Homework 1
STAT W4413: Nonparametric Statistics

DUE: Tuesday, February 2, 12:00 noon

(1) Please sign your home work with your name and UNI number.

(2) Homework must be submitted into the Statistics Homework Boxes room 904 on the 9th floor of SSW building.

(3) Homework is due Tuesday, February 2, 12:00 noon.

(4) No late homework, under any circumstances, will be accepted.

(5) At the end of semester, one of your lowest homework scores will be dropped before the final grade is calculated.

(6) Your submitted solutions should consist of (i) the hand written (or printout) of the results with all the details, (ii) printout of the relevant figures, and (iii) the printout of the source code.

1. Let $X_1, X_2, \dots, X_n$ be independently drawn from a differentiable CDF $F(x)$. Let $a \in \mathbb{R}$ be a fixed number. In the first lecture we mentioned

$$Z = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(X_i \leq a),$$

as an estimate of $F(a)$. In this problem we want to analyze some properties of this estimator. Calculate $E(Z)$ and $Var(Z)$. Characterize the probability mass function of Z.

2. Let X be a random variable with a continuous and strictly increasing CDF $F_X(x)$. Define a new random variable $W = F_X(X)$. Characterize the distribution of W?

3. Fitting linear/polynomial models. In this problem we consider a data set (X_1, Y_1), $(X_2, Y_2), \dots, (X_n, Y_n)$ that are assumed to be independent and identically distributed. We believe a model of the form $Y_i = g(X_i) + Z_i$ is true and we would like to characterize the function g. Suppose that Z_i's are independently drawn from $N(0, \sigma^2)$. Given X_i the probability density function of Y_i's are given by

$$f(Y_1, Y_2, \dots Y_n) = \frac{1}{(2\pi\sigma^2)^{n/2}} \mathrm{e}^{\frac{-\sum_{i=1}^{n}(Y_i - g(X_i))^2}{2\sigma^2}}$$

This is known as the likelihood of $Y_1, Y_2, \dots Y_n$. Usually our goal is to restrict the form of g and find the function g that maximizes the likelihood function.

(a) Show that maximizing the likelihood function with respect to g is equivalent to minimizing

$$\sum_{i=1}^{n}(Y_i - g(X_i))^2$$

with respect to g. Here σ is assumed to be known.

(b) Suppose that g has a linear form $g(X) = \alpha_0 + \alpha_1 X$. Write down two linear equations that the optimal values of α_0 and α_1 should satisfy.

(c) Suppose that g has a linear form $g(X) = \sum_{j=0}^{p-1} \alpha_j X^j$. Write down p linear equations that the optimal values of $\alpha_0, \dots, \alpha_{p-1}$ should satisfy (you can use matrix notation).

(d) Download the dataset provided on the courseworks (dataHW1.rtf). Solve the linear equations you derived in part (c) for $p = 2, 4, 10, 15$. Plot both the dataset and your estimated functions. Based on visual inspection, which value of p do you think provides a "good" fit. Why?

4. Let $X_1, X_2, \dots, X_5 \overset{iid}{\sim} N(\theta, 1)$. We would like to test for the null hypothesis $H_0 : \theta = 0$ versus $H_1 : \theta = 2$.We have the following test: If $|\sum_{i=1}^{5} X_i| > k$ reject H_0, otherwise accept H_0. Characterize the probability of type I and type II errors in terms of the parameter k. Plot probability of type II error versus the probability of Type I error as k varies in $[0, 3]$. Repeat this for testing the null hypothesis $H_0 : \theta = 0$ versus $H_1 : \theta = -1.8$ and compare the two curves.

5. As in the previous problem let $X_1, X_2, \dots, X_5 \overset{iid}{\sim} N(\theta, 1)$. However, we are now interested in testing $H_0 : \theta = 0$ versus $H_1 : \theta \neq 0$. Again we use a similar test: If $|\sum_{i=1}^{5} X_i| > k$ reject H_0, otherwise accept H_0. Suppose that we would like the significance level of the test to be 0.01. Calculate the value of k. Now write down the formula for the power of this test. Plot the power of test for $\theta \in [-1, 1]$ using a software of your choice.