Survival Data and Methods

David M. Rocke

March 11, 2026

Mean Residual Life

The mean lifetime with a survival distribution $f(x)$ is

$$\int_0^\infty xf(x)dx$$

For the exponential distribution we know that the mean is λ^{-1} The mean residual life after survival to time x is

$$mrl(x) = S^{-1}(x)\int_x^\infty (u-x)f(u)du$$

For the exponential, the mean residual life is also λ^{-1}

Period Life Table, 2019

Exact age	Male			Female		
	Death probability [a]	Number of lives [b]	Life expectancy	Death probability [a]	Number of lives [b]	Life expectancy
0	0.006081	100,000	76.23	0.005046	100,000	81.28
1	0.000425	99,392	75.69	0.000349	99,495	80.69
2	0.000260	99,350	74.73	0.000212	99,461	79.72
3	0.000194	99,324	73.75	0.000166	99,440	78.74
4	0.000154	99,305	72.76	0.000137	99,423	77.75
5	0.000142	99,289	71.77	0.000122	99,409	76.76
6	0.000135	99,275	70.78	0.000111	99,397	75.77
7	0.000127	99,262	69.79	0.000103	99,386	74.78
8	0.000117	99,249	68.80	0.000098	99,376	73.79
9	0.000104	99,238	67.81	0.000095	99,366	72.79
10	0.000097	99,227	66.81	0.000096	99,357	71.80
11	0.000106	99,218	65.82	0.000102	99,347	70.81
12	0.000145	99,207	64.83	0.000116	99,337	69.81
13	0.000220	99,193	63.84	0.000139	99,326	68.82
14	0.000324	99,171	62.85	0.000170	99,312	67.83
15	0.000437	99,139	61.87	0.000204	99,295	66.84
16	0.000552	99,096	60.90	0.000240	99,275	65.86
17	0.000676	99,041	59.93	0.000278	99,251	64.87
18	0.000806	98,974	58.97	0.000319	99,223	63.89
19	0.000939	98,894	58.02	0.000360	99,192	62.91

- The 2019 US standard mortality table estimates the expectation of life of females at birth as 81.28 years.
- At age 50, 95.6% of US females are still alive.
- The mean residual life at age 50 is 33.51 years (age $50 + 33.51 = 83.51$). At age 83, 56.4% are still alive.
- In 1850 an estimate of the expectation of life at birth for females is 39.4 years. At age 1, it is $1 + 49.3 = 50.3$
- But 44.7% of females lived to age 50 and the further expectation of life was 20.4 years, so to age 70.4. About 24% lived to age 70 and 10% to age 80.
- So it was not rare to live beyond age 39.

Actuarial Life Tables

An actuarial life table is constructed from age- and sex-specific death rates in a current year and the age- and sex-specific data on the percentage of current age people who survived from their birth year. These actuarial tables present life data for a hypothetical population in which at every age the death rates and survival from the birth year were the same as in the year the table was constructed.

Actuarial Life Tables

- This applies as if a (say) female person born 1/1/2019 would have a chance of dying between age 50 and 51 equal to the chance a female person born 1/1/1969 had of dying between the age of 50 and 51 even though the hypothetical event would take place between 1/1/2069 and 12/31/2069.
- In reality, a 50 year old person in 2019 would have a death rate depending on the current year (time effect) and the year of birth (cohort effect).
- The cohort effect in 2069 for those born in 2019 and the time effect for those aged 50 in 2069 are both unknown.

Actuarial Life Tables

- This is however the best we can do, and is in any case standard.
- Announcements that the life expectency in the US dropped in 2020 by one year is, however, not useful. This comes from applying the excess mortality due to COVID to each future year of someone born in 2020, which seems unlikely to be correct.
- These tables have substantial practical importance and it is unclear what SSA and others will make of this.

Other Parametric Survival Distributions

- Any density on $[0, \infty)$ can be a survival distribution, but the most useful ones are all skew right.
- The commonest generalization of the exponential is the Weibull.
- Other common choices are the gamma, log-normal, log-logistic, Gompertz, inverse Gaussian, and Pareto.
- Most of what we in biomedical statistics is non-parametric or semi-parametric, but sometimes these parametric distributions provide a useful approach.
- Engineering applications usually, but not always, use parametric distributions.

Weibull Distribution

$$
\begin{aligned}
f(x) &= \alpha\lambda x^{\alpha-1} e^{-\lambda x^{\alpha}} \\
h(x) &= \alpha\lambda x^{\alpha-1} \\
S(x) &= e^{-\lambda x^{\alpha}} \\
E(X) &= \Gamma(1+1/\alpha)/\lambda^{1/\alpha}
\end{aligned}
$$

When $\alpha = 1$ this is the exponential. When $\alpha > 1$ the hazard is increasing and when $\alpha < 1$ the hazard is decreasing. This provides more flexibility than the exponential.

Nonparametric Survival Analysis

- Mostly, we work without a parametric model.
- The first task is to estimate a survival function from data listing survival times, and censoring times for right-censored data.
- For example one patient may have relapsed at 10 months. Another might have been followed for 32 months without a relapse having occurred (censored).
- The minimum information we need for each patient is a time and a censoring variable which is 1 if the event occurred at the indicated time and 0 if this is a censoring time.

KM drug6mp Data

This is from a clinical trial in 1963 for 6-MP treatment vs. placebo for Acute Leukemia in 42 children. At the time, the 5-year survival rate from this disease was around 10%—it is now around 90%.

Pairs of children matched by remission status at the time of treatment (1 = partial or 2 = complete) and randomized to 6-MP or placebo. Followed until relapse or end of study. All of the placebo group relapsed, but some of the 6-MP group were censored (which means they were still in remission).

6-MP = 6-Mercaptopurine (Purinethol) is an anti-cancer ("antineoplastic" or "cytotoxic") chemotherapy drug used currently for Acute Lymphoblastic Leukemia (ALL). It is classified as an antimetabolite.

KM drug6mp Data

Clinical trial in 1963 for 6-MP treatment vs. placebo for Acute Leukemia in 42 children. Pairs of children matched by remission status at the time of treatment (1 = partial or 2 = complete) and randomized to 6-MP or placebo. Followed until relapse or end of study. All of the placebo group relapsed, but some of the 6-MP group were censored.

```
> library(KMsurv)
> data(drug6mp)
> drug6mp
   pair remstat t1 t2 relapse
1     1       1  1 10       1
2     2       2 22  7       1
3     3       2  3 32       0
```

KM drug6mp Data

```
drug6mp data

Description

The drug6mp data frame has 21 rows and 5 columns.

Format

This data frame contains the following columns:

pair          pair number
remstat       Remission status at randomization (1=partial, 2=complete)
t1            Time to relapse for placebo patients, months
t2            Time to relapse for 6-MP patients, months
relapse       Relapse indicator (0=censored, 1=relapse) for 6-MP patients
```

Descriptive Statistics

- The average time in each group is not useful. Some of the 6-MP patients have not relapsed at the time recorded, while all of the placebo patients have relapsed.
- The median time is not really useful either because so many of the 6-MP patients have not relapsed (12/21).
- Both are biased down in the 6-MP group. Remember that lower times are worse since they indicate sooner recurrence.

Descriptive Statistics

- We can compute the average hazard rate, which is the estimate of the exponential parameter: number of relapses divided by the sum of the times.
- For the placebo, that is just the reciprocal of the mean time $= 1/8.667 = 0.115$.
- For the 6-MP group this is $9/359 = 0.025$
- The estimated average hazard in the placebo group is 4.6 times as large (if the hazard is constant over time).

The Kaplan-Meier Product Limit Estimator

- The estimated survival function for the placebo patients is easy to compute. For any time t in months, $S(t)$ is the fraction of patients with times greater than t.
- For the 6-MP patients, we cannot ignore the censored data because we know that the time to relapse is greater than the censoring time.

