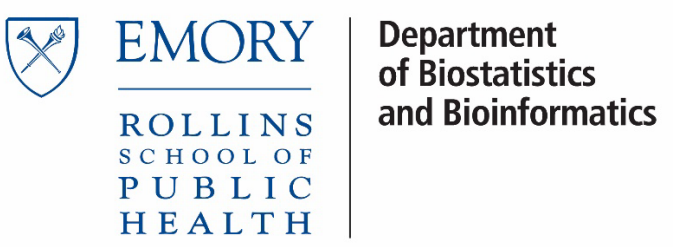


BIOS 522: Survival Analysis Methods

Reading 2:

The survival function

This week, we will define the survival function for time-to-event data. We will learn about two non-parametric methods for estimating the survival function: the empirical CDF when censoring is absent, and the Kaplan-Meier method when censoring is present.

Part 1. The survival function

A **survival random variable** T measures the time elapsed from an origin ("time zero") until the event of interest. T is positive ($T > 0$). T can either be discrete or continuous. In this class, with limited exceptions, we will focus on continuous survival times only.

A key estimand for a survival random variable is the **survival function** $S(t)$, also known as the **survival curve**. This function defines the probability that an individual will "survive" beyond a given length of time t, i.e.:

$$S(t) = \Pr(T > t)$$

Several key properties follow from this definition.

- $S(t) = \Pr(T > t) = 1 - \Pr(T \leq t) = 1 - F(t)$ where $F(t)$ is the **cumulative distribution function** (CDF) of T. The CDF and survival function are "opposites" of each other. thus, given the CDF, we can directly calculate $S(t)$.

- Since $S(t)$ is a probability, $0 \leq S(t) \leq 1$ for all t.

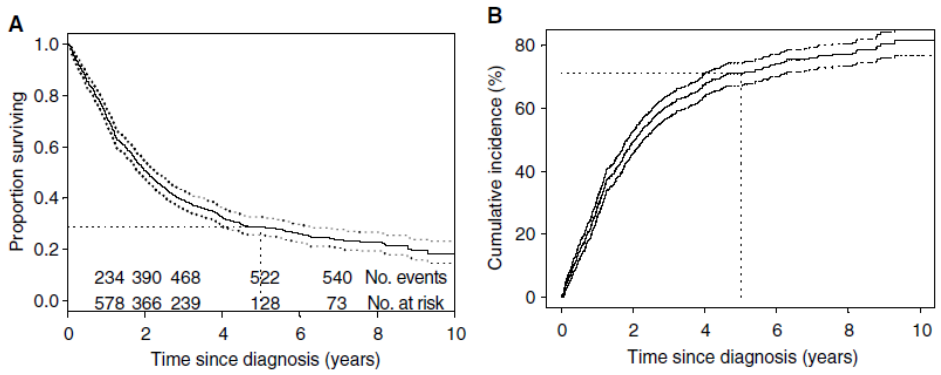
- Since everyone is still at-risk at time 0, $S(0) = 1$ (100% survival).

- Assuming everyone eventually fails, $S(\infty) = 0$. (0% survival after infinite follow-up.)

- $S(t)$ decreases or stays constant over time but never increases. If $t < u$, you can survive to time u only if you survive to time t, so $S(t) \geq S(u)$.

Much of survival analysis is dedicated to calculating and comparing survival functions. Graphical displays of survival functions are probably the most common method used to summarize data. We can use the survival function to estimate median survival or survival at a landmark time (e.g. 5-year survival).

Example: The figures below summarize the survival time distribution for 825 patients diagnosed with primary epithelial ovarian carcinoma between January 1990 and December 1999 at the Western General Hospital in Edinburgh. Figure A (left) summarizes the survival function $S(t)$, and Figure B (right) summarizes the cumulative incidence (i.e., CDF) $F(t)$. We can see that they are opposite one another. At each time point, they add up to 1 (equivalently, 100%). From Figure A, we see that median survival is approximately 2 years after diagnosis. The five-year survival probability is 29%.



