Survival Analysis

Session 6: The Kaplan-Meier Estimator

Jonas Schöley

@jschoeley
0000-0002-3340-8518
j.schoeley@uni-rostock.de

Recap: Survival Identities

In survival analysis we consider the random variable **"*X*: Time until event"**
x = 0.1 weeks, 2.3 weeks…

We express our knowledge about the distribution of *X* in any of these functions. Knowing any single function we can derive all other via the **Survival Identities**.

h(*x*): Hazard function
The instantaneous rate of new events at time *x* among those who did not experience the event yet.

$$h(x) = \lim_{h\to 0} P(x \le X < x+h \mid X \ge x)/h$$

f(*x*): Density function
The relative likelihood of experiencing the event around time *x*.

$F(x)={}_0\int^x f(x)\,dx$

$S(x)={}_x\int^{Inf} f(x)\,dx$

$h(x)=-S'(x)/S(x)$

$H(x) = {}_0\int^x h(x)\,dx$

F(*x*): Distribution function
aka Cumulative function
The probability of experiencing the event until time *x*.
F(*x*) = P(*X*≤*x*)

$S(x)=1-F(x)$

S(*x*): Survival function
The probability of *not* experiencing the event until time *x*.
S(*x*) = P(*X*>*x*)

$H(x) = -\log S(x)$

H(*x*): Cumulative Hazard
The integral of h(x).

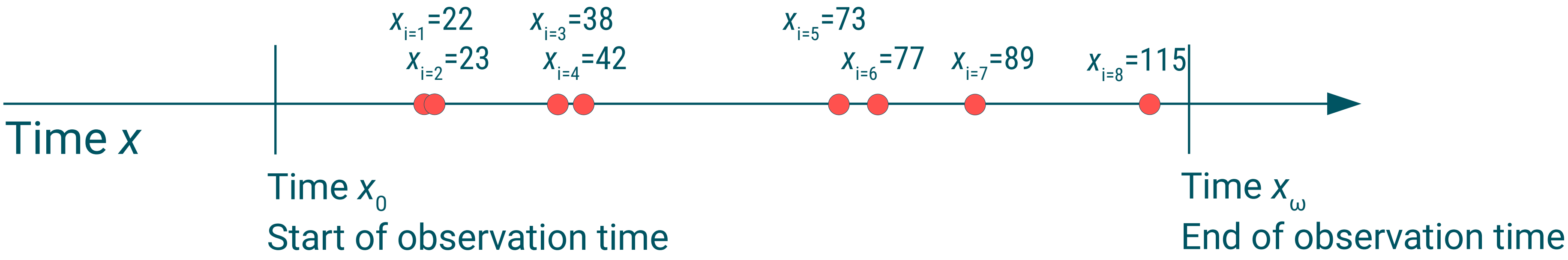
$S(x) = \exp(-H(x))$

What Does Survival Data Look Like?

In survival analysis we consider the random variable **"*X*: Time until event"**
x = 0.1 month, 2.3 months…

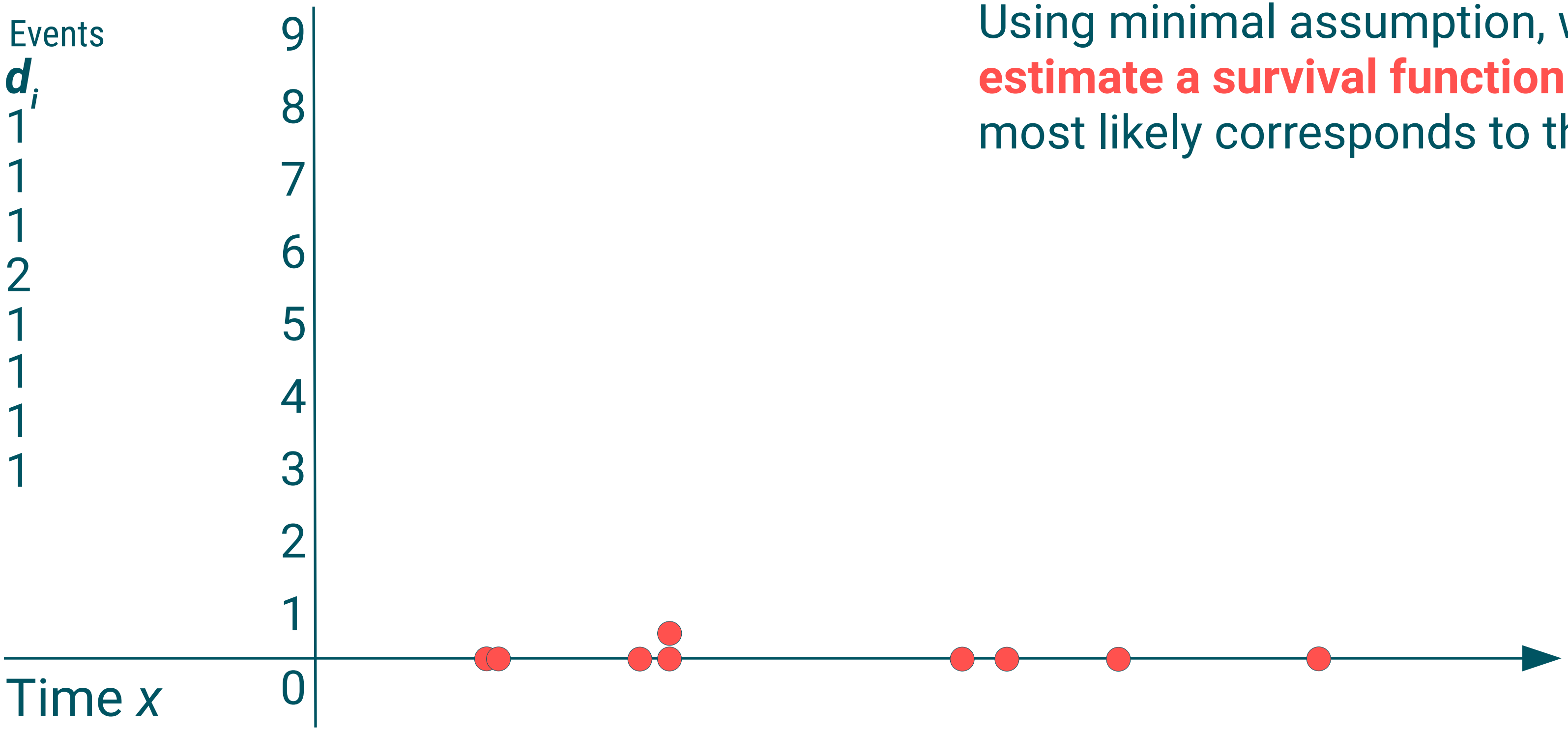
Who? 8 breast cancer patients
Start of observation? Cancer diagnosis
Event of interest? Death
Time unit? Months
End of observation? 10 years follow up

Observation index i	Observed event time x
1	22
2	23
3	38
4	42
5	73
6	77
7	89
8	115



Deriving the Kaplan-Meier Estimator

Index	Time	Events
i	x_i	d_i
1	22	1
2	23	1
3	38	1
4	42	2
5	73	1
6	77	1
7	89	1
8	115	1



Using minimal assumption, we want to **estimate a survival function S(x)** which most likely corresponds to the data .

Time x_i

Number of events d_i

Number of survivors n_i

Deriving the Kaplan-Meier Estimator

Index	Time	Events
i	x_i	d_i
1	22	1
2	23	1
3	38	1
4	42	2
5	73	1
6	77	1
7	89	1
8	115	1

