University of Ruhuna

Bachelor of Science Special Degree
Level I (Semester II) Examination - December 2016

Subject: Mathematics
Course unit: MSP3292 (Applied Statistics-III)

Time: One and half ($1\frac{1}{2}$) Hours

Answer 03 questions only, including the question one.

---

1. Consider the general (one-dimensional) regression model, as used in nonparametric regression, of the form

$$Y_j = m(X_j) + \epsilon_j, \quad j = 1, 2, \ldots, N;$$

where $\epsilon_j, j = 1, 2, \ldots, N$ are independent and identically distributed (iid) errors with $E(\epsilon_j/X = x) = 0$ and $\sigma^2_{\epsilon_j}(x) = var(\epsilon_j/X = x) < \infty$ for all $x \in \mathbb{R}$.

a) State one advantage and one disadvantage of nonparametric regression techniques over standard parametric regression techniques.

b) (i) State the conditions that satisfied by a kernel function K used in nonparametric kernel smoothing.

(ii) Defining the scaled kernel K_h, $h > 0$ for a kernel K, verify that K_h satisfies the conditions described in part b(i) above.

(iii) Show that the function $K : \mathbb{R} \longrightarrow \mathbb{R}$ given, in the usual notation, by $K(u) = (1 - |u|)^+$ defines a kernel. Sketch the graph of $K_{\frac{1}{2}}(u)$.

c) Consider the Priestly-Chao (PC) Kernel estimate of $m : [0, 1] \longrightarrow \mathbb{R}$ with bandwidth $h > 0$, given by

$$\hat{m}(x, h) = \frac{1}{N}\sum_{j=1}^{N}(K_h(x - X_j)Y_j), \quad x \in [0, 1]$$

for a kernel K in equidistance deterministic design. Show that, in the usual notation,

$$var(\hat{m}(x, h)) \approx \frac{\sigma^2_\epsilon}{Nh}\int K(u)^2\, du.$$

d) State the formula of Nadaraja-Watson Kernel estimate $\hat{m}(x, h)$ when X_j values are random (that is, in stochastic design).

e) In k-nearest neighbour (k-NN) regression, the prediction of Y of the regression model (that is the estimate $\hat{m}(x)$ at a point x_0) is given by the average of the values of Y of k neighbours closest to x_0.

(i) Obtain the formula for k-NN estimate of $\hat{m}(x_0)$ described above.

(ii) What is the behaviour of the variance of $\hat{m}(x_0)$ as k increases ?

(iii) State a formula for k-NN estimate of $\hat{m}(x, h)$ using a kernel function K.

Continued.

2. For one way ANOVA model $y_{ij} = \mu + \alpha_i + \epsilon_{ij}$; where $i = 1, 2, ..., s, \quad j = 1, 2, ..., n_i$, with the usual notation, show that $SS_T = SS_{Tr} + SS_{Error}$.

A researcher predicts that students will learn more effectively in a constant background sound than backgrounds of random sound or no sound at all. He randomly selected 3 groups of students and asked all students to learn a paragraph of text for 30 minutes. Group 1 was set to learn in a constant background sound. The groups 2 and 3 were allowed to learn the same paragraph of text for 30 minutes in a background of random sound and in a background of no sound at all, respectively.

Finally a test was held to measure the learning ability of students and the scores (out of 300) were as follows:

Group 1	Group 2	Group 3
176.2	107.4	104.1
212.1	113.3	93.7
188.4	84	108.3
206	108.3	117.4
200	110	95.2
184.5	149.5	117.1
193	100.1	109.4

a) Test whether the variances of scores in three backgrounds are same at 0.05 significance level.

b) Construct the ANOVA table and, obtain the relevant conclusion at 0.05 significance level.

---

3. a) Derive the formula to estimate a missing value in a randomized block design in the usual notation.

An experimenter has discussed some instances in a market research. In each experiment of this series four merchandising practices A, B, C and D are compared in three areas by applying a randomized block design.

The percentages of improvements of sales per month are recorded as follows with two missing percentages x and y.

	A	B	C	D
area 1	31.5	37	26	30
area 2	32.5	y	17	16.5
area 3	30	25	x	23.5

Estimate the missing percentages x and y.

b) In a random sample of 400 persons with low income group, 500 persons with average income group and 400 persons with high income group, a shopper had asked whether he or she favours a particular detergent or not. The results are as follows:

Decision

Income group	Favour the detergent	Do not favour
Low	232	168
Average	260	240
High	197	203

Test whether the proportion of favouring the detergent is same in all three income groups at 0.05 significance level.

---

4. a) It is believed that the IQ level of Advanced level students is of mean 80 with the standard deviation 5. In a particular area the authorities wanted to test whether the average IQ level is higher than the believed value. A test was carried out in that area with 35 students. By how much their average IQ level should exceed 80 to make the difference significant at 0.01 significance level.

b) The following are the weight gained in (kilo grams) of two random samples of young turkeys fed two different diets keeping all the other conditions similar.

Diet 1	16.3	10.1	10.7	13.5	14.9	11.8	14.3	10.2
	12	14.7	23.6	15.1	14.5	18.4	13.2	14
Diet 2	21.3	23.8	15.4	19.6	12	13.9	18.8	19.2
	15.3	20.1	14.8	18.9	20.7	21.1	15.8	16.2

Using a suitable non parametric test at 0.05 significance level, test whether the mean weight gained for Diet 2 is higher than that of Diet 1.

c) Suppose 4 identical dice are tossed 20 times. The number '3' is considered as the success. Observed frequencies of the number of successes are as shown below:

Number of successes	0	1	2	3	4
Observed frequencies	3	8	5	2	2

Check whether the dice are fair at 0.05 significance level.