Source: Clark et al. (2003) British Journal of Cancer
https://doi.org/10.1038/sj.bjc.6601118

Part 2. Empirical CDF

In the absence of censoring, estimating the survival function from data does not require specialized methods. We start by estimating the CDF. Recall that the CDF of a random variable T is:

$$F(t) = \Pr(T \leq t)$$

When we observe uncensored survival times $T_1, \ldots, T_n$ from n participants drawn from the same underlying distribution T, the **empirical CDF** is the right-continuous function:

$$\hat{F}(t) = \frac{1}{n}\sum_{i=1}^{n} I[T_i \leq t]$$

For a fixed value of t, $\hat{F}(t)$ is the proportion of the observed survival times that are less than or equal to t. Thus, the estimated survival function is:

$$\hat{S}(t) = 1 - \hat{F}(t)$$

This is a non-parametric estimator of the survival function because its shape is driven by the data and is not assumed to follow a particular form. Like the empirical CDF, it is also right-continuous.

Example: Researchers collected data on the remission times of 21 leukemia patients who received standard of care therapy. Leukemia eventually recurred in all of these people (no censoring). Their recurrence times (in weeks) from the time of initiation of standard of care therapy are listed below.

1, 1, 2, 2, 3, 4, 4, 5, 5, 8, 8, 8, 8, 11, 11, 12, 12, 15, 17, 22, 23

Source: Cox and Oakes (1984) Analysis of Survival Data

Calculate a point estimate for the probability of a leukemia patient experiencing recurrence at or before 10 weeks after initiation of therapy.

If the recurrence time is a random variable T, then we want to estimate:

$$F(10) = \Pr(T \leq 10)$$

Because there is no right censoring in the data, we can use the empirical CDF. This is the proportion of recurrence events occurring at or before 10 weeks:

$$\hat{F}(10) = \frac{13}{21} = 0.619$$

61.9% (13/21) of leukemia patients had experienced recurrence at or before 10 weeks after initiation of standard of care therapy.

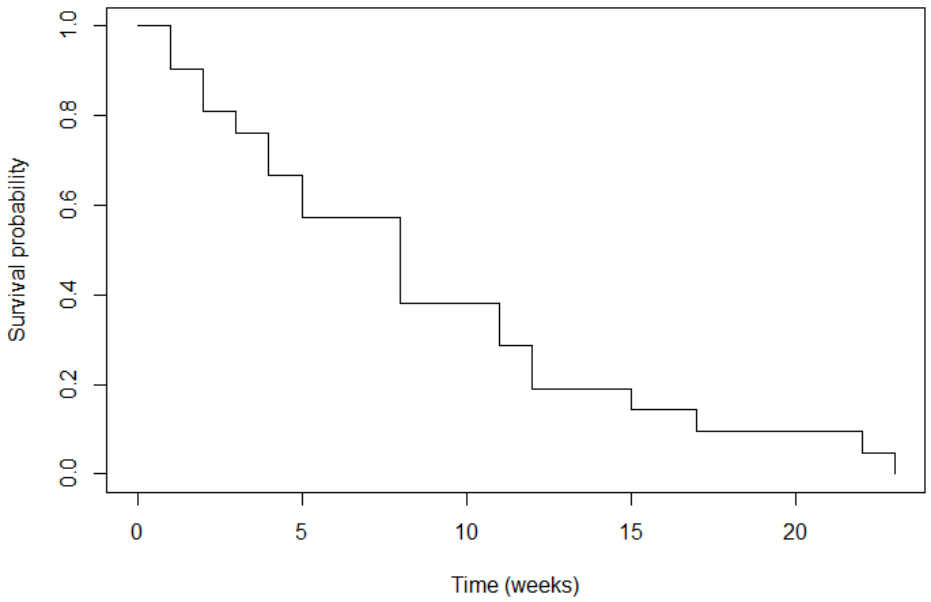
Calculate a point estimate for the probability of surviving recurrence-free more than 10 weeks after initiation of therapy.

$$\hat{S}(10) = 1 - \hat{F}(10) = 1 - 0.619 = 0.381$$

8 out of 21 patients (38.1%) had not yet experienced recurrence by 10 weeks. The estimated probability of surviving more than 10 weeks after initiation of standard of care therapy without experiencing recurrence is 38.1%.

