Applied Statistics - Spring 2015
Final Exam

June 1st, 2015

INSTRUCTIONS:

- You can bring slides, lecture notes and exercises/solutions used in the course. Other materials are NOT allowed.
- You can use a calculator. Other electronic devices are NOT allowed.
- The exam consists of 6 exercises.
- Time: 3 hours.
- Indicate your name, student number and the university on EACH answer sheet.
- Answers should be explained and clearly formulated. You should clearly and concisely indicate your reasoning and show all relevant work.

1. (10 points) Let (X, Y) be a bivariate random variable and

$$F(x, y) = \mathbb{P}(X \leq x, Y \leq y).$$

Based on a random sample $\{(X_1, Y_1), \ldots, (X_n, Y_n))\}$ from F,

(a) find an estimator of $F(x_0, y_0)$, for a given (x_0, y_0);

(b) derive an asymptotic $100(1-\alpha)\%$ confidence interval for $F(x_0, y_0)$.

2. (20 points) The Cramér-Von-Mises test statistics for a simple GoF test (that is $H_0 : F = F_0$. v.s. $H_1 : F \neq F_0$.) is given by

$$C_n = n \int_{-\infty}^{\infty} (\hat{F}_n(x) - F_0(x))^2 dF_0(x),$$

where $\hat{F}_n$ is the empirical distribution function of the random sample $\{X_i\}_{i=1}^n$ and F_0 is continuous. Under H_0, $C_n \xrightarrow{d} Y$, where Y is a continuous random variable. Let $\mathbb{P}(Y > y_{1-\alpha}) = \alpha$. A test with an asymptotic significance level α rejects H_0 if $C_n > y_{1-\alpha}$.

Suppose that the data $\{X_i\}_{i=1}^n$ is from a continuous distribution $F_1 \neq F_0$.

(a) Let $D_n = \sup_{t \in [0,1]} \left| \frac{1}{n} \sum_{i=1}^n I\{U_i \leq t\} - t \right|$, where U_i are i.i.d. from a uniform distribution on $[0, 1]$. Show that

$$\sup_{-\infty < x < \infty} \left| \hat{F}_n(x) - F_1(x) \right| \stackrel{d}{=} D_n.$$

(b) Assume that F_0 has positive density in $\mathbb{R}$. Show that

$$\lim_{n \to \infty} \mathbb{P}(C_n > y_{1-\alpha}) = 1.$$

Hint: $\sqrt{n} D_n = O_p(1)$, that is, for any small $\epsilon \in (0, 1]$, there exists $c < \infty$ such that when n is large enough, $\mathbb{P}(\sqrt{n} D_n > c) < \epsilon$.

3. (12 points) A study observed 15 nursing home patients with dementia. The number of aggressive behavior incidents was recorded each day for 12 weeks. A day is called a "*moon day*" if it is the day before, during, or after a full moon. Table 1 provides the average number of aggressive incidents during moon days and other days of each subject. Based on this data, apply the permutation test to answer the following question: Can full moon influence the behavior of dementia patients?

Patient	Moon days	Other days	Patient	Moon days	Other days
1	3.33	0.27	9	6	1.59
2	3.67	0.59	10	4.33	0.6
3	2.67	0.32	11	3.33	0.65
4	3.33	0.19	12	0.67	0.69
5	3.33	1.26	13	1.33	1.26
6	3.67	0.11	14	0.33	0.23
7	4.67	0.3	15	2	0.38
8	2.67	0.4			

Table 1: Aggressive behaviors of dementia patients

(a) Formulate the null and alternative hypotheses. Define an appropriate test statistics, T. When do you reject the null hypothesis, for large or small value of T?

(b) What is the total number of possible randomizations under H_0?

9 (c) Derive a scheme to obtain the p-value.

4. (12 points) Let $X_1, \ldots, X_n$ be a random sample from an unknown distribution F. Denote the empirical distribution function by $\hat{F}_n$. Let $\theta = h(F)$ be the quantity of interest, which is estimated by $\hat{\theta}_n = h(\hat{F}_n) = T(X_1, \ldots, X_n)$. Let $\hat{\theta}^*_{n1}, \ldots, \hat{\theta}^*_{nB}$ be a bootstrap sample, where $\hat{\theta}^*_{ni} = T(X^*_{i1}, \ldots, X^*_{in})$ and $\{X^*_{i1}, \ldots, X^*_{in}\}$ are i.i.d. from $\hat{F}_n$.

