Statistics 582, Problem Set 2 Solutions

Wellner; 1/17/2018

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 $(\hat{p}, \hat{q}, \hat{r})$.

Solution: A. The complete data is $\underline{N} \equiv (N_{OO}, N_{AA}, N_{AO}, N_{BB}, N_{BO}, N_{AB})$ with multinomial distribution $\text{Mult}_6(N; (r^2, p^2, 2rp, q^2, 2rq, 2pq))$. Thus

$$P(\underline{N} = \underline{n}) = \frac{N!}{n_{OO}!n_{AA}!n_{AO}!n_{BB}!n_{BO}!n_{AB}!} \cdot p^{2n_{AA}+n_{AO}+n_{AB}}\, q^{2n_{BB}+n_{BO}+n_{AB}}\, r^{2n_{OO}+n_{AO}+n_{BO}}\, 2^{n_{AO}+n_{BO}+n_{AB}}.$$

This is proportional to a $\text{Mult}_3(2N; (p, q, r))$ distribution, and hence the MLE's based on the complete data are

$$(\hat{p}, \hat{q}, \hat{r}) = \frac{1}{2N}(2N_{AA} + N_{AO} + N_{AB}, 2N_{BB} + N_{BO} + N_{AB}, 2N_{OO} + N_{AO} + N_{BO})\,.$$

This forms the basis of the "M - step"of an E-M algorithm. The incomplete data Y is $(N_A, N_B, N_O, N_{AB}) = (N_{AA} + N_{AO}, N_{BB} + N_{BO}, N_{OO}, N_{AB})$; thus

$$(N_{AA}|Y) = (N_{AA}|N_A) \sim \text{Binomial}(N_A, \frac{p^2}{p^2 + 2rp}), \qquad E(N_{AA}|Y) = N_A \frac{p}{p + 2r},$$

$$(N_{AO}|Y) = (N_{AO}|N_A) \sim \textit{Binomial}(N_A, \frac{2rp}{p^2 + 2rp}), \qquad E(N_{AO}|Y) = N_A \frac{2r}{p + 2r},$$

$$(N_{BB}|Y) = (N_{BB}|N_B) \sim Binomial(N_B, \frac{q^2}{q^2+2rq}), \qquad E(N_{BB}|Y) = N_B\frac{q}{q+2r},$$
$$(N_{BO}|Y) = (N_{BO}|N_B) \sim Binomial(N_B, \frac{2rq}{q^2+2rq}), \qquad E(N_{BO}|Y) = N_B\frac{2r}{q+2r}.$$

This gives the basis of the "E - step" for an E - M algorithm. Hence, starting from $(\hat{p}^{(0)}, \hat{q}^{(0)}, \hat{r}^{(0)}) = (1/3, 1/3, 1/3)$ say, we take

$$(\hat{p}^{(m+1)}, \hat{q}^{(m+1)}) = \frac{1}{2N}(2\hat{N}_{AA}^{(m)} + \hat{N}_{AO}^{(m)} + N_{AB}, 2\hat{N}_{BB}^{(m)} + \hat{N}_{BO}^{(m)} + N_{AB}),$$
$$\hat{r}^{(m+1)} = 1 - \hat{p}^{(m+1)} - \hat{q}^{(m+1)}$$

where

$$\hat{N}_{AA}^{(m)} \equiv N_A\frac{\hat{p}^{(m)}}{\hat{p}^{(m)} + 2\hat{r}^{(m)}}, \qquad \hat{N}_{AO}^{(m)} \equiv N_A - \hat{N}_{AA}^{(m)},$$
$$\hat{N}_{BB}^{(m)} \equiv N_B\frac{\hat{q}^{(m)}}{\hat{q}^{(m)} + 2\hat{r}^{(m)}}, \qquad \hat{N}_{BO}^{(m)} \equiv N_B - \hat{N}_{BB}^{(m)}.$$

B. For the given data, the E - M algorithm in A yields:

Iteration	$\hat{p}^{(m)}$	$\hat{q}^{(m)}$
0	.333	.333
1	.298	.111
2	.271	.094
3	.266	.093
4	.265	.093
5	.264	.093
6	.264	.093

Thus the estimator is $(\hat{p}, \hat{q}, \hat{r}) = (.264, .093, .642)$.

