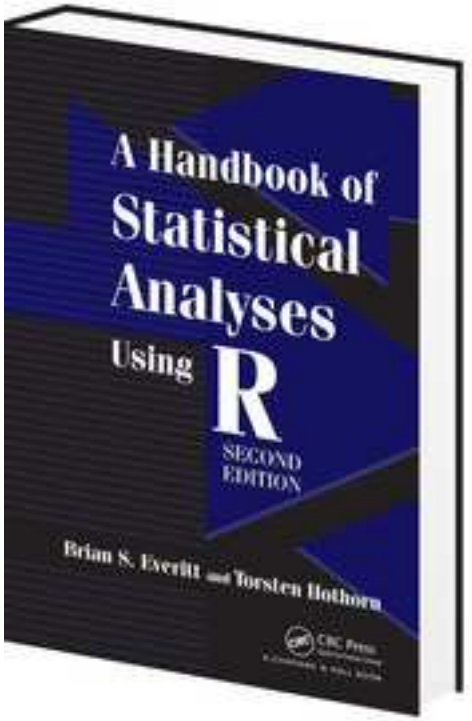
Analisi di sopravvivenza



fully perform a certain task in engineering. Such observations are generally referred to by the generic term *survival data* even when the endpoint or event being considered is not death but something else. Such data generally require special techniques for analysis for two main reasons:

1. Survival data are generally not symmetrically distributed – they will often appear positively skewed, with a few people surviving a very long time compared with the majority; so assuming a normal distribution will not be reasonable.
2. At the completion of the study, some patients may not have reached the endpoint of interest (death, relapse, etc.). Consequently, the exact survival times are not known. All that is known is that the survival times are greater than the amount of time the individual has been in the study. The survival times of these individuals are said to be *censored* (precisely, they are right-censored).

Of central importance in the analysis of survival time data are two functions used to describe their distribution, namely the *survival* (or *survivor*) *function* and the *hazard function.*

The survivor function, $S(t)$, is defined as the probability that the survival time, T, is greater than or equal to some time t, i.e.,

$$S(t) = \mathsf{P}(T \geq t)$$

When there are describing the survival experience of a group of individuals. When there are no censored observations in the sample of survival times, a non-parametric survivor function can be estimated simply as

$$\hat{S}(t) = \frac{\text{number of individuals with survival times} \geq t}{n}$$

where n is the total number of observations. Because this is simply a propor-

observations are present. In the presence of censoring, the survivor function is typically estimated using the *Kaplan-Meier* estimator (Kaplan and Meier,

stimatore Kaplan Meier

- permette di stimare la curva di sopravvivenza
  - valido anche nel caso censurato a destra
- permette di effettuare un test statistico non parametrico, simile al ChiQuadro, detto **logrank test** , tra due curve
- *non* permette l'analisi delle covariate

Of central importance in the analysis of survival time data are two functions used to describe their distribution, namely the *survival* (or *survivor*) *function* and the *hazard function.*

risks is the hazard function, $h(t)$, defined as the probability that an individual experiences the event in a small time interval, s, given that the individual has survived up to the beginning of the interval, when the size of the time interval approaches zero; mathematically this is written as

$$h(t) = \lim_{s \to 0} \mathsf{P}(t \leq T \leq t + s | T \geq t)$$

where T is the individual's survival time. The conditioning feature of this definition is very important. For example, the probability of dying at age 100 is very small because most people die before that age; in contrast, the probability of a person dying at age 100 who has reached that age is much greater.

relazioni matematiche

- il rischio h(t), la densità di probabilità d(t), la probabilità P(t) e la funzione di sopravvivenza S(t) sono legate tra loro matematicamente:

