STAT 511: Statistical Methods
Summer 2014 - Project 1(60 pts + 5 pts BONUS)
Due Friday, July 11

Note: This project requires the use of a spreadsheet to determine the experimental probabilities in addition to some computer software that will generate QQ plots.

Please do not include your raw data; however it is necessary to list the site that you used to obtain the experimental values and all parameters that were entered. In addition, the analysis procedure for how you obtained your experimental answer as well as the work for the theoretical answers needs to be provided.

(10 pts.) 1. Dice Experiment. Roll two dice 100 times (Chapter 2): You will need to get separate data sets for parts a) and b). From the experimental data obtained from http://www.random.org/dice or http://www.roll-dice-online.com/, any other online simulator of rolling 2 dice or a software package (bonus), determine the probability of each part. This should be done by a direct counting; no theoretical concepts should be used. If you use an online source, state the site that you used including all parameters. If you use statistical software to obtain the information, you need to include the package used and the procedure or code. Be sure to include both the experiential procedure and the theoretical calculations for each part.

a) What is the probability of getting a total of 11 or an absolute difference of 3? Compare the experimental probability with the theoretical value.

b) What is the probability of getting at least one 5 or an absolute difference of 3? Compare the experimental probability with the theoretical value.

(30 pts.) 2. Cards Experiment (Chapter 3). You will need to obtain separate data sets for each part. For the theoretical probability, you need to name the distribution and then calculate the theoretical value using that distribution (for the parameters, use the values for 1 repeat). For the experimental probability, state the site that you used to generate the data (or software package for bonus) and the methodology used to determine the probability (no theoretical concepts should be used). To compare the results, determine how many standard deviations (theoretical) the experimental value is from the theoretical and then state whether they are the same or not.

a) Draw 3 cards without replacement from one shuffled deck containing the 13 cards from one suit, repeat the process 99 times for a total of 100 trials. What is the probability that there is exactly one face card? (face cards are Jack or 11, Queen or 12 and King or 13)? Compare the experimental probability with the theoretical answer. To do this experiment visit http://www.random.org/playing-cards/ or http://www.randomizer.org/form.htm or any online simulator of a deck of cards or use any software package (bonus).

b) Repeat a) 100 times but this time with replacement (13 cards from one suit). Compare the experimental probability with the theoretical answer. To do this experiment, visit http://www.randomizer.org/form.htm or any appropriate online simulator or use any software package (bonus).

c) Perform the following experiment 100 times: Using the same deck as in b) (13 cards from one suit with replacement), we are interested in the number of trials that it will take ~~before~~ to draw the first face card. Calculate the probability that it will take exactly one trial for this to occur. To do this experiment, visit http://www.randomizer.org/form.htm or any appropriate online simulator or use any software package (bonus).

(20 pts.) 3. Assessing Normality (Chapters 4 and 5). Be sure to state what software package and/or web site that you use and any relevant parameters.

a) Normal distribution (note: The normal distribution is often called the Gaussian distribution) Randomly generate normal distributions either using your software package or at http://www.random.org/gaussian-distributions/ (or another online site) whichever is easier. Create and present two different QQ-plots for each of the following sample sizes: ten (10); hundred (100); one thousand (1000). Note: In JMP, it is better to use the QQ plot that is generated when you select continuous fit → normal which can be selected via the Diagnostic Plot. This also will display the axes labels on the plot; the normal probability axis goes from 0.1 to 0.9. The actual data points will be on the other axes and with the range for those data. For JMP, use Help → Books to get the theoretical background for the methodology used. Use the same parameters for all six of the plots and state what values that you are using. Please compare and contrast the plots. In JMP, please ignore the dotted lines in the plots. In your discussion, explain why all of the plots do not look normal when they are all derived from normal distributions. Be sure to state the formula that is being used to calculate the normal percentiles. (This is only necessary for part a) because I am assuming that you are using the same software package for all of the parts. If this assumption is not correct, you will need to state the formula used for each part where you use a software package.)

b) Using the data file "project1.xlsx", use a QQ-plot to determine which of the three samples is derived from a normal distribution. Explain your answer. Note: There may be 0, 1, 2, or 3 normal distributions here.

c) Can we use the QQ-plot to determine the mean and standard deviation of a sample that is normally distributed? Explain your answer. If yes, report the analysis method and determine the mean and standard deviation for each of the **normal** distributions found in b). Note: Most software packages will print out what the mean and standard deviation are; however, you still need to explain how they determine what the values are.

d) Uniform distribution: http://rechneronline.de/random-numbers/ (10 decimal places) or from your software package whichever is easier. Generate 10 uniform distributions on the interval [1,10] with 100 values each. Generate the following QQ plots and histograms when averaging the following number of columns: 1, 2, 6, 10. (Hint: After generating the 10 columns, use either a spread sheet or a software package to average the appropriate number of columns. Then use the new generated column as the data for the plots.) Which of any of these averages are normal distributions? Besides looking at the shape of the plot, compare the mean and standard deviation generated with the theoretical values (see p. 223 in the PROPOSITION box for details on how to generate the theoretical values).

e) BONUS (5 pts): Generate 100 random numbers from an asymmetric distribution. Please be explicit on how you are generating your data sets (program, which distribution, etc.). How many columns need to be averaged for the resulting distribution to be normal? Why is this number larger than the value in part d). Please include the QQ plotsand histogram for the average of 1 (initial distribution) and the one that you are stating is normal.