C. *Method 1: Direct calculation from the (incomplete) data* $\underline{Y} \equiv (N_A, N_B, N_O, N_{AB})$.
The likelihood of the observations $(N_A, N_B, N_O, N_{AB}) = (N_{AA} + N_{AO}, N_{BB} + N_{BO}, N_{OO}, N_{AB})$ is

$$\begin{aligned} l_N(p,q) &= N_A\log(p^2 + 2p(1-p-q)) \\ &\quad + N_B\log(q^2 + 2q(1-p-q)) \\ &\quad + N_O\log(1-p-q)^2 + N_{AB}\log(2pq)\,. \end{aligned}$$

Thus

$$\begin{aligned} -\frac{\partial^2}{\partial p^2}l_N(p,q) &= 2N_A\{\frac{1}{2p-p^2-2pq} + \frac{2(1-p-q)^2}{(2p-p^2-2pq)^2}\} \\ &\quad + N_B\frac{4q^2}{(2q-q^2-2pq)^2} \\ &\quad + \frac{N_{AB}}{p^2} + \frac{2N_O}{(1-p-q)^2}\,, \end{aligned}$$

$$
\begin{aligned}
-\frac{\partial^2}{\partial p \partial q} l_N(p,q) &= 2N_A\{\frac{1}{2p-p^2-2pq} - \frac{2p(1-p-q)}{(2p-p^2-2pq)^2}\} \\
&\quad + 2N_B\{\frac{1}{2q-q^2-2pq} - \frac{4q^2}{(2q-q^2-2pq)^2}\} \\
&\quad + \frac{2N_O}{(1-p-q)^2}\,,
\end{aligned}
$$

$$
\begin{aligned}
-\frac{\partial^2}{\partial q^2} l_N(p,q) &= N_A \frac{4p^2}{(2p-p^2-2pq)^2} \\
&\quad + 2N_B\{\frac{1}{2q-q^2-2pq} + \frac{2(1-p-q)^2}{(2q-q^2-2pq)^2}\} \\
&\quad + \frac{N_{AB}}{q^2} + \frac{2N_O}{(1-p-q)^2}\,.
\end{aligned}
$$

Since

$$
\begin{aligned}
E(N_A) &= N(p^2 + 2p(1-p-q))\,, \\
E(N_B) &= N(2q - q^2 - 2pq)\,, \\
E(N_{AB}) &= N(2pq), \\
\end{aligned}
$$

and

$$
E(N_O) = N(1-p-q)^2\,,
$$

it follows that

$$
I_{11}(p,q) = 2N\{1 + \frac{2r^2}{2p-p^2-2pq} - \frac{2q^2}{2q-q^2-2pq} + \frac{q}{p} + 1\},
$$

$$
I_{12}(p,q) = 2N\{2 - \frac{2p(1-p-q)}{(2p-p^2-2pq)^2} - \frac{2q(1-p-q)}{(2q-q^2-2pq)^2} + 1\}\,,
$$

$$
I_{22}(p,q) = 2N\{1 + \frac{2r^2}{2q-q^2-2pq} - \frac{2p^2}{2p-p^2-2pq} + \frac{p}{q} + 1\}
$$

and hence the estimated Fisher information matrix is

$$
\hat{I}(p,q) = N \begin{pmatrix} 9.01584 & 2.47553 \\ 2.47553 & 23.2541 \end{pmatrix}
$$

so that

$$
\hat{I}^{-1}(p,q) = \frac{1}{N} \begin{pmatrix} 0.114256 & -0.0121631 \\ -0.0121631 & 0.044298 \end{pmatrix} = 10^{-3} \cdot \begin{pmatrix} 0.262657 & -0.027961 \\ -0.027961 & 0.101834 \end{pmatrix}.
$$

Furthermore, since $\hat{r} = 1 - \hat{p} - \hat{q}$,

$$
Var(\hat{r}) = Var(\hat{p}) + Var(\hat{q}) + 2Cov(\hat{p}, \hat{q})\,,
$$

$$
\begin{aligned}
Cov(\hat{p}, \hat{r}) &= -Var(\hat{p}) - Cov(\hat{p}, \hat{q})\,, \\
Cov(\hat{q}, \hat{r}) &= -Var(\hat{q}) - Cov(\hat{p}, \hat{q})\,;
\end{aligned}
$$

and hence we estimate $Cov(\hat{p}, \hat{q}, \hat{r})$ by

$$\widehat{Cov}(\hat{p}, \hat{q}, \hat{r}) = 10^{-3} \cdot \begin{pmatrix} 0.262657 & -0.027961 & -0.234695 \\ -0.027961 & 0.101834 & -0.073873 \\ -0.234695 & -0.073873 & 0.308569 \end{pmatrix}.$$

Method 2: Via Louis's formula
Louis (1982) gives the formula

$$\hat{I}_Y(\theta) = E_\theta\{-\ddot{\mathbf{l}}_{\theta\theta}(\underline{X})|\underline{Y}\} - Cov_\theta(\dot{\mathbf{l}}_\theta(\underline{X}), \dot{\mathbf{l}}_\theta(\underline{X})|\underline{Y}).$$

I will apply this to the complete data model for $\underline{X} = (N_{AA}, N_{AO}, N_{BB}, N_{BO}, N_O, N_{AB})$ parameterized by $\theta = (p, q)$, and hence $r = 1 - p - q$. Thus the scores for p and q are given by

$$\dot{\mathbf{l}}_\theta(\underline{X}) = \begin{pmatrix} \dot{\mathbf{l}}_p(\underline{X}) \\ \dot{\mathbf{l}}_q(\underline{X}) \end{pmatrix} = \begin{pmatrix} \frac{N_p}{p} - \frac{N_r}{r} \\ \frac{N_q}{q} - \frac{N_r}{r} \end{pmatrix}$$

where

$$\begin{aligned} N_p &= 2N_{AA} + N_{AO} + N_{AB}, \\ N_q &= 2N_{BB} + N_{BO} + N_{AB}, \\ N_r &= 2N_{OO} + N_{AO} + N_{BO}. \end{aligned}$$

Furthermore, minus one times the matrix of second derivatives is

$$-\ddot{\mathbf{l}}_{\theta\theta}(\underline{X}) = \begin{pmatrix} \frac{N_p}{p^2} + \frac{N_r}{r^2} & \frac{N_r}{r^2} \\ \frac{N_r}{r^2} & \frac{N_q}{q^2} + \frac{N_r}{r^2} \end{pmatrix}.$$

To compute the terms in Louis's formula we need $E_\theta(\underline{X})|\underline{Y})$ and $Cov_\theta(\underline{X}|\underline{Y})$ where $\underline{Y} = (N_A, N_B, N_O, N_{AB})$. To this end we calculate

$$\begin{aligned} E(N_{AA}|N_A) &= N_A \frac{p^2}{p^2 + 2pr}, \quad E(N_{AO}|N_A) = N_A \frac{2pr}{p^2 + 2pr}, \\ E(N_{BB}|N_A) &= N_B \frac{q^2}{q^2 + 2qr}, \quad E(N_{BO}|N_B) = N_A \frac{2qr}{q^2 + 2qr}. \end{aligned}$$

Furthermore, since the conditional distribution of $\underline{X}$ given $\underline{Y}$ is given by

$$\text{Mult}_2\left(N_A, \left(\frac{p^2}{p^2 + 2pr}, \frac{2pr}{p^2 + 2pr}\right)\right) \cdot \text{Mult}_2\left(N_A, \left(\frac{q^2}{q^2 + 2qr}, \frac{2qr}{q^2 + 2qr}\right)\right) \cdot \delta_{X_5 = Y_3} \cdot \delta_{X_6 = Y_4},$$

the conditional covariance matrix of the complete data scores given $\underline{Y}$ is given by

$$\begin{aligned} &Cov_\theta(\dot{\mathbf{l}}_\theta(\underline{X})|\underline{Y}) \qquad (1) \\ &= \begin{pmatrix} Var_\theta(N_p/p - N_r/r|\underline{Y}) & Cov_\theta(N_p/p - N_r/r, N_q/q - N_r/r|\underline{Y}) \\ Cov_\theta(N_p/p - N_r/r, N_q/q - N_r/r|\underline{Y}) & Var_\theta(N_q/q - N_r/r|\underline{Y}) \end{pmatrix} \\ &= \begin{pmatrix} v_p\left(\frac{1}{p} + \frac{1}{r}\right)^2 + \frac{1}{r^2} v_q & v_p\left(\frac{1}{r^2} + \frac{1}{rp}\right) + v_q\left(\frac{1}{r^2} + \frac{1}{rq}\right) \\ v_p\left(\frac{1}{r^2} + \frac{1}{rp}\right) + v_q\left(\frac{1}{r^2} + \frac{1}{rq}\right) & v_q\left(\frac{1}{q} + \frac{1}{r}\right)^2 + \frac{1}{r^2} v_p \end{pmatrix} \end{aligned}$$

since, by noting that $Var_\theta(N_p|\underline{Y}) = Var_\theta(-N_{AO}|\underline{Y})$ and $Var_\theta(N_q|\underline{Y}) = Var_\theta(-N_{BO}|\underline{Y})$, we have

$$Var_\theta(N_p|\underline{Y}) = N_A \frac{p^2}{p^2+2pr} \cdot \frac{2pr}{p^2+2pr} \equiv v_p, \tag{2}$$

$$Var_\theta(N_q|\underline{Y}) = N_B \frac{q^2}{q^2+2qr} \cdot \frac{2qr}{q^2+2qr} \equiv v_q, \tag{3}$$

$$Var_\theta(N_r|\underline{Y}) = Var_\theta(N_p|\underline{Y}) + Var_\theta(N_q|\underline{Y}),$$
$$Cov_\theta(N_p, N_q)|\underline{Y}) = 0,$$
$$Cov_\theta(N_p, N_r|\underline{Y}) = -Var_\theta(N_p|\underline{Y}),$$
$$Cov_\theta(N_q, N_r|\underline{Y}) = -Var_\theta(N_q|\underline{Y}).$$