9 Suppose that there exists an unknown monotone transformation l such that $l(\hat{\theta}_n) - l(\theta)$ has a symmetric distribution around 0. Show that the $(1-\alpha)100\%$ percentile interval of θ is given by

$$[\hat{\theta}^*_{(\frac{\alpha B}{2})}, \hat{\theta}^*_{((1-\frac{\alpha}{2})B)}],$$

where $\hat{\theta}^*_{(i)}$ is the i-th order statistisc of the bootstrap sample and B is sufficiently large.

5. (16 points) Let $X_1, \ldots, X_n$ be a random sample from a distribution with continuous density f. For a given bandwidth h, the naive density estimator is given by

$$\hat{f}_n(x) = \frac{\hat{F}_n(x+h) - \hat{F}_n(x-h)}{2h},$$

where $\hat{F}_n$ is the empirical distribution function.

(a) Show that $\hat{f}_n(x)$ is a probability density.

(b) Show that if $h = h_n \to 0$ and $nh_n \to \infty$ as $n \to \infty$, then

$$\hat{f}_n(x) \xrightarrow{p} f(x), \qquad \text{as } n \to \infty.$$

6. (10 points) Consider the example of New Jersey Pick-It Lottery. The data set consists of the winning numbers and their payoffs for the first 245 drawings of the lottery, as represented by the circles in Figure 1.

(a) A nonparametric regression estimator is implemented for this data. What is computed in the function **hat_r**? Write down the formula.

```
# The vector X stores observations of winning number.
# The vector Y stores observations of payoff.

h=20;
L=function(x,i){k=dunif((x-X)/h,-1,1);  li=k[i]/sum(k);  return(li);}

hat_r=function(x)
{w=numeric(n);  for(i in 1:n) {w[i]=L(x,i)}  r_x=sum(Y*w);  return(r_x);}
```

(b) The black curve in Figure 1 plots the function **hat_r**. Since it is not smooth, a student suggests to increase the value of h, the smoothing parameter. Do you agree? If yes, give your reason. If not, what is your solution to obtain a smooth estimate?

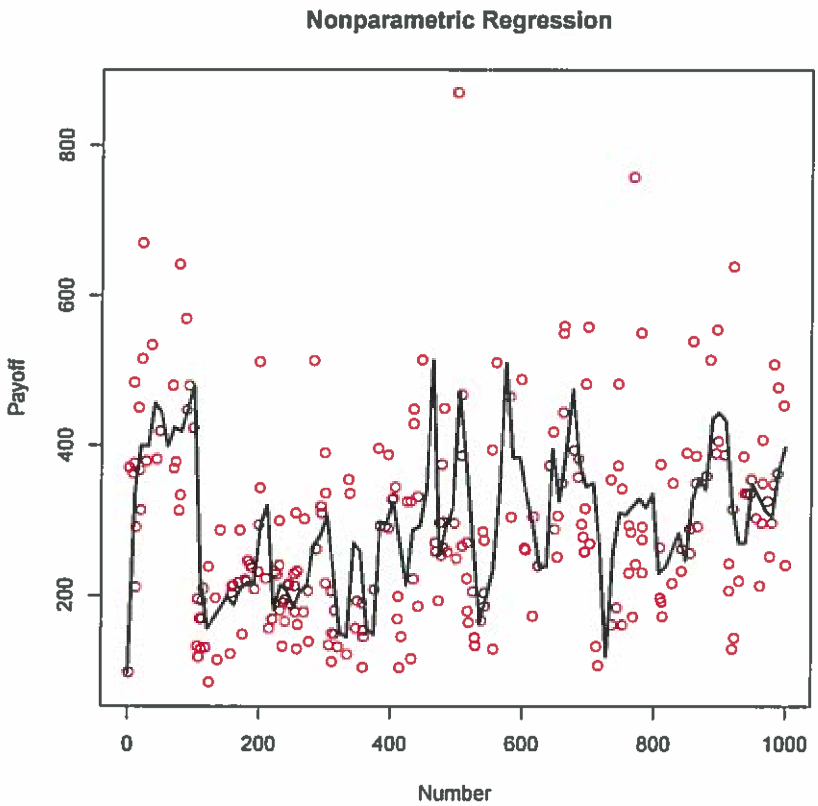

Figure 1: Nonparametric regression