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.

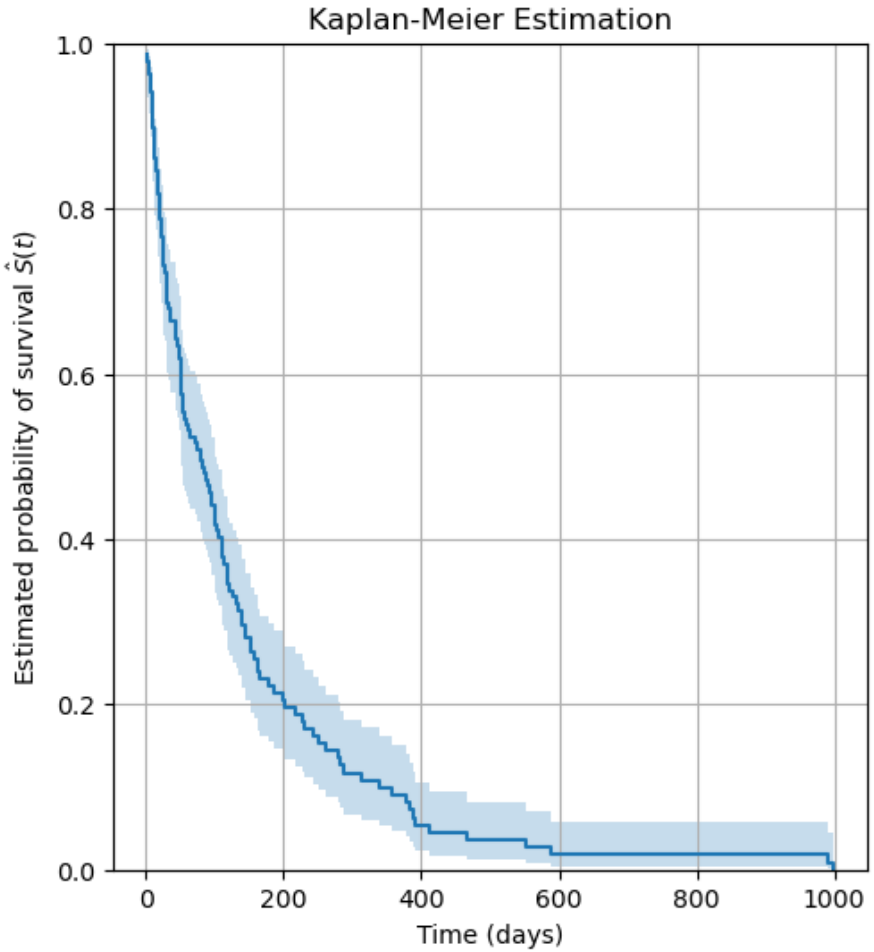
trt: 1=standard 2=test
celltype: 1=squamous, 2=smallcell, 3=adeno, 4=large
time: survival time
status: censoring status
karno: Karnofsky performance score (100=good)
diagtime: months from diagnosis to randomisation
ageA: in years
prior: prior therapy 0=no, 10=yes

```
from sksurv.datasets import load_veterans_lung_cancer
import pandas as pd

data_x, data_y = load_veterans_lung_cancer()
data_x

Age_in_years Celltype Karnofsky_score Months_from_Diagnosis Prior_therapy
      Treatment
0 69.0 squamous 60.0 7.0 no standard
1 64.0 squamous 70.0 5.0 yes standard
2 38.0 squamous 60.0 3.0 no standard
3 63.0 squamous 60.0 9.0 yes standard
4 65.0 squamous 70.0 11.0 yes standard
... ... ... ... ... ... ...
```

