MATH-449 - Biostatistics
EPFL, Spring 2021
Problem Set 2

1. Let N be an nonhomogeneous Poisson processes with deterministic intensity function $\alpha(t)$. Define $A(t) = \int_0^t \alpha(s)ds$. The following three points i)-ii) provide equivalent definitions of such a process:

   i)
   - $N(t) - N(s) \sim \text{Poisson}(A(t) - A(s))$ for $s < t$
   - $N(t) - N(s)$ is independent of $\mathcal{F}_s$ for $s < t$

   ii)

   $$P(N_{t+\delta} - N_t = 1|\mathcal{F}_t) = \alpha(t)\delta + o(\delta^2)$$
   $$P(N_{t+\delta} - N_t = 0|\mathcal{F}_t) = 1 - \alpha(t)\delta + o(\delta^2)$$

   as $\delta \longrightarrow 0^+$.

   Here, $\mathcal{F}$ is the filtration generated by N.

   a) Show that $M(t) = N(t) - A(t)$ is a martingale with respect to $\mathcal{F}$.*

   b) Show that the increments of M are uncorrelated, i.e. that, for $v \leq u \leq s \leq t$, †

   $$E[(M(t) - M(s))(M(u) - M(v))] = 0.$$

   Suppose that N is only recorded up to the (deterministic) time X, and define $N^\star(t) = N(\min\{t, X\})$. Thus, $N^\star$ is censored at X.

   c) Argue that $N^\star(t)$ is the observed number of jumps of N up to time t, and demonstrate that $N^\star$ satisfies the multiplicative intensity model‡.

   d) Suppose now that X is a random variable. Verify that the conclusion in c) holds when $\{X \leq t\} \in \mathcal{F}_t$ for each t, or equivalently, that $I(X \leq \cdot)$ is adapted to $\mathcal{F}$.§

2. In this problem we will use the definition of the optional variation process $[\cdot]$ from the lecture notes. Thus, we will need to take limits $[H](t) = \lim_{n\to\infty} \sum_{k=1}^n \big(H(kt/n) - H((k-1)t/n)\big)^2$ (in probability) of processes H. Let $\{N(t) : t \in [0, \tau]\}$ be a counting process.

   a) Show that the optional variation process $[N]$ is equal to N.

   Let λ be the intensity of N with respect to some filtration $\mathcal{F}$, so that $\Lambda(t) = \int_0^t \lambda(s)ds$ is the cumulative intensity, and $M = N - \Lambda$ is a martingale with respect to $\mathcal{F}$. Assume that $\int_0^\tau \lambda(s)^2 ds \leq K$ for some constant K.

   b) Show that $[M] = N$.¶

   Let H be a predictable process such that $\int_0^\tau H(s)^2\lambda(s)^2 ds \leq K'$ for some constant K'.

   c) Show that $\big[\int_0^\cdot H(s)dM(s)\big](t) = \int_0^t H(s)^2 dN(s)$

   d) What are the compensators of $[M]$ and $\big[\int_0^\cdot H(s)dM(s)\big]$?

*Hint: A Poisson distributed variable with parameter $\lambda > 0$ has mean λ.

†Note: this is true for any martingale M, not just the one from a).

‡Hint: start with definition ii).

§X is then called a *stopping time* with respect to $\mathcal{F}$. Heuristically, $\mathcal{F}_t$ contains enough information to determine whether X has occurred by t.

¶Hint: Use the inequality $\big(\int_a^b f(s)ds\big)^2 \leq (b-a)\int_a^b f(s)^2 ds$.

3. Suppose we follow n individuals over a study period. You will later learn about the *Kaplan-Meier estimator*[*], which is an estimator of the survival probability $P(T > t)$ as a function of t. The estimator takes the form[†]

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

where the product is over the observed failure times, and where $Z(t) = \sum_{i=1}^{n} Z_i(t)$ is the number of individuals at risk (i.e. alive and not censored) just before t, so that $Z_i(t)$ is 1 if subject i is at risk just before t, and 0 otherwise. In the lectures we will see that the Kaplan-Meier estimator is a consistent estimator under the independent censoring assumption.[‡]

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) The *Greenwood estimator*

$$\hat{\sigma}(t)^2 = \hat{S}(t)^2 \int_0^t \frac{dN(s)}{Z(s)(Z(s) - 1)}$$

estimates the variance (for each fixed t) of the Kaplan-Meier estimator. Show that, in the absence of censoring,

$$\hat{\sigma}(t)^2 = \frac{\hat{S}(t)(1 - \hat{S}(t))}{n},$$

which is the maximum likelihood estimator for the Bernoulli probability $S(t)$.

c) 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. To formulate his 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 and $D_i = 0$ otherwise, so that

$$\tilde{T}_i = T_i \text{ if } D_i = 1,$$
$$\tilde{T}_i < T_i \text{ if } D_i = 0,$$

he suggests to estimate the survival function using

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

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

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

---

[*] Kaplan and Meiers paper from 1958 remains the most cited statistics paper in history.

