Homework 2

Tian Jiang

Due date: Thursday, October 9

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

$$\begin{aligned}
\hat{S}_{\text{KM}}(t) &= \prod_{t_{(i)} \le t} \frac{n_i - d_i}{n_i} \\
&= \frac{n_1 - d_1}{n_1} \times \frac{n_2 - d_2}{n_2} \times \cdots \times \frac{n_{\max\{i:t_{(i)} \le t\}} - d_{\max\{i:t_{(i)} \le t\}}}{n_{\max\{i:t_{(i)} \le t\}}} \\
&= \frac{n_1 - d_1}{n_1} \times \frac{n_1 - d_1 - d_2}{n_1 - d_1} \times \cdots \times \frac{n_1 - \sum_{t_{(i)} \le t} d_i}{n_1 - \sum_{t_{(i+1)} \le t} d_i} \\
&= \frac{n_1 - \sum_{t_{(i)} \le t} d_i}{n_1} = \frac{n - \sum_{t_{(i)} \le t} d_i}{n} \\
&= \frac{\#\text{ individuals with survival times} \ge t}{\#\text{ individuals in the data set}} = \hat{S}_{\text{e}}(t)
\end{aligned}$$

   where $t_{(i)}$'s are ordered failure times, n_i is the number at risk at $t_{(i)}$ and d_i is the number of observed failures.

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.

$$\begin{aligned}
\hat{Var}\left(\hat{S}_{\text{KM}}(t)\right) &= (\hat{S}_{\text{KM}}(t))^2 \sum_{t_{(i)} \le t} \frac{d_i}{n_i(n_i - d_i)} = \left(\hat{S}_{\text{KM}}(t)\right)^2 \sum_{t_{(i)} \le t} \left(\frac{1}{n_i - d_i} - \frac{1}{n_i}\right) \\
&= \left(\hat{S}_{\text{KM}}(t)\right)^2 \left(\frac{1}{n_1 - d_1} - \frac{1}{n_1} + \frac{1}{n_2 - d_2} - \frac{1}{n_2} + \cdots + \frac{1}{n_{\max\{i:t_{(i)} \le t\}} - d_{\max\{i:t_{(i)} \le t\}}} - \frac{1}{n_{\max\{i:t_{(i)} \le t\}}}\right) \\
&= \left(\hat{S}_{\text{KM}}(t)\right)^2 \left(\frac{1}{n_1 - d_1} - \frac{1}{n_1} + \frac{1}{n_1 - d_1 - d_2} - \frac{1}{n_1 - d_1} + \cdots + \frac{1}{n_1 - \sum_{t_{(i)} \le t} d_i} - \frac{1}{n_1 - \sum_{t_{(i+1)} \le t} d_i}\right) \\
&= \left(\hat{S}_{\text{KM}}(t)\right)^2 \left(\frac{1}{n_1 - \sum_{t_{(i)} \le t} d_i} - \frac{1}{n_1}\right) = \left(\hat{S}_{\text{KM}}(t)\right)^2 \frac{\sum_{t_{(i)} \le t} d_i}{n_1(n_1 - \sum_{t_{(i)} \le t} d_i)} = \left(\frac{n - \sum_{t_{(i)} \le t} d_i}{n}\right)^2 \frac{\sum_{t_{(i)} \le t} d_i}{n(n - \sum_{t_{(i)} \le t} d_i)} \\
&= \frac{1}{n}\left(\frac{n - \sum_{t_{(i)} \le t} d_i}{n}\right)\left(\frac{\sum_{t_{(i)} \le t} d_i}{n}\right) = \frac{1}{n}\left(\frac{n - \sum_{t_{(i)} \le t} d_i}{n}\right)\left(1 - \frac{n - \sum_{t_{(i)} \le t} d_i}{n}\right) \\
&= \frac{1}{n}\hat{S}_{\text{KM}}(t)\left(1 - \hat{S}_{\text{KM}}(t)\right)
\end{aligned}$$

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

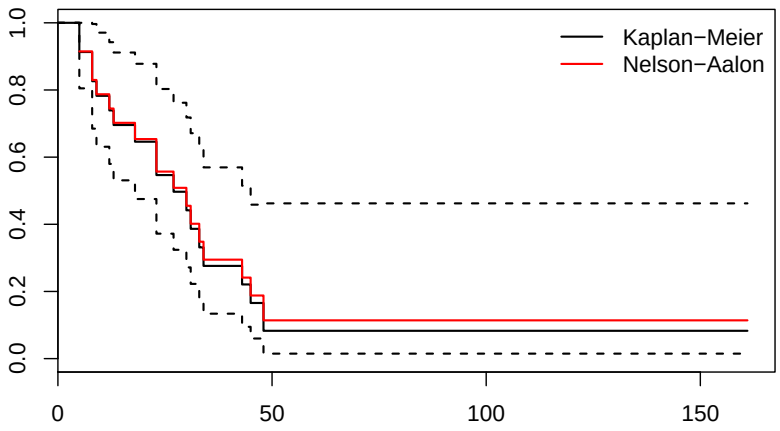
```
> library(survival)
> head(aml)
```

```
  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. `time` 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).

```
>   # Kaplan-Meier
>   km <- survfit(Surv(time, status) ~ 1, data = aml)
>
>   # Nelson-Aalon
>   cox <- coxph(Surv(time, status) ~ 1, data = aml)
>   H0 <- basehaz(cox)
>   plot(km, lwd = 1.5)
>   lines(H0$time, exp(-H0$hazard), col = 2, 's', lwd = 1.5)
>   legend("topright", bty = "n", lty = 1, lwd = 1.5, col = 1:2,
+          c("Kaplan-Meier", "Nelson-Aalon"))
```



4. The expected survival time for the Leukemia data in #(3) does not exist because the last observation is a censored event. Instead 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).

