Survival data analysis

Introduction to survival analysis

Survival analysis focuses on the analysis of the time to some event of interest, which is a common question of interest in biomedical research. This event represents a change on the status of the subject, that is, the subject acquires a trait that it did not have before. However, the focus of survival analysis is not studying if subjects had experienced the event or not, but the time from some fixed starting point to the occurrence of the event of interest. Some examples of survival times are time to death or the time a battery takes to run out.

Thus, in survival analysis the dependent variable is a continuous variable, T, which identifies the time to the occurence of the event of interest. This T is defined by two instants: the initial instant and the final instant. The initial instant is defined by the moment in which the individuals start being at risk of changing their status, whereas the final instant is defined by the moment in which this change happens. For example, if we think about the general mortality, the initial instant is birth and the final instant is death. Nevertheless, if we think about lethality (deaths caused by a disease), the initial instant is the moment in which the subject contracts the disease. Hence, it is not always easy to define initial or final instants.

Ideally, the initial instant should be the same for all subjects. This is possible in some cases, like in the example about the lifetime of a battery. We can fully charge a sample of batteries and turn them on all at once. Unfortunately, this is quite unusual in biomedical research (ordinarily, different individuals start being at risk at different moments). When this is not possible, we can at least choose an initial instant that is equivalent among all subjects. Some examples are the date of birth, the date of diagnosis of a disease or the date of liver transplantation in a group of transplanted patients. In these situations, the initial instant is equivalent although the cronological moments are different, as illustrated in Figure 1.

The initial instant and the event of interest define the variable that we should analyse. In the example about the liver transplantation, if the event of interest is death, our outcome variable is survival time from transplantation. Here we only study those phenomena in which the event of interest is unique and implies that the subject leaves the study. Hence, we exclude the situations in which the event can occur more than once or there are two or more events of interest.

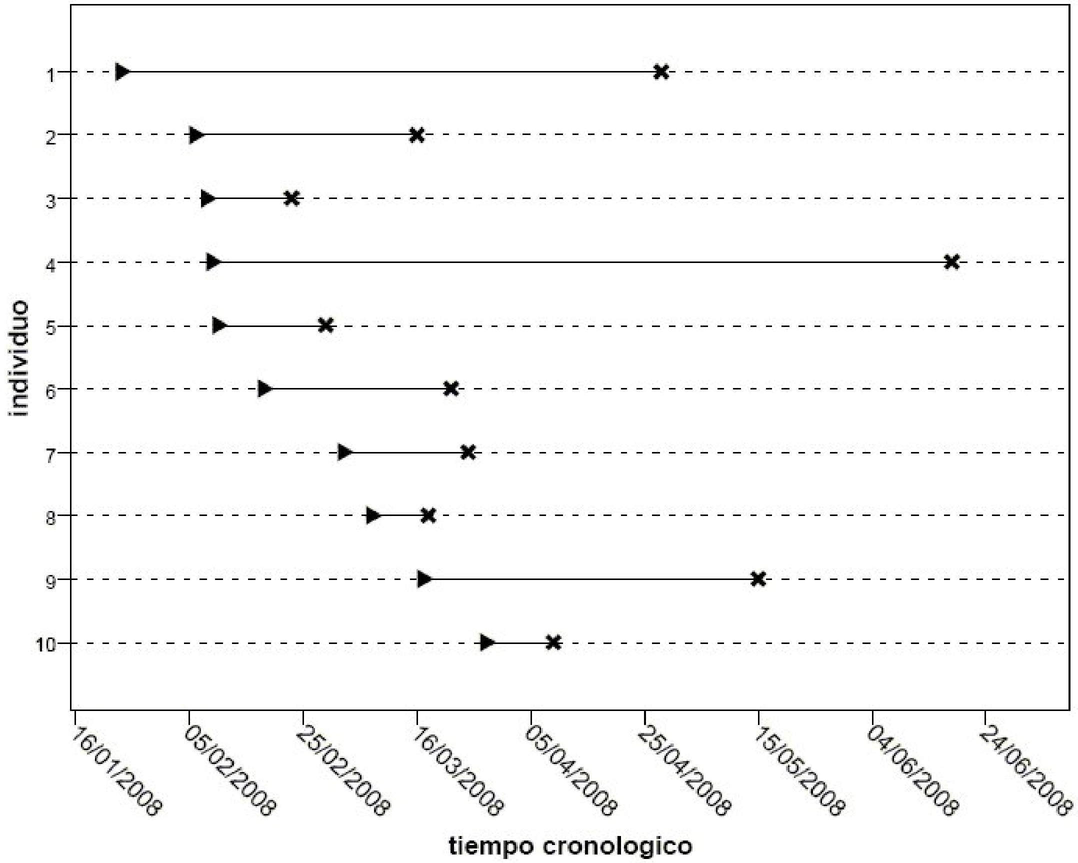


Figure 1: Survival times of 10 individuals

The random variable T

We call T the outcome of interest, the time to event. This random variable presents some particular features, which implies that we must use specific statistical methods to analyse it. We define t as the realization of the random variable T. Some properties of T are the following:

- T is non-negative, that is, $t_i \geq 0$ for any individual i.
- T usually follows an asymmetric probability distribution model. Frequently, some individuals present high values of T, which leads to a left-skewed probability distribution.
- Often, the event of interest is not observed in all subjects of the sample within the follow-up time of the study, as illustrated in Figure 2.

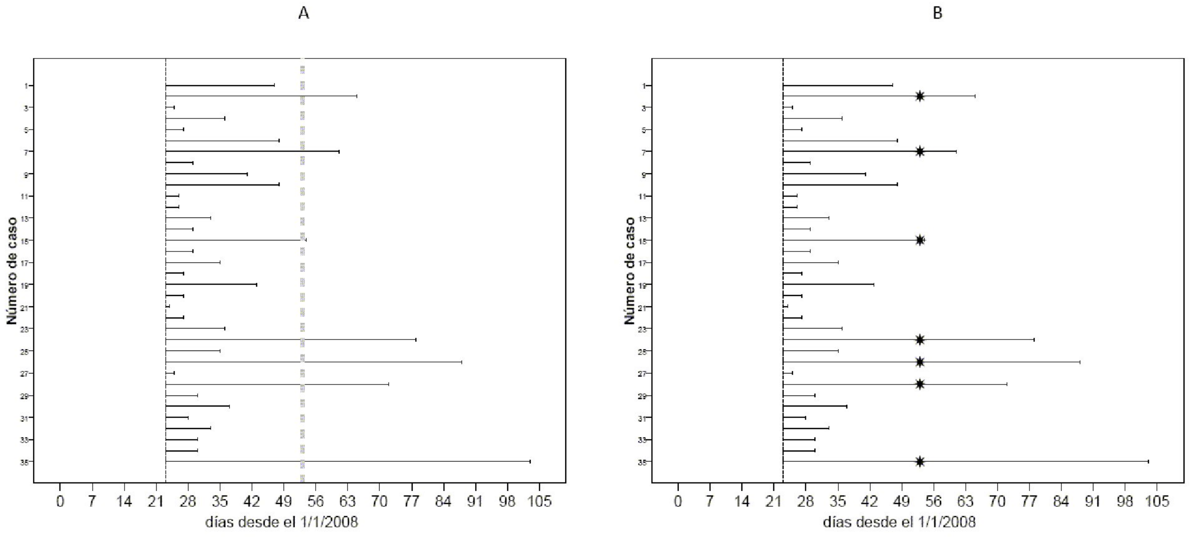


Figure 2: Design with the same initial instant and fixed follow-up period. The dashed line indicates the time-point when the follow-up ends. Stars indicate the censored times

Censoring

Censoring is a phenomenon related to incomplete information in which the value of an observation is only partially known. The determination of the survival time is incomplete. We distinguish between three types of censoring:

- Right censoring: the initial time is known and the subject is followed-up, but this follow-up ends before the occurrence of the event of interest. Thus, we assume that the event will happen after the last follow-up (to the right of the observed time), so the survival time is greater than the follow-up time.
- Left censoring: this happens when we know that the event of interest occurred prior to a certain follow-up time, but the exact time of occurrence is unknown (the event occurred before, or to the left, of the observed time).
- Interval censoring: the initial time is known but the follow-up of the subject has been intermittent and the event of interest has occurred between two follow-ups. Hence, interval censoring occurs when the event of interest is known to have occurred between two time-points.

