Applied Statistics - Spring 2013
Final Exam

June 3rd, 2013

INSTRUCTIONS:

- This is a OPEN NOTES exam. You can bring all the materials distributed in class, plus the book by L. Wasserman we used.
- In addition, you can bring any MANUSCRIPT PERSONAL NOTES and homework assignments (other books, or copies of books are not allowed).
- You can use a calculator. Cellphone, notebooks or similar devices are not allowed.
- You have 180 minutes (3 hours) to complete the exam.
- There are 4 problems, each graded from 0 to 25 points (totaling 100 points). There are 6 pages in total, including this one.
- The problems are not necessarily in order of difficulty. I recommend that you quickly read through all problems first, then do the problems in whatever order suits you best.
- The exam is to be done INDIVIDUALLY. Therefore discussion with your fellow colleagues is not allowed!
- A correct answer does not guarantee full credit, and a wrong answer does not guarantee loss of credit. You should clearly and concisely indicate your reasoning and **show all relevant work.** Your grade on each problem will be based on my best assessment of your level of understanding as reflected by what you have written. **JUSTIFY** your answers and be **CRITICAL** of your results.
- Please be organized in your write-up – I can't grade what I can't decipher!
- Remember to **IDENTIFY YOUR HANDOUT.**

P.I: A new procedure striving to reduce call duration at a customer service call center is being considered. The goal of the new procedure is to reduce the amount of time it takes reroute the incoming calls to the proper department. Suppose we want to test if the new procedure is more effective than the standard one. For this purpose five calls were answered using a *randomized* assignment of procedure: two callers were assigned the standard procedure and three callers were assigned the new procedure. Denote the rerouting times with the standard procedure by x_1 and x_2, and the rerouting times with the new procedure by y_1, y_2, and y_3. The collected call-duration data (in seconds) is summarized in the following table.

Standard Procedure	New Procedure
x_1 95	28 y_1
x_2 145	47 y_2
	135 y_3

We want to test the null hypothesis:

H_0 : both procedures have the same mean rerouting time,

against:

H_1 : the new procedure reduces the mean rerouting time .

Therefore it is appropriate to consider the test statistic

$$T = \frac{1}{2}(x_1 + x_2) - \frac{1}{3}(y_1 + y_2 + y_3) \ .$$

$\bar{x} - \bar{y}$

For the data collected the value of this statistic is simply $t_0 = (95+145)/2-(28+47+135)/3 = 50$. In what follows we will test the hypotheses using a randomization test.

(a) Show that the statistic $S = x_1 + x_2$ is equivalent to T, meaning these yield equivalent randomization tests.

(b) Compute the (exact) distribution of S under the null hypothesis (note that there are at most $\binom{5}{2}$ possible values for S).

(c) Compute the p-value for testing the null-hypothesis against the alternative. Can you reject the null-hypothesis with 95% confidence?

P.II: *Testing the symmetry of a distribution:* Let X be a random variable with an arbitrary cummulative distribution function F. We say that the distribution F is symmetric around zero if and only if $-X$ also has distribution F. That is

$$\forall t \in \mathbb{R} \quad \mathbb{P}(X \leq t) = \mathbb{P}(-X \leq t) \ .$$

Suppose we have i.i.d. samples $X_1, \ldots, X_n$ from a *continuous* distribution F, and want to test the null hypothesis

$$H_0 : \ F \text{ is symmetric around zero} \ ,$$

against the alternative

$$H_1 : \ F \text{ is } \textit{not} \text{ symmetric around zero} \ .$$

We can construct a simple statistic for this test based on the empirical cumulative distribution function. In particular define

$$S_n = \sup_{t \in \mathbb{R}} \left| \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{X_i \leq t\} \ - \ \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{-X_i \leq t\} \right| \ .$$

To make things simpler you may assume F is strictly monotone.

(a) Show that, under the null hypothesis, this statistic is distribution free. That is, if F satisfies the symmetry property then the distribution of S_n does not depend on the shape of F (**note:** to answer this you do not need to derive an analytic form for the distribution of S_n).

(b) Show that a test based on this statistic is consistent under the alternative H_1. In particular you must show that, under H_1, there is an $\epsilon > 0$ such that

$$\mathbb{P}_{H_1}(S_n > \epsilon) \to 1 \ ,$$

as $n \to \infty$, but under the H_0, for all $\epsilon > 0$, $\mathbb{P}_{H_0}(S_n > \epsilon) \to 0$ as $n \to \infty$. (**hint:** the Glivenko-Cantelli Theorem might be handy here).

(c) Suppose you observe data $x_1, \ldots, x_n$ and want to use this test. Explain how you can approximate the p-value for this test. That is, give a few lines of pseudo code that upon execution would result in an approximation of the p-value.

(d) Sometimes we just want to test if a distribution is symmetric around a certain unknown point. A distribution F is symmetric around the point t_0 if for X with distribution F

$$\forall t \in \mathbb{R} \quad \Pr(X - t_0 \leq t) = \Pr(-(X - t_0) \leq t) \ .$$

The problem is that generally we don't know t_0. Suggest a way to modify the above statistic so that we can test if F is symmetric around some unknown point. Argue if the modification you propose still retains the distribution-free property shown in (a).

P.III: Consider the usual regression setting. Let $(x_1, Y_1), \ldots, (x_n, Y_n)$ be n data points and suppose that

$$Y_i = r(x_i) + \epsilon_i \ ,$$

where ϵ_i's are i.i.d. random variables such that $\mathbb{E}[\epsilon_i] = 0$ and $\mathbb{E}[\epsilon_i^2] = \sigma^2$. Furthermore suppose $x_i = i/n$.

