# BIOSTAT/STAT 576
# Statistical Methods for Survival Data

## Problem Set 2
## Due: ~~April 9, 2007~~

1. Assume that $T_i \sim \exp(\lambda)$ and $C_i \mid T_i = t \sim \exp(\mu + \gamma t)$, for $i = 1, 2, \ldots, n$. Let $X_i = \min(T_i, C_i)$ and $\Delta_i = I(T_i \leq C_i)$. In addition, let $N_i(t) = I(X_i \leq t, \Delta_i = 1)$ and $Y_i(t) = I(X_i \geq t)$. Consider the Nelson-Aalen estimator of

$$\hat{\Lambda}(t) = \int_0^t \frac{dN(u)}{Y(u)}.$$

   (a) Calculate $\Lambda^*(t)$ such that $\hat{\Lambda}(t) \to_P \Lambda^*(t)$, as $n \to \infty$.

   (b) Plot $\Lambda^*(t)$ and $\Lambda_T(t) = \lambda t$.

2. Assume that the observed data consist of $n$ i.i.d. copies of $(X_i, \Delta_i)$, where $X_i = \min(T_i, C_i)$ and $\Delta_i = I(T_i \leq C_i)$, $i = 1, 2, \ldots, n$. $T_i$ and $C_i$ are continuous. Let $N_i(t) = I(X_i \leq t, \Delta_i = 1)$ and $N_i^C(t) = I(X_i \leq t, \Delta_i = 0)$. Consider the filtration defined by

$$\mathcal{F}_t = \sigma\{N_i(s), 0 \leq s \leq t; N_i^C(s), s \geq 0; i = 1, 2, \ldots, n\}$$

   (a) Show that $E[dN_i(t) \mid \mathcal{F}_{t-}] = [1 - N_i(t-) - N_i^C(\infty)]dF_1^s(t)/S_1^s(t)$, where $F_1^s(t) = \Pr\{X \leq t, \Delta = 1\}$ and $S_1^s(t) = \Pr\{X > t, \Delta = 1\}$. Find the compensator $A(t)$ for $N(t)$.

   (b) Let $M(t) = N(t) - A(t)$ and $U_n(t) = n^{-3/2} \int_0^t Y(u)dM(u)$, where $Y(t) = \sum_{i=1}^n I(X_i \geq t)$. Establish the weak convergence of $U_n(t) \Rightarrow U(t)$, and calculate $v(t)$ such that $< U_n, U_n > (t) \to_P v(t)$.

3. This is a small-scale simulation study on the validity and efficiency of the Kaplan-Meier estimator. You can choose any computer program you prefer. Attach your program codes. Follow this algorithm to conduct your simulation study:

   For the $i$th simulation, $i = 1, 2, \ldots, m$,

   (a) Simulate $n$ i.i.d copies of $T \sim \exp(1)$;

   (b) Simulate $n$ i.i.d. copies of $C \sim \exp(\theta)$, where $\theta$ is selected such that the censoring proportion $p$ to be $p = \Pr\{T > C\}$;

(c) Write your own codes to calculate both the MLE and the Kaplan-Meier estimate of $\hat{S}_i(t)$, their associated variances, standard errors and 95% confidence intervals, respectively;

(d) reiterate step (a)-(c).

Complete the following table from your simulation study for $m = 1000$.

| $n$ | $p$ | $t$ | $\hat{S}(t)$ | Bias | Cov. Prob. | SSE | MSE |
|---|---|---|---|---|---|---|---|
| 50 | 10% | 0.5 | MLE | | | | |
| | | | KM | | | | |
| | | 1.0 | MLE | | | | |
| | | | KM | | | | |
| | 30% | 0.5 | MLE | | | | |
| | | | KM | | | | |
| | | 1.0 | MLE | | | | |
| | | | KM | | | | |
| 100 | 10% | 0.5 | MLE | | | | |
| | | | KM | | | | |
| | | 1.0 | MLE | | | | |
| | | | KM | | | | |
| | 30% | 0.5 | MLE | | | | |
| | | | KM | | | | |
| | | 1.0 | MLE | | | | |
| | | | KM | | | | |

Let $S(t)$ be the true value. In this table,

- bias: $\sum_{i=1}^{m}[\hat{S}_i(t) - S(t)]/m$;
- cov. prob.: 95% nominal coverage probability $\sum_{i=1}^{m} I\{S(t) \in 95\% \text{ CI}\}]/m$;
- SSE (sample standard error): $\{\sum_{i=1}^{m}\{\hat{S}_i(t) - \bar{\hat{S}}(t)\}]^2/(m-1)\}^{1/2}$;
- MSE (mean standard error): $\sum_{i=1}^{m} \text{se}\{\hat{S}(t)\}/m$.

Summarize any findings that you notice in this table.