Homework 2

Tianjian Shi

Due date: Thursday, October 11

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

$$\hat{S}_{\text{KM}}(t) = \prod_{t_{(i)} \le t} \frac{n_i - d_i}{n_i} = \frac{n_1 - d_1}{n_1} \cdot \frac{n_2 - d_2}{n_2} \cdot \frac{n_3 - d_3}{n_3} \cdots \frac{n_t - d_t}{n_t} = \frac{n_1 - d_1}{n_1} \cdot \frac{n_2 - d_2}{n_1 - d_1} \cdot \frac{n_3 - d_3}{n_2 - d_2} \cdots \frac{n_t - d_t}{n_{t-1} - d_{t-1}}$$
$$= \frac{n_t - d_t}{n_1} = \frac{n_1 - \sum_{t_{(i)} \le t} d_i}{n_1} = \frac{n_{t+1}}{n_1} = \hat{S}_{\text{e}}(t)$$

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.

$$Var(\hat{S}_{\text{KM}}(t)) = Var(\frac{n_{t+1}}{n}) = \frac{Var(n_{t+1})}{n^2} = \frac{n \cdot (1 - \frac{n_{t+1}}{n}) \cdot \frac{n_{t+1}}{n}}{n^2} = \frac{\hat{S}_{\text{KM}}(t) \times \{1 - \hat{S}_{\text{KM}}(t)\}}{n}$$

Since $Var(n_{t+1})$ follows binomial distribution with parameters $(1 - \frac{n_{t+1}}{n})$ and $(\frac{n_{t+1}}{n})$.

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

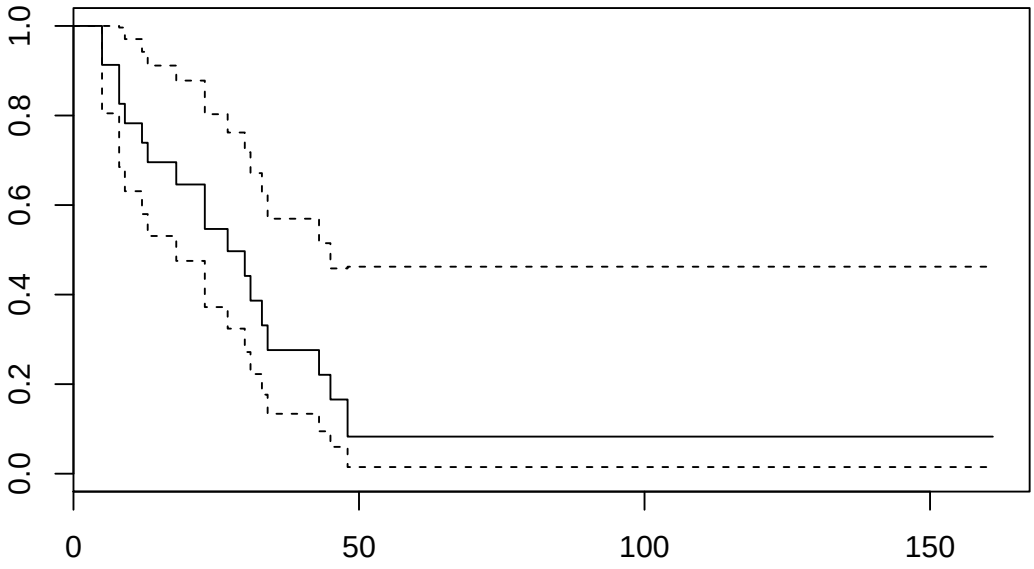
In here, each row represent one patient. `aml` is the observed survival time, `status` is the censoring indicator (1 = event, 0 = censored), and `x` is the treatment indicator. We will ignore the treatment indicator for now.

   a. Plot the Kaplan-Meier survival curve for the data.

```
> library(survival)
```

