STAT 513 fa 2020 Lec 11

Censored data and survival analysis

Karl B. Gregory

Censored time-to-event data

- There is an entire field of statistics devoted to analyzing time-to-event data, for example, the time to failure of electronic components, the time to death of terminally ill patients, or the time to alleviation of symptoms after a treatment. An unavoidable feature of time-to-event data is *censoring*—when the event time is not observed for every component/patient/subject. This could happen, for example, if a subject left a study before the event took place or if the study terminated before all subjects experienced the event.

- **Form of censored data:** Let $T_1, \ldots, T_n$ be independent random variables which represent times to an event and let $C_1, \ldots, C_n$ be censoring times. Suppose we observe the data pairs

$$(Y_1, \delta_1), \ldots, (Y_n, \delta_n),$$

where $\delta_1, \ldots, \delta_n$ are indicator variables such that

$$Y_i = \begin{cases} T_i & \text{if } \delta_i = 1 \\ C_i & \text{if } \delta_i = 0 \end{cases} \qquad \text{for } i = 1, \ldots, n.$$

So for censored observations $\delta_i = 0$ and for uncensored observations $\delta_i = 1$.

- **Right-censoring:** If the censoring is such that the event of interest occurs *after* the time of censoring, we refer to it as right-censoring. In this case, we have

$$\delta_i = \begin{cases} 1 & \text{if } T_i \leq C_i \\ 0 & \text{if } T_i > C_i \end{cases} \qquad \text{for } i = 1, \ldots, n,$$

so that

$$Y_i = \min\{T_i, C_i\} \quad \text{for } i = 1, \ldots, n.$$

- **Left-censoring:** This is another type of censoring whereby the event of interest occurs *before* the censoring time; but to understand left-censoring it may be useful to drop the time-to-event language of "before" and "after". Consider the following example: Suppose a measurement device can only give reliable measurements when the true measurement is greater than some value C.

Then if the device returns a measurement less than C, we might choose to discard the measurement and simply conclude that the true measurement is some value less than C and to mark this observation as censored. In this setting we would have

$$\delta_i = \begin{cases} 1 & \text{if } T_i \geq C \\ 0 & \text{if } T_i < C, \end{cases} \qquad \text{for } i = 1, \ldots, n,$$

so that

$$Y_i = \max\{T_i, C\} \text{ for } i = 1, \ldots, n.$$

Note that censoring can be a feature of any kind of data, not just of time-to-event data.

- **Right-censoring example:** The following data is taken from [1], and shows times in remission in weeks for two groups of leukemia patients; the event of interest is coming out of remission.

Group 1 (Treatment)	Group 2 (Placebo)
6, 6, 6, 7, 10,	1, 1, 2, 2, 3,
13, 16, 22, 23,	4, 4, 5, 5,
6+, 9+, 10+, 11+,	8, 8, 8, 8,
17+, 19+, 20+,	11, 11, 12, 12,
25+, 32+, 32+,	15, 17, 22, 23
34+, 35+	

The numbers marked with a "+" are censored values. The "+" indicates that these patients either were still in remission at the end of the study, were lost to follow-up (could not be contacted), or withdrew from the study. We would set $\delta_i = 0$ for those observations marked with a "+" and set $\delta_i = 1$ for the remaining observations.

Maximum likelihood estimation with censored data

- Given data like the leukemia data in the previous section, if we were to ignore the censoring, our analyses would be biased. Suppose we wished to estimate the mean time in remission for the treatment group. If we simply averaged all of the observed values, it is likely that we would underestimate the mean, since the true remission times of the censored patients were greater than the numbers we have in the table. How can we proceed? We find that we can incorporate the censoring into the likelihood function in order construct reliable estimators via the maximum likelihood approach.

- In this section we assume that $T_1, \ldots, T_n$ are independent random variables with cdf $F(\cdot;\theta)$ and pdf $f(\cdot;\theta)$. Moreover, we define the *survival function* as $S(\cdot;\theta) = 1 - F(\cdot;\theta)$.

