Applied Statistical Modeling 2

Lorenzoni Valentina
valentina.lorenzoni@santannapisa.it

Institute of Management - Scuola Superiore Sant' Anna, Pisa

7th May 2025

Example dataset

For both the practical application and the assignment, we will use datasets freely available online

Addicts dataset

The "addicts" dataset is also considered in Hosmer DW, Lemeshow S, May S ' book and consists of data related to a sample of drug users from a study by Caplehorn et al. "Methadone Dosage and Retention of Patients in Maintenance Treatment," Med J. Aust., 1991

Addicts dataset

The dataset contains data on 238 subjects identified by an anonymous code, and the Stata format can be downloaded from:

`http://web1.sph.emory.edu/dkleinb/allDatasets/surv2datasets/addicts.dta`

Addicts dataset

Data comprise time (in days) spent by heroin addicts from entry to departure from two different methadone clinics, there are also some covariates, namely, prison record and methadone dose, believed to affect the survival time

We will use these data to evaluate **retention in methadone treatment in a cohort of heroin addicts entering maintenance programmes to evaluate both understand "retention time" and to evaluate potential variables impacting on retention**

Addicts dataset

List of variables

- Subject ID
- Clinic (1 or 2)
- Survival status (0 = censored, 1 = departed from clinic)
- Survival time in days
- Prison record (0 = none, 1 = any)
- Methadone dose (mg/day)

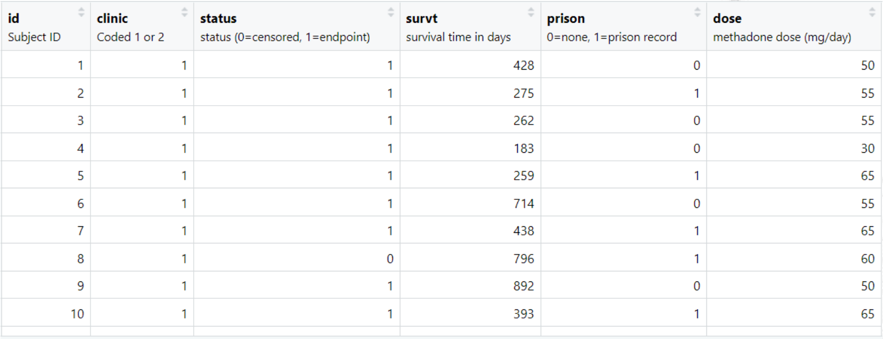
Analysis of survival data

Based on the concepts and approaches introduced in the first lecture, we will have a simple analysis of the *addicts* dataset using R to

- both go into specific examples of how to prepare and describe survival data, interpret results, perform formal tests to compare survival among groups, and
- learning how to do that in practice with R software

Analysis of Addicts in R

A simple preview of some records of the *addicts* dataset



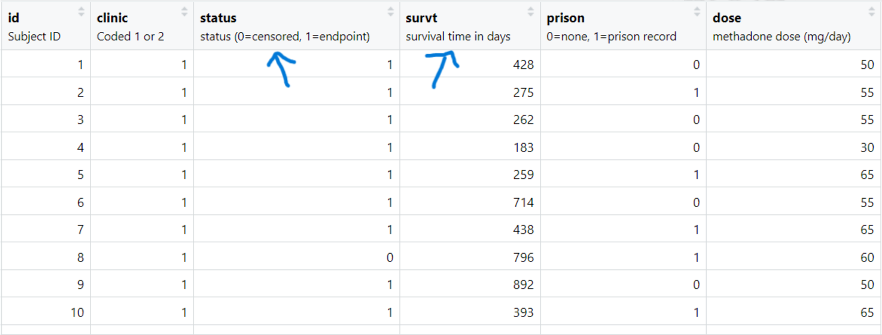
id	clinic	status	survt	prison	dose
Subject ID	Coded 1 or 2	status (0=censored, 1=endpoint)	survival time in days	0=none, 1=prison record	methadone dose (mg/day)
1	1	1	428	0	50
2	1	1	275	1	55
3	1	1	262	0	55
4	1	1	183	0	30
5	1	1	259	1	65
6	1	1	714	0	55
7	1	1	438	1	65
8	1	0	796	1	60
9	1	1	892	0	50
10	1	1	393	1	65