$$h(t) = d(t) \ / \ S(t)$$

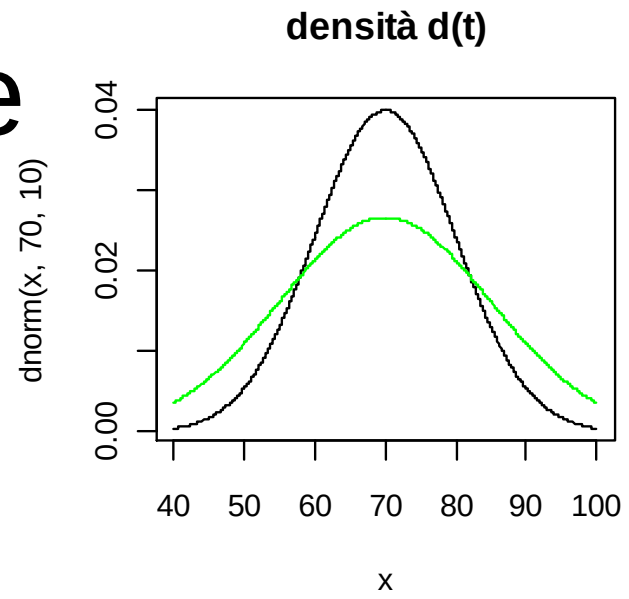
$$S = 1 - P$$

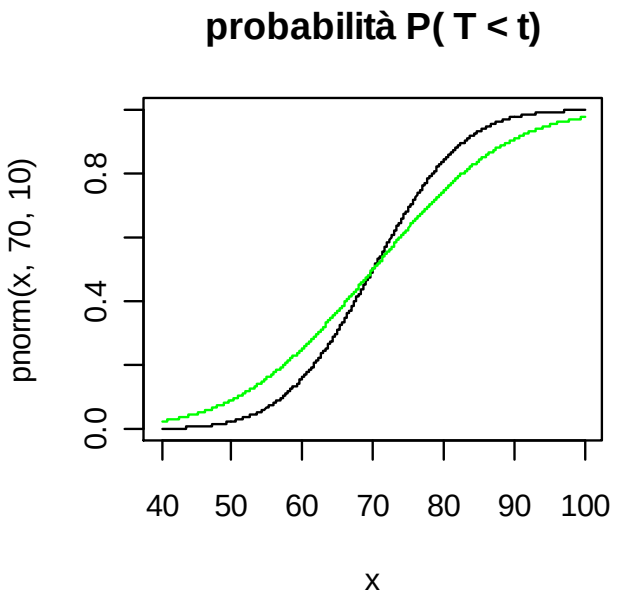
$$S(t) = \exp(-H(t))$$

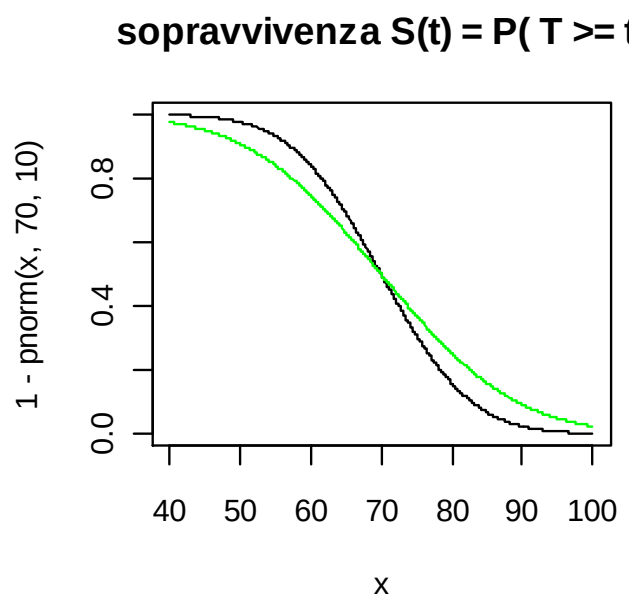
where $H(t)$ is known as the *integrated hazard* or *cumulative hazard*, and is defined as follows:

