

## Medical Statistics: Lecture 22

**Survival Analysis** refers to the analysis of **time to event data**. The time to an event is called the **failure time** and is treated as a random variable. Such data arise in many fields including medicine, engineering, economics, and many others. We will be concerned with applications in medicine.

Many medical studies involve the time to an event. Examples include

- the time to AIDS for HIV positive individuals,
- the time to death for cancer patients,
- the time to first myocardial infarction,
- the time from transplant to graft-versus-host disease

1

Let  $X$  denote a random variable interpreted as the waiting time until some random event.

**Survival function:**  $S(x) = P[X > x]$

If  $P[X = 0] = 0$ , then  $S(0) = 1$ .

For  $0 \leq x < x'$ ,  $S(x) \geq S(x')$ , so  $S$  is nonincreasing.

2

### Example

Here we consider an example of a medical study resulting in survival data.

Nahman et al. (1992) designed a study to see if the time to first exit-site infection (in months) for patients undergoing kidney dialysis differed for those with a surgically placed catheter and patients with a percutaneous catheter. The sample sizes were 43 and 76 for the patients with surgically placed and percutaneously placed catheters, respectively. The data for those with a surgically placed catheter are given below.

The variable **Time** refers to the time until either an infection was observed, or the time that the patient was observed without yet suffering an infection (**censored**). The variable **infection** indicates which of these is true, and takes value 1 if an infection was observed. Catheter failure was the primary reason for censoring.

3

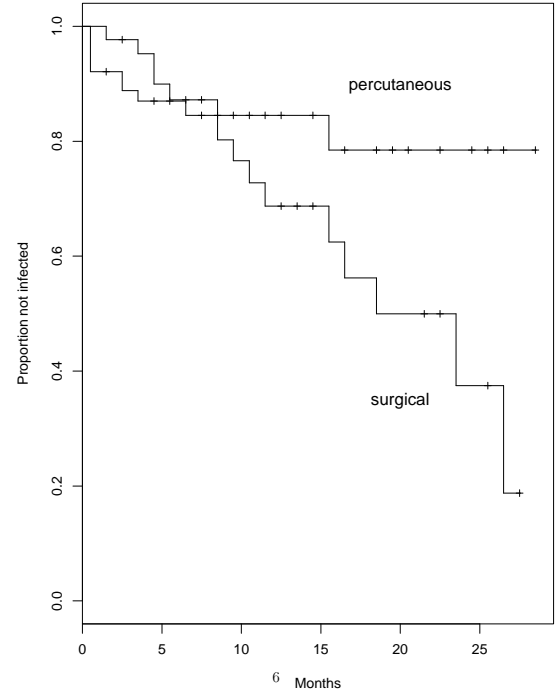
Time	infection
1.5	1
3.5	1
4.5	1
4.5	1
5.5	1
8.5	1
8.5	1
9.5	1
10.5	1
11.5	1
15.5	1
16.5	1
18.5	1
23.5	1
26.5	1
2.5	0
3.5	0
3.5	0
3.5	0
4.5	0
5.5	0
6.5	0
6.5	0
7.5	0
7.5	0
7.5	0
8.5	0
9.5	0
10.5	0
11.5	0
12.5	0
12.5	0
13.5	0
14.5	0
14.5	0
21.5	0
22.5	0
25.5	0
27.5	0

4

To compare the two methods of catheter placement we obtain a nonparametric estimate of the survival curve for time until infection for both groups. These are plotted on Figure 1. Aside from some early crossing, it appears that the time until infection tends to be much greater for percutaneous catheter placement, indicating that it might be a better choice.

5

Figure 1: Time until exit-site infection  
(product-limit estimate)



Assume that deaths in the dataset are recorded at  $D$  distinct time points,

$$t_1 < t_2 < t_3 < \dots < t_D.$$

We allow for the possibility of multiple deaths (ties) at any of the time points. Let  $t_{max}$  denote the maximum follow-up time for all subjects. Note that  $t_{max} \geq t_D$ .

