EXERCISE: UNBIASEDNESS, CLT

(1) The directory for this exercise includes a copy of the data file newyork.asc, which contains data on the populations of over 800 cities in New York State. It can be read in to R with **scan**(``newyork.asc'', skip=7) command, though that mixes the population numbers with some extraneous numbers in the second column. You'll need to reformat it and delete the second column.

We haven't discussed the central limit theorem yet, but it asserts that the distribution of sample averages of i.i.d. draws from a common distribution starts to be nearly a normal distribution if the sample size is large. In many contexts 20 observations is enough to make this "approximate normality result hold.

You are to generate 30 samples of size 20 as random draws, with replacement, from the full population of 804 cities. In R this can be done with the **sample**() function, with the option **replace**=TRUE. Take the sample means of these 30 samples and display their histogram. Also display a normal qq plot of the data (in R, use **qqnorm**()). The normal qq plot plots the quantiles of the data sample against the quantiles of a normal distribution of the same mean and variance. If the data were normal, it would be approximately a straight line with slope 1. When we discuss the CLT in lecture, we'll discuss why we get these results for this sample.

Date: September 22, 2020.

(2) As promised in the Tuesday lecture, you are to consider the model in which i.i.d. data are drawn from a $N(\mu, \mu^2)$ distribution. Write out the likelihood for this model as a function of μ and the sample mean and sample standard deviation. Again you are to generate 30 random samples of size 20, but this time from a $N(1, 1)$ distribution (i.e., the $\mu = 1$ case). For each sample, calculate the posterior mean of μ. This requires numerical integration (I think), but numerical integration in one dimension is pretty easy. You also have to be careful, as usual, about numerical underflow and overflow when exponentiating log likelihoods.

Code that does the hard parts of this in R is available as `mumuMain.R` in the exercise directory.

(a) Compare the mean, across the thirty samples, of the sample means (`mmean` in the provided R code), to the mean of posterior means (`mhat` in the R code). Probably the posterior means show more bias.

(b) Calculate the RMSE (root mean square error) across samples of the posterior means and the sample means. The RMSE will be the square root of the sum of squares of the differences between the estimators and the true μ, which here is $\mu = 1$.

(c) A better frequentist estimator, though not an unbiased one, uses both the sample mean and the sample standard deviation, since both should be close to μ according to the model. So you might use the average of the sample mean and the sample standard deviation as an estimator of μ. Calculate the bias and RMSE of that estimator. This should be doable using the `mmean` and `msigsq` vectors already computed in the `mumuMain.R` code.