To calculate the entire survival curve, we can repeat this process at all times t.

Example: Continuing with the leukemia example above, we can calculate and plot the survival function from one minus the empirical CDF. Note the stepwise appearance of the plot. The survival probability only changes at the time of a new failure (recurrence). It drops to a lower value at the time of failure. Large drops correspond to times when multiple people fail simultaneously.



The empirical CDF is closely related to the binomial distribution, because, at each time t, the number of observations that failed at or before t follows a binomial distribution with sample size n and probability $F(t)$.

By the law of large numbers, the empirical CDF is unbiased for the true CDF:

$$E\left[\hat{F}(t)\right] = F(t)$$

The variance of the empirical CDF at time t is:

$$Var\left(\hat{F}(t)\right) = \frac{1}{n}F(t)\left(1 - F(t)\right)$$

The above variance depends upon knowing the true CDF $F(t)$, so in practice we can estimate the **variance** by substituting $\hat{F}(t)$ for $F(t)$:

$$\widehat{Var}\left(\hat{F}(t)\right) = \frac{1}{n}\hat{F}(t)\left(1 - \hat{F}(t)\right)$$

By the central limit theorem (CLT), for large n, an approximate pointwise 95% **confidence interval** for $F(t)$ is:

$$\hat{F}(t) \pm 1.96\sqrt{\frac{1}{n}\,\hat{F}(t)\left(1 - \hat{F}(t)\right)}$$

This is a **Wald confidence interval**, sometimes referred to as a standard confidence interval.

Example: Using the leukemia data set provided earlier, calculate a 95% confidence interval for the probability of a leukemia patient experiencing recurrence at or before 10 weeks after initiation of standard of care therapy.

We can construct a Wald confidence interval for $F(10)$.

$$\hat{F}(10) \pm 1.96\sqrt{\frac{\hat{F}(10)(1-\hat{F}(10))}{n}} = 0.619 \pm 1.96\sqrt{\frac{0.619(1-0.619)}{21}}$$

The 95% confidence interval is (41.1%, 82.7%).

A limitation of the Wald interval method is that there is no guarantee that it will be bounded between 0 and 1. To ensure that the confidence limits for a probability are never negative, we can apply a **log transformation**. We start by constructing an interval for $\log\big(F(t)\big)$, centered at $\log\left(\hat{F}(t)\right)$, with variance $\widehat{Var}\left[\log\left(\hat{F}(t)\right)\right]$.

$$\log\left(\hat{F}(t)\right) \pm 1.96\sqrt{\widehat{Var}\left[\log\left(\hat{F}(t)\right)\right]}$$

We can estimate the variance of log-transformed $\hat{F}(t)$ via the Delta method (more to follow in your homework). Finally, we exponentiate the endpoints to calculate an interval for $F(t)$.

$$\exp\left\{\log\left(\hat{F}(t)\right) \pm 1.96\sqrt{\widehat{Var}\left[\log\left(\hat{F}(t)\right)\right]}\right\}$$

$$= \hat{F}(t)\exp\left\{\pm 1.96\sqrt{\widehat{Var}\left[\log\left(\hat{F}(t)\right)\right]}\right\}$$

The log transformation guarantees that our confidence interval will never be negative, but our confidence interval could exceed 1. This is not appropriate for a probability.

Another common transformation is the **log-log transformation**, also called the *complementary log-log*, or *"log-minus-log" transformation*. It has this form:

$$y = \log\{-(\log(x))\}$$

Or it can be undone to solve for x:

$$x = \exp(-\exp(y))$$

Note that for *any value of y,* $\exp(y)$ is positive, $-\exp(y)$ is negative, and $\exp(-\exp(y))$ is *between 0 and 1*. Very helpful for a probability!

Following the same procedure used for the log transformation, we can calculate a confidence interval centered at $\log\left(-\log\left(\hat{F}(t)\right)\right)$.

$$\log\left(-\log\left(\hat{F}(t)\right)\right) \pm 1.96\sqrt{\widehat{Var}\left[\log\left(-\log\left(\hat{F}(t)\right)\right)\right]}$$

Again, the variance term can be approximated via the Delta method. The last step is to apply the transformation $\exp(-exp(y))$ to get our final result.

$$\exp\left\{-\exp\left[\log\left(-\log\left(\hat{F}(t)\right)\right) \pm 1.96\sqrt{\widehat{Var}\left[\log\left(-\log\left(\hat{F}(t)\right)\right)\right]}\right]\right\}$$

$$= \exp\left\{\log\left(\hat{F}(t)\right)\exp\left(\mp 1.96\sqrt{\widehat{Var}\left[log\left(-log\left(\hat{F}(t)\right)\right)\right]}\right)\right\}$$

$$= \left[\hat{F}(t)\right]^{\exp\left(\mp 1.96\sqrt{\widehat{Var}\left[log\left(-log\left(\hat{F}(t)\right)\right)\right]}\right)}$$

This provides a log-log transformed interval of $F(t)$, from which we could readily calculate an interval for $S(t)$.

Part 3. Kaplan-Meier estimation

In data with censoring, we cannot calculate the empirical CDF directly because we do not observe the failure times of anyone who is right censored. There is valuable information, though, in knowing that an individual survived until being censored. We desire a rigorous method that leverages this information.

The **Kaplan-Meier estimator** combines information from samples with observed failure times and samples that were censored to form a single estimated survival curve $\hat{S}(t)$ meant to reflect the true underlying survival curve $S(t)$. In other words, using right-censored data T^*, we make inference about the distribution of the underlying random variable T.

The Kaplan-Meier estimator is a nonparametric estimator because it does not assume that the data fit any underlying parametric distribution. It is similar in spirit to the empirical CDF, but modified to handle censored data.

Before explaining how Kaplan-Meier estimation works, it is useful to think about another way to calculate the empirical CDF.

Example: Recall our leukemia example of 21 patients.

At $t = 1$ week, 2 patients experience recurrence, leaving 19 patients at risk. $\hat{S}(t = 1) = 19/21$ or 90.5%.

At $t = 2$ weeks, 2 of the 19 patients at risk experience recurrence. This leaves 17 patients at risk. $\hat{S}(t = 2) = 17/21$ or 81.0%.

Another way to estimate survival at 2 weeks is:

$$\begin{aligned}\hat{S}(t = 2) &= \widehat{\Pr}(T > 2) \\ &= \widehat{\Pr}(T > 2 \cap T > 1) \\ &= \widehat{\Pr}(T > 1)\widehat{\Pr}(T > 2|T > 1)\end{aligned}$$

Plugging in the numbers:

$$\begin{aligned}\hat{S}(t = 2) &= \left(\frac{19}{21}\right)\left(\frac{17}{19}\right) \\ &= \left(\frac{17}{21}\right) = 81.0\%\end{aligned}$$

Note how the 19 cancels. This logic extends to later time intervals. At $t = 3$ weeks, 1 of the 17 remaining patients at risk experiences recurrence. This leaves 16 patients at risk.

$$\begin{aligned}\hat{S}(t = 3) &= \left(\frac{19}{21}\right)\left(\frac{17}{19}\right)\left(\frac{16}{17}\right) \\ &= \left(\frac{16}{21}\right) = 76.2\%\end{aligned}$$

It is much simpler to calculate $16/21$ instead of multiplying many fractions. But the simplified calculation is only possible because the interior numerators and denominators cancel.

