EPI/BIOST 537 Lab

January 28, 2020

(Materials Modified from Those of Jeremy Roth, a Past TA)

What we will cover today

Nonparametric estimation

- Kaplan-Meier estimate of survival function (and median survival time)
- Nelson-Aalen estimate of cumulative hazard function
- A peek at the R code needed for parts of problem 2 on HW2

Kaplan-Meier Estimate of Survival

The Kaplan-Meier (KM) estimate of the survival function is commonly used

- It does not require us to specify/commit to a parametric model
- It accommodates right censoring
- It is pretty straightforward to calculate

Kaplan-Meier Estimate of Survival

What are the “ingredients” for the KM estimate?

For each distinct time point u_k (could be censoring or an event), we must determine

- The number of events at time u_k = d_k
- The number of individuals at risk at time u_k = n_k

Kaplan-Meier Estimate of Survival

Let's do an example by hand: suppose we have 11 people in our study and

Time = {1, 3, 3+, 4, 6+, 7, 7, 9, 11+, 12, 14+}

Time (t)	1	3	4	6	7	9	11	12	14
# at risk (n)	11	10	8	7	6	4	3	2	1
# events (d)	1	1	1	0	2	1	0	1	0
d/n									
1-d/n									
$S_{KM}(t)$									

Kaplan-Meier Estimate of Survival

Let's do an example by hand: suppose we have 11 people in our study and

Time = {1, 3, 3+, 4, 6+, 7, 7, 9, 11+, 12, 14+}

Time (t)	1	3	4	6	7	9	11	12	14
# at risk (n)	11	10	8	7	6	4	3	2	1
# events (d)	1	1	1	0	2	1	0	1	0
d/n	1/11 = 0.091								
1-d/n	10/11 = 0.909								
$S_{KM}(t)$	0.909								

Kaplan-Meier Estimate of Survival

Let's do an example by hand: suppose we have 11 people in our study and

Time = {1, 3, 3+, 4, 6+, 7, 7, 9, 11+, 12, 14+}

Time (t)	1	3	4	6	7	9	11	12	14
# at risk (n)	11	10	8	7	6	4	3	2	1
# events (d)	1	1	1	0	2	1	0	1	0
d/n	1/11 = 0.091	1/10 = 0.1							
1-d/n	10/11 = 0.909	9/10 = 0.9							
$S_{KM}(t)$	0.909	0.909*0.9 = 0.818							

Kaplan-Meier Estimate of Survival

Let's do an example by hand: suppose we have 11 people in our study and

Time = {1, 3, 3+, 4, 6+, 7, 7, 9, 11+, 12, 14+}

Time (t)	1	3	4	6	7	9	11	12	14
# at risk (n)	11	10	8	7	6	4	3	2	1
# events (d)	1	1	1	0	2	1	0	1	0
d/n	1/11 = 0.091	1/10 = 0.1	1/8 = 0.125						
1-d/n	10/11 = 0.909	9/10 = 0.9	7/8 = 0.875						
$S_{KM}(t)$	0.909	0.909*0.9 = 0.818	0.818*0.875 = 0.716						

Kaplan-Meier Estimate of Survival

Let's do an example by hand: suppose we have 11 people in our study and

Time = {1, 3, 3+, 4, 6+, 7, 7, 9, 11+, 12, 14+}

Time (t)	1	3	4	6	7	9	11	12	14
# at risk (n)	11	10	8	7	6	4	3	2	1
# events (d)	1	1	1	0	2	1	0	1	0
d/n	1/11 = 0.091	1/10 = 0.1	1/8 = 0.125	0/7 = 0					
1-d/n	10/11 = 0.909	9/10 = 0.9	7/8 = 0.875	1					
$S_{KM}(t)$	0.909	0.909*0.9 = 0.818	0.818*0.875 = 0.716	0.716*1 = 0.716					

