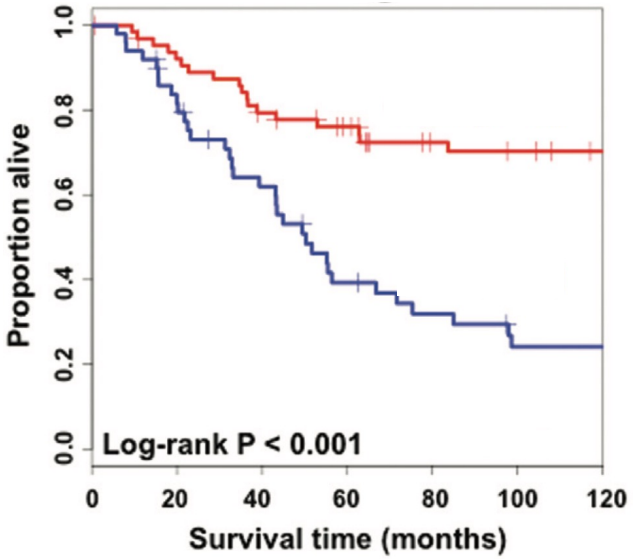
Survival Analysis Part One

Parcours recherche Clinique — UMR3 Advanced Biostatistics



François Grolleau

October 27, 2021

Université de Paris

METHODS Team

Centre de Recherche Epidémiologie et StatistiqueS
Université de Paris - Inserm UMR 1153

What is Survival Analysis?

aka “Time-To-Event Analysis”

Not just the analysis of survival data

Death or any other event (i.e hospital readmission, disease relapse…)

Not merely interested in whether the outcome happened
(≠ logistic regression), but also when it happened

Survival Analysis Challenges

Over the course of a long follow-up some patients drop out
e.g. as a result of annoying questionnaires or blood samples

Some other patients are simply lost to follow-up

Then you don't know if the outcome occurred (death registries are not always reliable and it's even more problematic for non-death outcomes)

→ Survival analysis handles **censoring**

Quiz

1. Survival analysis can be applied to
   - studies that run over many months/years
   - cross sectional studies

2. We are mainly interested to know **when/whether** a particular outcome has occurred for each of the patients involved over the study period

3. Examples of outcomes survival analysis deals with
   - onset of speech from birth
   - time to cancer relapse
   - hospital discharge after kidney transplant
   - sex at birth
   - treatment choice after myocardial infarction

Quiz

1. Survival analysis can be applied to
   - **studies that run over many months/years**
   - cross sectional studies

2. We are mainly interested to know **when**/whether a particular outcome has occurred for each of the patients involved over the study period

3. Examples of outcomes survival analysis deals with
   - **onset of speech from birth**
   - **time to cancer relapse**
   - **hospital discharge after kidney transplant**
   - sex at birth
   - treatment choice after myocardial infarction

The Kaplan Meier plot and Log-rank test

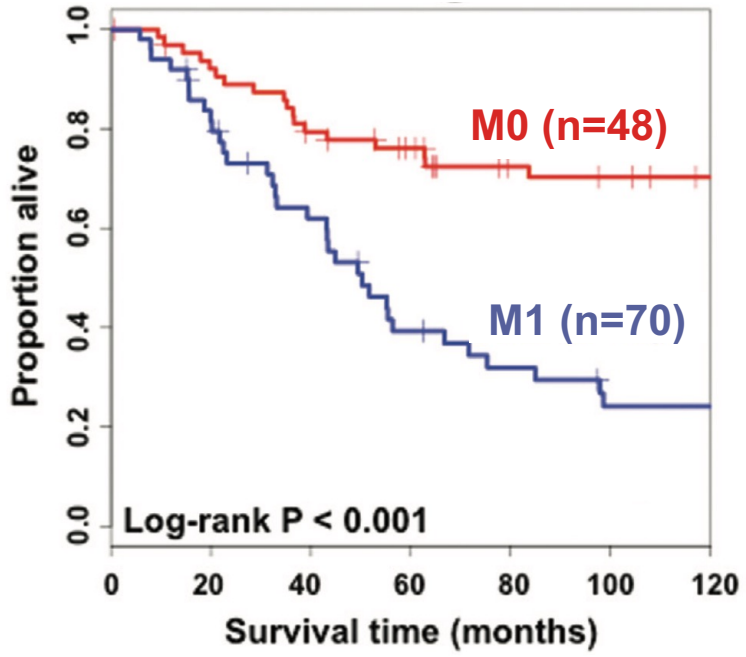
Like Newton & Liebnitz, Edward Kaplan & Paul Meier
had the same great idea at the same time independently. JASA 1958.

Estimates the probability of surviving at least to any given time point
= **“survival function”**

e.g. patients with metastatic breast cancer
or not, after tumorectomy

- Line drops each time a patient dies
- Vertical tick each time a patient is **censored**

- Log-rank test compares survival curves



How to calculate a Kaplan Meier table & plot by hand

Time (t) in days	Event
0 (study start)	8 patients recruited
1	2 patients die
4	1 patient dies
5	1 patient dies
6	1 patient drops out
9	1 patient dies and 1 drops out
22	1 patient dies

Time (t) in days	Nb of patients alive at time t	Nb of patients who died at time t	Proportion of patients surviving past time t	Probability of survival *past* time t
0 (study start)	8	0	(8-0)/8=1	1
1	8	2	(8-2)/8=0.75	0.75
4	6	1	(6-1)/6=0.83	0.75*0.83=0.623
5	5	1	(5-1)/5=0.8	0.623*0.8=0.498
6+	4	0	4/4=1	0.498*1=0.498
9	3	1	(3-1)/3=0.667	0.498*0.667=0.332
9+	2	0	2/2=1	0.332*1=0.332
22	1	1	(1-1)/1=0	0

Time (t) in days	Nb of patients alive at time t	Nb of patients who died at time t	Proportion of patients surviving past time t	Probability of survival *past* time t
0 (study start)	8	0	(8-0)/8=1	1
1	8	2	(8-2)/8=0.75	0.75
4	6	1	(6-1)/6=0.83	0.75*0.83=0.623
5	5	1	(5-1)/5=0.8	0.623*0.8=0.498
6+	4	0	4/4=1	0.498*1=0.498
9	3	1	(3-1)/3=0.667	0.498*0.667=0.332
9+	2	0	2/2=1	0.332*1=0.332
22	1	1	(1-1)/1=0	0