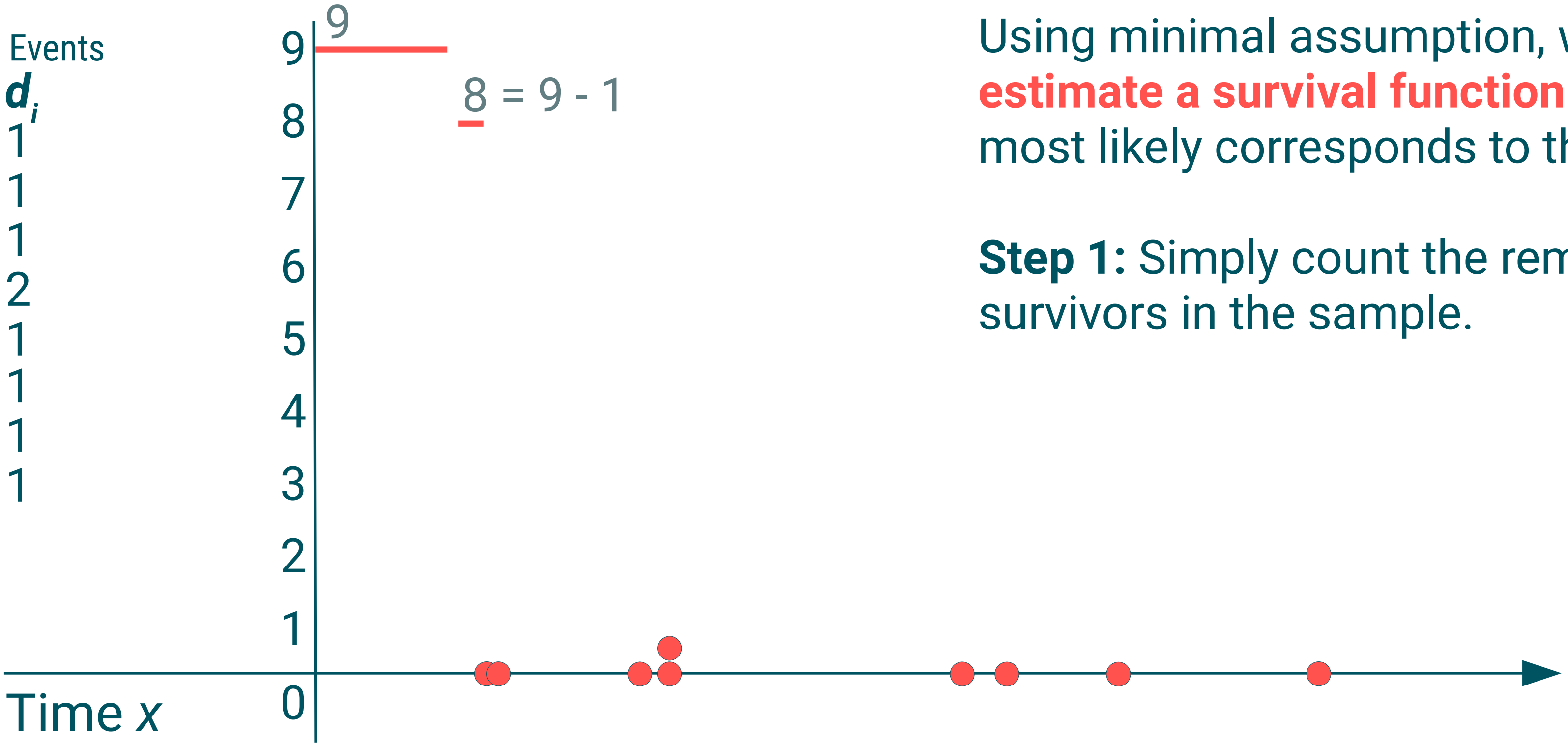
Time	x_i	0
Number of events	d_i	0
Number of survivors	n_i	9

Using minimal assumption, we want to **estimate a survival function S(x)** which most likely corresponds to the data ● .

Step 1: Simply count the remaining survivors in the sample.

Deriving the Kaplan-Meier Estimator

Index i	Time x_i	Events d_i
1	22	1
2	23	1
3	38	1
4	42	2
5	73	1
6	77	1
7	89	1
8	115	1



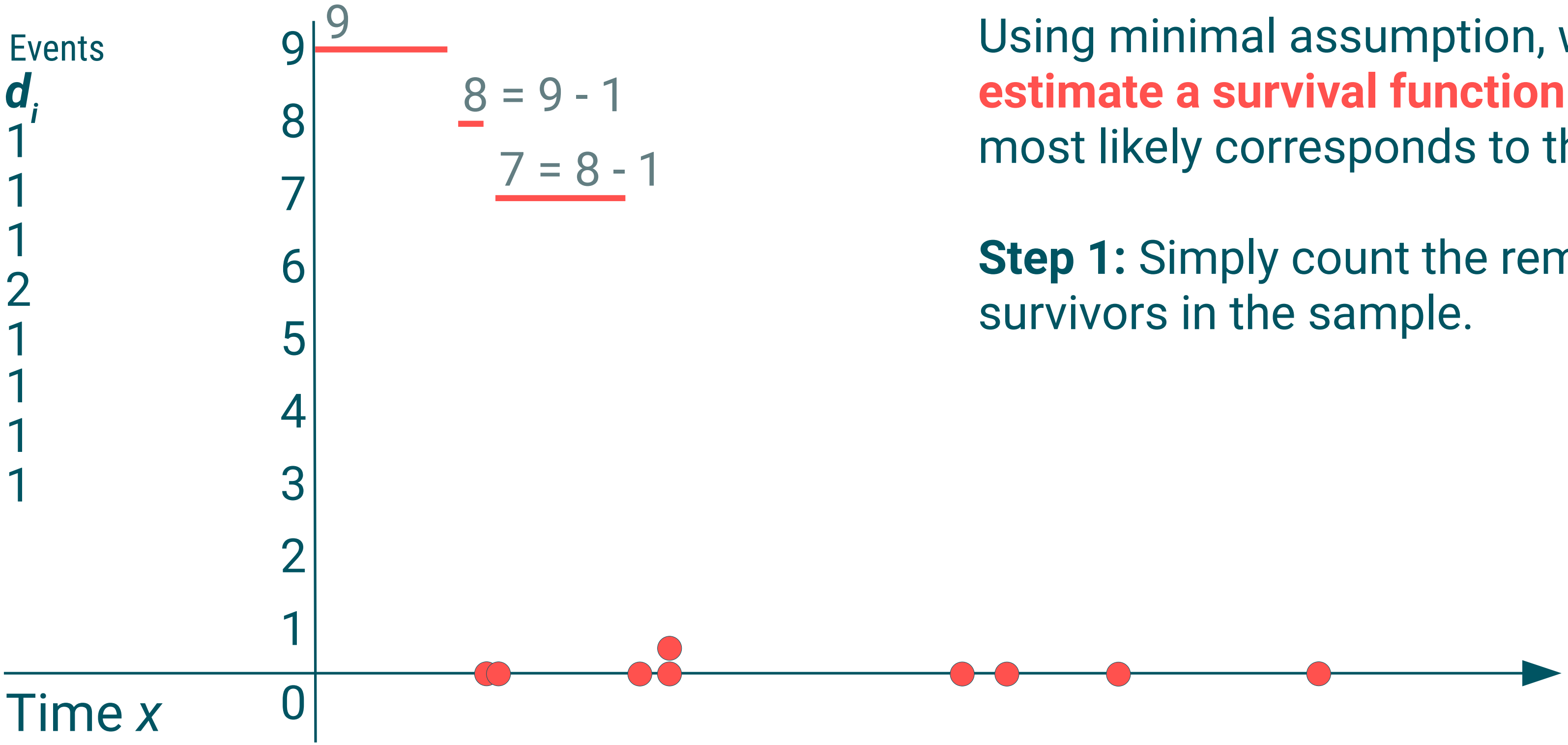
Time	x_i	0	22
Number of events	d_i	0	1
Number of survivors	n_i	9	8

Using minimal assumption, we want to **estimate a survival function S(x)** which most likely corresponds to the data ● .

Step 1: Simply count the remaining survivors in the sample.

Deriving the Kaplan-Meier Estimator

Index	Time	Events
i	x_i	d_i
1	22	1
2	23	1
3	38	1
4	42	2
5	73	1
6	77	1
7	89	1
8	115	1



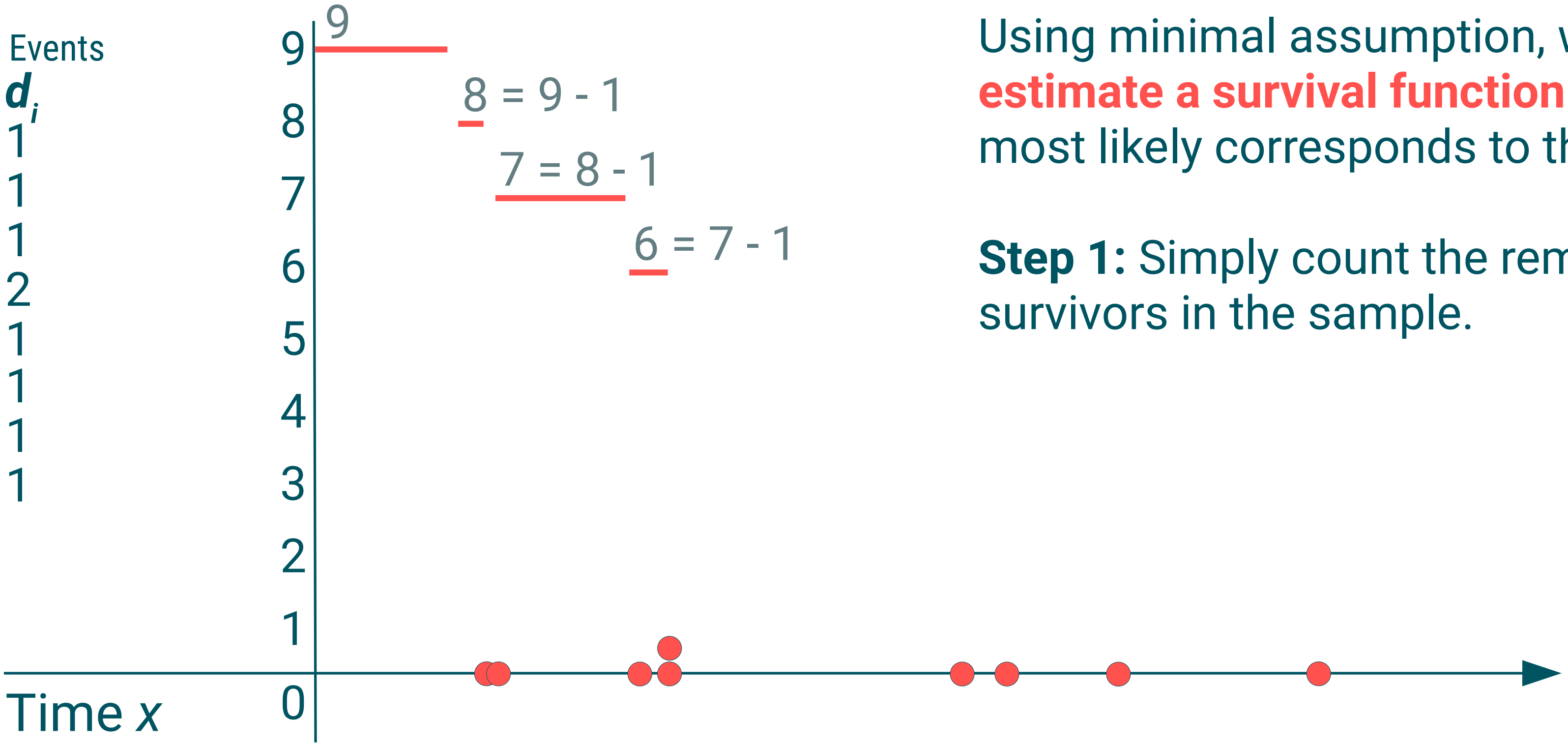
Time	x_i	0	22	23
Number of events	d_i	0	1	1
Number of survivors	n_i	9	8	7

Using minimal assumption, we want to **estimate a survival function S(x)** which most likely corresponds to the data ● .

Step 1: Simply count the remaining survivors in the sample.

Deriving the Kaplan-Meier Estimator

Index	Time	Events
i	x_i	d_i
1	22	1
2	23	1
3	38	1
4	42	2
5	73	1
6	77	1
7	89	1
8	115	1



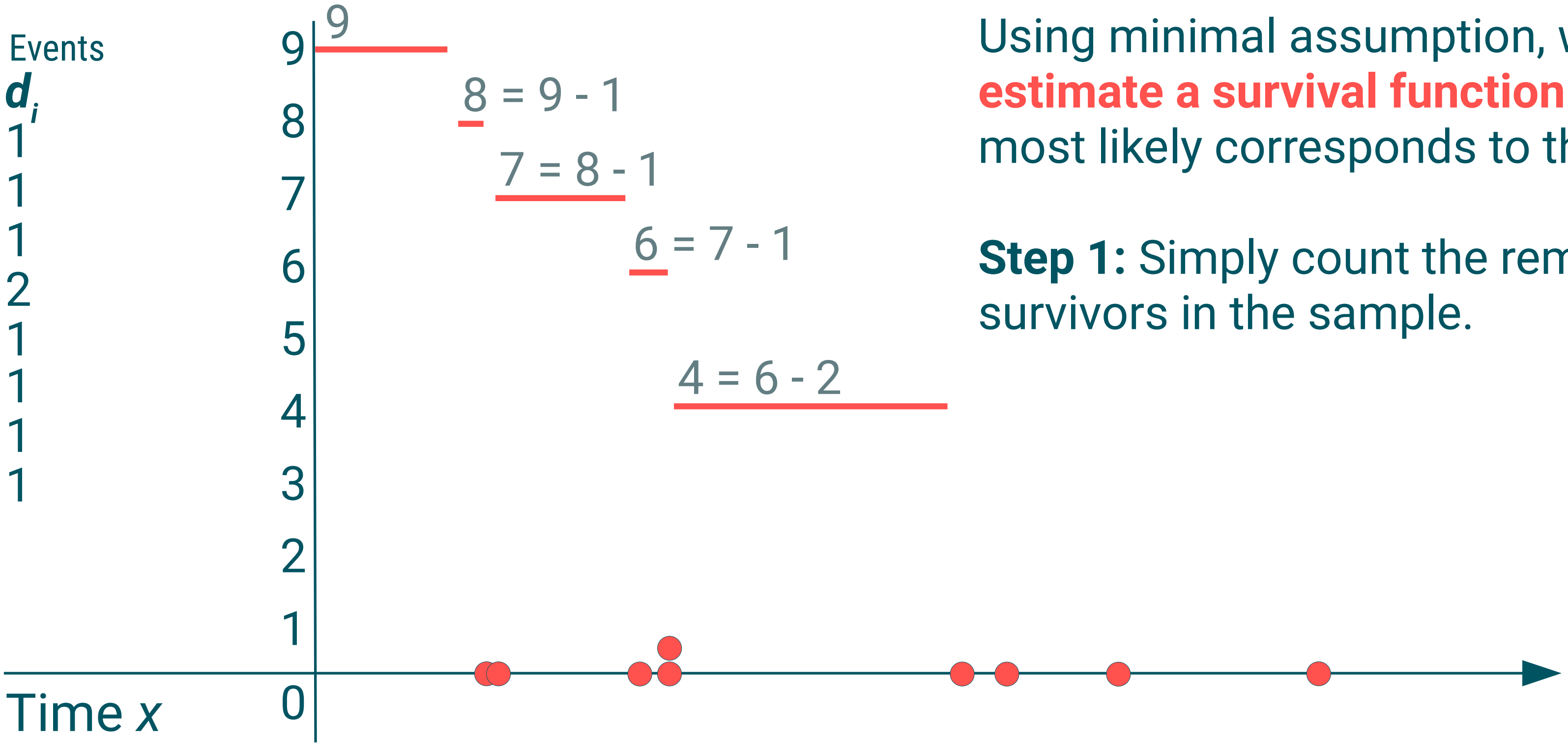
Time	x_i	0	22	23	38
Number of events	d_i	0	1	1	1
Number of survivors	n_i	9	8	7	6

Using minimal assumption, we want to **estimate a survival function S(x)** which most likely corresponds to the data ● .

Step 1: Simply count the remaining survivors in the sample.

Deriving the Kaplan-Meier Estimator

Index i	Time x_i	Events d_i
1	22	1
2	23	1
3	38	1
4	42	2
5	73	1
6	77	1
7	89	1
8	115	1



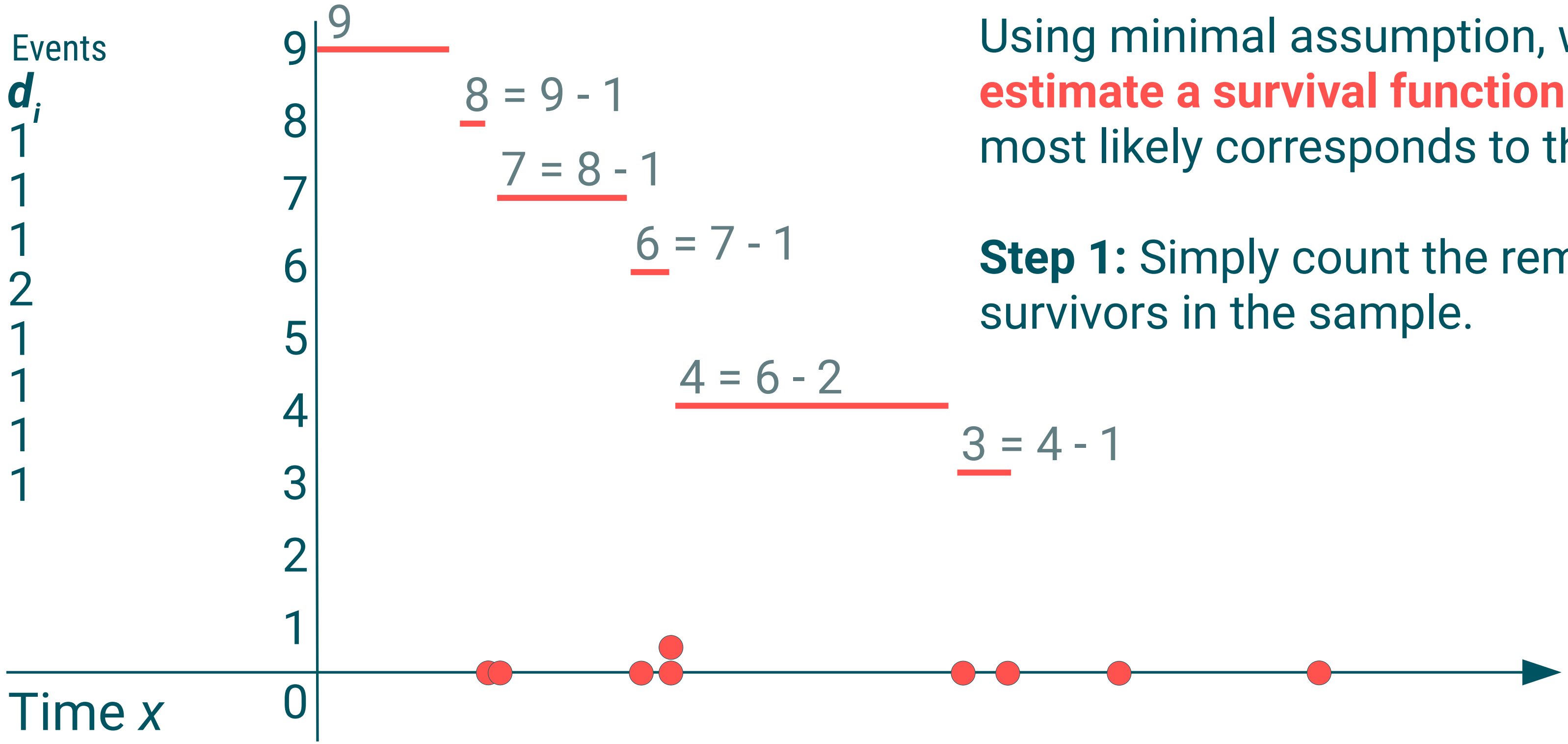
Time	x_i	0	22	23	38	42
Number of events	d_i	0	1	1	1	2
Number of survivors	n_i	9	8	7	6	4

Using minimal assumption, we want to **estimate a survival function S(x)** which most likely corresponds to the data ● .

Step 1: Simply count the remaining survivors in the sample.

Deriving the Kaplan-Meier Estimator

Index	Time	Events
i	x_i	d_i
1	22	1
2	23	1
3	38	1
4	42	2
5	73	1
6	77	1
7	89	1
8	115	1



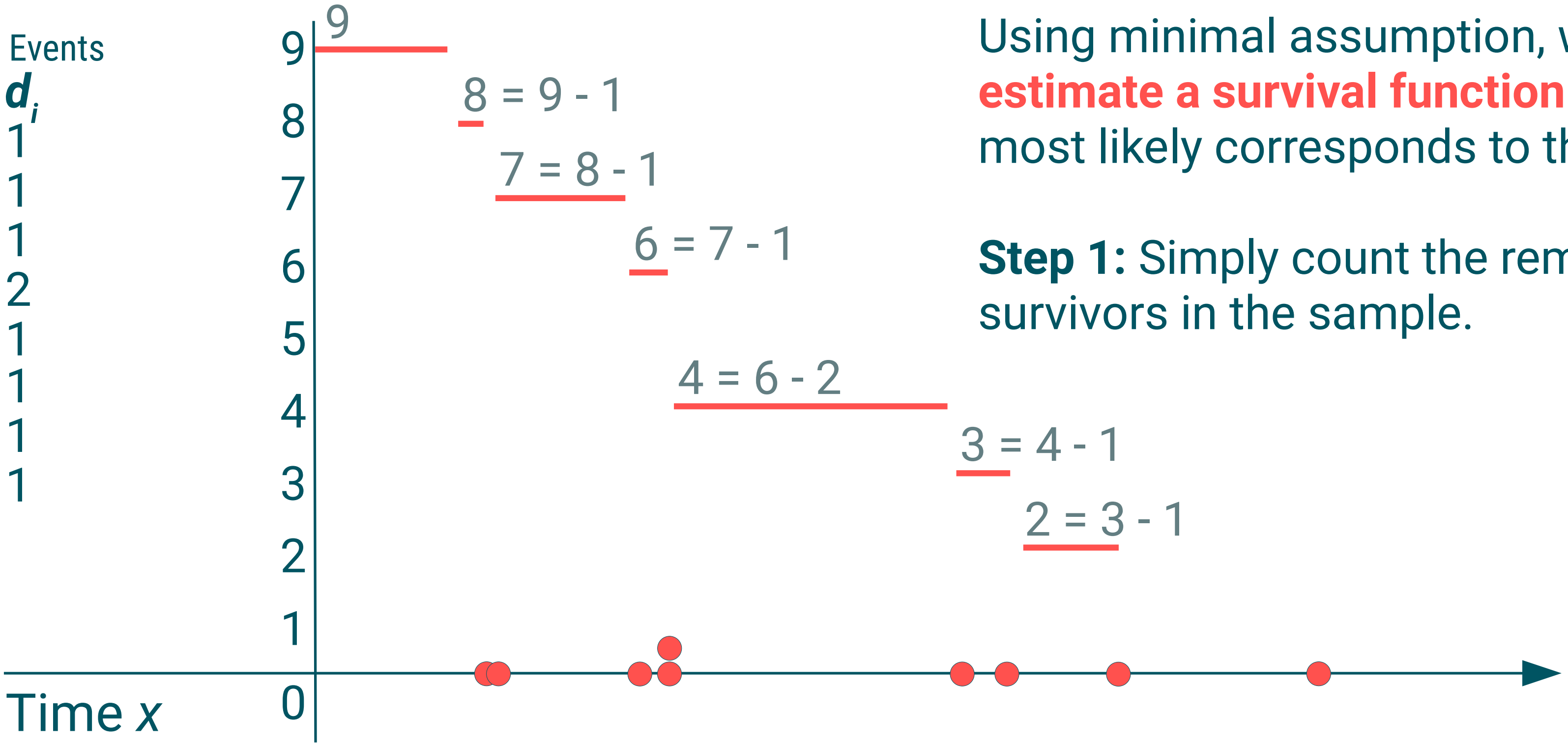
Time	x_i	0	22	23	38	42	73
Number of events	d_i	0	1	1	1	2	1
Number of survivors	n_i	9	8	7	6	4	3

Using minimal assumption, we want to **estimate a survival function S(x)** which most likely corresponds to the data ● .

Step 1: Simply count the remaining survivors in the sample.

Deriving the Kaplan-Meier Estimator

Index i	Time x_i	Events d_i
1	22	1
2	23	1
3	38	1
4	42	2
5	73	1
6	77	1
7	89	1
8	115	1