Kaplan-Meier Estimate of Survival

Let's do an example by hand: suppose we have 11 people in our study and

Time = {1, 3, 3+, 4, 6+, 7, 7, 9, 11+, 12, 14+}

Time (t)	1	3	4	6	7	9	11	12	14
# at risk (n)	11	10	8	7	6	4	3	2	1
# events (d)	1	1	1	0	2	1	0	1	0
d/n	1/11 = 0.091	1/10 = 0.1	1/8 = 0.125	0/7 = 0	2/6 = 0.333	1/4 = 0.25	0/3 = 0	1/2 = 0.5	0/1 = 0
1-d/n	10/11 = 0.909	9/10 = 0.9	7/8 = 0.875	1	4/6 = 0.667	3/4 = 0.75	1	1/2 = 0.5	1
$S_{KM}(t)$	0.909	0.909*0.9 = 0.818	0.818*0.875 = 0.716	0.716*1 = 0.716	0.716*0.667 = 0.477	0.477*0.75 = 0.358	0.358*1 = 0.358	0.358*0.5 = 0.179	0.179*1 = 0.179

Kaplan-Meier Estimate of Survival

Let's do another example by hand: suppose we have 8 people in our study and

Time = {1, 1+, 2, 4+, 5, 7, 9+, 10}

Time (t)	1	2	4	5	7	9	10
# at risk (n)	8	6	5	4	3	2	1
# events (d)	1	1	0	1	1	0	1
d/n	1/8 = 0.125	1/6 = 0.167	0/5 = 0	1/4 = 0.25	1/3 = 0.333	0/2 = 0	1/1 = 1
1-d/n	0.875	0.833	1	0.75	0.667	1	0
$S_{KM}(t)$	0.875	0.875*0.833 = 0.729	0.729*1 = 0.729	0.729*0.75 = 0.547	0.547*0.667 = 0.365	0.365*1 = 0.365	0.365*0 = 0

R Lab: Kaplan-Meier and Nelson-Aaelen Estimates

(Available on Canvas as **Discussion_3.Rmd** and **Discussion_3.html**)

R Lab:
Setting up

```
library(survival)
mp <- read.csv(file="Data/6mp.csv", header=TRUE)
head(mp, n=6)
```

```
##   pair time cens   treat
## 1    1    1    1 control
## 2    1   10    1    6-MP
## 3    2   22    1 control
## 4    2    7    1    6-MP
## 5    3    3    1 control
## 6    3   32    0    6-MP
```

```
surv.mp <- Surv(time=mp$time, event=mp$cens, type="right")
survfit.mp <- survfit(surv.mp ~ 1, data=mp, conf.type = "log-log")
survfit.by.treat.mp <- survfit(surv.mp ~ treat, data=mp, conf.type = "log-log")
```

