2014.10.21.

Nonparametric Statistics & Function Estimation Midterm Exam (solution)

※ Notice
- Your mobile devices should be turned off.
- The exam will take place from 11:40 till 1:20.
- All your answer should be written on this document.
- Raise your hand silently if you have any questions.

I declare that I will not cheat on the exam. _____

Department _____

Student number _____

Name _____

1. (25 pts) The following table gives the lengths (in inch) of sunfishes:

3.03	5.53	5.60	9.30	9.92	12.51	12.95	15.21	16.04	16.84

(a) (10 pts) Based on the Sign statistic, test if the median length is larger than 3.7 at the significance level 0.05 and find an exact confidence interval for the median with a confidence level of around 90%.

(b) (5 pts) Using an asymptotic method, find a 90% confidence interval for the median. (Use this expression : $X_{(r)} \equiv \frac{X_{([r])} + X_{([r]+1)}}{2}$, $[r]$ is the largest integer among those smaller than r. e.g.] $X_{(1.2)} \equiv \frac{X_{(1)} + X_{(2)}}{2}$)

(c) (5 pts) Test the same hypotheses based on the Wilcoxon signed rank statistic at the significance level 0.05. Use a large sample approximation.

(d) (5 pts) Describe a method to construct a 90% confidence interval for the median based on the Wilcoxon signed rank statistic.

(a) $H_0 : m = 3.7 \quad vs \quad H_1 : m > 3.7$

$s_n = 9,\ P(S_n \geq 9) = \frac{11}{1024} < 0.05$ where $S_n \sim B(10, 0.5)$

$\therefore$ *reject* H_0, the median is larger than 3.7.

$P(S_n \geq 8) = \frac{56}{1024} = \frac{0.1094}{2}$ where $S_n \sim B(10, 0.5)$

$\therefore (X_{(3)}, X_{(8)}) = (5.60, 15.21)$ is a 89.06% confidence interval for the median.

(b) $k_{0.1} = \frac{10}{2} + z_{0.05}\sqrt{\frac{10}{4}} \fallingdotseq 7.59$

$\therefore (X_{(3.41)}, X_{(7.59)}) = (\frac{5.6+9.3}{2}, \frac{12.95+15.21}{2}) = (7.45, 14.08)$

(c) $w_n = 2 + 3 + \cdots + 10 = 54$

$z_0 = \frac{54 - \frac{10(10+1)}{4}}{\sqrt{\frac{10(10+1)(2*10+1)}{24}}} \fallingdotseq 2.7 > 1.64 = z_{0.05}$

$\therefore$ *reject* H_0 the median is larger than 3.7.

(d) Define $W_{ij} = \frac{X_i + X_j}{2},\ 1 \leq i \leq j \leq 10 = n$

Find k_α such that $P(W_n \geq k_\alpha) = \frac{0.1}{2}$ or set $k_\alpha = \frac{n(n+1)}{4} + z_{0.05}\sqrt{\frac{n(n+1)(2n+1)}{24}}$.

Then, $(W_{(n-k_\alpha+1)}, W_{k_\alpha})$ is a 90% confidence interval for the median.

2. (20 pts) The data below shows Nitrogen Concentrations in 4 lakes in agricultural areas, and 5 lakes in a nearby natural area.

Agricultural :	156	225	369	451	
Natural :	89	155	290	331	401

(a) (10 pts) Calculate the Wilcoxon rank sum statistic and the Mann-Whitney statistic and compare the locations of the distributions for Nitrogen Concentrations at the significance level 0.05, assuming the shapes of the distribution are the same. Use an asymptotic method.

(b) (5 pts) Construct a 95% confidence interval for the difference of the locations. Use an asymptotic method.

(c) (5 pts) What is your estimate for the difference of the locations?

(a)

89 155 156 225 290 331 369 401 451

$w_n = 1+2+5+6+8 = 22$

$mw_N = 22 - \frac{5(5+1)}{2} = 7$

$H_0 : \Delta = 0 \quad vs \quad H_1 : \Delta \neq 0$

$z_0 = \frac{7 - \frac{4\times 5}{2}}{\sqrt{\frac{4\times 5\times(4+5+1)}{12}}} \fallingdotseq -0.73 > -z_{0.025}$ do not reject H_0. No SHIFT!

