Survival Analysis

Robert Winslow

GalvanizeU

July 11th, 2017

Survival analysis

Motivation

Survival analysis is a way to describe how long things last.

It is often used to study human lifetimes. . .

and it also applies to *survival* of mechanical and electronic components, or more generally to intervals in time before an event.

Survival analysis

Examples:

- 5-year mortality rate
- LD50 in pharmacology
- How long a politician stays in office
- MTTF - Mean time to failure of a mechanical component (like a hard drive)
- Length of time people remain unemployed after a job loss

Survival analysis

Survival curves

The fundamental concept in survival analysis.

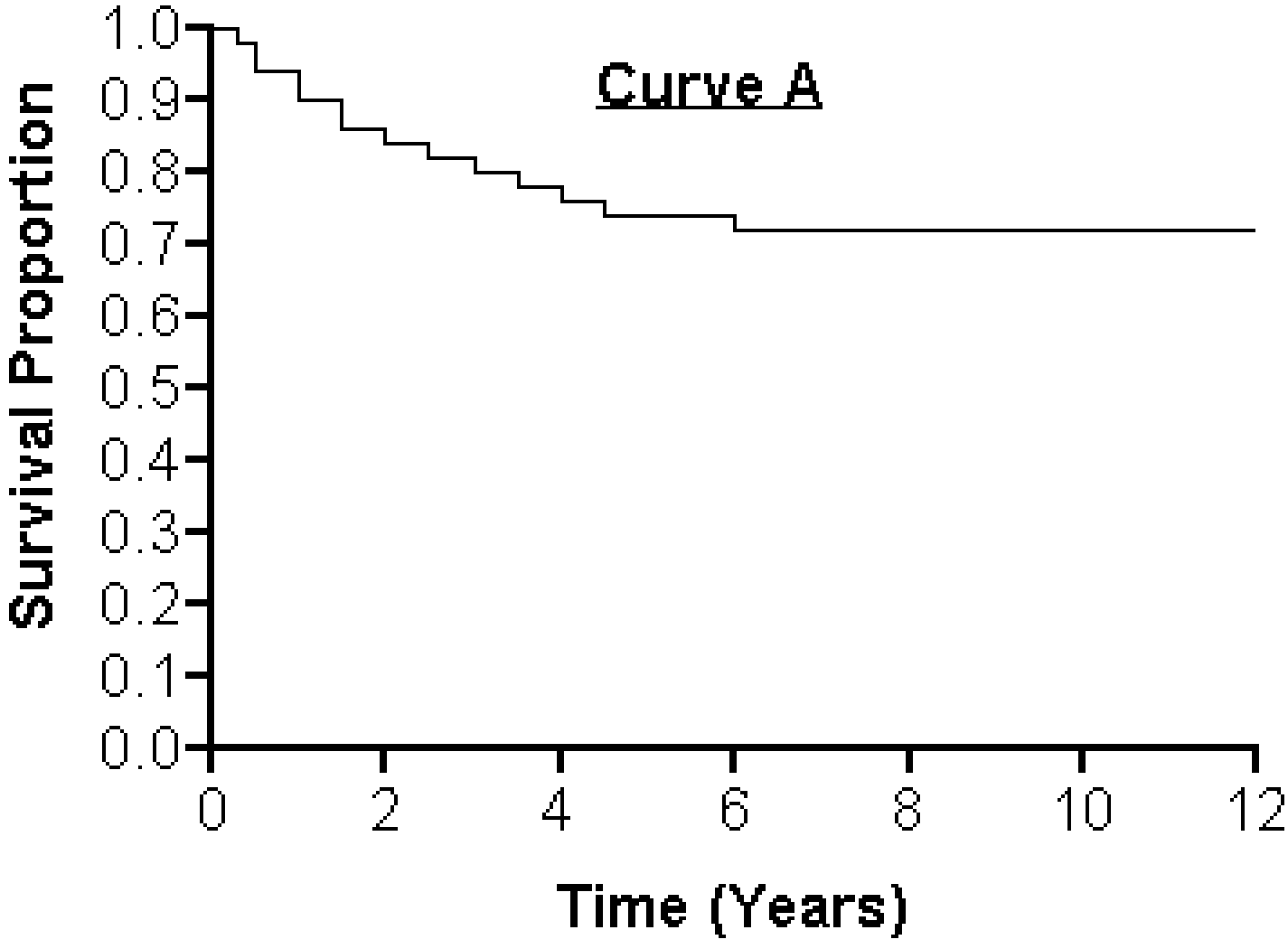
$S(t)$ is a function that maps from a duration t to the probability of surviving longer than t.

