Nonparametric Statistics and Bootstrapping
Midterm Exam #1

Due Date: March 15, 2018 - By 11:59pm

All your answers must be written on a separate sheet, properly typeset and submitted in the form of a report, in pdf format. No MS Word report will be accepted. No handwritten report will be accepted either, and I mean none. Make sure your report is in pdf format, and is uploaded to the designated dropbox folder by the deadline. In fact, all reports prepared with RMarkDown are preferred, although nice pdf from Latex or Word are also acceptable. Be sure to properly indicate which question you are answering, and write your name on the title page of your report.

Exercise 1: (42.5 points)

The statistician at a store has collected a random sample of 11 ratings given by their customers. The ratings are given as Likert-type entries. The randomly selected customers responded to the invitation: "Please rate your experience at our store", and their responses could be one of

$$\{\texttt{Very Dissatisfied}, \texttt{Dissastified}, \texttt{Neutral}, \texttt{Satisfied}, \texttt{Very Satisfied}\}$$

Their responses were then coded into pseudo-numbers as follows

Very Dissatisfied	Dissatisfied	Neutral	Satisfied	Very Satisfied
◯	◯	◯	◯	◯
1	2	3	4	5

The random sample obtained by the statistician at this store is: *Very Satisfied*, *Very Dissatisfied*, *Dissatisfied*, *Neutral*, *Neutral*, *Very Satisfied*, *Satisfied*, *Very Satisfied*, *Very Satisfied Satisfied*, *Very Satisfied* The general manager at this store is claiming that typical sentiment of his customers about her store is more than neutral. The statistician is primarily interested in running a sign test to assess her manager's claim

1. Translate the raw data into pseudo numbers
2. Specify clearly the "parameter" in the nonparametric setting that best captures what must measured to assess the manager's claim. Denote that quantity θ.
3. Write down the null and alternative hypotheses of the hypothesis testing task translates the manager's claim

4. Write down the formula of the appropriate test statistic B for the sign test to be used to assess the manager's claim [*Hint: Use the exclusion approach on ties, so that used sample size is reduced]*
5. Write down the sampling distribution of B
6. At significance level $\alpha = 0.05$, write down the rejection region $\texttt{RR}_{0.05}$.
7. Compute $B_{\texttt{obs}}$, the observed value of the test statistic for the data collected by this store's statistician.
8. Provide your final decision on this test at significance level $\alpha = 0.05$.
9. Compute the Pvalue for this test and comment on what it says.
10. Provide a 95% lower confidence bound for θ.
11. Use the `R` function `SIGN.test()` to perform the same test performed step by step earlier.
12. A gentleman named `George` decided to perform a traditional t-test on this data.
    - Specify what parameter `George` will base his test on.
    - Write the null and alternative hypotheses used by `George`.
    - Perform `George`'s test using the `R` command `t.test`.
    - What conclusion should `George` draw about the typical sentiment (rating) of this store?
    - Comment and compare the two conclusions.
    - Which test should you rely on in this case? Explain your answer in great details with all supporting theories and tests wherever necessary.

Exercise 2: (42.5 points)

In an effort to improve their web presence and effectiveness, the data scientist at a company commissioned the collection of number of clicks per thousand visits on several random selected occasions. Two different designs are considered, and it is of interest to check if the distribution of the number of clicks from one design to the other. Specifically, the data scientist would like to test if the distribution of the number of clicks Y in the second design is positively shifted from that of the first design. The data collected by the data scientist is

$$\mathbf{x} = (614, 653, 982, 761, 662, 931, 842, 733, 728)$$

$$\mathbf{y} = (969, 796, 893, 886, 964, 962, 992, 617, 955, 936, 992)$$

Consider using the Wilcoxon-Mann-Whitney test procedure to solve this problem.

1. Specify the null and alternative hypothesis at play here, based on the description/statement of the problem given earlier. Be sure to first clearly define the parameter of interest

2. Indicate clearly the assumptions made by the test procedure
3. Write down the expression of the test statistic to be used for this test
4. Compute $U_{\texttt{obs}}$, the observed value of the Mann-Whitney test statistic U for this dataset
5. Deduce $W_{\texttt{obs}}$, the observed value of the test statistic W for this dataset
6. Find the critical value w_α at significance level $\alpha = 0.05$.
7. Write down RR, the rejection region for this test.
8. Based on the above, provide your decision on the upper sided test on the water transfer.
9. Using the R command pwilcox, compute the Pvalue for the upper sided test.
10. Perform the test directly in R using the following command

```
wilcox.test(X, Y, alternative='greater',  conf.int=F)
```

11. Compute $W^\star_{\texttt{obs}}$, the observed value of the test statistic $W^\star$ for this dataset
12. Write down RR, the rejection region for this test for the approximate large sample test
13. Provide your conclusion on the test
14. Compute the Pvalue for the approximate test
15. Perform the test of normality on both X and Y using shapiro.test(X)
16. Perform the traditional two sample t-test using t.test(X, Y, alternative='less')
17. Comment all the nuances between to two families of procedures.

Exercise 3: (15 points)

Let $\Delta > 0$ be positive real number. Consider the Wilcoxon-Mann-Whitney upper tail test $H_0 : \Delta \leq 0$ vs $H_a : \Delta > 0$ aimed at testing the difference between the distribution F of X and the distribution G of Y, based on $X_1, \cdots, X_m$ and $Y_1, \cdots, Y_n$. The distribution-free Rank Sum Test created by Wilcoxon, Mann and Whitney can be summarized simply as follows: Let $N = m + n$, and let $S_1, S_2, \cdots, S_n$ denote the ranks assigned to $Y_1, \cdots, Y_n$ respectively in the joint sample of size $N = m + n$. Define the following statistic

$$W = \sum_{j=1}^{n} S_j.$$

Now, recall that the Mann-Whitney statistic is given by

$$U = \sum_{i=1}^{m} \sum_{j=1}^{n} h(X_i, Y_j),$$

where

$$h(X_i, Y_j) = \begin{cases} 1 & \texttt{if } X_i < Y_j, \\ 0 & \texttt{otherwise.} \end{cases}$$

1. Show that
$$\mathbb{E}_0(W) = \frac{n(m+n+1)}{2}$$
2. Show that
$$\mathbb{V}_0(W) = \frac{mn(m+n+1)}{12}$$
3. Show that
$$W = U + \frac{n(n+1)}{2}$$
4. Show that
$$\mathbb{E}_0(U) = \frac{mn}{2}$$
5. Show that
$$\mathbb{V}_0(U) = \frac{mn(m+n+1)}{12}$$

Exercise 4: (10 points)

Consider the upper tail $H_0 : \theta \leq 0$ vs $H_a : \theta > 0$ sign test with the test statistic $B = \sum_{i=1}^{n} \psi_i$ where $\psi_i = \mathbb{1}(Z_i > \theta)$, where $\mathbb{P}(\psi_i = 1) = \mathbb{P}(Z_i > \theta) = p$. With the assumption that the n observations are iid, answer the following questions.

1. Show that B has binomial distribution and specify its parameters.
2. Write down the null distribution of B.
3. Find and write down the expression of the expected value of B under the hypothesis of no treatment effect.
4. Find and write down the expression of the variance of B under the hypothesis of no treatment effect.
5. Write down the expression of the power of the upper tail sign test when the alternative has p_0.