INDIAN INSTITUTE OF TECHNOLOGY BOMBAY

EP219 Data analysis and interpretation
Assignment 2
Dated : 19 - 8 - 2018

— — — — — — — — — — — — — — — — — —

Initial steps

1. Take a look at the data swachhbharat.xlsx (from data.gov.in) that shows the state and district wise data of the number of villages in a district and number of villages that have been declared as open defecation free (ODF). This data shows the progress towards creating accessible toilets everywhere in India.

2. For a sample of data $(x_1, x_2, ..., x_N)$, we can define the sample mean

$$\overline{x} = \frac{1}{N} \sum_{i=1}^{N} x_i.$$

We can also define the sample variance as,

$$\overline{\sigma}^2 = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \overline{x})^2$$

where $\overline{\sigma}$ is the sample standard deviation.

You assignment

1. Extract the full data set to a numpy or pandas array.

2. We will focus on Uttar Pradesh (UP)for this assignment. For each district in UP we can find the fraction of villages that are ODF. Make a histogram of the fraction of villages that are declared ODF in each district in UP (x-axis is fraction of ODF villages, y-axis is number of districts).

3. Find the mean, variance and standard deviation of the district-wise ODF fraction. Write your own functions to calculate these quantities (do not use in-built python library functions, you can however check your answer against the in-built functions).

4. Clearly mark the sample mean (using a vertical line) and the sample standard deviation (using a horizontal line) on your histogram.

Deadline

1. Add all plots to your report along with a discussion. Upload your code and report to your website by Monday, August 27th 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