Kaplan-Meier analysis: introduction

The Kaplan-Meier estimator is given by the following formula:

$$\hat{S}(t) = \prod_{t_i \leq t} (1 - \hat{q}_i) = \prod_{t_i \leq t} \left(1 - \frac{d_i}{n_i}\right)$$

$\hat{S}(t)$ denoting the survival function at a time t, n_i i is the number of subjects at risk at time t_i and lastly d_i is the number of individuals who 'fail' at time t_i. More technical detail found in the book 'Applied Survival Analysis Using R', essentially in chapter three.

We note that for confidence bounds (as displayed on the subsequent plots), the variance is obtainedn by the formula below (more details in chapter three):

$$var\left(log\left[-log\hat{S}(t)\right]\right) \approx \frac{1}{\left[log\hat{S}(t)\right]^2} \sum_{t_i \leq t} \frac{d_i}{n_i(n_i - d_i)}$$

'Veteran' dataset

veteran: dataset of 137 observations x 8 variables form a two-treatment randomized trial for lung cancer.

trt: 1=standard 2=test
celltype: 1=squamous, 2=smallcell, 3=adeno, 4=large
time: survival time
status: censoring status
karno: Karnofsky performance score (100=good)
diagtime: months from diagnosis to randomisation
ageA: in years
prior: prior therapy 0=no, 10=yes

```
> head(veteran)
  trt celltype time status karno diagtime age prior
1   1 squamous   72      1    60        7  69     0
2   1 squamous  411      1    70        5  64    10
3   1 squamous  228      1    60        3  38     0
4   1 squamous  126      1    60        9  63    10
5   1 squamous  118      1    70       11  65    10
6   1 squamous   10      1    20        5  49     0
```