Kaplan-Meier analysis



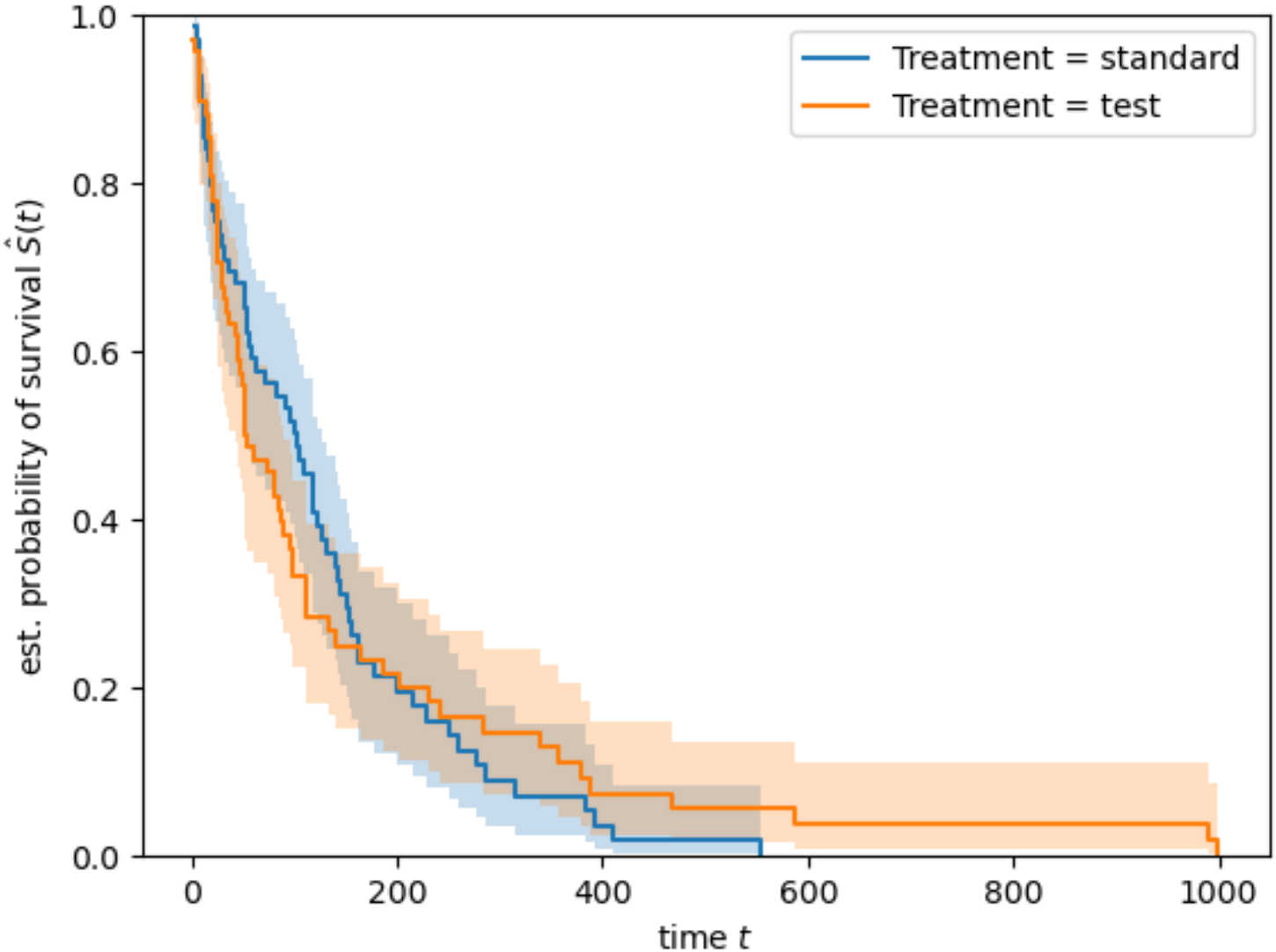
Summary Table

Time	Survival Probability
0.0	0.9854
30.0	0.7004
60.0	0.5309
90.0	0.464
120.0	0.3377
150.0	0.2718
180.0	0.2224
210.0	0.1882
240.0	0.1621
270.0	0.1351
300.0	0.1081
330.0	0.099
360.0	0.081
390.0	0.054
420.0	0.036
450.0	0.036
480.0	0.027
510.0	0.027
540.0	0.027
570.0	0.018

