STA 032 Spring 2016
R Project - Due Friday, June 8^{th} by 9pm on Canvas.

Instructions

* Complete 4 of the problems below in R Markdown. Clearly indicate which problems you choose.

* You may work in pairs (groups of 2) or individually.

* Use complete sentences, proper grammar, check spelling and punctuation.

* All code should be organized and placed in the appendix, unless needed in the problem.

* You will be graded on accuracy, format, and presentation. This is a report.

Problems

I. Create functions which perform the following tasks:

(a) Takes in a vector, and subtracts the mean and divides by the standard deviation (I.e., for every x_i finds $(x_i - \bar{x})/s$). Then returns the standard deviation of the result. Test the function on the following vector: `X = 1:100`.

(b) Takes in a vector and finds the values which are $(\bar{x} - 2s, \bar{x} + 2s)$, where s is the sample standard deviation, and returns both values, with labels of "lower" and "upper" respectively. Test the function on the following vector: `X = 1:100`

(c) Takes in a vector, and calculates the mean after removing any observations that are more than 3 standard deviations from the mean. Test the function on the following vector: Test the function on the following vector: `X = c(1:100,200,300)`

II. The purpose of this problem is to simulate a fair coin flip, and to see how many flips it takes for the probability of a head to be approximately 0.50.

(a) Use the function `sample` to flip a fair coin 20 times, and find the probability that you flipped a "head" based on the 20 flips.

(b) Use an `sapply` to repeat (a) for the following values of n: 10, 100, 1000, 10000, 100000. Show the probabilities for all 5 values of n.

(c) The error of a coin flip is the absolute value of the estimated probability minus the true probability, i.e
error $= |0.50 - \hat{P}(head)|$
Find the error for your simulations from (c).

(d) What happens to the error as n increases, and why? Explain your answer.

III. This problem will use R to find all possible orderings of 7 objects, and probabilities associated with them. Consider your vector of possible values to be:
`values = as.character(1:7)`

(a) Use the function `sample` to draw from `values` 7 times (without replacement), and return this vector. Notice it is a vector, with 7 values. Display your particular draw.

(b) Repeat (a) 100000 times using an `sapply`. Notice the result has 7 rows, and 100000 columns, where each column is a specific random draw. Use this result to find how many of your orderings begin with the character `1`.

(c) Use your samples from (b) to find the probability that a random ordering started with a `3` and ended with a `7`.

(d) The function `paste` can collapse a vector of many characters into a single character with the following command:
`one.order = paste(one.draw,collapse = "")`
Modify the above and use it with an `sapply` to find how many unique orderings of 7 values there are (assuming order matters and no repetitions are allowed).

IV. The goal of this problem is to simulate a binomial random variable. Consider a class with 40 students, and the probability that a student does not turn in a homework is 0.05 (a "success"). Assume all students are independent of all other students, and the probability does not change.

(a) Use `sample` to simulate drawing 40 students who either do, or do not, turn in their homework, and then find the total (out of 40) who did not turn in their homework. You should return one number, X = total # of students out of 40 who did not turn in their homework.

(b) Repeat (a) 1000000 times (you should have 1000000 values for "number of successes",or X), plot a histogram of your result **(do not print out the 1000000 values!!)**. Is the distribution symmetric? Explain.

(c) Find the average of the number of successes in 40 trials **and** the standard deviation based on your simulation from I(b).

(d) Estimate the probability that all students turned in their homework based on your simulation from I(b).

(e) Estimate the probability that at least two students did not turn in their homework based on your simulation from I(b).

(f) What is the median number of students who will forget their homework based on your simulation from I(b)?

V. On Canvas you will find the file `crime.csv`. It has two columns, one of which is the percentage of individuals in the county with at least a high-school diploma (column `dip`), and the other is the crime rate per 100,000 residents for the counties (column `rate`). Consider Y to be crime rate, and X to be percentage with high school diploma.
Use R to complete the following tasks:

(a) Plot a scatter plot of Y and X, being sure to label the axes and give a main title.

(b) Calculate the estimated regression line.

(c) Interpret the slope and intercept (if appropriate) in terms of the problem.

(d) Does there appear to be outliers in the plot from (a)? If so, identify them in R (for example, list the pair (X,Y) that are outliers, or equivalently the row).

(e) Create a QQ plot (normal probability plot) of the residuals. Does it appear that they are normally distributed? Explain.

(f) Create a plot of the errors vs. the fitted values ($\hat{Y}_i$'s). Does it appear the variance of the errors is constant? Explain.

(g) Find the 95% confidence interval for the slope, and interpret it in terms of the problem. Does the interval suggest there is a significant linear relationship? Explain.