- **Right-censoring with random censoring times:** Suppose $C_1, \ldots, C_n$ are independent random variables with cdf G and pdf g; in addition, suppose $C_1, \ldots, C_n$ are independent of $T_1, \ldots, T_n$ and that G does not depend on the parameter θ. Then we find that the likelihood function of the observed data can be written

$$L(\theta; (Y_1, \delta_1), \ldots, (Y_n, \delta_n)) = \prod_{i \in \mathcal{U}} f(Y_i;\theta) \prod_{i \in \mathcal{C}} S(Y_i;\theta) \times K((Y_1, \delta_1), \ldots, (Y_n, \delta_n)), \tag{1}$$

where K is some function of the data which does not involve the parameter θ, and $\mathcal{U} = \{i : \delta_i = 1\}$ are the indices of the uncensored observations and $\mathcal{C} = \{i : \delta_i = 0\}$ are the indices of the uncensored observations.

Derivation of the likelihood: In a few steps, we may derive the conditional densities of $Y_i|\delta_i = 1$ and $Y_i|\delta_i = 0$ as

$$Y_i|\delta_i = 1 \sim f(y;\theta)[1 - G(y)]/P(T_i \leq C_i)$$
$$Y_i|\delta_i = 0 \sim g(y)[1 - F(y;\theta)]/P(T_i > C_i).$$

To construct the likelihood, we consider the "probability" of each pair (Y_i, δ_i), which is a little awkward because Y_i is continuous and δ_i is discrete. Ignoring this awkwardness and simply denoting by $p(Y_i, \delta_i)$ the joint "density" (which is not really a density or a probability mass function) of (Y_i, δ_i) and by $p(\delta_i)$ the marginal probability mass function of δ_i, we may write

$$p(Y_i, \delta_i) = p(Y_i|\delta_i)p(\delta_i) = \begin{cases} f(y;\theta)[1 - G(y)] & \text{when } \delta_i = 1 \\ g(y)[1 - F(y;\theta)] & \text{when } \delta_i = 0. \end{cases}$$

The quantity $p(Y_i, \delta_i)$ is the contribution to the likelihood of the data pair (Y_i, δ_i). So we see that the likelihood function can be written as

$$\begin{aligned} L(\theta; (Y_1, \delta_1), \ldots, (Y_n, \delta_n)) &= \prod_{i=1}^{n} \{f(Y_i;\theta)[1 - G(Y_i)]\}^{\delta_i} \{g(Y_i)[1 - F(Y_i;\theta)]\}^{1-\delta_i} \\ &= \prod_{i=1}^{n} f(Y_i;\theta)^{\delta_i}[1 - F(Y_i;\theta)]^{1-\delta_i} \underbrace{\prod_{i=1}^{n} g(Y_i)^{1-\delta_i}[1 - G(Y_i)]^{\delta_i}}_{=:K((Y_1,\delta_1),\ldots,(Y_n,\delta_n))} \\ &= \prod_{i\in\mathcal{U}} f(Y_i;\theta) \prod_{i\in\mathcal{C}} S(Y_i;\theta) \times K((Y_1, \delta_1), \ldots, (Y_n, \delta_n)), \end{aligned}$$

where we see that the function K does not involve the parameter θ.

- **Right-censoring with a fixed censoring time:** Suppose $C_1 = \cdots = C_n = C$ for some constant $C > 0$. Following similar steps we will get that the likelihood function of the observed data is

$$L(\theta; (Y_1, \delta_1), \ldots, (Y_n, \delta_n)) = \prod_{i\in\mathcal{U}} f(Y_i;\theta) \prod_{i\in\mathcal{C}} S(C;\theta) \times K((Y_1, \delta_1), \ldots, (Y_n, \delta_n)),$$

where K is some function of the data which does not involve the parameter θ. Note that all the censored data take the same value C.

