HW3 - Errorbars and correlation

Electronic submission due 11:59pm, Saturday 3/12

To complete this homework, you need to download one csv file, 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 (for a total of 41 years). The data file contains 123 rows and 12 columns. Each column represents a month from Jan to Dec. The first 41 rows are the number of new measles cases in each year during that period, the next 41 rows are for mumps, and the remaining 41 rows are chicken pox. The rows are ordered by the years in chronical order.

Complete the python script skeleton to analyze the data for the following tasks. For your information, data has been loaded with the Pandas package and reorganized into a Numpy 3-d array of shape (3, 41, 12), where the first dimension represents the three diseases in the order mentioned above. We will refer to this 3-d array as D for convenience, and D_{ijk} = D[I, j, k] be an element of D with indices i, j, and k. Several other variables are also defined for your convenience.

Q1 (10 pts). Calculate the mean of the number of cases per year for each disease, and estimate the 95% confidence interval of the mean (Lec4-1.pptx slide #4). Plot as an errorbar. (Use marker='d', linestyle='', capsize=5 to show a figure similar to **example Figure 1** on the next page.)

Q2 (10 pts). For each disease, calculate the fraction of cases occurred in each month of the year. Specifically, you will end up with a matrix ***C*** of size 3 x 12, where each row is for a disease, and the value in the i-th row and k-th column, ***Cik,*** is the total number of cases of disease *i* occurred in month *k* divided by the total number of cases of disease *i*. Mathematically, Cik = sum_j(Dijk) / sum_j sum_k(Dijk). (Note: use matrix multiplication instead of for loops for this if you can.) Plot the vectors as three lines in one graph. (See **example figure 2**.)

Q3.1 (8 pts) Scatter plot the mean monthly **measles** cases occurred in each month of the year against that of the **mumps** cases. In other words, you are scatter plotting two vectors, x and y, each of which has 12 values. The i-th value of x represents the average number of measles cases per year in month i. Similarly, the i-th value of y is the average number of mumps cases pear year in month i. (See **example figure 3**.) Annotating the figure with months is *optional* (lecture2 slides #27).

Q3.2 (7 pts) Calculate the Pearson correlation coefficient as well as the spearman correlation coefficient between the mean monthly **measles** cases and mean monthly **mumps** cases (the two vectors x and y you calculated in Q3.1), Display the values (with a precision 0.0001) in the figure (decide the x and y positions of display ad hoc from your figure).

Q4.1 (8 pts) Scatter plot the total number of **measles** cases in each year against that of **mumps** cases. That is, you are scatter plotting two vectors, x, and y, each of which has 41 values, representing the number of measles or mumps cases in each year (1931, 1932, etc.) (See **example figure 4**.)

Q4.2 (7 pts) Calculate the Pearson correlation coefficient as well as the spearman correlation coefficient between the annual measles cases and mumps cases (the two vectors x and y you calculated in Q4.1), Display the values (with a precision 0.0001) in the figure.

Q5.1 (5 points) Scatter plot the monthly **measles** cases against the monthly **mumps** cases. Each dot in the scatter plot represents one month and there is a total of 41 x 12 = 492 months.

So you x and y should each has 492 values. **See Fig 5.1.** Calculate the Pearson correlation coefficient as well as the spearman correlation coefficient between them, and display the values (with a precision 0.0001) in the figure.

Q5.2 (5 points) Repeat Q5.1, but plot both x and y axis in **logarithmic** scale, and calculate the Pearson as well as Spearman correlation coefficients using the log values). **See Fig 5.2.**

Challenge question 1 (0 point – no submission needed): compare figure 5.1 with figure 3 and figure 4, what can you say about the correlation between the occurrences of the two diseases?

Challenge question 2 (0 points – no submission needed): compare the correlation values in Fig 5.1 and 5.2, what is your observation and what have you learned (about the data, and about the difference between Pearson and Spearman correlation)?

Q6 (Bonus - 10 pts) Calculate and show the correlation matrix between each of the 12 months for the number of **mumps** cases. Formally, you have a matrix ***M*** of size 41 x 12, where ***Mij*** is the number of mumps cases in year ***i*** and month ***j***. You need to calculate a matrix ***C*** of size 12 x 12 (using np.corrcoef), where ***Cij*** is the correlation between the i-th column of ***M*** and the j-th column of ***M***. Use plt.imshow(C) to display the matrix, and plt.colorbar() to show the color map. Changing the months from 0-11 to 1-12 is optional but can be done with xticks and yticks as usual: xticks(range(12), range(1,13)). (See **example Fig 6**.)