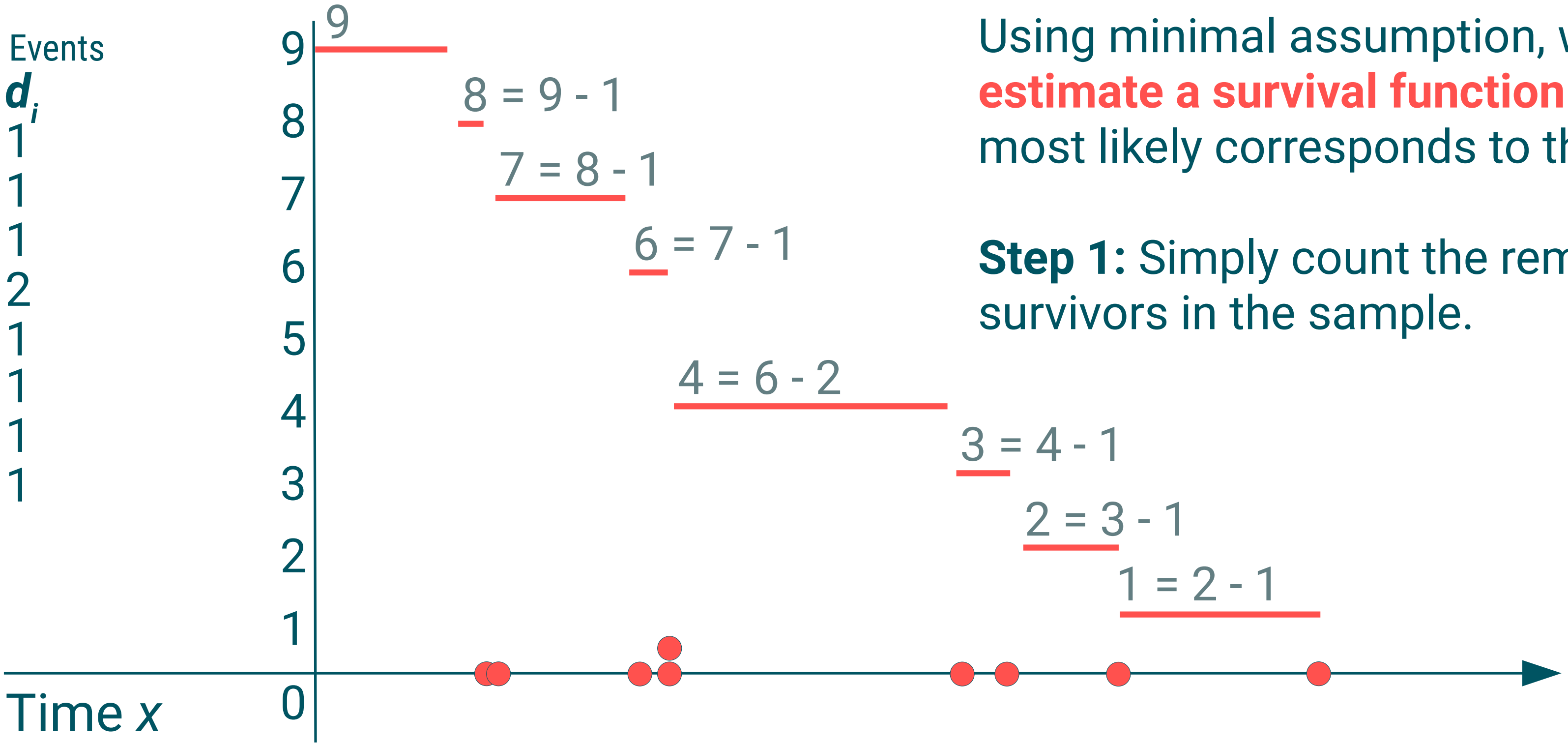
Time	x_i	0	22	23	38	42	73	77
Number of events	d_i	0	1	1	1	2	1	1
Number of survivors	n_i	9	8	7	6	4	3	2

Using minimal assumption, we want to **estimate a survival function S(x)** which most likely corresponds to the data ● .

Step 1: Simply count the remaining survivors in the sample.

Deriving the Kaplan-Meier Estimator

Index i	Time x_i	Events d_i
1	22	1
2	23	1
3	38	1
4	42	2
5	73	1
6	77	1
7	89	1
8	115	1



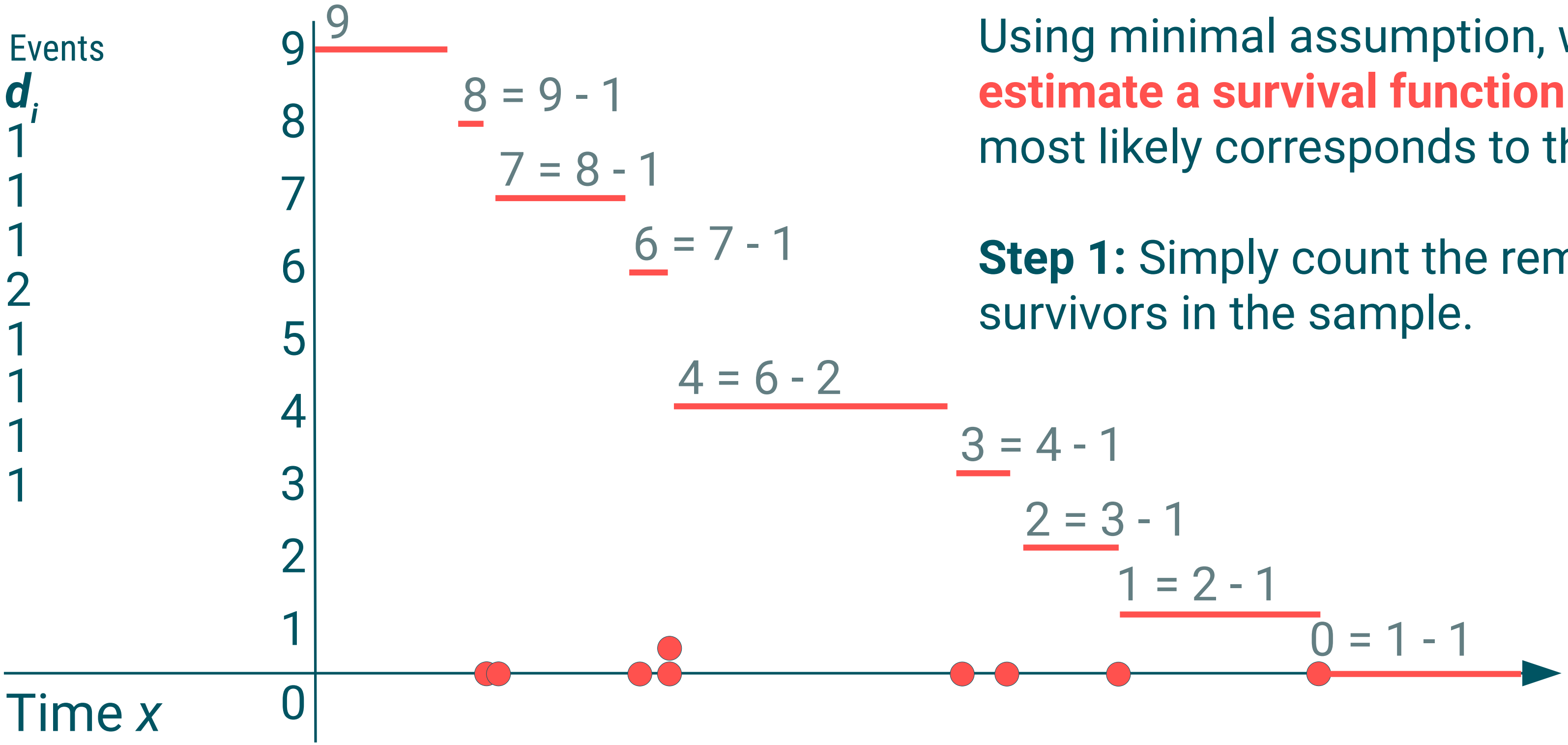
Time	x_i	0	22	23	38	42	73	77	89
Number of events	d_i	0	1	1	1	2	1	1	1
Number of survivors	n_i	9	8	7	6	4	3	2	1

Using minimal assumption, we want to **estimate a survival function S(x)** which most likely corresponds to the data ● .

Step 1: Simply count the remaining survivors in the sample.

Deriving the Kaplan-Meier Estimator

Index i	Time x_i	Events d_i
1	22	1
2	23	1
3	38	1
4	42	2
5	73	1
6	77	1
7	89	1
8	115	1



Time	x_i	0	22	23	38	42	73	77	89	115
Number of events	d_i	0	1	1	1	2	1	1	1	1
Number of survivors	n_i	9	8	7	6	4	3	2	1	0

Using minimal assumption, we want to **estimate a survival function S(x)** which most likely corresponds to the data ● .

