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.

Probability Density 1: X~Gamma. Using R, generate a random variable *X* that has 10,000 random Gamma pdf values. A Gamma pdf is completely describe by *n* (a size parameter) and lambda (λ , a shape parameter). Choose any *n* greater 3 and an expected value (λ) between 2 and 10 (you choose).

Probability Density 2: Y~Sum of Exponentials. Then generate 10,000 observations from the sum of *n* exponential pdfs with rate/shape parameter (λ). The *n* and λ must be the same as in the previous case. (e.g., mysum=rexp($10000,\lambda$)+rexp($10000,\lambda$)+..)

Probability Density 3: Z~ Exponential. Then generate 10,000 observations from a single exponential pdf with rate/shape parameter (λ).

NOTE: The Gamma distribution is quite common in data science. For example, it is used to model failures for multiple processes when each of those processes has the same failure rate. The exponential is used for constant failure rates, service times, etc.

5 points.

1a. Calculate the empirical expected value (means) and variances of all three pdfs.

1b. Using calculus, calculate the expected value and variance of the Gamma pdf (X). Using the moment generating function for exponentials, calculate the expected value of the single exponential (Z) and the sum of exponentials (Y)

1c-e. Probability. For pdf Z (the exponential), calculate empirically probabilities a through c. Then evaluate through calculus whether the memoryless property holds.

a. $P(Z> \lambda | Z> \lambda/2)$ b. $P(Z> 2\lambda | Z> \lambda)$ b. $P(Z>3\lambda | Z> \lambda)$

5 points. Loosely investigate whether P(YZ) = P(Y) P(Z) by building a table with quartiles and evaluating the marginal and joint probabilities.

	1st Quartile Y	2d Quartile Y	3d Quartile Y	4th Quartile Y	Sum
1st Quartile Z					
2d Quartile Z					
3d Quartile Z					
4th Quartile Z					
Sum					

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?

Problem 2

You are to register for Kaggle.com (free) and compete in the House Prices: Advanced Regression Techniques competition. https://www.kaggle.com/c/house-prices-advanced-regression-techniques . 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 LU 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 5^{th} and 95^{th} percentiles using the cumulative distribution function (CDF). Also generate a 95% confidence interval from the empirical data, assuming normality. Finally, provide the empirical 5^{th} percentile and 95^{th} 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. Provide a screen snapshot of your score with your name identifiable.**