In the above example, we broke time up into successive intervals. Each interval ends at a time when failures occur (1, 2, 3 weeks and so on). We calculate the probability of surviving past each interval conditional on having survived at least until the start of the interval. We then multiply these conditional probabilities to estimate the probability of survival.

Note that individuals who have failed before an interval starts are not included in the calculation of subsequent conditional probabilities.

Example: Continuing with our previous leukemia calculation, the conditional probability of surviving past three weeks given survival past two weeks is:

$$\widehat{\Pr}(T > 3|T > 2) = 16/17$$

This calculation excludes the 4 individuals who failed by two weeks.

To accommodate censoring, we can apply a similar reasoning. As we will see in the Kaplan-Meier approach, individuals only contribute to the conditional probability if they are still under observation and at-risk during that interval. After someone is censored, they do not contribute to any subsequent intervals.

The Kaplan-Meier estimator

The **Kaplan-Meier or KM estimator** of $S(t)$ is:

$$\hat{S}(t) = \prod_{j:t_j \leq t} \hat{q}_j$$

where $\hat{q}_j$ is the conditional probability of surviving past time t_j given that one has already survived past t_{j-1}. We estimate these conditional probabilities using data from the n_j individuals **at risk** at the start of the interval, i.e. immediately following t_{j-1}. This is also known as the risk set, and it excludes individuals who have previously failed or been censored.

The interval ends with d_j failures and c_j censored observations at time t_j. The proportion of those at risk who fail during the interval is d_j/n_j. The proportion who survive is $1 - d_j/n_j$. Thus, we can replace $\hat{q}_j$ in our KM estimator:

$$\hat{S}(t) = \prod_{j:t_j \leq t} \left(1 - \frac{d_j}{n_j}\right)$$

The Kaplan-Meier estimator is also known as the **product limit estimator**. That is because our estimate $\hat{S}(t)$ is the product of conditional survival probabilities in all intervals that end before time t.

Defining the Kaplan-Meier intervals

In practice, we define as many intervals as unique failure/censoring times, so that each failure/censoring time can be at the end of an interval. Consider data with unique, ordered, failure/censoring times $t_1, \ldots, t_J$. Each of these times becomes a boundary for a new interval.

We can more formally define the risk set as $R(t) = \{i: T_i^* \geq t\}$. It is the set of all individuals who are at risk immediately before time t, and so includes anyone who fails at time t. At time t_{j-1}, the size of the risk set is n_{j-1}. Immediately after t_{j-1}, the size of the risk set drops to $n_j = n_{j-1} - d_{j-1} - c_{j-1}$, excluding anyone who failed or was censored at t_{j-1}. The size of the risk set is thus a left-continuous function, i.e., the number at risk from $\left(t_{j-1}, t_j\right] = n_j$.

In contrast, the Kaplan-Meier estimate is a right-continuous function. That is because it is defined as the product of all intervals j such that $t_j \leq t$. Thus, at

time t_j, a new interval is included, and the function drops. This is best seen in an example.

Example: *A total of 12 hemophiliacs, all 40 years of age or younger at the time of HIV seroconversion in the 1980s, were followed from primary AIDS diagnosis until death. (It would be preferable to use the time which AIDS developed as the starting point rather than time of diagnosis, but this information was not known.) For most of the patients, treatment was not available. Two of the patients were lost to follow-up. Their follow-up times (in months) are listed in sorted order:*

2, 3+, 6, 6, 8, 10+, 15, 15, 16, 27, 30, 32

Based on this sample, we seek to make inference about the survival function S(t) of the population of hemophiliacs under the age of 40 who were diagnosed with AIDS in the 1980s. The origin time is the time of primary AIDS diagnosis. Their follow-up times (in months) are listed in the table on the right in sorted order.

Use the Kaplan-Meier estimator to obtain a point estimate of survival 7 months after primary AIDS diagnosis. Interpret your point estimate in the context of the problem.

We are asked to calculate $\hat{S}(t = 7)$. We must first divide time into intervals such that no one leaves the cohort except at the end of the interval. We will construct intervals that end before time $t = 7$. Since after $t = 6$, the next observed follow-up time is at $t = 8$, we stop after three intervals.

Unique failure/ censoring time t_j	**Number at risk n_j during $(t_{j-1}, t_j]$**	**Number of deaths d_j at t_j**	**Number censored c_j at t_j**	**Conditional survival probability $\hat{q}_j$**	**Kaplan-Meier estimate $[t_j, t_{j+1})$**
$t_0 = 0$					$t = [0,2)$ $\hat{S}(t) = 1$
$t_1 = 2$	$t = (0,2]$ $n_1 = 12$	$d_1 = 1$	$c_1 = 0$	$\hat{q}_1 = \left(1 - \frac{1}{12}\right)$	$t = [2,3)$ $\hat{S}(t) = \widehat{q_1}$
$t_2 = 3$	$t = (2,3]$ $n_2 = 11$	$d_2 = 0$	$c_2 = 1$	$\hat{q}_2 = 1$	$t = [3,6)$ $\hat{S}(t) = \widehat{q_1 q_2}$
$t_3 = 6$	$t = (3,6]$ $n_3 = 10$	$d_3 = 2$	$c_3 = 0$	$\hat{q}_3 = \left(1 - \frac{2}{10}\right)$	$t = [6,8)$ $\hat{S}(t) = \widehat{q_1 q_2 q_3}$

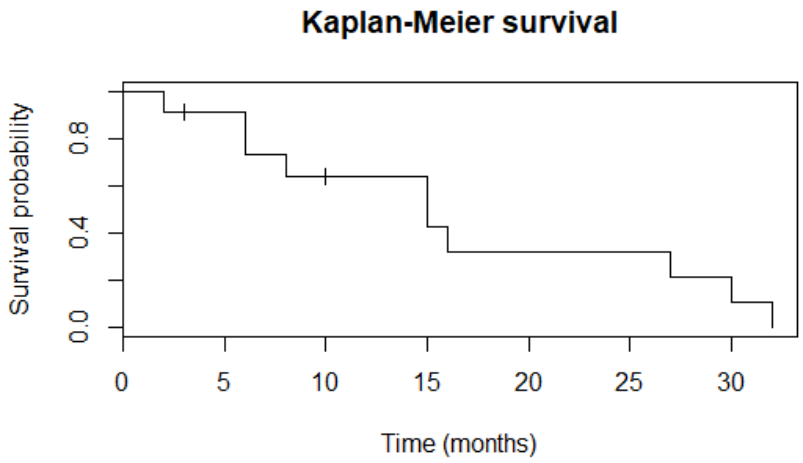
In the previous table, note how each row considers time *prior to the failure* time t_j to calculate how many people are at risk. But then the impact of this row on the Kaplan-Meier estimator via $\hat{q}_j$ only occurs *after the failure* time t_j.

The Kaplan-Meier estimate is the product of the $\hat{q}_j$ terms for each interval.

$$\hat{S}(t=7) = \prod_{j:t_j \leq 7} \hat{q}_j = \left(1 - \frac{1}{12}\right) \times 1 \times \left(1 - \frac{2}{10}\right) = 0.733$$

An estimated 73.3% of hemophiliacs are alive 7 months after primary AIDS diagnosis.

The survival probability using this data is plotted below. Note that censored observations are marked with small vertical lines.



Features of the Kaplan-Meier estimator

The Kaplan-Meier estimator $\hat{S}(t)$:

- Has a step function appearance.
- Is equal to one up to the first death time.
- It only drops at the time of failure. It does not drop when individuals are censored. (Note how $\hat{q}_2 = 1$ in the AIDS example.)
- Drops to 0 if the last event is a death. (Has poor fit in tails.)
- In the absence of censoring, the Kaplan-Meier simplifies to the empirical CDF.

