University of Cincinnati

Statistical Computing-Stat 6045 Fall 2020

**Midterm Take Home Exam**

**(Due Tuesday 6 pm, November 10, 2020)**

Submit a copy of your code, results, and plots on Canvas. In addition you can also send an electronic version of all of your code to ([alex.konomi@uc.edu](mailto:alex.konomi@uc.edu)). Make sure the code is not platform dependent and can be run at any other machine in any subdirectory (have a separate file with the code that I can run without results).

**1)** (13 points) (a) Write a function to create a contingency table of adjacent k-tuples in a string of characters from the set A,C,G,T. For example, with k=2 and with the string ’CAGACAAAAC’, you would want to produce the following table (excluding the zero components):

AA AC CA AG GA

3 2 2 1 1

(b) To check your function, run it on a simulated string of 10,000 characters, drawn uniformly and independently from the set A,C,G,T .

(c) Build a function that can draw 10,000 times from the set (A,C,G,T) without any entries AAA and CG (have zero replicates of AAA and CG)

**2)** (10 points) *data(airquality)* contains measurements of ozone concentration in New York from May to September 1983, together with other relevant variables.

* 1. Plot the ozone concentration over time. Look at how different plotting types ("l", "h") affect the appear- ance. The EPA standard for ozone until recently was 120ppb, 140ppb was moderate nonattainment and 160ppb serious nonattainment. Indicate these with horizontal lines in appropriate colors. Use the text() function to annotate the severe ozone days with their dates.
  2. Ozone is produced by chemical reactions in the air that require sunlight. Draw a scatterplot of ozone and solar radiation.
  3. Perhaps wind or temperature are responsible for the shape of the plot. Use the coplot function to draw scatterplots of ozone and solar radiation for different levels of wind speed, of temperature, and of both at once. Do the same thing with the trellis command xyplot.

**3)** (12.5 points) In [statistics](https://en.wikipedia.org/wiki/Statistics), Wilcoxon **signed-rank test** is a [non-parametric](https://en.wikipedia.org/wiki/Non-parametric_statistics) [statistical hypothesis test](https://en.wikipedia.org/wiki/Statistical_hypothesis_testing) used when comparing two related samples, matched samples, or repeated measurements on a single sample to assess whether their population mean ranks differ (i.e. it is a [paired difference test](https://en.wikipedia.org/wiki/Paired_difference_test)). It can be used as an alternative to the [paired Student's t-test](https://en.wikipedia.org/wiki/Student%27s_t-test), *t*-test for matched pairs, or the *t*-test for dependent samples when the population cannot be assumed to be [normally distributed](https://en.wikipedia.org/wiki/Normally_distributed). A Wilcoxon signed-rank test is a nonparametric test that can be used to determine whether two dependent samples were selected from populations having the same distribution. Unlike the [*t*-test](https://en.wikipedia.org/wiki/T-test) it does not require the assumption of [normal distributions](https://en.wikipedia.org/wiki/Normal_distribution). It is nearly as efficient as the *t*-test on normal distributions. Following the attached steps of algorithms and the data from two groups:

1. Build a function that calculates the W value (as it is described in the attached rtf document).
2. Build a function that calculates the p-value using exact distribution with 20,000 replications as it is explained in the attached rtf document. Calculate the p-value without using the normal approximation
3. Use the normal approximation and calculate the p-value. Compare with (c). Are these numbers similar?

* Please do not use the R build in Wilcoxon test. However, you can compare it with the results you get.

**4.)** (12.5 points) Students were recently polled on 6 separate items measuring the availability and cost of texts at the University bookstore. Their responses are 1=“Strongly Disagree”, 2=”Disagree”, 3=”Neutral”, 4=”Agree”, 5=”Strongly Agree”, 6=”Missing”. These coded answers for variables Y1-Y6 are stored in the Excel worksheet, along with student IDs.

(a) Read the data into SAS as the data set Survey.

(b) Use arrays and DO loops to create the character variables Q1-Q6; these variables correspond to the descriptive labels provided for the numeric variables Y1-Y6.

(c) Calculate the sum of Y1-Y6 after this recoding; don’t worry about missing data.

(d) Create a format Likertfmt, which will print the character strings shown above in place of each of the coded values 1-5, and ”Missing” for missing values. Print a frequency table for variable Y1 using the Likertfmt in your table.

**5)** (12 points) In [mathematics](https://en.wikipedia.org/wiki/Mathematics), **Buffon's needle problem** is a question first posed in the 18th century by Georges-Louis Leclerc, Comte de Buffon and it is used to estimate the value of numerically. For more information on this please check <https://en.wikipedia.org/wiki/Buffon%27s_needle_problem>. The Buffon-Laplace needle problem is an extension in a grid two dimensional space with rectangle grids and triangular grids: <http://mathworld.wolfram.com/Buffon-LaplaceNeedleProblem.html> and [http://mathworld.wolfram.com/CleanTileProblem.html](http://mathworld.wolfram.com/CleanTileProblem.html%20). If you like to read more information you can look at the paper in the Journal of The American Statistician “Sharpening Buffon's Needle”: <https://www.jstor.org/stable/pdf/2683484.pdf>

Write a function in R or in SAS (depending on the preference) that you can use in a Monte Carlo simulation study to compute the value of for triangular grid (If you want you can do the rectangular for 9/12 points). Show the distribution (here you must simulate multiple times and show the results--with a histogram or a kernel will be fine) of the estimated for n=20,50,100, and 10,000. Make a 2-dimensional plot of one single process for n=100.