Here we only focus on right censoring since this is the most common situation we find in biomedical research.

Non-informative censoring occurs when the cause of censoring is unrelated to the event of interest. This condition must be satisfied in all the analysis techniques described in this document.

Studies with right censoring can be classified in different types according to how we define the follow-up period. If the initial instant is the same for all individuals and the study design considers some fixed follow-up time C, the survival time for the censored observations must be equal to this constant (see Figure 2). Other studies are designed such that the follow-up time is enough for observing a pre-specified proportion of events. Figure 3 is an example of such study designs.

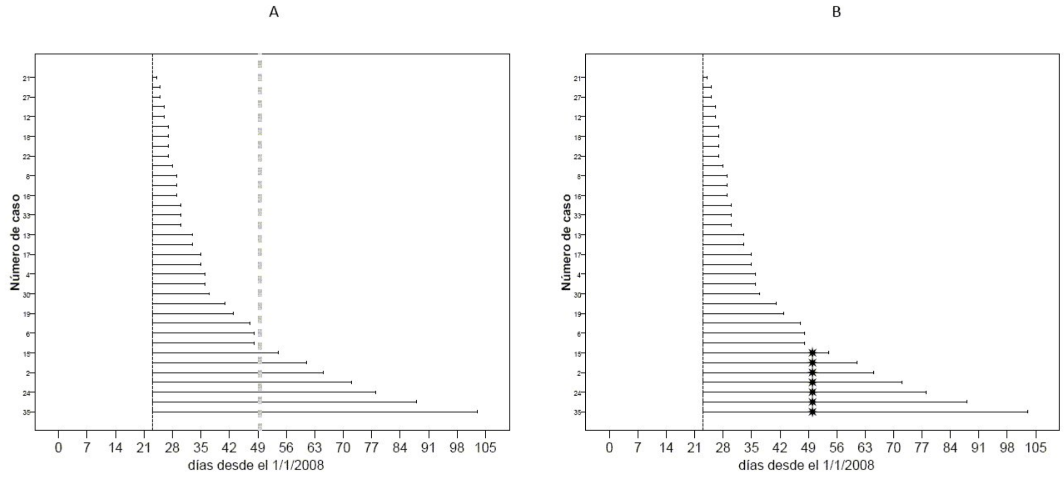


Figure 3: Study design that ends the follow-up when the 80% of the sample has experienced the event. The dashed line indicates the time-point when the follow-up ends. Stars indicate the censored times

The most common design in biomedical research consists of a recruitment period, in which the individuals presenting the initial event (for example, the diagnosis of the disease of interest) are included in the study, then the sample is followed-up during some pre-specified time period, such that censoring times are random (see Figure 4).

In presence of right censoring, the information collected for each subject consists of two variables, (y_i, δ_i). y_i represents the observed time for each individual i, whereas δ_i is the event indicator. Thus, $\delta_i = 1$ means that subject i has experienced the event, so in this case the survival time equals the observed time ($y_i = t_i$). $\delta_i = 0$ means that subject i has not experienced the event (that is, it has been censored) so the observed time is lower than the survival time ($y_i < t_i$).

Apart from the censored observations resulting from the study design, there are more situations in which the survival time cannot be determined and also led to censored observations.

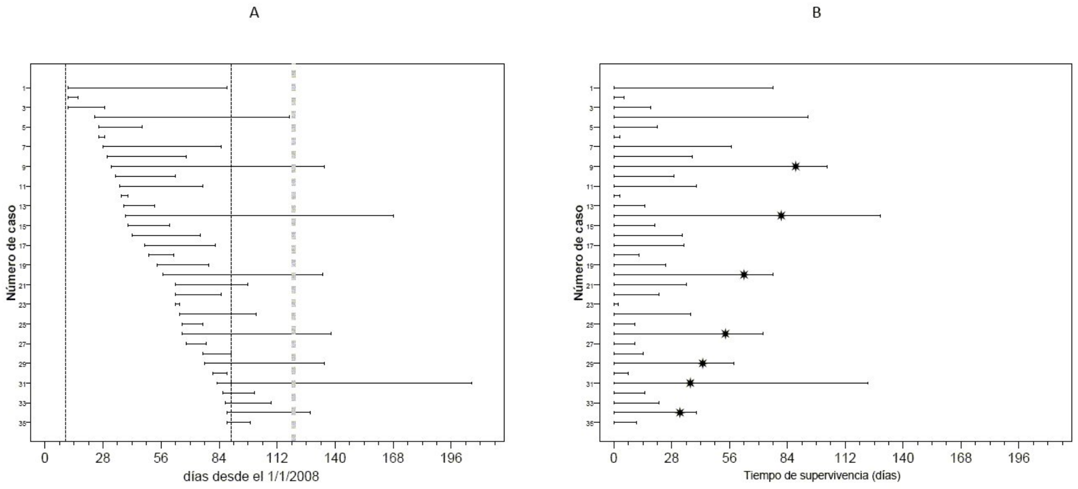


Figure 4: Study design with a 80-day recruitment period and 30 days of follow-up. The dashed line indicates the time-point when the follow-up ends. Stars indicate the censored times

These observations are related to loss to follow up. These subjects can be lost to follow up for a wide range of reasons: change of address, accidental death, death unrelated to the disease under study, refusal to continue in the study, etc. Here we assume that any censoring caused by a loss to follow-up is non-informative.

Functions related to survival data analysis

As we have seen before, our variable of interest is T, the time from some fixed starting point to the occurrence of a given event. There are some functions related to T that help us understand and interpret it.

- Probability distribution function, $F(t)$:

  $$F(t) = P(T \leq t),$$

  which is defined for any $t \geq 0$. $F(t)$ is the probability that a subject experiences the event before t. When the event of interest is death, it is called **cumulative mortality function**.

- Density function, $f(t)$:

$$f\left(t\right) = \frac{\delta F\left(t\right)}{\delta t}.$$

- Survival function, $S(t)$:

  $$S\left(t\right) = P\left(T > t\right) = 1 - P\left(T \leq t\right) = 1 - F\left(t\right).$$

  $S(t)$ is defined as the probability that an individual survives longer than t (where surviving means not having experienced the event). Since it is a probability, its domain is $[0, 1]$ (the same domain than that of the probability distribution function). $S(t)$ verifies $S(0) = 1$ (the probability of surviving to $t = 0$ is 1) and $S\left(\infty\right) = 0$. In Figure 5 we can see that $S(t)$ is a non-increasing function, that is, given two time-points t_1 and t_2 such that $t_1 < t_2$, we have $S\left(t_2\right) \leq S\left(t_1\right)$.