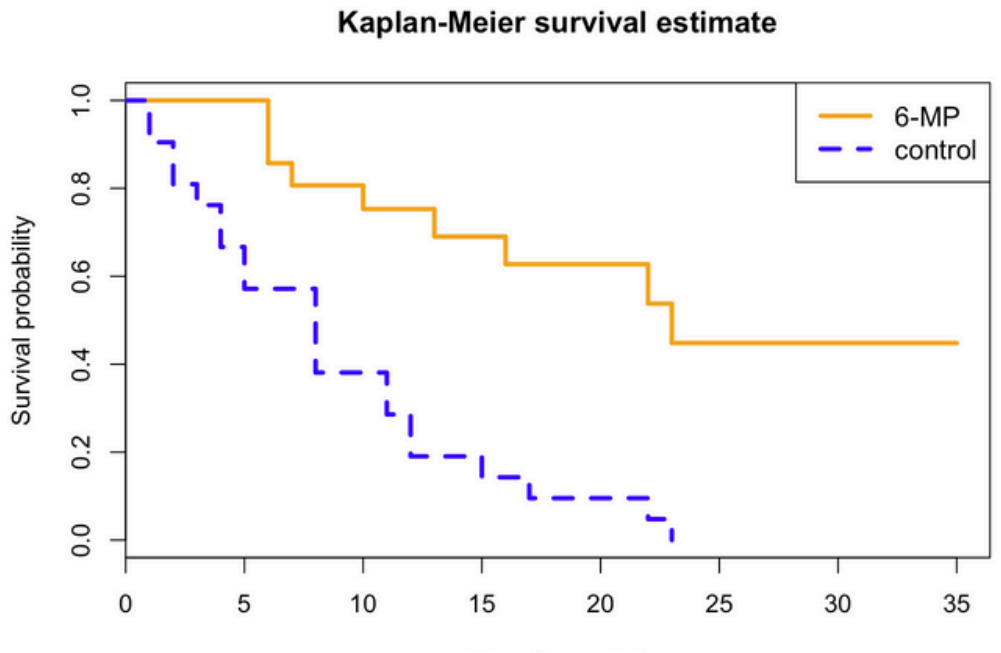
R Lab:
Kaplan-Meier Estimate of Survival

```
summary(survfit.by.treat.mp)
```

```
## Call: survfit(formula = surv.mp ~ treat, data = mp, conf.type = "log-log")
##
##                 treat=6-MP
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##     6     21       3    0.857  0.0764        0.620        0.952
##     7     17       1    0.807  0.0869        0.563        0.923
##    10     15       1    0.753  0.0963        0.503        0.889
##    13     12       1    0.690  0.1068        0.432        0.849
##    16     11       1    0.627  0.1141        0.368        0.805
##    22      7       1    0.538  0.1282        0.268        0.747
##    23      6       1    0.448  0.1346        0.188        0.680
##
##                 treat=control
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##     1     21       2   0.9048  0.0641      0.67005        0.975
##     2     19       2   0.8095  0.0857      0.56891        0.924
##     3     17       1   0.7619  0.0929      0.51939        0.893
##     4     16       2   0.6667  0.1029      0.42535        0.825
##     5     14       2   0.5714  0.1080      0.33798        0.749
##     8     12       4   0.3810  0.1060      0.18307        0.578
##    11      8       2   0.2857  0.0986      0.11656        0.482
##    12      6       2   0.1905  0.0857      0.05948        0.377
##    15      4       1   0.1429  0.0764      0.03566        0.321
##    17      3       1   0.0952  0.0641      0.01626        0.261
##    22      2       1   0.0476  0.0465      0.00332        0.197
##    23      1       1   0.0000     NaN           NA           NA
```

R Lab:
Kaplan-Meier Estimate of Survival

```
plot(survfit.by.treat.mp,
    conf.int=FALSE,
    main="Kaplan-Meier survival estimate",
    ylab="Survival probability", xlab="Time (in weeks)",
    col=c("orange", "blue"),
    lty=c("solid", "dashed"),
    lwd=c(3, 3))
legend("topright",
        levels(mp$treat),
        col=c("orange", "blue"),
        lty=c("solid", "dashed"),
        lwd=c(3, 3), cex=1.1)
```