Combining these pieces and computing, we find that

$$\hat{I}(p,q) = N \begin{pmatrix} 8.99267 & 2.45797 \\ 2.45797 & 23.2005 \end{pmatrix}$$

so that

$$\hat{I}^{-1}(p,q) = \frac{1}{N} \begin{pmatrix} 0.114518 & -0.0121325 \\ -0.0121325 & 0.0443878 \end{pmatrix} = 10^{-3} \cdot \begin{pmatrix} 0.26326 & -0.027891 \\ -0.027891 & 0.102041 \end{pmatrix}.$$

Furthermore, since $\hat{r} = 1 - \hat{p} - \hat{q}$,

$$Var(\hat{r}) = Var(\hat{p}) + Var(\hat{q}) + 2Cov(\hat{p}, \hat{q}),$$

$$Cov(\hat{p}, \hat{r}) = -Var(\hat{p}) - Cov(\hat{p}, \hat{q}),$$
$$Cov(\hat{q}, \hat{r}) = -Var(\hat{q}) - Cov(\hat{p}, \hat{q});$$

and hence we estimate $Cov(\hat{p}, \hat{q}, \hat{r})$ by

$$\widehat{Cov}(\hat{p}, \hat{q}, \hat{r}) = 10^{-3} \cdot \begin{pmatrix} 0.26326 & -0.027891 & -0.235369 \\ -0.027891 & 0.102041 & -0.074150 \\ -0.235369 & -0.074150 & 0.309095 \end{pmatrix}.$$

See the end of this solution set for the Mathematica code I used to carry out these calculations.

2. Consider nonparametric maximum likelihood estimation of F in the right-censored data problem considered in class, but extend the argument to include ties as follows: (a) When there are ties, let the distinct Z's be denoted by $T_1 < \ldots < T_k$.Let $m_1, \ldots, m_k$ and $n_1, \ldots, n_k$ be defined by $m_j \equiv \#$ of $Z_i\Delta_i = T_j$, $n_j \equiv \#$ of $Z_i(1-\Delta_i) = T_j$, and let $p_j \equiv \Delta F(T_j) = F(T_j) - F(T_j-)$, $j = 1, \ldots, k$, $p_{k+1} = 1 - F(T_k)$. Show that the likelihood (for F) is

$$L(F|\underline{Z}, \underline{\Delta}) = \prod_{i=1}^{k} p_i^{m_i} (\sum_{j=i+1}^{k+1} p_j)^{n_i}.$$

(b) By defining $\lambda_i = p_i / \sum_{j=i}^{k+1} p_j$ for $i = 1, \ldots, k$ and $\lambda_{k+1} = 1$, and rewriting the likelihood in terms of the λ_i's, show that the likelihood is maximized by

$$\hat{\lambda}_i = m_i / \sum_{j=i}^{k} (m_j + n_j) = \frac{n \Delta \mathbb{H}_n^{uc}(T_i)}{n(1 - \mathbb{H}_n(T_i-))}.$$

and hence that the nonparametric MLE of F is (again) the Kaplan - Meier estimator

$$1 - \widehat{F}_n(t) = \prod_{s \le t} (1 - \Delta \widehat{\Lambda}_n(s)).$$

(c) Compute $1 - \widehat{F}_n$ for the following data (lengths of remission in weeks for the 6-MP group of leukemia patients from Lawless (1982), *Statistical Models and Methods for Survival Data*, page 5):

$$\begin{array}{llllllllllll} 6, & 6, & 6, & 6+, & 7, & 9+, & 10, & 10+, & 11+, & 13, & 16, \\ 17+, & 19+, & 20+, & 22, & 23, & 25+, & 32+, & 32+, & 34+, & 35+ \end{array}$$

here $+$ indicates censoring ($\Delta = 0$).

Solution: (a) When there are ties, let the distinct Z's be denoted by $T_1 < \cdots < T_k$ Let $m_1, \ldots, m_k$ and $n_1, \ldots, n_k$ be defined by $m_j = \#\{i \le n : \ Z_i \Delta_i = T_j\}$, $n_j = \#\{i \le n : \ Z_i(1-\Delta_i) = T_j\}$, and let $p_j \equiv \Delta F(T_j)$, $j = 1, \ldots, k$, $p_{k+1} = 1 - F(T_k-)$. Then the likelihood (for F) is

$$L(F|\underline{Z}, \underline{\delta}) = \prod_{i=1}^{k} p_i^{m_i} \left(\sum_{j=i+1}^{k} p_j \right)^{n_i}.$$

