STAT 621 HOMEWORK 2
Due: Monday Sept. 16

1. Data below (and on Blackboard as `hamilton.txt`) are measurements of Hamilton depression scale factors on nine subjects diagnosed with anxiety and depression. Measurements were taken pre (X) and post (Y) therapy. (From Hollander, Wolfe and Chicken)

```
        x       y
1    1.83   0.878
2    0.50   0.647
3    1.62   0.598
4    2.48   2.050
5    1.68   1.060
6    1.88   1.290
7    1.55   1.060
8    3.06   3.140
9    1.30   1.290
```

   (a) Is there evidence to suggest that therapy reduces the Hamilton depression scale factor, in general? Use the Wilcoxon test at significance level $\alpha = .05$.

   (b) Change the value of X_3 from 1.62 to 16.2. How does this outlying observation affect the calculation and result of the test?

   (c) Construct an example in which changing one observation has a marked effect on the final decision of the hypothesis test.

   (d) What do parts b and c suggest about the relative robustness of the signed rank test to outliers?

2. Consider testing $H_0 : \theta = 0$ vs. $H_A : \theta > 0$ with the Wilcoxon signed rank test on the following observed differences Z_i,

```
2.5   3.7   0.0   -0.6   4.7   0.0   1.4   0.0   1.9   5.2
```

   Compute the test statistic W and the p-value based on

   (a) Discarding the 0-values and reducing n accordingly

   (b) The conservative approach to 0-values

   Compare the results.

3. An experiment compared two methods of wastewater treatment for removing benzene, a common industrial solvent. Both treatments were applied to three samples having the same initial benzene concentration. The benzene concentration after treatment was as follows

Treatment	1	1	1	2	2	2
Concentration	7.8	6.8	6.2	4.1	6.5	5.9

(a) Use a permutation test to determine if the mean benzene concentration is lower for treatment 2 compared to treatment 1. State your hypotheses, define your test statistic, derive the null distribution of your statistic, report the p-value and state your conclusion.

(b) Repeat the test but this time use a randomization test. Compare your results.

4. A study measured invertebrate species richness (number of distinct species) in a number of lakes in Idaho. Other lake characterstics of the lake were recorded including area (square km). Data appear below.

Species	21	30	32	37	44	49	58	62	75	78	85
Area	1.2	2.4	0.9	3.1	2.5	1.6	3.0	2.5	2.1	3.0	4.0

Suppose interest is in testing whether there is a nonzero correlation between species richness and area. Specifically we want to test

$$H_0 : \rho = 0 \qquad \text{vs.} \qquad H_1 : \rho \neq 0$$

where ρ is the true correlation.

(a) Imagine using a permutation test. Suppose that your test statistic is r, the usual sample correlation. Explain how you would derive the null distribution of the statistic for a permutation test (you don't have to actually do it).

(b) Test the hypothesis using a randomization test, approximating the null distribution by simulation. Note the R command to compute the sample correlation between vectors `x` and `y` is `cor(x,y)`.

5. Consider the species data from the previous problem. Use the bootstrap to approximate the sampling distribution of the sample correlation. Use a bootstrap sample of size $B \geq 200$. Construct a histogram, calculate the bootstrap mean and standard deviation, and report a 95% confidence interval for ρ. Are your results consistent with the conclusion in the previous problem?

6. The data set `income.txt` on blackboard contains two columns: x= the number of adults in a household and y = combined income of the household. Consider two estimators of the mean income per person,

$$\widehat{\theta}_1 = \frac{1}{n}\sum \frac{y_i}{x_i} \qquad \text{and} \qquad \widehat{\theta}_2 = \frac{\overline{y}}{\overline{x}}$$

Use the jackknife to estimate the bias and variance of these two estimators. Which would you say is the better estimator?

7. For the income data in the previous problem, consider fitting a polynomial relationship between y =income and x =household size. Use 5-fold cross validation to choose the polynomial (i.e., linear, quadratic, cubic, ...) with the smallest predictive error.