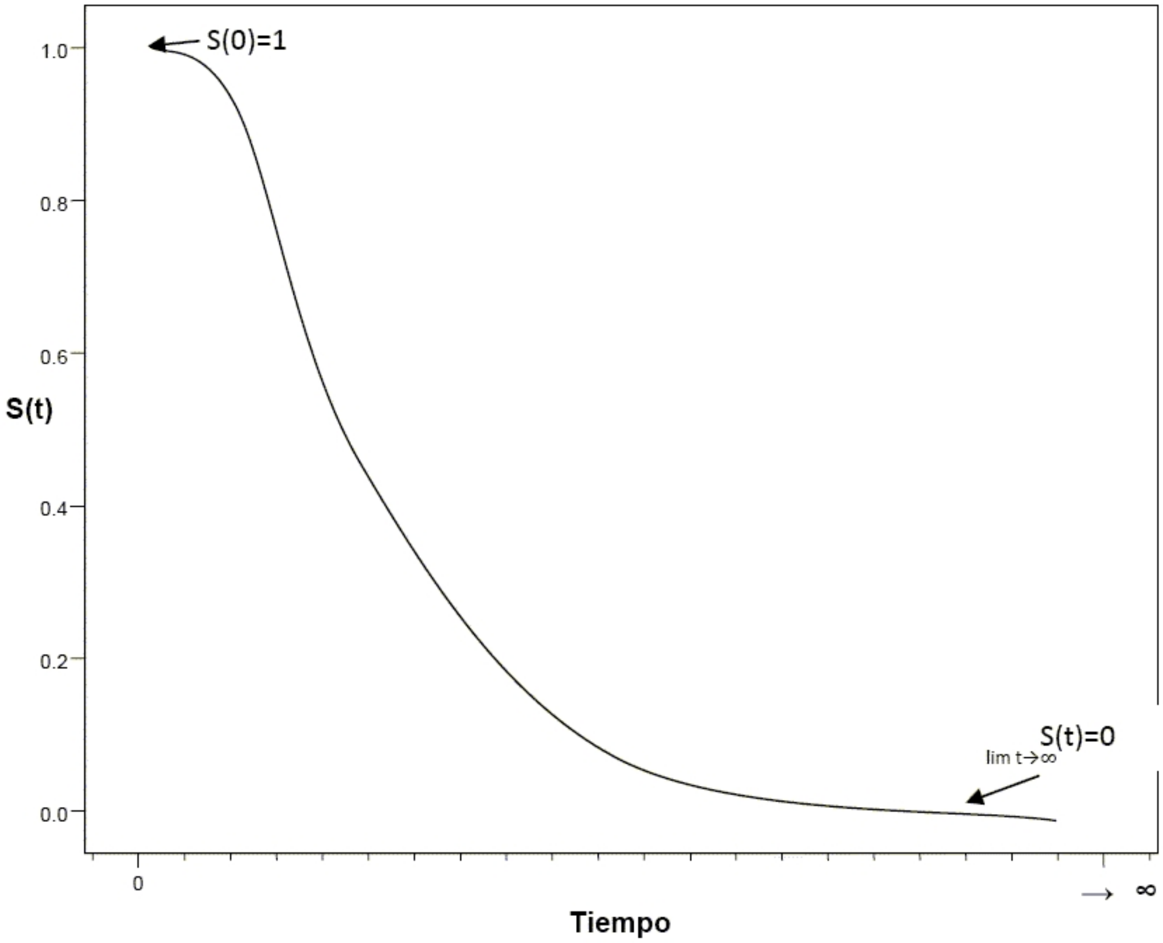


Figure 5: Example of survival function

For example, let us assume that we have data about the times to death from a sample of some population of interest. The initial time is birth and the event of interest is death. As these are annual data, graphs will be stepped. In Figure 6 we can see the density function $f(t)$ of these data, which shows the change in the cumulative probability of dying for different ages. Notice that the probability of dying during the first year of life is quite high, then it becomes almost zero with a small increase during adolescence. Then we observe a minor increase in this probability, which grows higher from age 50. The probability of dying shows a maximum at age 85 and falls thereafter. If we accumulate this function we obtain $F(t)$ (Figure 7) and its complementary is the survival function $S(t)$ (Figure 8).

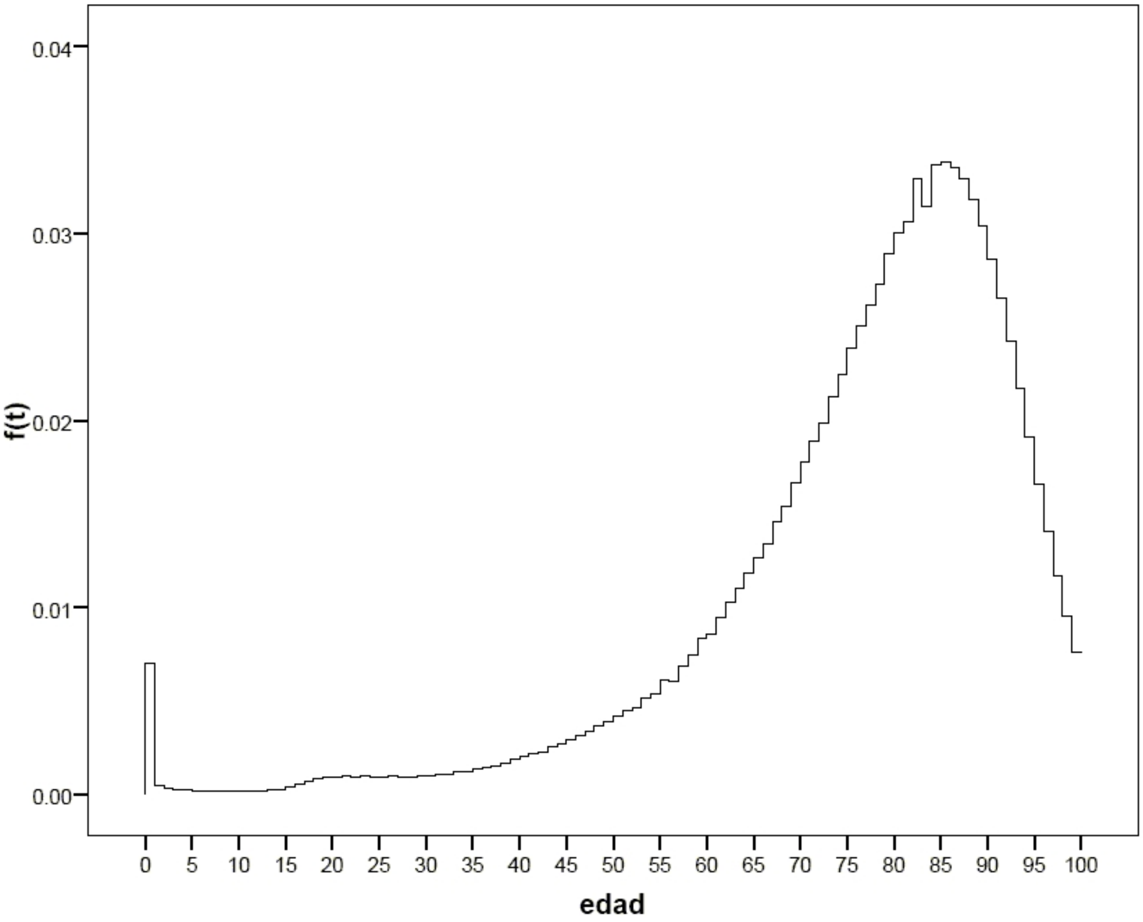


Figure 6: Example of density function

- Hazard function, $h(t)$: It is defined as the hazard of presenting the event of interest at time t, given that the event has not occurred before t. It quantifies the instantaneous hazard of event taking into account the number of deaths in a certain time interval as well as the number of subjects at risk of experiencing the event at the beginning of that interval (notice that subjects that had already experienced the event cannot