Setting $\lambda_i \equiv p_i / \sum_{j=i}^{k+1} p_j$, $\lambda_{k+1} = 1$ yields

$$\sum_{j=i}^{k+1} p_j = \prod_{j=1}^{i-1} (1 - \lambda_j), \qquad 1 - \lambda_i = \frac{\sum_{j=i+1}^{k+1} p_j}{\sum_{j=i}^{k+1} p_j},$$

and hence

$$\begin{aligned} L(F|\underline{Z}, \underline{\Delta}) &= \prod_{i=1}^{k} \left(\frac{p_i}{\sum_{j=i}^{k+1} p_j} \right)^{m_i} (\sum_{j=i}^{k+1} p_j)^{m_i} \left\{ \frac{\sum_{j=i+1}^{k+1} p_j}{\sum_{j=i}^{k+1} p_j} \sum_{j=i}^{k+1} p_j \right\}^{n_i} \\ &= \prod_{i=1}^{k} \lambda_i^{m_i} (1 - \lambda_i)^{n_i} \left(\sum_{j=i}^{k+1} p_j \right)^{m_i + n_i} \\ &= \prod_{i=1}^{k} \lambda_i^{m_i} (1 - \lambda_i)^{n_i} \left(\prod_{j=1}^{i-1} (1 - \lambda_j) \right)^{m_i + n_i} \\ &= \prod_{i=1}^{k} \lambda_i^{m_i} (1 - \lambda_i)^{n_i + \sum_{j=i+1}^{k} (m_j + n_j)} \\ &= \prod_{i=1}^{k} \lambda_i^{m_i} (1 - \lambda_i)^{r_i - m_i} \end{aligned}$$

where $r_i \equiv \sum_{j=i}^{k}(m_j + n_j)$.
(b) In view of the binomial form of this expression for each i, we know that it is maximized for each i by

$$\hat{\lambda}_i \;=\; \frac{m_i}{r_i} = \frac{m_i}{\sum_{j=i}^{k}(m_j+n_j)} = \frac{n\Delta\mathbb{H}_n^{(uc)}(T_i)}{n(1-\mathbb{H}_n(T_i-))},$$

for $i = 1, \ldots, k$. Then

$$\hat{p}_i = \prod_{j=1}^{i-1}(1-\hat{\lambda}_j)\hat{\lambda}_i, \qquad i = 1, \ldots, k+1.$$

as before. Note that $\hat{p}_{k+1} > 0$ if $n_k > 0$. Thus the nonparametric MLE's $\hat{\Lambda}_n$ and $\hat{F}_n$ of Λ and F are the Nelson-Aalen and Kaplan-Meier (or product-limit) estimators

$$\hat{\Lambda}_n(t) = \int_{[0,t]} \frac{d\mathbb{H}_n^{(uc)}(s)}{1-\mathbb{H}_n(s-)}$$

and $1 - \hat{F}_n(t) = \prod_{s\le t}(1-\Delta\hat{\Lambda}_n(s))$.
(c) For the given data, there are 16 distinct times T_i: these are 6, 7, 9, 10, 11, 13, 16, 17, 19, 20, 22, 23, 25, 32, 34, 35, with ties at 6, 10, and 32. If we let $r_i \equiv n(1-\mathbb{H}_n(T_i-))$ and $d_i = n\Delta\mathbb{H}_n^{(uc)}(T_i)$ then we obtain the following table and calculated values of the estimator:

Table 1:

T_i	r_i	d_i	$1-\frac{d_i}{r_i}$	$\prod_{j\le i}(1-\frac{d_i}{r_i})$	$\widehat{Var}(\hat{F})$	$\widehat{Var}_{GW}(\hat{F})$
6	21	3	6/7	.8571	.004998	.005831
7	17	1	16/17	.8067	.006679	.007558
9	16	0	1	.8067	.006679	.007558
10	15	1	14/15	.7529	.008338	.009283
11	13	0	1	.7529	.008338	.009283
13	12	1	11/12	.6902	.010314	.011419
16	11	1	10/11	.6275	.011778	.013008
17	10	0	1	.6275	.011778	.013008
19	9	0	1	.6275	.011778	.013008
20	8	0	1	.6275	.011778	.013008
22	7	1	6/7	.5378	.014556	.016444
23	6	1	5/6	.4482	.015688	.018115
25	5	0	1	.4482	.015688	.018115
32	4	0	1	.4482	.015688	.018115
34	2	0	1	.4482	.015688	.018115
35	1	0	1	.4482	.015788	.018115

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 $\hat{p}$.
(c) Use this problem and similar considerations as in the previous problem to formulate a version of Example 6.1 in the lecture notes when ties are present and show that the nonparametric MLE continues to be the empirical measure $\mathbb{P}_n$.

