Contents

§1 Survival model

§1.1 The Hazard and Survival functions

Survival or lifetime data can arise in many areas, e.g. lifetime of components in industry, survival time of cancer patients and time in unemployment. The major complication in analysing survival data is when the study ends before all lifetimes are observed. This form of censoring is called *right censoring*.

1. The survival time for a population T is a rv with a density function $f(t)$ and distribution function $F(t) = \Pr(T \leq t)$ which gives the probability that the event has occured by duration t.

2. The *survivor function* is the fraction still surviving at time t

$$S(t) = \Pr(T \geq t) = 1 - F(t) = \int_t^\infty f(x)dx$$

   if T is continuous.

3. The *hazard function* or *hazard density* measures the instantaneous risk, the probabilty of failing at t, given survival at t.

   If T are discrete and take values $t_1 < t_2 < \ldots$ at probabilities $\Pr(T = t_j) = f_j, \quad j = 0, 1, 2, \ldots$ then

$$h(t_j) = \frac{\Pr(T = t_j)}{\Pr(T \geq t_j)} = \frac{f_j}{f_j + f_{j+1} + \ldots}.$$

   If T is continuous then

$$h(t) = \lim_{\delta t \to 0} \frac{\Pr(t < T \leq t + \delta t | \, T \geq t)}{\delta t}$$

Since

$$\begin{aligned}
\Pr(T > t+\delta t) &= \Pr(T \geq t)\,[1 - \Pr(t < T \leq t+\delta t |\, T \geq t)],\\
1 - F(t+\delta t) &\simeq (1-F(t))(1-h(t)\delta t),\\
1 - F(t+\delta t) &\simeq (1-F(t)) - h(t)\delta t(1-F(t)),\\
h(t)(1-F(t)) &\simeq \frac{F(t+\delta t)-F(t)}{\delta t} = \frac{f(t)\delta t}{\delta t} = f(t)
\end{aligned}$$

Taking limits on $\delta t \to 0$, we get

$$\begin{aligned}
h(t) &= \frac{f(t)}{1-F(t)} = \frac{f(t)}{S(t)} = -\frac{d}{dt}\ln(1-F(t)) = -\frac{d}{dt}\ln S(t) \quad \text{and}\\
S(t) &= \exp\left[-\int_0^t h(x)dx\right] = \exp[-H(t)].
\end{aligned}$$

4. The *cumulative hazard* function is

$$H(t) = \int_0^t h(x)dx = -\ln[S(t)] = -\ln[1-F(t)].$$

5. The *expectation of life* in demographic studies is

$$\mu = \int_0^\infty t f(t)dt = -\int_0^\infty t\, dS(t) = [-tS(t)]_0^\infty + \int_0^\infty S(t)dt = \int_0^\infty S(t)dt$$

since $S(t) = 1 - F(t)$ implies $S'(t) = -f(t)$.

Example: The simplest survival time model is the exponential distribution model, i.e. $T \sim \mathcal{E}xp(\lambda)$ with $E(T) = \dfrac{1}{\lambda}$,

$$f(t) = \lambda e^{-\lambda t},\ F(t) = 1 - e^{-\lambda t},\ S(t) = e^{-\lambda t},\ h(t) = \frac{f(x)}{S(x)} = \lambda,\ H(t) = \lambda t.$$

Note that the hazard function is constant. In fact, $h(t)$ is constant iff T follows exponential distribution.

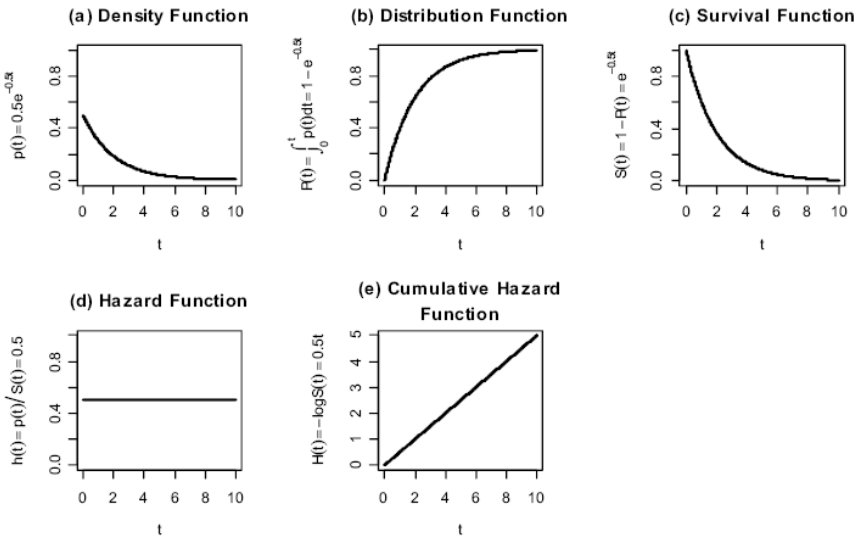


Figure 6. Several representations of an exponential distribution with rate parameter $\lambda = 0.5$.

