| Exercise 3: Survival analysis |
|---|

At the end of this exercise you should be able to:

- a. Understand the indications for survival analysis
- b. Be able to do a simple survival analysis in EpiData Analysis

An example of a survival analysis using the Kaplan-Meier method

We observe people over time, starting in 2007 into 2009 and note whether or not they develop an event (whatever it may be) during the observation time as shown for 26 individuals in the following graph:

Year of observation
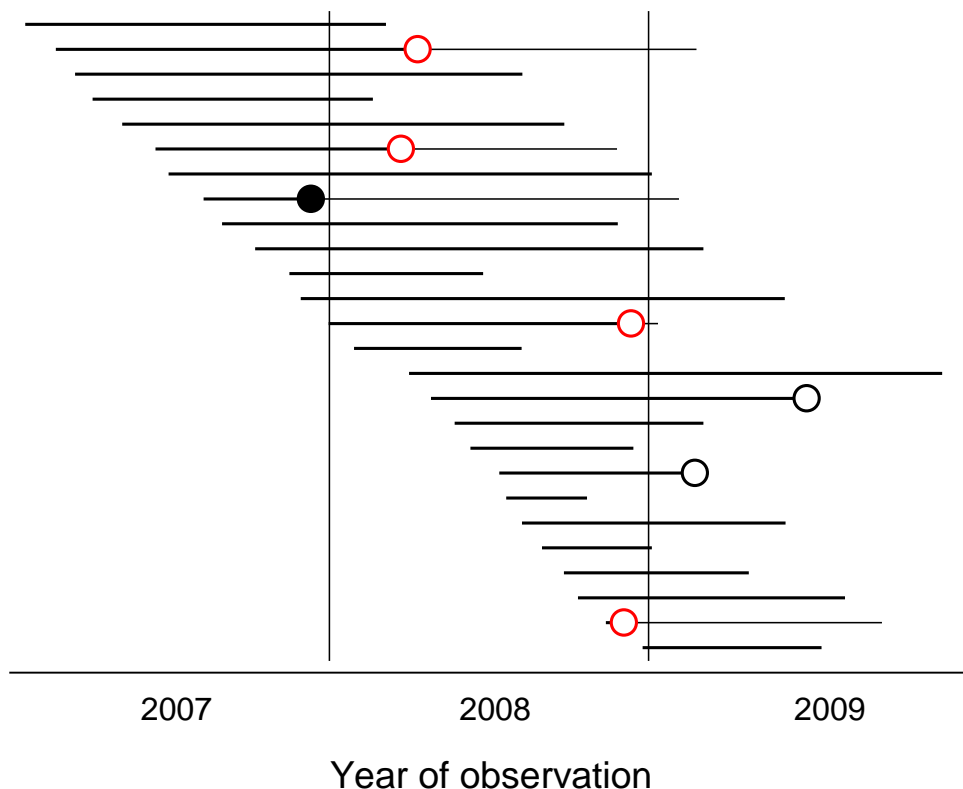
We can think of this setting like of an institution, such as a prison, where inmates enter the prison and are discharged at some time. During incarceration some may develop tuberculosis. If our interest is focused on the year 2008, we have two measures of the magnitude of the problem.

We might calculate the incidence rate in the year 2008. The numerator is 4 cases. The denominator is person-time of observation. As the following table shows, we know the date of entry and exit from the institution, and the date the event occurs among those who had and event:

ID	Entry date	Exit date	Event	Event date	Obs start	Obs end	Obs days
A	20-01-2007	04-03-2008	No		01-01-2008	04-03-2008	63
B	24-02-2007	23-02-2009	Yes	11-04-2008	01-01-2008	11-04-2008	101
C	18-03-2007	07-08-2008	No		01-01-2008	07-08-2008	219