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

[‡] 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.

4. A statistics student did an internship with a power company, where she was hired to analyze the time it took before cracks developed in the company's new turbine prototype. 41 turbines were observed in a testing facility for two months, where engineers had carefully recorded the time it took before noticeable cracks were discovered. The power company had installed machines that could make turbines rotate to simulate the rotation they could experience in a real-world environment.

   The following outcomes were recorded:

   - Some of the turbines developed cracks before the two months were over.
   - Some turbines were observed for the whole two months without any cracks.
   - For a significant number of turbines, the machines that enforced the rotation stopped working during the study. The power company didn't have the resources to repair these machines, making the turbines they rotated unobserved.

   The power company was interested in the time it took before cracks developed if the rotation enforcing machines did not stop. The turbines whose machines stopped were considered censored. The time it took for the machines to fail was thought to be unrelated to the time it took before cracks developed so that the observed (non-censored) turbines were representative of all turbines.

   a) Classify the above outcomes of turbine i using the variables $\tilde{T}_i$ and D_i from the lectures.

      Being familiar with survival analysis, the student calculated the Kaplan-Meier curve, which estimates the survival probability as a function of t. The estimate along with approximate 95% confidence intervals (for each fixed t), $\hat{S}(t) \pm 1.96\hat{\sigma}(t)$, is plotted in Figure 1

   b) Based on the plot, find the probability of a turbine being crack-free after 30 days, with a 95% confidence interval. Use the plot to estimate the 10th percentile of the survival times, along with a 95% confidence interval.

      The Kaplan-Meier estimator estimates the true survival probability $P(T > t)$ as long as the censoring is independent.

   c) Given the information provided thus far in this example, argue that the censoring is independent.

      After talking with some of the engineers, the student learned that the machines provided different rotational speeds to the turbines. She reasoned that the machines that provided higher rotational speed: 1) could make cracks appear faster due to increased stress on the turbines, and 2) were more likely to stop working during the study (thus leading to censoring) due to increased stress on the machines. She learned that the machines could broadly be categorised into two groups; those that provided fast rotational speed and those that provided slow rotational speed.

      After thinking about the problem for a bit, she realised that the result from b) could provide a misleading picture of the survival probability. However, she also found that she could use the extra information about the machines' rotation speeds to improve her statistical analysis.

   d) Can you guess what she did?[†]

[†]Hint: what do we know about the censoring?

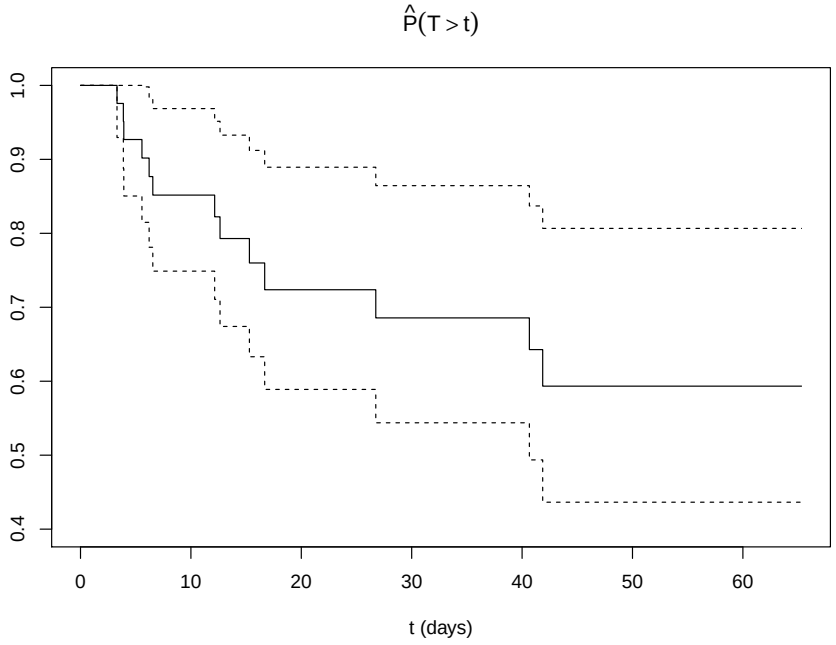

Figure 1: