

### A simple example of a Kaplan-Meier curve (January 24, 2008).

In response to a query, I wanted to write up a simple example of how to calculate survival probabilities when you have censored data. It is adapted from Chapter 6 of my book, [Statistical Evidence in Medical Trials](#). I have updated and simplified the example, for possible use in a second edition of the book, if I am so lucky.

Survival data models provide interpretation of data representing the time until an event occurs. In many situations, the event is death, but it can also represent the time to other bad events such as cancer relapse or failure of a medical device. It can also be used to denote time to positive events such as pregnancy.

Survival data models also incorporate one of the complexities of "time to event" data, the fact that not all patients experience the event during the time frame of the study. So, if we are doing a five year mortality study, we have the problem of those stubborn patients who refuse to die during the study period. Other patients may move out of town halfway through the study and are lost to follow-up. In a study of medical devices, sometimes the device continues to work up to a certain time, but then has to be removed, not because the device failed, but because the patient got healthier and no longer needed the device.

These observations are called censored observations. With censored observations, the actual time of the event is unknown but we do know that it would not be any earlier than the time that the last evaluation or follow-up visit was done. These censored observations provide partial information. They influence our estimates of survival probability up to the last evaluation or follow-up, but do not provide any information about survival probabilities beyond that point. To disregard this information is dangerous and could seriously bias your results.

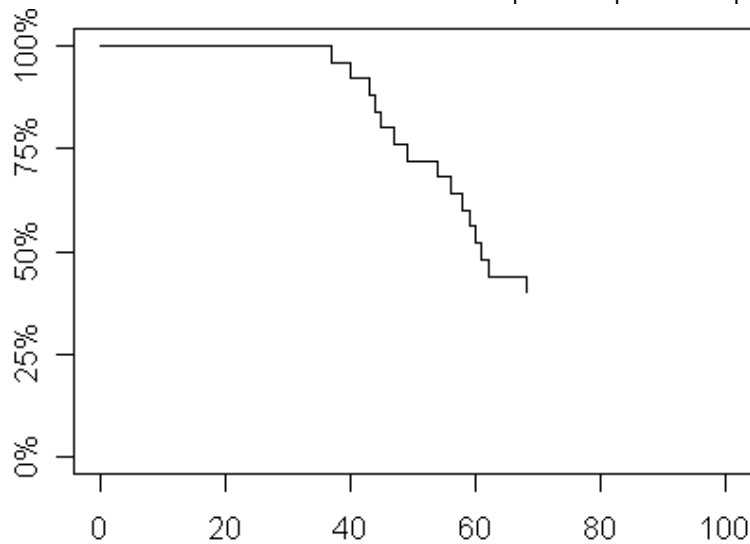
The following data represents survival time for a group of fruit flies and is a subset of a larger data set found at the [Data and Story Library \(DASL\)](#). The data set has been slightly modified to simplify some of these explanations.

There are 25 flies in the sample, with the first fly dying on day 37 and the last fly dying on day 96. If you wanted to estimate the survival probability for this data, you would draw a curve that decreases by 4% (1/25) every time a fly dies.

Day	Prob
37	96%
40	92%
43	88%
44	84%
45	80%
47	76%
49	72%
54	68%
56	64%
58	60%
59	56%
60	52%
61	48%
62	44%
68	40%
70	36%
71	32%
72	28%
73	24%
75	20%
77	16%
79	12%
89	8%
94	4%
96	0%

Here's a graph of these probabilities over time.





By the way, you might be tempted to ignore the ten flies who escaped. But that would seriously bias your results. All of these flies were tough and hardy flies who lived well beyond the median day of death. If you pretended that they didn't exist, you would seriously underestimate the survival probabilities. The median survival time, for example, of the 15 flies who did not escape, for example, is only 54 days which is much smaller than the actual median.

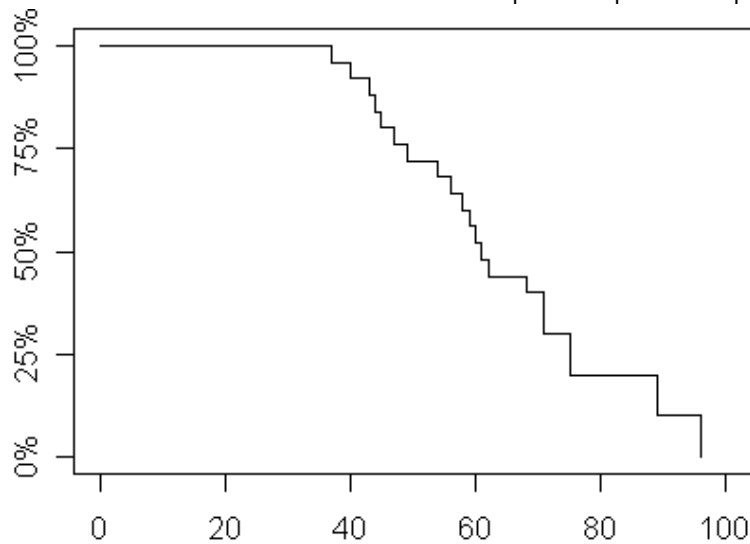
Let's look at a third experiment, where the screen cover is left open and all but four of the remaining flies escape. It turns out that those four remaining flies who didn't bug out will allow us to still get reasonable estimates of survival probabilities beyond 70 days. Here is the data and the survival probabilities.