Let

$t = (t_1, t_2, \ldots, t_n)^T$ be the observed survival times of n patients

n_j be the number of patients at risk just before t_j

d_j be the number of events at t_j

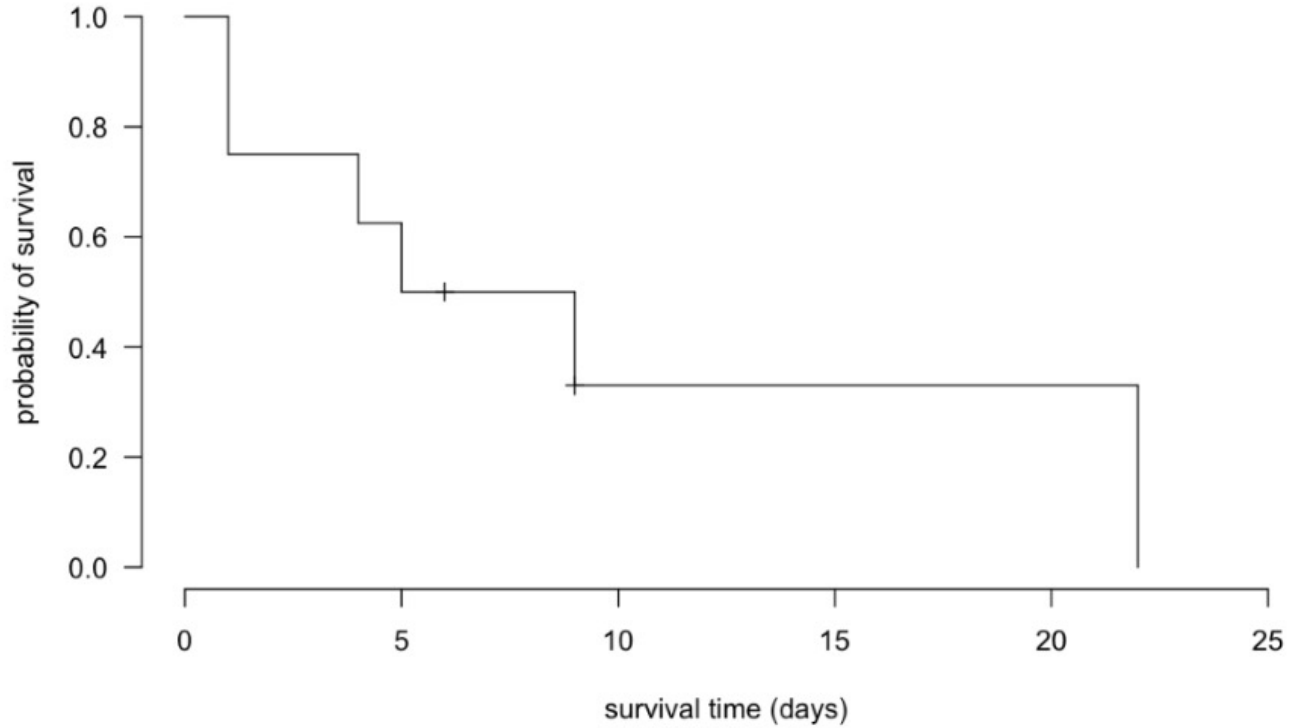
then

the proportion of patients surviving past time t_j is $\frac{n_j - d_j}{n_j}$

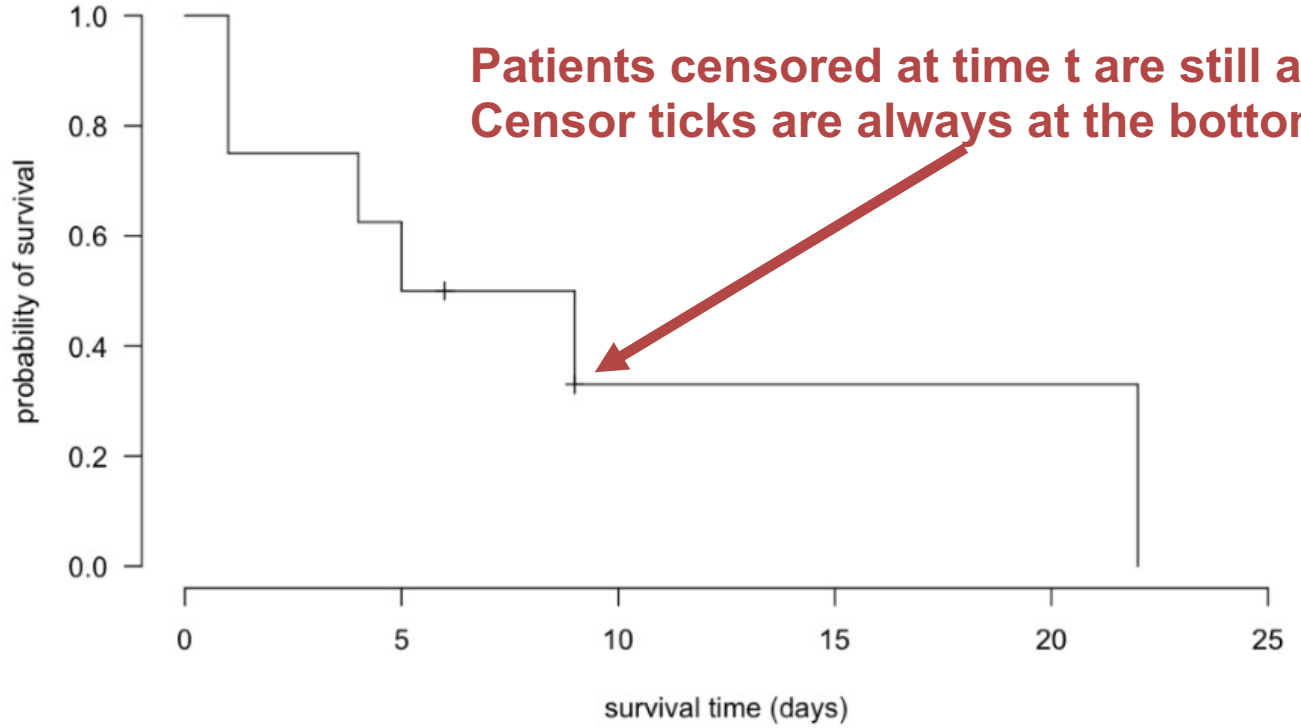
If the events occur independently from one another

then the probability of survival past time t_j is $\hat{S}(t_j) = \prod_{i=1}^{j} \frac{n_i - d_i}{n_i}$

Time (t) in days	Nb of patients alive at time t	Nb of patients who died at time t	Proportion of patients surviving past time t	Probability of survival *past* time t
0 (study start)	8	0	(8-0)/8=1	1
1	8	2	(8-2)/8=0.75	0.75
4	6	1	(6-1)/6=0.83	0.75*0.83=0.623
5	5	1	(5-1)/5=0.8	0.623*0.8=0.498
6+	4	0	4/4=1	0.498*1=0.498
9	3	1	(3-1)/3=0.667	0.498*0.667=0.332
9+	2	0	2/2=1	0.332*1=0.332
22	1	1	(1-1)/1=0	0

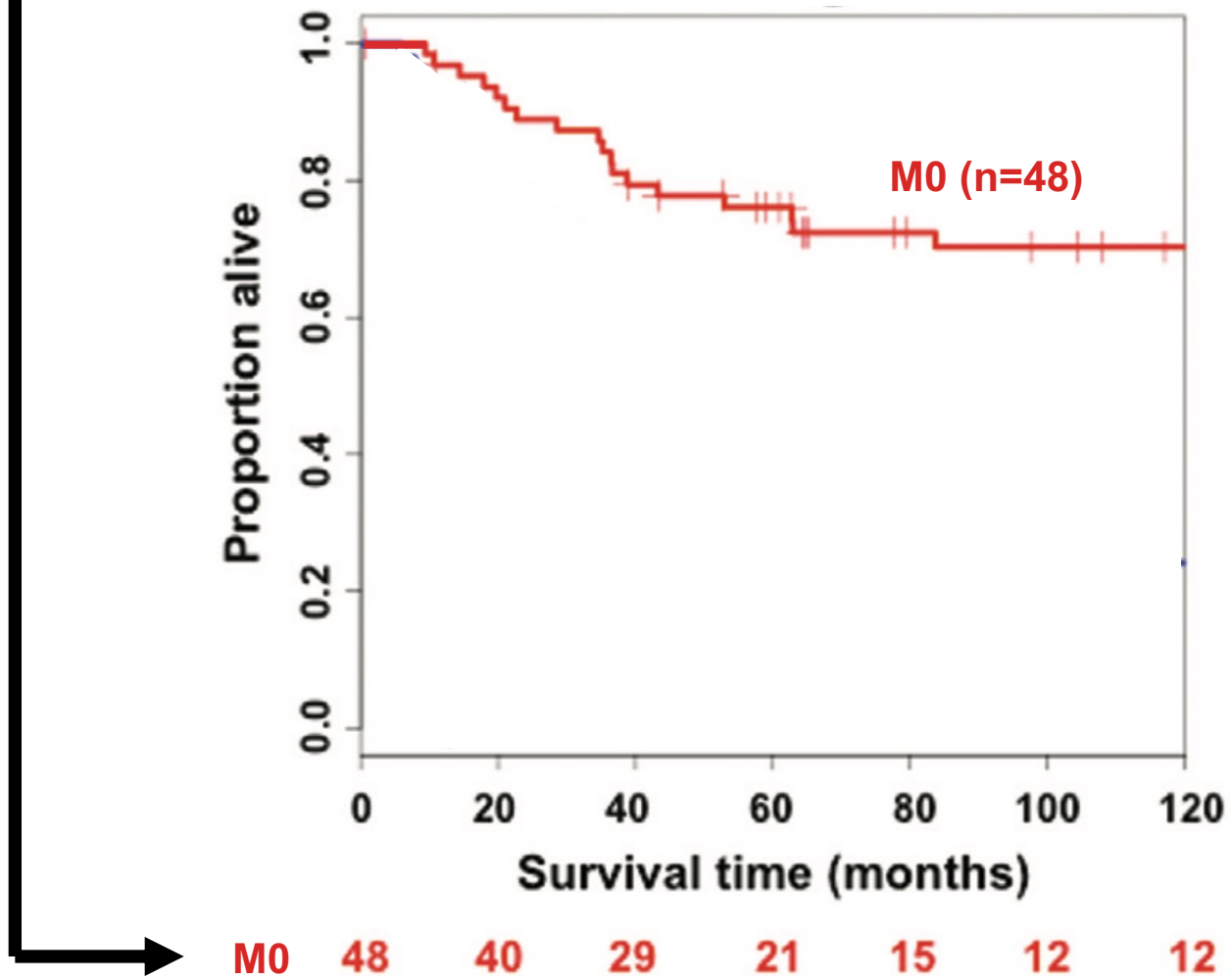


Time (t) in days	Nb of patients alive at time t	Nb of patients who died at time t	Proportion of patients surviving past time t	Probability of survival *past* time t
0 (study start)	8	0	(8-0)/8=1	1
1	8	2	(8-2)/8=0.75	0.75
4	6	1	(6-1)/6=0.83	0.75*0.83=0.623
5	5	1	(5-1)/5=0.8	0.623*0.8=0.498
6+	4	0	4/4=1	0.498*1=0.498
9	**3**	**1**	**(3-1)/3=0.667**	**0.498*0.667=0.332**
9+	**2**	**0**	**2/2=1**	**0.332*1=0.332**
22	1	1	(1-1)/1=0	0