The Kaplan-Meier Product Limit Estimator

- For any time t in months, we know that 6-MP patients with times greater than t have not relapsed, and those with relapse time less than t have relapsed, but we don't know if patients with censored time less than t have relapsed or not.
- The procedure we usually use is the Kaplan-Meier product-limit estimator of the survival function.

- The Kaplan-Meir estimator is a step function (like the empirical cdf), which changes value only at the event times, not at the censoring times.
- At each event time t, we compute the at-risk group size Y, which is all those observations whose event time or censoring time is at least t.
- If d of the observations have an event time (not a censoring time) of t, then the group of survivors immediately following time t is reduced by the fraction

$$\frac{Y-d}{Y} = 1 - \frac{d}{Y}$$

If the event times are t_i with events per time of d_i $(1 \leq i \leq k)$, then

$$\hat{S}(t) = \prod_{t_i < t} [1 - d_i / Y_i]$$

where Y_i is the set of observations whose time (event or censored) is $\geq t_i$, the group at risk at time t_i.

If there are no censored data, and there are n data points, then just after (say) the third event time

$$\begin{aligned}
\hat{S}(t) &= \prod_{t_i<t}[1 - d_i/Y_i] \\
&= [\frac{n-d_1}{n}][\frac{n-d_1-d_2}{n-d_1}][\frac{n-d_1-d_2-d_3}{n-d_1-d_2}] \\
&= \frac{n-d_1-d_2-d_3}{n}
\end{aligned}$$

the usual empirical cdf estimate.

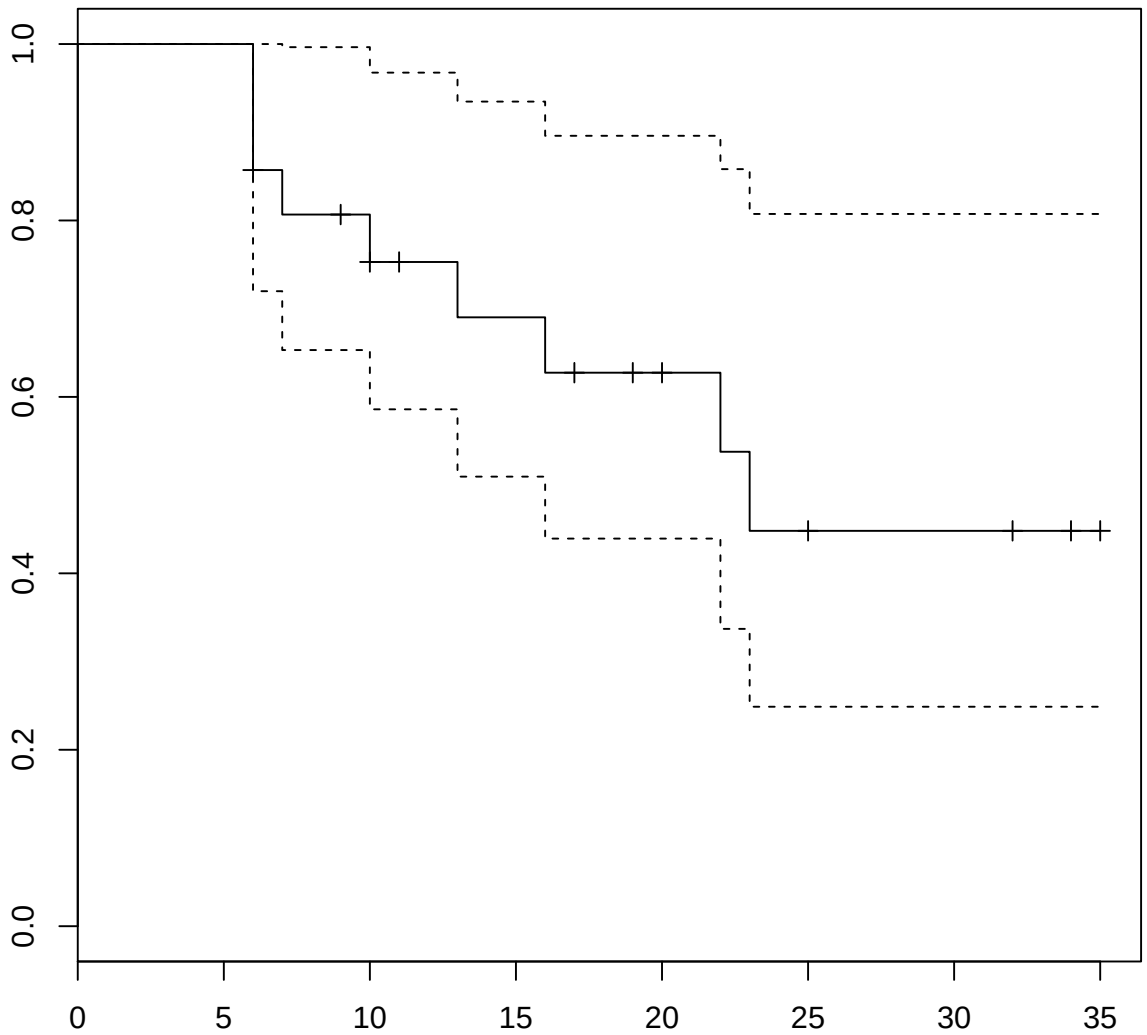
```
require(KMsurv)
data(drug6mp)
plot(survfit(Surv(drug6mp$t2,drug6mp$relapse)~1))
title("Kaplan-Meier Survival Curve for 6-MP Patients")

time12 <- c(drug6mp$t1,drug6mp$t2)
cens12 <- c(rep(1,21),drug6mp$relapse)
treat12 <- rep(1:2,each=21)
pairs12 <- rep(1:21,2)

plot(survfit(Surv(time12,cens12)~treat12),col=1:2)
title("Kaplan-Meier Survival Curve for 6-MP and Placebo Patients")

plot(survfit(Surv(time12,cens12)~treat12),conf.int=T,col=1:2)
title("Kaplan-Meier Survival Curve for 6-MP and Placebo Patients")
```