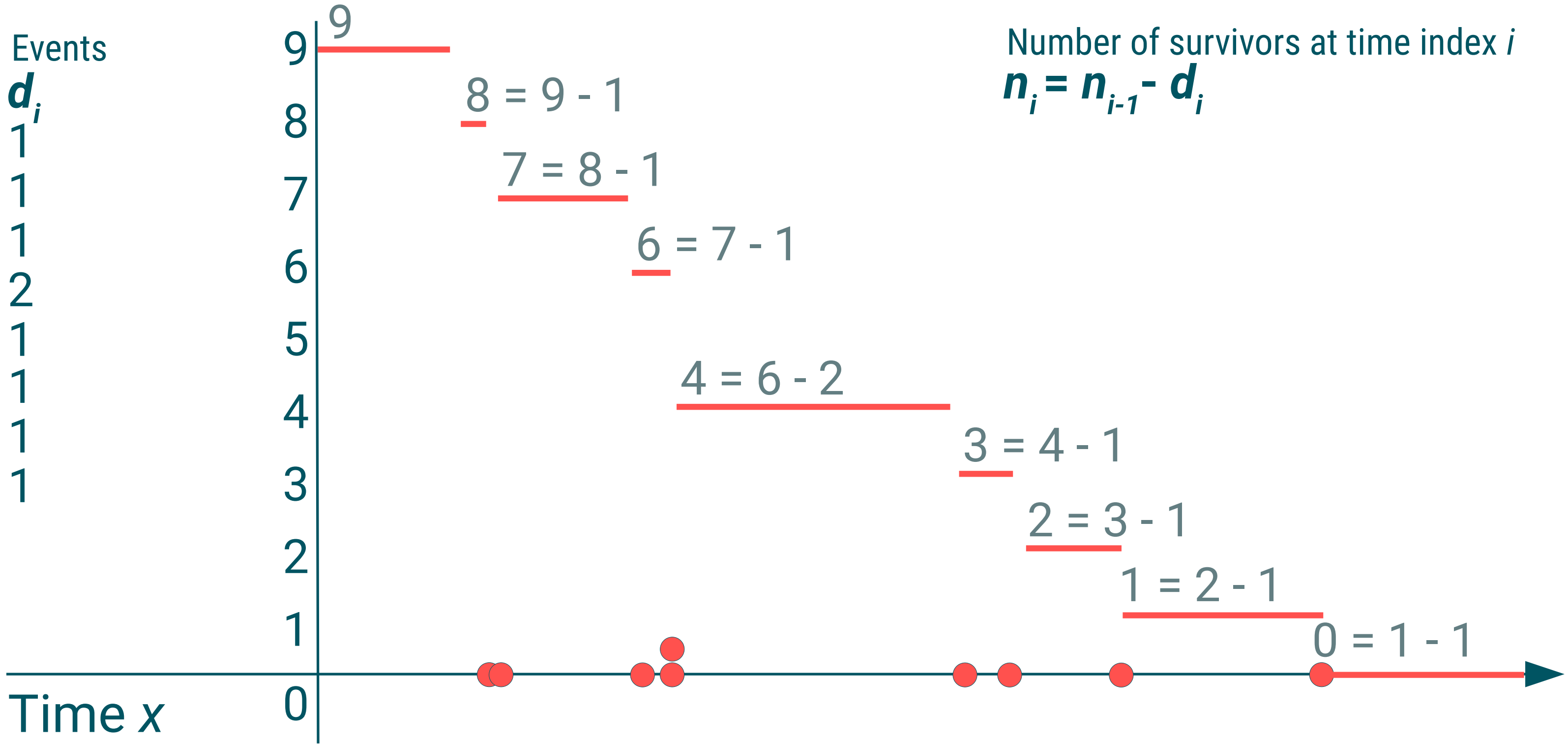
Step 1: Simply count the remaining survivors in the sample.

Deriving the Kaplan-Meier Estimator

Index	Time	Events
i	x_i	d_i
1	22	1
2	23	1
3	38	1
4	42	2
5	73	1
6	77	1
7	89	1
8	115	1

Number of survivors at time index *i*

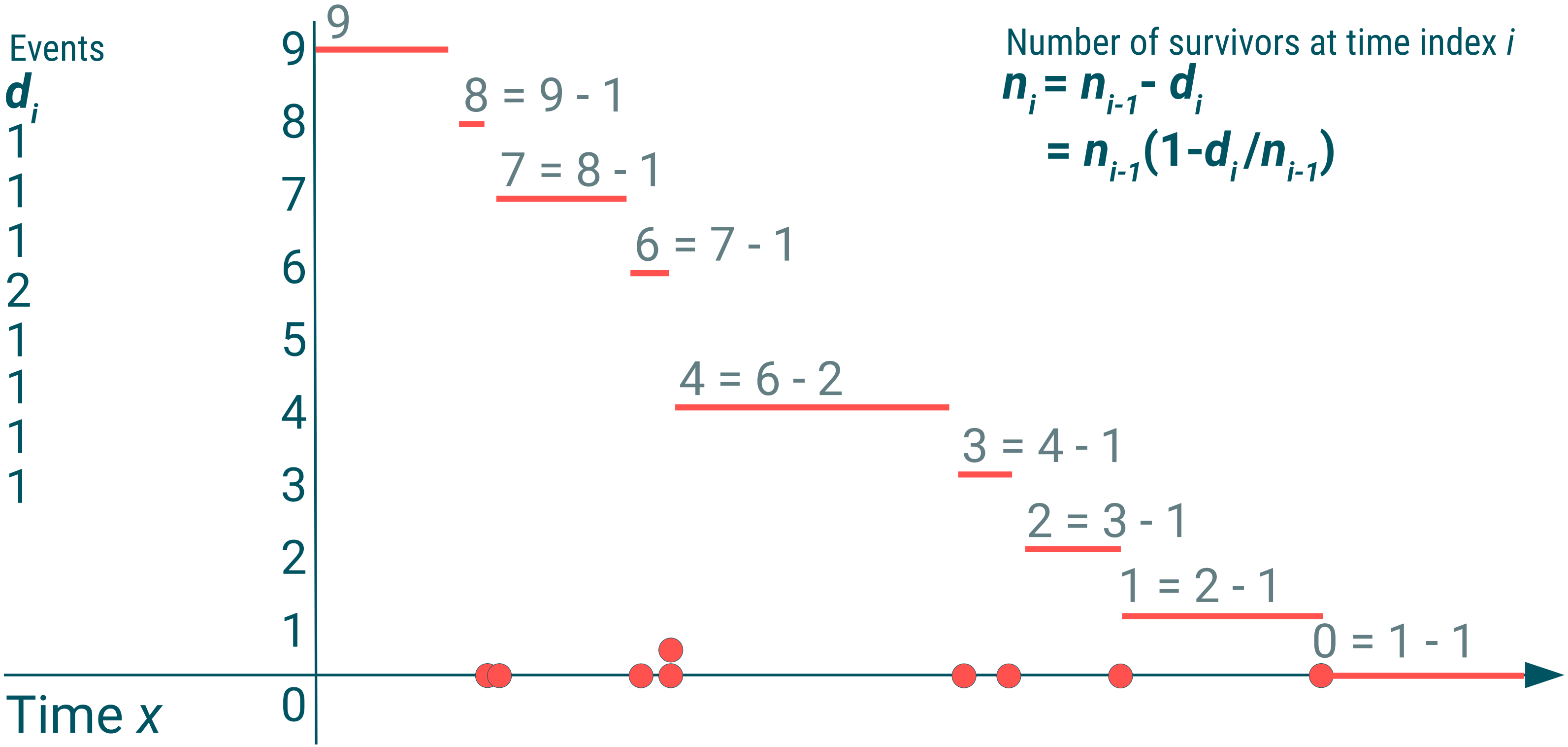
$$n_i = n_{i-1} - d_i$$



Time	x_i	0	22	23	38	42	73	77	89	115
Number of events	d_i	0	1	1	1	2	1	1	1	1
Number of survivors	n_i	9	8	7	6	4	3	2	1	0

Deriving the Kaplan-Meier Estimator

Index i	Time x_i	Events d_i
1	22	1
2	23	1
3	38	1
4	42	2
5	73	1
6	77	1
7	89	1
8	115	1