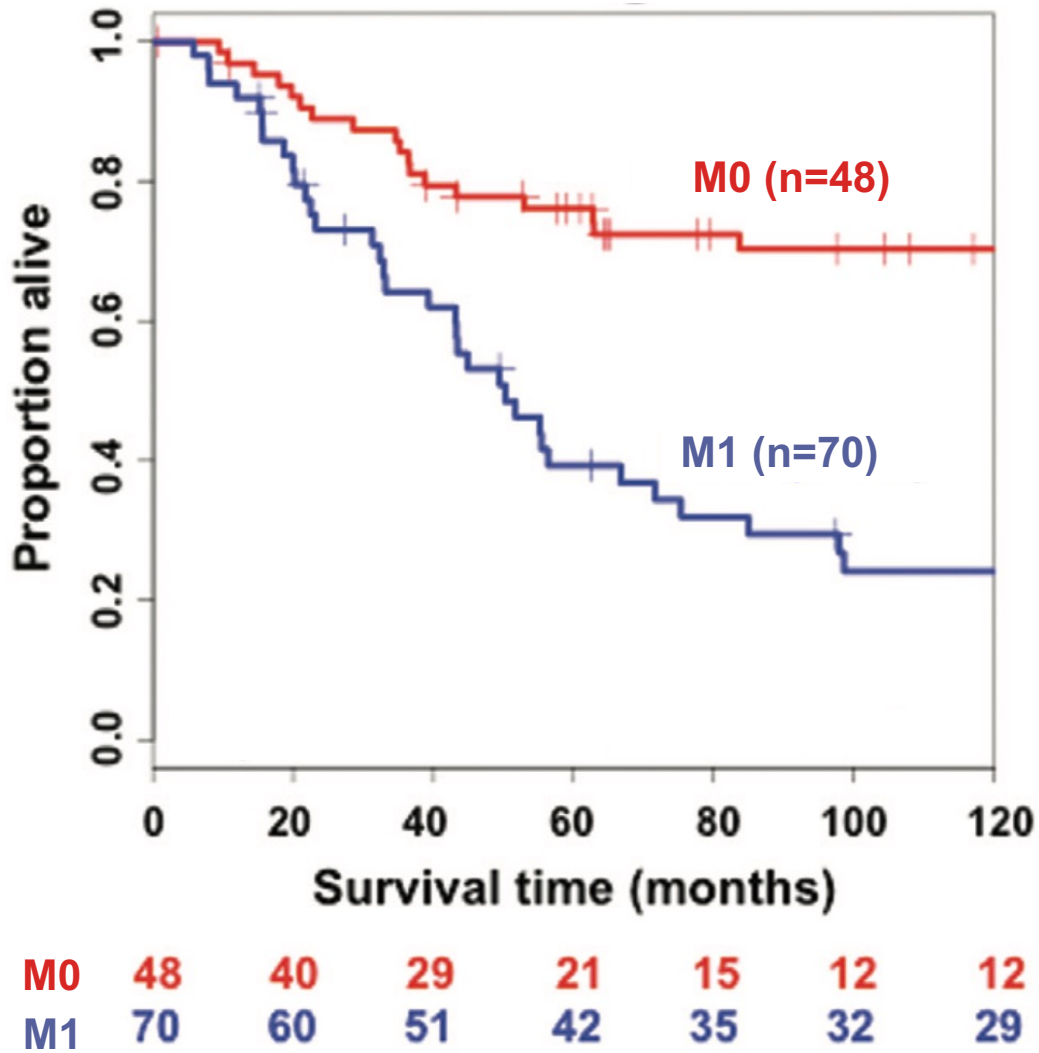
M0 Patients

Time (t) in months	Nb of patients alive at time t	Nb of patients who died at time t	Proportion of patients surviving past time t	Probability of survival *past* time t
0 (study start)	48	0	(48-0)/48=1	1
10	48	1	(48-1)/48=0.98	1*0.98



M1 Patients

Time (t) in months	Nb of patients alive at time t	Nb of patients who died at time t	Proportion of patients surviving past time t	Probability of survival *past* time t
0 (study start)	70	0	(70-0)/70=1	1
8	70	2	(70-2)/70=0.97	1*0.97



The Log-rank test

To compare the survival of groups

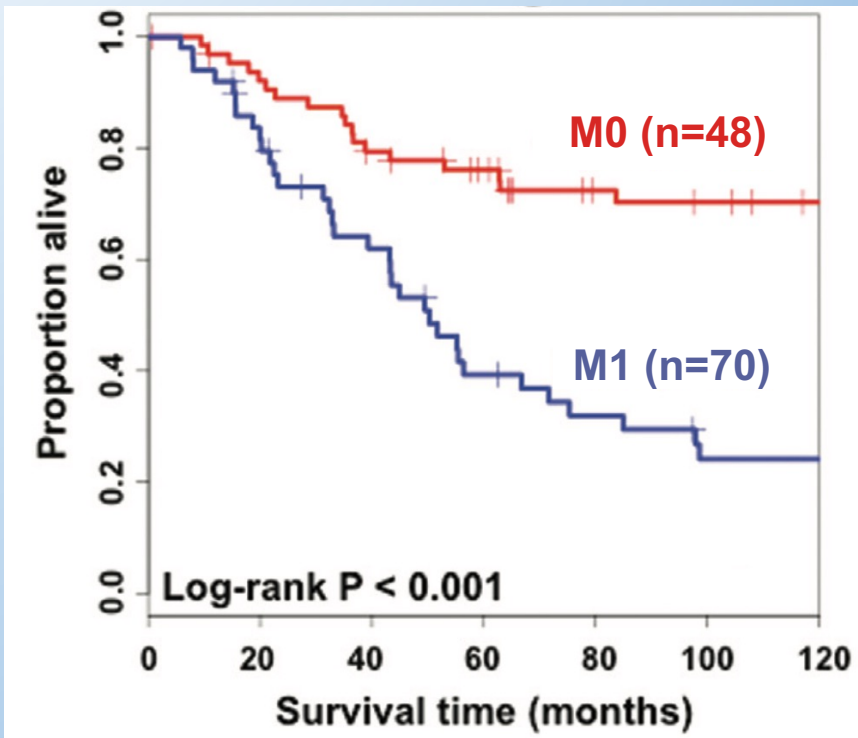
Takes the whole follow up period into account

Does not require us to know anything about the shape of the survival curve or the distribution of survival times

Test the null hypothesis that there is no difference between the populations in the probability of an event at any time point

Generalisation of the Chi-square test (df = nb groups - 1)

See Bland and Altman. The log rank test. BMJ. 2004.

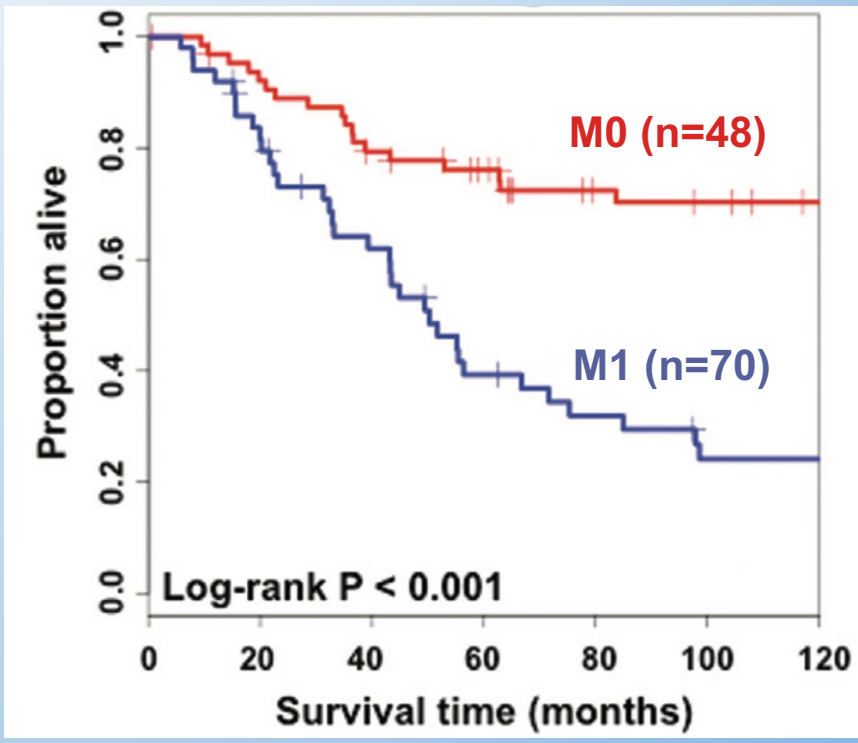


Kaplan Meier and Log-rank test assumptions

1. Survival probabilities are the same for subjects recruited early and late in the study
   i.e. does not hold if some patients were followed up with between 1940 and 1950 and others between 2010 and 2020

2. The events happened at the time specified
   i.e. does not hold if event status was assessed every month and we don't know when the events really occurred

3. Censoring is unrelated to prognosis
   i.e. does not hold if patients drop out when they are more likely to die



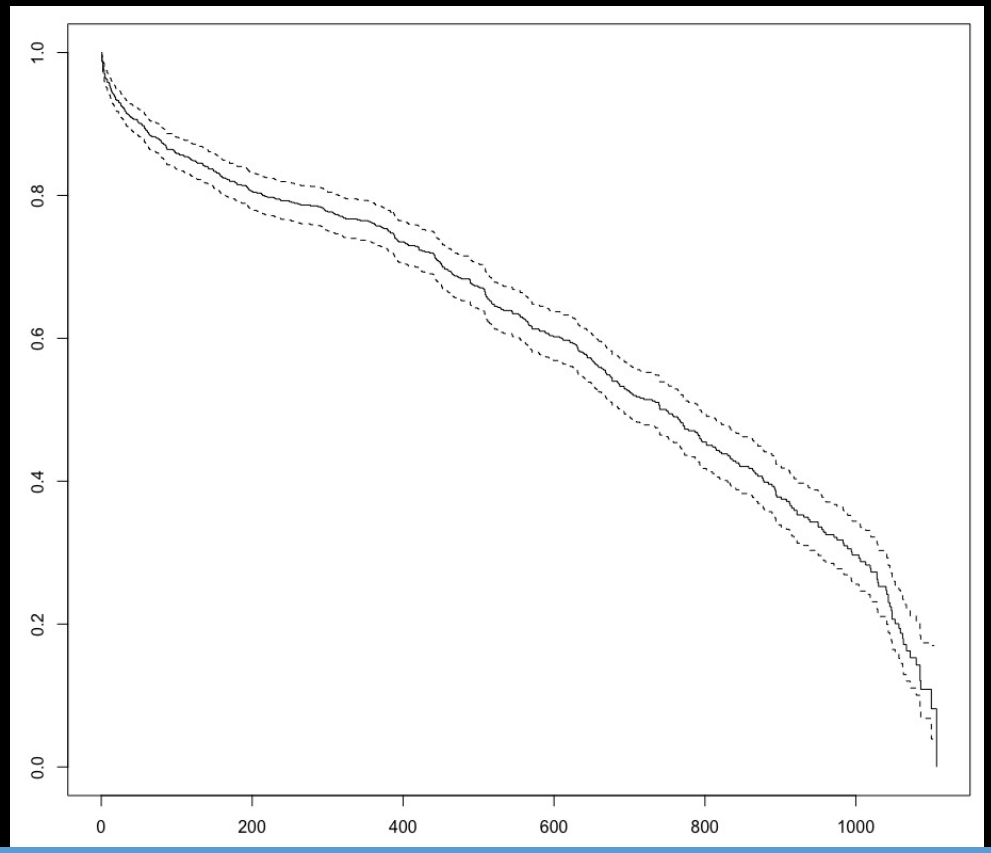
Let’s make our first Kaplan-Meier plot and tables in R