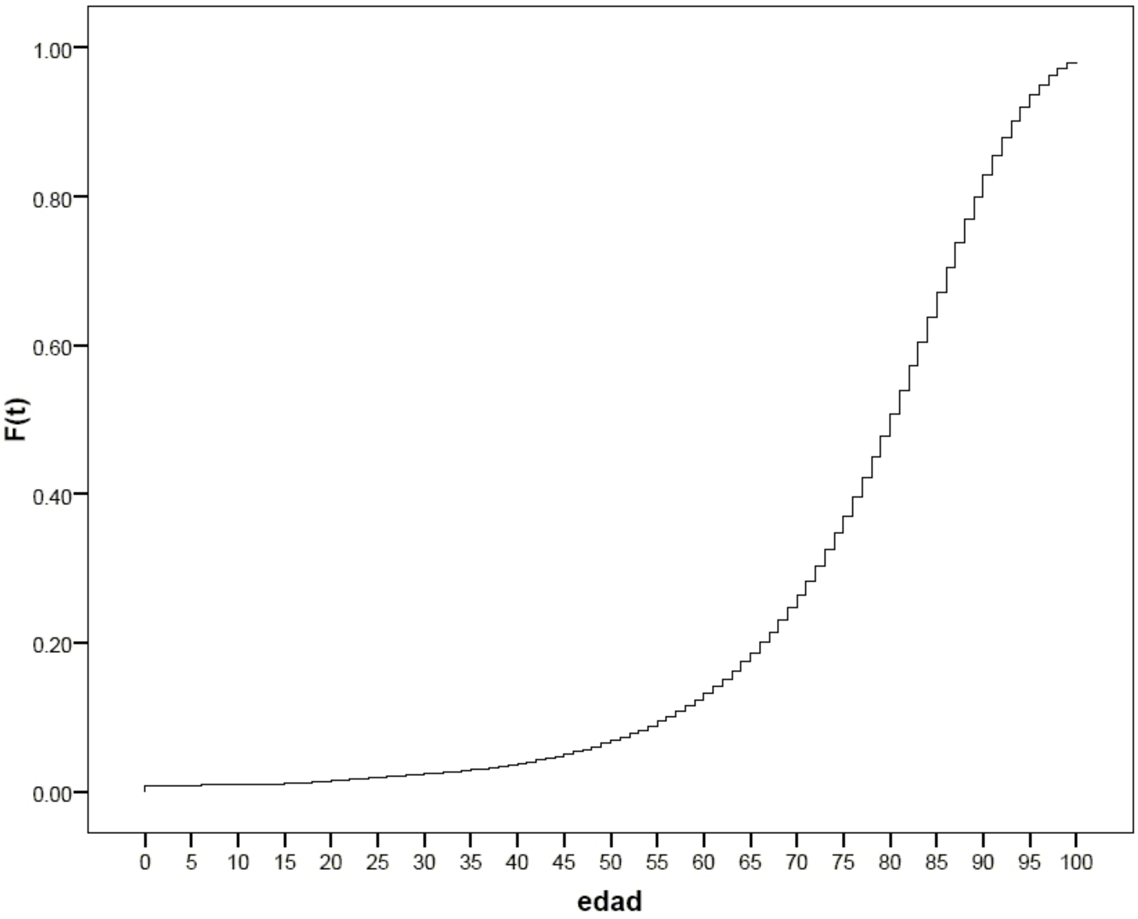


Figure 7: Example of distribution function

experience it again). For example, if survival decreases a 5% in one year, we have $S(t) - S(t+1) = 0.05$. If at the beginning of this interval survival is high, the instantaneous risk will be low, whereas this risk will be high if only a small fraction of the individuals has not experienced the event at the beginning of the interval. The hazard function is defined as follows:

$$h(t) = \lim_{\Delta t \to 0} \frac{P(T \leq t + \Delta t | T \geq t)}{\Delta t} = \frac{f(t)}{S(t)}.$$

Notice that $h(t)$ can be expressed as the ratio between the density function and the survival function. The hazard function is not a probability, but an infinitessimal indicating the intensity of mortality, so it can take values between 0 and ∞. In Figure 9 we show $h(t)$ for the data in the previous example.

If $h(t)$ is constant, for any t, a fixed percentage of individuals that had not dead at time t will experience the event in the following instant. This is an usual situation in the industrial

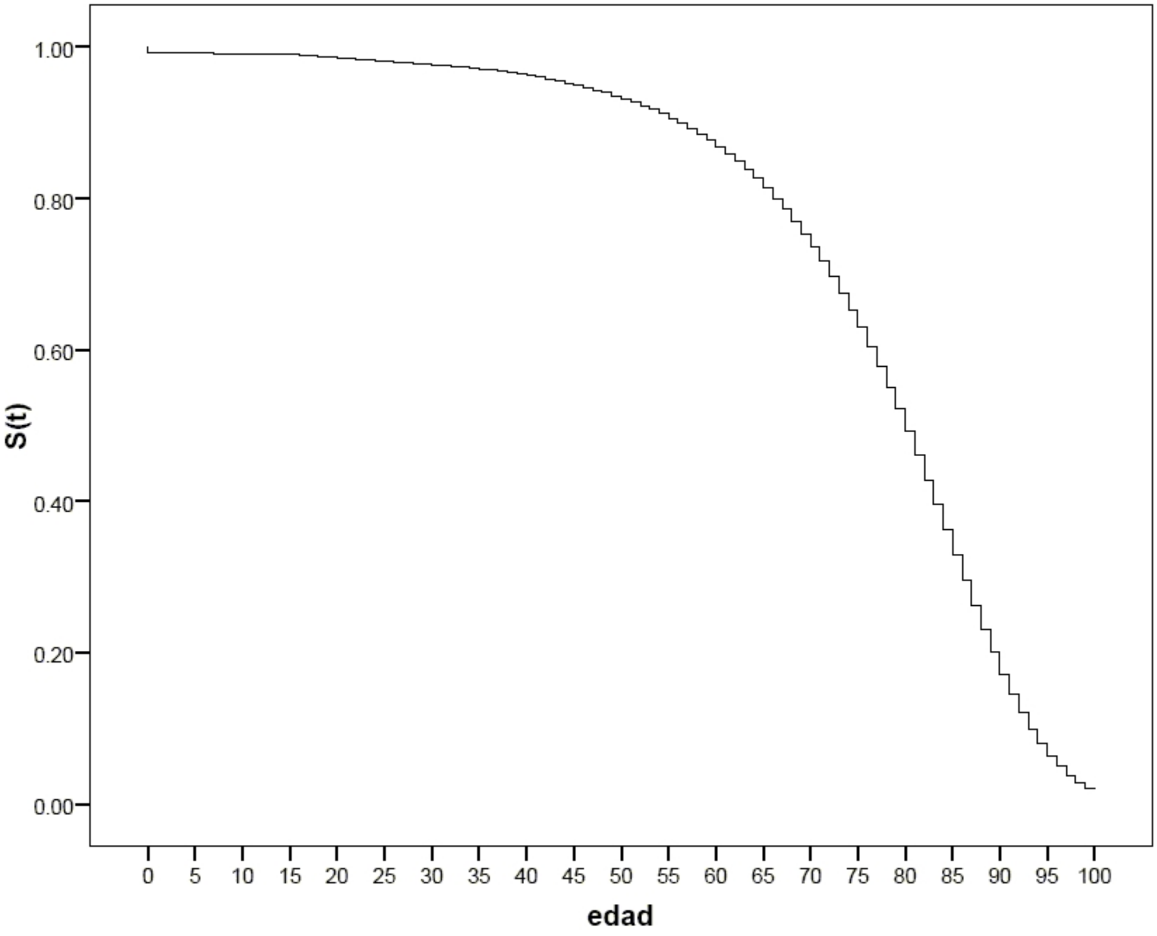


Figure 8: Example of survival function

context but not very common in biomedical research. For example, in patients diagnosed with leukemia the starting value of $h(t)$ tends to lightly increase just after the diagnosis, and then it increases rapidly since the greater the survival time, the worst the prognosis and the higher the probability of dying in the following instant. Other survival problems present the opposite situation, where $h(t)$ is high at first, but as individuals survive in time, their hazard of death tends to decrease. An example of these situations is the survival of post-surgical patients, in which the risk of post-surgical complications decreases over time.