A Kaplan-Meier plot is the most common figure shown in a paper with time-to-event data.

Variance and confidence intervals for the Kaplan-Meier estimator

$S(t)$ is the true population survival function. $\hat{S}(t)$ is an estimate of the true population survival function calculated using the information in a sample of right-censored observations. A second sample would likely result in a different survival curve. To capture this sampling variability, we can calculate the variance (or standard error) of $\hat{S}(t)$.

Greenwood's formula is used to estimate the variance of the Kaplan-Meier estimator. The variance is calculated with the same time intervals used in the Kaplan-Meier estimator:

$$\widehat{Var}\ \hat{S}(t) = \left(\hat{S}(t)\right)^2 \sum_{j:t_j \le t} \frac{d_j}{n_j(n_j - d_j)}$$

A sum is taken over all time intervals j that end before time t. At later time points, if most individuals have already failed or been censored, the variance can be large.

The standard error is the square root of the Greenwood's formula variance. By the central limit theorem (CLT), for large n, an approximate pointwise Wald 95% confidence interval for $S(t)$ is:

$$\hat{S}(t) \pm 1.96 \sqrt{\widehat{Var}(\hat{S}(t))}$$

Though this confidence interval is for a probability, there is no guarantee that it will be bounded between 0 and 1. It is common to report a log transformed interval or a log-log transformed interval. This follows the same structure as for the empirical CDF. We will apply the Delta method to estimate the variance terms in an upcoming homework assignment.

Example: Continuing with the previous example, use Greenwood's formula to approximate the variance of the estimated survival function 7 months after primary AIDS diagnosis. Construct a Wald-type 95% confidence interval for the true survival function 7 months after primary AIDS diagnosis.

We are asked to estimate $\widehat{Var}(\hat{S}(t = 7))$ using Greenwood's formula.

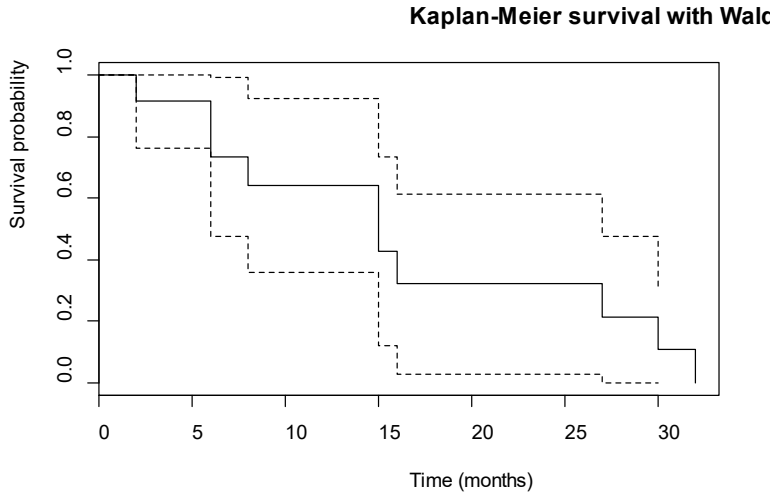
$$\widehat{Var}\left(\hat{S}(t)\right) = \hat{S}(t)^2 \sum_{j:t_j \le 7} \frac{d_j}{n_j\left(n_j - d_j\right)}$$
$$= 0.733^2 \left(\frac{1}{12(12-1)} + \frac{0}{11(11-0)} + \frac{2}{10(10-2)}\right) = 0.0175$$

The variance is 0.0175, and the standard error is $\sqrt{0.0175}$.