Challenge question 3 (0 point): Why Jan data and Dec data are not correlated (while other neighboring months are highly correlated)?

Q7 (Bonus: 10 pts). Calculate and plot the average probability of each disease occurring in each month. Take mumps cases as an example: you start with calculating a matrix **F** of size 41 x 12, where **Fij** is the probability of mumps cases in year ***i*** occurring in month ***j***, i.e, **Fij = Mij / Mi•,** where **Mi•** is the sum of the i-th row of the matrix M defined in Q5. (Double check that the sum of each row in **F** should be equal to 1.0). Then you would calculate the mean for each column of **F** and obtain a vector of size 12. Repeat this for the other two diseases and plot the three vectors in the same figure. (Alternatively: work with the three-dimensional data array to get a 3x12 matrix instead of three separate vectors.) (See **example Fig 7**.)

Challenge question 4 (0 point): What is the difference between Fig 2 and Fig 6 (what are they plotting)? Under what conditions do you expect to see bigger differences between them?

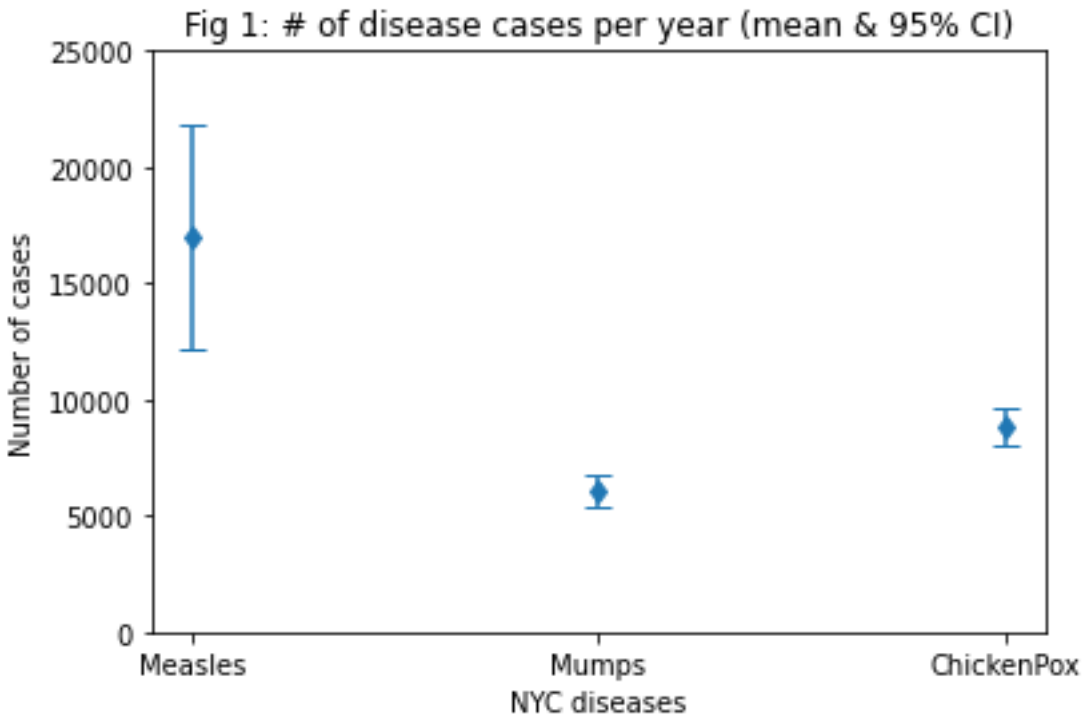
Fig 1: # of disease cases per year (mean & 95% CI)
25000
20000
15000
10000
5000
0
Number of cases
Measles
Mumps
ChickenPox
NYC diseases