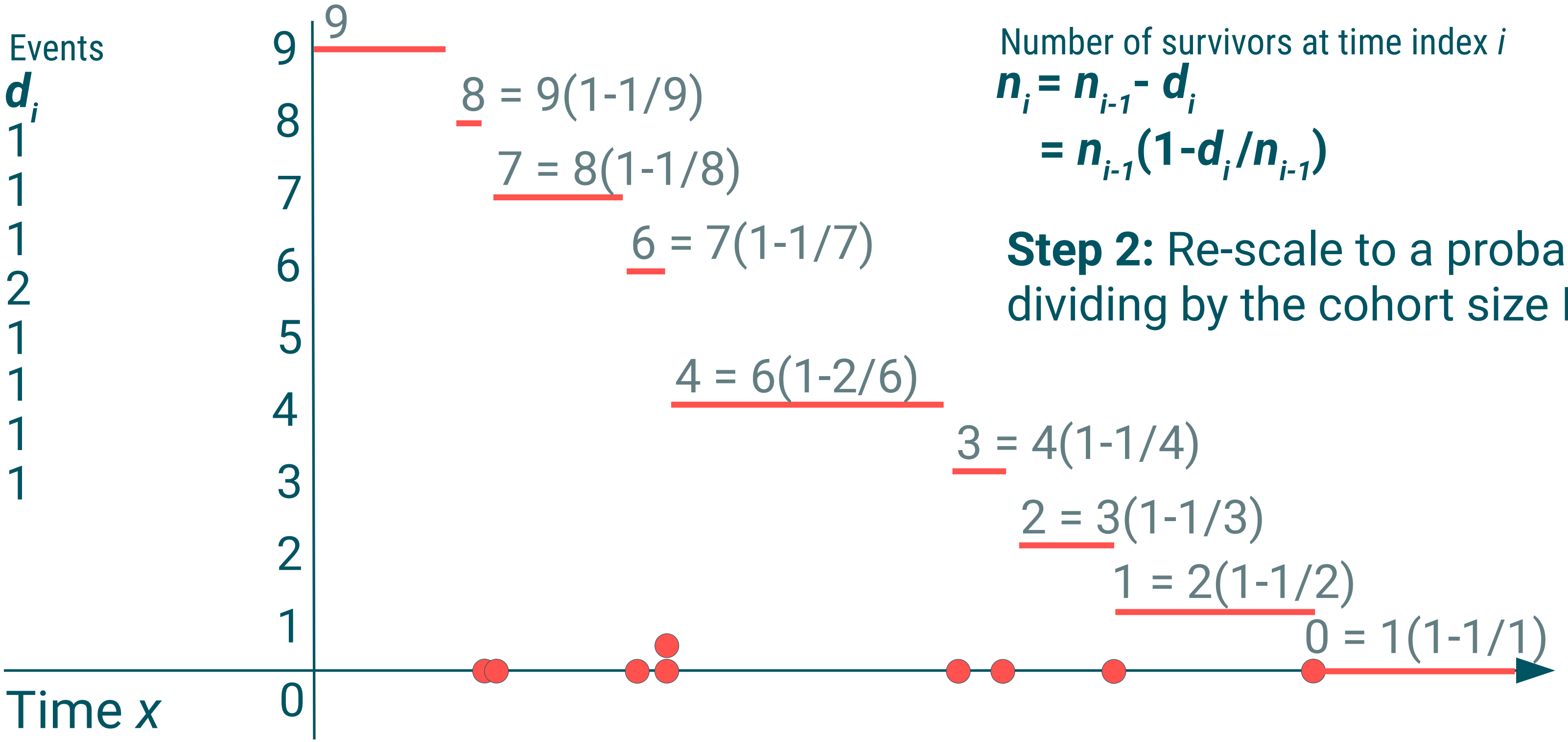
Number of survivors at time index i

$$n_i = n_{i-1} - d_i$$
$$= n_{i-1}(1 - d_i / n_{i-1})$$

Time	x_i	0	22	23	38	42	73	77	89	115
Number of events	d_i	0	1	1	1	2	1	1	1	1
Number of survivors	n_i	9	8	7	6	4	3	2	1	0

Deriving the Kaplan-Meier Estimator

Index i	Time x_i	Events d_i
1	22	1
2	23	1
3	38	1
4	42	2
5	73	1
6	77	1
7	89	1
8	115	1



Number of survivors at time index i

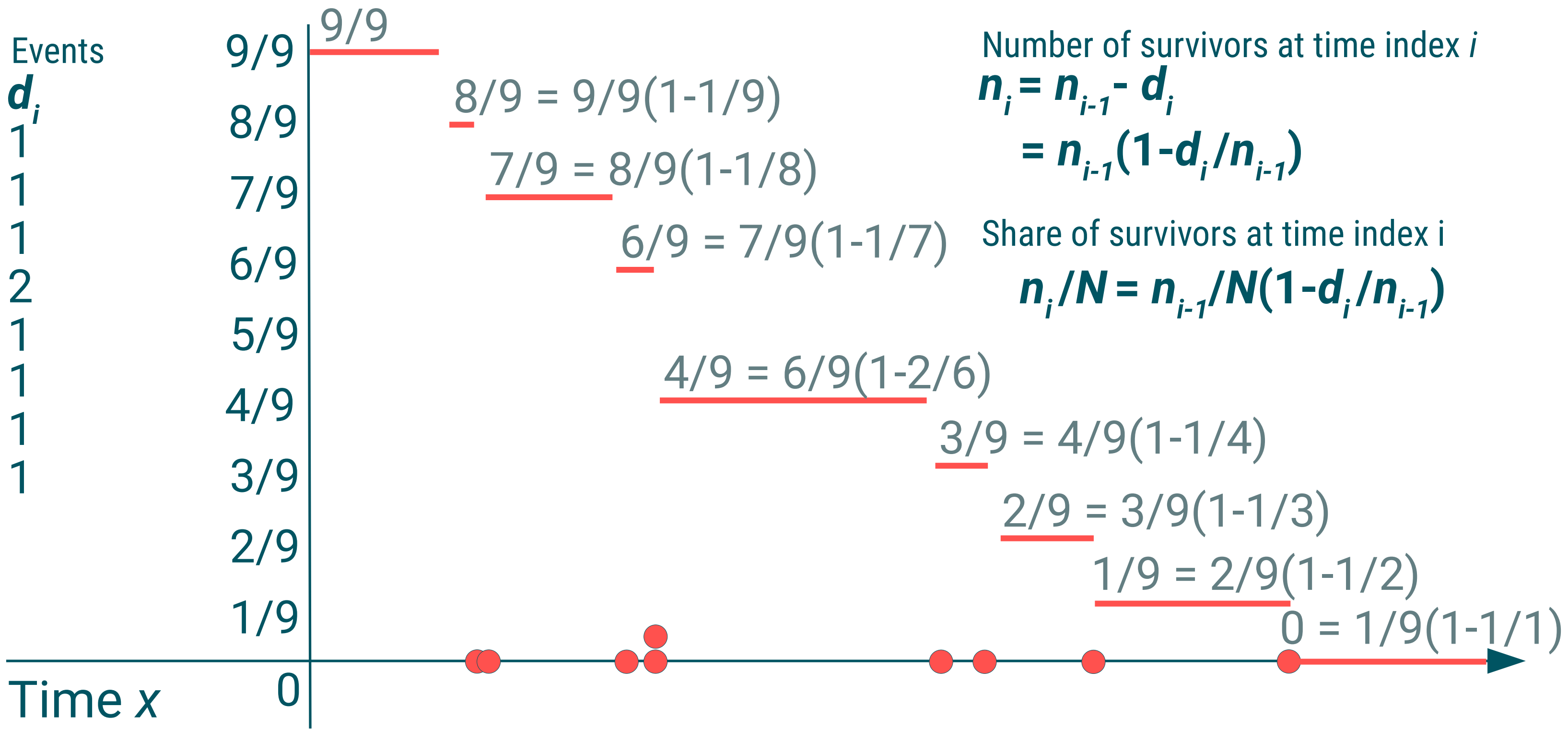
$$n_i = n_{i-1} - d_i$$
$$= n_{i-1}(1 - d_i/n_{i-1})$$

Step 2: Re-scale to a probability by dividing by the cohort size N.

Time	x_i	0	22	23	38	42	73	77	89	115
Number of events	d_i	0	1	1	1	2	1	1	1	1
Number of survivors	n_i	9	8	7	6	4	3	2	1	0

Deriving the Kaplan-Meier Estimator

Index i	Time x_i	Events d_i
1	22	1
2	23	1
3	38	1
4	42	2
5	73	1
6	77	1
7	89	1
8	115	1



Time	x_i	0	22	23	38	42	73	77	89	115
Number of events	d_i	0	1	1	1	2	1	1	1	1
Number of survivors	n_i	9	8	7	6	4	3	2	1	0

