Assignment 1

(Due 11pm on Monday, 19 October)

Instructions:

- This assignment consists of 8 questions, to be completed independently by each student.
- Questions 1 – 4 (Q1 – 4) are True/False questions requiring explanations.
- Questions 5 – 8 (Q5 – 8) are problem-solving questions requiring detailed solutions.
- It will count for 20% of assessment.
- Each of Q1 – 4 consists of parts (a) – (c). For each part, choose "T" if the statement is true, or "F" if false.
- Justify your choice of T or F, including correcting false statements.
- Marking scheme for each part (a) – (c) of Q1 – 4:
  - 1 mark for a correct choice of T or F, and 0 mark for incorrect choice;
  - 3 marks for convincing reasons, 1 or 2 marks for partially correct reasons, and 0 mark for incorrect or irrelevant reasons;
  - 4 marks maximum for each part; 12 marks for each of Q1 – 4.
- For Questions 5 – 8, work out the details and show the steps to solve each problem, including the right theory and methods used, appropriate formulae to calculate the answers, and the steps of calculations.
- The marks for Q5 – 8 are indicated in each part of the questions.
- The maximum total mark of the assignment is 100.
- Submit a pdf file of your answers in **typed** (not handwritten) contexts by Sunday 11pm, 13 October 2019.
- Your TA will advise you on how to submit your answers.

Rules for use of R programme:

- If a question indicates to use R, present relevant input/output with R-commands in your answers – which must be in your own words.
- For any question (or part of a question) with no mention of using R, your submitted answers should not rely on R.

True/False questions

Question 1 [12 marks]

The following statements are correct:

(a) Let X denote a symmetric random variable about $a \in \mathbb{R} = (-\infty, \infty)$ with a cumulative distribution function (cdf) $F(x)$. Then

$$F(2a - x) = 1 - F(x) \text{ holds for all } x \in \mathbb{R}$$

if and only if $F(x)$ is a continuous function on $\mathbb{R}$.

(b) Randomly select (b_1, b_2, b_3) from distinct numbers $\{a_1, a_2, \ldots, a_{10}\}$ without replacement. Then the distribution of the random variable

$$X = b_1 b_2 + b_1 b_3 + b_2 b_3 - 3 b_1 b_2 b_3$$

can be determined by

$$\Pr(X = x) = \frac{\text{No. } \{(b_1, b_2, b_3):\ b_1 b_2 + b_1 b_3 + b_2 b_3 - 3 b_1 b_2 b_3 = x;\ b_1 < b_2 < b_3\}}{120}$$

for each possible value x of X.

(c) In a test of the null hypothesis H_0 against the alternative H_1, if the p-value of the test is 0.05, then there is 95% chance to accept a correct hypothesis H_1.

Question 2 [12 marks]

Let $X_1, \ldots, X_n$ be independent continuous random variables with a common median θ.

(a) Define $Y_i = I_{\{X_i > 0\}}$, $i = 1, \ldots, n$. Then each Y_i has a parametric distribution, but the sign test for the null hypothesis $H_0 : \theta = 0$ based on $X_1, \ldots, X_n$ is nonparametric.

(b) The assumption of symmetric distributions for $X_1, \ldots, X_n$ ensures the symmetry of the Wilcoxon signed-rank test statistic T^+, but has no affects on the rejection rule of the null hypothesis $H_0 : \theta = 0$.

(c) To construct a nonparametric confidence interval of θ, it is necessary to first find a point estimate of θ.

Question 3 [12 marks]

Let T^+ denote the Wilcoxon signed rank statistic from a random sample of symmetric random variables $X_1,\ldots,X_n$ of size $n>10$, θ the common median of $X_1,\ldots,X_n$, and R_i the rank of X_i for T^+, $i=1,\ldots,n$. The following statements are true:

(a) $\Pr(T^+=9)=2^{3-n}$ under $H_0:\theta=0$.

(b) $\Pr(T^+\geq 30)\leq 0.5$ under $H_0:\theta=0$.

(c) If $X_{(5)}<0$, where $X_{(1)}<\cdots<X_{(n)}$ are the order statistics of $X_1,\ldots,X_n$, then the Walsh averages $W_{ij}<0$ for at least 15 pairs $\{(i,j):1\leq i\leq j\leq n\}$.

Question 4 [12 marks]

Given two independent random samples $(X_1,\ldots,X_m)$ and $(Y_1,\ldots,Y_n)$ with medians θ_X and θ_Y respectively, if $(Y_1,\ldots,Y_n)$ have mostly smaller values but a substantially wider range than those of $(X_1,\ldots,X_m)$, then the following statements are reasonable:

(a) The sample $(Y_1,\ldots,Y_n)$ is likely to have a smaller median but a greater variance than those of $(X_1,\ldots,X_m)$.

(b) The Wilcoxon rank sum test is likely to reject the null hypothesis $H_0:\theta_X=\theta_Y$ in favor of the alternative $H_1:\theta_X>\theta_Y$.

(c) The Ansari-Bradley rank test is likely to reject $H_0:\mathrm{Var}(X)=\mathrm{Var}(Y)$ in favor of the alternative $H_1:\mathrm{Var}(X)<\mathrm{Var}(Y)$.