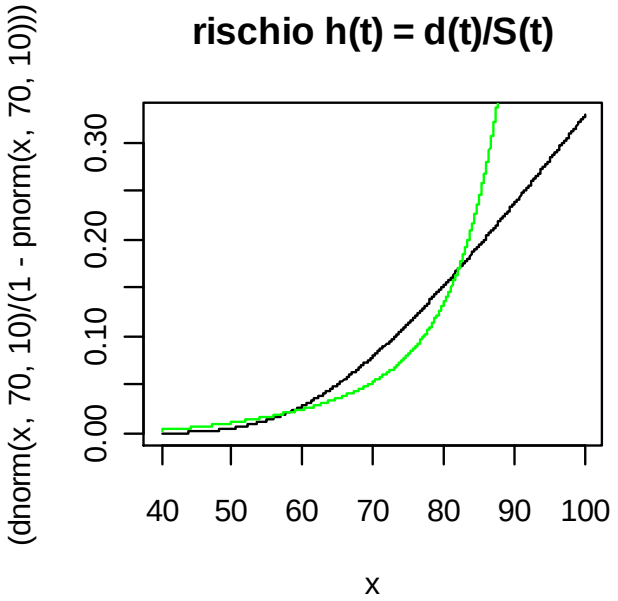
$$H(t) = \int_0^t h(u)du,$$

Normale

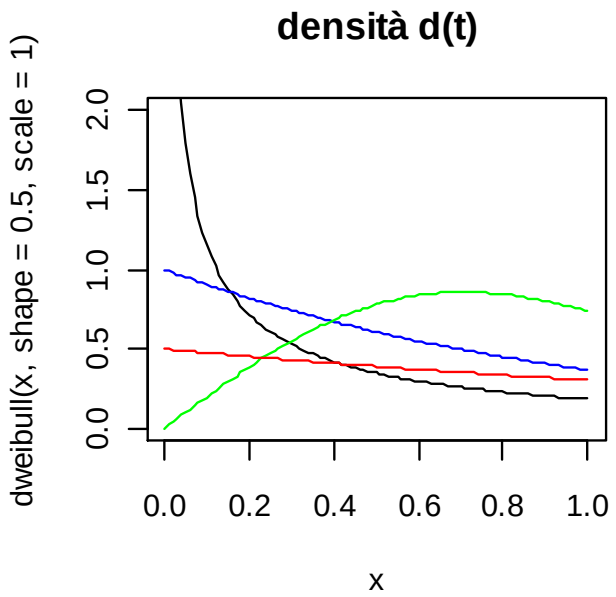

densità d(t)
dnorm(x, 70, 10)
0.00
0.02
0.04
40
50
60
70
80
90
100
x

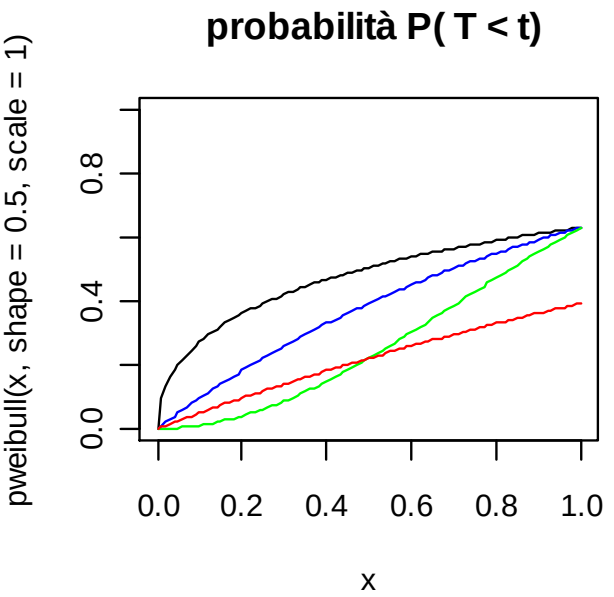

probabilità P(T < t)
pnorm(x, 70, 10)
0.0
0.4
0.8
40
50
60
70
80
90
100
x

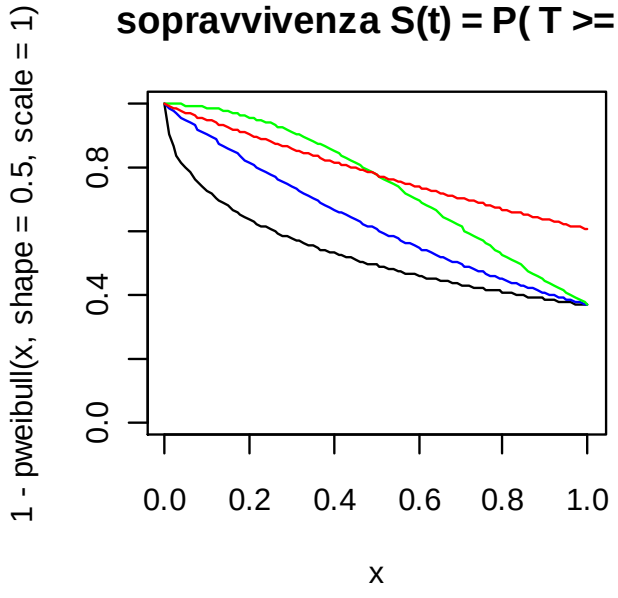

sopravvivenza S(t) = P(T >= t)
1 - pnorm(x, 70, 10)
0.0
0.4
0.8
40
50
60
70
80
90
100
x