- Cumulative hazard function, $H(t)$: It is obtained from the integration of $h(t)$ (Figure 10).

This function is related to the survival function because it can be shown that

$$H\left(t\right) = -\ln\left(S\left(t\right)\right),$$

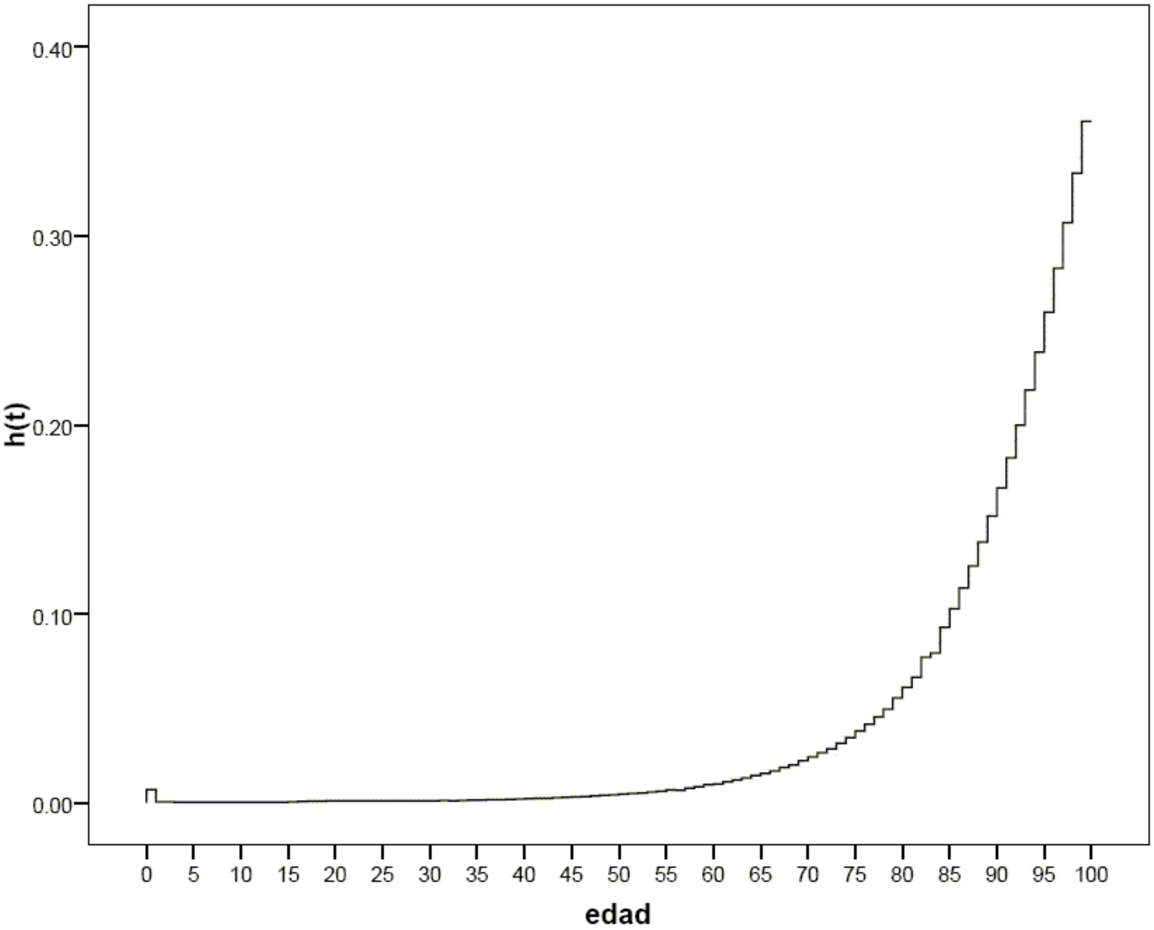


Figure 9: Example of hazard function

and so

$$S\left(t\right)=e^{-H\left(t\right)}.$$

Goals of survival analysis

Since T is a continuous variable, we could use the standard regression techniques to describe or model T and estimate its expected value, when the event of interest is supposed to happen. However, T is usually right-censored, which makes standard methods inappropriate. Moreover, in survival analysis the interest lies in estimating the probability of survival for different values of t or the instantaneous hazard, since they allow us to model the mechanism that leads to the occurrence of the event of interest. For all these reasons, the main goal of survival analysis is to describe and interpret the survival function $S(t)$ and/or its corresponding hazard function $h(t)$. A secondary objective is to compare survival functions

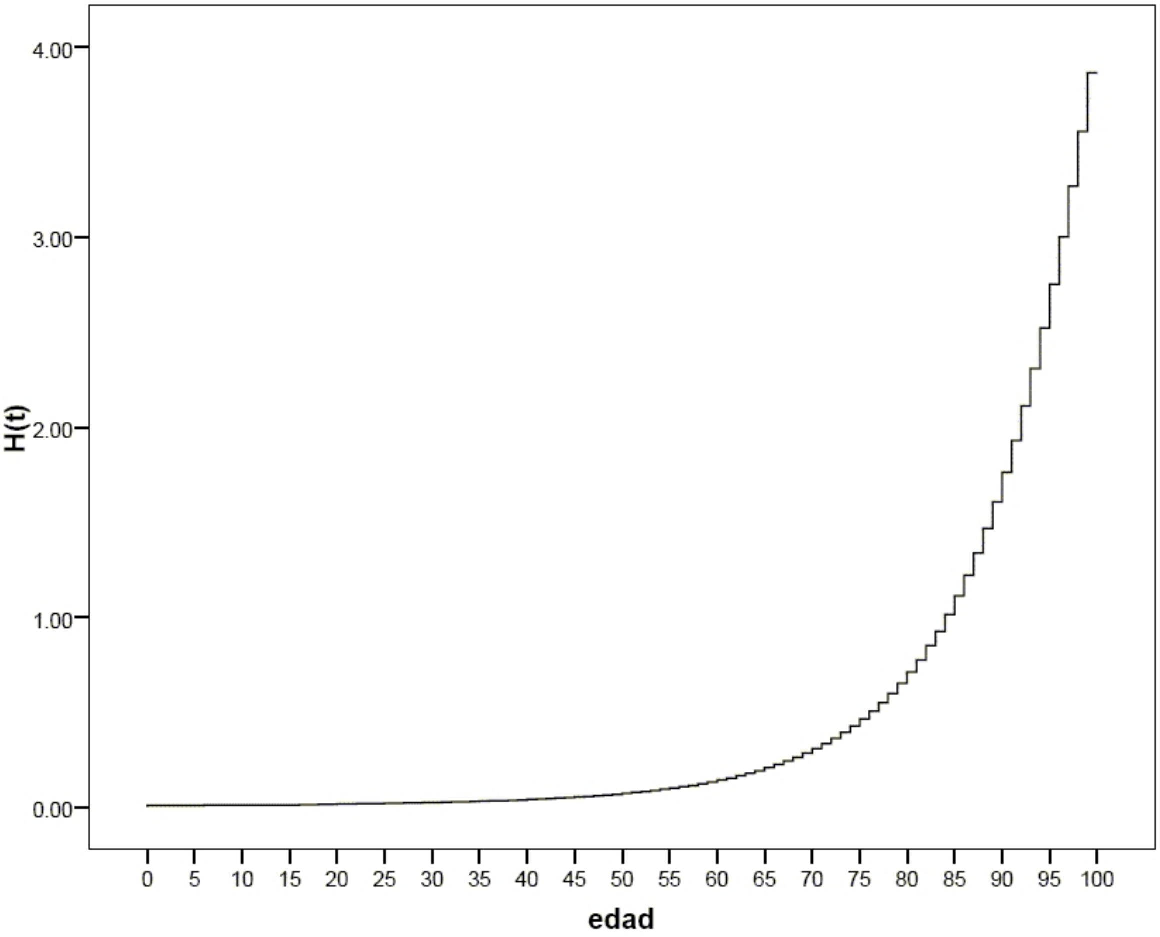


Figure 10: Example of cumulative hazard function

among two or more different populations or groups.