Solution: (a) Our goal is to show that

$$n\sum_{j=1}^k \hat{p}_j \log p_j \le n\sum_{j=1}^k \hat{p}_j \log \hat{p}_j$$

with equality if and only if $\underline{p} = \hat{\underline{p}}$. Subtracting the right side from the left side and dividing by n, we see that we want to show that

$$\sum_{j=1}^k \hat{p}_j \log\left(\frac{p_j}{\hat{p}_j}\right) \le 0.$$

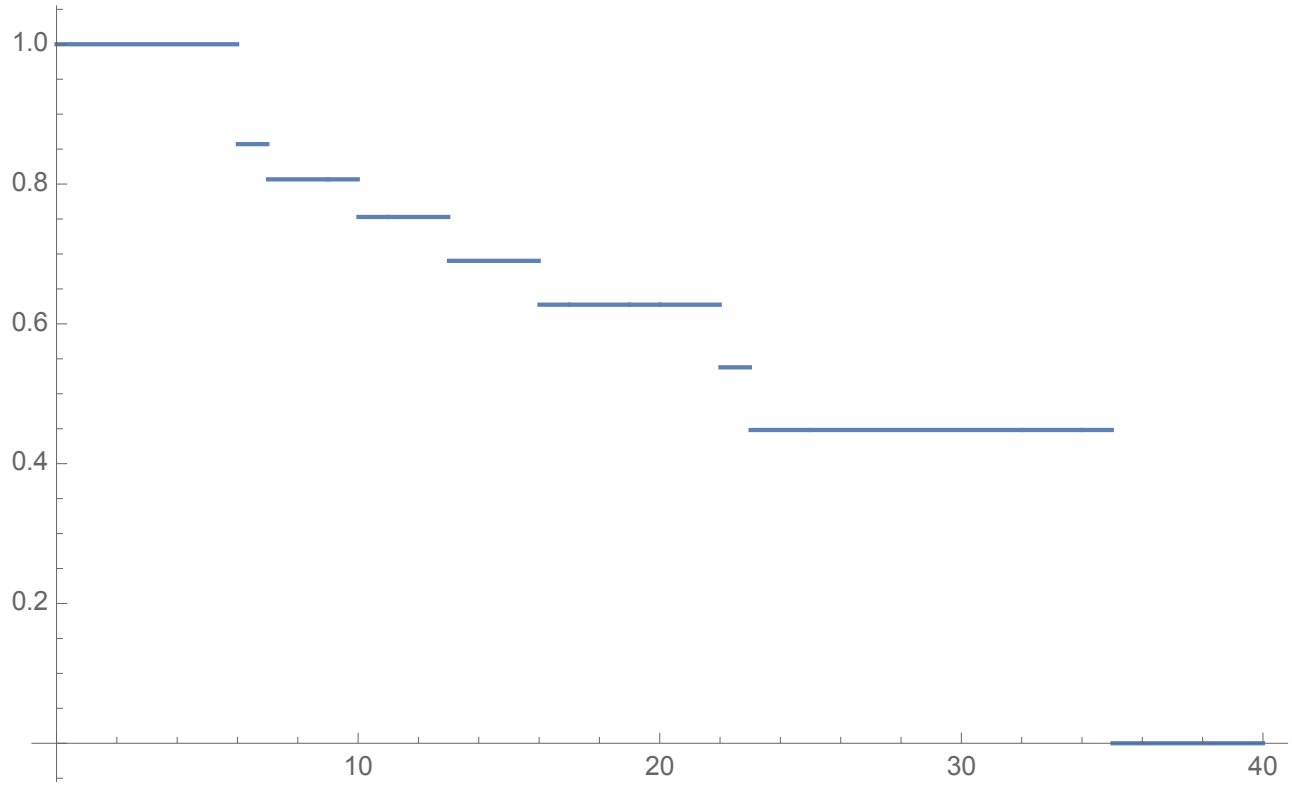

Figure 1: Plot of $1 - \widehat{F}_n(t)$, 6-MP group, leukemia remission times

But since log is a concave function, Jensen's inequality yields

$$\begin{aligned} \sum_{j=1}^{k} \hat{p}_j \log\left(\frac{p_j}{\hat{p}_j}\right) &\leq \log\left(\sum_{j=1}^{k} \hat{p}_j \left(\frac{p_j}{\hat{p}_j}\right)\right) \\ &= \log\left(\sum_{j=1}^{k} p_j\right) = \log(1) = 0. \end{aligned}$$

(b) Note that in the above argument we have shown that

$$l_n(\underline{p}) - l_n(\underline{\hat{p}}) = -nK(\underline{\hat{p}}, \underline{p}) \leq 0$$

since $K(P, Q) \geq 0$ for all P, Q. Thus $l_n(\underline{p})$ is maximized by $\underline{p} = \underline{\hat{p}}$.