- **Left-censoring with a fixed censoring time:** Suppose $C_1 = \cdots = C_n = C$ for some constant $C > 0$. Following similar steps we will get that the likelihood function of the observed data is

$$L(\theta; (Y_1, \delta_1), \ldots, (Y_n, \delta_n)) = \prod_{i\in\mathcal{U}} f(Y_i;\theta) \prod_{i\in\mathcal{C}} F(C;\theta) \times K((Y_1, \delta_1), \ldots, (Y_n, \delta_n)),$$

where K is some function of the data which does not involve the parameter θ. Note that all the censored data take the same value C.

- **Exercise:** Suppose you observe some censored data $(Y_1, \delta_1), \ldots, (Y_n, \delta_n)$ such that $Y_i = \min\{T_i, C_i\}$, where C_i is a random censoring time and $T_i \sim \text{Exponential}(\lambda)$, with C_i and T_i independent. The variable δ_i is defined such that $\delta_i = 1$ if $T_i \leq C_i$ and $\delta_i = 0$ otherwise.

  (i) Find an expression for the maximum likelihood estimator of λ based on the censored data.

  (ii) Assume for a moment that the times in remission of the treatment group in the leukemia data set follow an exponential distribution and compute the maximum likelihood estimator of the mean time in remission.

  Answers:

  (i) We have

$$f(Y_i; \lambda) = \frac{1}{\lambda} \exp\left[-\frac{Y_i}{\lambda}\right]$$
$$S(Y_i; \lambda) = 1 - F(Y_i; \lambda) = 1 - \left(1 - \exp\left[-\frac{Y_i}{\lambda}\right]\right) = \exp\left[-\frac{Y_i}{\lambda}\right].$$

  Plugging these into the likelihood expression (1), we have

$$\begin{aligned} L(\lambda; (Y_1, \delta_1), \ldots, (Y_n, \delta_n)) &\propto \prod_{i \in \mathcal{U}} \frac{1}{\lambda} \exp\left[-\frac{Y_i}{\lambda}\right] \prod_{i \in \mathcal{C}} \exp\left[-\frac{Y_i}{\lambda}\right] \\ &= \left(\frac{1}{\lambda}\right)^{n_U} \exp\left[-\frac{\sum_{i \in \mathcal{U}} Y_i}{\lambda}\right] \exp\left[-\frac{\sum_{i \in \mathcal{C}} Y_i}{\lambda}\right], \end{aligned}$$

  where we use the $\propto$ symbol in order to omit from the right-hand side the function K of the observed data which does not involve the parameter λ, and where n_U is the number of uncensored observations. Continuing to ignore K, the log-likelihood can be written

$$\ell(\lambda; (Y_1, \delta_1), \ldots, (Y_n, \delta_n)) = -n_U \log \lambda - \frac{\sum_{i \in \mathcal{U}} Y_i}{\lambda} - \frac{\sum_{i \in \mathcal{C}} Y_i}{\lambda}.$$

  Taking the derivative of this function with respect to λ and setting it equal to zero gives the maximum likelihood estimator of λ. We have

$$\frac{\partial}{\partial \lambda} \ell(\lambda; (Y_1, \delta_1), \ldots, (Y_n, \delta_n)) = -\frac{n_U}{\lambda} + \frac{\sum_{i \in \mathcal{U}} Y_i}{\lambda^2} + \frac{\sum_{i \in \mathcal{C}} Y_i}{\lambda^2},$$

  so that the maximum likelihood estimator of λ is given by

$$\hat{\lambda} = \frac{\sum_{i \in \mathcal{U}} Y_i + \sum_{i \in \mathcal{C}} Y_i}{n_U} = \frac{1}{n_U} \sum_{i=1}^{n} Y_i.$$

  (ii) The following R code computes the maximum likelihood estimator:

```
Y <- c(6,6,6,7,10,13,16,22,23,6,9,10,11,17,19,20,25,32,32,34,35)
d <- c(1,1,1,1,1,1,1,1,1,0,0,0,0,0,0,0,0,0,0,0,0)
nU <- sum(d)
lambda.hat <- sum(Y)/nU
```

  We get $\hat{\lambda} = 39.88889$.