Kaplan-Meier Estimate of Survival

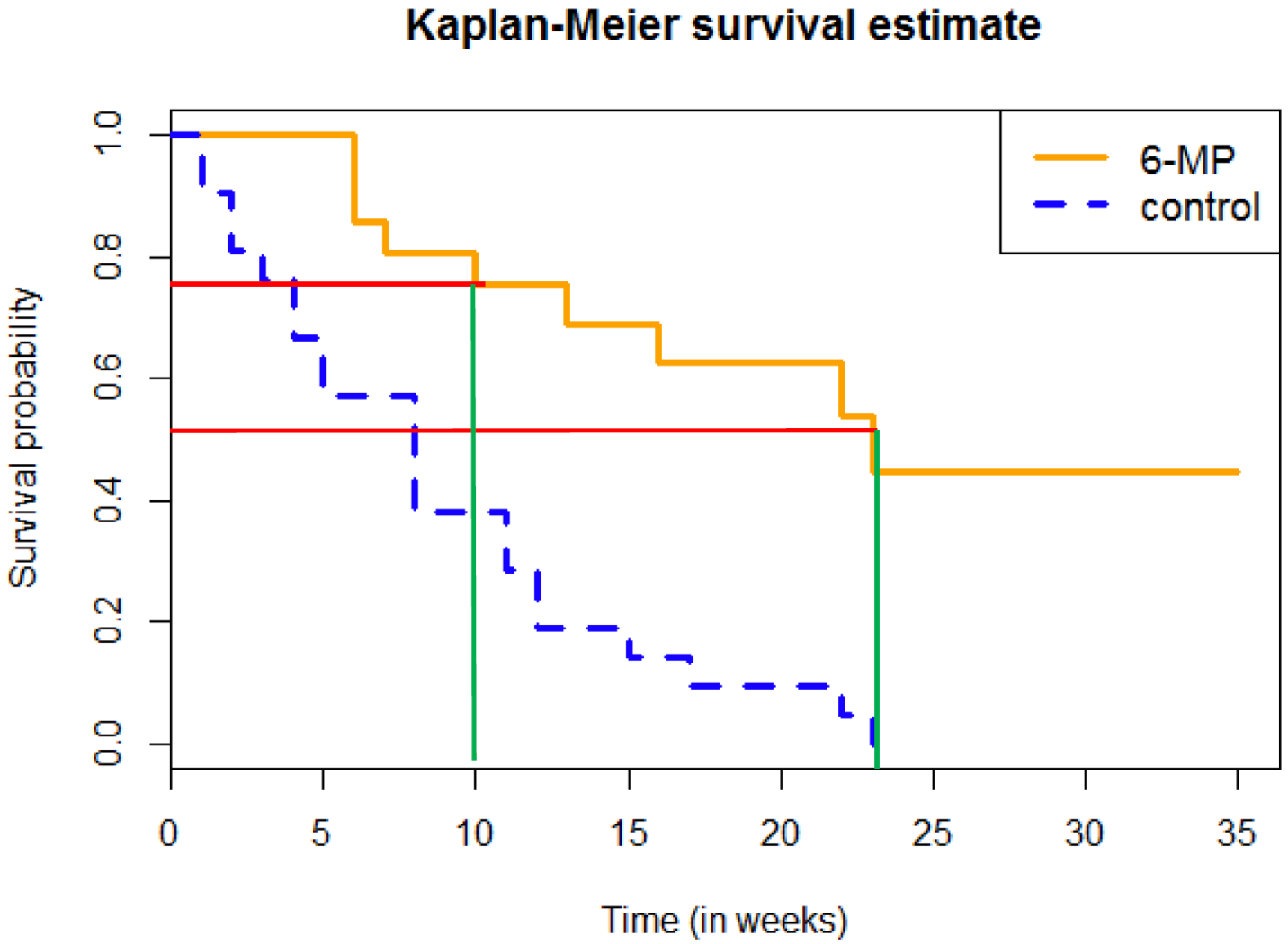
How do we get the estimates we want from this plot?

- Estimates of survival time at a specific $\hat{S}(t)$
  - E.g., median survival time: t such that $\hat{S}(t)$=0.5
- $\hat{S}(t)$ at a specific time
  - E.g., 5-year survival time: $\hat{S}(5\ years)$
- Confidence interval estimates for (1) survival time and (2) $\hat{S}(t)$

R Lab:
Kaplan-Meier Estimate of Survival

For 6-MP group:

- 25th percentile of survival is about 10 weeks
- Median survival is about 23 weeks