Get ready as follows

```
getwd() # tells you what your current working directory is

# load the dataset from your current working directory
hfdata <- read.csv(file = "hf_survival_data.csv", header=TRUE, sep=',')
dim(hfdata) # check the dataset dimensions
head(hfdata) # check the dataset first 6 rows
hfdata[35:41,] # check rows 35 to 41 of the dataset

install.packages("survival")

library(survival) # this is the cornerstone command for survival analysis in R

gender <- as.factor(hfdata[,"gender"]) # R calls categorical variables factors
fu_time <- hfdata[,"fu_time"] # continuous variable (numeric)
death <- hfdata[,"death"] # binary variable (numeric)
```

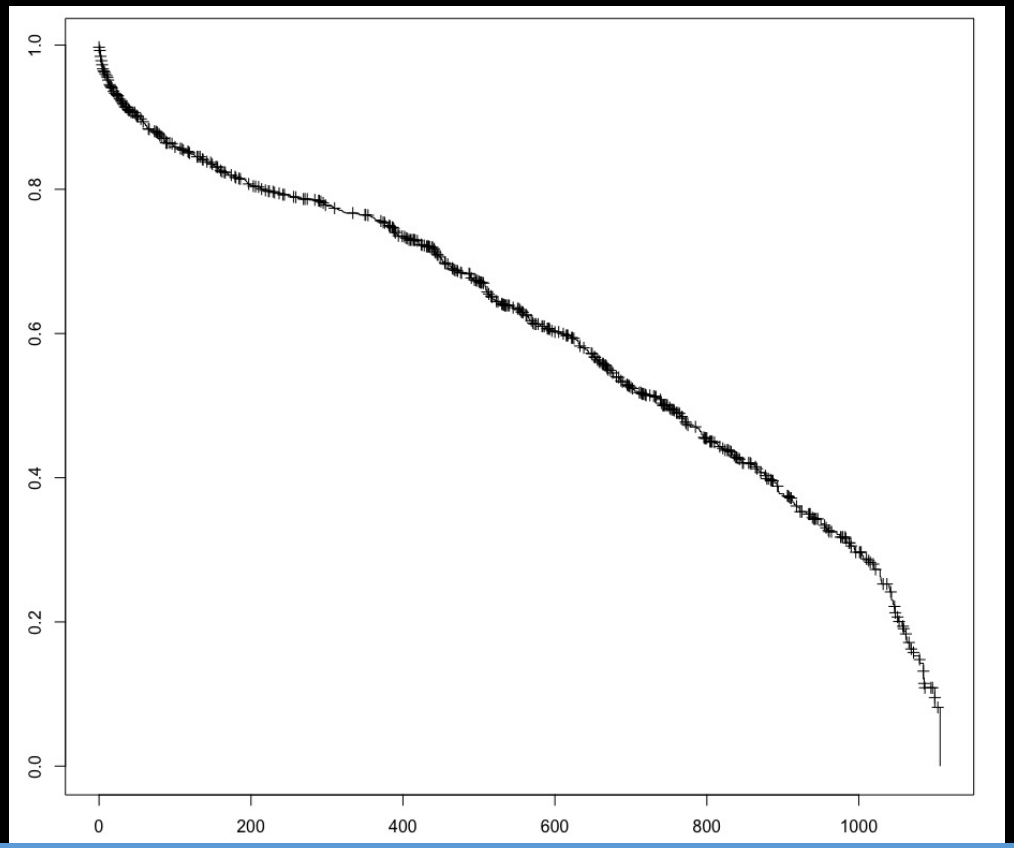
Kaplan-Meier Plot in R

```
km_fit <- survfit(Surv(fu_time, death) ~ 1)
plot(km_fit)
```



Kaplan-Meier Plot in R

```
plot(km_fit, conf.int = F, mark.time = T)
```



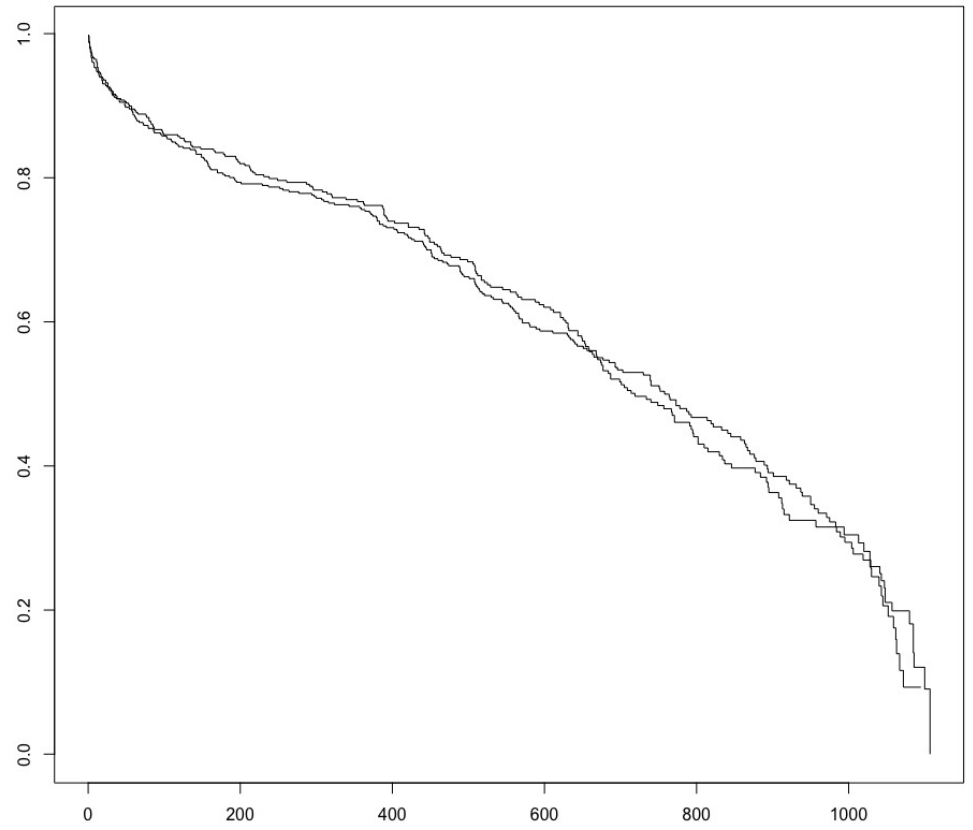
Kaplan-Meier Table in R

```
summary(km_fit, times = c(1:7,30,60,90*(1:10)))

    Call: survfit(formula = Surv(fu_time, death) ~ 1)

 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    1    992      12    0.988 0.00346        0.981        0.995
    2    973       7    0.981 0.00435        0.972        0.989
    3    963       5    0.976 0.00489        0.966        0.985
    4    954       6    0.970 0.00546        0.959        0.980
    5    945       5    0.964 0.00590        0.953        0.976
    6    938       1    0.963 0.00598        0.952        0.975
    7    933       1    0.962 0.00606        0.951        0.974
   30    865      39    0.921 0.00865        0.905        0.939
   60    809      28    0.891 0.01010        0.871        0.911
   90    770      24    0.864 0.01117        0.843        0.887
  180    698      43    0.815 0.01282        0.790        0.841
  270    653      24    0.787 0.01363        0.760        0.814
  360    619      21    0.761 0.01428        0.733        0.789
  450    525      44    0.705 0.01554        0.675        0.736
  540    429      47    0.639 0.01681        0.607        0.673
  630    362      32    0.589 0.01765        0.556        0.625
  720    266      43    0.514 0.01876        0.479        0.552
  810    190      31    0.448 0.01979        0.411        0.488
  900    126      26    0.378 0.02098        0.339        0.421
```