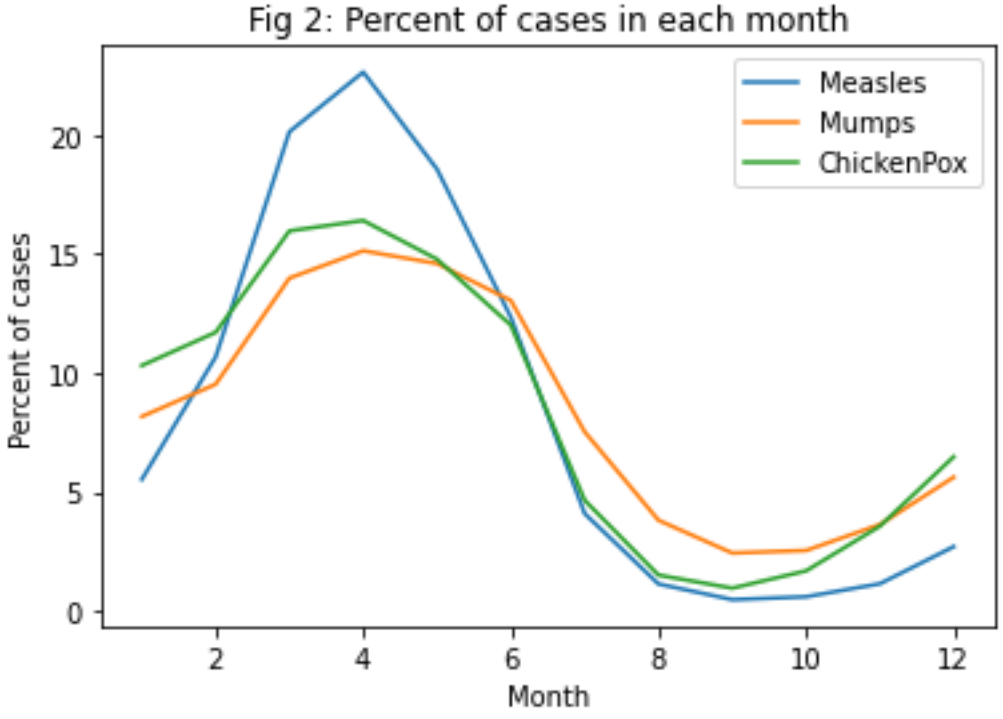
Fig 2: Percent of cases in each month
Measles
Mumps
ChickenPox
20
15
10
5
0
Percent of cases
2
4
6
8
10
12
Month

