Homework 2

Steven Chiou

Due date: Thursday, October 11

1. **Show that (algebraically) in the absence of censoring $\hat{S}_{\text{KM}}(t) = \hat{S}_{\mathbf{e}}(t)$.**

   Without loss of generality, we assume $t_1 < t_2 < \ldots < t_n$ and $\Delta_i = 1$ for $i = 1, \ldots, n$. Under the assumption of no ties, we have $d_1 = \ldots = d_n = 1$ and $n_1 = n$, $n_2 = n-1, \ldots, n_i = n - i - 1$.

$$\begin{aligned}
\hat{S}_{km}(t) &= \prod_{i:t_i let} \left(1 - \frac{d_i}{n_i}\right) = \prod_{i:t_i<t} \left(\frac{n_i - d_i}{n_i}\right) \\
&= \frac{n-1}{n} \times \frac{n-2}{n-1} \times \ldots \times \frac{n-i}{n-i-1} \\
&= \frac{n-i}{n},
\end{aligned}$$

   which is the expression for the empirical survival estimator, $\hat{S}_{\text{e}}(t)$.

2. **In the absence of censoring, show that the Greenwood Formula (page 30 on note 2) can be reduced to**

$$\frac{\hat{S}_{\text{KM}}(t) \times \{1 - \hat{S}_{\text{KM}}(t)\}}{n}.$$

   You might assume there are no ties among the observations.

   Continue with the assumption outlined in #1, we can simplify the Greenwood formula as follows.

$$\begin{aligned}
\hat{S}^2_{km}(t) \cdot \sum_{i:t_i \leq t} \frac{d_i}{n_i(n_i - d_i)} &= \left(\frac{n-i}{n}\right)^2 \cdot \left(\frac{1}{n(n-1)} - \frac{1}{(n-1)(n-2)} - \ldots - \frac{1}{(n-i-1)(n-i)}\right) \\
&= \left(\frac{n-i}{n}\right)^2 \cdot \left(\frac{1}{n-i} - \frac{1}{n}\right) \\
&= \left(\frac{n-i}{n}\right)^2 \cdot \frac{i}{n(n-i)} = \frac{i(n-i)}{n^3} = \frac{\hat{S}_{\text{KM}}(t) \times \{1 - \hat{S}_{\text{KM}}(t)\}}{n}.
\end{aligned}$$

3. **Consider the Leukemia data from the `survival` package:**

```
> library(survival)
> head(aml)
```

```
  time status          x
1    9      1 Maintained
2   13      1 Maintained
3   13      0 Maintained
4   18      1 Maintained
5   23      1 Maintained
6   28      0 Maintained
```

In here, each row represent one patient. `aml` is the observed survival time, `status` is the censoring indicator (1 = event, 0 = censored), and `x` is the treatment indicator. We will ignore the treatment indicator for now.

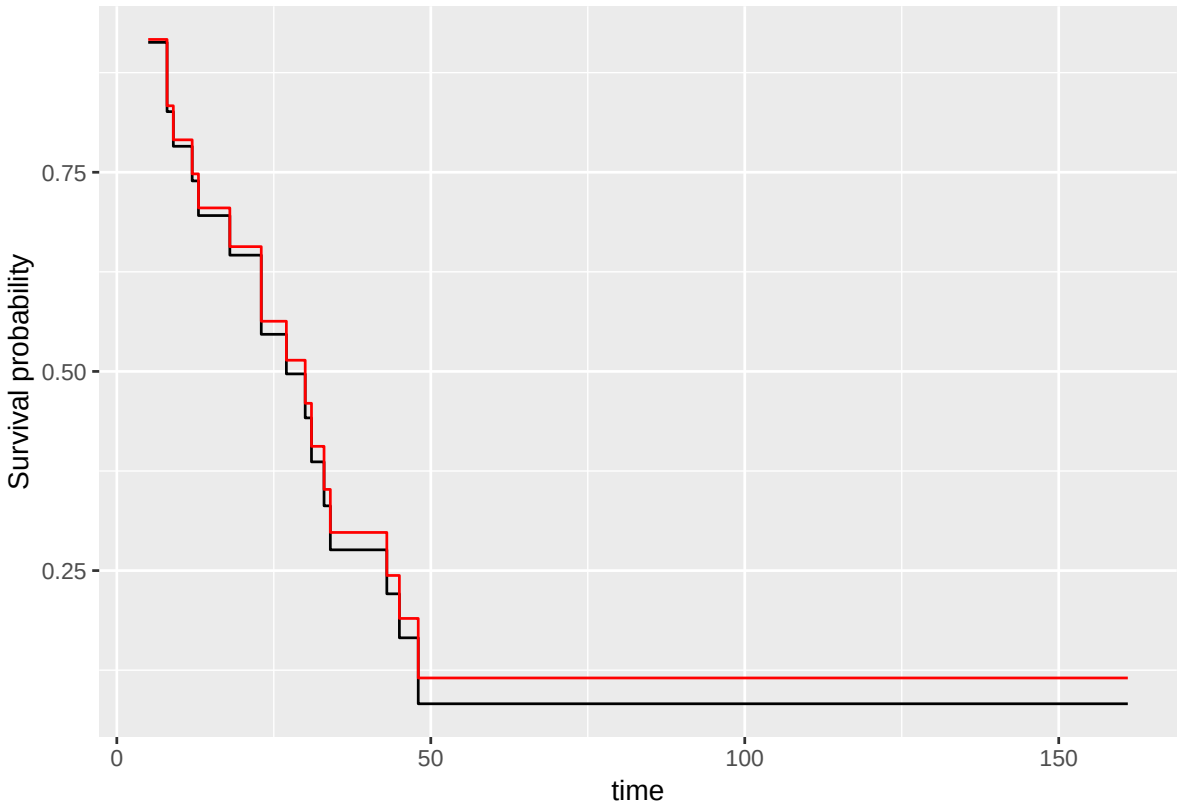
   a. **Plot the Kaplan-Meier survival curve for the data.**
   b. **Add the Nelson-Aalen survival curve to the Kaplan-Meier plot from** (3a)**.**

Load libraries:

```
> library(tidyverse)
> library(survival)
```

The following gives both the Kaplan-Meier survival curve (black) and the Nelson-Aalen survival curve (red).

```
> dat <- aml %>% arrange(time) %>% select(-x) %>% group_by(time) %>%
+   summarize(di = sum(status), ni = length(status)) %>%
+   mutate(ni = rev(cumsum(rev(ni))), KM = cumprod(1 - di / ni), Na = cumsum(di / ni))
> ggplot(data = dat, aes(x = time)) +
+   geom_step(aes(y = KM), show.legend = TRUE) +
+   geom_step(aes(y = exp(-Na)), col = 2) + ylab("Survival probability")
```



4. **The expected survival time for the Leukemia data in #(3) does not exist because the last observation is a censored event. An alternative is to lookInstead of looking at the expected survival time, an alternative is to look at the restricted mean survival time. Compute $E(T|T < 161)$ based on the survival curve in (3a).**

```
> sum(diff(c(0, dat$time)) * c(1, dat$KM[-18]))
```

```
[1] 36.36439
```

5. **Let $N_i(t)$ be the number of events over time interval $(0, t]$ for the ith patient in #(3). Let $N(t) = \sum_{i=1}^{n} N_i(t)$ be the aggregated counting process.**

   a. **Plot $N(t)$.**
   b. **Plot $M(t)$, where $M(t) = N(t) - \hat{H}(t)$ and $\hat{H}(t)$ is the Nelson-Aalen estimator for the cumulative hazard function.**

   Omit.