D	07-04-2007	18-02-2008	No		01-01-2008	18-02-2008	48
E	11-05-2007	24-09-2008	No		01-01-2008	24-09-2008	267
F	18-06-2007	24-11-2008	Yes	23-03-2008	--	--	
G	03-07-2007	02-01-2009	No		01-01-2008	31-12-2008	365
H	12-08-2007	03-02-2009	Yes	11-12-2007	01-01-2008	31-12-2008	365
I	02-09-2007	24-11-2008	No		01-01-2008	24-11-2008	328
J	10-10-2007	02-03-2009	No		01-01-2008	31-12-2008	365
K	18-11-2007	23-06-2008	No		01-01-2008	23-06-2008	174
L	01-12-2007	03-06-2009	No		01-01-2008	31-12-2008	365
M	02-01-2008	10-01-2009	Yes	11-12-2008	02-01-2008	11-12-2008	344
N	31-01-2008	06-08-2008	No		31-01-2008	06-08-2008	188
O	03-04-2008	30-11-2009	No		03-04-2008	31-12-2008	272
P	28-04-2008	05-07-2009	Yes	30-06-2009	28-04-2008	31-12-2008	247
Q	25-05-2008	02-03-2009	No		25-05-2008	31-12-2008	220
R	12-06-2008	12-12-2008	No		12-06-2008	12-12-2008	183
S	15-07-2008	02-03-2009	Yes	22-02-2009	15-07-2008	31-12-2008	169
T	23-07-2008	20-10-2008	No		23-07-2008	20-10-2008	89
U	10-08-2008	04-06-2009	No		10-08-2008	31-12-2008	143
V	02-09-2008	02-01-2009	No		02-09-2008	31-12-2008	120
W	27-09-2008	23-04-2009	No		27-09-2008	31-12-2008	95
X	13-10-2008	11-08-2009	No		13-10-2008	31-12-2008	79
Y	14-11-2008	23-09-2009	Yes	03-12-2008	14-11-2008	03-12-2008	19
Z	26-12-2008	15-07-2009	No		26-12-2008	31-12-2008	5

Days	4833
Cases	4
Cases/1000 person-days	0.828
Cases/100 person-years	30.2

Each person contributes person-time of observation, starting earliest from the beginning of the year 2008 or later if entry into the system was later. Person-time is contributed until exit from the institution or up to the point of the event if either happened in the year 2008. If the event happens later, observation time ends at the right-censoring point of 31 December 2008. Individual F developed the event in 2007 and although still in the system in 2008 does not contribute any person-time of observation in 2008. Individuals P and S developed the event only in 2009 and are thus not counted in 2008 and they contribute to person-time of observation through the end of the year 2008. Thus summed up, the 25 of the 26 inmates contributed 4,833 days of observation time which gives a case rate of 0.8 per 1,000 person-days of observation, or annualized, 30.2 cases per 100 observation years.

Survival analysis

The second way to look at the size of the problem is to ask what the probability for an individual is to "survive" the year 2008 without developing the event.

We exclude the person who had the event already in 2007 and sort the persons by the time of event or censoring (discharge or latest end of 2008). We note the number at risk of the

beginning of the interval, then the number who got censored, then how many were left after censoring, and finally how many get an event during the interval:

At risk at beginning	Censored	At risk at end	Event at end	Proportion surviving	Survival
25	1	24	0		
24	0	24	1	23/24	0.958
23	1	22	0		0.958
22	1	21	0		0.958
21	1	20	0		0.958
20	0	20	1	(19/20)*0.958	0.910
19	1	18	0		0.910
18	1	17	0		0.910
17	0	17	1	(16/17)*0.910	0.857
16	1	15	0		0.857
15	1	14	0		0.857
14	1	13	0		0.857
13	1	12	0		0.857
12	1	11	0		0.857
11	1	10	0		0.857
10	1	9	0		0.857
9	1	8	0		0.857
8	1	7	0		0.857
7	1	6	0		0.857
6	1	5	0		0.857
5	1	4	0		0.857
4	0	4	1	(3/4)*0.857	0.643
3	3	0	0		0.643

At each point where an event (not censoring) happens, we calculate the survival probability by dividing the number "surviving" after the event by the number at risk at the beginning of the interval when the event occurred.