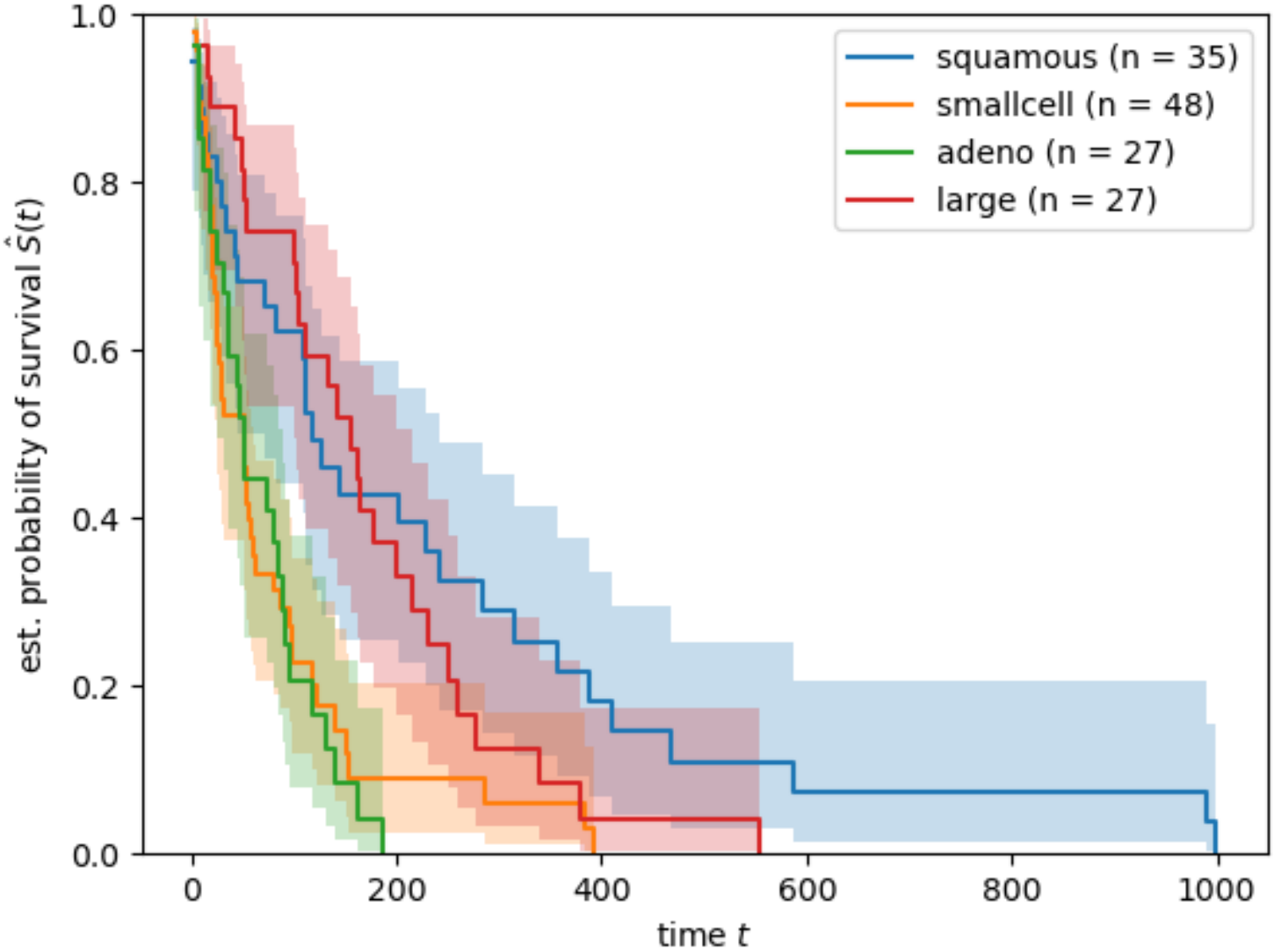
Main observations

- 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.

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`