4. We showed in class that the nonparametric maximum likelihood estimator of F in the (right) censored data problem, possibly with ties, is the Kaplan-Meier (product limit) estimator $\widehat{F}_n(t)$ given by

$$1 - \widehat{F}_n(t) = \prod_{s \leq t} (1 - \Delta\widehat{\Lambda}_n(s))$$

where $\widehat{\Lambda}_n(t)$ is the *Nelson-Aalen* estimator of

$$\Lambda(t) \equiv \Lambda_F(t) \equiv \int_0^t \frac{1}{1 - F_-} dF\,,$$

given by

$$\widehat{\Lambda}_n(t) = \int_0^t \frac{1}{1 - \mathbb{H}_n(s-)} d\mathbb{H}_n^{uc}(s) \qquad 0 \le s \le t\,.$$

Here

$$\mathbb{H}_n^{uc}(t) = \frac{1}{n}\sum_{i=1}^n \Delta_i 1_{[Z_i \le t]}\,, \qquad \mathbb{H}_n(t) = \frac{1}{n}\sum_{i=1}^n 1_{[Z_i \le t]}$$

are the sub-empirical distribution function of the uncensored observations and the marginal empirical distribution of all the Z's uncensored or censored.
(a) In class I gave a sketch of proof that

$$\sqrt{n}(\widehat{F}_n(t) - F(t)) \Rightarrow (1 - F(t))\mathbb{B}(C(t))$$

as a process uniformly in $t \in [0, \tau]$ for any $\tau < \tau_H$ (i.e. for any τ with $1 - H(\tau) = (1 - F(\tau))(1 - G(\tau)) > 0$, where $\mathbb{B}$ is a standard Brownian motion process and where

$$C(t) \equiv \int_0^t \frac{1}{(1 - H_-(s))^2} dH^{uc}(s)\,, \qquad 0 \le s \le t$$

Thus we have, for each fixed $t < \tau$,

$$\sqrt{n}(\widehat{F}_n(t) - F(t)) \to_d N(0, (1 - F(t))^2 C(t))$$

Suggest an estimator of $C(t)$ and hence an estimator of $(1 - F(t))^2 C(t)$.
(b) Show that your estimator of $(1 - F(t))^2 C(t)$ is consistent.
(c) Use the estimator you suggest in (b) to obtain an approximate 90% confidence interval for $F(15)$ based for the data given in problem 2 above.

Solution: (a) In this case there are ties in the data. A table giving the distinct time points T_i together with the numbers at risk and the number of deaths at each time point, together with the successive terms of the product and the resulting Kaplan-Meier estimator was given in problem 1. The last two columns of the table give two variance estimates: column 6 gives the variance estimator from (b) below; column 7 gives the usual Greenwood estimator (cf. part the notes handed out in class on 14 January and Kalbfleisch and Prentice (1980), pages 12 - 14). (b) A natural estimator of

$$C(t) = \int_{[0,t]} \frac{1}{(1 - H(s-))^2} dH^{(uc)}(s)$$

is

$$\begin{aligned} \hat{C}_n(t) &= \int_{[0,t]} \frac{1}{(1 - \mathbb{H}_n(s-))^2} d\mathbb{H}_n^{(uc)}(s) \\ &= n\int_{[0,t]} \frac{1}{R_n(s)^2} d\left(n\mathbb{H}_n^{(uc)}(s)\right) \end{aligned}$$

where $R_n(s) \equiv n(1 - \mathbb{H}_n(s-))$. Note that in the Mathematica program accompanying the solution set the quantity labeled "Cest" is $n^{-1}\hat{C}_n(t) =$

$\int_{[0,t]} R_n(s)^{-2} d(n\mathbb{H}_n^{(uc)}(s))$.

(c) To see that $\hat{C}_n(t) \to_p C(t)$ note that

$$\|\mathbb{H}_n^{(uc)} - H^{(uc)}\|_\infty = \sup_{0<t<\infty} |\mathbb{H}_n^{(uc)}(t) - H^{(uc)}(t)| \to_{a.s.} 0, \tag{4}$$

$$\|\mathbb{H}_n - H\|_\infty = \sup_{0<t<\infty} |\mathbb{H}_n(t) - H(t)| \to_{a.s.} 0 \tag{5}$$

by the Glivenko-Cantelli theorem. Therefore we can write

$$\begin{aligned}
\hat{C}_n(t) - C(t) &= \int_{[0,t]} \frac{d\mathbb{H}_n^{(uc)}(s)}{(1-\mathbb{H}_n(s-))^2} - \int_{[0,t]} \frac{dH^{(uc)}(s)}{(1-H(s-))^2} \\
&= \int_{[0,t]} \left(\frac{1}{(1-\mathbb{H}_n(s-))^2} - \frac{1}{(1-H(s-))^2} \right) d\mathbb{H}_n^{(uc)}(s) \\
&\quad + \int_{[0,t]} \frac{1}{(1-H(s-))^2} d(\mathbb{H}_n^{(uc)}(s) - H^{(uc)}(s)) \\
&= \int_{[0,t]} \frac{(1-H(s-))^2 - (1-\mathbb{H}_n(s-))^2}{(1-\mathbb{H}_n(s-))^2(1-H(s-))^2} d\mathbb{H}_n^{(uc)}(s) \\
&\quad + \int_{[0,t]} \frac{1}{(1-H(s-))^2} d(\mathbb{H}_n^{(uc)}(s) - H^{(uc)}(s)) \\
&= \int_{[0,t]} \frac{[(1-H(s-)) - (1-\mathbb{H}_n(s-))][(1-H(s-) + (1-\mathbb{H}_n(s-))]}{(1-\mathbb{H}_n(s-))^2(1-H(s-))^2} d\mathbb{H}_n^{(uc)}(s) \\
&\quad + \int_{[0,t]} \frac{1}{(1-H(s-))^2} d(\mathbb{H}_n^{(uc)}(s) - H^{(uc)}(s)) \\
&\equiv I_n(t) + II_n(t)
\end{aligned}$$