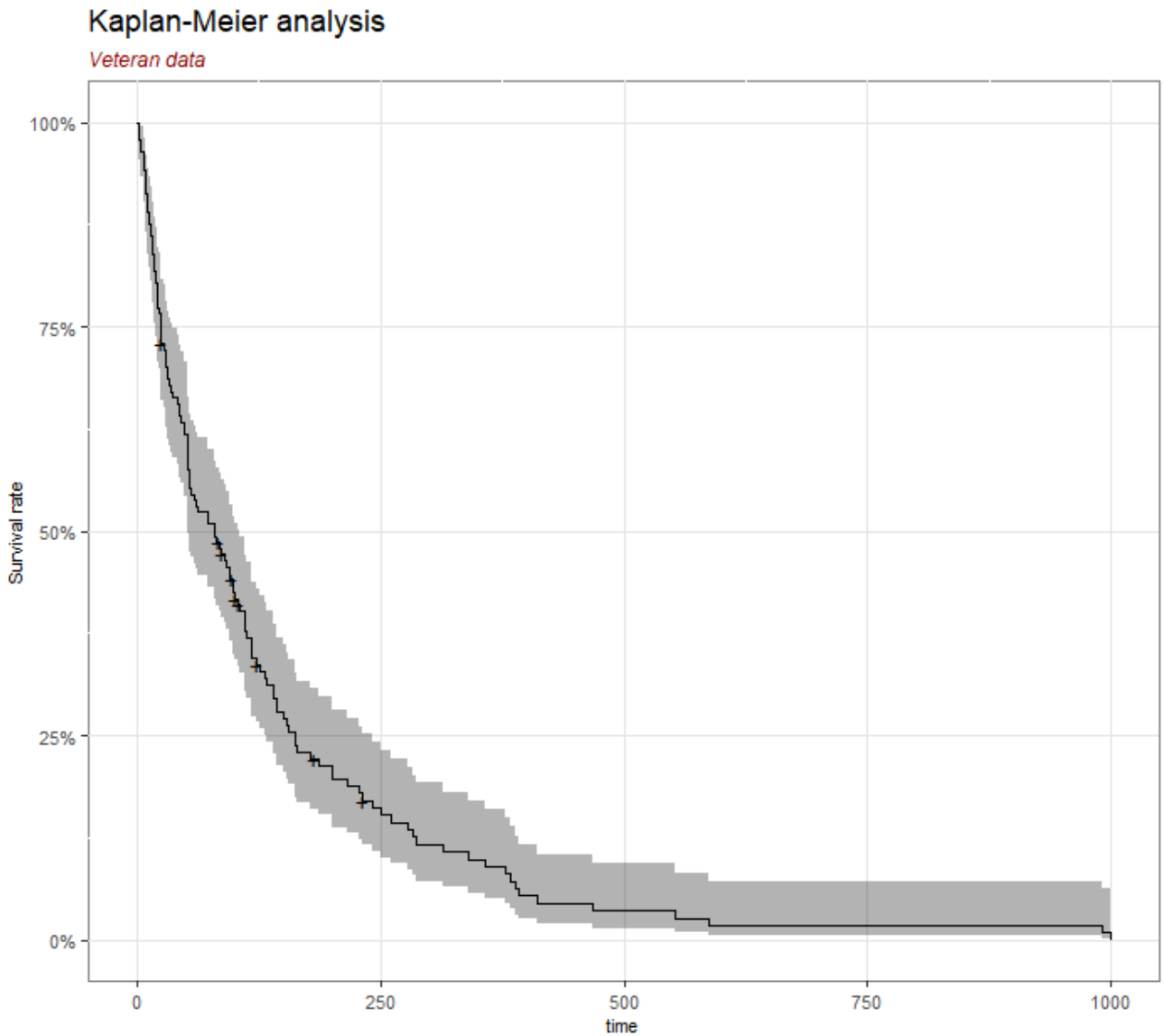
Summary

From this initial summary which has class "summary survfit", we will make a dataframe ready for ploting withing the ggplot2 environment.

```
> survival.30.dataset = summary(kma_1, times = c(1, (1:33)*30))
> survival.30.dataset
Call: survfit(formula = Surv(time, status) ~ 1, data = veteran)

 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    1    137       2    0.985  0.0102       0.96552        1.0000
   30     97      39    0.700  0.0392       0.62774        0.7816
   60     73      22    0.538  0.0427       0.46070        0.6288
   90     62      10    0.464  0.0428       0.38731        0.5560
  120     43      15    0.346  0.0414       0.27345        0.4372
  150     34       8    0.280  0.0395       0.21240        0.3693
  180     27       7    0.222  0.0369       0.16066        0.3079
  210     23       3    0.197  0.0355       0.13814        0.2802
  240     19       3    0.171  0.0338       0.11613        0.2520
  270     16       3    0.144  0.0319       0.09338        0.2223
  300     13       3    0.117  0.0295       0.07147        0.1917
  330     12       1    0.108  0.0285       0.06439        0.1813
  360     10       2    0.090  0.0265       0.05061        0.1602
...
```

Kaplan-Meier analysis using "autoplot()"