§1.2 Censoring Mechanism

1. **Type I (Fixed) censoring:** a sample of n units is followed for a fixed in advance time τ, and the number of 'death' during time τ is random. The probability that unit i will survive after time τ_i is $S(\tau_i)$.

2. **Type II censoring:** a sample of n units is followed as long as necessary until d units which is fixed in advance are dead, and the total duration of the study is random.

3. **Random censoring:** A more general design has a potential censoring time C_i and a potential lifetime T_i which are assumed to be independent rv and we observe $Y_i = \min(C_i, T_i)$. An indicator variable $\omega_i = 1$ if the observation is terminated by death and 0 by censoring.

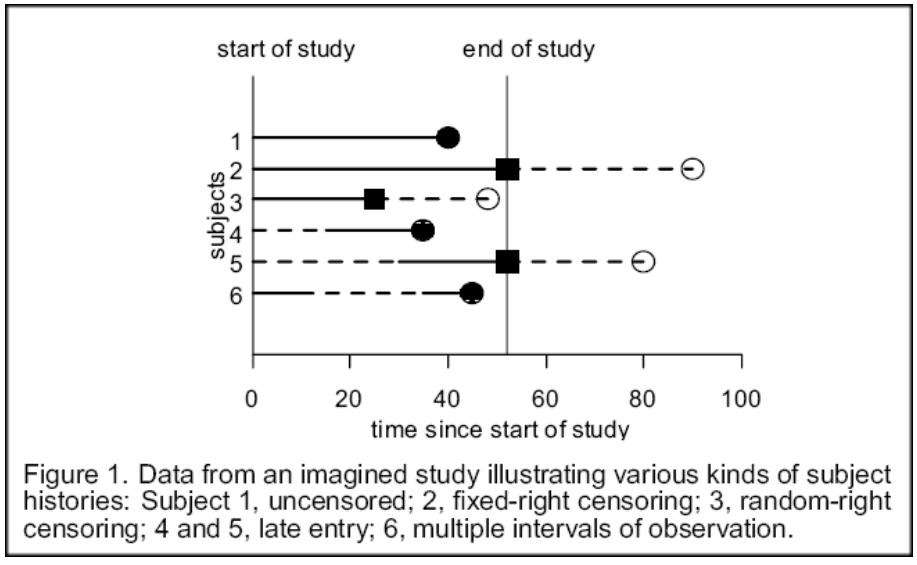


Figure 1. Data from an imagined study illustrating various kinds of subject histories: Subject 1, uncensored; 2, fixed-right censoring; 3, random-right censoring; 4 and 5, late entry; 6, multiple intervals of observation.

The basic assumption is that the censoring should not provide any information regarding the survival of a unit beyond the censoring time and so they are called *non-informative* censoring.

§1.3 The Kaplan Meier (KM) Estimator

The KM estimator is a nonparametric estimator for $S(t)$. It estimates $S(t)$ without imposing a distribution on T. If failures can occur at $t_1 < t_2 < \ldots$ and h_j is the hazard function at time t_j such that

$$h_j = h(t_j) = \frac{\Pr(T = t_j)}{\Pr(T \geq t_j)} = \frac{f_j}{f_j + f_{j+1} + \ldots}.$$

where $f_j = \Pr(T = t_j)$. Hence

$$1 - h_j = \frac{f_{j+1} + f_{j+2} + \ldots}{f_j + f_{j+1} + \ldots}.$$

Then, for $t_j < t \leq t_{j+1}$,

$$\begin{aligned} S(t) &= \sum_{k:t_k \geq t} f_k = f_{j+1} + f_{j+2} + \ldots \\ &= \frac{f_2 + f_3 + \ldots}{f_1 + f_2 + \ldots} \times \frac{f_3 + f_4 + \ldots}{f_2 + f_3 + \ldots} \times \cdots \times \frac{f_{j+1} + f_{j+2} + \ldots}{f_j + f_{j+1} + \ldots} \\ &= \prod_{k:t_k \leq t} (1 - h_k) = (1 - h_1)(1 - h_2) \ldots (1 - h_j) \end{aligned}$$

