STA3007 Nonparametric Statistics

Assignment 2

(Due 11pm on Monday, 30 November 2020)

Instructions:

- This test consists of 6 questions, to be completed independently by each student.
- Questions 1 – 3 (Q1 – 3) are True/False questions requiring explanations.
- Questions 4 – 6 (Q4 – 6) are problem-solving questions requiring detailed solutions.
- It will count for 20% of assessment.
- Each of Q1 – 3 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 – 3:
  * 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 – 3.
- For Questions 4 – 6, 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 Q4 – 6 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 Monday 11pm, 30 November 2020.
- 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 for submission.
- 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 true in a one-way layout model with k treatments and sample sizes $n_1,\ldots,n_k$:

(a) If $n_1=\cdots=n_k=n$, then it requires $(nk)!\big/\left[k!(n!)^k\right]$ rank assignments to determine the exact distribution of the Kruskal-Wallis test statistic H.

(b) If $k=3$ and $n_1=n_2=n_3=2$, then the distribution of the Jonckheere-Terpstra test statistic J for ordered alternatives can be determined by 15 rank assignments.

(c) If the Jonckheere-Terpstra test rejects the null hypothesis $H_0:\tau_1=\cdots=\tau_k$ at the level α of significance, so will do the Kruskal-Wallis test.

Question 2 [12 marks]

Refer to multiple comparisons in a one-way layout in Lecture Notes Section 5.

(a) A multiple comparison procedure is a special case of hypothesis test.

(b) Under $H_0:\tau_1=\cdots=\tau_k$, if the W_{ij} in (5.15) is asymptotically normal, then

$$W_{ij}^* \sim \frac{Z_i - Z_j}{\sqrt{(n_i+n_j)/(2n_in_j)}} \quad \text{approximately for large samples,}$$

where $Z_i \sim N(0,1/n_i)$ are independent random variables, $i=1,\ldots,k$.

(c) If $k=3$ and the Steel-Dwass-Critchlow-Fligner (SDCF) two-sided all-treatment multiple comparison procedure decides $\tau_1=\tau_2$, $\tau_1=\tau_3$, $\tau_2\neq\tau_3$ at $\alpha=0.05$ exact, then $\Pr(\text{The decision})=0.05$ under $H_0:\tau_1=\tau_2=\tau_3$.

Question 3 [12 marks]

Consider a two-way layout model with k treatments and n blocks.

(a) $\sum_{j=1}^{g_i} t_{i,j}^3 - k < \sum_{j=1}^{g_i} t_{i,j}\left(t_{i,j}-1\right)\left(t_{i,j}+1\right)$ holds in (6.5) of Lecture Notes.

(b) The Page test statistic L for ordered alternatives is always greater than $2\mathrm{E}_0[L]/3$.

(c) For unbalanced incomplete data, if each block has an equal number of observations, then $A_j = aR_j + b$ for some constants a and b, where A_j is defined in (6.21) and R_j is the sum of in-block ranks for treatment j, $j=1,\ldots,k$.

Problem-solving questions

Question 4 [20 marks]

The data on two independent random samples are recorded below:

$X = (X_1, \ldots, X_{11}) = (7.5, 3.5, 16.8, 3.8, 4.6, 2.8, 6.1, 15.7, 8.8, 9.2, 9.8)$

$Y = (Y_1, \ldots, Y_{10}) = (14.6, 2.5, 11.6, 2.2, 12.3, 12.7, 10.1, 14.1, 15.5, 13.5)$

Assume the location-scale parameter model for the data.

(a) Test the null hypothesis $H_0 : \mathrm{Var}(X) = \mathrm{Var}(X)$ against $H_1 : \mathrm{Var}(X) \neq \mathrm{Var}(X)$ by the Miller's Jackknife test at appropriate level of significance (present the values of S_i, T_j, A_i, B_j defined for the test statistic Q). [6]

(b) Test the null hypothesis of no difference in location and/or dispersion parameters between the two samples by the Lepage test at appropriate level of significance (show the ranks and scores used to calculate the test statistic). [6]

(c) Calculate the values of the empirical distribution functions $F_{11}(t)$ and $G_{10}(t)$ of X and Y, respectively, at ordered values $Z_{(1)} < \cdots < Z_{(21)}$ of combined (X, Y) to find the value of the two-sample Kolmogorov-Smirnov test statistic J.