Parametric estimation of the survival function

- If $T_1, \ldots, T_n$ come from a distribution with a cdf $F(\cdot;\theta)$ that is known up to a parameter θ, then the survival function is $S(\cdot;\theta) = 1 - F(\cdot;\theta)$. In order to estimate the survival function, we only need to estimate the parameter θ, with, for example, the maximum likelihood estimator $\hat{\theta}$. Then our estimator of the survival function is simply $S(\cdot;\hat{\theta})$.

- **Exercise:** Make a plot of the estimated survival functions for the treatment and placebo group of the leukemia data under the assumption that the times in remission follow Exponential(λ) distributions with $\lambda = \lambda_{\mathrm{T}}$ for the treatment group and $\lambda = \lambda_{\mathrm{P}}$ for the placebo group.

  Answer: From previous work, the maximum likelihood estimator of λ_{T} is $\hat{\lambda}_{\mathrm{T}} = 39.88889$. So the estimated survival function for the treatment group is given by

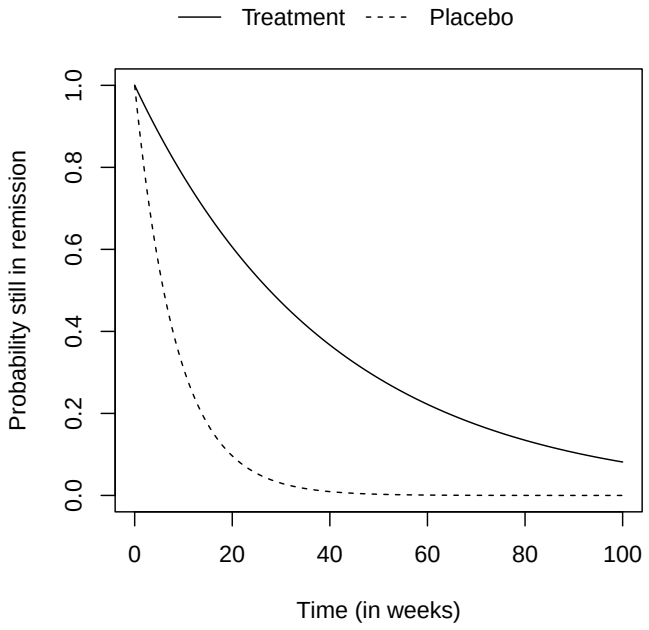
$$S_{\mathrm{T}}(t;\hat{\lambda}_{\mathrm{T}}) = \exp\left[-\frac{t}{39.88889}\right].$$

  There are no censored observations in the placebo group, so the maximum likelihood estimator of λ_{P} is $\hat{\lambda}_{\mathrm{P}} = 8.55$, which is the mean of the observed times in remission. So the estimated survival function for the placebo group is given by

$$S_{\mathrm{P}}(t;\hat{\lambda}_{\mathrm{P}}) = \exp\left[-\frac{t}{8.55}\right].$$

