Computational Mathematics

Your final is due by the end of the last week of class. You should post your solutions to your GitHub account or RPubs. You are also expected to make a short presentation via YouTube and post that recording to the board. This project will show off your ability to understand the elements of the class.

Problem 1.

Using R, set a random seed equal to 1234 (i.e., set.seed(1234)). Generate a random variable X that has 10,000 continuous random uniform values between 5 and 15.Then generate a random variable Y that has 10,000 random *normal* values with a mean of 10 and a standard deviation of 2.89.

Probability. Calculate as a minimum the below probabilities a through c. Assume the small letter "x" is estimated as the median of the X variable, and the small letter "y" is estimated as the median of the Y variable. Interpret the meaning of all probabilities.

5 points a. P(X>x | X>y) b. P(X>x & Y>y) c. P(X<x | X>y)

5 points. Investigate whether P(X>x & Y>y)=P(X>x)P(Y>y) by building a table and evaluating the marginal and joint probabilities.

5 points. Check to see if independence holds by using Fisher's Exact Test and the Chi Square Test. What is the difference between the two? Which is most appropriate? Are you surprised at the results? Why or why not?

Problem 2

You are to register for Kaggle.com (free) and compete in the Regression with a Crab Age Dataset competition. https://www.kaggle.com/competitions/playground-series-s3e16 I want you to do the following.

5 points. Descriptive and Inferential Statistics. Provide univariate descriptive statistics and appropriate plots for the training data set. Provide a scatterplot matrix for at least two of the independent variables and the dependent variable. Derive a correlation matrix for *any* three quantitative variables in the dataset. Test the hypotheses that the correlations between each pairwise set of variables is 0 and provide an 80% confidence interval. Discuss the meaning of your analysis. Would you be worried about familywise error? Why or why not?

5 points. Linear Algebra and Correlation. Invert your correlation matrix from above. (This is known as the precision matrix and contains variance inflation factors on the diagonal.) Multiply the correlation matrix by the precision matrix, and then multiply the precision matrix by the correlation matrix. Conduct **LDU** decomposition on the matrix.

5 points. Calculus-Based Probability & Statistics. Many times, it makes sense to fit a closed form distribution to data. Select a variable in the Kaggle.com training dataset that is skewed to the right, shift it so that the minimum value is absolutely above zero if necessary. Then load the MASS package and run fitdistr to fit an exponential probability density function. (See https://stat.ethz.ch/R-manual/R-devel/library/MASS/html/fitdistr.html). Find the optimal value of λ for this distribution, and then take 1000 samples from this exponential distribution using this value (e.g., rexp(1000, λ)). Plot a histogram and compare it with a histogram of your original variable.   Using the exponential pdf, find the 5th and 95th percentiles using the cumulative distribution function (CDF).   Also generate a 95% confidence interval from the empirical data, assuming normality. Finally, provide the empirical 5th percentile and 95th percentile of the data. Discuss.

10 points. Modeling. Build some type of *multiple* regression model and **submit your model** to the competition board. Provide your complete model summary and results with analysis. **Report your Kaggle.com user name and score.**