STAT 645: Biostatistics - Assignment 1

Due Friday, ~~September 07~~, 11:55pm CT

Due date: August 28

1. For the income by degree and gender data set, contained in the file "inc_deg_data.csv" (Course Content/Data/incdeg):

   (a) Make side-by-side box plots of income, with separate boxes for each of female arts (gender = 0, degree = 0), female science (gender = 0, degree = 1), male arts (gender = 1, degree = 0), and male science (gender = 1, degree = 1). Include labels on the x-axis to indicate which box goes with which category.

   (b) Report the mean, median, standard deviation, and first and third quartiles of income.

   (c) Report the mean, median, standard deviation, and first and third quartiles of income, now with income expressed in dollars (rather than 1,000's of dollars).

   (d) Report the mean, median, standard deviation, and first and third quartiles of income (in 1,000's of dollars), now excluding the minimum and maximum values.

2. Set your random seed to be 101 (do `set.seed(101)`). Create a 100×5 matrix of random realizations from the standard normal distribution (normal with mean 0 and standard deviation 1).

   (a) Report the column means (a vector of length 5). Demonstrate how you would do this (i) using the `apply` function and (ii) using vector / matrix arithmetic.

   (b) Make a histogram of the row ranges; i.e., compute the range (maximum minus minimum) for each row, and make a histogram of the resulting 100 ranges.

3. Consider the gamma distribution with shape and scale parameters both equal 2; this corresponds to a mean of 4 and a variance of 8. Simulate samples of size $n = 10, 30, 90$ from this distribution, repeating $B = 1000$ times. For each simulated data set, compute the sample mean. Thus, you will have $B = 1000$ sample means for each of the three sample sizes. For each sample size, draw a probability histogram (as opposed to a frequency histogram...you can do this by setting `probability = TRUE` as an option to the `hist` function). Overlay the normal curve that would apply if the central limit theorem could be assumed to hold. Report the resulting three figures as a single three-panel figure.

4. In R create a matrix, named A, with 5 rows and 4 columns, such that the first three rows are random numbers generated from normal$(0, 1)$ distribution while the last two rows contain random numbers generated from Uniform$(-2, 2)$. Create another matrix, named B, with 5 rows and 4 columns, such that the all elements are random draw from the Beta$(2, 1)$ distribution. For creating A and B, use set.seed(101) and set.seed(102), respectively.

   (a) Provide the code to obtain the column sum of A (sum of all entries for each each column).

   (b) Provide the code to obtain $A + B$, then print the (4, 2) and (4, 4)th entries of this sum.

   (c) Provide the code to obtain AB^T, then print the (4, 2) and (4, 4)th entries of this multiplication.

   (d) Obtain the inverse of $B^T A$, and also obtain the determinant of $B^T A$.