Estimation of the survival function

We consider two ways of estimating $S(t)$ under the assumption that all subjects in the sample follow the same survival function:

- Parametric method: we assume that T follows some probability distribution model, such as exponential or Weibull model. Then we estimate $S(t)$ by simply estimating the parameters of the distribution model. If the model is correct, this kind of estimations will be more precise that the ones obtained with the non-parametric method.
- Non-parametric method: we do not assume any parametric probability distribution model for T. We make an empirical estimation of the survival function, which is known

as the **Kaplan-Meier survival curve**. These estimations are less precise than the ones obtained with parametric models, but the main advantage of this approach is that we do not have to assume any parametric model when it is unknown. It is one of the most widely used methods in survival analysis.

The Kaplan-Meier method: estimation of $S(t)$ with right-censored data

Let's illustrate the Kaplan-Meier method with an example. Suppose that a research group sampled 200 patients from the population of interest to participate in a study to investigate the time to the development of a certain complication after a new treament. This study was carried out in two hospitals, A and B. Each hospital recruited 100 patients and followed them from the administration of the treatment. Hospital A followed the patients for a year, whereas hospital B did so for two years.

After a year from the beginning of the study, 25 patients from hospital A and 20 from hospital B had experienced the complication. Two patients out of the remaining 80 patients from hospital B experienced the complication in the second year of follow-up. These data are described in Table 1:

	Hospital A		Hospital B	
Follow-up year	# individuals at risk	# events	# individuals at risk	# events
1	100	25	100	20
2	75	??	80	2

Table 1: Example: estimation of $S(t)$ with censored data

Since the number of subjects that presented the complication during the second year is unknown, we cannot calculate the probability of surviving (not experiencing the event) after 2 years using the data from both hospitals. Some alternative approaches to compute this probability are the following:

- Exclude the information from hospital A since this is incomplete:

$$S\left(1\right)=\frac{100-20}{100}=0.8,$$

$$S\left(2\right)=\frac{100-22}{100}=0.78.$$

  Notice that with this approach we are excluding relevant (althought incomplete) information from hospital A.

- Use all the complete information available: use data from both hospitals to estimate 1-year survival and use data from hospital B to estimate 2-year survival:

$$S\left(1\right)=\frac{200-\left(20+25\right)}{200}=0.775,$$

$$S\left(2\right)=\frac{100-22}{100}=0.78.$$

It seems to be a good way of using all the available information, but it can lead to inconsistent results, like in this example, in which $S(1) < S(2)$.

- Estimate survival separately for each period (year) and then combine these estimations:

$$S\left(1\right)=\frac{200-\left(20+25\right)}{200}=0.775,$$

$$S\left(2|T>1\right)=\frac{78}{80}=0.975.$$

Notice that the probability of surviving to the second year (in hospital B) has been conditioned to having survived to the first year. Thus, the probability of surviving (not experiencing the complication) through all the 2-year period is:

$$S\left(2\right)=S\left(2|t>1\right)\cdot S\left(1\right)=0.78\cdot 0.975=0.756.$$

With this approach we obtain an estimate that is consistent with $S(1) = 0.78$ and we have used all the information, whether censored or not, we had available. This method allows us to calculate the cumulative probability of surviving in a period (divided into intervals) by multiplying the conditionate probabilities corresponding to each interval (for this reason, this method is called **product limit estimator**).

The application of this method to pre-specified intervals is known as **life table estimation**, whereas if we use the times observed in the sample instead of time intervals, this method is called **Kaplan-Meier estimation**.

Survival analysis with R

Data management

The R package `survival` contains functions to perform survival analysis.

```
library(survival)
```

The example dataset we use, `aml`, is included in the `survival` package. This dataset contains the survival times from 23 patients with acute myelogenous leukemia (AML). The main goal of the study was investigate whether the standard course of chemotherapy should be extended for additional cycles.

```
data(aml)
head(aml)
```

```
  time status          x
1    9      1 Maintained
2   13      1 Maintained
3   13      0 Maintained
4   18      1 Maintained
5   23      1 Maintained
6   28      0 Maintained
```

The dataset contains 3 variables:

- time: time to death of last follow-up, in months (survival or censoring time)
- status: censoring status (1=death–the event of interest has occurred, 0=alive–the event of interest has not occurred, the patient has been censored)
- x: maintenance chemotherapy given? (Nonmaintained=standard course of chemotherapy, Maintained=additional cycles of chemotherapy)

In survival analysis, we usually create a `Surv` object. It indicates if survival times are observed or censored (+):

```
t <- Surv(aml$time, aml$status)
t
```

```
 [1]   9   13   13+  18   23   28+  31   34   45+  48  161+    5    5    8    8
[16]  12   16+  23   27   30   33   43   45
```

```
class(t)
```

```
[1] "Surv"
```

Kaplan-Meier estimator

Here we will see how to estimate and plot survival curves. The Kaplan-Meier estimation for the entire sample is calculated using the following code:

```
km.fit <- survfit(t~1, data=aml)
summary(km.fit)
```

```
Call: survfit(formula = t ~ 1, data = aml)

 time n.risk n.event survival std.err lower 95% CI upper 95% CI
```

```
    5     23       2   0.9130  0.0588        0.8049        1.000
    8     21       2   0.8261  0.0790        0.6848        0.996
    9     19       1   0.7826  0.0860        0.6310        0.971
   12     18       1   0.7391  0.0916        0.5798        0.942
   13     17       1   0.6957  0.0959        0.5309        0.912
   18     14       1   0.6460  0.1011        0.4753        0.878
   23     13       2   0.5466  0.1073        0.3721        0.803
   27     11       1   0.4969  0.1084        0.3240        0.762
   30      9       1   0.4417  0.1095        0.2717        0.718
   31      8       1   0.3865  0.1089        0.2225        0.671
   33      7       1   0.3313  0.1064        0.1765        0.622
   34      6       1   0.2761  0.1020        0.1338        0.569
   43      5       1   0.2208  0.0954        0.0947        0.515
   45      4       1   0.1656  0.0860        0.0598        0.458
   48      2       1   0.0828  0.0727        0.0148        0.462
```

The output generated consists of a table containing the following columns:

- `time`: event times in our sample (notice that this output excludes censored times)
- `n.risk`: number of subjects at risk immediately before `time`
- `n.event`: number of events occurred at `time`
- `survival`: Kaplan-Meier estimate of the survival curve at `time` – probability of surviving until `time`
- `std.err`: Standard error of `survival`
- `lower 95% CI` and `upper 95% CI`: 95% confidence interval for `survival`

If we want to estimate the survival function separately for two or more groups, we just have to change the `1` in the formula for the grouping variable:

```
km.fit.x <- survfit(t~x, data=aml)
summary(km.fit.x)
```

```
Call: survfit(formula = t ~ x, data = aml)

                x=Maintained
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    9     11       1    0.909  0.0867       0.7541        1.000
   13     10       1    0.818  0.1163       0.6192        1.000
   18      8       1    0.716  0.1397       0.4884        1.000
   23      7       1    0.614  0.1526       0.3769        0.999
   31      5       1    0.491  0.1642       0.2549        0.946
   34      4       1    0.368  0.1627       0.1549        0.875
   48      2       1    0.184  0.1535       0.0359        0.944
```

```
                x=Nonmaintained
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    5     12       2   0.8333  0.1076       0.6470        1.000
    8     10       2   0.6667  0.1361       0.4468        0.995
   12      8       1   0.5833  0.1423       0.3616        0.941
   23      6       1   0.4861  0.1481       0.2675        0.883
   27      5       1   0.3889  0.1470       0.1854        0.816
   30      4       1   0.2917  0.1387       0.1148        0.741
   33      3       1   0.1944  0.1219       0.0569        0.664
   43      2       1   0.0972  0.0919       0.0153        0.620
   45      1       1   0.0000     NaN           NA           NA
```