7

Let  $Y_i$  denote the number of subjects at risk in the instant just before  $t_i$ , and let  $d_i$  denote the number of deaths at time  $t_i$ .

The **product-limit estimator** (Kaplan and Meier, 1958) is the standard nonparametric estimator of  $S(t)$ . We denote the product-limit estimator by  $\hat{S}(t)$  and define it over two intervals.

For  $t \in [0, t_1)$ ,  $\hat{S}(t) = 1$

For  $t \in [t_1, t_{max}]$ ,  $\hat{S}(t) = \prod_{t_i \leq t} [1 - \frac{d_i}{Y_i}]$

$\hat{S}(t)$  is not well defined for  $t > t_{max}$ .

8

As an example consider the data from section 1.2. Here we record the remission time (time until relapse) for children with acute leukemia. The trial included matched pairs and randomizing within each pair to 6-mercaptopurine (6-MP) or placebo. Patients were selected who had a complete or partial remission induced by treatment with prednisone.

In the example to follow we compute the product-limit estimator for the 6-MP children.

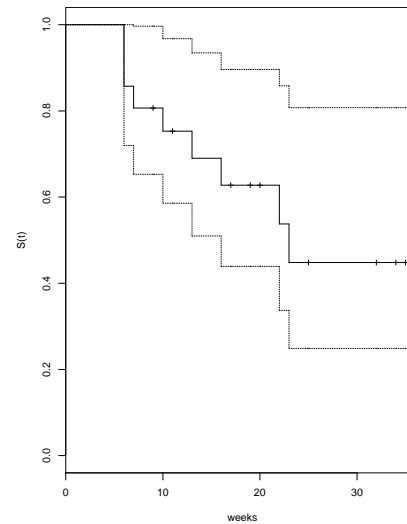
Table 1: Construction of Product-Limit Estimator for 6-MP Patients

$t_i$	$d_i$	$Y_i$	$\hat{S}(t)$
6	3	21	$[1-3/21]=.857$
7	1	17	$.857(1-1/17)=.807$
10	1	15	$.807(1-1/15)=.753$
13	1	12	$.753(1-1/12)=.690$
16	1	11	$.690(1-1/11)=.628$
22	1	7	$.628(1-1/7)=.538$
23	1	6	$.538(1-1/6)=.448$

Table 2:  $\hat{S}$  and  $\hat{V}$  for product-limit estimator

$t$	$\hat{S}(t)$	$\hat{V}[\hat{S}(t)]^{1/2}$
$0 \leq t < 6$	1.000	0.000
$6 \leq t < 7$	0.857	0.076
$7 \leq t < 10$	0.807	0.087
$10 \leq t < 13$	0.753	0.096
$13 \leq t < 16$	0.690	0.107
$16 \leq t < 22$	0.628	0.114
$22 \leq t < 23$	0.538	0.128
$23 \leq t < 35$	0.448	0.135

Figure 2: Estimated survival curve with 95% pointwise confidence intervals for 6-MP subjects



We can also use the product-limit estimator to estimate the  $p$ th quantile  $x_p$ . The median would be  $x_{.5}$

$$x_p = \min\{t : S(t) \leq 1 - p\}$$

The obvious estimator is then

$$\hat{x}_p = \min\{t : \hat{S}(t) \leq 1 - p\}.$$

Consider a Dutch study of 90 patients with cancer of the larynx. The data give the time from treatment to death or censoring, and patients are classified according the stage of the disease. Stages range from the least severe stage I (cancer is localized), to the most severe stage IV (cancer has spread around the body and to lymph nodes).

