, you need to download the three csv files, which contain the monthly totals of the number of new cases of measles, mumps, and chicken pox, respectively, for New York City during the years 1931-1971. Each

data file contains 41 rows and 13 columns: the first column represents the year (1931 to 1971), and the remaining 12 columns are the number of new cases for each month from January to December. Note that in the chicken pox file, the year is not ordered (i.e., the rows are not chronically ordered), unlike in the other two files. (24 pts)

a. Plot the total number of measles cases in each year. (5 pts)

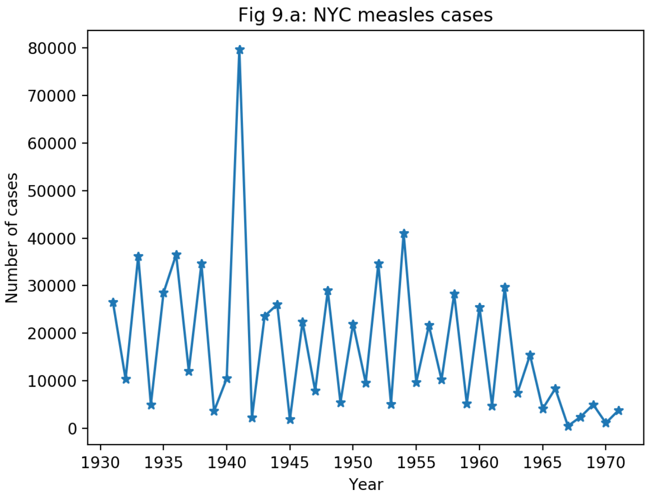


Fig 9.a: NYC measles cases

b. Bar plot the average number of mumps cases for each month of the year. (5 pts)

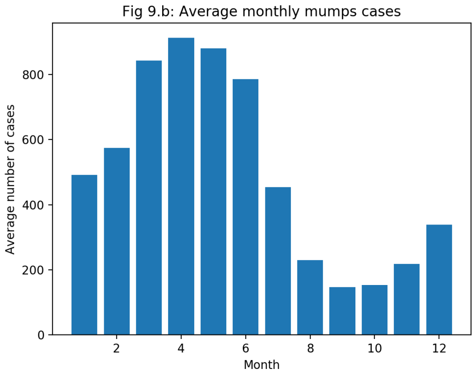
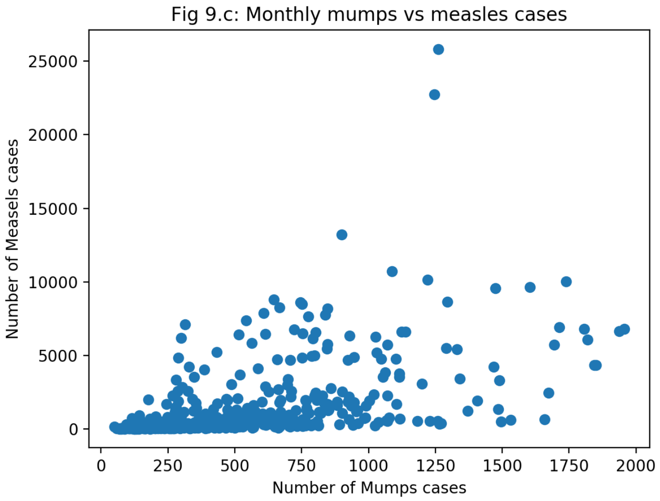
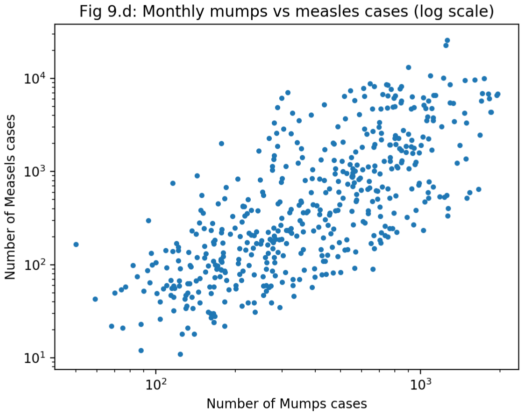


Fig 9.b: Average monthly mumps cases

c. Scatter plot the monthly mumps cases against the measles cases. Each dot in the plot represents one month and there is a total of 41 x 12 months. (5pts)



d. Similar to the previous question, but plot both x and y axis in logarithm scale (using loglog function) (5 pts)



e. In some cases, why would we want to plot the y axis in logarithm scale instead of linear scale? (4 pts)