MASENO UNIVERSITY

UNIVERSITY EXAMINATIONS 2015/2016

FIRST YEAR FIRST SEMESTER EXAMINATIONS FOR THE DEGREE OF MASTER OF SCIENCE IN APPLIED STATISTICS

CITY CAMPUS

MAS 812: NON PARAMETRIC METHODS

Date: 20th February, 2015 Time: 2.00 - 5.00 pm

INSTRUCTIONS:

- Answer ANY THREE questions.

Question 1 (20 Marks)

a) Suppose we have a random of size n from a population with cumulative distribution function $F_0(x)$ and $S_n(x)$ defines its empirical distribution function. The statistic $D_n = \sup_x |S_n(x) - F_0(x)|$ is called the Kolmogorov-Smirnov test statistic. The directional deviations are defined as $D_n^+ = \sup_x [S_n(x) - F_0(x)]$ and $D_n^- = \sup_x [F_0(x) - S_n(x)]$. Provide a simple proof that D_n, D_n^+ and D_n^- are completely distribution free for any continuous and completely specified F_X **(10 Marks)**

b) Two types of corn (golden and green-striped) carry recessive genes. When these were crossed, a first generation was obtained, which was consistently normal (neither golden nor green-striped). When this generation was allowed to self-fertilize, four distinct types of plants were produced: normal, golden, green-striped, and golden-green-striped. In 1200 plants, this process produced the following distribution:

Normal: 670

Golden: 230

Green-striped: 238

Golden-green-striped: 62

A monk named Mendel wrote an article theorizing that in a second generation of such hybrids, the distribution of plant types should be in a 9:3:3:1 ratio. Are the above data consistent with the good monk's theory?

(10 Marks)

Question 2 (20 Marks)

The following data are $n = 15$ time intervals (recorded in minutes and decimal fractions of a minute) between successive customers arriving at a particular service point:

$$5.50, 2.38, 7.82, 1.51, 0.31,$$
$$8.78, 2.32, 1.55, 3.36, 0.09,$$
$$2.85, 5.39, 5.94, 2.92, 0.73$$

a) Using a Kolmogorov-Smirnov test, examine whether it is reasonable to assume that these data have arisen from an exponential distribution with mean equal to 4.0 **(10 Marks)**

b) Obtain a 90% confidence region for the true underlying distribution function. Present these in both tabular and graphical forms and augment the latter by also plotting both the null and empirical functions. **(8 Marks)**

c) Based on your results in (a) and (b) report your conclusions. **(2 Marks)**

Question 3 (20 Marks)

Let $X_1, \ldots, X_n$ be a random sample from some continuous distribution and suppose that we want to test the null hypothesis $H_0: M = M_0$ vs the alternative $H_1: M \neq M_0$, where M denotes the population median and M_0 its value under H_0. Consider the differences $D_i = X_i - M_0,\ i = 1, \ldots, n$. The sign test statistic is defined as K = the number of plus signs amongst the n $D_i^{'}s$. For the signed rank test an additional assumption that the underlying distribution of the data is symmetric is made. The $|D_i|$ are ranked from smallest to largest, and the ranks are assigned the original signs of the differences D_i. The signed rank test statistic is then defined as R^+ = the sum of the ranks of the positive differences.

a) Compute and tabulate the complete exact null distributions of K and of R^+ when the sample size is $n = 5$.

b) Five randomly selected students took a particular test twice (before and after a training course) and obtained the following scores:

Student	First attempt	Second attempt
1	32	38
2	39	47
3	27	31
4	37	34
5	28	35

Is the population median score for second attempts greater than that for first attempt? (Use both the sign test and Wilcoxon signed-rank test and use the exact null distributions computed in part (a) above to calculate p-values for the observed values of the test statistics). Clearly state your conclusions.

Question 4 (20 Marks)

a) Let $X_1, \dots, X_n$ be a random sample from a continuous distribution with cumulative density function F_X.

i) Determine when the variance of the empirical distribution function $S_n(x)$ a maximum and what is the maximum value? **(6 Marks)**

ii) State what happens to the variance as $n \to \infty$ **(1 Mark)**

b) Let $X_1, \dots, X_n$ be a random sample from a continuous distribution with cumulative density function F_X

i) Show that

$$Cov(S_n(s), S_n(t)) = \frac{1}{n}[F_X(u) - F_X(s)F_X(t)]$$

where $u = \min(s,\ t)$ for $s \neq t$. **(10 Marks)**

ii) Are $S_n(s)$ and $S_n(t)$ negatively or positively correlated? **(3 Marks)**

Hint: Express the empirical distribution function, $S_n(x)$ as

$$S_n(x) = \frac{1}{n}\sum_{i=1}^{n} \delta_i(x)$$

where

$$\delta_i(x) = \begin{cases} 1, & X_i \leq x \\ 0, & X_i > x \end{cases}$$

Question 5 (20 Marks)

a) The following data, presented in numerical order, are a random sample of $n = 25$ waiting times experienced by people phoning a particular helpline.

0.61,	0.86,	0.92,	1.14,	1.70
1.71,	2.24,	3.03,	3.25,	3.34
3.87,	4.69,	5.26,	5.44,	5.96
6.18,	6.64,	10.30,	12.10,	12.65
17.07,	17.78,	20.87,	28.58,	29.58

i) Find a point estimate and 90% confidence interval for the population upper quartile **(8 Marks)**

ii) Based on the results of part (i), are the data consistent with having come from a distribution with an upper quartile equal to 20.0? State your reason **(2 Marks)**

b) In a psychological experiment, the research question of interest is whether a rat "learned" its way through a maze during 64 trials. Suppose the timeordered observations on number of correct choices by the rat on each trial are as follows:

0, 1, 2, 1, 1, 2, 3, 2, 2, 2, 1, 1, 3, 2, 1, 2,
1, 2, 2, 1, 1, 2, 2, 1, 4, 3, 1, 2, 2, 1, 2, 2,
2, 2, 3, 2, 2, 3, 4, 3, 2, 3, 3, 2, 3, 3, 2, 3,
3, 2, 3, 4, 3, 3, 4, 2, 3, 3, 4, 3, 4, 4, 4, 4

Test these data for randomness against the alternative of a tendency to cluster, using the dichotomizing criterion that 0, 1, or 2 correct choices indicate no learning, while 3 or 4 correct indicate some learning. **(10 Marks)**