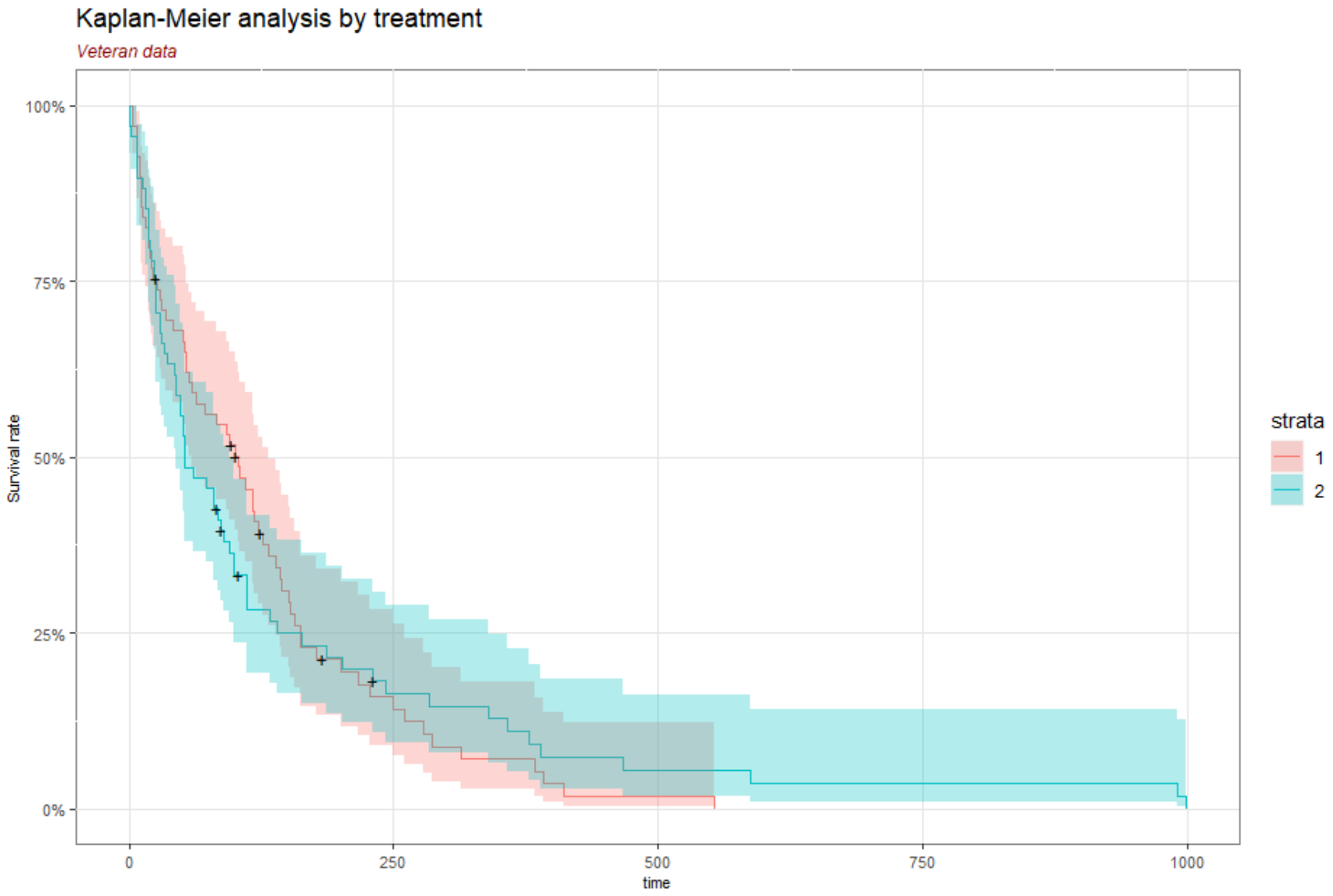


	time	surv
1	1	0.9854
2	30	0.7004
3	60	0.5382
4	90	0.464
5	120	0.3458
6	150	0.2801
7	180	0.2224
8	210	0.1967
9	240	0.1711
10	270	0.1441
11	300	0.1171
12	330	0.1081
13	360	0.09
14	390	0.063
15	420	0.045
16	450	0.045
17	480	0.036
18	510	0.036
19	540	0.036
20	570	0.027

