Statistics 582, Problem Set 2

Wellner; 1/17/2006

Reading: Lehmann and Casella, TPE, Chapter 6, section 6.4, especially pages 455 - 461. Chapter 4, sections 5 - 6.
Due: Wednesday, January 24, 2007.

1. Human beings can be classified into one of four blood groups (phenotypes) O, A, B, AB. The inheritance of blood groups is controlled by three genes, O, A, B, of which O is recessive to A and B. If r, p, q are the gene probabilities in the population of O, A, B respectively, then the probabilities of the six possible combinations (genotypes) in random mating (where two individuals drawn at random from the population contribute one gene each) are shown in the following table:

Phenotype	Genotype	probability
O	OO	r^2
A	AA	p^2
A	AO	$2rp$
B	BB	q^2
B	BO	$2rq$
AB	AB	$2pq$

We observe among N individuals the phenotype frequencies N_O, N_A, N_B, N_{AB}, and wish to estimate the gene probabilities from such data. A simple approach is to regard the observations as incomplete, the complete data set being the genotype frequencies N_{OO}, N_{AA}, N_{AO}, N_{BB}, N_{BO}, N_{AB}.
A. Derive the EM algorithm for estimation of (p, q, r).
B. Estimate (p, q, r) from $N_O = 176$, $N_A = 182$, $N_B = 60$, $N_{AB} = 17$.
C. Estimate the covariance matrix of the estimator $(\hat{p}, \hat{q}, \hat{r})$.

2. Lehmann and Casella, TPE, Problem 4.9, page 504.

3. Suppose, as in Example 4.3.10, that $\underline{X}_1, \ldots, \underline{X}_n$ are i.i.d. $\text{Mult}_k(1, \underline{p})$ so that $\underline{N}_n = \sum_{i=1}^n \underline{X}_i \sim \text{Mult}_k(n, \underline{p})$.
(a) Use Jensen's inequality to show that the log-likelihood

$$l_n(\underline{p}|\underline{X}) = \sum_{j=1}^k N_j \log p_j + \sum_{i=1}^n \log\left(\frac{1!}{X_{i1}! \cdots X_{ik}!}\right)$$

is maximized by $\hat{\underline{p}} = \underline{N}_n/n$. [Hint: write the first term of $l_n(\underline{p}|\underline{X})$ as $n \sum_{j=1}^k \hat{p}_j \log p_j$.]
(b) Relate $l_n(\underline{p})$ to $K(\hat{\underline{p}}, \underline{p})$ and hence show again that the maximizing value of $\underline{p}$ is $\hat{p}$.

4. Suppose that the "complete data" X is given by three independent multinomial random vectors,

$$N(1) \equiv (N_{ij}(1) : i = 1, \ldots, r; j = 1, \ldots, s) \sim \text{Mult}_{rs}(n_1; p = (p_{ij}, i = 1, \ldots, r, j = 1, \ldots, s),$$

$$N(2) \equiv (N_{ij}(2) : i = 1, \ldots, r; j = 1, \ldots, s) \sim \text{Mult}_{rs}(n_2; p = (p_{ij}, i = 1, \ldots, r, j = 1, \ldots, s),$$

$$N(3) \equiv (N_{ij}(3) : i = 1, \ldots, r; j = 1, \ldots, s) \sim \text{Mult}_{rs}(n_3; p = (p_{ij}, i = 1, \ldots, r, j = 1, \ldots, s)\,.$$

Suppose that the "incomplete data" Y consists of $N(1), (N_{i\cdot}(2) : 1 \le i \le r), (N_{\cdot j}(3) : 1 \le j \le s)$.

A. What are the distributions of $N(1), (N_{i\cdot}(2) : 1 \le i \le r)$ and $(N_{\cdot j}(3) : 1 \le j \le s)$?

B. Find the conditional distribution(s) of X given Y.

C. Suggest an EM - algorithm for estimation of p.

5. (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

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

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

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

Furthermore, show that the survival function $1 - H(z) = P(Z > z)$ is given by $1 - H(z) = (1 - F(z))(1 - G(z))$ and also $H(z) = H^{(uc)}(z) + H^{(c)}(z)$.

B. Suppose that the cumulative hazard function corresponding to F is defined by

$$\Lambda_F(x) = \int_{[0,x]} \frac{1}{1 - F(y-)} dF(y).$$

Show that this can be expressed in terms of H and H_{uc} as

$$\Lambda_F(x) = \int_{[0,x]} \frac{1}{1 - H(y-)} dH^{(uc)}(y).$$

C. If $\mathbb{H}_n^{(uc)}(z) = n^{-1} \sum_{i=1}^n \delta_i 1\{Z_i \le z\}$ and $\mathbb{H}_n(z) = n^{-1} \sum_{i=1}^n 1\{Z_i \le z\}$, suggest an estimator of Λ_F based on the observed (Z_i, δ_i)'s.

6. **Optional bonus problem.** Lehmann and Casella, TPE, Problem 4.15, page 506.