STAT 330/530
Statistical Computing with SAS
California Polytechnic State University, San Luis Obispo
Lab 2

Some of these problems may be more challenging than others. Please feel free to work with others or speak with me if you need help.

It should all be done in one script, but feel free to use as many DATA and PROC steps as you like along the way.

Fun with Dates and Sampling Distributions!

1. `Project Euler` (https://projecteuler.net/) presents a set of challenging problems that requires programming to solve. This project is gaining in popularity and promise - job candidates are even listing on their resume how many `Project Euler` problems they have solved. For this exercise, we will examine `Problem #19`: how many Sundays fell on the first of the month during the twentieth century (January 1, 1901 to December 31, 2000)?

   (a) In your SAS code, create a new observation for each instance in which a Sunday fell on the first of the month such that you can see the date on which it occurs. Print your results, and include the statement `format date_variable MMDDYY10.;` in your `proc print` so that the dates are readable.

2. In this exercise we are going to generate data from the *uniform* distribution (https://en.wikipedia.org/wiki/Uniform_distribution_(continuous)). The Uniform(a, b) distribution is a distribution in which all values in the interval (a, b) are equally likely. Let's consider a population that is Uniform(0, 10). For this distribution, we know

$$\mu = \frac{1}{2}(a+b) = \frac{1}{2}(0+10) = 5$$

$$\sigma = \sqrt{\frac{1}{12}(b-a)^2} = \sqrt{\frac{1}{12}(10-0)^2} = 2.89$$

   Some additional notes:

   - You can use the SAS function `rand("uniform",0,10)` to generate data from Uniform(0, 10).
   - SAS generates random numbers based on a *seed*. If you do not set the seed, then every time you submit your SAS code you will see different results. If you want to keep your results consistent for each time you submit, then you must set the seed. You can do so at the beginning of the data step with `call streaminit(insert #);`.

   (a) Generate a *single* random sample from the uniform(0, 10) distribution for two different sample sizes: $n = 50$ and $n = 150$.

(b) For each sample size, compute the mean and standard deviation of the sample. Print these results.

(c) In a comment in your SAS code, comment on how your sample means and sample standard deviations compare to the theoretical values above.

3. The Central Limit Theorem describes the behavior of *many sample means*: as the sample size n increases, the distribution of *sample means* becomes approximately normal with a mean equal to μ and and standard deviation equal to $\sigma/\sqrt{n}$ (regardless of the shape of the underlying population). Let's consider the behavior of 500 sample means from samples of size $n = 50$ and $n = 150$.

   (a) Do the math to determine the theoretical mean and standard deviation of *many sample means* for samples of size $n = 50$ and $n = 150$. Describe these results in a comment in your SAS code.

   (b) Write a *new* data step that builds upon your work in question 2 to generate 500 random samples from the uniform(0, 10) distribution for two different sample sizes: $n = 50$ and $n = 150$. For each sample, record the sample mean (*note: you do not need to record the values from the individual samples*). These 500 sample means (from the two sample sizes) represent a simulation of *many* sample means.

   (c) For each of the two sample sizes, calculate the mean and standard deviation of the 500 sample means. Print these results.

   (d) In a comment in your SAS code, comment on how the mean of your sample means and the standard deviation of your sample means compare to the theoretical values you calculated in (a).