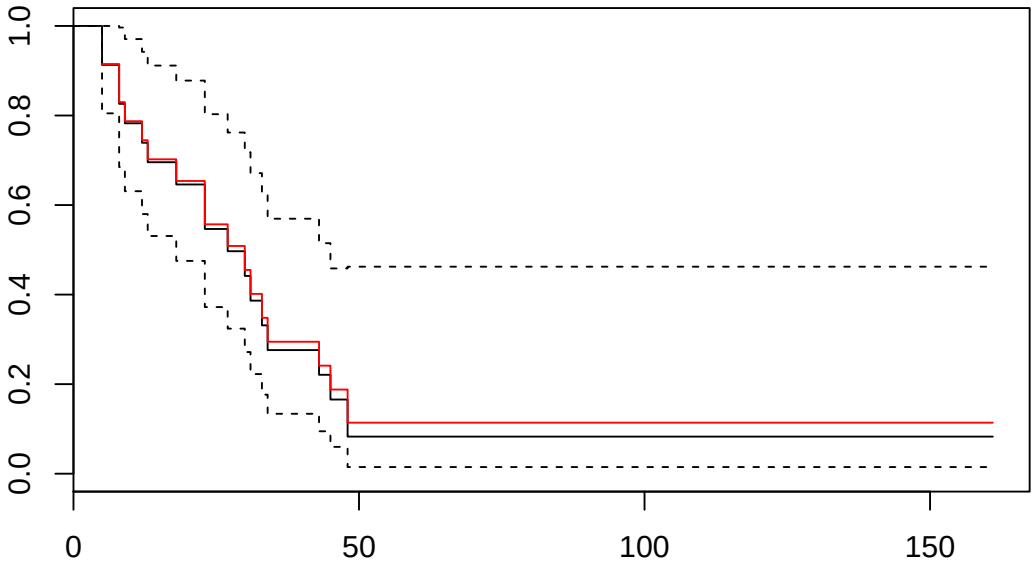
```
Warning: package 'survival' was built under R version 3.4.4
```

```
> km <- survfit(Surv(time, status) ~ 1, dat = aml)
> plot(km)
```



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

```
> cox <- coxph(Surv(time, status) ~ 1, dat = aml)
> H0 <- basehaz(cox)
> plot(km)
> lines(H0$time, exp(-H0$hazard), 's', col = 2)
```



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 look 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).

   I calculate the integration of the survival function. (I cannot figure out how to do it automatically in R, so I typed in each interval by hand.) The result is shown as follows:

```
> diff_time<-c(5,3,1,3,1,0,5,5,4,3,0,1,2,1,9,2,3,113,0)
> Surv_rate<-c(1,km$surv)
> Surv_rate
```

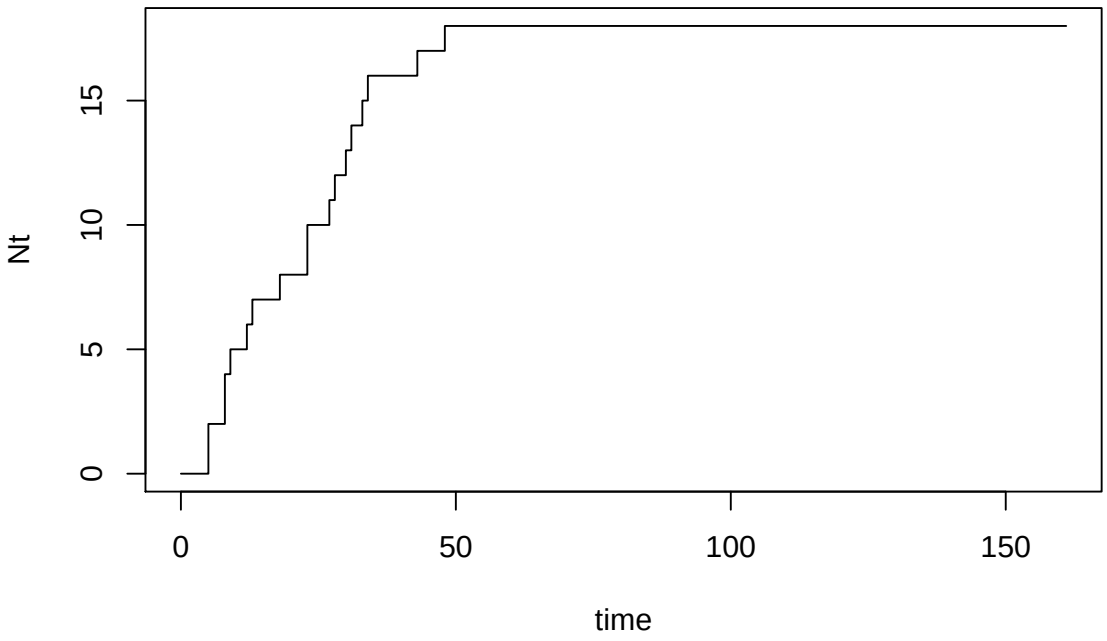
```
 [1] 1.00000000 0.91304348 0.82608696 0.78260870 0.73913043 0.69565217
 [7] 0.69565217 0.64596273 0.54658385 0.49689441 0.49689441 0.44168392
[13] 0.38647343 0.33126294 0.27605245 0.22084196 0.16563147 0.08281573
[19] 0.08281573
```

```
> sum(diff_time*Surv_rate)
```

```
[1] 36.36439
```