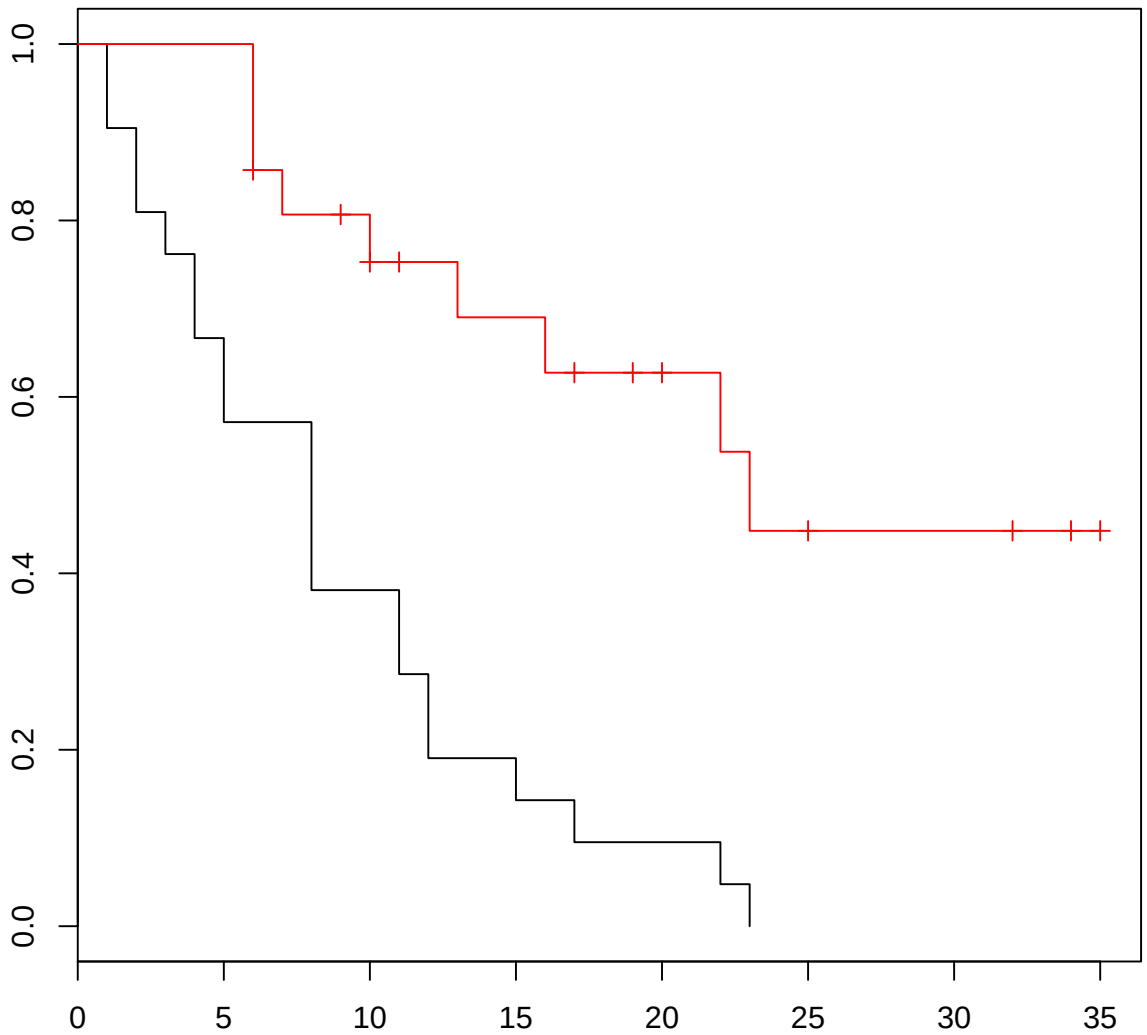
Time	At Risk	Relapses	Censored	KM Factor	KM Curve
6	21	3	1	0.857	0.857
7	17	1	0	0.941	0.807
9	16	0	1	1	0.807
10	15	1	1	0.933	0.753
11	13	0	1	1	0.753
13	12	1	0	0.917	0.690
16	11	1	0	0.909	0.627
17	10	0	1	1	0.627
19	9	0	1	1	0.627
20	8	0	1	1	0.627
22	7	1	0	0.857	0.538
23	6	1	0	0.833	0.448
25	5	0	1	1	0.448
32	4	0	2	1	0.448
34	2	0	1	1	0.448
35	1	0	1	1	0.448

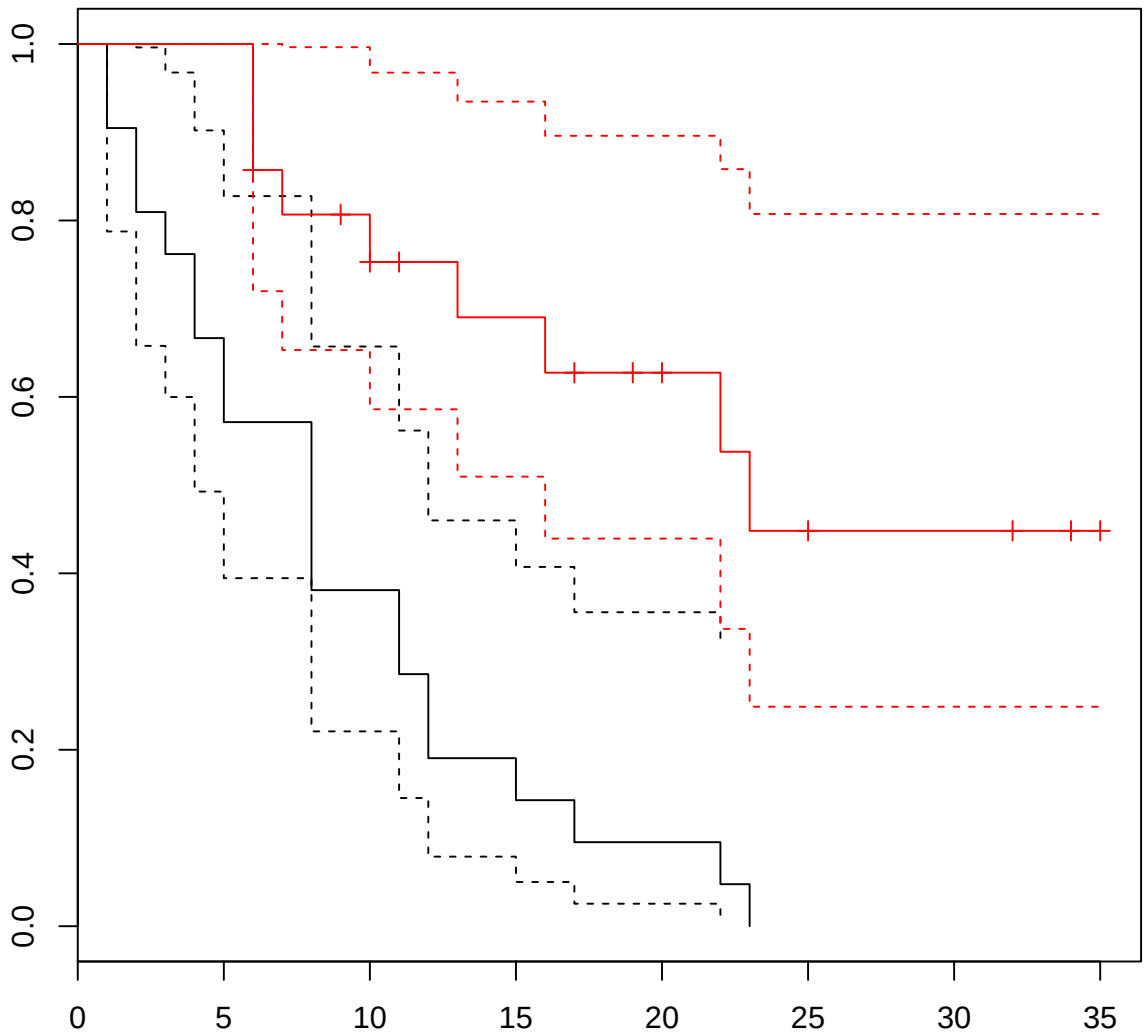
For the 6-MP patients at time 6 months, there are 21 patients at risk. At $t = 6$ there are 3 relapses and 1 censored observations. The Kaplan-Meier factor is $(21 - 3)/21 = 0.857$. The number at risk for the next time $(t = 7)$ is $21 - 3 - 1 = 17$.

