Instructions

Submission. All responses to these problems must be received via email in my inbox by Friday, January 28th by 5pm. A response should be in the form of a working, self-sufficient R script. Because this is the only file that needs to be turned in, you should place all information to identify whose submission it is inside comments in the script.

Resources. The only external objects/data/functionality that you are allowed to use are R packages from CRAN. However, you do not need any at all.

Solutions. Textual solutions to the problems should be placed as a comment after the code used to compute the answer. Graphical solutions should be comprised of the code used to generate the plots.

Incentive Structure. The questions will each be worth different point values. Partial work can receive partial credit at my discretion. Two people will receive Starbucks gift cards (5$). The decision rule used to draw the two winners will be as follows:

1. All valid problem sets will be graded and raw points assigned to each person.
2. Each point is worth one lottery ticket with your name on it.
3. The total number of points possible is irrelevant except that it places an upper limit on the number of lottery tickets you can receive.
4. The winner of the first gift card will be drawn from all of the lottery tickets.
5. The remainder of this winner's lottery tickets will be discarded.
6. Another winner will be drawn from the remaining tickets using the same procedure.

Consequently, there is incentive to compete even if you know you won't have time to work each problem out.

Assistance. You **may not** ask any other first-year student any questions regarding this problem set except "are you done yet?" until after the deadline. You **may** ask other students in the program for their help and thoughts, but you must disclose the nature of this assignment! You **should** ask me questions since I have solutions already prepared!

Solutions. Where possible, the provided solutions reflect an approach to solving the problem which is like the material we covered in class. This does not mean there are not more concise or more efficient ways of solving the problem. For example, loops are employed here more than they should be in practice.

Example

The following is an example question and the example code that might correspond to a full credit response.

Example Question

Consider a sequence such that the i^{th} term, a_i, is given by

$$a_i = log\left((i^{i-1})^{1/2}\right).$$

Find the sum of the first 50 terms, that is, find

$$\sum_{i=1}^{50} a_i.$$

Example Response

```
################################################################################
### AUTHOR: J. P. Olmsted --- jpolmsted@NOSPAM.gmail.com
###
### DATE: Sat Jan 22 13:48:16 2011
################################################################################

### Example Problem

vI <- (1:50)
sum(log(sqrt(vI ^ {vI - 1})))

### Answer: 2107.463
```

./hw_example.R

Problems

Problem 1

Consider random samples of size 10 from the F distribution with 3 and 7 degrees of freedom. Use 1,000 such samples to calculate the standard deviation of the sampling distribution of the sample mean. (**10 points**)

Problem 2

Consider the 26 letters of the Modern English alphabet as defined here (`http://en.wikipedia.org/wiki/English_alphabet`). Assume the following definitions:

Letter Value The integer value of the index corresponding to a particular letter given the standard *a, b, c, d, ...* order of the letters.[1]

Text Sum The sum of the sequence of letter values corresponding to a sequence of letters comprising a portion of text where all whitespace and punctuation is ignored.

Find the *Text Sum* corresponding to the main body of text on `http://www.rochester.edu/college/gradstudents/jolmsted/research/` which begins "My primary academic interests ..." and ends "... original spatial voting model." (**20 points**)

Hint. This problem requires "string manipulation".

Problem 3

Construct a function to compute whether a given integer input is "odd" or "even". Use this function to confirm that the sum of the odd numbers between the 1 and 20, inclusive, is 100. (**5 points**)

Problem 4

Consider $x = (x_1, x2, \ldots, x_{1,000,000})$ and $y = (y_1, y_2, \ldots, y_{1,000,000})$ where

$$x_i \sim N(0, \sigma = 1),$$

and

$$y_i \sim N(x_i, \sigma = s).$$

Find some value of s such that $\rho(x, y) \in (0.3, 0.4)$, where ρ is the sample correlation between the vectors. You may consider s a solution if three runs of the data-generating process produce ρ's that satisfy the constraint. (**20 points**)

[1] So, LV(a) = 1 and LV(z) = 26 where LV is the letter value function which maps a letter to its letter value.

Note. Technically, ρ is a random variable because x and y are random variables and it has non-zero density over the unit interval. However, for an appropriate value of s, the use of 1,000,000 draws ensures that we can constrain the sampling distribution of ρ to be sufficiently within the desired interval that we can ignore the chance that we will draw a sample which results in a correlation in the interval sometimes and not in others. That is, if you run the data generating process for a fixed s several times and you get a ρ that satisfies this constraint you can count this as an answer.

Problem 5

Base `R` and `R` packages provide datasets that don't need to be loaded from external files. They can be accessed with the function `data()`. Load the `airmiles` dataset and the `discoveries` dataset. The "data" that these two provide are two appropriately named vectors that are also `timeseries` objects. Use `ts.intersect()` to merge the time series and then coerce this object to a dataframe object.

1. Code a new variable in the dataframe object such that it takes on a value of 1 if both the number of `airmiles` is odd and the number of `discoveries` is odd, the value 2 if both of the other variables of interest are even, and the value 3 if the variables of interest are mixed. Call this variable "yeartype". (**5 points**)
2. If `yeartype` were a factor variable, what would the summary statistics for that variable be? (**5 points**)
3. If `yeartype` were a numeric variable, what would the summary statistics for that variable be? (**5 points**)
4. Calculate summary statistics conditional on the value of yeartype. (**5 points**)
5. Plot a graph of `discoveries` vs. `airmiles` with the following characteristics:
   - the axes and title are given informative names (**5 points**)
   - each (x, y) point has a different character based on its `yeartype` (**5 points**)
   - each (x, y) point has a different color based on whether the corresponding `airmiles` value is greater than the overall average over each year or less than the overall average. Equality can be assigned to either group. (**5 points**)

Problem 6

Consider a sequence $b = (b_1, b_2, \ldots, b_k)$ where

$$b_i = i + k.$$

Find the k that maximizes the sum of the sequence subject to the ceiling L: $\sum_{i=1}^{k} b_i < L$. This is a strict inequality! Find these k's for $L \in \{10, 100, 1000\}$. (**15 points**)