Consider the Nadaraya-Watson estimator $\hat{r}_n^{(h)}$, with bandwidth h and kernel

$$K(x) = c\frac{|x|(1-|x|)}{1+x^2}1\{-1 \leq x \leq 1\} \ ,$$

where $c = \frac{1}{(\pi/2+\log 2)-2}$.

(a) Write the formula of the weights $\ell(x) = (\ell_1(x), \ldots, \ell_n(x))$ so that $\hat{r}_n^{(h)}(x) = \sum_{j=1}^n \ell_j(x) Y_j$ (you should write the formula for an arbitrary kernel K).

(b) Suppose the bandwidth is $h = 2/n$. Explicitly write the smoothing matrix L (recall that $L_{ij} = \ell_j(x_i)$).

(c) Compute the effective degrees of freedom for the smoothing matrix L in (b). That is, compute $\nu = \text{trace}(L)$.

(d) The answer to the previous question might seem puzzling to you, as the number of effective degrees of freedom is rather small. However, as we seen when studying the quality of an estimate of the variance σ^2, the *fitted degrees of freedom* play a prominent role. Compute the fitted degrees of freedom when $h = 2/n$, that is

$$\tilde{\nu} = \text{trace}(L^T L) = \sum_{i=1}^n \|\ell(x_i)\|^2 \ .$$

Does this value seem somewhat more reasonable/intuitive?

(e) Define

$$R(h) = \frac{1}{n}\sum_{i=1}^n (Y_i - \hat{r}_n^{(h)}(x_i))^2 \ ,$$

that is, the average of the squared residuals. Compute the expected value of $R(h)$.

(f) Suppose you take $\hat{h} = \arg\min_h R(h)$ and compute your final regression estimator as

$$\hat{r}_n(x) = r_n^{(\hat{h})}(x_i) \ .$$

Is this a reasonable approach? Carefully justify your answer.

Hint: The answers to c) and d) should involve only n.

P.IV: In gene expression studies one often wants to identify genes that are up-regulated. In the following question you will consider a simplified model closely related to that problem.

Let S^* be an unknown subset of $\{1, \ldots, n\}$ (the up-regulated genes), and assume you make measurements of the form

$$Y_i = \mu \mathbf{1}\{i \in S^*\} + \sigma \epsilon_i \ , \quad i = 1, \ldots, n \ ,$$

where $\mu > 0$ and $\sigma > 0$ are known, and $\{\epsilon_i\}_{i=1}^n$ are i.i.d. standard normal random variables. Our goal is to estimate S^* from these measurements, in other words, to construct an estimator $\hat{S}(Y)$ of S^*. A reasonable performance metric to consider is the *symmetric set difference* size, defined as

$$d(\hat{S}, S^*) = |S^* \Delta \hat{S}| = |S^* \setminus \hat{S}| \ + \ |\hat{S} \setminus S^*| \ .$$

In words, this is the number of entries of S^* we failed to identify plus the number of entries we erroneously included in $\hat{S}$. Suppose we have reason to believe that are exactly s entries in S^*. Let $\mathcal{C}_s$ be the class of subsets of $\{1, \ldots, n\}$ with exactly s entries.

(a) A first idea that comes to mind is to use maximum likelihood estimation. Show that the maximum likelihood estimator of S is given by

$$\hat{S} = \arg\min_{S \in \mathcal{C}_s} \sum_{i=1}^n (Y_i - \mu \mathbf{1}\{i \in S\})^2 \ .$$

(b) Show that

$$\frac{1}{n} \sum_{i=1}^n (\mu \mathbf{1}\{i \in \hat{S}\} - \mu \mathbf{1}\{i \in S^*\})^2 = \frac{\mu^2}{n} d(\hat{S}, S^*) \ .$$

(c) Using the result of b) and the MPLE Theorem from class (reproduced in the exam for your convenience) show that

$$\mathbb{E}[d(\hat{S}, S^*)] \leq \frac{8 s \sigma^2 \log n}{\mu^2} \ .$$

(d) Now suppose you wanted to remove the assumption that S^* has exactly s entries. Propose a suitable modification of the above methodology, and show that the new estimator (denoted by $\tilde{S}$) achieves the following bound

$$\mathbb{E}[d(\tilde{S}, S^*)] \leq \frac{8 \sigma^2 s (\log 2 + \log n)}{\mu^2} \ .$$

Hint: Note that $\binom{n}{k} \leq n^k$.

Theorem 1. *Let*

$$Y_i = r^*(x_i) + \epsilon_i \quad , \epsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2) \ ,$$

where x_i are deterministic. Let $\mathcal{R}$ be a class of models such that there is a map $c : \mathcal{R} \to [0, \infty]$ satisfying

$$\sum_{r \in \mathbb{R}} 2^{-c(r)} \leq 1 \ .$$

Define

$$\hat{r}_n = \arg\min_{r \in \mathbb{R}} \left\{ \frac{1}{n} \sum_{i=1}^{n} (Y_i - r(x_i))^2 + \frac{4\sigma^2 c(r) \log 2}{n} \right\} \ .$$

Then

$$\frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left[(\hat{r}_n(x_i) - r^*(x_i))^2 \right] \leq 2 \min_{r \in \mathbb{R}} \left\{ \frac{1}{n} \sum_{i=1}^{n} (r(x_i) - r^*(x_i))^2 + \frac{4\sigma^2 c(r) \log 2}{n} \right\} \ .$$