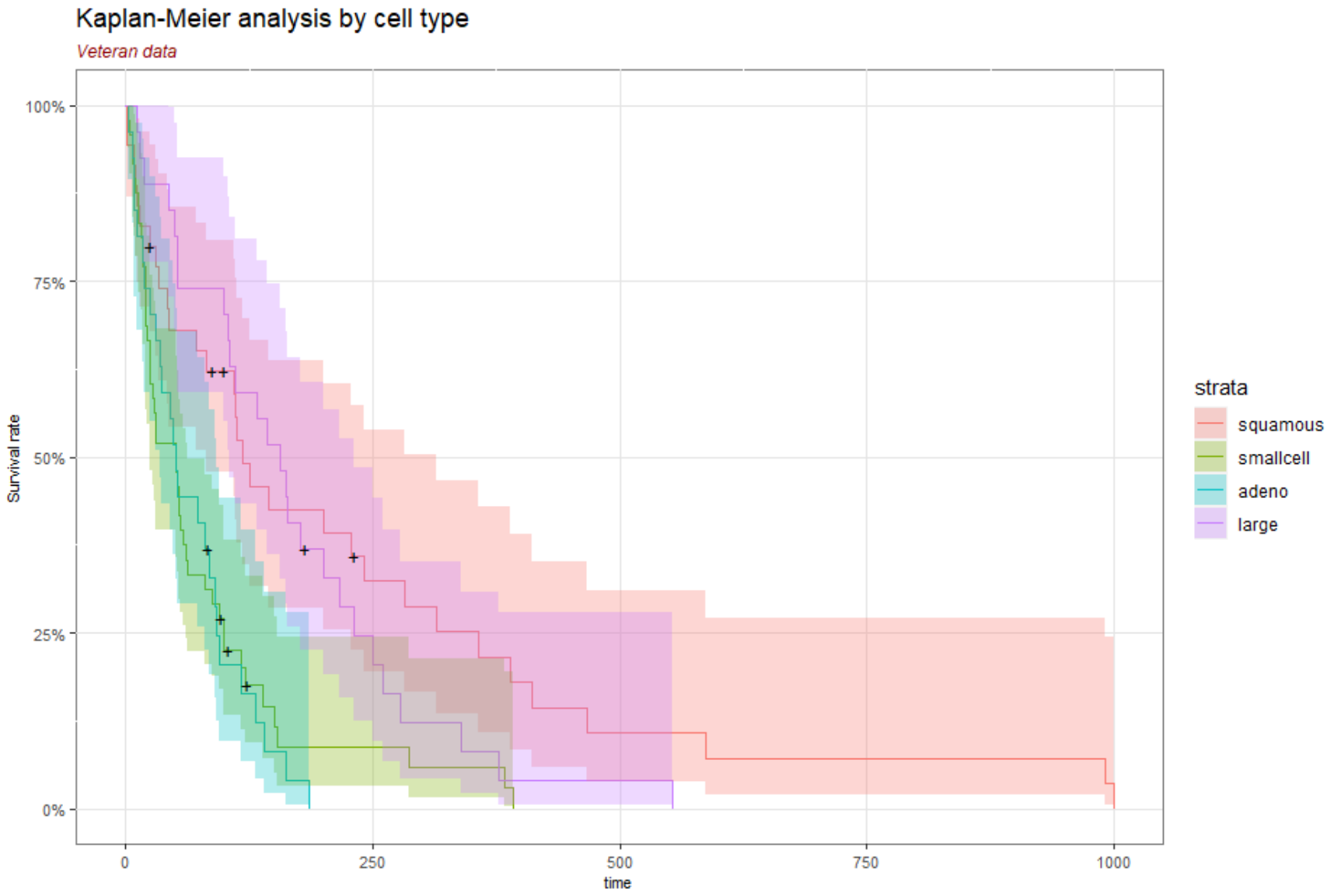
Main observations

- At the first month (after 30 days), the survival rate or probability of survival is about 70%.
- There seems to be some kind of breakup point at 6 months (after 180 days) as the slope gets less steep.
- After on year, the survival rate is lower than 10%. A patient has a 10% or less probability of surving one year.
- Information about censoring (a vertical line on the Kaplan-Maier survival function) is obtained in R using "autoplot()".

Kaplan-Meier analysis by treatment



Kaplan-Meier analysis by cell type



Main observations

- Treatment stratum 2 has the overall better survival rate with a better suvival curve (overall less steep)
- Cell type "squamous" has the overall better survival rate compared to small cell, adeno and lage.
- Stratification or dividing the initial sample into homogeneous subsamples help us refine the analyses.
- Kaplan-Meier estimator is a non-parametric inferential method. Cox proportional hazard regression allows for multiple explanatory variables. We will explore this family of regression methods in as next topic.

