Statistics 582, Problem Set 2

Wellner; 1/14/2009

Reading: Lehmann and Casella, TPE, Chapter 6, section 6.4,
especially pages 455 - 461 and 504-508.
Chapter 4, sections 5 - 6.

Due: Wednesday, January 21, 2009.

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. Lehmann and Casella, TPE, Problem 4.12, page 505.

4. Lehmann and Casella, TPE, Problem 4.16, page 506. (It seems to me that ζ_i in the third line of the problem statement should be just ζ.)

5. 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 $N(1), (N_{i\cdot}(2) : 1 \le i \le r), (N_{\cdot j}(3) : 1 \le j \le s)$. Thus $N(1)$ gives cell counts for a two-way table, while $(N_{i\cdot}(2) : 1 \le i \le r)$ and $(N_{\cdot j}(3) : 1 \le j \le s)$ give additional information on the marginal distributions of the table. (If n_2 and n_3 are very large relative to n_1, we might regard the marginal distributions as "known".)

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.

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