MATH-449 - Biostatistics
EPFL, Spring 2025
Problem Set 5 - Answer Key

Consider the following filtrations from the lectures:

$$\begin{aligned}
\mathcal{F}_t &= \sigma(L_0, N(u), Z(u); 0 \leq u \leq t) && \text{(the information we have)}\\
\mathcal{F}^c_t &= \sigma(L_0, N^c(u); 0 \leq u \leq t) && \text{(what we want to make inference about)}\\
\mathcal{G}_t &= \sigma(L_0, N^c(u), C(u); 0 \leq u \leq t) && \text{(auxiliary filtration to define independent censoring)}
\end{aligned}$$

where $C(t) = I(t \leq T^*)$ and T^* is the censoring time, $N^c(t) = I(t \geq T)$ and T is the event time, $N(t) = I(t \geq \tilde{T}, D = 1)$ with $\tilde{T} = \min(T, T^*)$ and $D = I(T < T^*)$, and $Z(t) = I(t \leq \tilde{T})$. Thus, $\mathcal{F}_t \subseteq \mathcal{G}_t \supseteq \mathcal{F}^c_t$. L_0 is the (possibly empty) set of covariates known at time $t = 0$.

1. From the lectures we recall that the censoring is independent if the compensator Λ^c of N^c with respect to $\mathcal{F}^c$ is also the compensator of N^c with respect to $\mathcal{G}$. This can be rephrased as

$$E[N^c(t)|\mathcal{G}_t] = E[N^c(t)|\mathcal{F}^c_t]. \tag{1}$$

Thus, (1) hold if and only if we have independent censoring. We will often focus on the intensity λ^c instead of the cumulative intensity $\Lambda^c(t) = \int_0^t \lambda^c(s)ds$.

In the lectures, you learned that the independent censoring assumption in this context is that the intensity of the observed counting process N with respect to the observed information $\mathcal{F}$ is

$$\lambda(t) = Z(t)\alpha(t), \tag{2}$$

where $\alpha(t)$ is the hazard function: [1]

$$\alpha(t) = \alpha(t, L_0) = \lim_{h \to 0+} \frac{1}{h} P(t \leq T < t + h | t \leq T, L_0).$$

It turns out that (2) defines the intensity of N with respect to $\mathcal{F}$ if and only if [2]

$$\lim_{h \to 0+} \frac{1}{h} P(t \leq T < t + h | t \leq T, L_0) = \lim_{h \to 0+} \frac{1}{h} P(t \leq T < t + h | t \leq \tilde{T}, L_0). \tag{3}$$

a) Show that independent censoring holds if $T \perp\!\!\!\perp T^* | L_0$, i.e. if we have random censoring when conditioning on L_0.

Solution Note that $\{t \leq \tilde{T}\}$ is equivalent to $\{t \leq T, t \leq T^*\}$. We thus have

$$P(t \leq T < t+h | t \leq \tilde{T}, L_0) = P(t \leq T < t+h | t \leq T, t \leq T^*, L_0) = P(t \leq T < t+h | t \leq T, L_0)$$

where we used random censoring for the last equality, so random censoring implies that (3) holds, and thus that (2) holds, which is the independent censoring assumption in this setting.

2. Consider the counting process N^c and suppose that the intensity $\lambda^{\mathcal{G}}$ of N^c with respect to $\mathcal{G}$ is predictable with respect to $\mathcal{F}^c$. Show that independent censoring is satisfied. [3]

Solution By the innovation theorem, $\lambda^{\mathcal{F}^c}(t) = E[\lambda^{\mathcal{G}}(t)|\mathcal{F}^c_{t-}]$. Since $\lambda^{\mathcal{G}}$ is predictable with respect to $\mathcal{F}^c$, $\lambda^{\mathcal{G}}(t)$ is measurable with respect to $\mathcal{F}^c_{t-}$. Hence $E[\lambda^{\mathcal{G}}(t)|\mathcal{F}^c_{t-}] = \lambda^{\mathcal{G}}(t)$. Thus, $\lambda^{\mathcal{F}^c}(t) = \lambda^{\mathcal{G}}(t)$, i.e. independent censoring is satisfied.

[1] Of course, we may remove L_0 from the conditioning set if $L_0 = \varnothing$.

[2] See Fleming and Harrington (1991) for a proof.

[3] Hint: Use the innovation theorem and the fact that, if X is predictable with respect to $\mathcal{F}$, then $X(t)$ is measurable with respect to $\mathcal{F}(t-)$.

