2014.12.16.

Nonparametric Statistics & Function Estimation
Final 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) Answer to the questions. => Midterm exam

(a) (10 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

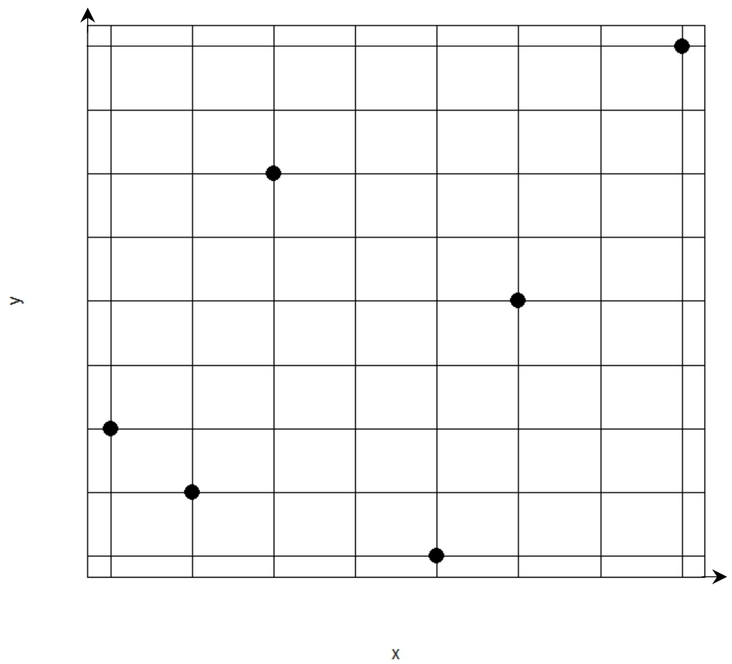
Test if the median length is larger than 3.7 based on the Sign statistic and the Wilcoxon signed rank statistic, respectively, at the significance level 0.05.

(b) (8 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

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.

(c) (7 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.

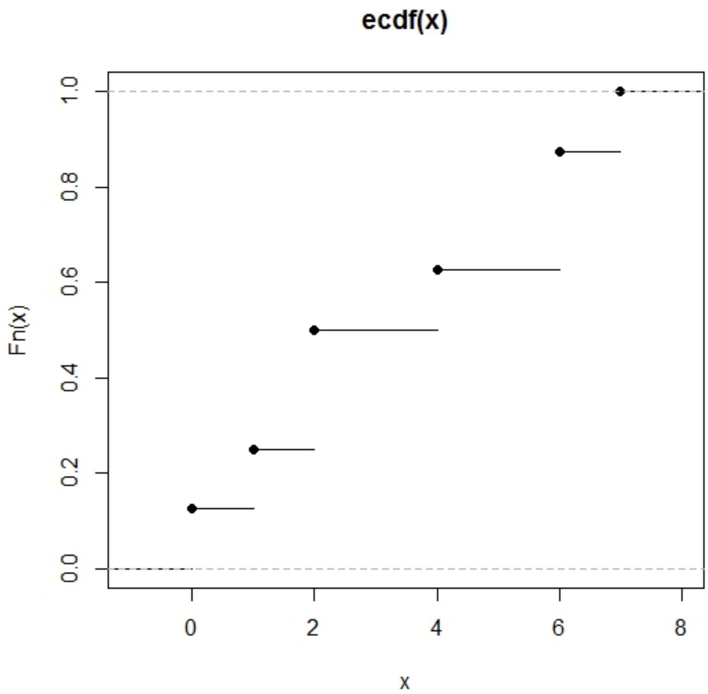


Calculate the Spearman correlation coefficient and the Kendall's tau correlation coefficient, respectively.

2. (15 pts : 5 pts each) Let $F(x)$ be a cumulative distribution function. We observe the following data from F.

1 0 6 2 4 2 6 7

(1) Calculate and draw the empirical distribution function from this sample.



(2) Find 95% pointwise confidence intervals for $F(x)$ at $x=1.5$ and $x=3.5$. Which one is broader?

$$x=1.5\ :\ \widehat{F}_n(1.5)\pm z_{0.025}\sqrt{\frac{\widehat{F}_n(1.5)(1-\widehat{F}_n(1.5))}{n}}=0.25\pm 0.30=(-0.05,0.55)$$

$$x=3.5\ :\ \widehat{F}_n(3.5)\pm z_{0.025}\sqrt{\frac{\widehat{F}_n(3.5)(1-\widehat{F}_n(3.5))}{n}}=0.25\pm 0.35=(-0.10,0.60)$$

The latter ($x=3.5$) is broader.

P.S. One may truncate the intervals with 0.

(3) Calculate and draw a 95% confidence band for F.

$$k=\sqrt{\frac{1}{2\cdot 8}\log\frac{2}{0.05}}=0.4802$$

Therefore, $\widehat{F}_n(x)\pm 0.4802$ is a confidence band. (truncation at 0 and 1 is needed.)

3. (10 pts) The following data is a sample from a continuous distribution. We want to test if the true distribution is $U(0,2)$.