where

$$\begin{aligned}
|I_n(t)| &\leq 2\frac{\sup_{0<s\leq t} |\mathbb{H}_n(s-) - H(s-)|}{(1-\mathbb{H}_n(t-))^2(1-H(t-))^2} \int_{[0,t]} d\mathbb{H}_n^{(uc)}(s) \\
&\leq 2\frac{\sup_{0<s\leq t} |\mathbb{H}_n(s-) - H(s-)|}{(1-\mathbb{H}_n(t-))^2(1-H(t-))^2} \cdot 1 \\
&\to_{a.s.} 0 \cdot \frac{1}{(1-H(t-))^4} \cdot 1 = 0
\end{aligned}$$

if $1 - H(t-) > 0$ by (5). Also,

$$\begin{aligned}
|II_n(t)| &\leq |\int_{[0,t]} \frac{1}{(1-H(s-))^2} d(\mathbb{H}_n^{(uc)}(s) - H^{(uc)}(s)| \\
&= |n^{-1} \sum_{i=1}^n \left\{ \frac{\Delta_i 1_{[0,t]}(Z_i)}{(1-H(Z_i-)^2} - E\left(\frac{\Delta 1_{[0,t]}(Z)}{(1-H(Z-))^2} \right) \right\}| \\
&\to_{a.s.} 0
\end{aligned}$$

by the strong law of large numbers where we again use $1 - H(t-) > 0$. Thus $|\hat{C}_n(t) - C(t)| \leq |I_n(t)| + |II_n(t)| \to_{a.s.} 0$. Assuming that $1 - \hat{F}_n(t) \to_p 1 - F(t)$ this yields

$$(1-\hat{F}_n(t))^2 \hat{C}_n(t) \to_p (1-F(t))^2 C(t).$$

(d) An approximate 90% confidence interval for $F(15)$ is given by

$$\hat{F}_n(15) \pm z_{.95} n^{-1/2} (1 - \hat{F}_n(15)) \sqrt{\hat{C}_n(15)}.$$

where $P(N(0,1) > z_{.95}) = .05$. For the data given I compute $1 - \hat{F}_n(15) = 0.6902$, $n^{-1}\hat{C}_n(15) = 0.02165$, and hence an approximate 90% confidence interval for the point estimator $\hat{F}_n(15) = 1 - .6902 = .3098$ is given by

$$\begin{aligned} & 0.3098 \pm 1.64485(.6902)(.02165)^{1/2} \\ = \;& 0.3098 \pm 1.64485(0.10156) \\ = \;& 0.3098 \pm 0.1671 = (0.1427, 0.4769). \end{aligned} \tag{6}$$

(Alternatively, an approximate 90% confidence interval for $1 - \hat{F}_n(15)$ is $.6902 \pm 0.1671 = (.5231, 0.8573)$.)

It turns out that the variance estimator based on $\hat{C}_n$ is *not* the usual one for the Kaplan-Meier estimator: instead the usual Greenwood formula for estimation of $C(t)$ is

$$\widehat{C}_n^{GW}(t) = \int_{[0,t]} \frac{d\mathbb{H}_n^{(uc)}(s)}{(1 - \mathbb{H}_n(s-)(1 - \mathbb{H}_n(s-) - \Delta\mathbb{H}_n^{(uc)}(s))}.$$

This yields $n^{-1}\widehat{C}_n^{GW}(15) = 0.02395$ and the resulting value of $\widehat{Var}_{GW}(\hat{F}_n(t))$ at $t = 15$ is 0.01141 (rather than $.6902^2 \cdot 0.02165 = 0.01031$ as in (6)). This leads to the slightly wider confidence interval

$$0.3098 \pm 1.64485(.6902)(0.02395)^{1/2} = .3098 \pm 0.17569 = (0.1341, 0.4855). \tag{7}$$

(Alternatively, an approximate 90% confidence interval for $1 - \hat{F}_n(15)$ is $.6902 \pm 0.17569 = (.5145, 0.8659)$.)

See Kalblfeisch and Prentice page 15 for a brief discussion of alternatives involving transformations to stay in the range $[0, 1]$ and to improve the normal approximation.