R code (1/4) - Load libraries and data cleaning

```r
#load libraries
library(survival)
library(ggplot2)
library(gridExtra)

# load data and head of the dataset
data(veteran)
head(veteran)

# "+" indicates censoring (incomplete information due to the event of interest
      not being observed)
km = with(veteran, Surv(time, status))
head(km, 100)

# Kaplan-Meier estimates of the probability of survival over time
kma_1 = survfit(Surv(time, status) ~ 1, data=veteran)
# max time: 999 days (about 33 months (30 days))
survival.30.dataset = summary(kma_1, times = c(1, (1:33)*30))
# convert summary to data.frame for plotting
cols = lapply(1:15 , function(x) survival.30.dataset[x])
df = do.call(data.frame, cols)
# clean initial data frame
df = df[, 1:6]
# table to be displayed next to the graph as a second graph
df2 = df[1:20, c(2,6)]
df2$surv = round(df2$surv, 4)
```

R code (2/4) - K-M Analysis

```
# Plot using 'autoplot()' and ggplot2 customization (and with information about
    censoring)
p3 = autoplot(kma_1) +
  labs(title = 'Kaplan-Meier analysis',
       subtitle = 'Veteran data',
       y="Survival rate", x="time") +
  theme(axis.text=element_text(size=8),
        axis.title=element_text(size=8),
        plot.subtitle=element_text(size=9, face="italic", color="darkred"),
        panel.background = element_rect(fill = "white", colour = "grey50"),
        panel.grid.major = element_line(colour = "grey90"))

p4 = tableGrob(df2) # to have a table with time and survival rate
grid.arrange(p3, p4, ncol = 2, nrow = 1, widths = c(6, 2))
```

R code (3/4) - Analysis by treatment

```r
# Analysis by treatment

# Kaplan-Meier estimates of the probability of survival over time
kma_3 = survfit(Surv(time, status) ~ trt, data=veteran)
# max time: 999 days (about 33 months (30 days))
survival.30.dataset.celltype = summary(kma_3, times = c(1, (1:33)*30))

# plotting
autoplot(kma_3) +
  labs(title = 'Kaplan-Meier analysis by treatment',
       subtitle = 'Veteran data',
       y="Survival rate", x="time") +
  theme(axis.text=element_text(size=8),
        axis.title=element_text(size=8),
        plot.subtitle=element_text(size=9, face="italic", color="darkred"),
        panel.background = element_rect(fill = "white", colour = "grey50"),
        panel.grid.major = element_line(colour = "grey90"))
```

R code (4/4) - Analysis by cell type

```r
# Analysis by cell type

# Kaplan-Meier estimates of the probability of survival over time
kma_2 = survfit(Surv(time, status) ~ celltype, data=veteran)
# max time: 999 days (about 33 months (30 days))
survival.30.dataset.celltype = summary(kma_2, times = c(1, (1:33)*30))

# plotting
autoplot(kma_2) +
  labs(title = 'Kaplan-Meier analysis by cell type',
       subtitle = 'Veteran data',
       y="Survival rate", x="time") +
  theme(axis.text=element_text(size=8),
        axis.title=element_text(size=8),
        plot.subtitle=element_text(size=9, face="italic", color="darkred"),
        panel.background = element_rect(fill = "white", colour = "grey50"),
        panel.grid.major = element_line(colour = "grey90"))
```

References

Survival Analysis with R, by Joseph Rickert, 2017-09-25, link to the article on R-views:
https://rviews.rstudio.com/2017/09/25/survival-analysis-with-r/

Applied Survival Analysis Using R, Dirk F. Moore, 2016, Springer, ISBN 978-3-319-31245-3 (e-book)

Introduction to Survival Analysis using scikit-survival, link to the article:
`https://scikit-survival.readthedocs.io/en/stable/user_guide/00-introduction.html`