Assignment 3 - MAT 581

1. Make a table of values and probabilities for the Wilcoxon signed rank statistic T^+ for the case when $n = 4$, similar to what we did in class for $n = 3$.

2. Laureysens et al. (2004) measured metal content in the wood of 13 poplar clones growing in a polluted area, once in August and once in November. Concentrations of aluminum (in micrograms of Al per gram of wood) are shown below.

Clone	August	November	August-November
Columbia River	18.3	12.7	-5.6
Fritzi Pauley	13.3	11.1	-2.2
Hazendans	16.5	15.3	-1.2
Primo	12.6	12.7	0.1
Raspalje	9.5	10.5	1.0
Hoogvorst	13.6	15.6	2.0
Balsam Spire	8.1	11.2	3.1
Gibecq	8.9	14.2	5.3
Beaupre	10.0	16.3	6.3
Unal	8.3	15.5	7.2
Trichobel	7.9	19.9	12.0
Gaver	8.1	20.4	12.3
Wolterson	13.4	36.8	23.4

There are two nominal variables: time of year (August or November) and poplar clone (Columbia River, Fritzi Pauley, etc.), and one measurement variable (micrograms of aluminum per gram of wood). The differences are somewhat skewed; the Wolterson clone, in particular, has a much larger difference than any other clone.

Make a comparative boxplot of the concentraion of aluminum for the samples collected in August and in November.

Analyze the data to see if there is a difference in the concentraion of aluminum in samples collected at different times of the year, using:

a. paired t-test

b. Wilcoxon signed-rank test

c. a Fisher rank test

d. permutation

e. bootstrap

Also. provide a confidence interval for the difference in concentration of aluminum found.

3. Here are some data on Wright's FST (a measure of the amount of geographic variation in a genetic polymorphism) in two populations of the American oyster, Crassostrea virginica. McDonald et al. (1996) collected data on FST for six anonymous DNA polymorphisms (variation in random bits of DNA of no known function) and compared the FST values of the six DNA polymorphisms to FST values on 13 proteins from Buroker (1983). The biological question was whether protein polymorphisms would have generally lower or higher FST values than anonymous DNA polymorphisms. McDonald et al. (1996) knew that the theoretical distribution of FST for two populations is highly skewed, so they analyzed the data with a non-parametric test. Carry out explorative data analysis, do the test that compares the two groupsand provide a confidence interval for the difference in median of the two groups.

gene	class	FST
CVJ5	DNA	-0.006
CVB1	DNA	-0.005
6Pgd	protein	-0.005
Pgi	protein	-0.002
CVL3	DNA	0.003
Est-3	protein	0.004
Lap-2	protein	0.006
Pgm-1	protein	0.015
Aat-2	protein	0.016
Adk-1	protein	0.016
Sdh	protein	0.024
Acp-3	protein	0.041
Pgm-2	protein	0.044
Lap-1	protein	0.049
CVL1	DNA	0.053
Mpi-2	protein	0.058
Ap-1	protein	0.066
CVJ6	DNA	0.095
CVB2m	DNA	0.116
Est-1	protein	0.163

4. Using a non parametric test to check if the women's height are related to age in a pairwise fashion. Carry out the six possible pairwise comparisons as well, providing confidence intervals for each pair difference. For example, you need to compare 20-29 vs. 30-39 and 20-29 vs. 40-49 etc. As usual, start the analysis but doing exploratory data analysis.

20-29:	63.75	68.25	62.25	67.25
30-39:	64.75	67.5	64.75	66.5
40-49:	68.5	64.25	64.5	66
50-59:	65.25	64.75	67.5	

5. Using the Kolmogorov-Smirnov two-sample test, show that the t statistic with large degrees of freedom can be approximated by the standard normal distribution. How large should the degrees of freedom be? Hint: Do simulations on a large sample.

6. Compute Kendall's tau and Spearman rho together with their corresponding 95% confidence intervals, for the first sample problem in the folder NP-tests.