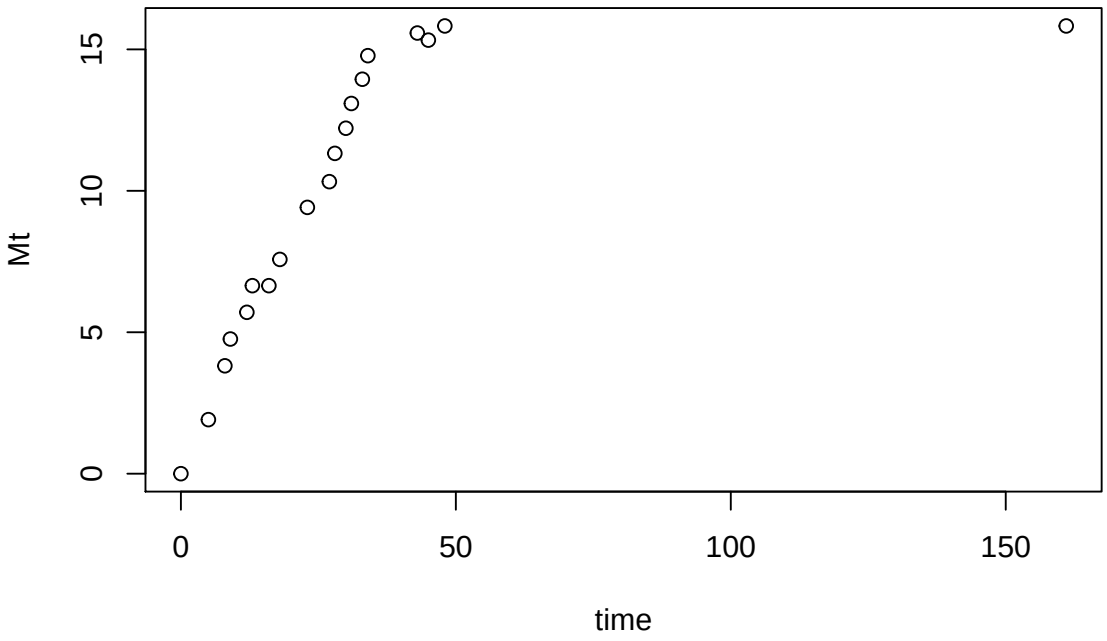
5. Let $N_i(t)$ be the number of events over time interval $(0, t]$ for the ith patient in #(3). Let $N(t) = \sum_{i=1}^{n} N_i(t)$ be the aggregated counting process.

```
> time <- c(0,HO$time)
> Nt <- c(0,2,4,5,6,7,7,8,10,11,12,13,14,15,16,17,17,18,18)
> plot(time,Nt, type="s")
```