(b) $D_{ij} = X_i - Y_j$

	[,1]	[,2]	[,3]	[,4]	[,5]
[1,]	67	1	-134	-175	-245
[2,]	136	70	-65	-106	-176
[3,]	280	214	79	38	-32
[4,]	362	296	161	120	50

$k_{0.05} = \frac{20}{2} + z_{0.025}\sqrt{\frac{50}{3}} \fallingdotseq 18.00$

$(D_{(3)}, D_{(18)}) = (-175, 280)$ is a 95% confidence interval.

(c) $\hat{\Delta} = median[D_{ij}] = \frac{D_{(10)} + D_{(11)}}{2} = \frac{50+67}{2} = 58.5$

3. (10 pts) There are three kinds of electric bulbs. One wants to know if the lifetimes of them are all the same or not. The following table gives a tested result for some selected samples:

Bulb A	Bulb B	Bulb C
12(3)	13(4)	11(2)
17(8)	15(6)	30(15)
10(1)	18(9)	28(13)
14(5)	16(7)	26(11)
	21(10)	29(14)
		27(12)
Rank Bulb A의 합 : (17)	Rank Bulb B의 합 : (36)	Rank Bulb C의 합 : (67)

Test if there is a difference among groups using a nonparametric method at the significance level 0.05.

$H_0 : \triangle_A = \triangle_B = \triangle_C \; VS \; H_1 :$ not H_0

$$KW_N = \frac{12}{N(N+1)} \sum_{i=1}^{K} \frac{R_{i\cdot}^2}{n_i} - 3(N+1) = 5.9808 < 5.991 = \chi^2_{0.05}(2) \quad \therefore don't \; reject \; H_0$$

NO DIFFERENCE!

4. (15 pts : 5 pts each) We have four data points as follows:

2 4 6 3

We want to test if the median (m) is positive at the significance level is 0.1.

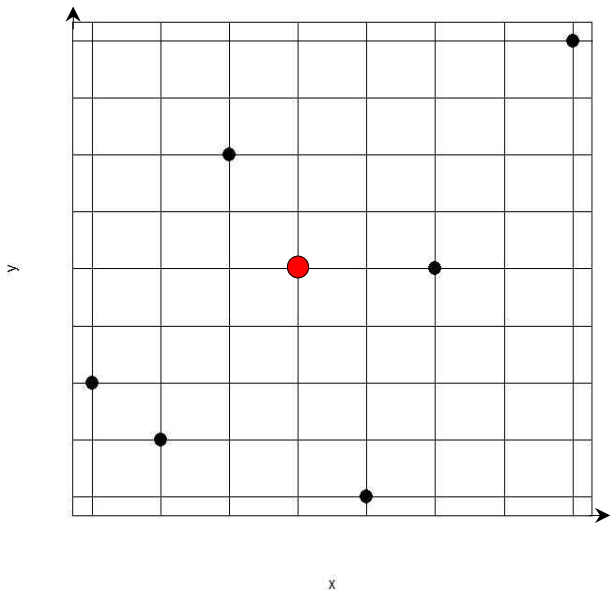
(a) Calculate the observed value of the Wilcoxon signed rank statistic (w_0).
(b) Calculate $P(W_n \geqq w_0)$ under $m = 0$ where W_n is the Wilcoxon signed rank test statistic from a sample of size n.
(c) What do you conclude from (b)? Answer with reasoning.

(a) $w_0 = 1+2+3+4 = 10$

(b) $P(W_n \geqq w_0 | m = 0) = P(W_n = w_0 | m = 0) = (\frac{1}{2})^4 = \frac{1}{16}$

(c) $H_0 : m = 0 \quad vs \quad H_1 : m \neq 0$. $P(W_n \geqq w_0 | m = 0)$ is the p-value and is smaller than 0.1. Therefore, we reject the null hypotheses We conclude $m > 0$.

5. (20 pts) The next figure is a scatter plot with two variables. We observe six points. Note that the horizontal and vertical lines are all equally spaced.



(a) (15 pts) Calculate the Pearson correlation coefficient, the Spearman correlation coefficient and the Kendall's tau correlation coefficient, respectively.

(b) (5 pts) Test if there is a monotonic relationship between two variables at the significance level 0.05.