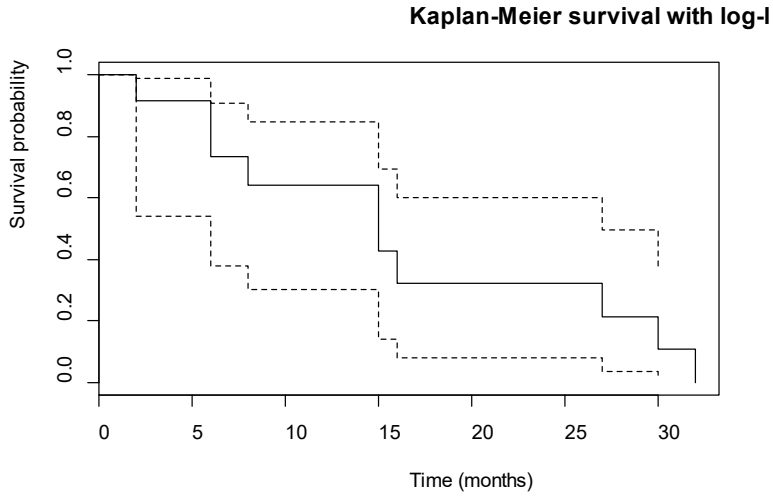
We are also asked to construct a 95% confidence interval for $S(t = 7)$.

$$\hat{S}(t = 7) \pm 1.96\ \widehat{SE}(\hat{S}(t = 7)) = 0.733 \pm 1.96\sqrt{0.0175} = (0.474, 0.993)$$

A 95% confidence interval for the survival probability of hemophiliacs 7 months after primary AIDS diagnosis is (47.4%, 99.3%). This interval is wide because of the small sample size. The limits are added to the plot below.



We may prefer to report a confidence interval with log-log transformed limits, as these are guaranteed to stay between 0 and 1. Observe how, unlike the Wald limits, the transformed limits are asymmetric around the point estimate.



Part 4. Censoring

Informative and non-informative censoring

The Kaplan-Meier estimator, like most other methods in survival analysis, relies on an assumption known as **non-informative (or independent) censoring**.

Non-informative censoring means that, at each censoring time, those who are censored have the same prognosis as those who remain under observation. Being censored at time c tells us only that $T > c$, and nothing more. Formally, the time-to-event variable T is independent of the censoring variable C (i.e., $T_i \perp C_i$). Sometimes, censoring may depend upon covariates but T and C are **conditionally independent** within levels of the covariates.

All of the methods in this course assume independent or conditionally independent censoring. Unfortunately, we are unable to check whether the assumption of non-informative censoring is valid based on the data alone. This is a so-called "untestable assumption" exactly because the survival times for censored individuals are missing from our data set.

Fortunately, censoring is non-informative in the following cases:

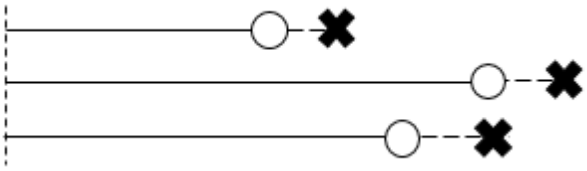
- **Administrative censoring** occurs when observation of each individual ends at a predetermined time under the control of the investigators; since $C = c$ is the same for everyone, T and C are independent.

- Censoring times vary across individuals, but it is reasonable to assume that they are **random** and thus unassociated with T.

Different censoring mechanisms can occur within the same study as long as all censoring is independent.

Censoring is not independent if those left under observation have systematically different failure times than those who are censored. The independent censoring assumption would be violated if, for example, individuals who are lost to follow-up are too sick to make it to their appointments. These are thus the individuals most likely to fail soon. Censoring in this case is said to be **informative (dependent)**. When censoring is dependent, Tsiatis (1975) showed that it is impossible to identify the distribution of T from the data without further assumptions.

In the figure below, people are censored but shortly afterwards fail. This is an example of dependent censoring because censoring carries prognostic information.



Part 5. Looking ahead

Next week, we will study how to compare survival curves using the log-rank test. We will begin analyzing survival data in R.