  The following R code makes a plot of the two survival functions:

```
Y <- c(6,6,6,7,10,13,16,22,23,6,9,10,11,17,19,20,25,32,32,34,35)
delta <- c(1,1,1,1,1,1,1,1,1,0,0,0,0,0,0,0,0,0,0,0,0)
nU <- sum(delta)
lambda.hat <- sum(Y)/nU

Y.P <- c(1,1,2,2,3,4,4,5,5,8,8,8,8,11,12,12,15,17,22,23)
lambda.P.hat <- mean(Y.P)

t.seq <- seq(0,100,length=200)
S.T <- exp( - t.seq / lambda.hat)
S.P <-  exp( - t.seq / lambda.P.hat)

plot(S.T ~ t.seq,type="l",lty=1,ylim=c(0,1),xlab="Time (in weeks)",
        ylab="P(still in remission)")
lines(S.P ~ t.seq,lty=2)

x.pos <- grconvertX(.5,from="nfc",to="user")
y.pos <- grconvertY(.95,from="nfc",to="user")
legend(x = x.pos, y = y.pos, legend=c("Treatment","Placebo"),
            lty=c(1,2),xpd=NA,bty="n",horiz=TRUE,xjust=.5)
```



Nonparametric estimation of the survival function

- We may not be comfortable assuming any specific distribution for the event times $T_1, \ldots, T_n$. In this case, we cannot simply estimate a parameter in order to get an estimator of the survival function. In this section we introduce some non-parametric estimators of the survival function; by "non-parametric" we mean that we do not try to estimate a parameter, but rather the values of the function itself.

- **Life-table estimator for leukemia data:** One non-parametric way to estimate the survival function is to break the observation period into intervals of equal length, and then to record for each interval the number of events that occurred in the interval, the number of subjects that were censored during the interval, and the number of subjects under observation at the beginning of the interval. The table can then be used to estimate the probability of occurrence of the event in each interval, and these probabilities can be used to construct an estimate of height of the survival function over each interval. For the treatment group of the Leukemia data, if the observation period is broken into 5-week intervals, we have the following *life table*:

Interval	# out of remission	# cens.	# uncens. at beginning of int.	$\hat{h}$	$\hat{S}$
$[0, 5)$	0	0	21	0/21	1.000
$[5, 10)$	4	2	21	4/21	0.810
$[10, 15)$	2	2	15	2/15	0.702
$[15, 20)$	1	2	11	1/11	0.638
$[20, 25)$	2	1	8	2/8	0.478
$[25, 30)$	0	1	5	0/5	0.478
$[30, 35)$	0	3	4	0/4	0.478
$[35, 40)$	0	1	1	0/1	0.478

The $\hat{h}$ column contains for each interval the estimated probability that a subject experiences the event in that interval. Each entry is given by the number of events occurring in the interval divided by the number of subjects under observation at the beginning of the interval. The "h" stands for "hazard", which can be regarded as the rate of occurrence of the event at a given time (in a given interval).

The $\hat{S}$ column contains the values of the estimated survival function over the intervals. The following gives a general form of the life table and explains how it is used to construct an estimator of the survival function.

- **General form of life table:** Suppose the observation period is broken into K intervals

$$[t_0, t_1), \ldots, [t_{K-1}, t_K), \quad \text{with} \quad 0 = t_0 < t_1 < \cdots < t_K.$$

For $k = 1, \ldots, K$, let

$d_k = \#$ observed "deaths" in the interval $[t_{k-1}, t_k)$,

$c_k = \#$ subjects censored during the interval $[t_{k-1}, t_k)$, and

$n_k = \#$ subjects "alive" and not yet censored at the beginning of the interval $[t_{k-1}, t_k)$.

The subjects which are "alive" and not yet censored at a given time are referred to as "subjects at risk", as they are still in the study and have not yet experienced the "death" event. With this notation the life table has the form

Interval	# deaths	# censored	# at risk	$\hat{h}$	$\hat{S}$
$[t_0, t_1)$	d_1	c_1	n_1	d_1/n_1	$1 - d_1/n_1$
$[t_1, t_2)$	d_2	c_2	n_2	d_2/n_2	$(1 - d_2/n_2)(1 - d_1/n_1)$
$[t_2, t_3)$	d_3	c_3	n_3	d_3/n_3	$\prod_{j=1}^{3}(1 - d_j/n_j)$
	$\vdots$				
$[t_{K-1}, t_K)$	d_K	c_K	n_K	d_K/n_K	$\prod_{j=0}^{K}(1 - d_j/n_j)$

