Econ. 240A, Spring 2000 D. McFadden

PROBLEM SET 4 (Estimation)

(Due Wed., Feb. 23, with discussion in section on March 1,3)

1. You have a random sample i = 1,...,n of observations x_i drawn from a normal distribution with unknown mean μ and known variance 1. Your prior density p(μ) for μ is normal with mean zero and variance 1/k, where k is a number you know. You must choose an estimate T of μ. You have a quadratic loss function $C(T,\mu) = (T - \mu)^2$. (a) What is the density of the observations, or likelihood, $f(\mathbf{x},\mu)$? (b) What is the posterior density $p(\mu|\mathbf{x})$? (c) What is the Bayes risk $R(T(\mathbf{x})|\mathbf{x})$? (d) What is the optimal estimator $T^*(\mathbf{x})$ that minimizes Bayes risk?

2. A simple random sample with n observations is drawn from an exponential distribution with density $\lambda\cdot\exp(-\lambda x)$. (a) What is the likelihood function $f(\mathbf{x},\lambda)$? (b) What is the maximum likelihood estimator for λ? (c) If you have a prior density $\alpha\cdot\exp(-\alpha\lambda)$ for λ, where α is a constant you know, what is the posterior density of λ? What is the optimal estimator that minimizes Bayes risk if you have a quadratic loss function. (d) Using characteristic functions, show that the exact distribution of $W = 2n\lambda\bar{x}$, where $\bar{x}$ is the sample mean, is chi-square with 2n degrees of freedom. Use this to find the exact sampling distribution of the maximum likelihood estimator.

3. If h(t) is a convex function and $t = T(\mathbf{x})$ is a statistic, then Jensen's inequality says that $\mathbf{E}h(T) \geq h(\mathbf{E}T)$, with the inequality strict when h is not linear over the support of T. When h is a concave function, $\mathbf{E}h(T) \leq h(\mathbf{E}T)$. If T is an unbiased estimator of a parameter σ^2, what can you say about $T^{1/2}$ as an estimator of σ and exp(T) as an estimator of $\exp(\sigma^2)$?

4. A simple random sample i = 1,...,n is drawn from a binomial distribution b(K,1,p); i.e., $K = k_1 + ... + k_n$ is the count of the number of times an event occurs in n independent trials, where $k_i = 1$ with (unknown) probability p and $k_i = 0$ with probability 1-p for I = 1,...,n. Which of the following statistics are <u>sufficient</u> for the parameter p: a. $(k_1,...,k_n)$; b. $(k_1^2,[k_2+...+k_n]^2)$; c. $f \equiv K/n$; d. $(f,[k_1^2+...+k_n^2])$; e. $[k_1^2+...+k_n^2]$?

5. You want to estimate mean consumption from a random sample of households i = 1,...,n. You have two alternative income measures, C_{1i} which includes the value of in-kind transfers and C_{2i} which excludes these transfers. You believe that the sample mean m_1 of C_{1i} will overstate economic consumption because in-kind transfers are not fully fungible, but the sample mean m_2 of C_{2i} will understate economic consumption because these transfers do have value. After some investigation, you conclude that $0.7\cdot m_1 + 0.3\cdot m_2$ is an unbiased estimator of mean economic consumption; i.e., an

in-kind transfer that costs a dollar has a value of 70 cents to the consumer because it is not fully fungable. Your friend Dufus proposes instead the following estimator: Draw a random number between 0 and 1, report the estimate m_2 if this random number is less than 0.3, and report the estimate m_1 otherwise. Is the Dufus estimator unbiased? Is it as satisfactory as your estimator? (Hint: Does it pass the test of ancillarity?)

6. Suppose T(**x**) is an unbiased estimator of a parameter θ, and that T has a finite variance. Show that T is *inadmissible* by demonstrating that $(1-\lambda)\cdot T(\mathbf{x}) + \lambda\cdot 17$ for λ some small positive constant has a smaller mean square error. (This is called a *Stein shrinkage estimator*. The constant 17 is obviously immaterial, zero is often used.)

7. In Problem Set 2, you investigated some of the features of the data set nyse.txt, located in the class data area, which contains 7806 observations from January 2, 1968 through December 31, 1998 on stock prices. The file contains columns for the date in yymmdd format (DAT), the daily return on the New York Stock Exchange, including distributions (RNYSE), the Standard & Poor Stock Price Index (SP500), and the daily return on U.S. Treasury 90-day bills from the secondary market (RTB90). Define an additional variable GOOD which is one on days when RNYSE exceeds RTB90, and zero otherwise. The variable GOOD identifies the days on which an overnight buyer of the NYSE portfolio makes money. For the purposes of this exercise, make the maintained hypothesis that these observations are independent, identically distributed draws from an underlying population; i.e., suspend concerns about dependence in the observations across successive days or secular trends in their distribution.

a. Estimate **E**(GOOD). Describe the finite sample distribution of your estimator, and estimate its sample variance. Use a normal approximation to the finite sample distribution (i.e., match the mean and variance of the exact distribution) to estimate a 90 percent confidence bound.

b. Estimate the population expectation μ of RNYSE employing the sample mean, and alternately the sample median. To obtain estimates of the distribution of these estimators, employ the following procedure, called the *bootstrap*. From the given sample, draw a *resample* of the same size *with replacement*. (To do this, draw 7806 random integers k = floor(1+7806*u), where the u are uniform (0,1) random numbers. Then take observation k of RNYSE for each random integer draw; some observations will be repeated and others will be omitted. Record the resample mean and median. Repeat this process 100 times, and then estimate the mean and variance of the 100 bootstrap resample means and medians. Compare the bootstrap estimate of the precision of the sample mean estimator with what you would expect if RNYSE were normally distributed. Do confidence statements based on an assumption of normality appear to be justified? Compare the bootstrap estimates of the precision of the mean and median estimators of μ. Does choice of the sample mean rather than the sample median to estimate μ appear to be justified?