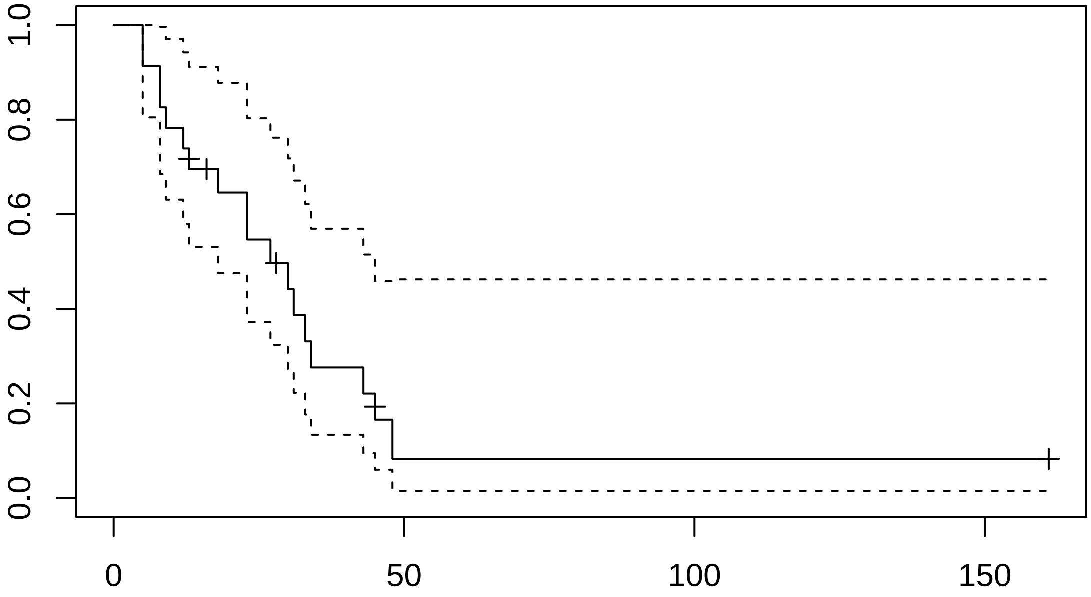
This allows us to describe the survival experience of each group. See Table 2 for conclusions about the 12-month survival.

	Maintained	Nonmaintained
# deaths at $t = 12$	0	1
# deaths until $t = 12$	1	5
$S(12)$	0.909	0.583

Table 2: Some results about the 12-month survival in the AML data

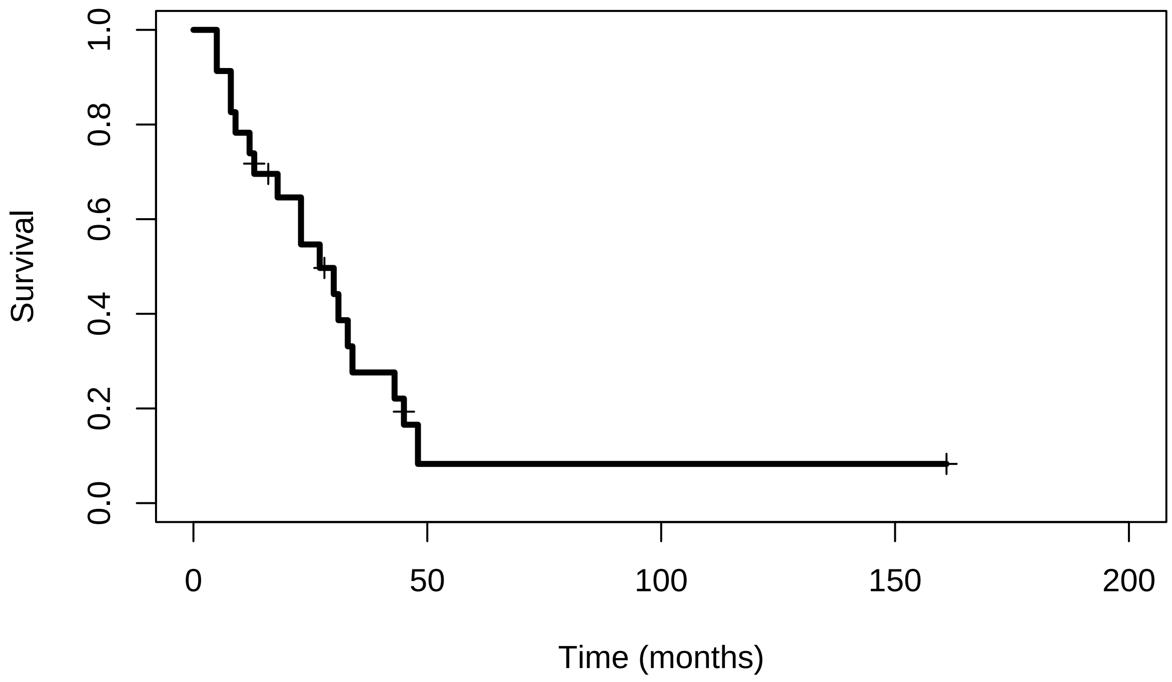
We can also plot the survival curves:

```r
# entire sample: curve + confidence interval
plot(km.fit, mark.time=TRUE)
```

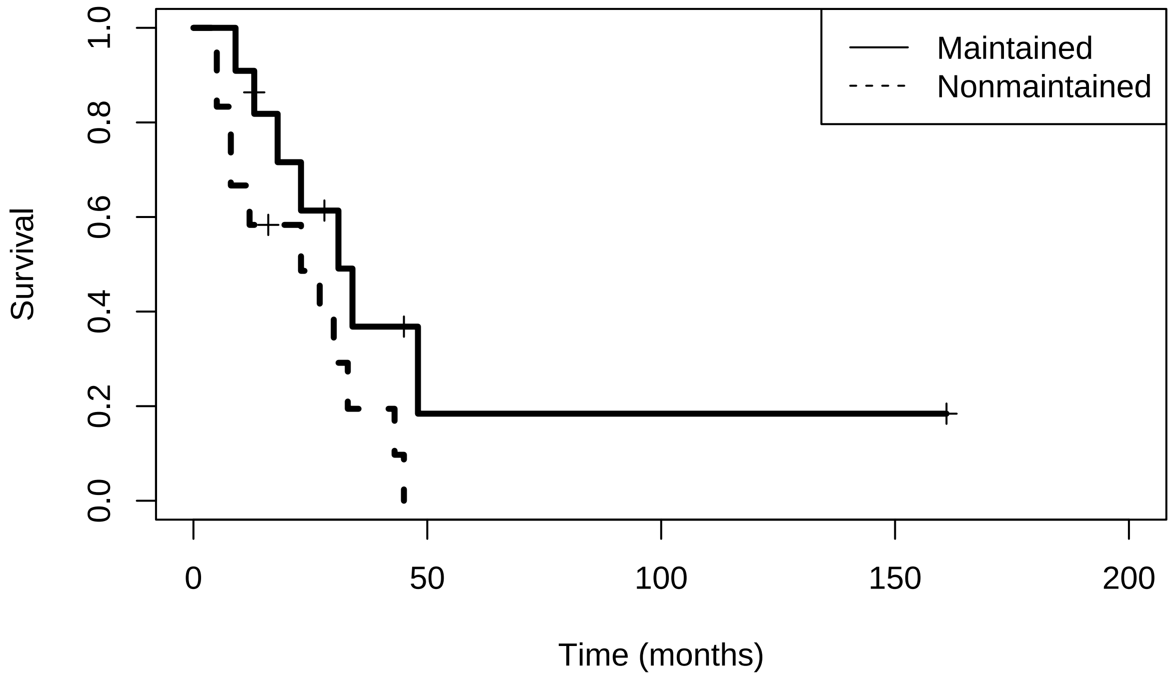


```
# mark.time=TRUE adds marks in censored times
```

```
# remove confidence interval and enhance the plot
plot(km.fit, mark.time=TRUE, conf.int=FALSE, xlab="Time (months)",
     ylab="Survival", xlim=c(0, 200), lwd=3)
```