since $f_1 + f_2 + \cdots = 1$. This result states that to survive to time t_j one must first survive t_1, then survive t_2 given that one survived t_1, and so on, finally surviving to t_j given survival up to t_{j-1}. Then the KM estimator for $S(t)$ is

$$\widehat{S}(t) = \prod_{j:t_j \leq t} (1 - \hat{h}_j) = \prod_{j:t_j \leq t} \left(1 - \frac{d_j}{r_j}\right)$$

where $\hat{h}_j = \left(\frac{d_j}{r_j}\right)$ is the ML estimator for h_j when there are d_j failures among the r_j subjects at risk at time t_j. If individuals are *censored* at t_j then these are included in the r_j count but they do not appear in the failure count at any stage.

Example: (Leukaemia) Two samples of 21 patients each were used in the study. One sample was treated with 6-mercaptopurine and the other sample received a placebo. The time of remission is given in weeks. An asterisk denotes censored data.

6-MCP	6*	6	6	6	7	9*	10*	10	11*	13	16	17*	19*	20*	22	23	25*	32*	32*	34*	35*
Control	1	1	2	2	3	4	4	5	5	8	8	8	8	11	11	12	12	15	17	22	23

There are $j = 17$ failure times. The combined data is

Time t_j	1	2	3	4	5	6	7	8	10	11	12	13	15	16	17	22	23	Total
T1																		
Death d_{1j}						3	1	0	1	0		1	0	1	0	1	1	9
Censor c_{1j}						${}_{6}1$	0	${}_{9}1$	${}_{10}1$	${}_{11}1$		0	0	0	3	0	5	12
Risk r_{1j}	21	21	21	21	21	21	17	16	15	13	12	12	11	11	10	7	6	
Hazard h_{1j}	$\frac{0}{21}$					$\frac{3}{21}$	$\frac{1}{17}$	$\frac{0}{16}$	$\frac{1}{15}$	$\frac{0}{13}$		$\frac{1}{12}$	$\frac{0}{11}$	$\frac{1}{11}$	$\frac{0}{10}$	$\frac{1}{7}$	$\frac{1}{6}$	
$1 - h_{1j}$	1					$\frac{18}{21}$	$\frac{16}{17}$	1	$\frac{14}{15}$	1		$\frac{11}{12}$	1	$\frac{10}{11}$	1	$\frac{6}{7}$	$\frac{5}{6}$	
Survival S_{1j}	1					$\frac{18}{21}$	$\frac{96}{119}$	$\frac{96}{119}$	$\frac{192}{255}$	$\frac{192}{255}$		$\frac{176}{255}$	$\frac{176}{255}$	$\frac{32}{51}$	$\frac{32}{51}$	$\frac{64}{119}$	$\frac{160}{357}$	
T2																		
Death d_{2j}	2	2	1	2	2			4		2	2		1		1	1	1	21
Risk r_{2j}	21	19	17	16	14	12	12	12	8	8	6	4	4	3	3	2	1	
Hazard h_{2j}	$\frac{2}{21}$	$\frac{2}{19}$	$\frac{1}{17}$	$\frac{2}{16}$	$\frac{2}{14}$			$\frac{4}{12}$		$\frac{2}{8}$	$\frac{2}{6}$		$\frac{1}{4}$		$\frac{1}{3}$	$\frac{1}{2}$	1	
$1 - h_{2j}$	$\frac{19}{21}$	$\frac{17}{19}$	$\frac{16}{17}$	$\frac{14}{16}$	$\frac{12}{14}$			$\frac{8}{12}$		$\frac{6}{8}$	$\frac{4}{6}$		$\frac{3}{4}$		$\frac{2}{3}$	$\frac{1}{2}$	0	
Survival S_{2j}	$\frac{19}{21}$	$\frac{17}{21}$	$\frac{16}{21}$	$\frac{14}{21}$	$\frac{12}{21}$			$\frac{8}{21}$		$\frac{6}{21}$	$\frac{4}{21}$		$\frac{3}{21}$		$\frac{2}{21}$	$\frac{1}{21}$	$\frac{0}{21}$	
All																		
Death d_j	2	2	1	2	2	3	1	4	1	2	2	1	1	1	1	2	2	30
Risk r_j	42	40	38	37	35	33	29	28	23	21	18	16	15	14	13	9	7	
Hazard h_j	$\frac{2}{42}$	$\frac{2}{40}$	$\frac{1}{38}$	$\frac{2}{37}$	$\frac{2}{35}$	$\frac{3}{33}$	$\frac{1}{29}$	$\frac{4}{28}$	$\frac{1}{23}$	$\frac{2}{21}$	$\frac{2}{18}$	$\frac{1}{16}$	$\frac{1}{15}$	$\frac{1}{14}$	$\frac{1}{13}$	$\frac{2}{9}$	$\frac{2}{7}$	
$E_{1j} = r_{1j}h_j$	$\frac{42}{42}$	$\frac{42}{40}$	$\frac{21}{38}$	$\frac{42}{37}$	$\frac{42}{35}$	$\frac{63}{33}$	$\frac{17}{29}$	$\frac{64}{28}$	$\frac{15}{23}$	$\frac{26}{21}$	$\frac{26}{18}$	$\frac{12}{16}$	$\frac{11}{15}$	$\frac{11}{14}$	$\frac{10}{13}$	$\frac{14}{9}$	$\frac{12}{7}$	19.3
$E_{2j} = r_{2j}h_j$	$\frac{42}{42}$	$\frac{38}{40}$	$\frac{17}{38}$	$\frac{32}{37}$	$\frac{28}{35}$	$\frac{36}{33}$	$\frac{12}{29}$	$\frac{48}{28}$	$\frac{8}{23}$	$\frac{16}{21}$	$\frac{12}{18}$	$\frac{4}{16}$	$\frac{4}{15}$	$\frac{3}{14}$	$\frac{3}{13}$	$\frac{4}{9}$	$\frac{2}{7}$	10.7