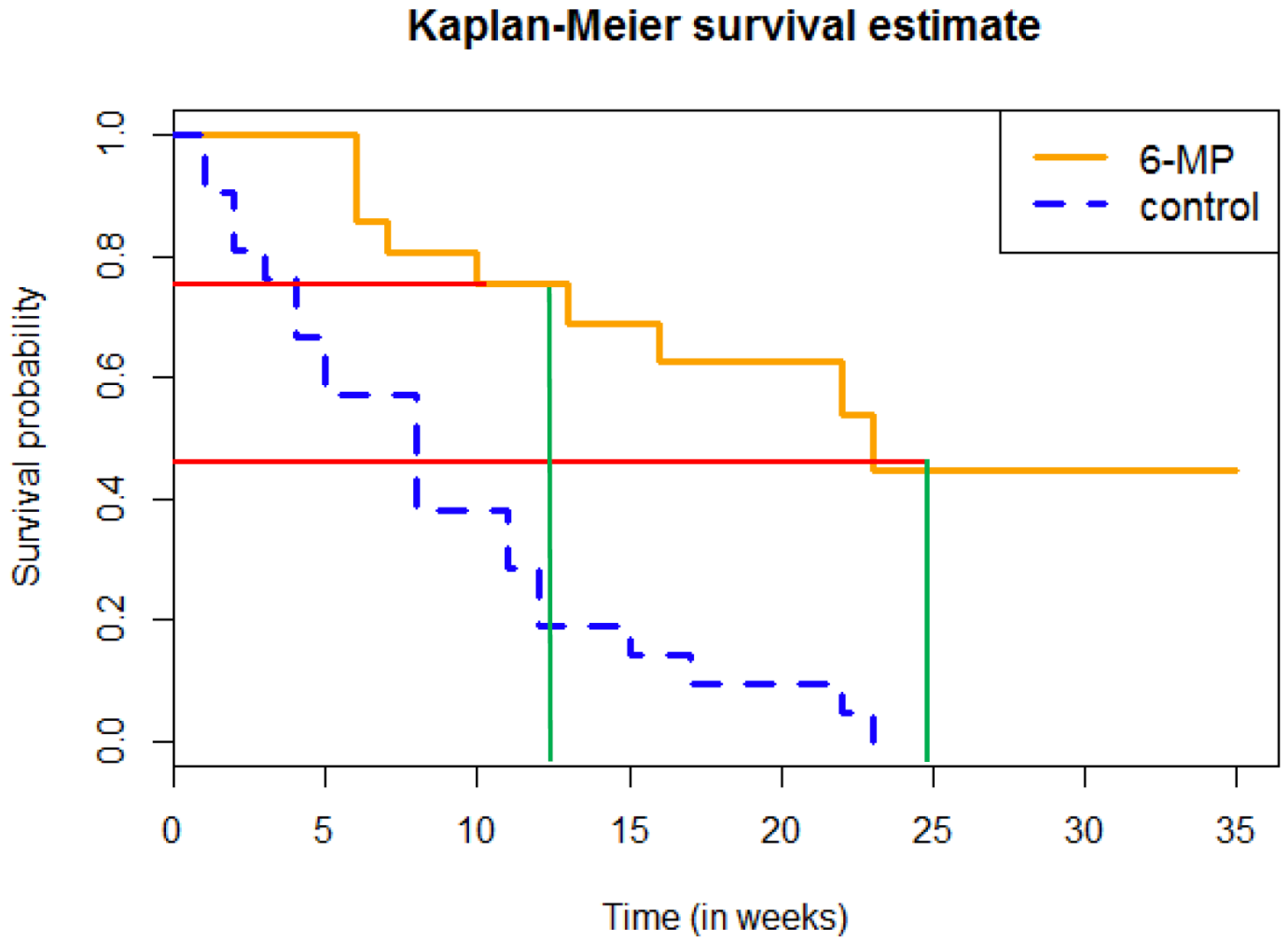


Recall that the 25th percentile of the survival distribution is the value below which 25% of the observations may be found.

R Lab:
Kaplan-Meier Estimate of Survival

For 6-MP group:

- S(12) ≈ 0.75
- S(25) ≈ 0.45



Nelson-Aalen Estimate of CumHaz

The Nelson-Aalen (NA) estimate of the cumulative hazard function is commonly used

- It does not require us to specify/commit to a parametric model
- It accommodates right censoring
- It is pretty straightforward to calculate

Nelson-Aalen Estimate of CumHaz

What are the “ingredients” for the NA estimate?

For each distinct time point u_k (could be censoring or an event), we must determine

- The number of events at time u_k = d_k
- The number of individuals at risk at time u_k = n_k
- (Same as before!)

Nelson-Aalen Estimate of CumHaz

Let's do an example by hand: suppose we have 11 people in our study and

Time = {1, 3, 3+, 4, 6+, 7, 7, 9, 11+, 12, 14+}

Time (t)	1	3	4	6	7	9	11	12	14
# at risk (n)	11	10	8	7	6	4	3	2	1
# events (d)	1	1	1	0	2	1	0	1	0
d/n									
$H_{NA}(t)$									

Nelson-Aalen Estimate of CumHaz

Let's do an example by hand: suppose we have 11 people in our study and

Time = {1, 3, 3+, 4, 6+, 7, 7, 9, 11+, 12, 14+}

Time (t)	1	3	4	6	7	9	11	12	14
# at risk (n)	11	10	8	7	6	4	3	2	1
# events (d)	1	1	1	0	2	1	0	1	0
d/n	1/11 = 0.091								
$H_{NA}(t)$	0.091								