[Questions 5 – 8 start from next page]

Problem-solving questions

Question 5 [17 marks]

A company has adopted a new technology to produce a certain type of products. The numbers of such products made by 11 factories of the company before and after using the new technology are recorded in the table below:

Factory	Before	After
1	525	614
2	718	805
3	650	590
4	387	455
5	882	938
6	936	1050
7	584	540
8	256	356
9	630	721
10	462	489
11	535	490

Let X and Y represent the numbers of products before and after using the new technology, respectively, and θ the median of the difference $Z = Y - X$.

Based on the data provided in the above table, perform the following analyses:

(a) Calculate the exact p-value of testing the null hypothesis $H_0 : \theta = 0$ against the alternative $H_1 : \theta > 0$ by the sign test. [2 marks]

(b) Obtain an exact confidence interval of θ with a target at least 90% confidence level based on the sign statistic. [3 marks]

(c) Determine if there is sufficient evidence at the 5% level that the new technology is effective to increase the production of the company by the Wilcoxon signed rank test using the exact p-value from enumeration. [5 marks]

(d) Estimate the median θ and obtain its exact confidence interval with a target at least 95% confidence level based on the Wilcoxon signed ranks. [4 marks]

(e) Compare the p-values of the tests and the confidence intervals of θ from the sign test statistic and the Wilcoxon signed ranks in parts (a) – (d), and comment on the differences between the two methods. [3 marks]

Question 6 [15 marks]

Let X_1, X_2 be independent random variables with densities $f_1(x), f_2(x)$ respectively, and R_1, R_2 the Wilcoxon signed ranks of X_1, X_2 respectively. Define

$$S = I_{\{X_1>0\}} + 2I_{\{X_2>0\}} \quad \text{and} \quad T^+ = R_1 I_{\{X_1>0\}} + R_2 I_{\{X_2>0\}},$$

(a) Calculate the probabilities:

$$\Pr(S=2), \ \Pr(X_1>0, R_1=2, X_2<0), \ \Pr(X_1<0, R_2=2, X_2>0) \ \text{ and } \ \Pr(T^+=2)$$

with $f_1(x) = 0.5 I_{\{|x|\le 1\}}$ and $f_2(x) = e^{-2|x|}$. [7 marks]

(b) Repeat Part (a) with $f_1(x) = f_2(x) = f(x) = I_{\{-0.5\le x<0\}} + 2(1-x)^3 I_{\{0\le x\le 1\}}$. [5 marks]

(c) Comment on the results of Parts (a) and (b). [3 marks]

Question 7 [10 marks]

Two independent samples are given by

$$(X_1,\ldots,X_6) = (1,-3,3,12,8,-1) \quad \text{and} \quad (Y_1,Y_2,Y_3,Y_4) = (-1,6,1,12)$$

Denote the ordered values of $(X_1,\ldots,X_6,Y_1,\ldots,Y_4)$ by $Z_1 \le \cdots \le Z_{10}$.

(a) Let W be the two-sample Wilcoxon rank sum statistic, w the value of W observed from the two samples given above, and $(r_1,\ldots,r_{10})$ the ranks of $(Z_1,\ldots,Z_{10})$ with average ranks assigned to ties.

Calculate the value of w and find all 4-tuples (r_i, r_j, r_k, r_l) such that $r_i + r_j + r_k + r_l = w$ and $i<j<k<l$. Then determine $\Pr(W=w)$ conditional on observed ties under the null hypothesis of no treatment effect. [4 marks]

(b) Let C be the Ansari-Bradley test statistic for the two-sample dispersion problem and $(a_1,\ldots,a_{10})$ the scores of $(Z_1,\ldots,Z_{10})$ for C with average scores assigned to ties.

Calculate the observed value c of C and find all 4-tuples (a_i, a_j, a_k, a_l) with $i<j<k<l$ such that $a_i + a_j + a_k + a_l = c$. Then determine $\Pr(C=c)$ conditional on observed ties under the null hypothesis of equal dispersion between the two samples. [6 marks]

Question 8 [10 marks]

Two independent samples $X = (X_1, \ldots, X_{14})$ and $Y = (Y_1, \ldots, Y_{16})$ are recorded below:

X =(6.17, 4.78, 3.99, 5.65, 3.87, 4.43, 4.82, 6.68, 4.46, 6.95, 3.02, 4.22, 4.21, 3.97)

Y = (9.94, 7.08, 7.14, 5.82, 9.60, 10.09, 8.66, 4.74, 4.14, 10.92, 5.61, 6.47, 5.20, 8.21, 3.55, 9.81)

Use R to carry out the following statistical analyses based on the samples X and Y (show the R-commands and output):

(a) Under the location-shift model, find the *p*-value of the Wilcoxon rank sum test to determine the level of evidence for sample Y to have a greater location parameter than sample X. [3 marks]

(b) Under the location-scale parameter model, find the *p*-value of the Ansari-Bradley rank test to determine the level of evidence for different dispersions between the two samples X and Y. [3 marks]

(c) Let $X + 2 = (X_1 + 2, \ldots, X_{14} + 2)$ denote the sample by adding 2 to each X_i, $i = 1, \ldots, 14$. Repeat the test in Part (b). Comment on the results in Parts (b) and (c). [4 marks]