You may note that censoring during an interval is assumed not to affect survival probability during that interval, but censored individuals are also subtracted from those at risk for the next interval. [1-3] This assumption is a simplification because censoring may, under certain circumstances, indeed be affecting the survival probability during the remaining interval. In any case, however, taking both the occurrence of the event and censoring into account for each interval following an event is a much more appropriate way to calculate survival than the proportion with an event among all who entered the cohort or by removing those censored from the cohort.

Graphically, we summarize the Kaplan-Meier survival probability as a step graph, sometimes showing the points when an individual was censored:

Example Kaplan-Meier Survival Plot
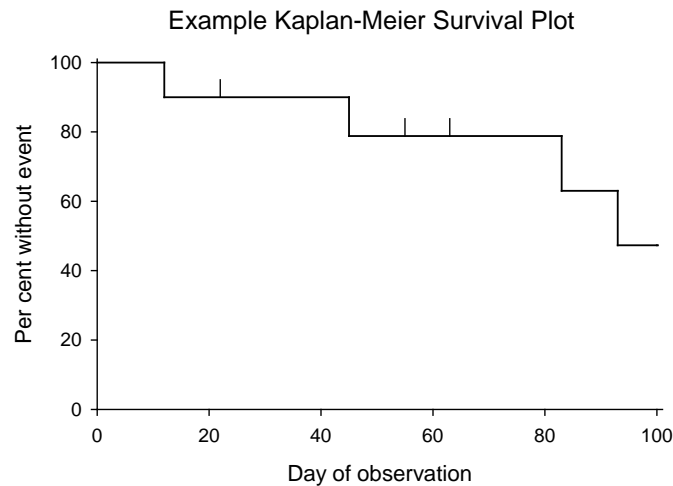
1. Kaplan E L, Meier P. Nonparametric estimation from incomplete observations. J Am Stat Ass 1958;53:457-81.

2. Bland J M, Altman D G. Survival probabilities (the Kaplan-Meier method). (Statistics notes). BMJ 1998;317:1572.

3. Fink S A, Brown R S, Jr. Survival analysis. Gastroenterol Hepatol 2006;2:380-3.

Survival analysis in EpiData Analysis

The command for a Kaplan-Meier survival analysis in EpiData Analysis is:

```
lifetable outcome interval
```

or

```
lifetable outcome startdate enddate
```

There are a multitude of options available both lifetable-specific options and general graph options. Use the Help file to test out various options until you have the survival plot that suits your needs.
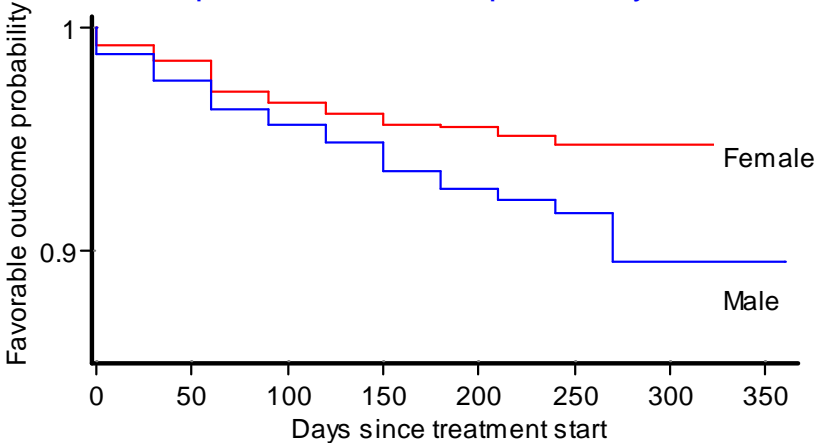
Tasks

The purpose of the exercise is to demonstrate quantitatively the probability of remaining without an event.

We use to this end a real dataset from a tuberculosis program, although we have removed any identifier and most variables, and retained to simplify your work only records of patients who have an exact date of treatment start and an exact date of treatment end. The dataset is provided as a supplementary file d_ex03_required.rec.

1) Define a binomial outcome, where "favorable" is cured or treatment completed and all other outcomes are "unfavorable"

2) Do the lifetable analysis only for new sputum smear-positive cases and show the survival probability stratified by sex

Kaplan-Meier survival probability

EpiData course: Exercise D_EX03