Print only selected time points

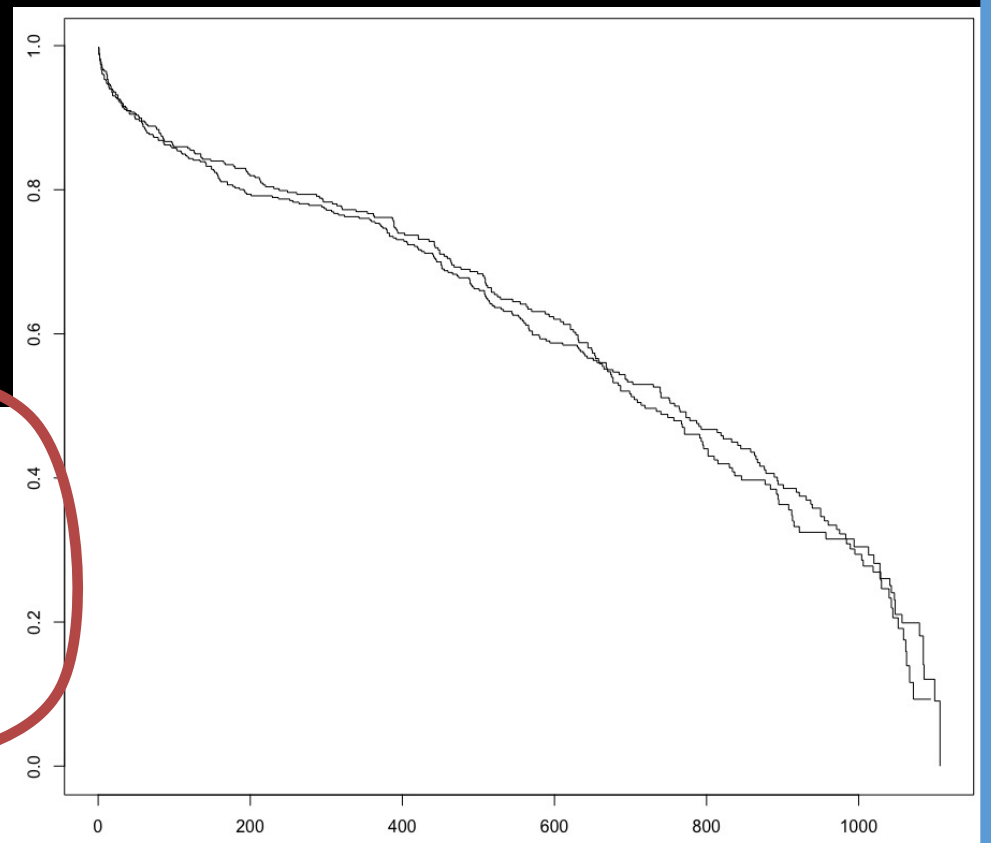
Splitting the curve by gender

```
km_gender_fit <- survfit(Surv(fu_time, death) ~ gender)
plot(km_gender_fit)
```



Log rank test to compare survival by gender

```
survdiff(Surv(fu_time, death) ~ gender) # runs log-rank test by gender


Call:
survdiff(formula = Surv(fu_time, death) ~ gender)

            N Observed Expected (O-E)^2/E (O-E)^2/V
gender=1 548      268      271    0.0365     0.082
gender=2 452      224      221    0.0448     0.082

 Chisq= 0.1  on 1 degrees of freedom, p= 0.8
```



Number of patient in each group

Number of observed death in each group

Number of expected death in each group under H0

χ^2 df = n groups - 1 P-value

Compare the survival times for patients 65 and over with those under 65

```
age_65plus <- ifelse(hfdata[,"age"]>=65,1,0) # dichotomise age
table(age_65plus) # inspect the numbers - always a good idea
age_65plus
table(hfdata$age, age_65plus) # check - an even better idea...
```

```
     age_65plus
       0  1
  29   2  0
  30   1  0
  33   1  0
  36   1  0
   ▪  ▪  ▪
  62  11  0
  63  18  0
  64  10  0
  65   0 16
  66   0 16
  67   0  7
  68   0 18
   ▪  ▪  ▪
  98   0  2
  99   0  6
  100  0  1
  101  0  1
  102  0  1
```

Data preparation worked OK

Log rank test to compare survival by age category

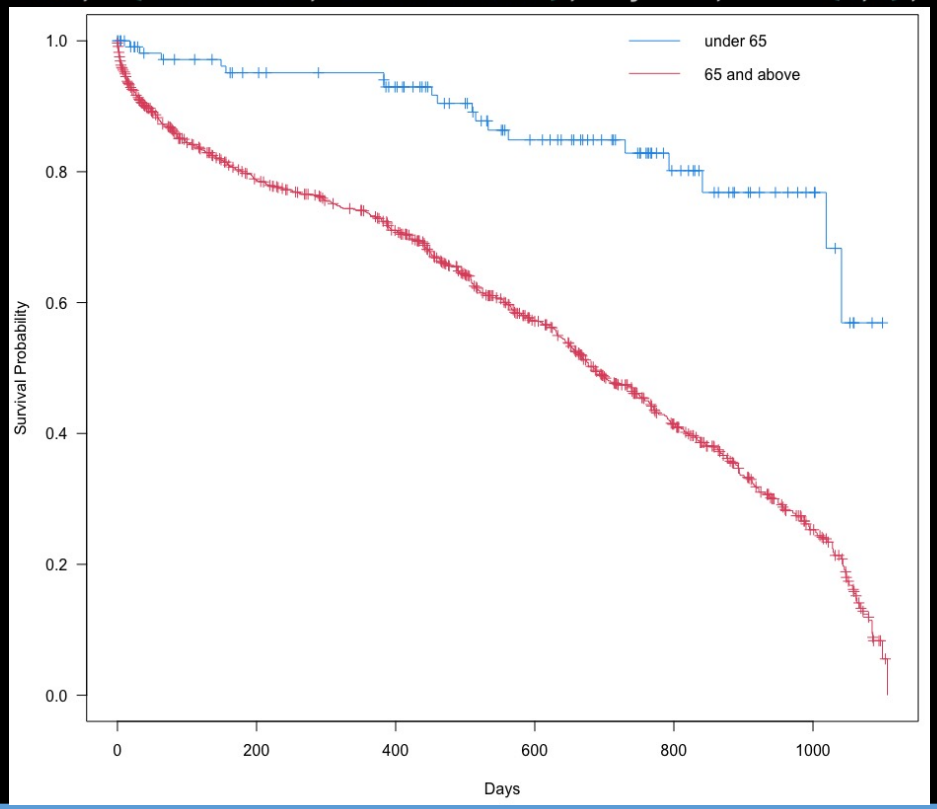
```
survdiff(Surv(fu_time, death) ~ age_65plus) # runs log-rank test by age>=65


        Call:
        survdiff(formula = Surv(fu_time, death) ~ age_65plus)

                       N Observed Expected (O-E)^2/E (O-E)^2/V
        age_65plus=0 115       18       67     35.85      41.7
        age_65plus=1 885      474      425      5.65      41.7

         Chisq= 41.7  on 1 degrees of freedom, p= 1e-10
```

NOT SO FAST!

we need to plot the survival curves to interpret the P-value

Let's make nice Kaplan-Meier Plots

```r
# plot survival curves by age category (65 and above vs under 65)
survfit.obj<-survfit(Surv(fu_time, death) ~ age_65plus)
plot(survfit.obj, col=c(4,2), mark.time=TRUE,
     xlab = "Days", ylab="Survival Probability",
     las=1)

# add a legend
legend(700, 1.05, c("under 65","65 and above"), lty = 1, col=c(4,2), bty="n")
```

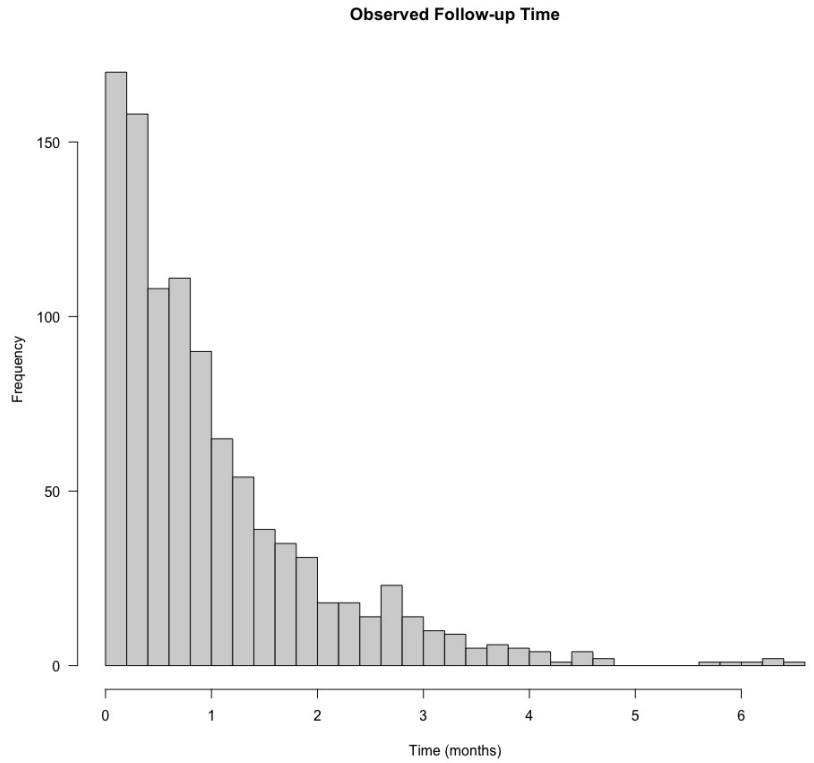


Key Concepts 1/3: Censoring

Let

T be the time to event,
C be the time to censorship,
X be the observed follow-up time

$$X_i = min(T_i, C_i).$$



Key Concepts 1/3: Censoring

The survival time of an individual is censored when the outcome has not been observed for that individual

- Right censoring
  - i.e. patient known to have survived up to a timepoint (end of the study, drop out, lost to follow-up)
- Left censoring
  - i.e. patient known to have died before a given timepoint
- Interval censoring
  - i.e. patient known to have died between two timepoints

