Indian Institute of Technology Bombay

EP219 Data analysis and interpretation
Assignment 3
Dated : 30 - 9 - 2018

— — — — — — — — — — — — — — — —

Take a look at the two data files unemploymentrate.csv and crimerate.csv (from data.gov.in). These files show the state-wise (and union territory) unemployment rate (in percentage) (we will denote this as U) and crimerate (crimes per 100,000 in column J) (we will denote this as C). For this assignment we will study the correlation between unemployment and crime. We will only be interested in the year 2016.

1. Find the mean unemployment rate and crime-rate by averaging each rate separately across states/UTs.

2. Find the standard deviations of the unemployment rate (σ_U) and crime-rate (σ_C). Explain clearly the meaning of these standard deviations.

3. Make a 1-D histogram of the unemployment rate and a 1-D histogram of the crime rate. Clearly mark the mean and the standard deviation for each histogram.

4. Now for each state we can consider the pair of observations of (U_i, C_i), where i denotes an index for each state/union territory. Make a scatter plot of these pairs of variables. Add the plot to your report.

5. Also make a 2-D histogram of the pairs (U_i, C_i). Use a colormap for the heights of the histogram.

6. Show that the correlation between two data samples $\{X_i,\ Y_i\}$ $i = 1, ..N$ can be estimated as:

$$C_{XY}^{\text{est}} = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \overline{X})(Y_i - \overline{Y})$$

   Here $\overline{X}$ is the sample average of X_i (similarly $\overline{Y}$ is the sample average of Y_i). You may assume that each pair is independent of the others. Write up the proof in your report.

7. Find the estimated correlation coefficient

$$\rho \equiv \sqrt{C_{XY}^{\text{est}}}/\sqrt{\sigma_X \sigma_Y}.$$

8. Find the correlation coefficient between unemployment rate and crime-rate. What do to the sign and magnitude of the correlation coefficient tell you? What can you say about the relationship between crime and unemployment?

Deadline

1. Upload your code and report to your website by Monday, October 8th at 10 am.

Notes:

- Make sure to label all your plots, axes, title etc. Install latex so that you can use latex symbols in the plot legends.
- Try to experiment with histogram bins, axes range, colors, linestyles, plot markers, displaying multiple plots on the same image, saving plots to pdfs etc.
- Comment your code with detailed comments! Uncommented code will receive no credit.
- Try to follow best programming practices in python. https://gist.github.com/sloria/7001839