0.42 0.84 0.12 1.95 1.91 1.75 1.79 1.99 1.08 0.12

Test it at the significance level 0.05. (Divide (0,2) into 3 sub-intervals with equal lengths)

category	(0,2/3)	(2/3,4/3)	(4/3,2)
observed frequency	3	2	5
expected frequency	10/3	10/3	10/3

$H_0 : F \sim U(0,2) \quad vs \quad \cancel{H_0}$

$\chi_0^2 = \frac{(3-10/3)^2}{10/3} + \frac{(2-10/3)^2}{10/3} + \frac{(5-10/3)^2}{10/3} = 1.4 < 5.99 = \chi_{0.05}^2(2)$

We do not reject H_0.

4. (20 pts : 10 pts each) Let $X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$ and $Corr(X_1, X_2) = \rho$. Suppose that we have a sample of size 100, $X_1, \ldots, X_{100}$, and the sample correlation coefficient r is 0.73.

(a) Describe how to estimate the variance of r by Bootstrapping.

Lecture note

(b) Let $\widehat{\sigma^2}$ be the Bootstrap estimate for the variance of r obtained in (a). Suppose that we want to find a 95% confidence interval for the population correlation coefficient ρ. Describe how to obtain the "normal interval", the "pivotal interval" and the "percentile interval" for ρ.

Lecture note

5. (30 pts : 6 pts each) Let (X, Y) be a pair of two random variables and

(0.5,1) (1,2) (1.5,3) (2,2) (2,3) (3,2) (3,4) (4,3)

be a random sample from it. We consider nonparametric estimation for the density function of X, $f_X(x)$ and the regression function $m(x) = E(Y|X=x)$. (Use the Epanechnikov kernel $K(u) = 0.75(1-u^2)I(|u| \leqq 1)$.)

(a) Calculate $\hat{f}_X(1.5)$ and $\hat{f}_X(3)$ with the bandwidth $h=1$.

$$\hat{f}_X(1.5) = \frac{1}{8 \cdot 1}(K(0.5) + K(0) + K(0.5) + K(0.5)) = 0.3047$$

$$\hat{f}_X(3) = \frac{1}{8 \cdot 1}(K(0) + K(0)) = 0.1875$$

(b) Calculate $\hat{f}_X(1.5)$ and $\hat{f}_X(3)$ with the bandwidth $h=1.5$.

$$\hat{f}_X(1.5) = \frac{1}{8 \cdot 1.5}(K(2/3) + K(1/3) + K(0) + K(1/3) + K(1/3)) = 0.2639$$

$$\hat{f}_X(3) = \frac{1}{8 \cdot 1.5}(K(2/3) + K(2/3) + K(0) + K(0) + K(2/3)) = 0.2292$$

(c) Based on the Nadaraya-Watson (local constant) estimation, calculate $\hat{m}(1.5)$ and $\hat{m}(3)$ with the bandwidth $h=1$.

$$\hat{m}(1.5) = \frac{K(0.5)2 + K(0)3 + K(0.5)2 + K(0.5)3}{K(0.5) + K(0) + K(0.5) + K(0.5)} = 2.5385$$

$$\hat{m}(3) = \frac{K(0)2 + K(0)4}{K(0) + K(0)} = 3$$

(d) Based on the Nadaraya-Watson (local constant) estimation, calculate $\hat{m}(1.5)$ and $\hat{m}(3)$ with the bandwidth $h=1.5$.

$$\hat{m}(1.5) = \frac{K(2/3)1 + K(1/3)2 + K(0)3 + K(1/3)2 + K(1/3)3}{K(2/3) + K(1/3) + K(0) + K(1/3) + K(1/3)} = 2.3158$$

$$\hat{m}(3) = \frac{K(2/3)2 + K(2/3)3 + K(0)2 + K(0)4 + K(2/3)3}{K(2/3) + K(2/3) + K(0) + K(0) + K(2/3)} = 2.8485$$

(e) Suppose that the bandwidth h is very large enough. What would happen on $\widehat{m}(x)$?

An estimated function $\widehat{m}(\cdot)$ becomes flat. (i.e., constant function).
Particularly, $\widehat{m}(\cdot) \to \overline{Y}$ as $h \to \infty$.
The local constant regression estimator acts like the global constant regression estimator $\overline{Y}$.

(Rough derivation :
Note that the NW estimator with bandwidth h is defined as

$$\widehat{m}(x) = \mathrm{argmin}_{\beta_0} \sum_{i=1}^{n} (Y_i - \beta_0)^2 K_h(X_i - x)$$

If h is large enough, then $K_h(X_i - x) \approx c \ \ \forall x$ for some constant $c > 0$.
Therefore,

$$\widehat{m}(x) \approx \mathrm{argmin}_{\beta_0} \sum_{i=1}^{n} (Y_i - \beta_0)^2 = \overline{Y})$$

6. (5 extra pts) It is well known that the Nadaraya-Watson (local constant) estimator suffers from the “boundary effect (problem)”. Explain about this phenomenon briefly and suggest how to improve the Nadaraya-Watson (local constant) estimator.

Lecture note

<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