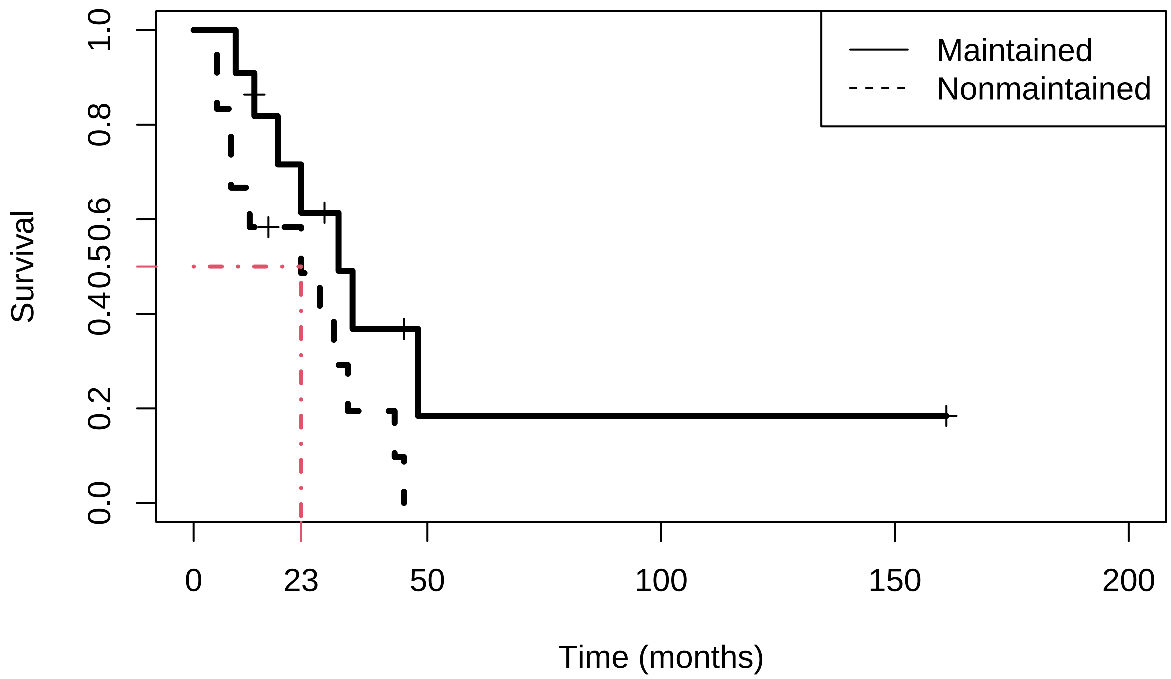


```
# according to x
plot(km.fit.x, mark.time=TRUE, lty=1:2, xlab="Time (months)",
     ylab="Survival", xlim=c(0, 200), lwd=3)
legend("topright", levels(aml$x), lty=1:2)
```



From the survival curves we can also calculate some quantiles (such as the median or the first or third quantile). The median survival time is defined as t such that $S(t) = 0.5$. Let us see how we can calculate the median from the Kaplan-Meier estimate: we draw a straight line in $S(t) = 0.5$ and look for the t in which this straight line crosses the survival curve. Let us do it for the Nonmaintained group:

```
plot(km.fit.x, mark.time=TRUE, lty=1:2, xlab="Time (months)",
     ylab="Survival", xlim=c(0, 200), lwd=3)
legend("topright", levels(aml$x), lty=1:2)
lines(c(0,23), c(0.5, 0.5), lty=4, col=2, lwd=2)
lines(c(23, 23), c(0.5, -1), lty=4, col=2, lwd=2)
axis(1, 23, col=2)
axis(2, 0.5, col=2)
```



Thus, the median survival time is 23 months. If we look at the table we generated previously, we see that 23 is the first value such that survival is lower than 0.5. The `quantile` function applied to a `Surv` object gives us (by default) the median, the first and the third quantiles.

```
quantile(km.fit.x)
```

```
$quantile
                25 50 75
x=Maintained    18 31 48
x=Nonmaintained  8 23 33

$lower
                25 50 75
x=Maintained    13 18 34
x=Nonmaintained  5  8 27

$upper
                25 50 75
x=Maintained    NA NA NA
x=Nonmaintained 30 NA NA
```

This function also gives us the 95% confidence intervals of these measures (`$lower` and

`$upper`). `NA` values indicate that either the survival curve or its corresponding 95% confidence interval do not fall to the desired quantile.

Comparing survival curves

Now we can consider the problem of comparing survival curves among groups. The **logrank test** is one of the most widely used tests for comparing survival curves. This test can only be applied if the following conditions are satisfied:

- The observations of the groups we want to compare are independent.
- Censoring pattern is similar in all groups.
- The hazards of event are proportional among groups (in particular, the survival curves do not cross).

The last condition means that, for any t, the hazard of the population A is equal to that of the population B multiplied by a constant θ:

$$h_A(t) = \theta \cdot h_B(t)$$

The parameter θ is known as **hazard ratio**. This last assumption is called **proportional hazards assumption**. In terms of survival:

$$S_A(t) = S_B(t)^{\theta}.$$

Thus, comparing survival curves is equivalent to the following hypothesis test:

$$\begin{aligned} H_0 &: \theta = 1 \\ H_A &: \theta \neq 1 \end{aligned}$$

This can be tested with the function `survdiff`:

```
survdiff(t~x, data=aml) # logrank test
```

```
Call:
survdiff(formula = t ~ x, data = aml)

                 N Observed Expected (O-E)^2/E (O-E)^2/V
x=Maintained    11        7    10.69      1.27       3.4
x=Nonmaintained 12       11     7.31      1.86       3.4

 Chisq= 3.4  on 1 degrees of freedom, p= 0.07
```

We obtain $p = 0.07$. If we set the type I error α to 5%, we do not reject the null hypothesis, so we cannot conclude that survival is different between the patients treated with standard course of chemotherapy and those treated with additional cycles.

Sample size

Computing the sample size in a survival analysis can be complicated because the number of factors to account for as the lost of follow-up or drop outs that will be in the sample.

Furthermore, commonly there are two different periods in a study:

- the **enrollment period**. Subjects are entering in the study sequentially.
- the **follow-up period**. No more subjects enter in the study and recruited subjects are follow up to the end of the study.

Most of the approaches to compute sample size in survival analysis need to specify the times of the enrollment and follow-up period.

It is also assumed that every subject is randomly assigned to one of the treatment groups.

Here, we will use the function *nSurvival(lambda1,lambda2,Ts,Tr,beta)* from *gsDesign* package. The arguments are:

- *lambda1* and *lambda2*. Event hazard rate for placebo (or baseline) and treatment group respectively.
- *Ts*. Manimujm study duration.
- *Tr*. Enrollment period duration.
- *beta*. The opposite of the power

Let us suppose that we want to detect as significant a hazard ratio of 2 with a power of 80%. The length of the study is 3 years with an enrollment period of 1 year.

```r
library(gsDesign)
ss <- nSurvival(lambda1 = 2, lambda2 = 1, Ts = 3, Tr = 1, beta=0.2)
ss$n
```

```
[1] 67.55026
```

The sample size necessary is 68 subjects.