At time 7 months, there are 17 patients at risk. At $t = 7$ there is 1 relapse and 0 censored observations. The Kaplan-Meier factor is $(17 - 1)/17 = 0.941$. The Kaplan Meier estimate is $0.857 \times 0.941 = 0.807$. The number at risk for the next time $(t = 9)$ is $17 - 1 = 16$.

```
time12 <- c(drug6mp$t1,drug6mp$t2)
cens12 <- c(rep(1,21),drug6mp$relapse)
treat12 <- rep(1:2,each=21)
pairs12 <- rep(1:21,2)

print(survdiff(Surv(time12,cens12)~treat12))

           N Observed Expected (O-E)^2/E (O-E)^2/V
treat12=1 21       21     10.7      9.77      16.8
treat12=2 21        9     19.3      5.46      16.8

 Chisq= 16.8  on 1 degrees of freedom, p= 4.17e-05

print(survdiff(Surv(time12,cens12)~treat12+strata(pairs12)))

           N Observed Expected (O-E)^2/E (O-E)^2/V
treat12=1 21       21     13.5      4.17      10.7
treat12=2 21        9     16.5      3.41      10.7

 Chisq= 10.7  on 1 degrees of freedom, p= 0.00106
```


Kaplan−Meier Survival Curve for 6−MP Patients
1.0
0.8
0.6
0.4
0.2
0.0
0
5
10
15
20
25
30
35


Kaplan−Meier Survival Curve for 6−MP and Placebo Patients
1.0
0.8
0.6
0.4
0.2
0.0
0
5
10
15
20
25
30
35


Kaplan−Meier Survival Curve for 6−MP and Placebo Patients
1.0
0.8
0.6
0.4
0.2
0.0
0
5
10
15
20
25
30
35

Package Survival

```
Surv

Create a survival object, usually used as a response variable in a model formula.

Usage

Surv(time, event)

Arguments

time   for right censored data, this is the follow up time.

event   The status indicator, normally 0=alive, 1=dead.
        Also TRUE/FALSE (TRUE = death) or 1/2 (2=death).
        The event indicator can be omitted,
          in which case all subjects are assumed to have an event.
-----
Surv(drug6mp$t2,drug6mp$relapse)
```

Package Survival

```
survfit

This function creates survival curves from either a formula
(e.g. the Kaplan-Meier), a previously fitted Cox model,
or a previously fitted accelerated failure time model.

Usage

survfit(formula, ...)

Arguments

formula    either a formula or a previously fitted model
-----
plot(survfit(Surv(drug6mp$t2,drug6mp$relapse)~1))
plot(survfit(Surv(time12,cens12)~treat12))
```

Package Survival

```
survdiff

Tests if there is a difference between two or more survival curves.

Usage

survdiff(formula, data, subset, na.action, rho=0)

Arguments

formula     a formula expression as for other survival models,
            of the form Surv(time, status) ~ predictors.
            A strata term may be used to produce a stratified test.

rho         Type of test. Default is the Mantel-Haenszel test.
-------
print(survdiff(Surv(time12,cens12)~treat12))
print(survdiff(Surv(time12,cens12)~treat12+strata(pairs12)))
```

Bone Marrow Transplant Data

- Copelan et al. (1991) study of allogeneic (from a donor) bone marrow transplant therapy for acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL).
- Possible intermediate events are graft vs. host disease (GVHD), an immunological rejection response to the transplant, and platelet recovery, a return of platelet count to normal levels. One or the other, both in either order, or neither may occur.
- End point events are relapse of the disease or death.
- Any or all of these events may be censored.

KMsurv bmt data

```
The bmt data frame has 137 rows and 22 columns.

This data frame contains the following columns:

group   Disease Group 1-ALL, 2-AML Low Risk, 3-AML High Risk
t1      Time To Death Or On Study Time
t2      Disease Free Survival Time (Time To Relapse, Death, Or End Of Study)
d1      Death Indicator 1-Dead 0-Alive
d2      Relapse Indicator 1-Relapsed, 0-Disease Free
d3      Disease Free Survival Indicator 1-Dead Or Relapsed, 0-Alive Disease Free)
ta      Time To Acute Graft-Versus-Host Disease
da      Acute GVHD Indicator 1-Developed Acute GVHD 0-Never Developed Acute GVHD)
tc      Time To Chronic Graft-Versus-Host Disease
dc      Chronic GVHD Indicator 1-Developed Chronic GVHD
          0-Never Developed Chronic GVHD
tp      Time To Platelet Recovery
dp      Platelet Recovery Indicator 1-Platelets Returned To Normal,
          0-Platelets Never Returned to Normal
```

KMsurv bmt data

```
z1      Patient Age In Years
z2      Donor Age In Years
z3      Patient Sex: 1-Male, 0-Female
z4      Donor Sex: 1-Male, 0-Female
z5      Patient CMV Status: 1-CMV Positive, 0-CMV Negative
z6      Donor CMV Status: 1-CMV Positive, 0-CMV Negative
z7      Waiting Time to Transplant In Days
z8      FAB: 1-FAB Grade 4 Or 5 and AML, 0-Otherwise
z9      Hospital: 1-The Ohio State University, 2-Alferd , 3-St. Vincent,
          4-Hahnemann
z10     MTX Used as a Graft-Versus-Host- Prophylactic: 1-Yes 0-No
```

Bone Marrow Transplant Example

- We concentrate for now on disease-free survival (t2 and d3) for the three risk groups, ALL, AML Low Risk, and AML High Risk.
- We will construct the Kaplan-Meier survival curves, compare them, and test for differences.
- We will construct the cumulative hazard curves and compare them.
- We will estimate the hazard functions, interpret, and compare them.
- Then we will introduce the Cox proportional hazards model.

Survival Function

$$\hat{S}(t) = \prod_{t_i < t} [1 - d_i / Y_i]$$

where Y_i is the group at risk at time t_i.
The estimated variance of $\hat{S}(t)$ is (Greenwood's formula)

$$\hat{V}[\hat{S}(t)] = \hat{S}(t)^2 \sum_{t_i < t} \frac{d_i}{Y_i(Y_i - d_i)}$$

which we can use for confidence intervals for a survival function or a difference of survival functions.

Cumulative Hazard

$$h(t) = -\frac{d \ln S(t)}{dt}$$

The cumulative hazard function is

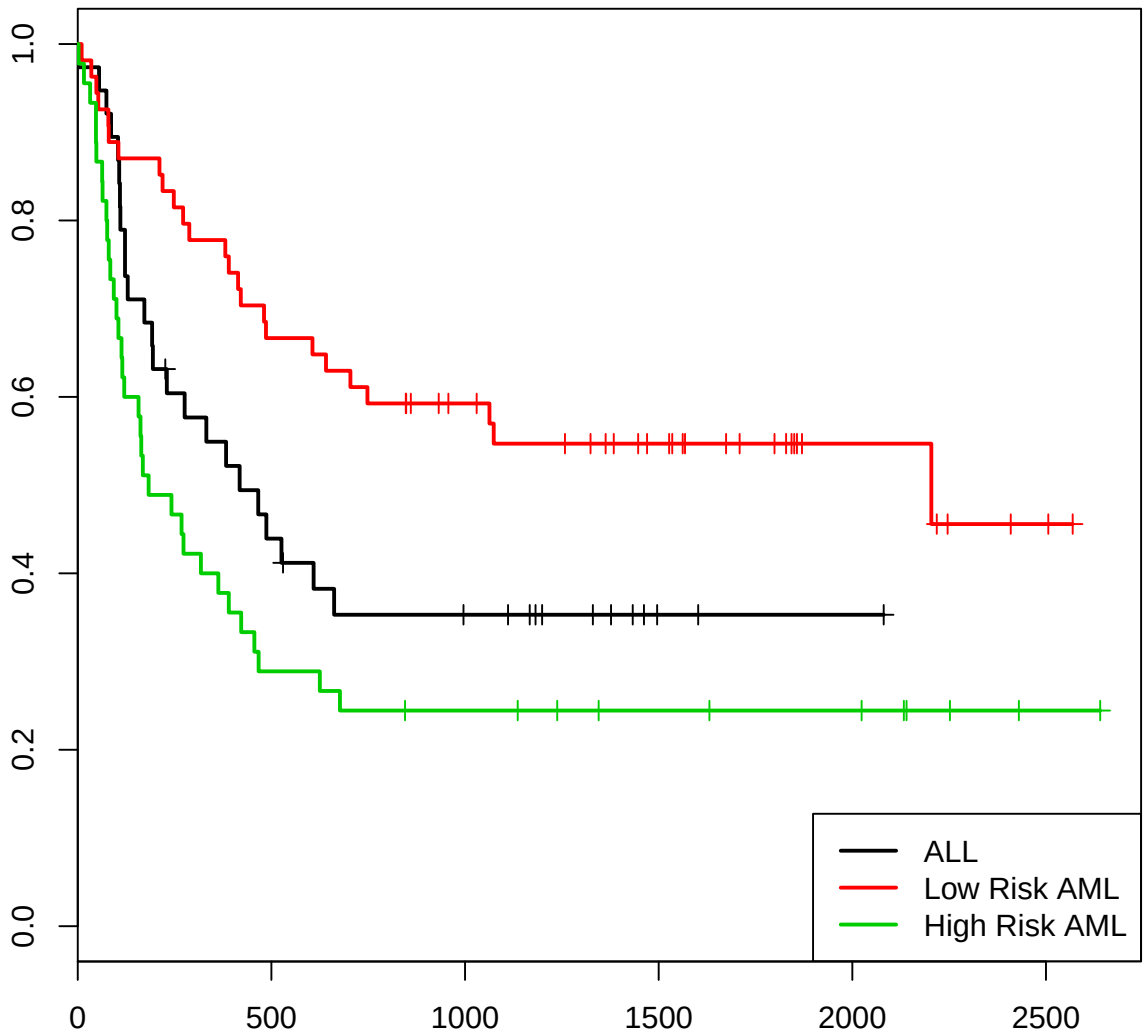
$$\begin{aligned} H(t) &= \int_0^t h(t)dt \\ &= -\ln S(t) \\ \hat{H}(t) &= -\ln \hat{S}(t) \end{aligned}$$

```
> library(KMsurv)
> library(survival)
> data(bmt)
> dfsurv <- Surv(bmt$t2,bmt$d3)
```