3. Suppose the intensity $\lambda^{\mathcal{G}}(t)$ with respect to $\mathcal{G}$ is $I(T \geq t)(2 - I(T^* \geq t))$.

   a) Sketch $\lambda^{\mathcal{G}}$ for the scenario $T^* < T$.

   b) Which of the following is true: The short-term risk of death for a censored individual is

      (i) higher than

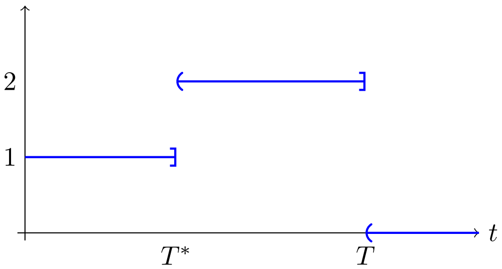
      (ii) the same as

      (iii) lower than

      the short-term risk of death among the subjects that are alive and not censored. Can you think of an example where this is the case?

   c) Show that independent censoring is not satisfied in this situation.

   Solution



   a)

   b) $\lambda(t)dt$ is heuristically the short-term risk of death from t to $t + dt$. We see that this risk is lower (i.e. 1) before T^*, and higher (i.e. 2) between T^* and T, hence the answer is (i). This could e.g. be the case if subjects receive extra care or closer monitoring during the follow-up period than they would otherwise.

   c) We calculate $\lambda^{\mathcal{F}^c}(t)$ using the innovation theorem. We have

$$\begin{aligned}\lambda^{\mathcal{F}^c}(t) &= E\big[\lambda^{\mathcal{G}}(t)|\mathcal{F}^c_{t-}\big]\\ &= E\big[I(T \geq t)(2 - I(T^* \geq t))|\mathcal{F}^c_{t-}\big]\\ &= I(T \geq t)\Big(2 - E\big[I(T^* \geq t)|\mathcal{F}^c_{t-}\big]\Big),\end{aligned}$$

   where we used the fact that $I(T \geq t)$ is predictable with respect to $\mathcal{F}^c$ to obtain the last line. Comparing the expressions for $\lambda^{\mathcal{G}}$ and $\lambda^{\mathcal{F}^c}$ we see that we have independent censoring if and only if

$$E\big[I(T^* \geq t)|\mathcal{F}^c_{t-}\big] = I(T^* \geq t) \text{ for all } t. \tag{4}$$

   However, (4) does not hold, since $I(T^* \geq t)$ is not predictable with respect to $\mathcal{F}^c$.

4. Let N be a counting process with jump times $T_1, T_2, \cdots$. Argue that the (Stieltjes) integral $\int_0^t H(s)dN(s)$ is equal to $\sum_{T_j \leq t} H(T_j)$. [4]

   Solution We use the definition $\int_0^t H(s)dN(s) = \sum\limits_{j: t_j \in [0,t]} H(t_j) \cdot (N(t_j) - N(t_j-))$, and make a few observations:

   - $N(t)$ is a right-continuous counting process, which means that $N(t) - N(t-) = 1$ if N jumps at t, and zero otherwise.
   - The t_j's that contribute to the sum is exactly the event times.
   - The function H is evaluated at the event times.

[4] Hint: Use the definition $\int_0^t H(s)dN(s) = \sum\limits_{j: t_j \in [0,t]} H(t_j) \cdot (N(t_j) - N(t_j-))$, where $N(t-) = \lim\limits_{s>0, s\to 0} N(t - s)$.

These three points lead to the result $\sum\limits_{j:t_j \in [0,t]} H(t_j) \cdot (N(t_j) - N(t_j-)) = \sum_{T_j \leq t} H(T_j)$.

5. Calculate the Nelson-Aalen estimator $\hat{H}(t)$, $t \geq 0$, for the data set below by hand.

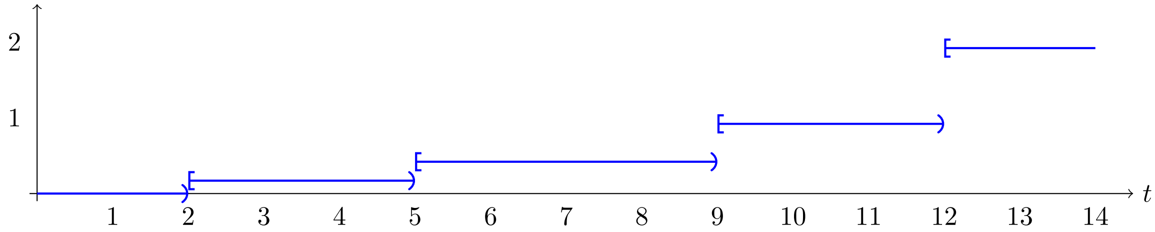
i	$\tilde{T}_i$	D_i
1	2	1
2	2.5	0
3	5	1
4	5.5	0
5	9	1
6	12	1

Draw the result, and use "(",")","[","]" to indicate the continuity properties of $\hat{H}$ at the jump times (here "[" at a point indicates continuous from the right at that point, "(" at a point indicates not continuous from the right at that point, "]" at a point indicates continuous from the left at that point, and ")" at a point indicates not continuous from the left at that point,).