Analysis of Addicts data in R

There are several packages in R that could be used to perform survival analysis and graph survival curves, and we will see some of them, for today's class, we will use survival and ggsurfit

Analysis of Addicts data in R

As we discussed in the first lecture the most important thing in the analysis of survival data is related to the identification of the variables of interest



id	clinic	status	survt	prison	dose
Subject ID	Coded 1 or 2	status (0=censored, 1=endpoint)	survival time in days	0=none, 1=prison record	methadone dose (mg/day)
1	1	1	428	0	50
2	1	1	275	1	55
3	1	1	262	0	55
4	1	1	183	0	30
5	1	1	259	1	65
6	1	1	714	0	55
7	1	1	438	1	65
8	1	0	796	1	60
9	1	1	892	0	50
10	1	1	393	1	65

Analysis of Addicts data in R

!! In this case, the variables of interest, which are *status* and *survt*, are "ready to be used", as for example, because time is already recorded in the dataset as the difference between the date of the event/date of the last follow-up without the need to have some previous calculus, in some cases we need to "prepare data" to have the time variable and the variable related to the event in the correct format to be used from software

Analysis of Addicts data in R

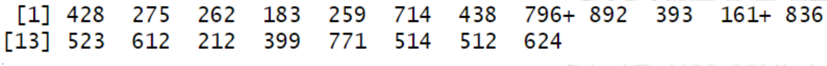
Once variables of interest have been identified and prepared, we need to create an R object containing the variable of interest in order to let R properly "understand" and use survival data, this could be done using the function Surv the by typing:

```
Surv(survt,status)
```

Analysis of Addicts data in R

We can have a look on the object created, and for the ease of reading, we can ask R to see just the first twenty entries of the object, typing:

```
Surv(survt,status) [1:20]
```



```
 [1] 428  275  262  183  259  714  438  796+ 892  393  161+ 836
[13] 523  612  212  399  771  514  512  624
```

R just created an object of length equal to the number of units in the file, and made up of the survival time for each subject, which is complemented by a *plus* for those subjects experiencing the event of interest, that in our case is treatment failure

Analysis of Addicts data in R

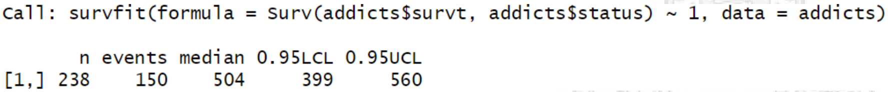
Once we set up the data, it is possible to estimate the survival function using the Kaplan-Meier estimator with the function survfit, which requires a survival object and eventual additional variables (to be added after the tilde) to be considered to estimate the survival function within different groups.

As we just want to estimate the survival function overall, we need to add 1 after the tilde:

```
survfit(Surv(survt, status)  1, data=addicts)
```

Analysis of Addicts data in R

The output of the survfit function provides the estimated median survival time according to the Kaplan-Meier approach and its 95% confidence interval



```
Call: survfit(formula = Surv(addicts$survt, addicts$status) ~ 1, data = addicts)

      n events median 0.95LCL 0.95UCL
[1,] 238    150    504     399     560
```

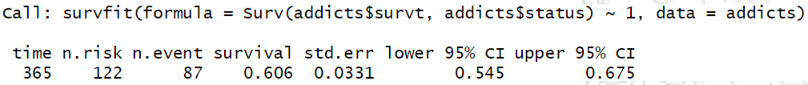
The estimate of median survival time corresponds to the survival time when the survival function is equal to 50%, in our case we estimated a median retention on tretment of about 504 days (less than 1 year and a half) with a 95% confidence interval varying between 399 an560 days