The last command creates a survival object from the time variable t2 (disease-free survival) and the associated status variable d3. This is usually the first step in computer analysis of survival data.

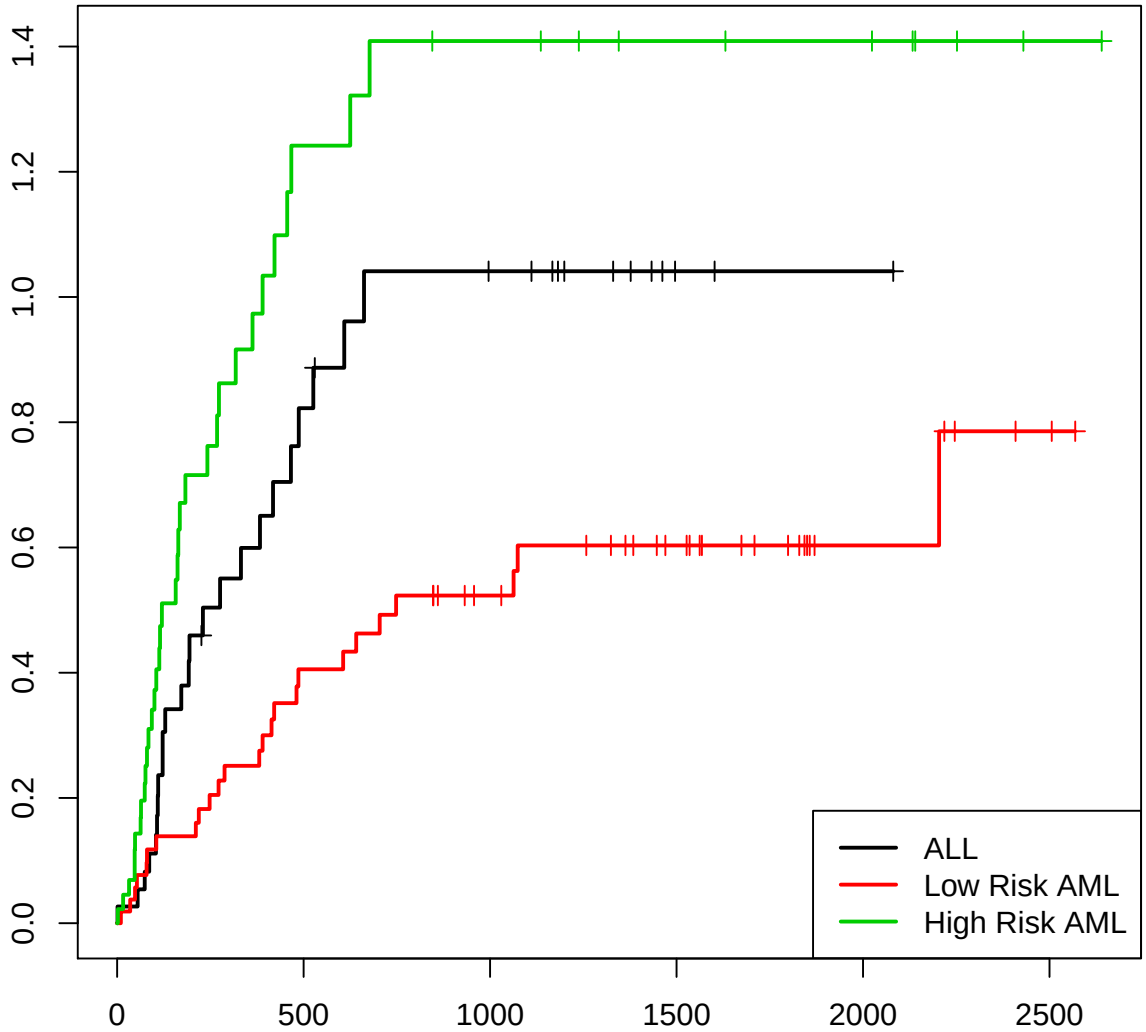
```
> plot(survfit(dfsurv~group,data=bmt),col=1:3,lwd=2)
> title("Disease-Free Survival for Three Groups")
> legend("bottomright",c("ALL","Low Risk AML","High Risk AML"),col=1:3,lwd=2)
```

This plots the estimated survival curves for the three groups on the same graph in three colors with associated legend.


Disease−Free Survival for Three Groups
1.0
0.8
0.6
0.4
0.2
0.0
0
500
1000
1500
2000
2500
ALL
Low Risk AML
High Risk AML

```
> plot(survfit(dfsurv~group,data=bmt),col=1:3,lwd=2,fun="cumhaz")
> title("Disease-Free Cumulative Hazard for Three Groups")
> legend("bottomright",c("ALL","Low Risk AML","High Risk AML"),col=1:3,lwd=2)
```

This plots the cumulative hazards for the three groups.


Disease−Free Cumulative Hazard for Three Groups
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
0
500
1000
1500
2000
2500
ALL
Low Risk AML
High Risk AML

```
> survdiff(dfsurv~group,data=bmt)
         N Observed Expected (O-E)^2/E (O-E)^2/V
group=1 38       24     21.9     0.211     0.289
group=2 54       25     40.0     5.604    11.012
group=3 45       34     21.2     7.756    10.529

 Chisq= 13.8  on 2 degrees of freedom, p= 0.00101
```

This tests whether the three groups could have a common survival function. Note that group is treated as a factor even though it is numeric. This is the Mantel-Haenszel test.