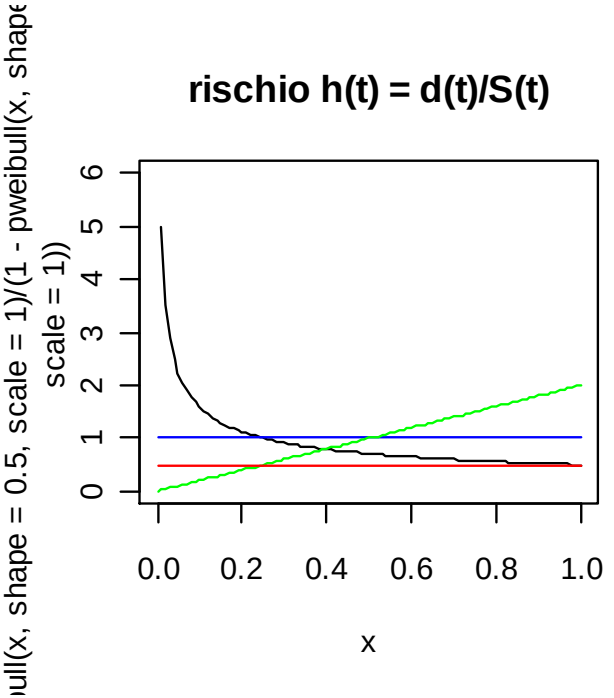

rischio h(t) = d(t)/S(t)
(dnorm(x, 70, 10)/(1 - pnorm(x, 70, 10)))
0.00
0.10
0.20
0.30
40
50
60
70
80
90
100
x

Weibull


densità d(t)
dweibull(x, shape = 0.5, scale = 1)
0.0
0.5
1.0
1.5
2.0
0.0
0.2
0.4
0.6
0.8
1.0
x


probabilità P(T < t)
pweibull(x, shape = 0.5, scale = 1)
0.0
0.4
0.8
0.0
0.2
0.4
0.6
0.8
1.0
x


sopravvivenza S(t) = P(T >= t)
1 - pweibull(x, shape = 0.5, scale = 1)
0.0
0.4
0.8
0.0
0.2
0.4
0.6
0.8
1.0
x


rischio h(t) = d(t)/S(t)
ull(x, shape = 0.5, scale = 1)/(1 - pweibull(x, shape
scale = 1))
0
1
2
3
4
5
6
0.0
0.2
0.4
0.6
0.8
1.0
x

9.2.3 *Cox's Regression*

When the response variable of interest is a possibly censored survival time, we need special regression techniques for modelling the relationship of the response to explanatory variables of interest. A number of procedures are available but the most widely used by some margin is that known as *Cox's proportional hazards model*, or *Cox's regression* for short. Introduced by Sir

include in (9.1). The problem is overcome in the proportional hazards model proposed by Cox (1972) by allowing the form of dependence of $h(t)$ on t to remain unspecified, so that

$$\log(h(t)) = \log(h_0(t)) + \beta_1 x_1 + \ldots + \beta_q x_q$$

where $h_0(t)$ is known as the *baseline hazard function*, being the hazard function for individuals with all explanatory variables equal to zero. The model can be rewritten as

$$h(t) = h_0(t) \exp(\beta_1 x_1 + \ldots + \beta_q x_q).$$

la funzione di rischio baseline

- talvolta è comodo scegliere una forma particolare per $h_0(t)$ (**forma parametrica**)
  - es. "Weibull"
    - `library (survival)`
- talvolta (ad es. se le covariate cambiano con il tempo) è comodo (**forma non parametrica**) non specificare $h_0(t)$
  - es. "Breslow / Nelson-Aalen"
    - `library (timereg)`