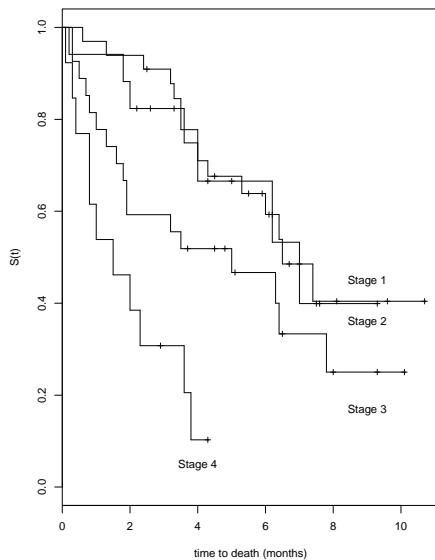
```
#### Variables in larynx data are
## stage: stage of cancer at time of first treatment
## time: death time or time on study (months)
## age: age at diagnosis
## year: year of diagnosis
## death: 1 for death, 0 for right-censored
```

```
### Let's look at the first 5 lines
```

```
larynx[1:5,]
```

```
      stage time age year death
1         1  0.6  77   76      1
2         1  1.3  53   71      1
3         1  2.4  45   71      1
4         1  2.5  57   78      0
5         1  3.2  58   74      1
```

Figure 3: Estimated survival curves by stage of cancer in larynx example



Estimate the median survival time for stages 3 and 4.

Now consider a study by the Gastrointestinal Tumor Study Group. 45 patients were randomized to each of two arms and followed for about eight years, to see if there was a survival difference between chemotherapy and chemotherapy combined with radiotherapy for locally unresectable gastric cancer.

```

### variables are
## time: time on study
## death: indicator of death or censoring
## treatment: 1=chemo, 0=chemo+rad

```

```

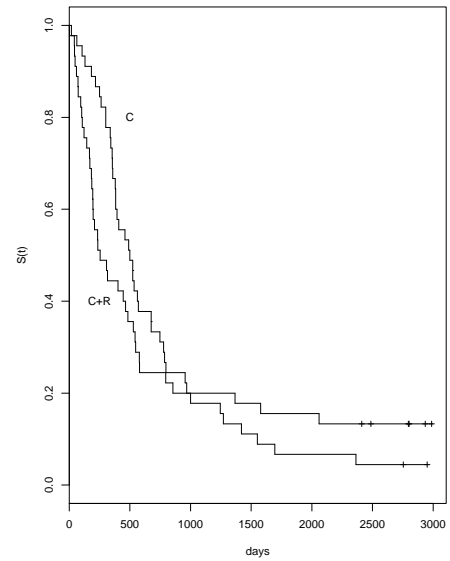
### look at first five lines

```

	time	death	treatment
1	1	1	1
2	63	1	1
3	105	1	1
4	129	1	1
5	182	1	1

17

Figure 4: Estimated survival curves by treatment in GI example



18

So far we have focused on estimating  $S(t)$  without conditioning on any covariates. However, it is often the case that variables which might influence survival are recorded such as gender, smoking history, dietary habits, blood pressure, heart rate, or physical activity level.

In some cases the primary interest is to find which of these covariates are **risk factors**, and in other cases we just wish to control for them because they are **confounders** for our real aim of assessing a treatment effect.

19

A function that plays a critical role in survival analysis is the *hazard function*,

$$h(t) = \frac{P[t \leq X < t + \Delta t \mid X \geq t]}{\Delta t}.$$

where  $\Delta t$  can be thought of as a very small positive number. The hazard function  $h(t)$  describes the instantaneous death rate for those who survive until time  $t$ .

The **Cox model** is a popular model in medical applications, that allows the use of covariates in modeling the hazard function.

20

Let  $h(t | z_1, z_2, \dots, z_p)$  denote the hazard rate at time  $t$  for an individual with values of  $p$  covariates  $z_1, z_2, \dots, z_p$ . The basic Cox model is

$$h(t | z_1, z_2, \dots, z_p) = h_0(t)e^{\sum_{j=1}^p \beta_j z_j}$$

where  $h_0(t)$  is an arbitrary **baseline hazard rate**, and  $\beta_1, \beta_2, \dots, \beta_p$  are the **risk coefficients**.

The resulting survival function is then

$$S(t | z_1, \dots, z_p) = [S_0(t)]e^{-\sum_{j=1}^p \beta_j z_j}$$

where  $S_0(t)$  is a function determined by  $h_0(t)$ .

