Long Problem Set 2

For each problem below, complete any requested calculations and answer any accompanying questions. Your responses are evaluated on the appropriateness of your approach and the insightfulness of your analysis. Be sure to consider significant figures when interpreting the output from R (which, as with a calculator, often ignores such niceties).

You will share your answers to these problems in two ways: a text document that contains your written responses to the questions, and a `.R` script file that contains your code (use comments to separate your code by problem). Save your script file using `yourlastname_LPS02.R` as a file name and share it with by email. You may use any program you wish for your text document.

Your answers to the following questions are due by 4:00 pm on Tuesday, September 11th.

1. In LPS01 you analyzed data on the concentration of NOX collected at a station along London's Marlybone Road. Reload the data and create four vectors: one for the spring months (March, April, and May), one for the summer months (June, July, and August), one for the fall months (September, October, and November), and one for the winter months (December, January, and February). For each discuss evidence for or against the claim that the data are normally distributed. See the schedule for August 29th if you need to download the data, which is in `LPS01.RData`.

2. Using the data in the file `MM.RData` giving the colors of the candies in 30 1.69-oz bags of M&Ms, report the probability that a randomly selected bag will have (a) more than 18 brown M&Ms, (b) fewer than 12 red M&Ms, (c) between 10 and 15 blue M&Ms, and (d) either fewer than 5 or more than 10 orange M&Ms. You may assume that the experimental mean, $\bar{X}$, and experimental standard deviation, s, are appropriate estimates for the population's mean, μ, and standard deviation, σ. See the schedule for August 29th if you need to download the data, which is in `MM.RData`.

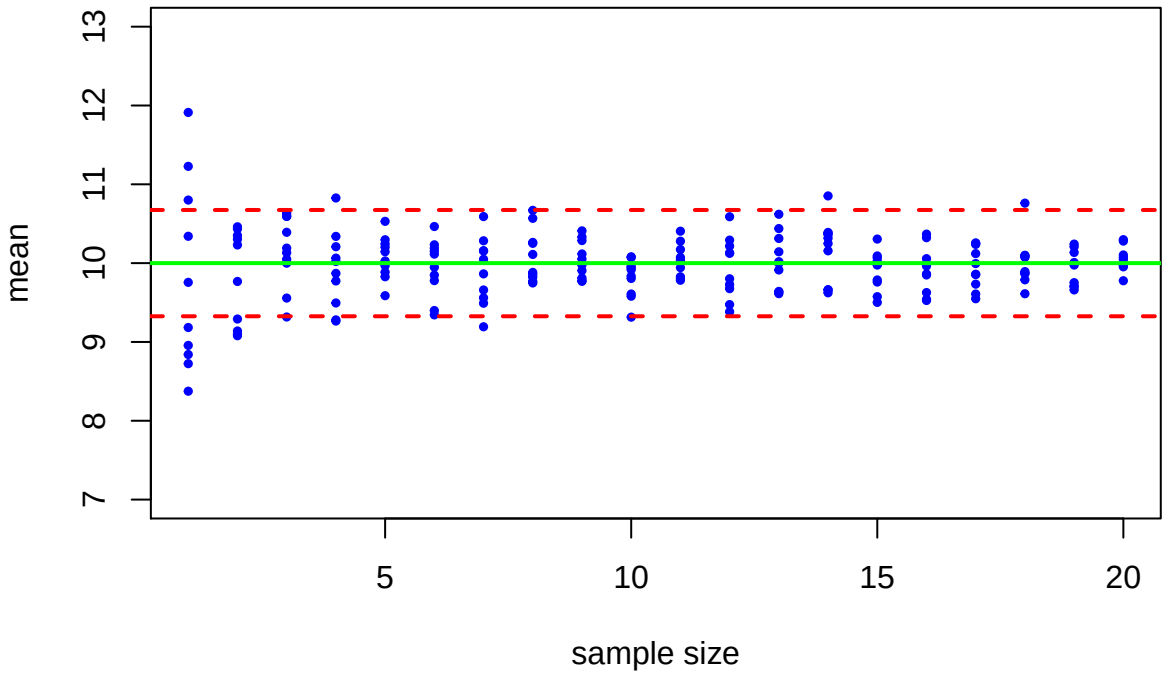
3. The file `DistoNorm.RData` contains 26 vectors, each of which is a random sample of 100 values drawn from a uniform distribution with a minimum of 0 and a maximum of 1; these vectors are identified using the letters "a,", "b," ... "z". The file also contains a vector with the averages for each of the other 26 vectors. Partition the plot window into two rows of three columns each (see the document "Creating Plots Using R's Base Graphics" for details on how to do this). Pick a four letter word that uses four different letters and plot histograms for the data in the vectors for each of your word's letters. Next combine your four letters into a single vector—check your vector to ensure that it has 400 values—and plot its histogram. Finally, plot a histogram for the vector "avg". Discuss how your results provide support for the central limit theorem.

4. The script in the file `SimSample.R` defines the function `simsample`, which simulates the drawing of random samples from a parent population. The function takes four arguments:

   - `mean`: the true mean of the parent population
   - `stdev`: the true standard deviation of the parent population
   - `maxsize`: the largest sample to draw from the parent population; samples are drawn with all sizes from 1 to maxsize
   - `reps`: the number of individual samples drawn for each possible sample size

   The function returns a plot that shows the mean for each sample drawn from the parent population—a total of maxsize × reps samples—a solid green line that marks the parent population's mean, and two dashed red lines that span the middle 50% of the parent population's values. An example of the code and the output is shown here

```
simout = simsample(mean = 10, stdev = 1, maxsize = 20, reps = 10)
```



When assigned to an object—as done above using `simout = simsample()`—the function also returns a list of each sample's size and its average, as shown here

```
str(simout)
```

```
## List of 2
##  $ samplesize: int [1:200] 1 2 3 4 5 6 7 8 9 10 ...
##  $ average   : num [1:200] 11.23 10.44 10.59 9.77 10.3 ...
```

Try experimenting with how the size of a sample drawn at random from a normally distributed parent population affects your confidence in reporting the parent population's true mean. Summarize your findings in one or two **well-written** paragraphs supported, as appropriate, with figures and/or tables.