If you know the cumulative distribution function (CDF), then it is easy:

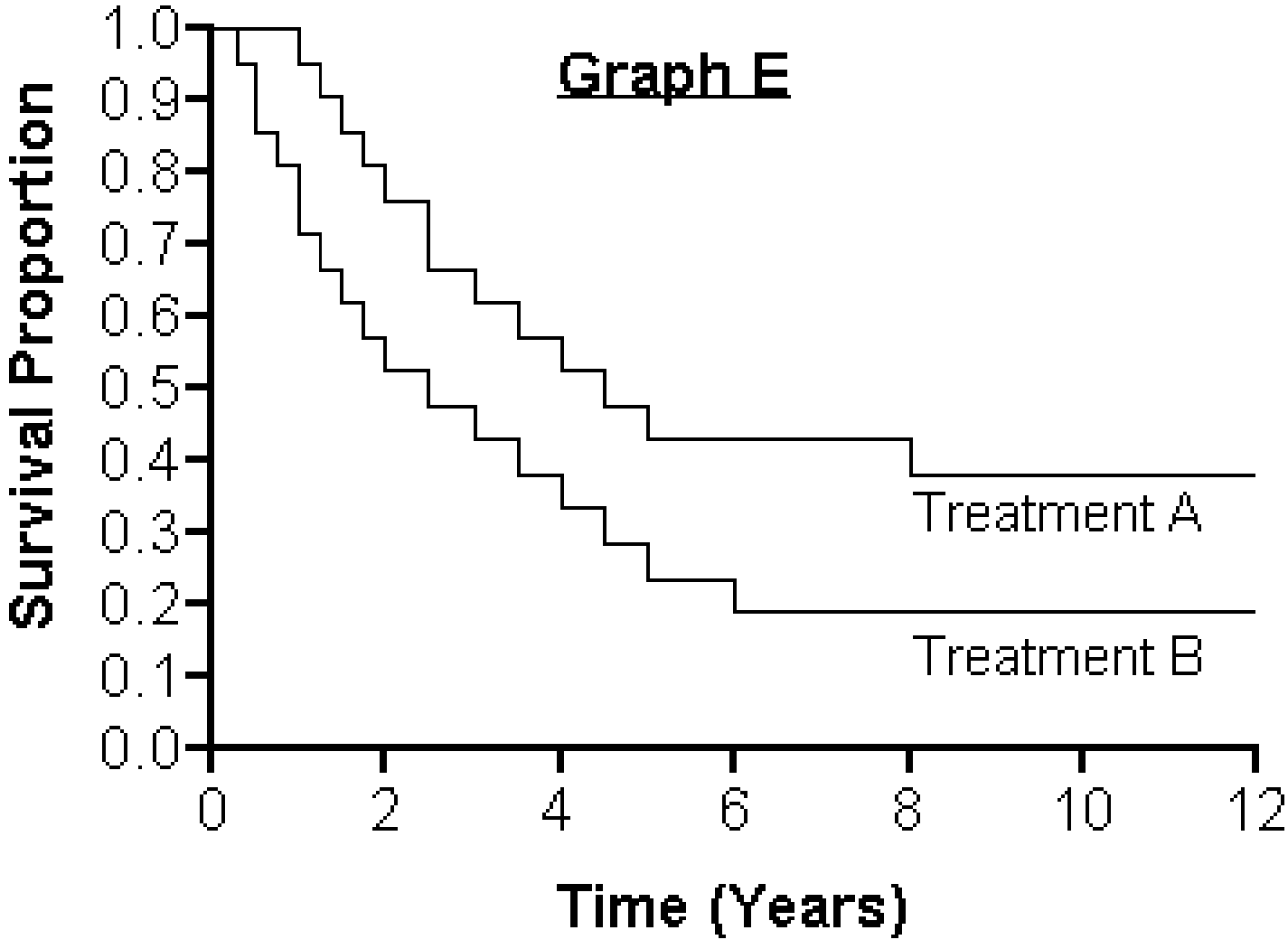
$$S(t) = 1 - CDF(t)$$

Where $CDF(t)$ is the probability of a lifetime less than or equal to t.

Survival analysis



Survival analysis



Survival analysis

We can use the survival function to estimate the expectated survival rate of the whole population.

We can also break up our data based on certain attributes and compare their survival functions.

If the different groups have notably different survival functions, then perhaps there is a true difference between those populations.

(Like in a drug trial.)

Survival analysis

Hazard function

From the survival curve we can derive the *hazard function*.

The hazard function is also known as the *failure rate*.

Survival analysis

Failure rate $=$

$$\lambda(t) = \frac{S(t) - S(t+1)}{S(t)}$$

Interpretation:

The proportion of failures occuring at time t (and not before).

Survival analysis

If we have the *CDF* function, then we can compute the survival and hazard functions directly.

Let's make estimation more realistic (and more difficult):

In general, we do not have access to the *CDF*.

Survival analysis

In particular, we usually have incomplete data.

This can be a good thing!

Example:

Estimating the outcomes of a drug trial.

Incomplete *CDF* means that there are people who are still alive.

We want to estimate their survival probability, typically based on an intervention like a new drug.

Survival analysis

Censorship is the term used to describe the data that is incomplete.

If 40% of respondents have not reached the 'failure' event, then the censorship rate is 40%.

Survival analysis

If we have incomplete data, then how can we get the survival function?

Kaplan-Meier estimation can help.

One of the central algorithms in survival analysis.

Cited over 50,000 times since publishing in 1958.

General idea: we use our data to estimate the hazard function, then we convert the hazard function to the survival function.

Survival analysis

Kaplan-Meier general approach:

We consider, for each t, the number of failures at t.

We then compare that to the number of still-surviving entities at t.

This ratio (computed from our data) gives the KM estimate.