Nelson-Aalen Estimate of CumHaz

Let's do an example by hand: suppose we have 11 people in our study and

Time = {1, 3, 3+, 4, 6+, 7, 7, 9, 11+, 12, 14+}

Time (t)	1	3	4	6	7	9	11	12	14
# at risk (n)	11	10	8	7	6	4	3	2	1
# events (d)	1	1	1	0	2	1	0	1	0
d/n	1/11 = 0.091	1/10 = 0.1							
$H_{NA}(t)$	0.091	0.091 + 0.1 = 0.191							

Nelson-Aalen Estimate of CumHaz

Let's do an example by hand: suppose we have 11 people in our study and

Time = {1, 3, 3+, 4, 6+, 7, 7, 9, 11+, 12, 14+}

Time (t)	1	3	4	6	7	9	11	12	14
# at risk (n)	11	10	8	7	6	4	3	2	1
# events (d)	1	1	1	0	2	1	0	1	0
d/n	1/11 = 0.091	1/10 = 0.1	1/8 = 0.125						
$H_{NA}(t)$	0.091	0.091 + 0.1 = 0.191	0.191 + 0.125 = 0.316						

Nelson-Aalen Estimate of CumHaz

Let's do an example by hand: suppose we have 11 people in our study and

Time = {1, 3, 3+, 4, 6+, 7, 7, 9, 11+, 12, 14+}

Time (t)	1	3	4	6	7	9	11	12	14
# at risk (n)	11	10	8	7	6	4	3	2	1
# events (d)	1	1	1	0	2	1	0	1	0
d/n	1/11 = 0.091	1/10 = 0.1	1/8 = 0.125	0/7 = 0					
$H_{NA}(t)$	0.091	0.091 + 0.1 = 0.191	0.191 + 0.125 = 0.316	0.316 + 0 = 0.316					

Nelson-Aalen Estimate of CumHaz

Let's do an example by hand: suppose we have 11 people in our study and

Time = {1, 3, 3+, 4, 6+, 7, 7, 9, 11+, 12, 14+}

Time (t)	1	3	4	6	7	9	11	12	14
# at risk (n)	11	10	8	7	6	4	3	2	1
# events (d)	1	1	1	0	2	1	0	1	0
d/n	1/11 = 0.091	1/10 = 0.1	1/8 = 0.125	0/7 = 0	2/6 = 0.333	1/4 = 0.25	0/3 = 0	1/2 = 0.5	0/1 = 0
$H_{NA}(t)$	0.091	0.091 + 0.1 = 0.191	0.191 + 0.125 = 0.316	0.316 + 0 = 0.316	0.316 + 0.333 = 0.649	0.649 + 0.25 = 0.899	0.899 + 0 = 0.899	0.899 + 0.5 = 1.399	1.399 + 0 = 1.399

Nelson-Aalen Estimate of CumHaz

Let's do another example by hand: suppose we have 8 people in our study and

Time = {1, 1+, 2, 4+, 5, 7, 9+, 10}

Time (t)	1	2	4	5	7	9	10
# at risk (n)	8	6	5	4	3	2	1
# events (d)	1	1	0	1	1	0	1
d/n	1/8 = 0.125	1/6 = 0.167	0/5 = 0	1/4 = 0.25	1/3 = 0.333	0/2 = 0	1/1 = 1
$H_{NA}(t)$	0.125	0.125 + 0.167 = 0.292	0.292 + 0 = 0.292	0.292 + 0.25 = 0.542	0.542 + 0.333 = 0.875	0.875 + 0 = 0.875	0.875 + 1 = 1.875

R Lab:
Nelson-Aaelen Estimates

```
# In control group
mp.control.only <- mp[mp$treat == "control", ]
surv.control.mp <- Surv(time=mp.control.only$time, event=mp.control.only$cens, type="right")
na.mp.control.only <- basehaz(coxph(surv.control.mp ~ 1, data=mp.control.only))
na.mp.control.only
```

```
##        hazard time
## 1  0.09761905    1
## 2  0.20580618    2
## 3  0.26462971    3
## 4  0.39379638    4
## 5  0.54214803    5
## 6  0.92750156    8
## 7  1.19535870   11
## 8  1.56202537   12
## 9  1.81202537   15
## 10 2.14535870   17
## 11 2.64535870   22
## 12 3.64535870   23
```