Key Concepts 2/3: The Survival Function

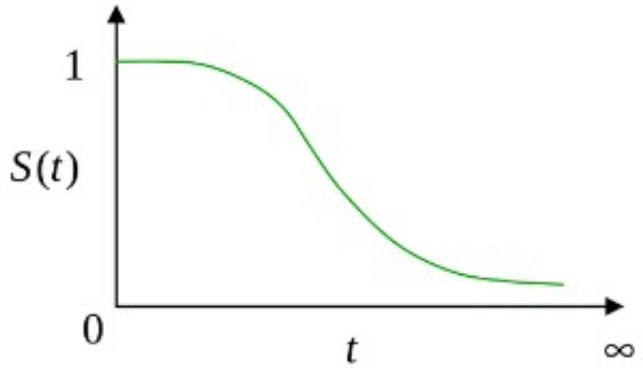
Let T be a random variable corresponding to survival times; $T \geq 0$

The survival function $S(t)$ is the probability of surviving at least to any given time point t

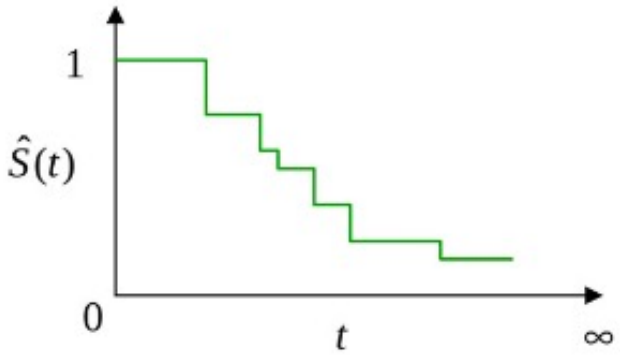
$$S(t) = P(T \geq t)$$

Proprieties of $S(t)$

- Monotonically decreasing
- $S(0) = 1$
- $\lim_{t \to \infty} S(t) = 0$

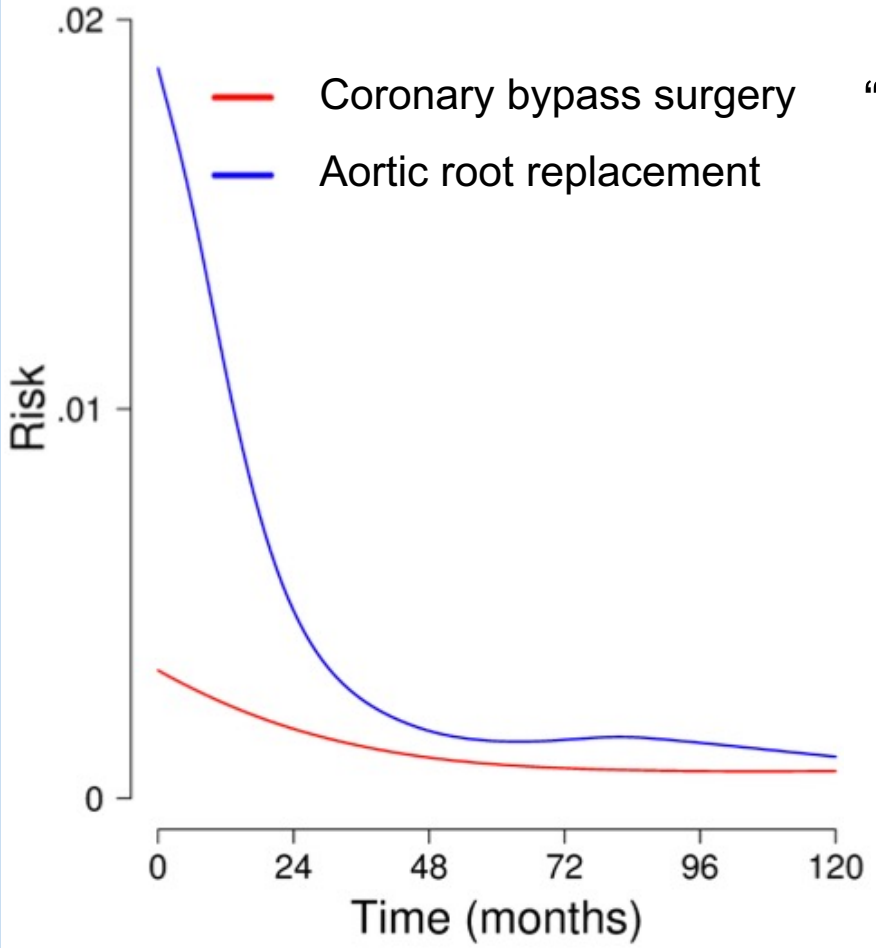


True unknown function



Kaplan Meier estimation

Key Concepts 3/3: The Hazard Function



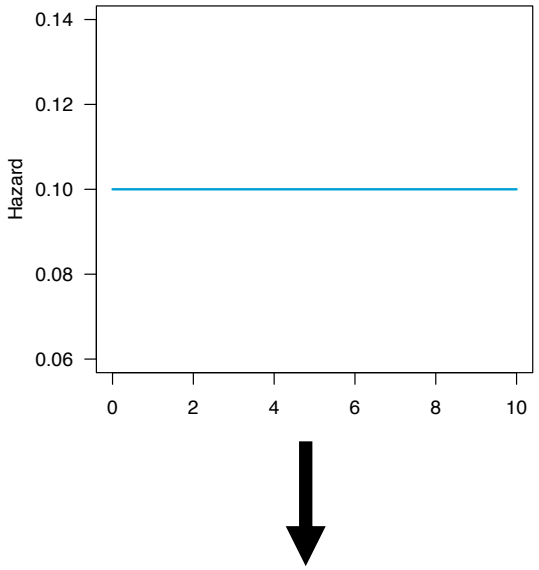
The hazard function $\lambda(t)$ is the "instantaneous rate of occurrence of the event at time t"

$\lambda(t)$ can never be directly measured but assuming it exists is useful for further computations…