Analysis of Addicts data in R

By asking a summary of the object created with survfit we can also obtain estimates of the survival function at a given time point, for example typing:

summary(survfit(Surv(survt, status) $\sim$ 1, data=addicts),times=365.25)

we obtain an estimate of the survival function and its 95% confidence interval at one year that is about 61% with a 95% confidence interval comprised between about 55% and 68%, at one year a total of 122 are still on treatment while till that time a total of 87 treatment withdrawn were recorded



```
Call: survfit(formula = Surv(addicts$survt, addicts$status) ~ 1, data = addicts)

 time n.risk n.event survival std.err lower 95% CI upper 95% CI
  365    122      87    0.606  0.0331        0.545        0.675
```

Analysis of Addicts data in R

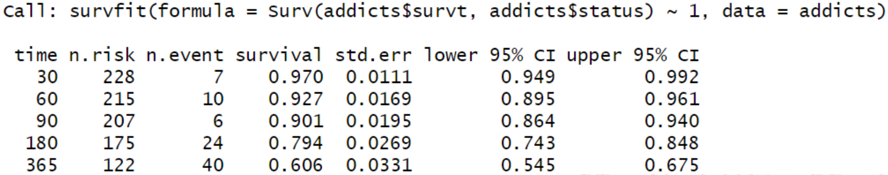
While typing :

```
summary(survfit(Surv(survt, status) ~ 1,
data=addicts),times=c(30,60,90,180,365.25))
```

allows obtaining an estimate of the survival function and its 95% confidence interval at different time points that could be specified depending on the specific interest

Analysis of Addicts data in R

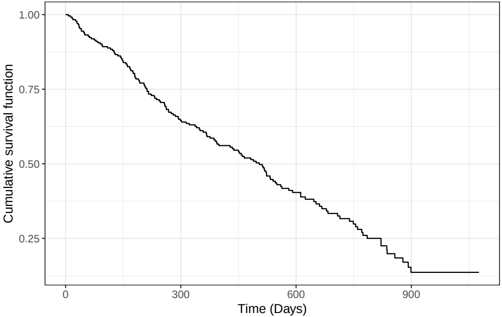
From the example we undrestand that retention on treatment is 97% (95% CI: 95%-99%) at 1 month, it decreases to abot 93% (95% CI: 90%-96%) at 2 months, 90% (95% CI: 86%-94%) at 3 months and drop to about 80% (95% CI: 74%-85%) at 6 months



```
Call: survfit(formula = Surv(addicts$survt, addicts$status) ~ 1, data = addicts)

 time n.risk n.event survival std.err lower 95% CI upper 95% CI
   30    228       7    0.970  0.0111        0.949        0.992
   60    215      10    0.927  0.0169        0.895        0.961
   90    207       6    0.901  0.0195        0.864        0.940
  180    175      24    0.794  0.0269        0.743        0.848
  365    122      40    0.606  0.0331        0.545        0.675
```

Analysis of Addicts data in R

To have a nice plot of the survival curves estimated using the Kaplan-Meier function, we can use the survfit2



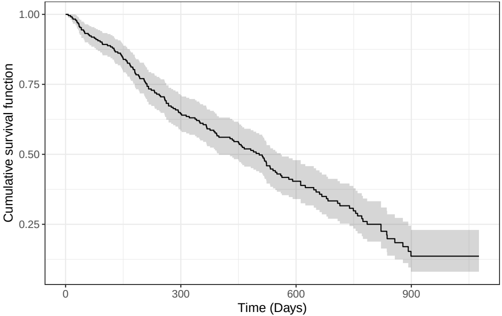
Analysis of Addicts data in R

The survival function plotted has a typical monotone pattern and allows identify values of the survival function along the lenght of the follow-up time considered
m

