

**Lab 3: Examine the New York City area precipitation characteristics**  
**Due Wednesday, September 29<sup>th</sup>**

**Project 1. Daily total Rainfall**

New York daily total rainfall amount for each month is in a worksheet called NYPrecip. Notice the total number of days for each month is different, 1736 days for January, March, May, July, August, October, and December, 1680 days for April, June, September, November, and 1582 days for February. Don't forget to account for this when you are averaging over the month. If you get sick of scrolling down to highlight to the bottom of the data set you can skip ahead to the fifth paragraph of Project 2 for a description on how to select a range without the mouse. Do whatever analysis you feel necessary to answer the following questions, but make sure to support your answers with evidence from the data.

Questions:

- On average, which month has the most daily rainfall amount? The least?
- On average, which month has the most rainy days? The least?
- Which month has the most daily rainfall amount fluctuations? The least?
- Given a 20 day period, what is the probability of having exactly 10 days raining for each month? (Remember the binomial distribution from last week, you can also use Excel's built in function BINOMDIST with "False" for the "Cumulative" entry)
- If we define rainfall amount above 20 mm/day as the heavy rainfall days, which month is most likely to have heavy rainfall? The least?

**Project 2. Use anomalous daily precipitation to examine the central limit theorem**

The data given are the anomalous daily precipitation for New York area (meaning the long term monthly average precipitation has been subtracted from each day's rainfall amount). This way, we take out the possible seasonality effect as examined in the first part. You have a total of 20454 days. A special note for large data sets: when typing a range into a formula if there is only one type of data in the column (i.e. you haven't added a row of means or anything) you can reference the entire column as A:A.

First compute the relative frequency distribution of the rainfall anomaly and plot the probability density distribution. Recall last week when we found the maximum and minimum to write a column of upper bin limits and then used the frequency function (with it's special execute command) to calculate the frequency distribution. This time we should use more than 8 bins, something like 15-20 bins is probably better because there is a large range in the data but as you will see the variance is still relatively small. Another change from last week is that we will make an additional column with the center value from each bin (as opposed to the upper limit) to use for X values in our plots and for calculating the normal distribution. We want density functions so don't forget to divide by the number of samples to bring the probability values between zero and one.

Next we want to compute the normal distribution corresponding to the mean and standard deviation of the daily anomaly precipitation. (That means you should calculate those parameters first!) The normal distribution is given by the following equation shown in lecture:

$$f(y) = k \cdot e^{-c(y-m)^2} \text{ where } k = \frac{1}{\sigma\sqrt{2\pi}}, c = \frac{1}{2\sigma^2}$$

$\mu$  is the mean and  $\sigma$  is standard deviation. Use this equation to calculate a normal distribution using the center bin values (not the upper bin limits that we used for the frequency function). Plot this on the same plot with the PDF we just created.

- a. Can you approximate daily rainfall distribution using the normal distribution, why or why not?

Next we will compute the 7 day running averages of the daily anomaly precipitation and form a weekly precipitation series. A running mean uses a window of constant size centered around a point to calculate a value for each point that is a mean of all values within that window. In other words you will create a time series that is the same size as the original (except missing a few days at the beginning and the end) where each value is replaced by an average of the seven days around it. It has the effect of smoothing the data and will remove fluctuations that have a variation shorter than the window over which we are averaging. In any case I know that sounds confusing, but hopefully after we calculate a running mean you will understand how it works. A running mean of seven days has to have, at each point, three days on either side. This means we can't start our running mean time series until there are three days before the day we are calculating an average for, i.e. we start the series on day 4 (or window size/2 rounded down to the nearest integer).

In a blank column choose a cell for day four and enter a formula that calculates the average of the original series for days 1-7. Now we need to highlight all cells until three before the end which we can do by typing in the range of cells (For example if you have day one in row 1 type "J4:J20450") in the "Name Box" which should be located to the left of the box where we type in equations. Believe me you don't want to scroll down to the bottom of the data set! Now go to the "Edit" menu, choose "Fill-Down" and Excel will paste down your function and create the weekly time series. Since our weekly time series is actually the mean of our original variable, we should calculate the standard error which is given by the standard deviation divided by the square root of our averaging period. Calculate the PDF and Normal Distribution (we still use standard deviation, not standard error) for the weekly precip series and plot the two series like the daily values.

- b. Can we use the same bin limits and or Probability Distribution that we calculated earlier? Why or Why not?

Compute the 15 day, 30 day, and 60 day running averages of the daily anomaly precipitation and for each calculate the PDF and Normal Distribution and plot the two series.

- c. How does the probability density function change as the averaging period increases? Why?

Calculate the standard error.

- d. How does the standard error change? How is the standard error different from the standard deviation and why? Do you think that the standard error is realistic in this example?