Survival Analysis - part 1

- Often the response of interest in a study is the length of time, T, until an event occurs.
- This could be the time from birth until death, the time from transplant surgery until the new organ fails, the time until progression from one stage of a disease to another, or length of remission from disease, etc.
- When death is the event, T is called the survival time, and this is the name used for T in other situations as well.
- The **survival function** $S(t)$ gives the probability of survival beyond a given time, i.e.
$$S(t) = P(T > t)$$
- This probability decreases from one at $T = 0$ to zero when $T = \infty$.

Example

Time of remission (weeks) of leukemia patients, treated with 6-mercaptopurine (sample 1), and placebo (sample 2) (Freireich et al, *Blood*, 1963)

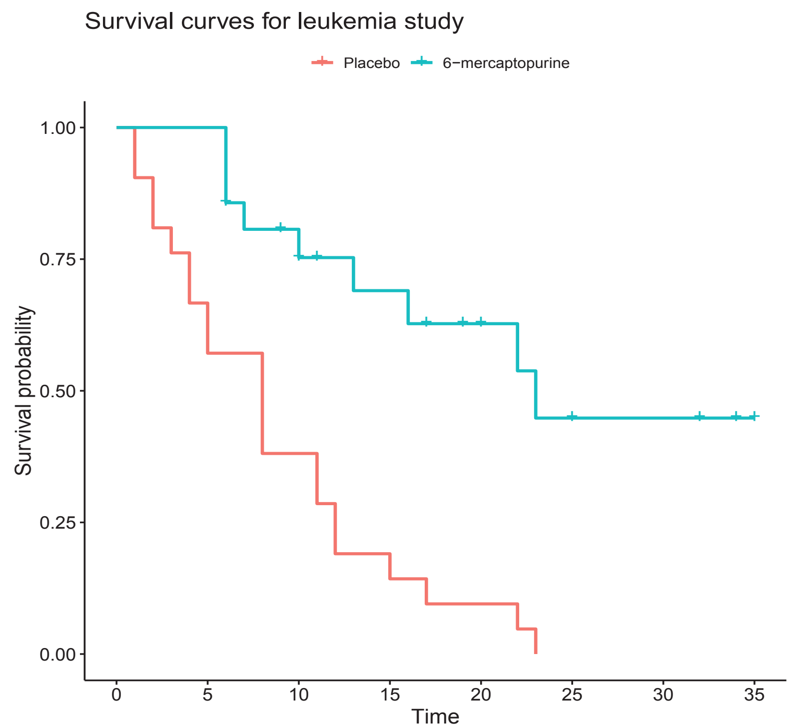
Sample 1	(6)	6	6	6	7	(9)	(10)
	10	(11)	13	16	(17)	(19)	(20)
	22	23	(25)	(32)	(32)	(34)	(35)
sample 2	1	1	2	2	3	4	4
	5	5	8	8	8	8	11
	11	12	12	15	17	22	23

- A feature of survival data is that there are often **censored** values, typically denoted by bracketed values or with the + symbol. For example, the first time, (6) in Sample 1, is censored, indicating that the survival time for that individual is **at least** 6 weeks. It might have been denoted 6+.

- Subjects may be censored because they are lost to observation, because they move away, quit the trial, die from other causes, or have not died before the end of a study.

- In this example, we would like to determine whether a treatment prolongs survival, i.e. whether the survival curve is shifted to the right relative to the control.

- The Kaplan-Meier estimate of the survival curve is a step function which decreases at each observed failure time, sometime including ticks at censoring times.



- Upper curve (treatment group) shows longer remission.

- Lower curve falls to zero as everyone in this group ended remission. Upper curve does not fall to 0, because the longest time for the treatment group is censored.

- **When there is no censoring**, survival at t is estimated by the proportion surviving beyond t

$$\hat{S}(t) = \frac{\#\ subjects\ with\ T > t}{total\ sample\ size}.$$

- For the control group

Time	No. failures	No. survivors	$\hat{S}(t)$
0	0	21	1
1	2	19	19/21
2	2	17	17/21
3	1	16	16/21
4	2	14	14/21
5	2	12	12/21
8	4	8	8/21
11	2	6	6/21
12	2	4	4/21
15	1	3	3/21
17	1	2	2/21
22	1	1	1/21
23	1	0	0
Total	21		

When there is censoring, another approach is required.

- where $t_i, i = 1, 2, \ldots$ are the unique ordered survival times (but not including censoring times), we can write

$$P(T > t_i) = P(T > t_{i-1})P(T > t_i | T > t_{i-1})$$

  or

$$S(t_i) = S(t_{i-1})P(T > t_i | T > t_{i-1})$$

- The second term is estimated by the proportion of those at risk at t_i who survive past t_i.

- The number at risk at t_i is the overall sample size n, minus the number of deaths or failures before t_i, minus the number censored before t_i.

- The calculations are summarized below for the treatment group.

Kaplan Meier example

Time	No. at risk	No. of failures	No. surviving	Prop. surv.	$\hat{S}(t)$
6	21	3	18	18/21	.857
7	17	1	16	16/17	.857(16/17)=.807
10	15	1	14	14/15	.753
13	12	1	11	11/12	.690
16	11	1	10	10/11	.627
22	7	1	6	6/7	.538
23	6	1	5	5/6	.448

- Note that when the last observation is censored the survival curve does not drop to zero.

Some computer programs will also give standard errors and confidence intervals.

```
 leuktr.km=survfit(leuktr.Surv~1)
> print(leuktr.km)
Call: survfit(formula = leuktr.Surv ~ 1)

records   n.max n.start  events  median 0.95LCL 0.95UCL
     21      21      21       9      23      16      NA
> summary(leuktr.km)
Call: survfit(formula = leuktr.Surv ~ 1)

 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
```

- Note that the standard error gets larger as time goes on, and that the confidence intervals are very large due to the small sample size.