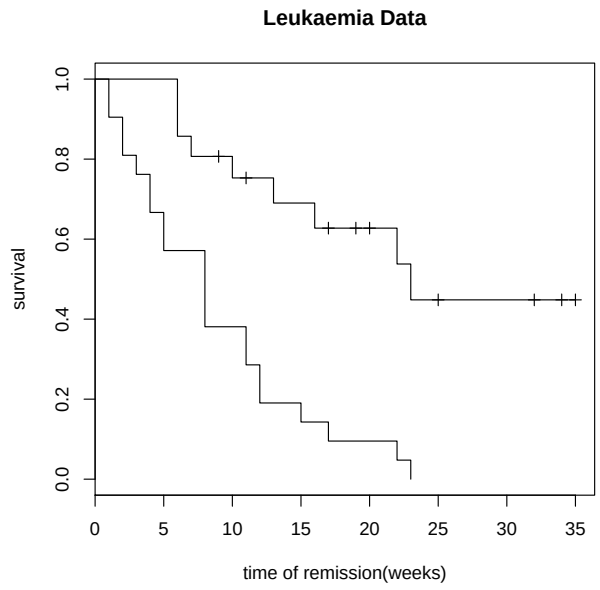
```
>#load package survival
> t=c(6,6,6,6,7,9,10,10,11,13,16,17,19,20,22,23,25,32,32,34,35,
```

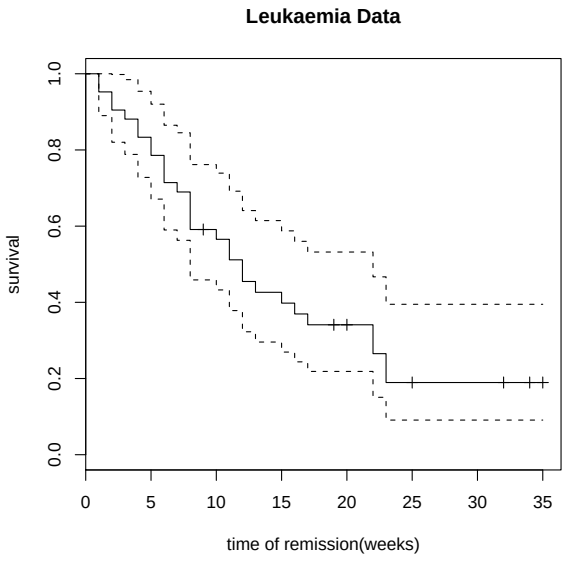
```
+ 1,1,2,2,3,4,4,5,5,8,8,8,8,11,11,12,12,15,17,22,23)
> w=c(0,1,1,1,1,0,0,1,0,1,1,0,0,0,1,1,0,0,0,0,0,
+ 1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1)  #0 censor
> c=c(rep(0,21),rep(1,21)) #0 treatment
> l=log(t)
> surv1=survfit(Surv(t,w)~c)
> surv0=survfit(Surv(t,w)~1)
> summary(surv1)
Call: survfit(formula = Surv(t, w) ~ c)

                c=0
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    6     21       3    0.857  0.0764        0.720        1.000
    7     17       1    0.807  0.0869        0.653        0.996
   10     15       1    0.753  0.0963        0.586        0.968
   13     12       1    0.690  0.1068        0.510        0.935
   16     11       1    0.627  0.1141        0.439        0.896
   22      7       1    0.538  0.1282        0.337        0.858
   23      6       1    0.448  0.1346        0.249        0.807

                c=1
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    1     21       2   0.9048  0.0641      0.78754        1.000
    2     19       2   0.8095  0.0857      0.65785        0.996
    3     17       1   0.7619  0.0929      0.59988        0.968
    4     16       2   0.6667  0.1029      0.49268        0.902
    5     14       2   0.5714  0.1080      0.39455        0.828
    8     12       4   0.3810  0.1060      0.22085        0.657
   11      8       2   0.2857  0.0986      0.14529        0.562
   12      6       2   0.1905  0.0857      0.07887        0.460
   15      4       1   0.1429  0.0764      0.05011        0.407
   17      3       1   0.0952  0.0641      0.02549        0.356
   22      2       1   0.0476  0.0465      0.00703        0.322
   23      1       1   0.0000      NA           NA           NA

> summary(surv0)
```

```
Call: survfit(formula = Surv(t, w) ~ 1)
```

