# ERV-5086

# Assignment 4

Use the rough starting Python notebook *Miami\_Rain\_Data.ipynb*, the precipitation data file *MIA\_J-D\_T\_precip\_inches.csv*, and the Old Faithful Geyser waiting time data (faithful.csv) which are in a module on your Canvas page, for the following exercises.

The overall idea here is to a) compute descriptive statistics and observe histograms and q-q plots for the rainfall data and its log transformation, b) generate realizations of samples from known normal distributions with the mean and variance of the sample data and its log transformation but many more samples than the original data and plot histograms and q-q plots, c) plot histograms and q-q plots for the original and log-transformed Old Faithful data, and d) test the hypothesis that the samples from the real data, the realizations, and the Old Faithful data, come from a normal distribution. Finally, you’ll compute the rank of the rainfall data and the return period.

One of the principal objectives of this Exercise is that you enhance the skeletal code provided by rearranging it so that you convince yourself of the power of the test for non-normality and provide comments describing what the code is doing. To that end, augment the code with comments (in the form *#Comment*… to maintain code operability) that explain what each line of code does and what the output means. I consider “Import” self-explanatory. Enhance all plots with axis labels.

## Exercise 1

Select a month from the data file instead of the Annual Total that I analyzed. I used a[: , -1] (: = all rows, -1 = column “before” the first column) to select the last data column. There should be 14 columns: (Year, J, F, M, A, M, J, J, A, S, O, N, D, Total) numbered 0—13 in Python. Please compute and plot the time series of your month’s fraction of the annual total for each year.

## Exercise 2

Compute descriptive statistics and observe histograms and q-q plots for your rainfall data and its log transformation. You will recall from class that we considered the statistics of both the original data and the log transformed data (simply by applying *np.log10(…)* or *np.log(…)* to the data). As provided, the code still mixes these. For example, the ….*describe* statement for the descriptive statistics uses *np.log10(…)*, while the subsequent line that computes the standard deviation from the variance is still hard-coded for the untransformed data. Clean this up throughout the code and provide well-labeled results for both the untransformed and log10 transformed data.

## Exercise 3

Generate realizations of samples from known normal distributions with the mean and variance of your selected rainfall data and its log transformation but many more samples than the original data and plot histograms and q-q plots

Where it says

#Here we plot sample from the known normal distribution. Note how much better the q-q plot gets with more samples!

Change it to “…how much worse the q-q plot gets with fewer sample!” and provide an example with fewer samples for visual comparison of the histograms and q-q plots.

## Exercise 4

Read and run the notebook available for download at <https://docs.scipy.org/doc/scipy/lite/notebooks/index.html?path=hypothesis_normaltest.ipynb> to understand this statistical test for normality of a distribution and the closely related Monte-Carlo variant for small samples. The *null* hypothesis (the hypothesis you can reject if the probability --- the *p* value --- of the computed statistic is small) is that the data were drawn from a normal distribution.

Use the normal test on your rainfall data and on your large-sample realization from the normal distribution with the same mean and variance. Use the same test on the Old Faithful waiting time data. This is the third column in the csv file. It may be simplest to cut the header line from the file before uploading it. Describe your observations and conclusions from the statistical test and consider and describe those results in the context of the histograms and q-q plots.

## Exercise 5

Use the last part of the script to rank your rainfall amounts and compute the return period. You can find a summary of the procedure here: <https://serc.carleton.edu/quantskills/methods/quantlit/RInt.html>. The very last line of the code plots ranked rainfall amounts as a function of the log10 of the return period in years. Thus, 2 is 100 years on the x axis. It would be easier to read and better to plot this as a semilog plot where the actual values (at least 1, 10, and 100 years are shown on the axis) but the spacing of the x-axis ticks is logarithmic. Please do that.