Nelson-Aalen Survival Function Estimate

The point hazard at time t_i can be estimated by d_i/Y_i which leads to the estimate of the cumulative hazard

$$\hat{H}(t) = \sum_{t_i<t} d_i/Y_i$$

which has approximate variance

$$\hat{V}[\hat{H}(t)] = \sum_{t_i<t} \frac{(d_i/Y_i)(1-d_i/Y_i)}{Y_i} \approx \sum_{t_i<t} \frac{d_i}{Y_i^2}$$

giving an alternate estimate of the survival function

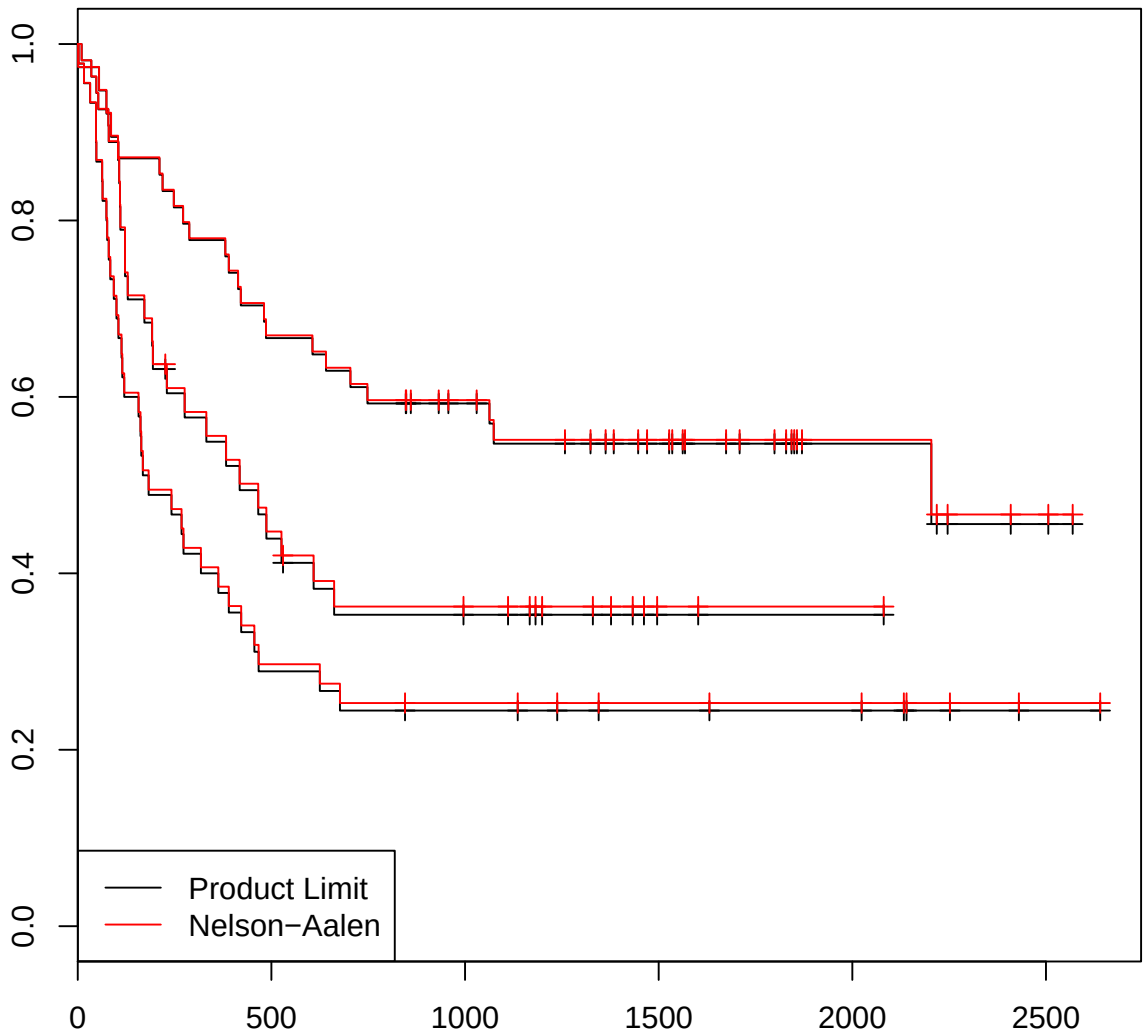
$$\hat{S}_{NA}(t) = \exp[-\hat{H}(t)]$$

KM and NA Survival Function Estimates

$$
\begin{aligned}
\hat{S}_{KM}(t) &= \prod_{t_i<t}[1-d_i/Y_i] \\
\hat{V}[\hat{S}_{KM}(t)] &= \hat{S}(t)^2 \sum_{t_i<t} \frac{d_i}{Y_i(Y_i-d_i)} \\
\hat{S}_{NA}(t) &= \exp[-\sum_{t_i<t} d_i/Y_i] \\
&= \prod_{t_i<t} \exp(-d_i/Y_i) \\
&\approx \prod_{t_i<t}[1-d_i/Y_i]
\end{aligned}
$$

The product limit estimate and the Nelson-Aalen estimate often do not differ by much. The latter is considered more accurate in small samples and also directly estimates the cumulative hazard. The `"fleming-harrington"` method reduces to Nelson-Aalen when the data are unweighted. We can also estimate the cumulative hazard as the negative log of the KM survival function estimate.

```
nafit <- survfit(dfsurv~group,type="fleming-harrington",data=bmt)

plot(survfit(dfsurv~group,data=bmt))
lines(nafit,col=2)
legend("bottomleft",c("Product Limit","Nelson-Aalen"),col=1:2,lwd=1)
title("Two Survival Function Estimates for Three Groups")
```

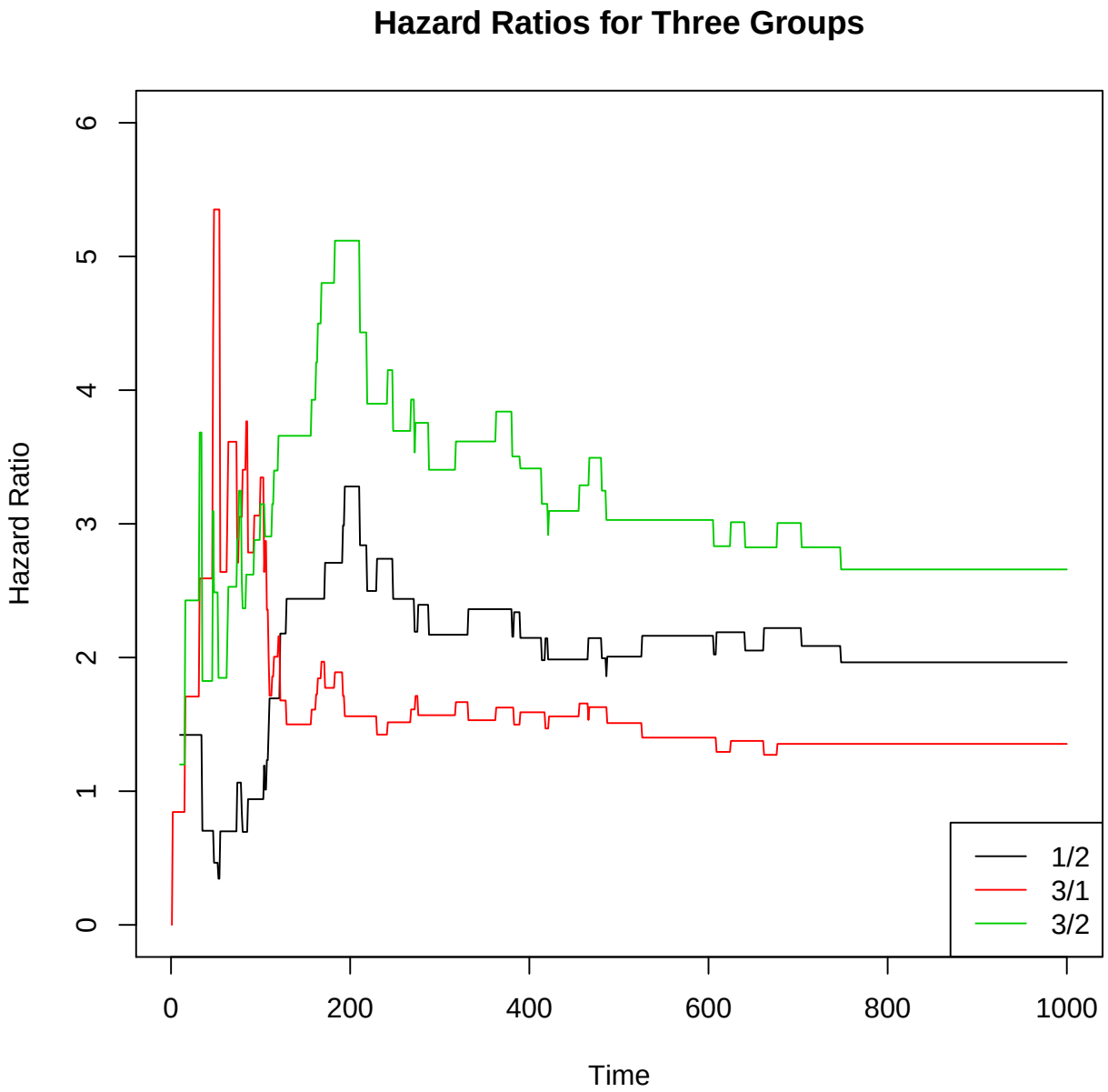

Two Survival Function Estimates for Three Groups
1.0
0.8
0.6
0.4
0.2
0.0
0
500
1000
1500
2000
2500
Product Limit
Nelson−Aalen