To understand where the entries in the $\hat{h}$ and $\hat{S}$ columns come from, we note that for any $k = 1, \ldots, K$ we may write

$$\begin{aligned}
S(t_k) &= P(T > t_k) \\
&= P(T > t_k | T > t_{k-1}) P(T > t_{k-1}) \\
&= P(T > t_k | T > t_{k-1}) P(T > t_{k-1} | T > t_{k-2}) P(T > t_{k-2}) \\
&= \prod_{j=1}^{k} P(T > t_j | T > t_{j-1}) \times P(T > t_0) \\
&= \prod_{j=1}^{k} [1 - P(T \le t_j | T > t_{j-1})] \\
&= \prod_{j=1}^{k} [1 - P(t_{j-1} < T \le t_j | T > t_{j-1})] \\
&= \prod_{j=1}^{k} [1 - P(t_{j-1} \le T < t_j | T \ge t_{j-1})] \quad \text{(since } T \text{ is continuous)} \\
&= \prod_{j=1}^{k} (1 - h_j),
\end{aligned}$$

where we define

$$h_j = P(t_{j-1} \leq T < t_j | T \geq t_{j-1}), \quad \text{for } j = 1, \ldots, K.$$

One way to estimate $h_1, \ldots, h_K$ is with

$$\hat{h}_j = \frac{d_j}{n_j}, \quad \text{for } j = 1, \ldots, K.$$

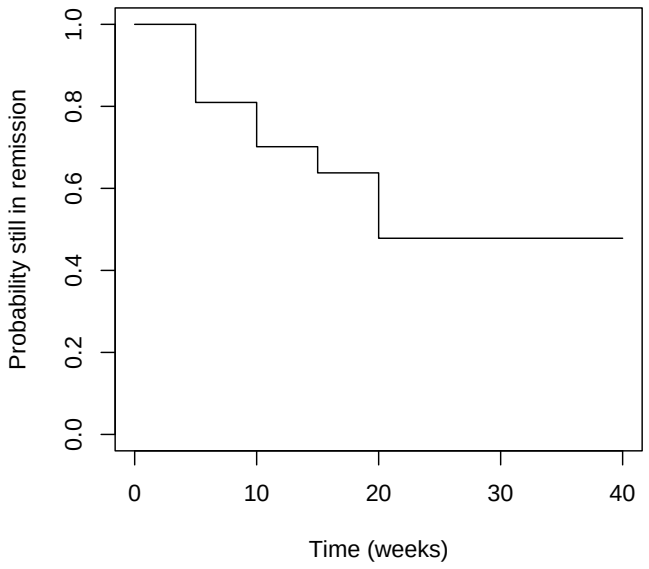
Then the life table estimator of S is defined as

$$\hat{S}(t) = \prod_{j=1}^{k} (1 - \hat{h}_j) \text{ for } t \in [t_{k-1}, t_k).$$

This life table estimator assumes that if a subject is censored in an interval, then the subject "survived" until the end of the interval.