Fig 3: Mean monthly cases of Measles vs Mumps

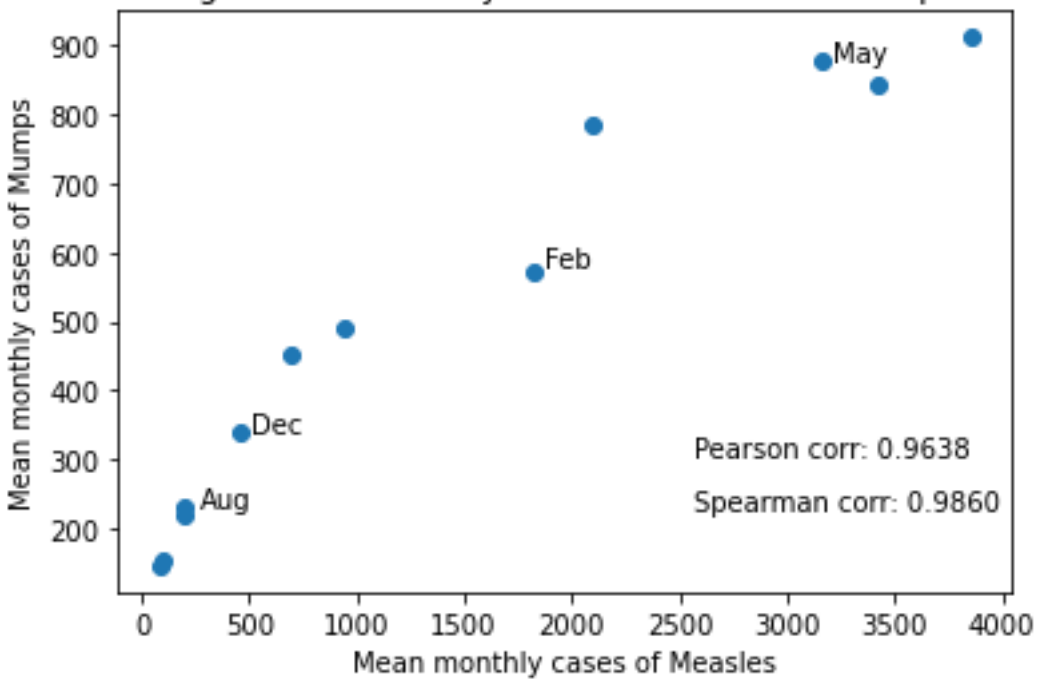

Fig 4: Annual cases of Measles vs Mumps

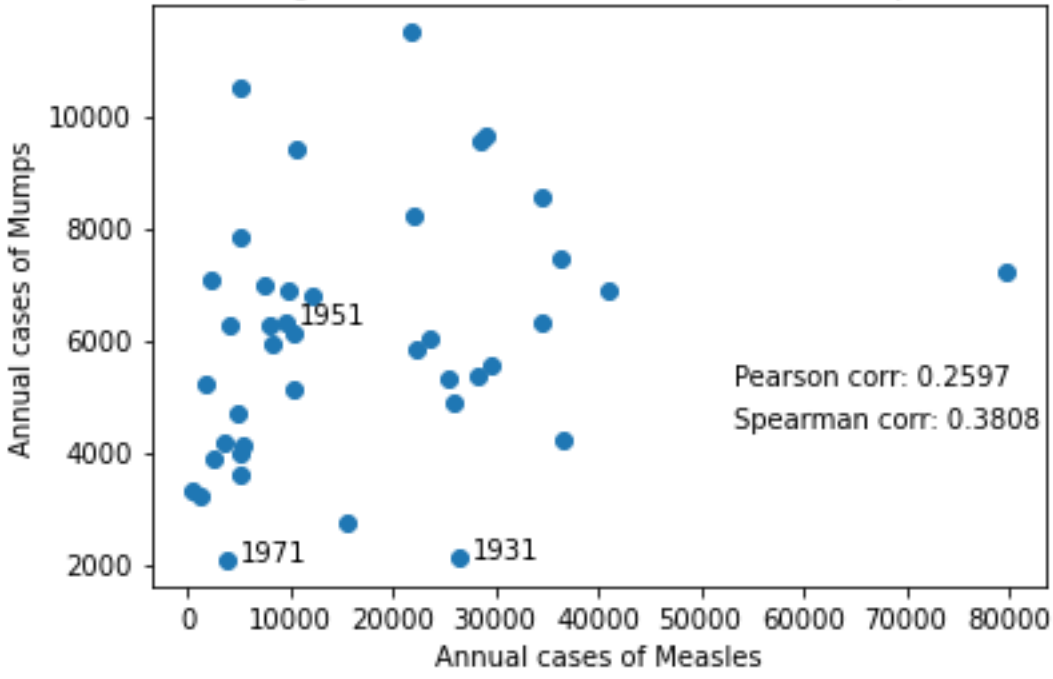

Fig 5.1: Monthly cases of Measles vs Mumps

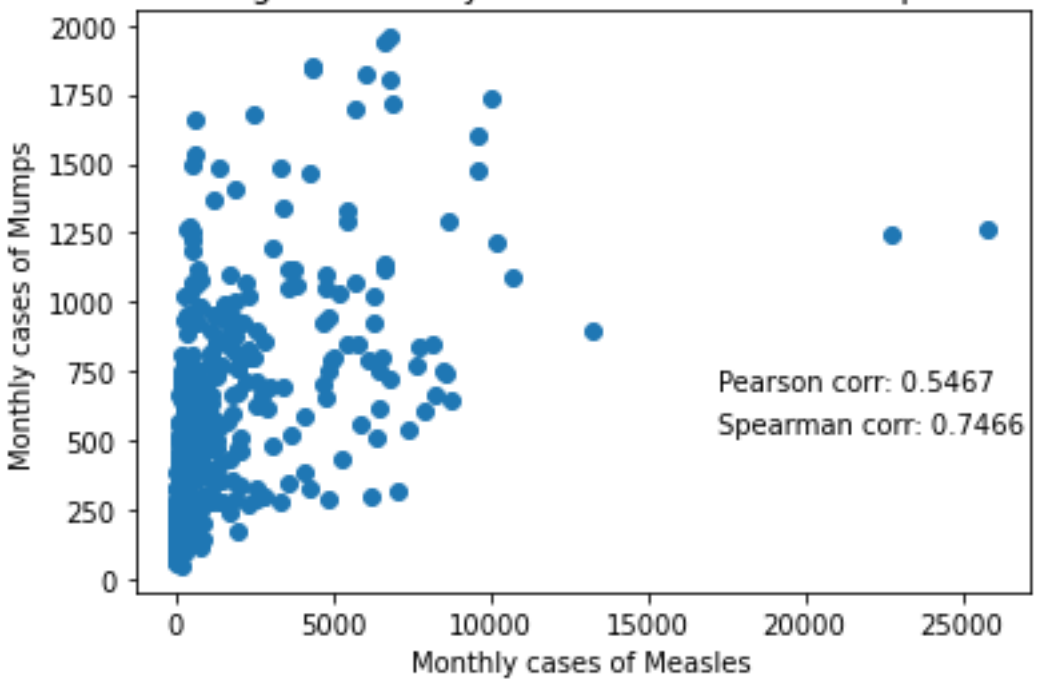