A visual analysis of the survival curves also help understanding the speed of decrease of the relapse timen and allows identify values of the survival function along the lenght of the follow-up time considered

Analysis of Addicts data in R

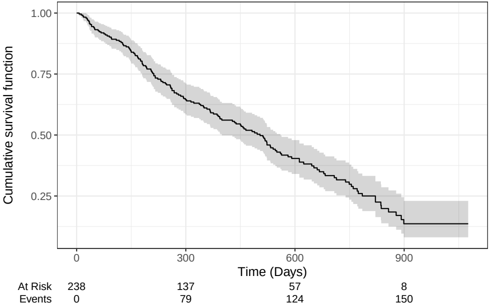
It is possible to add 95% confidence interval for the curve just by adding add_confidence_interval() to the command used before



The 95% confidence interval becomes wider as the follow-up time goes on

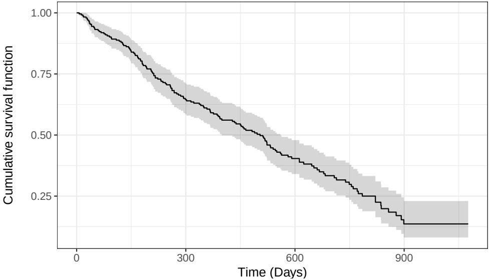
Analysis of Addicts data in R

Adding also add_risktable() to the command used before, we can show the risk table below the plot that allows clearly identifying the number of subjects at risk and the number of events that occurred over different time points



	0	300	600	900
At Risk	238	137	57	8
Events	0	79	124	150

Analysis of Addicts data in R

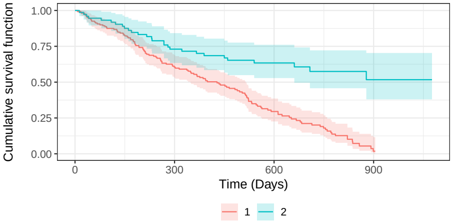


	0	300	600	900
At Risk	238	137	57	8
Events	0	79	124	150

In line with what was discussed in the first lecture, the number of subjects at risk decreased over time, according to both subjects "lost" because having experienced the events and of subjects censored, and this is also the reason of the widening on the confidence interval over time that reflect the increasing uncertainty around the estimate of the survival function

Analysis of Addicts data in R

The survfit2 function could also by applied to a Surv object, requiring the survival function to be estimated by different groups identified by a specific variable to compare $S(t)$ over groups



1				
At Risk	163	87	30	1
Events	0	60	99	122
2				
At Risk	75	50	27	7
Events	0	19	25	28

Analysis of Addicts data in R

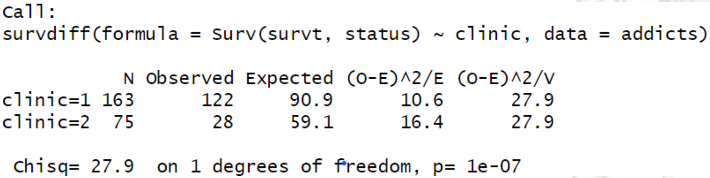
The survival curves plotted seem to suggest a better survival for subjects treated in Clinic 2 as the curve lies above that of Clinic 1 all over the follow-up time

However, we need to do a formal hypothesis test to claim a real difference between the curves

Analysis of Addicts data in R

It is possible to test for the difference between (curves of) survival function using the survfidiff function, which by default performs the Log-Rank test

survdiff(Surv(survt, status) ∼ clinic, data = addicts)