The following R code computes the columns of the life table and plots the life table estimator of the survival function for the treatment group of the leukemia data:

```
Y <- c(6,6,6,7,10,13,16,22,23,6,9,10,11,17,19,20,25,32,32,34,35)
delta <- c(1,1,1,1,1,1,1,1,1,0,0,0,0,0,0,0,0,0,0,0,0)

# define intervals:
t <- seq(0,40,by=5)
K <- length(t) - 1

h <- numeric(K)
S <- numeric(K+1)
n <- numeric(K)
cens <- numeric(K)
S[1] <- 1
d <- numeric(K)
for( k in 1:K)
{

  # number of uncensored events in interval:
  d[k] <- sum((Y[which(delta==1)] >= t[k]) & ( Y[which(delta==1)] < t[k+1]))

  # number subjects censored in interval:
  cens[k] <- sum((Y[which(delta==0)] >= t[k]) & ( Y[which(delta==0)] < t[k+1]))

  # number uncensored subjects at beginning of interval:
  n[k] <- sum(Y >= t[k])

  # hazard estimate over interval:
  h[k] <- d[k]/n[k]

  # survival function estimate:
  S[k+1] <- S[k] * ( 1 - h[k] )

}

life.table <- cbind(t[1:K],t[-1],d,cens,n,h,S[-1])

plot(NA,xlim=range(t),ylim=c(0,1), xlab = "Time (weeks)",
        ylab = "Probability still in remission")
y.vals <- as.vector(t(cbind(S[-1],S[-1])))
x.vals <- as.vector(t(cbind(t,t)))[-c(1,2*(K+1))]
lines(x = x.vals, y = y.vals)
```



- **Kaplan-Meier estimator:** The Kaplan-Meier estimator is the life table estimator under a certain choice of the intervals $[t_0, t_1), \ldots, [t_{K-1}, t_K)$. Specifically, if the data $(Y_1, \delta_1), \ldots, (Y_n, \delta_n)$ are observed and $U_{(1)} < \cdots < U_{(K-1)}$ denote the unique uncensored event times, then the Kaplan-Meier estimator is the life table estimator based on the intervals defined by

$$\begin{aligned} t_0 &= 0 \\ t_1 = U_{(1)} &< \cdots < t_{K-1} = U_{(K-1)} \\ t_K &= \infty. \end{aligned}$$

This results in a life table like this one:

Interval	# deaths	# censored	# at risk	$\hat{h}$	$\hat{S}$
$[0, U_{(1)})$	d_1	c_1	n_1	d_1/n_1	$1 - d_1/n_1$
$[U_{(1)}, U_{(2)})$	d_2	c_2	n_2	d_2/n_2	$(1 - d_2/n_2)(1 - d_1/n_1)$
$[U_{(2)}, U_{(3)})$	d_3	c_3	n_3	d_3/n_3	$\prod_{j=1}^{3}(1 - d_j/n_j)$
	$\vdots$				
$[U_{(K-1)}, \infty)$	d_K	c_K	n_K	d_K/n_K	$\prod_{j=0}^{K}(1 - d_j/n_j)$

From the values in this table the Kaplan-Meier estimator can be expressed as

$$\hat{S}_{\text{KM}}(t) = \begin{cases} 1 & \text{for } t < U_{(1)} \\ \prod_{j:U_{(j)} \leq t} \left(1 - \frac{d_j}{n_j}\right) & \text{for } t \geq U_{(1)}. \end{cases}$$

The reason for the piecewise definition of the estimator is that $\hat{S}_{\text{KM}}(t)$ must take a value of 1 for $t \in [0, U_{(1)})$, which we can see by noting that we must have $d_1 = 0$, since no events may occur before the first one occurs! So we have $1 - d_1/n_1 = 1$.

- **Example:** The leukemia data contain tied event times. The Kaplan-Meier choices of $t_0 < t_1 < \cdots < t_K$ for the treatment group of the leukemia data are