$$\lambda(t) = \lim_{dt \to 0} \frac{P(t \leq T < t + dt \mid T \geq t)}{dt}$$

$\lambda(t)$ is not a probability but a risk

Relationship Between the Hazard and Survival Functions



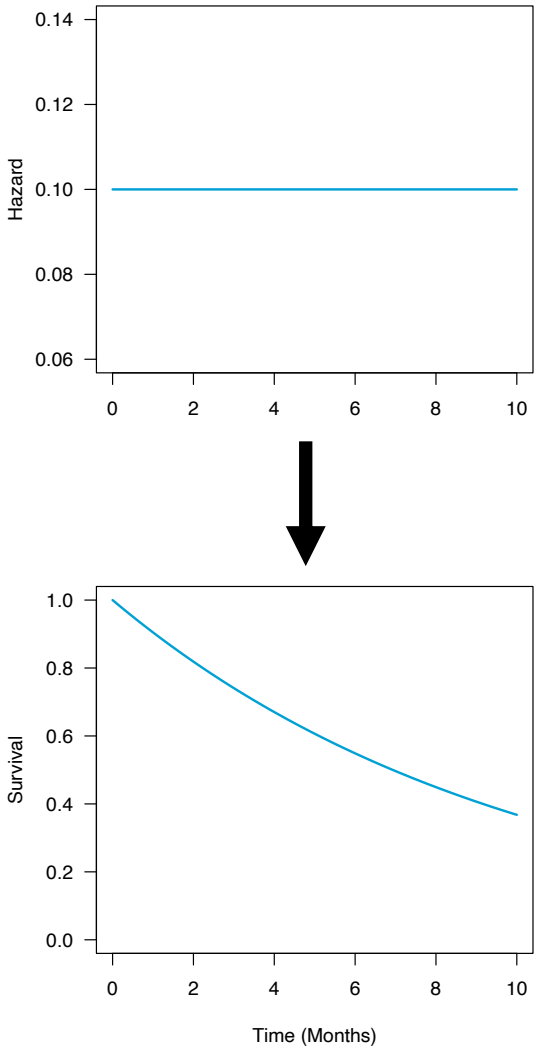
$$\lambda(t) = -\frac{d}{dt}\log S(t)$$

and therefore

$$S(t) = \exp\{-\int_0^t \lambda(x)\,dx\}$$

If interested, more details here,
https://data.princeton.edu/wws509/notes/c7s1

Relationship Between the Hazard and Survival Functions



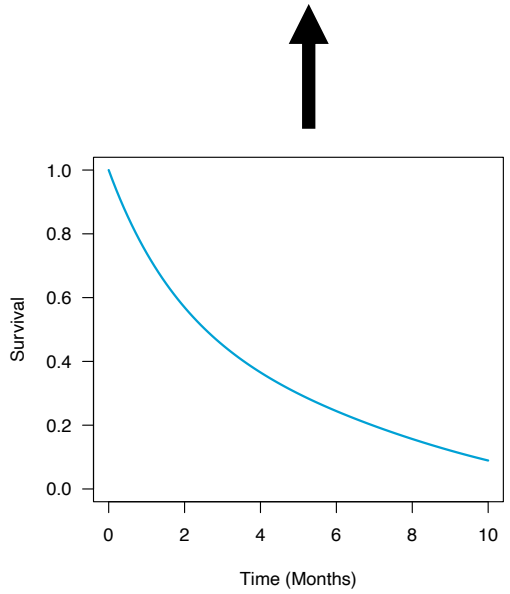
$$\lambda(t) = -\frac{d}{dt}\log S(t)$$

and therefore

$$S(t) = \exp\{-\int_0^t \lambda(x)\,dx\}$$

If interested, more details here,
https://data.princeton.edu/wws509/notes/c7s1

Relationship Between the Hazard and Survival Functions



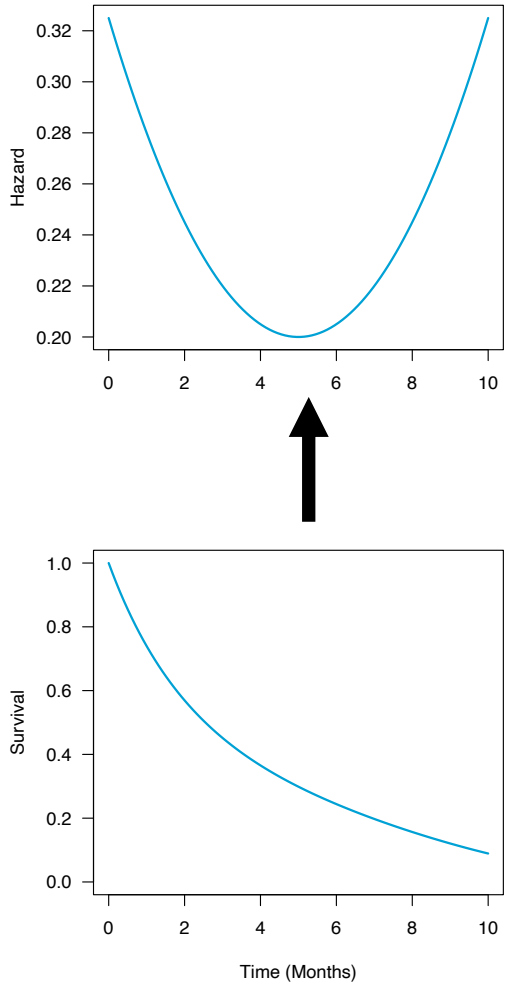
$$\lambda(t) = -\frac{d}{dt}\log S(t)$$

and therefore

$$S(t) = \exp\{-\int_0^t \lambda(x)\,dx\}$$

If interested, more details here,
https://data.princeton.edu/wws509/notes/c7s1

Relationship Between the Hazard and Survival Functions



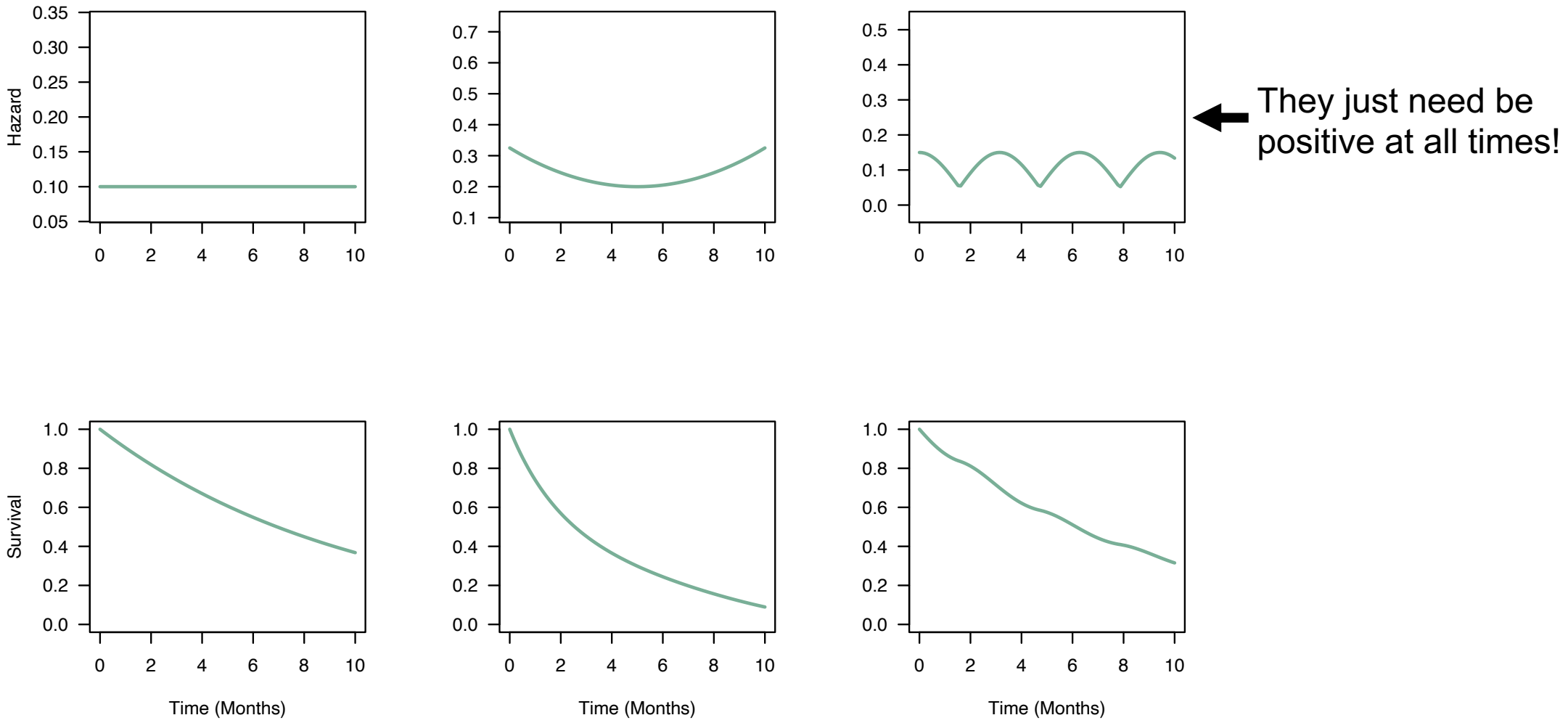
$$\lambda(t) = -\frac{d}{dt}\log S(t)$$

and therefore

$$S(t) = \exp\{-\int_0^t \lambda(x)\,dx\}$$

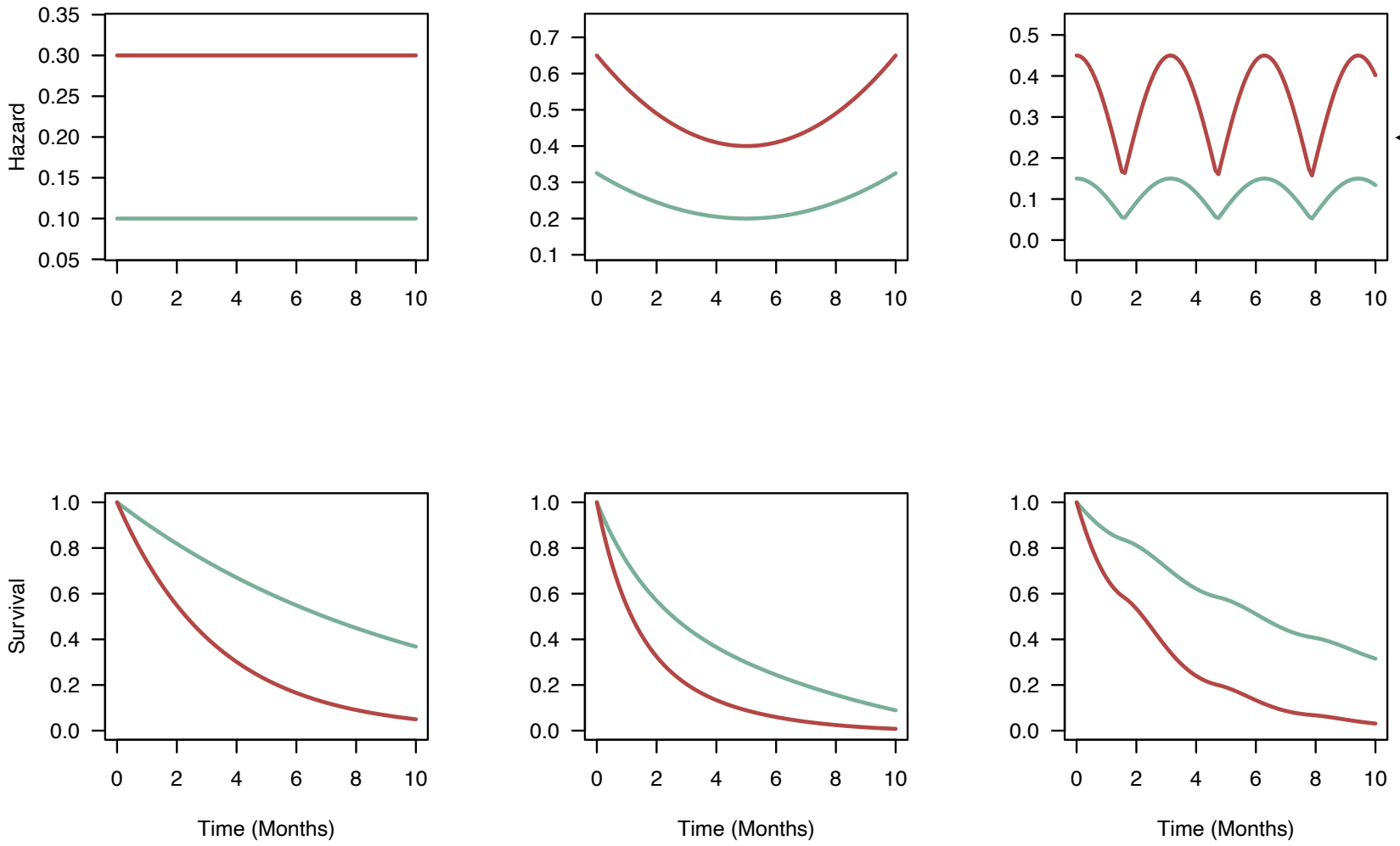
If interested, more details here,
https://data.princeton.edu/wws509/notes/c7s1

Hazard Functions Can Take Any Shape…



They just need be positive at all times!