time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
1	42	2	0.952	0.0329	0.890	1.000
2	40	2	0.905	0.0453	0.820	0.998
3	38	1	0.881	0.0500	0.788	0.985
4	37	2	0.833	0.0575	0.728	0.954
5	35	2	0.786	0.0633	0.671	0.920
6	33	3	0.714	0.0697	0.590	0.865
7	29	1	0.690	0.0715	0.563	0.845
8	28	4	0.591	0.0764	0.459	0.762
10	23	1	0.565	0.0773	0.433	0.739
11	21	2	0.512	0.0788	0.378	0.692
12	18	2	0.455	0.0796	0.323	0.641
13	16	1	0.426	0.0795	0.296	0.615
15	15	1	0.398	0.0791	0.269	0.588
16	14	1	0.369	0.0784	0.244	0.560
17	13	1	0.341	0.0774	0.219	0.532
22	9	2	0.265	0.0765	0.151	0.467
23	7	2	0.189	0.0710	0.091	0.395

```
> plot(surv1,xlab="time of remission(weeks)", ylab="survival",
  main="Leukaemia Data")
```





An estimate of the variance of $S(t)$ is given by *Greenwood's formula*:

$$\widehat{V}ar[\widehat{S}(t)] = [\widehat{S}(t)]^2 \sum_{t_i \le t} \frac{d_i}{r_i(r_i - d_i)}$$

and $\widehat{S}(t) \pm 1.96\sqrt{\widehat{V}ar[\widehat{S}(t)]}$ gives a point-wise 95-percent confidence envelope around the estimated survival function. Proof is asked in the assignment. For example,

$$\begin{aligned}
\text{Var}[\widehat{S}_2(5)] &\simeq [\widehat{S}_2(5)]^2 \sum_{j:t_j \le 5} \frac{d_j}{r_j(r_j - d_j)} \\
&= \frac{12^2}{21^2}\left[\frac{2}{21(21-2)} + \frac{2}{19(19-2)} + \frac{1}{17(17-1)} + \right. \\
&\qquad \left. \frac{2}{16(16-2)} + \frac{2}{14(14-2)}\right] \\
&= 0.1080^2
\end{aligned}$$

The *log rank test* can be used to compare two survival curves. Under the null hypothesis of no difference between the two survival curves, the test statistic is

$$\chi^2_{\text{logrank}} = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} \overset{H_0}{\sim} \chi^2_1$$

where the O_1 and O_2 are the total numbers of observed events in groups 1 and 2, respectively, and E_1 and E_2 the total numbers of expected events.

```
> survdiff(Surv(t,w)~c)
Call:
survdiff(formula = Surv(t, w) ~ c)

     N Observed Expected (O-E)^2/E (O-E)^2/V
c=0 21        9     19.3      5.46      16.8
```

```
c=1 21        21      10.7       9.77       16.8

 Chisq= 16.8  on 1 degrees of freedom, p= 4.17e-05
```

Since

$$\chi^2_{\text{logrank}} = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} = \frac{(9 - 19.3)^2}{19.3} + \frac{(21 - 10.7)^2}{10.7} = 16.8$$

and the p-value =0.000, we reject H_0 of no difference between the two survival curves.