For simplicity, let's assume that we have only one covariate in the model. The Cox model is often called a **proportional hazards model** because, if we look at the **relative risk** or **hazard ratio**, for two individuals with covariate values  $z$  and  $z^*$ , we obtain,

$$\frac{h(t | z)}{h(t | z^*)} = e^{\beta(z-z^*)}$$

which is a constant (does not depend on  $t$ ). Thus, the hazard rates are proportional for all  $t$ .

As a final example, we will look at a study of two treatments in a Veteran's Administration Lung Cancer study of 137 patients. The variables are given below:

time-study time for patient

status- 1 for death 0 for censored

age- age at time of treatment in years

cell type - 1 = squamous, 2=smallcell, 3=adenocarcinoma, 4=largecell

karno=Karnofsky performance score (0=dead,100=good)

diagtime=months from diagnosis to randomization

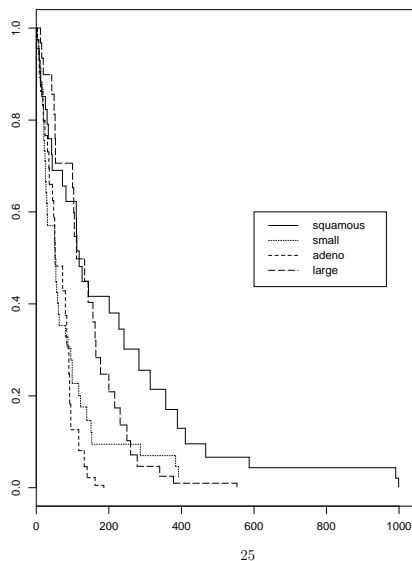
treat- treatment group (0-standard, 1-test)

prior: prior therapy (0=no,1=yes)

The fitted Cox model is given below.

	beta	exp(beta)	se(beta)	z	p
treat	0.295	1.343	0.207	1.419	0.160
age	-0.009	0.991	0.009	-0.936	0.350
Karn	-0.032	0.968	0.005	-5.958	<0.001
diag.time	0.000	1.000	0.009	0.008	0.999
cell2	0.862	2.367	0.275	3.129	0.002
cell3	1.20	3.307	0.300	3.974	<0.001
cell4	0.401	1.494	0.282	1.419	0.016
prior	0.071	1.074	0.232	0.308	0.076

Figure 5: Estimated survival curves in days by cell type in stratified Cox model at mean covariate value



25

Let's now look at a Dutch study of 90 patients with cancer of the larynx. The data give the time from treatment to death or censoring, and patients are classified according the stage of the disease. Stages range from the least severe stage I (cancer is localized), to the most severe stage IV (cancer has spread).

26

```
#### Variables in larynx data are
## stage: stage of cancer at time of first treatment
## time: death time or time on study (months)
## age: age at diagnosis
## year: year of diagnosis
## death: 1 for death, 0 for right-censored
### Let's look at the first 5 lines
```

	stage	time	age	year	death
1	1	0.6	77	76	1
2	1	1.3	53	71	1
3	1	2.4	45	71	1
4	1	2.5	57	78	0
5	1	3.2	58	74	1

27

```
### Cox model treat stage as continuous variable
```

	beta	exp(beta)	se(beta)	z	p
age	0.0228	1.02	0.0144	1.58	0.11000
stage	0.5014	1.65	0.1395	3.59	0.00032

```
### Cox model using dummy variables for stage 2
### stage 3 and stage 4 to measure relative risk
### compared to stage 1
```

	beta	exp(beta)	se(beta)	z	p-value
age	0.0187	1.019	0.0143	1.304	0.190
year	-0.0182	0.982	0.0765	-0.238	0.812
S2	0.1515	1.164	0.4648	0.326	0.743
S3	0.6445	1.905	0.3562	1.809	0.070
S4	1.7324	5.654	0.4359	3.974	<0.001

28

Figure 5 : Estimated survival curves by for 60 year-olds with stage 2 and stage 3 cancer of the larynx