Then use the R program to obtain the *p*-value of the test and decide whether there is sufficient evidence for general differences between the distributions of X and Y at appropriate level of significance. [6]

(d) Comment on the following issues based on the results of parts (a) – (c):

1) The overall differences between the two samples;
2) Whether the location-scale parameter model is appropriate, and why. [2]

Question 5 [24 marks]

In a one-way layout with data $\{X_{ij}, i=1,\ldots,n_j; j=1,\ldots,5\}$, the values of

$$U_{uv} = \sum_{i=1}^{n_u}\sum_{j=1}^{n_v} I_{\{X_{iu}<X_{jv}\}} = \text{No. } \{(i,j): X_{iu} < X_{jv}, i=1,\ldots,n_u; j=1,\ldots,n_v\},\ 1 \le u < v \le 5,$$

are provided below, where $(n_1,\ldots,n_5)=(6,5,7,4,6)$:

U_{uv}				
	v			
u	2	3	4	5
1	26	36	20	32
2		28	11	18
3			16	16
4				10

Let $\tau_1,\ldots,\tau_5$ denote the effects of the 5 treatments and the null hypothesis of interest is $H_0: \tau_1 = \cdots = \tau_5$.

Carry out the following tests in parts (a) – (c) at appropriate level α of significance.

(a) Test H_0 against ordered alternatives $H_1: \tau_1 \le \tau_2 \le \tau_3 \le \tau_4 \le \tau_5$ with at least one strict inequality by the Jonckheere-Terpstra test using both the exact rejection rule by R and the large-sample normal approximation. [7]

(b) Test H_0 against umbrella alternatives $H_1: \tau_1 \le \tau_2 \le \tau_3 \ge \tau_4 \ge \tau_5$ with at least one strict inequality by the Mack-Wolfe test with known peak $p=3$ using both the exact rejection rule by R and the large-sample normal approximation. [7]

(c) Test H_0 against umbrella alternatives $H_1: \tau_1 \le \cdots \le \tau_p \ge \cdots \ge \tau_5$ with at least one strict inequality by the Mack-Wolfe test with unknown peak p using R. [7]

(d) Based on the results in parts (a) – (c), what alternatives (ordered or umbrella) have the strongest support from the data according to the level of significance? Are the test results of (a) – (c) contradictive or consistent? Explain why. [3]

Question 6 [20 marks]

Consider a balanced incomplete block design (BIBD) with k treatments, n blocks, each treatment appearing in p blocks, s treatments observed in each block, and λ pairs of treatments available in each block.

Let $c_{ij} = 1$ if treatment j is available in block i, otherwise $c_{ij} = 0$, r_{ij} denote the rank of X_{ij} in block i with $r_{ij} = 0$ if $c_{ij} = 0$, and $R_j = r_{1j} + \cdots + r_{nj}$.

(a) Let D denote the Durbin-Skillings-Mack test statistic for general alternatives in BIBD. Prove $\mathrm{E}[D] = k - 1$ using

$$\mathrm{E}[r_{ij}] = \frac{s+1}{2} I_{\{c_{ij}=1\}} \quad \text{and} \quad \mathrm{Var}(r_{ij}) = \frac{(s+1)(s-1)}{12} I_{\{c_{ij}=1\}} \qquad [8]$$

(b) The following table presents the data $\{X_{ij}\}$ in an incomplete block design:

Block	Treatment				
	1	2	3	4	5
1	21	15	17	–	28
2	25	–	19	35	32
3	39	32	35	44	–
4	–	22	16	24	30
5	38	34	–	45	42

Test the null hypothesis $H_0 : \tau_1 = \cdots = \tau_k$ against general alternatives at appropriate level of significance. [8]

(c) Given that $q_{0.1} = 3.479$ for $k = 5$, decide the differences between treatment effects $\tau_1, \ldots, \tau_5$ based on the data in part (b) by the Skillings-Mack two-sided all-treatment multiple comparison procedure for BIBD with $\alpha = 0.1$. [4]