Two population of patients with proportional hazards at all times



Hazard Ratios can be derived from Cox regression

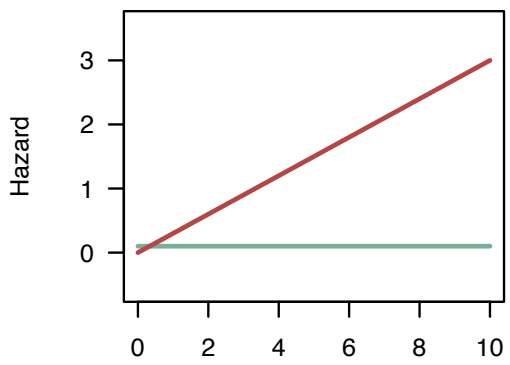
As opposed to Relative Risks, Hazard Ratios are provided independent of time

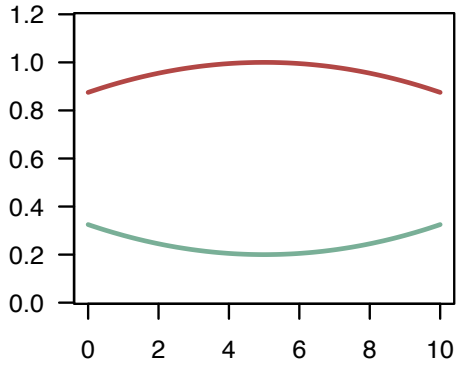
e.g., HR for death

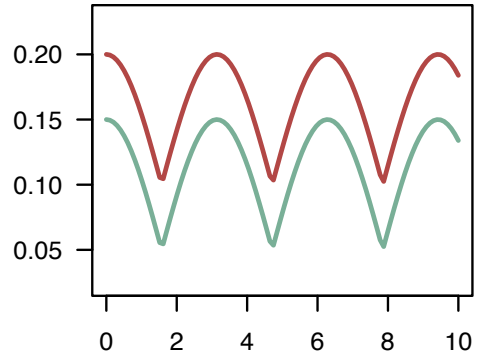
vs

RR for death at day 60

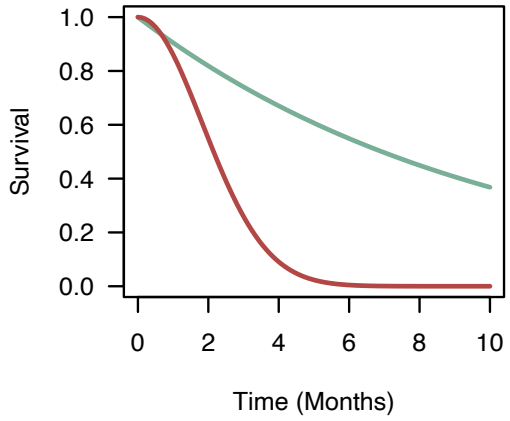
Two population of patients with NON proportional hazards functions

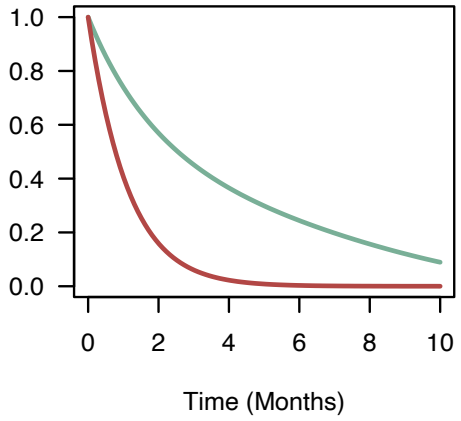


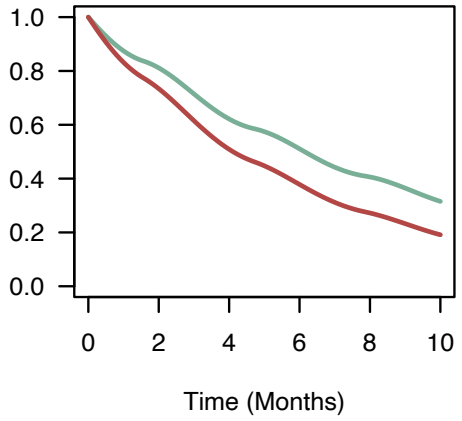




Hazard Ratios from Cox regression are incorect!







Cox Model and Proportional Hazard assumption will be covered next week ☺

References

- Suvival Analysis in R for Public Health by Alex Bottle (Imperial College London). https://www.coursera.org/learn/survival-analysis-r-public-health
- Bland M (2015) An Introduction to Medical Statistics (4th ed). Oxford University Press.
- Moore D (2016) Applied Survival Analysis Using R. Springer.

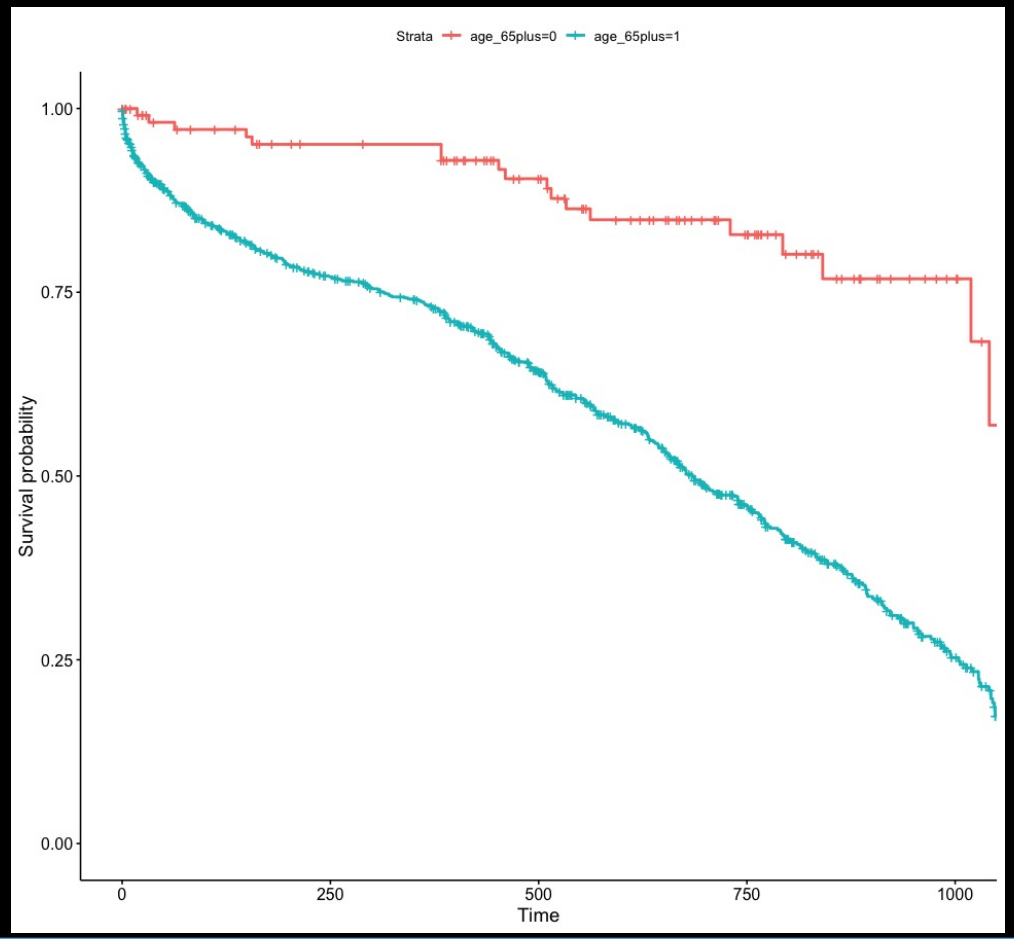
Let’s make our first Kaplan-Meier plot and tables in R

Get ready as follows

```
getwd() # tells you what your current working directory is

# load the dataset from your current working directory
hfdata <- read.csv(file = "hf_survival_data.csv", header=TRUE, sep=',')
dim(hfdata) # check the dataset dimensions
head(hfdata) # check the dataset first 6 rows
hfdata[35:41,] # check rows 35 to 41 of the dataset

install.packages("survival")
install.packages("ggplot")
install.packages("survminer")

library(survival) # this is the cornerstone command for survival analysis in R
library(ggplot2) # newer package that does nice plots
library(survminer) # newer package that does nice kaplan meir tables and plots

gender <- as.factor(hfdata[,"gender"]) # R calls categorical variables factors
fu_time <- hfdata[,"fu_time"] # continuous variable (numeric)
death <- hfdata[,"death"] # binary variable (numeric)
```

Let’s make nice Kaplan-Meier Plots

```
survfit.obj<-survfit(Surv(fu_time, death) ~ age_65plus)
splots<-ggsurvplot(survfit.obj, data=hfdata) # need to fill the data argument
splots # first basic ggsurvplot
```



We can do better…

```
splots<-ggsurvplot(survfit.obj, data=hfdata,
                   ggtheme = theme_survminer() +
                             theme(plot.title = element_text(hjust = 0.5)),
                   title    = "Heart Failure Prognosis",
                   font.title=12,
                   legend.title = "Age 65 and over",
                   legend.labs = c("No", "Yes"),
                   pval=TRUE, pval.method = T,
                   risk.table = TRUE,
                   risk.table.fontsize = 5,
                   break.time.by = 120,
                   tables.theme = theme_cleantable(),
                   tables.y.text = F)

splots$table <- splots$table + labs(title = "", subtitle = "No. at risk")
splots<- splots + labs(x  = "Days", y = "Overall Survival")
```

We can do better…

splots