(a)

Spearman

ranks of pts : (1,3) (2,2) (3,5) (4,1) (5,4) (6,6)

Therefore, $r_s = 1 - \frac{6}{6(6+1)(6-1)}(2^2+0^2+2^2+3^2+1^2+0^2) = \frac{17}{35} = 0.4857$

Kendall's tau

# of concordant pair : C = 10 , # of discordant pair : D = 5

Therefore, $\tau = \frac{10-5}{10+5} = \frac{1}{3} = 0.3333$

Pearson

The pearson correlation coefficient is location-invariant. Therefore, we can assign any coordinate values to pts without loss of generality. If we set the red point as the origin, the coordinate values of pts is expressed as:

(-3,-2) (-2,-3) (-1,2) (1,-4) (2,0) (4,4)

Then, $r = 0.5531$

(b) H_0 : there is no monotonic relationship vs H_1 : not H_1

$z_0 = r_s / \sqrt{1/(7-1)} \fallingdotseq 1.19$ Therefore, do not reject H_0.

6. (10 pts : 5 pts each) Answer the next questions.

(a) We observe (X_i, Y_i), $i = 1, \ldots, n$. S_i and R_i are the corresponding ranks of X_i and Y_i, respectively. Prove the next equation with appropriate reasoning.

$$\sum_{i=1}^{n}(S_i - \frac{n+1}{2})(R_i - \frac{n+1}{2}) = \sum_{i=1}^{n}(S_i - \frac{n+1}{2})^2 - \frac{1}{2}\sum_{i=1}^{n}(S_i - R_i)^2$$

(b) When is a nonparametric method useful? What are the advantages of a nonparametric approach over a parametric one? Describe briefly.

(a) $$\begin{aligned}\sum_{i=1}^{n}(S_i - \frac{n+1}{2})(R_i - \frac{n+1}{2}) &= \sum_{i=1}^{n}(S_i - \frac{n+1}{2})(S_i - \frac{n+1}{2} - (S_i - R_i)) \\ &= \sum_{i=1}^{n}(S_i - \frac{n+1}{2})^2 - \sum_{i=1}^{n}(S_i - \frac{n+1}{2})(S_i - R_i) \\ &= \sum_{i=1}^{n}(S_i - \frac{n+1}{2})^2 - \sum_{i=1}^{n}S_i(S_i - R_i) \quad (\because \sum_{i=1}^{n}R_i = \sum_{i=1}^{n}S_i)\end{aligned}$$

Now, it is enough to show that $\sum_{i=1}^{n}S_i(S_i - R_i) = \frac{1}{2}\sum_{i=1}^{n}(S_i - R_i)^2$.

$$\begin{aligned}2\sum_{i=1}^{n}S_i(S_i - R_i) &= 2\sum_{i=1}^{n}S_i^2 - \sum_{i=1}^{n}2S_iR_i \\ &= \sum_{i=1}^{n}S_i^2 - \sum_{i=1}^{n}2S_iR_i + \sum_{i=1}^{n}R_i^2 \quad (\because \sum_{i=1}^{n}R_i^2 = \sum_{i=1}^{n}S_i^2) \\ &= \sum_{i=1}^{n}(S_i - R_i)^2\end{aligned}$$

(b) A nonparametric approach is robust to the distribution of a population. In many cases, a parametric approach relies on some strict assumptions like the normality. If those assumptions fail, the parametric approach is not reliable. We do not need to worry much about such things if we use a nonparametric method.

<Table 1 : Standard normal distribution>

$P(0 < Z < z)$ where $Z \sim N(0,1)$

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817

<Table 2 : Chi-square distribution>

χ^2_α satisfying $P(X > \chi^2_\alpha) = \alpha$ where $X \sim \chi^2(r)$

	0.990	0.975	0.950	0.050	0.025	0.010
1	0.0002	0.0010	0.0039	3.8415	5.0239	6.6349
2	0.0201	0.0506	0.1026	5.9915	7.3778	9.2103
3	0.1148	0.2158	0.3518	7.8147	9.3484	11.3449
4	0.2971	0.4844	0.7107	9.4877	11.1433	13.2767
5	0.5543	0.8312	1.1455	11.0705	12.8325	15.0863
6	0.8721	1.2373	1.6354	12.5916	14.4494	16.8119
7	1.2390	1.6899	2.1673	14.0671	16.0128	18.4753
8	1.6465	2.1797	2.7326	15.5073	17.5345	20.0902
9	2.0879	2.7004	3.3251	16.9190	19.0228	21.6660
10	2.5582	3.2470	3.9403	18.3070	20.4832	23.2093