Solution We augment the table to make the calculation easier:

i	$\tilde{T}_i$	D_i	$Z(\tilde{T}_i)$	$\Delta\hat{A}(\tilde{T}_i)$	$\hat{A}(\tilde{T}_i)$
1	2	1	6	1/6	0.17
2	2.5	0	5	0	0.17
3	5	1	4	1/4	0.42
4	5.5	0	3	0	0.42
5	9	1	2	1/2	0.92
6	12	1	1	1	1.92

We have now included a column $Z(T_i)$, which is the number at risk right before $\tilde{T}_i$, and a column $\Delta\hat{A}(\tilde{T}_i)$ which shows the increments of the Nelso-Aalen estimator at each $\tilde{T}_i$; when $D_i = 0$ there is no event, and the Nelson-Aalen estimator has an increment of 0, and when $D_i = 1$ at $\tilde{T}_i$ the Nelson-Aalen estimator has an increment of $1/Z(\tilde{T}_j)$. The rightmost column contains the Nelson-Aalen estimator, which is the cumulative sum of the column to the left of it (up to two decimal places). The plot is shown below:



6. Suppose we follow n individuals over a study period. We will now consider an estimator of the survival probability $P(T > t)$ as a function of t. To formulate the estimator, we introduce the variables $\{\tilde{T}_i, D_i\}_{i=1}^n$, where $D_i = 1$ if subject i dies in the study period (so that $\tilde{T}_i = T_i$) and $D_i = 0$ if subject i is censored at $\tilde{T}_i$(so that $T_i > \tilde{T}_i$). The estimator, which is called the Kaplan-Meier estimator, then takes the form[5]

$$\hat{S}(t) = \prod_{j:T_j \leq t, D_j=1} \Big(1 - \frac{1}{Z(T_j)}\Big),$$

[5] As in the lectures, we only consider the case without ties; the estimator looks slightly different if some event times are tied.

so that the product is over the observed failure times, and $Z(t) = \sum_{i=1}^{n} Z_i(t)$ is the number of individuals at risk (i.e. alive and not censored) just before t. Here, $Z_i(t)$ is 1 if subject i is at risk just before t, and 0 otherwise. [6]

a) Suppose there is no censoring, i.e. that all individuals are followed up over the entire study period. Show that then $\hat{S}(t) = 1 - \hat{F}(t)$, where $\hat{F}$ is the *empirical distribution function*

$$\hat{F}(t) = \frac{1}{n}\sum_{i=1}^{n} I(T_i \leq t).$$

b) A student (not enrolled in MATH-449 - Biostatistics) gets inspired by the relationship between the Kaplan-Meier estimator and the empirical distribution function. He reasons that, if he modifies the sample by just removing the subjects that are censored during the follow-up period, he can estimate the survival function by 1 minus the empirical distribution function of the modified sample. His proposed estimator is

$$\hat{S}^\star(t) = 1 - \hat{F}^\star(t),$$

where $\hat{F}^\star(t) = \frac{1}{n^\star}\sum_{i=1}^{n} I(T_i \leq t, D_i = 1)$, and $n^\star = \sum\limits_{i=1}^{n} I(D_i = 1)$.

Argue that the estimator $\hat{S}^\star$ will fail to estimate S in the presence of censoring, even if we have independent censoring.

Solution

a) When there is no censoring the estimator reduces to $\hat{S}(t) = \prod_{j:T_j \leq t}\left(1 - \frac{1}{Z(T_j)}\right)$. Now, order the event times such that $0 < T_1 < T_2 < \cdots$. Since we have no censoring, the process $Z(t)$ will jump with a length of 1 for every event time. Now, fix t, and let $T_{j'} = \max\{T_j : T_j \leq t\}$. The Kaplan-Meier estimator is thus a telescoping product, which reduces to

$$\begin{aligned}
\prod_{j:T_j \leq T_j'}\left(1 - \frac{1}{Z(T_j)}\right) &= \prod_{j:T_j \leq T_{j'}} \frac{Z(T_j) - 1}{Z(T_j)} \\
&= \frac{Z(T_1) - 1}{Z(T_1)} \cdot \frac{Z(T_2) - 1}{Z(T_2)} \cdot \frac{Z(T_3) - 1}{Z(T_3)} \cdots \frac{Z(T_{j'}) - 1}{Z(T_{j'})} \\
&= \frac{1}{Z(T_1)} \cancel{\frac{Z(T_1) - 1}{Z(T_2)}} \cdot \cancel{\frac{Z(T_2) - 1}{Z(T_3)}} \cdots \cancel{\frac{Z(T_{j'-1}) - 1}{Z(T_{j'})}} \big(Z(T_{j'}) - 1\big) \\
&= \frac{Z(T_{j'}) - 1}{Z(T_1)}.
\end{aligned}$$