Fig 5.2: Monthly cases of Measles vs Mumps

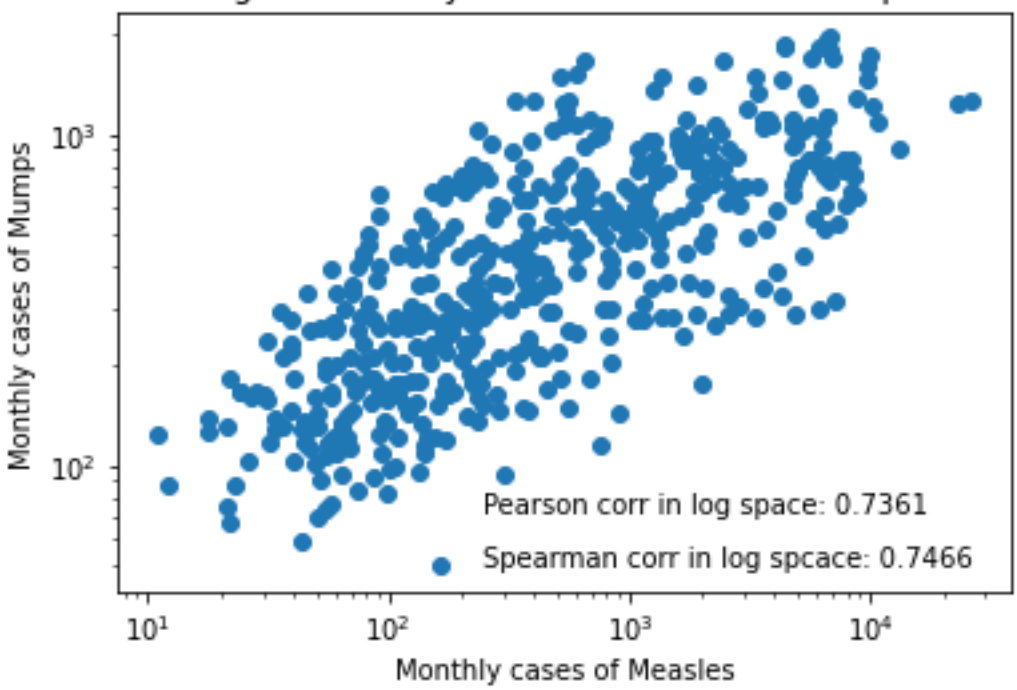

Fig 6: Correlation between monthly Mumps cases

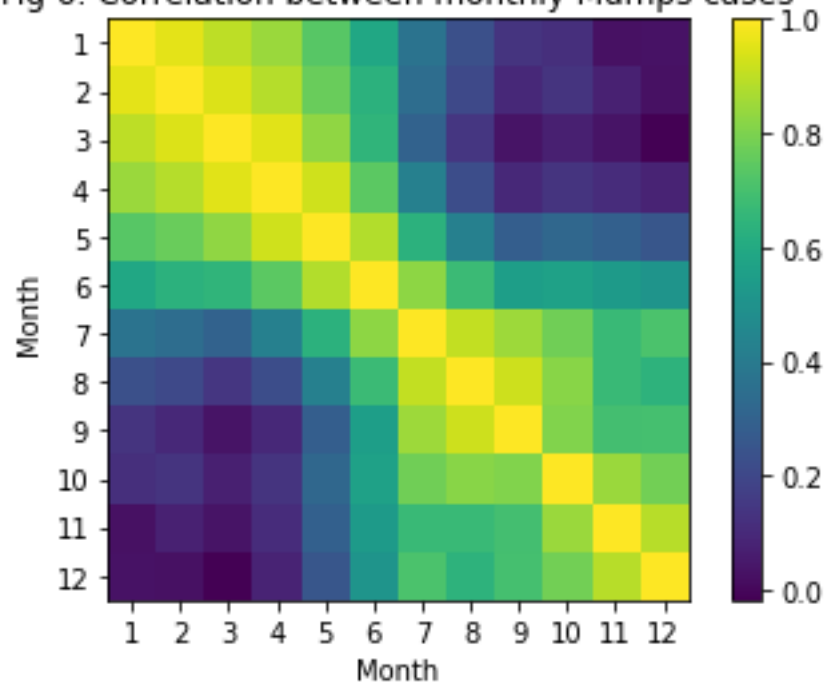

Fig 7: Mean % of disease cases in each month