Day	Prob
37	96%
40	92%
43	88%
44	84%
45	80%
47	76%
49	72%
54	68%
56	64%
58	60%
59	56%
60	52%
61	48%
62	44%
68	40%
70+	?
71	30%
70+	?
70+	?
75	20%
70+	?
70+	?
89	10%
70+	?
96	0%

What you do with the six escaped flies is to allocate their survival probabilities equally among the four flies who didn't bug out. This places a great responsibility among each of those four remaining flies since each one is now responsible for 10% of the remaining survival probability, their original 4% plus 6% more which represents a fourth of the 24% survival probability that was lost with the six escaping flies.

Another way of looking at this is that the six flies who escaped influence the denominator of the survival probabilities up to day 70 and then totally drop out of the calculations for any further survival probabilities. Because the denominator has been reduced, the jumps at each remaining death are much larger.

Here is a graph of the survival probability estimates from the third experiment.

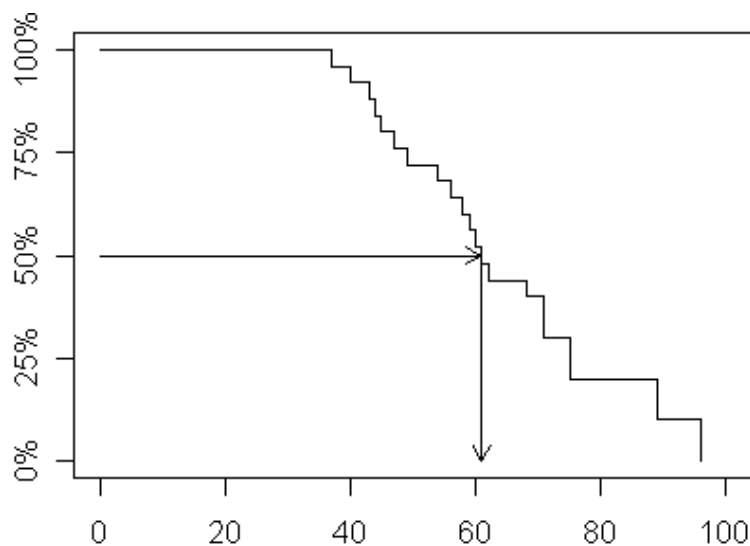


These survival probabilities differ only slightly from the survival probabilities in the original experiment. This works out because the mechanism that caused us to lose information on six of the fruit flies was independent of their ultimate survival.

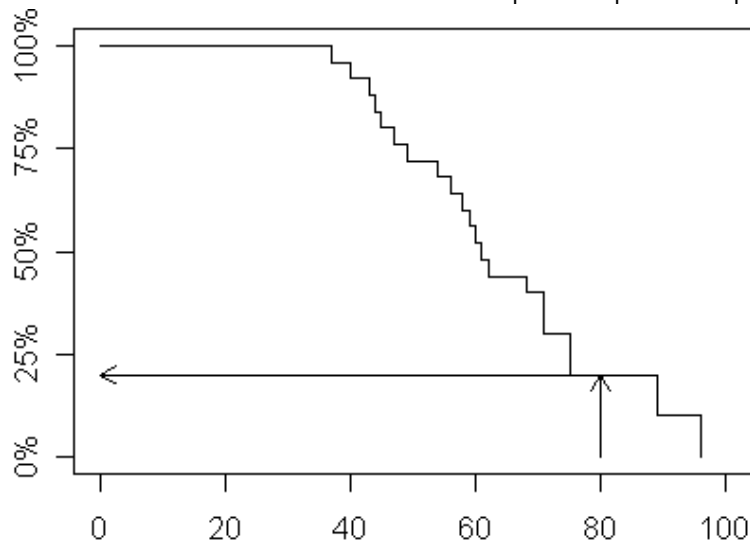
If the censoring mechanism were somehow related to survival prognosis, then you would have the possibility of serious bias in your estimates. Suppose for example, that only the toughest of flies (those with the most days left in their short lives) would have been able to escape. The flies destined to die on days 70, 71, 72, and 73, were already on their deathbeds and unable to fly at all, much less make a difficult escape. Then these censored values would not be randomly interspersed among the remaining survival times, but would constitute some of the larger values. But since these larger values would remain unobserved, you would underestimate survival probabilities beyond the 70th day.

This is known as informative censoring, and it happens more often than you might expect. Suppose someone drops out of a cancer mortality study because they are abandoning the drugs being studied in favor of laetrile treatments down in Mexico. Usually, this is a sign that the current drugs are not working well, so a censored observation here might represent a patient with a poorer prognosis. Excluding these patients would lead to an overestimate of survival probabilities.

When you see a survival curve in a research paper, there are two ways to interpret it. First, you can get an estimate of the median (or other percentiles) by projecting horizontally until you intersect with the survival curve and then head down to get your estimate. In the survival curve we have just looked at, you would estimate the median survival as slightly more than 60 days.



You can also estimate probabilities for survival at any given time by projecting up from the time and then moving to the left to estimate the probability. In the example below, you can see that the 80 day survival probability is a little bit less than 25%.



This webpage was written by Steve Simon on 2008-xx-xx, edited by Steve Simon, and was last modified on 2008-07-14. Send feedback to [ssimon at cmh dot edu](mailto:ssimon@cmh.edu) or click on the email link at the top of the page. [Category: Survival analysis](#)