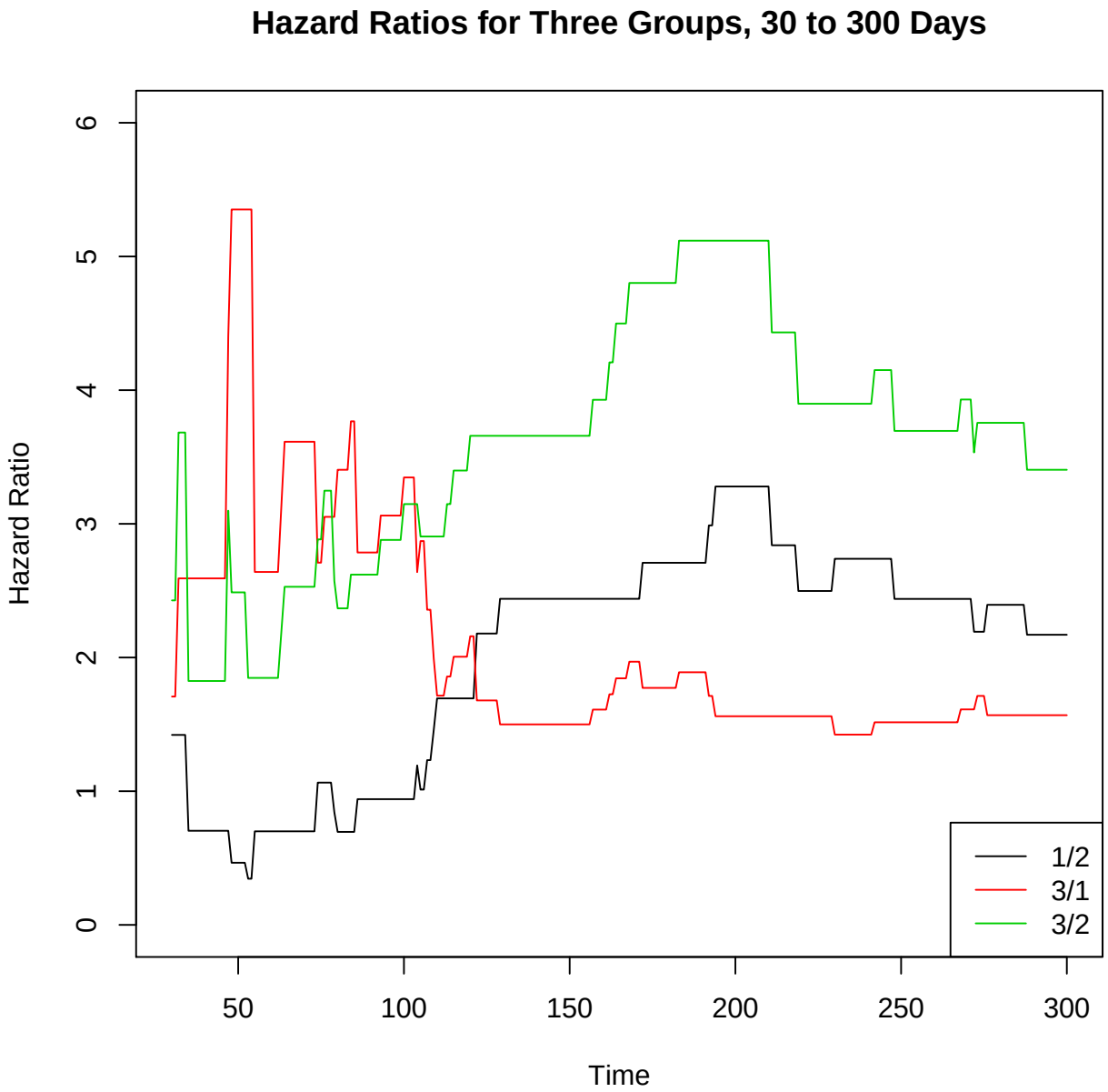
Nelson-Aalen Survival Function Estimate

The Nelson-Aalen estimate of the cumulative hazard is usually used for estimates of the hazard and often the cumulative hazard.

If the hazards of the three groups are proportional, that means that the ratio of the hazards is constant over t. We can test this using the ratios of the estimated cumulative hazards, which also would be proportional.

```
nafit <- survfit(dfsurv~group,type="fleming-harrington",data=bmt)
timevec <- 1:1000
sf1 <- stepfun(nafit[1]$time,c(1,nafit[1]$surv))
sf2 <- stepfun(nafit[2]$time,c(1,nafit[2]$surv))
sf3 <- stepfun(nafit[3]$time,c(1,nafit[3]$surv))
cumhaz1 <- -log(sf1(timevec))
cumhaz2 <- -log(sf2(timevec))
cumhaz3 <- -log(sf3(timevec))

plot(timevec,cumhaz1/cumhaz2,type="l",ylab="Hazard Ratio",xlab="Time",ylim=c(0,6))
lines(timevec,cumhaz3/cumhaz1,ylab="Hazard Ratio",xlab="Time",col=2)
lines(timevec,cumhaz3/cumhaz2,ylab="Hazard Ratio",xlab="Time",col=3)
legend("bottomright",c("1/2","3/1","3/2"),col=1:3,lwd=1)
title("Hazard Ratios for Three Groups")
```


Hazard Ratios for Three Groups
Hazard Ratio
0
1
2
3
4
5
6
0
200
400
600
800
1000
Time
1/2
3/1
3/2


Hazard Ratios for Three Groups, 30 to 300 Days
Hazard Ratio
0
1
2
3
4
5
6
50
100
150
200
250
300
Time
1/2
3/1
3/2

