Statistics 582, Problem Set 1

Wellner; 1/7/98

Reading: Lehmann, TPE, Chapter 6, sections 6.1 and 6.2. Chapter 4, sections 2 - 5.
Due: Wednesday, January 14, 1998.

1. This is a continuation of problem 10.4:
   A. For the data in problem 10.4, compute the maximum likelihood estimator of $q(\theta) \equiv F_\theta^{-1}(1/2)$ and a 95% confidence interval.
   B. For the data in problem 10.4, compute the nonparametric estimator of $F_\theta^{-1}(1/2)$ and a 95% confidence interval. Compare with the confidence interval computed in A.

2. (Right censored data). Suppose that $X, X_1, \ldots, X_n$ are i.i.d. survival times with unknown distribution function F, that $Y, Y_1, \ldots, Y_n$ are i.i.d. censoring times with unknown distribution function G, assumed to be independent of the X_i's, and that we can observe only the iid pairs $(Z_1, \delta_1), \ldots, (Z_n, \delta_n)$ where $Z_i \equiv X_i \wedge Y_i$ and $\delta_i \equiv 1_{[X_i \le Y_i]}$; also let $Z \equiv X \wedge Y$ and $\delta = 1_{[X \le Y]}$.
   A. Show that the joint distribution of (Z, δ) is given by

   $$P(Z \le z, \delta = 1) = \int_{(0,z]} (1 - G(x-))dF(x)$$

   where $G(x-) \equiv \lim_{y \uparrow x} G(y)$, and

   $$P(Z \le z, \delta = 0) = \int_{(0,z]} (1 - F(y))dG(y).$$

   Furthermore, show that the survival function $1 - H(z) = P(Z > z)$ is given by $1 - H(z) = (1 - F(z))(1 - G(z))$.
   B. Suppose we assume a parametric family of densities $\{f_\theta : \theta \in \Theta\}$ for the distribution of the survival time X. Find general expressions for the density of the observations (Z, δ) (assuming that the censoring distribution G also has density g which does not depend on θ), and for the scores for θ based on observation of one (Z, δ) pair.
   C. Suppose that the parametric family $\{f_\theta : \theta \in \Theta\}$ in part B is the Weibull family with survival functions $\overline{F}_\theta(x) = \exp(-(x/\alpha)^\beta)$ with $\theta = (\alpha, \beta) \in R^{+2}$. Derive an asymptotic likelihood based test of the hypothesis that the X_i are Rayleigh (i.e. $\beta = 2$), versus the alternative that they are Weibull distributed with $\beta \ne 2$.

3. Suppose that $(X,Y),(X_1,Y_1),\ldots,(X_n,Y_n)$ are i.i.d. with bivariate normal distribution $N_2(\mu,\Sigma)$ where $\mu \in R^2$ and

$$\Sigma = \begin{pmatrix} \sigma^2 & \sigma\tau\rho \\ \sigma\tau\rho & \tau^2 \end{pmatrix}$$

where $\sigma^2 > 0$, $\tau^2 > 0$, and $\rho \in (-1,1)$.
A. If we assume that $\mu_1 = \mu_2 \equiv \theta$ and Σ is known, what is the MLE of θ?
B. If we assume that μ is known and $\sigma^2 = \tau^2 \equiv \theta$, what is the MLE of θ?
C. What is the asymptotic distribution of the estimator you found in B?
D. Under the same assumption as in B, what is the MLE of ρ?
E. What is the asymptotic distribution of the estimator you found in D?

4. **Optional bonus problem:** Suppose that X takes on values in a fixed finite set of k points $\{x_1,\ldots,x_k\}$ with a parametric distribution P_θ: $P_\theta(X = x_j) = p_j(\theta)$, $j = 1,\ldots,k$, $\theta \in \Theta \subset R^d$ with $d < k$. Let $X_1,\ldots,X_n$ be i.i.d. with the distribution of X, and let $\mathbb{P}_n$ be the empirical measure of the X_i's: $\mathbb{P}_n = n^{-1}\sum_{i=1}^n \delta_{X_i}$.
A. Compute $K(\mathbb{P}_n, P_\theta)$.
B. Show that $K(\mathbb{P}_n, P_\theta) = -l(\theta|\underline{X}) + C(\underline{X})$ where $C(\underline{X})$ does not depend on θ.
C. Use the result of B to show that (in this case) the maximum likelihood estimator $\widehat{\theta}_n$ of θ (if it exists) can be expressed as

$$\begin{aligned} \widehat{\theta}_n &= \text{argmax}_\theta l(\theta|\underline{X}) \\ &= \text{argmin}_\theta K(\mathbb{P}_n, P_\theta)\,. \end{aligned}$$

D. Can you extend A - C to an arbitrary parametric family $\mathcal{P} \equiv \{P_\theta : \theta \in \Theta\}$ with densities $p_\theta = dP_\theta/d\mu$ with respect to a dominating measure μ (i.e. a parametric family not necessarily concentrated on a fixed finite set)?