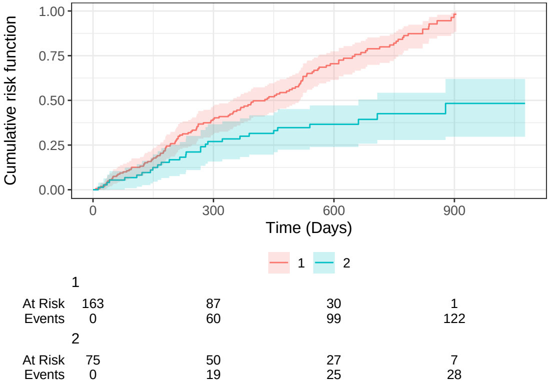
```
Call:
survdiff(formula = Surv(survt, status) ~ clinic, data = addicts)

           N Observed Expected (O-E)^2/E (O-E)^2/V
clinic=1 163      122     90.9      10.6      27.9
clinic=2  75       28     59.1      16.4      27.9

 Chisq= 27.9  on 1 degrees of freedom, p= 1e-07
```

Analysis of Addicts data in R

The test compare observed and expected events over the two grous along the follow-up time assigning equal weights to all the possible timepoints as we described in the first lecture, the P-value obtained that is quite equal to zero ($1e-7$) allows concluding that there is a statistically significant difference between curves, and in particular subjects in Clinic 2 exhibit a better survival as compared to those in Clinic 1; that could be interpreted in terms of a greater effectiveness of Clinic 2 in retaining subjects on tretament and that may be attributed to different factors that we are not able to evaluate, such as possible better approaches used in Clinic 2 or even different characteristics of subjects that could impact on tretment retention

Analysis of Addicts data in R

We could also have a plot of the hazard function to read data in terms of risk of treatment failure, asking the ggsurvfit to plot the "risk" instead of the "survival" function



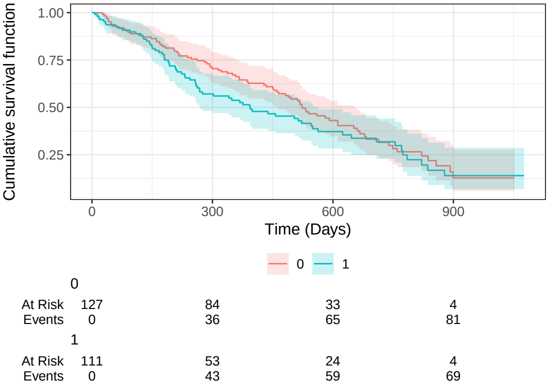
1				
At Risk	163	87	30	1
Events	0	60	99	122
2				
At Risk	75	50	27	7
Events	0	19	25	28

Analysis of Addicts data in R

Considering risk, interpretation of data remains quite the same, unless in this case we can interpret the plot reasoning in terms of hazard, and say that the plot seems to suggest a greater risk of treatment failure in Clinic 1 as compared to Clinic 2

Analysis of Addicts data in R

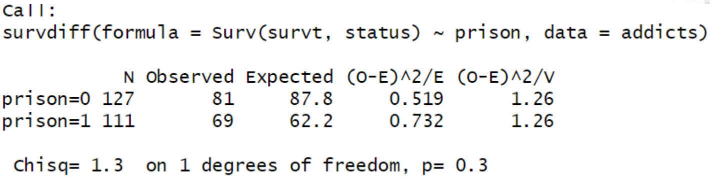
Similarly to what we have done with respect to subjects treated in the two clinics, we could try to use the same approach to assess if being imprisoned affects survival



0				
At Risk	127	84	33	4
Events	0	36	65	81
1				
At Risk	111	53	24	4
Events	0	43	59	69

Analysis of Addicts data in R

Using the Log-Rank test
survdiff(Surv(survt, status) $\sim$ prison, data = addicts)



```
Call:
survdiff(formula = Surv(survt, status) ~ prison, data = addicts)

           N Observed Expected (O-E)^2/E (O-E)^2/V
prison=0 127       81     87.8     0.519      1.26
prison=1 111       69     62.2     0.732      1.26

 Chisq= 1.3  on 1 degrees of freedom, p= 0.3
```

In this case, both looking at Kaplan-Meier curves, which are quite overlapping, and at the result of the Log-Rank test with P-value=0.3, suggest no difference in the survival function according to history of impsisonment