Thus,

$$\begin{aligned}
1 - \hat{S}(t) &= 1 - \frac{Z(T_{j'}) - 1}{Z(T_1)} \\
&= \frac{Z(T_1) - Z(T_{j'}) + 1}{Z(T_1)}.
\end{aligned}$$

[6] In the lectures we will see that the Kaplan-Meier estimator is a consistent estimator under the independent censoring assumption. By consistent we mean that, for any $\epsilon > 0$, $\lim\limits_{n\to\infty} P\big(\sup\limits_{s \leq \tau} |\hat{S}(s) - S(s)| \geq \epsilon\big) = 0$, where τ is the end of the study period.

Now, $Z(t) = \sum_{i=1}^n Z_i(t) = \sum_{i=1}^n I(T_i \geq t)$ in the case of no censoring. We thus have that $Z(T_1) = n$, and

$$\begin{aligned} Z(T_1) - Z(T_{j'}) + 1 &= n - \sum_{i=1}^n I(T_i \geq T_{j'}) + 1 \\ &= n - \sum_{i=1}^n \Big(1 - I(T_i < T_{j'})\Big) + 1 \\ &= \sum_{i=1}^n I(T_i < T_{j'}) + 1 \\ &= \sum_{i=1}^n I(T_i \leq T_{j'}) \\ &= \sum_{i=1}^n I(T_i \leq t). \end{aligned}$$

We thus have that

$$\begin{aligned} 1 - \hat{S}(t) &= \frac{Z(T_1) - Z(T_{j'}) + 1}{Z(T_1)} \\ &= \frac{\sum_{i=1}^n I(T_i \leq t)}{n} \\ &= \hat{F}(t). \end{aligned}$$

b) Clearly enough, removing individuals that are observed in parts of the study period, but that did not experience events while under observation, will not provide a coherent estimation strategy. After all, the subjects that were observed in parts of the study period without dying provide useful information regarding the survival experience; they could in principle have died while under observation, but didn't. Moreover, there is no guarantee that all subjects in a given study will die before the end of the study (this typically doesn't happen), but the estimator $\hat{S}^\star$ is defined for such a population.

More rigorously, we may consider the limiting distribution of $\hat{F}^\star$:

$$\begin{aligned} \lim_{n\to\infty} \hat{F}^\star(t) &= \lim_{n\to\infty} \frac{1}{n^\star} \sum_{i=1}^n I(T_i \leq t, D_i = 1) \\ & \lim_{n\to\infty} \frac{1/n}{n^\star/n} \sum_{i=1}^n I(T_i \leq t, D_i = 1) \end{aligned}$$

Now, $n^\star/n = \frac{1}{n}\sum\limits_{i=1}^n I(D_i = 1) \to P(D = 1)$, and $\frac{1}{n}\sum_{i=1}^n I(T \leq t, D = 1) \to P(T \leq t, D = 1)$ (using the law of large numbers, assuming independent observations), which gives

$$\begin{aligned} \lim_{n\to\infty} \hat{F}^\star &= \frac{P(T \leq t, D = 1)}{P(D = 1)} \\ &= P(T \leq t | D = 1), \end{aligned}$$

and thus $\hat{S}^\star(t) = 1 - \hat{F}^\star(t)$ estimates $P(T \geq t | D = 1) = P(T \geq t | T < \tilde{T})$, the survival function among the subjects that eventually died during the study, which is not what we are interested in.

7. Prove the following result:

Theorem 1 (identification under independent censoring) *Under independent censoring, the intensity of the right-censored counting process N_i can be written as*

$$\lambda_i(t)dt = Z_i(t)\alpha_i(t)dt$$

where $Z_i(t) = I(t \leq \tilde{T}_i)$ and α_i is the hazard of the "complete" counting process

$$\lambda_i^c(t)dt = Z_i^c(t)\alpha_i(t)dt$$

where $Z_i^c(t) = I(t \leq T_i)$.

As a hint: use the Innovation theorem:

Theorem 2 (Innovation theorem) *An intensity $\lambda_i^{\mathcal{F}''}(t)$ with respect to a filtration $\{\mathcal{F}_t''\}$ such that $\{\mathcal{F}_t'\} \supseteq \{\mathcal{F}_t''\}$, satisfies*

$$\lambda_i^{\mathcal{F}''}(t) = \mathbb{E}(\lambda_i^{\mathcal{F}'}(t) \mid \mathcal{F}_{t-}'').$$

Solution See slide 145 in the lecture notes.