```
> dat <- tibble(T = km$time, surv = km$surv)
> dat %>% add_row(T = 0, surv = 1,.before = 1) %>% mutate(diff = lead(T) -T) %>%
+   mutate(prod = cumsum(surv*diff)) %>%
+   summarise('Expected Survival Time' = max(prod, na.rm = TRUE))
```

```
# A tibble: 1 x 1
  `Expected Survival Time`
                     <dbl>
1                     36.4
```

Or, use the restricted mean computed by `survfit` function

```
> print(km,print.rmean=getOption("survfit.print.rmean"), rmean = "individual")
```

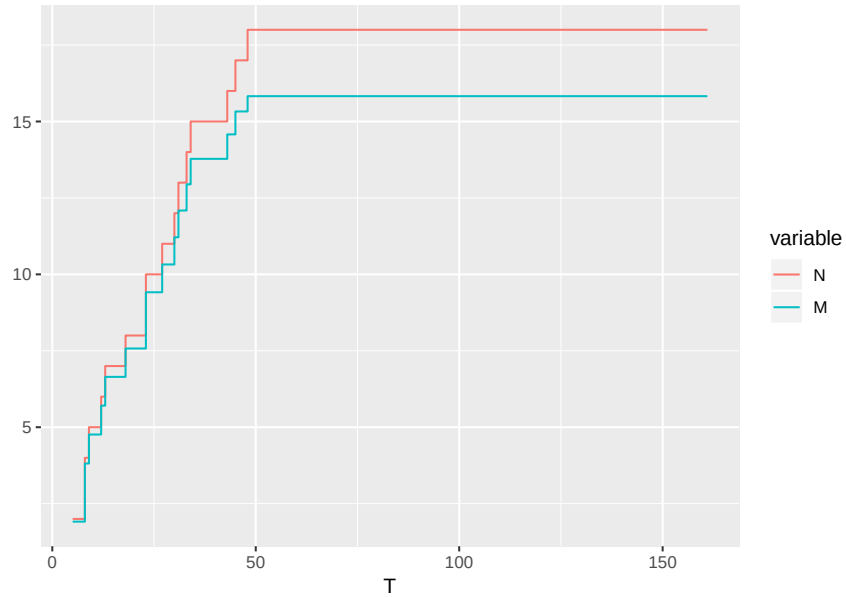
```
Call: survfit(formula = Surv(time, status) ~ 1, data = aml)

        n     events     *rmean *se(rmean)     median    0.95LCL
    23.00      18.00      36.36       9.85      27.00      18.00
  0.95UCL
    45.00
  * restricted mean with variable upper limit
```

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.

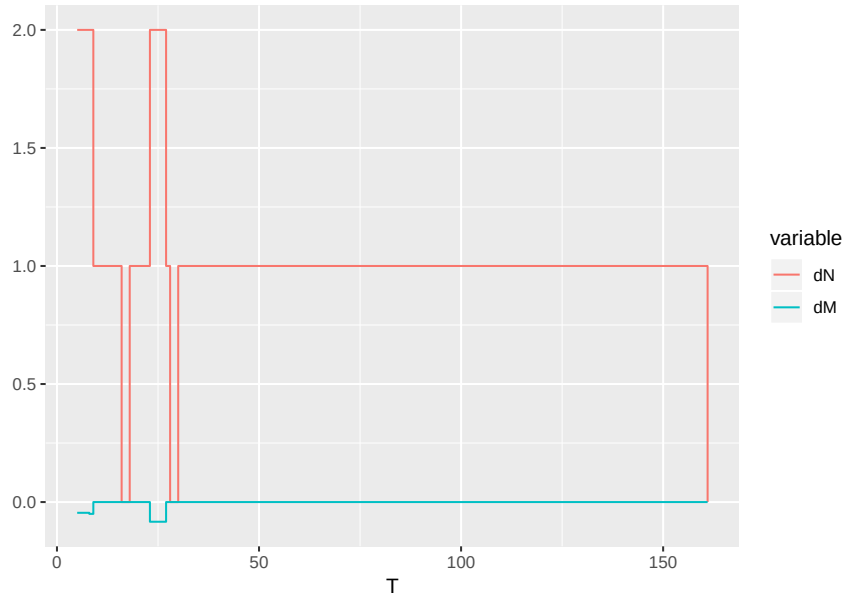
   a. Plot $N(t)$.
   b. Plot $M(t)$, where $M(t) = N(t) - \hat{H}(t)$.

```
> tibble(T = km$time, dN = km$n.event, H = H0$hazard) %>%
+   mutate(N = cumsum(dN), M = N - H) %>% select(T, N, M) %>% melt(id=c("T")) %>%
+   ggplot + geom_step(aes(x = T, y = value, col = variable)) + ylab("")
```



$dM(t) = dN(t) - h(t)Y(t)dt$

```
> tibble(T = km$time, dN = km$n.event, Y = km$n.risk, H = H0$hazard) %>%
+   add_row(T = 0, H = 0,.before = 1) %>%
+   mutate(h = lag(lead(H) - H), dM = dN - h*Y) %>%
+   filter(T != 0) %>% select(T, dN, dM) %>% melt(id=c("T")) %>%
+   ggplot + geom_step(aes(x = T, y = value, col = variable)) + ylab("")
```



$\widehat{dM}$ curve is almost the horizon since we use the estimated hazard. dM is a martingale difference sequence.