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

Consider the following filtrations from the lectures:

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

where $C(t) = I(t \le T^*)$ and T^* is the censoring time, $N^c(t) = I(t \ge T)$ and T is the event time, $N(t) = I(t \ge \tilde{T}, D = 1)$ with $\tilde{T} = \min(T, T^*)$ and $D = I(T < T^*)$, and $Z(t) = I(t \le \tilde{T})$. Thus, $\mathcal{F}_t \subseteq \mathcal{G}_t \supseteq \mathcal{F}_t^c$. 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

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

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]

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

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

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

   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.

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]

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

   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?

---

[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 $\mathcal{F}(t-)$.

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

4. 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,).

5. 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[4]

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

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. [5]

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.

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

[5] 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 $\epsilon > 0$, $\lim_{n\to\infty} P\big(\sup_{s\leq\tau} |\hat{S}(s) - S(s)| \geq \epsilon\big) = 0$, where τ is the end of the study period.

6. 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*

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