Number of survivors at time index *i*

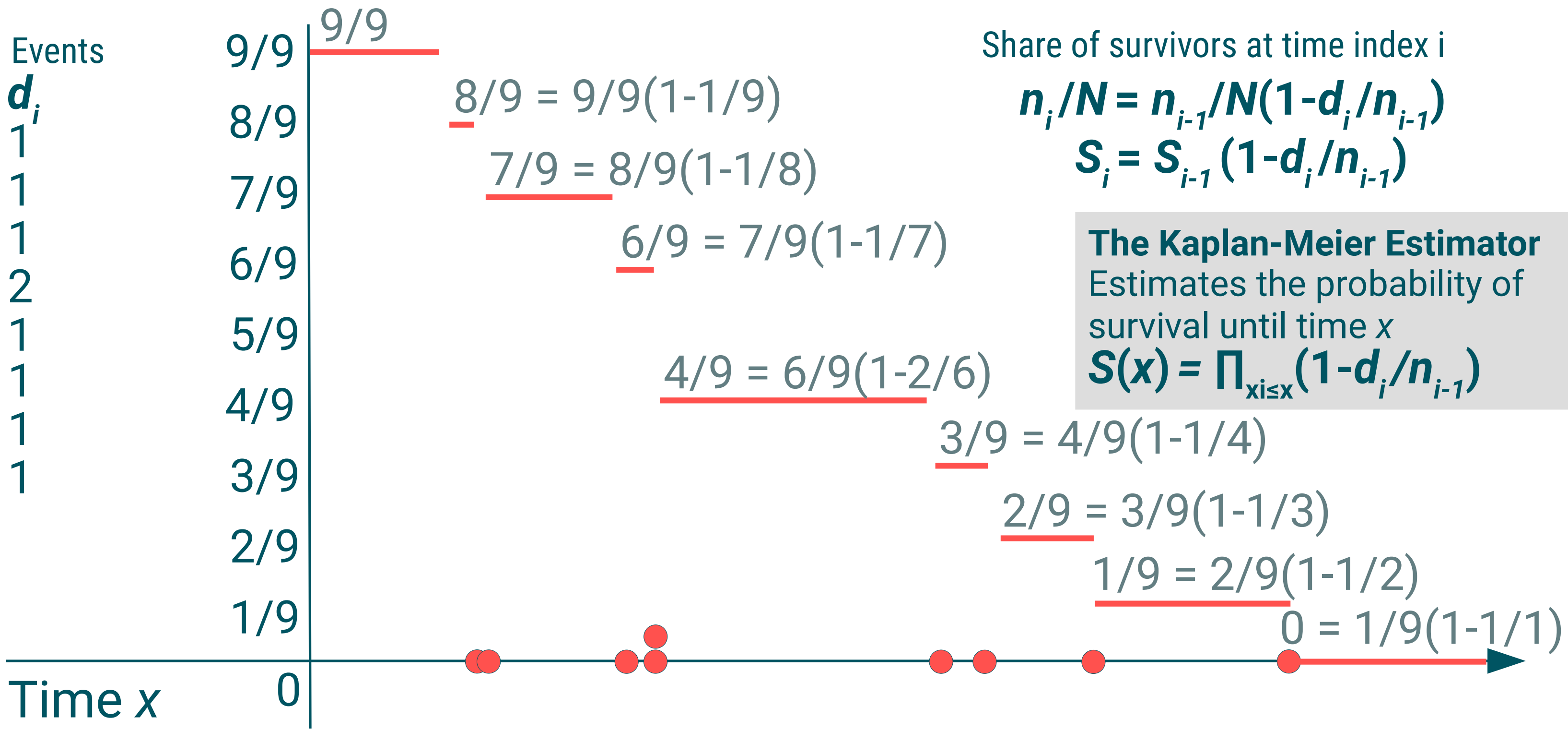
$$n_i = n_{i-1} - d_i$$
$$= n_{i-1}(1 - d_i/n_{i-1})$$

Share of survivors at time index i

$$n_i/N = n_{i-1}/N(1 - d_i/n_{i-1})$$

Deriving the Kaplan-Meier Estimator

Index i	Time x_i	Events d_i
1	22	1
2	23	1
3	38	1
4	42	2
5	73	1
6	77	1
7	89	1
8	115	1



Time	x_i	0	22	23	38	42	73	77	89	115
Number of events	d_i	0	1	1	1	2	1	1	1	1
Number of survivors	n_i	9	8	7	6	4	3	2	1	0

Share of survivors at time index i

$$n_i/N = n_{i-1}/N(1-d_i/n_{i-1})$$

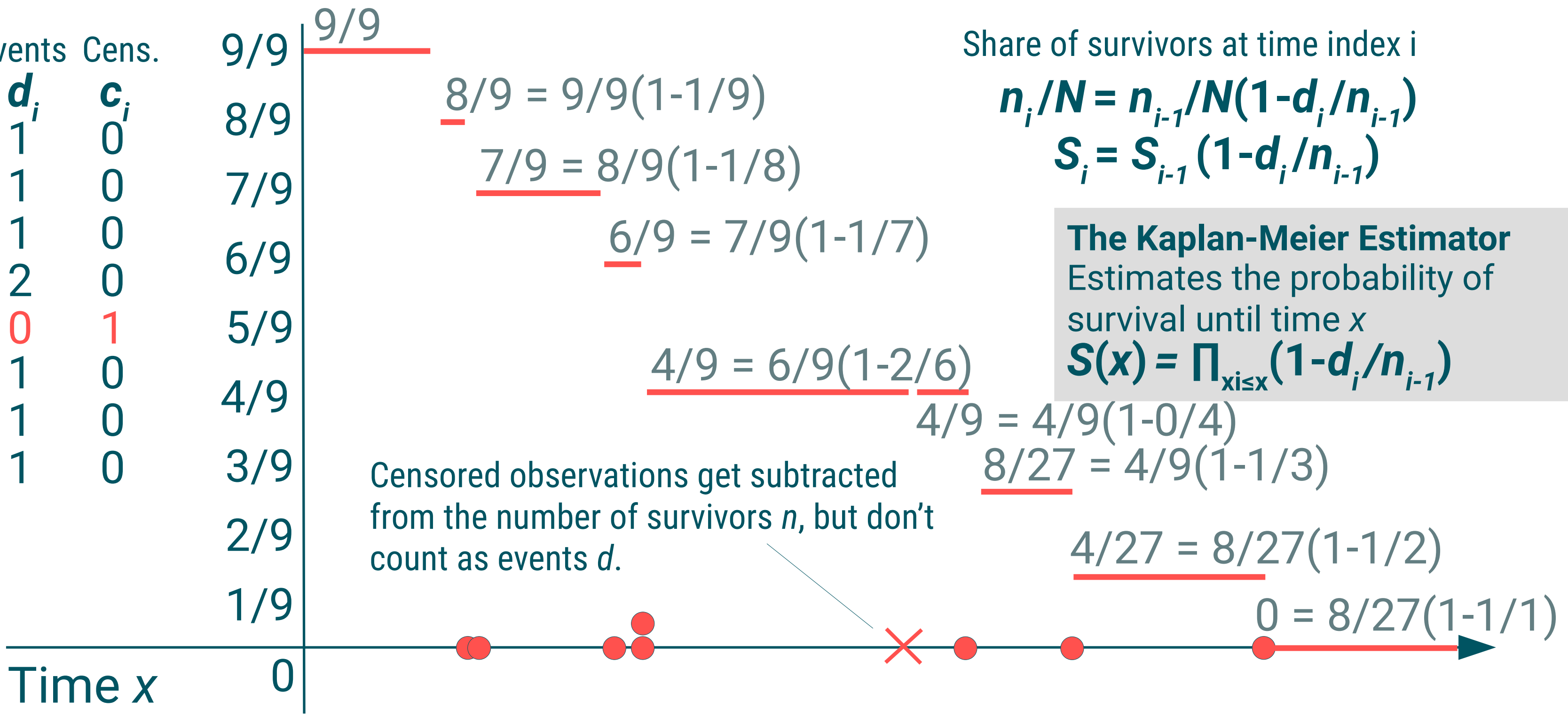
$$S_i = S_{i-1}(1-d_i/n_{i-1})$$

The Kaplan-Meier Estimator
Estimates the probability of survival until time x

$$S(x) = \prod_{x_i \leq x}(1-d_i/n_{i-1})$$

Deriving the Kaplan-Meier Estimator

Index i	Time x_i	Events d_i	Cens. c_i
1	22	1	0
2	23	1	0
3	38	1	0
4	42	2	0
5	73	0	1
6	77	1	0
7	89	1	0
8	115	1	0



Share of survivors at time index i

$$n_i/N = n_{i-1}/N(1-d_i/n_{i-1})$$
$$S_i = S_{i-1}(1-d_i/n_{i-1})$$

The Kaplan-Meier Estimator
Estimates the probability of survival until time *x*

$$S(x) = \prod_{x_i \leq x}(1-d_i/n_{i-1})$$

Time	x_i	0	22	23	38	42	73	77	89	115
Number of events	d_i	0	1	1	1	2	0	1	1	1
Number of censorings	c_i		0	0	0	0	1	0	0	0
Number of survivors	n_i	9	8	7	6	4	3	2	1	0

Recap

For the Kaplan-Meier Estimator read

Klein & Moeschberger (2003). Survival Analysis. pp. 91–93.

You can find the original paper here

Kaplan & Meier (1958). Nonparametric Estimation from Incomplete Observations. 10.2307/2281868

Homework

Using R, produce a Kaplan-Meier plot related to the topic of your eventual seminar paper (“Hausarbeit”).

You don’t need to have all the data for your topic yet, but you need to find “some” related data. Be prepared to present your plot to the group. Think about study time start and end, event of interest, and censoring. You may compare multiple groups, but a KM-plot for a single group is fine as well.

Materials for this lecture

github.com/jschoeley/survival_analysis-ur-ss22

Jonas Schöley

@jschoeley

0000-0002-3340-8518

j.schoeley@uni-rostock.de