

Homework 2

Tian Jiang

Due date: Thursday, October 9

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

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

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.

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

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

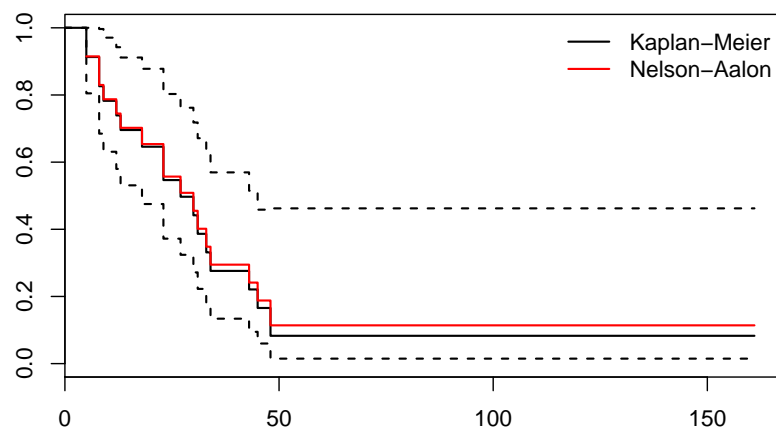
```
> library(survival)
> head(aml)
```

	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.

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

```
> # Kaplan-Meier
> km <- survfit(Surv(time, status) ~ 1, data = aml)
>
> # Nelson-Aalen
> 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-Aalen"))
```



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('E[T|T<161]' = max(prod, na.rm = TRUE))
```

```
# A tibble: 1 x 1
  `E[T|T<161]`
    <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)`

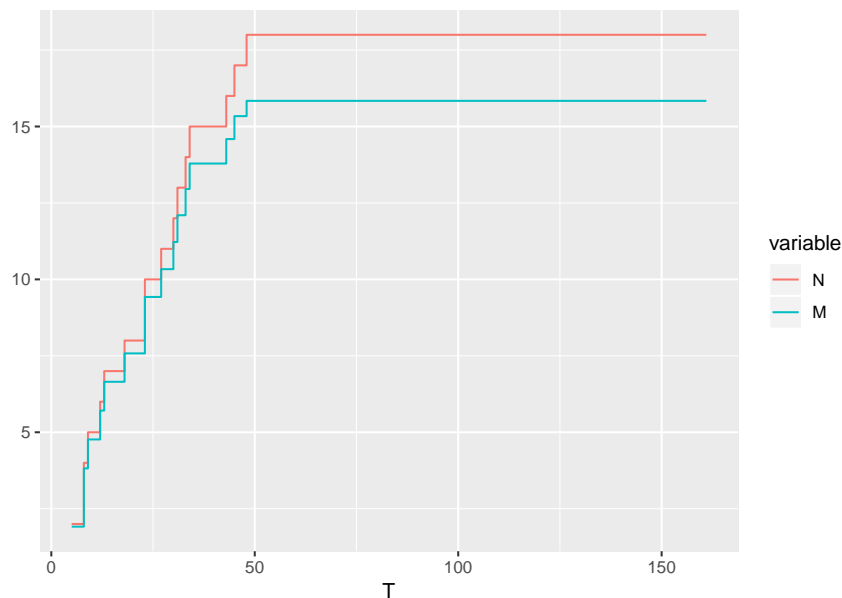
	n	events	*rmean	*se(rmean)	median	0.95LCL
	23.00	18.00	36.36	9.85	27.00	18.00
0.95UCL						
	45.00					

* restricted mean with variable upper limit

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

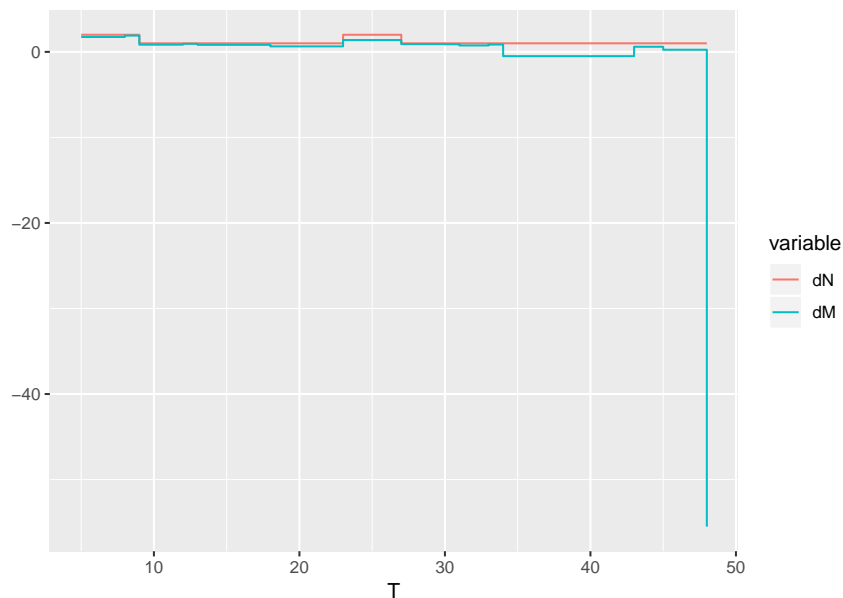
- Plot $N(t)$.
- Plot $M(t)$, where $M(t) = N(t) - \hat{H}(t)$.

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



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

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



Zoom in for the part before the last observed event time ($t=48$)

```
> dat %>% mutate(dM = replace(dM, T == 48, 0.25)) %>% melt(id=c("T")) %>%
+   ggplot() + geom_step(aes(x = T, y = value, col = variable)) +
+   geom_point(aes(x = 48, y = 0.25), shape = 1, size = 1.5) +
+   geom_point(aes(x = 48, y = 1), shape = 1, size = 1.5) + ylab("")
```