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.

```
> Ht <- c(0,H0$hazard)
> Mt<- Nt-Ht
> plot(time,Mt)
```

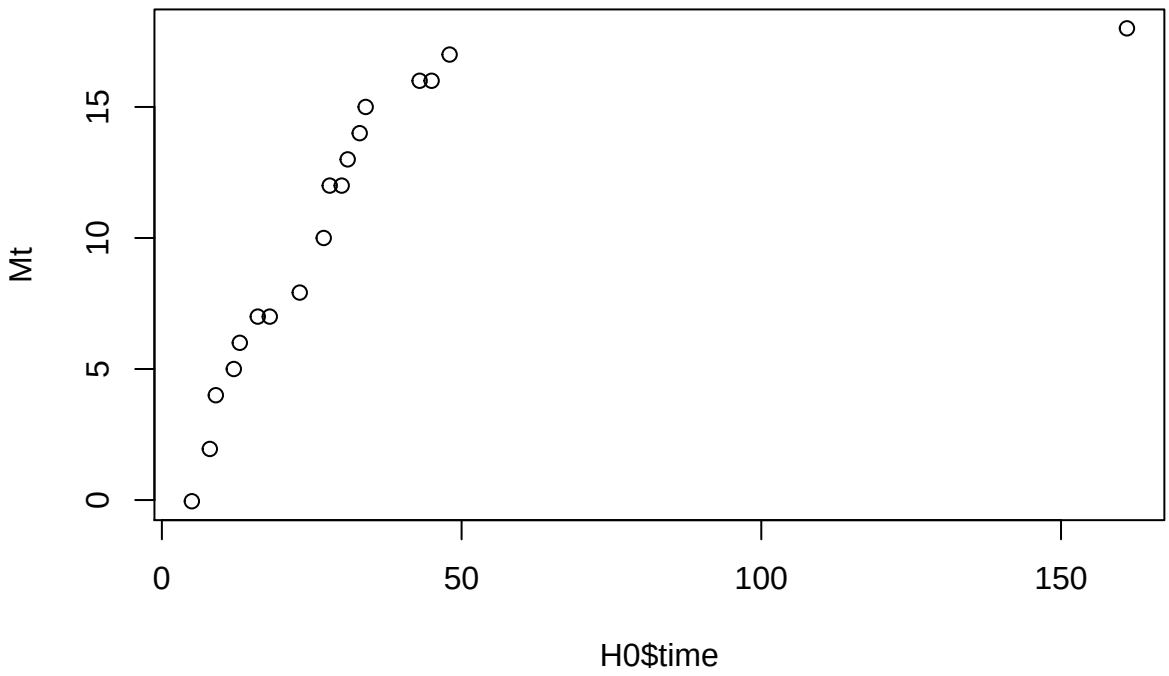


- Note on 5b: After giving some thought, I think it is more meaningful to plot $dM(t) = dN(t) - \hat{h}(t)dt$. Both plots will receive full credit for 5b.

Personally, I don't like this statement. A plot of $dM(t)$ will look nicer since all the points are near 0. However, if we examine carefully in this case, we will find almost all points of $dM(t)$ are above 0, which exactly explains why this $M(t)$ function is (almost) increasing. But ideally, we should have $dM(t)$ randomly dispersed around zero. As a result, a plot of $dM(t)$ will still be abnormal, though it looks more preferable at first glance.

Based on our discussion in Tuesday class, I directly calculate $M(t)$ from the definition: $M(t) = N(t) - \int h(t)Y(t)dt$, and use $\Delta H(t)$ as the approximation for $\int h(t)dt$. That gives the following result:

```
> Ht2 <- c(0,H0$hazard[1:17])
> delta_Ht <- H0$hazard-Ht2
>
> Nt <- c(2,4,5,6,7,7,8,10,11,12,13,14,15,16,17,17,18,18)
> Mt <- Nt-delta_Ht*km$n.risk
>
> plot(H0$time,Mt)
```



The two graphs are almost the same, the difference (I guess) is due to approximation error.