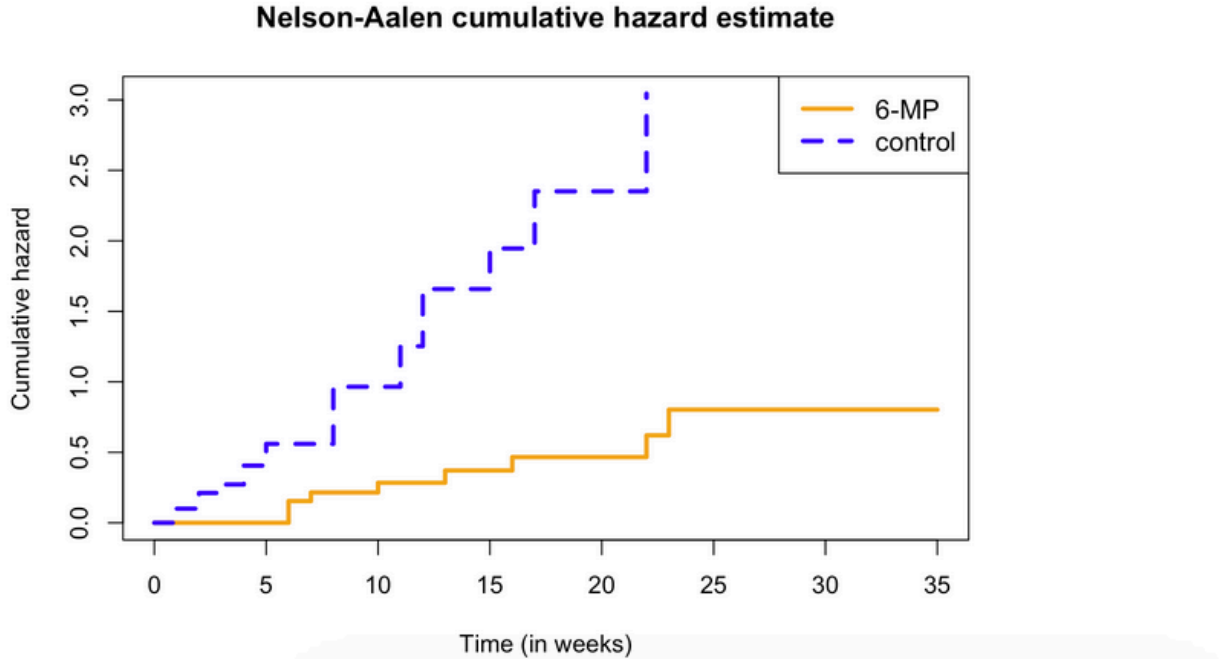
R Lab:
Nelson-Aaelen Estimates

```
  # In treatment group
mp.treatment.only <- mp[mp$treat == "6-MP", ]
surv.treatment.mp <- Surv(time=mp.treatment.only$time, event=mp.treatment.only$cens, type="right")
na.mp.treatment.only <- basehaz(coxph(surv.treatment.mp ~ 1, data=mp.treatment.only))
na.mp.treatment.only
```

```
##       hazard time
## 1  0.1502506    6
## 2  0.2090742    7
## 3  0.2090742    9
## 4  0.2757408   10
## 5  0.2757408   11
## 6  0.3590742   13
## 7  0.4499832   16
## 8  0.4499832   17
## 9  0.4499832   19
## 10 0.4499832   20
## 11 0.5928404   22
## 12 0.7595071   23
## 13 0.7595071   25
## 14 0.7595071   32
## 15 0.7595071   34
## 16 0.7595071   35
```

Nelson-Aalen Estimate of CumHaz

```
plot(survfit.by.treat.mp,
     fun="cumhaz",
     conf.int=FALSE,
     main="Nelson-Aalen cumulative hazard estimate",
     ylab="Cumulative hazard", xlab="Time (in weeks)",
     col=c("orange", "blue"),
     lty=c("solid", "dashed"),
     lwd=c(3, 3))
legend("topright",
         levels(mp$treat),
         col=c("orange", "blue"),
         lty=c("solid", "dashed"),
         lwd=c(3, 3), cex=1.1)
```



The Rest of Homework 2

See materials for next week's discussion section!