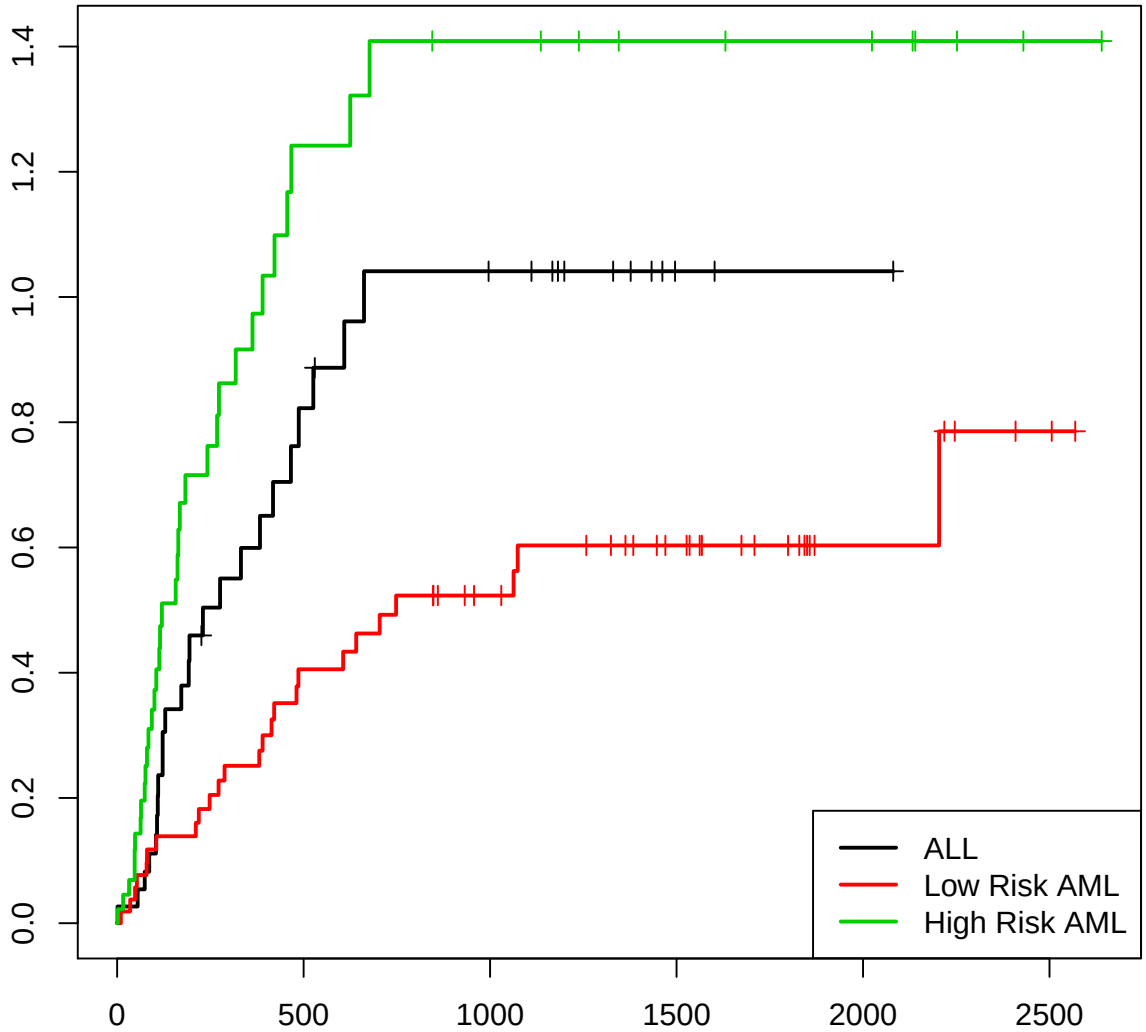
The Nelson-Aalen estimate of the cumulative hazard is usually used for estimates of the hazard. Since the hazard is the derivative of the cumulative hazard, we need a smooth estimate of the cumulative hazard, which is provided by smoothing the step-function cumulative hazard.

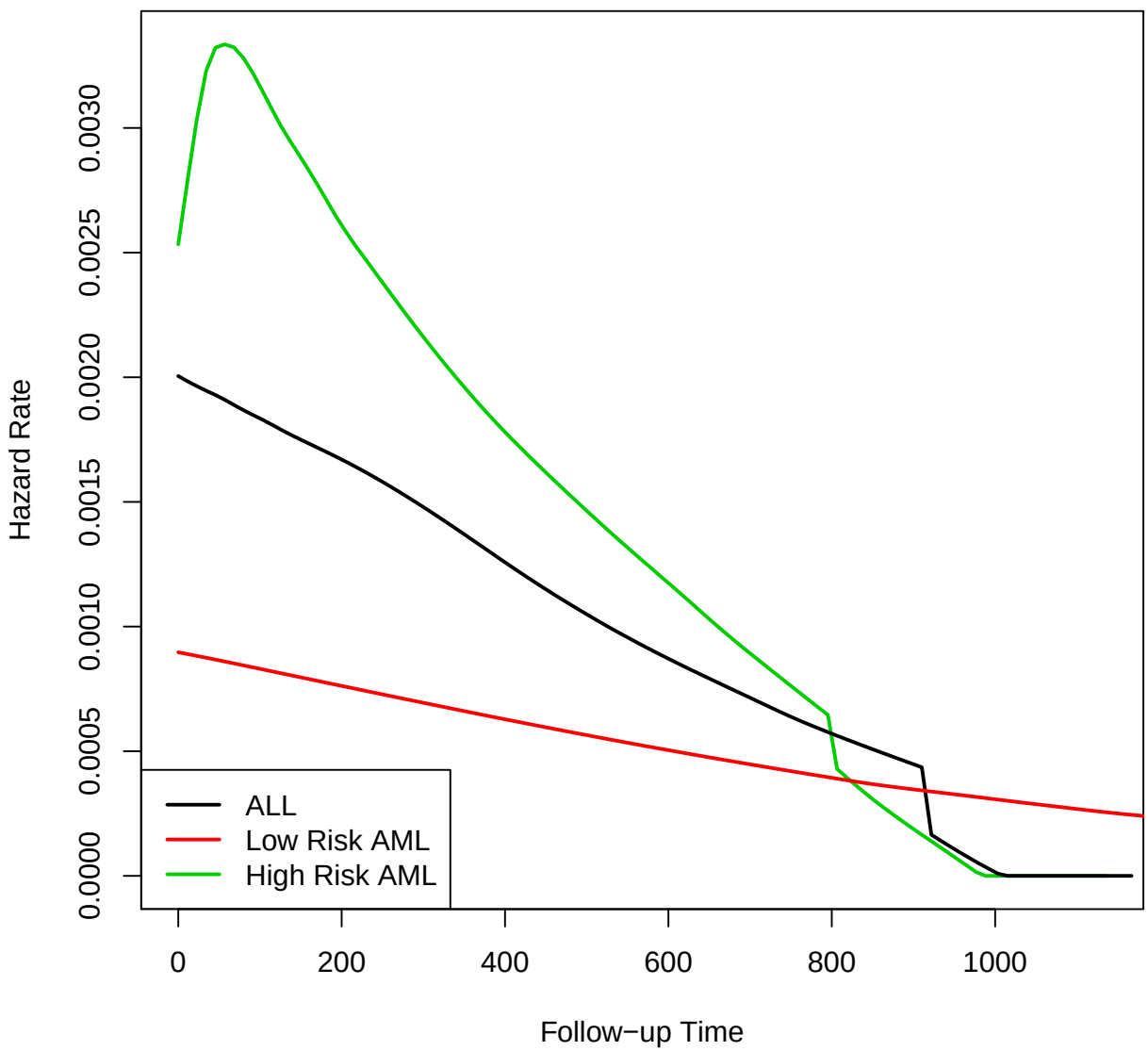
The R package `muhaz` handles this for us. What we are looking for is whether the hazard function is more or less the same shape, increasing, decreasing, constant, etc. Are the hazards "proportional"?

```
> library(muhaz)

> plot(muhaz(bmt$t2,bmt$d3,bmt$group==3),lwd=2,col=3)
> lines(muhaz(bmt$t2,bmt$d3,bmt$group==1),lwd=2,col=1)
> lines(muhaz(bmt$t2,bmt$d3,bmt$group==2),lwd=2,col=2)
> legend("bottomleft",c("ALL","Low Risk AML","High Risk AML"),col=1:3,lwd=2)
> title("Smoothed Hazard Rate Estimates for Three Groups")
```

Group 3 was plotted first because it has the highest hazard. We could also have set the ylim value in plot.

We will see that except for an initial blip in the high risk AML group, the hazards look roughly proportional . They are all strongly decreasing.


Disease−Free Cumulative Hazard for Three Groups
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
0
500
1000
1500
2000
2500
ALL
Low Risk AML
High Risk AML


Smoothed Hazard Rate Estimates for Three Groups
Hazard Rate
0.0000
0.0005
0.0010
0.0015
0.0020
0.0025
0.0030
0
200
400
600
800
1000
Follow−up Time
ALL
Low Risk AML
High Risk AML