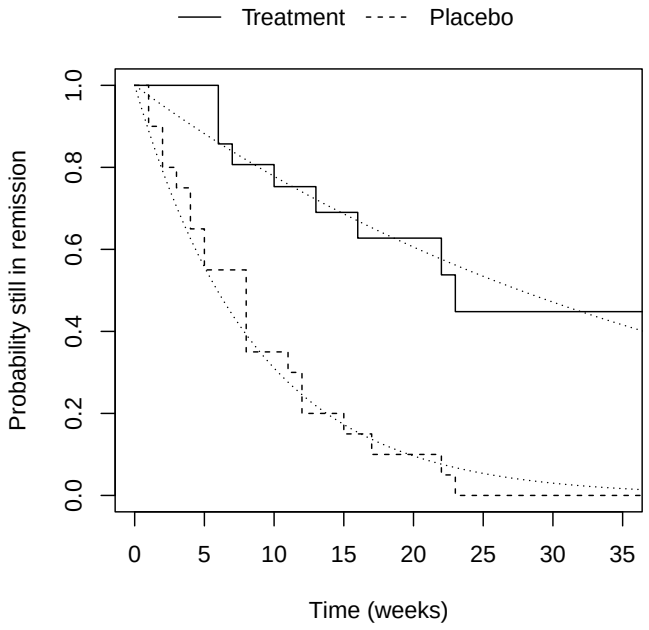
$$t_0 = 0,\ t_1 = 6,\ t_2 = 7,\ t_3 = 10,\ t_4 = 13,\ t_5 = 16,\ t_6 = 22,\ t_7 = 23,\ t_8 = \infty.$$

- **Exercise:** Compute the Kaplan-Meier estimator of the survival function for the two groups in the leukemia data set. Make a plot showing the two functions. In addition, overlay the parametric estimates of the survival functions obtained under the assumption that the times in remission follow exponential distributions.

  Answer: The following R code defines a function to compute the Kaplan-Meier life table. The function is applied to the data from the two groups of the leukemia study:

```
KM <- function(Y,delta)
{
  # define KM intervals:
  t <- c(0,sort(unique(Y[which(delta==1)])),2*max(Y))
  K <- length(t) - 1

  h <- numeric(K)
  S <- numeric(K+1)
  n <- numeric(K)
  cens <- numeric(K)
  S[1] <- 1
  d <- numeric(K)
  for( k in 1:K)
  {
    # number of uncensored events at beginning of interval:
    d[k] <- sum( (Y[which(delta==1)] >= t[k]) & ( Y[which(delta==1)] < t[k+1]) )

    # number subjects censored in interval:
    cens[k] <- sum( (Y[which(delta==0)] >= t[k]) & ( Y[which(delta==0)] < t[k+1]) )

    # number uncensored subjects at beginning of interval:
    n[k] <- sum(Y >= t[k])

    # hazard estimate over interval:
    h[k] <- d[k]/n[k]

    # survival function estimate:
    S[k+1] <- S[k] * ( 1 - h[k] )
  }

  life.table <- as.data.frame(cbind(t[1:K],d,cens,n,h,S[-1]))

  colnames(life.table) <- c("time","n.events","n.cens","n.risk","h","S")
```

```
  # output for making plots:
  y.vals <- as.vector(t(cbind(S[-1],S[-1])))
  x.vals <- as.vector(t(cbind(t,t)))[-c(1,2*(K+1))]

  output <- list( life.table = life.table,
                                    x.vals = x.vals,
                                    y.vals = y.vals )

  return(output)

}

Y <- c(6,6,6,7,10,13,16,22,23,6,9,10,11,17,19,20,25,32,32,34,35)
delta <- c(1,1,1,1,1,1,1,1,1,0,0,0,0,0,0,0,0,0,0,0,0)

Y.P <- c(1,1,2,2,3,4,4,5,5,8,8,8,8,11,12,12,15,17,22,23)
delta.P <- rep(1,length(Y.P))

KM.out <- KM(Y,delta)
KM.out.P <- KM(Y.P,delta.P)

plot(NA,xlim=range(t),ylim=c(0,1), xlab = "Time (weeks)",
            ylab = "Probability still in remission")
lines(x = KM.out$x.vals, y = KM.out$y.vals)
lines(x = KM.out.P$x.vals, y = KM.out.P$y.vals , lty=2)

x.pos <- grconvertX(.5,from="nfc",to="user")
y.pos <- grconvertY(.95,from="nfc",to="user")
legend(x = x.pos, y = y.pos, legend=c("Treatment","Placebo"),
lty=c(1,2),xpd=NA,bty="n",horiz=TRUE,xjust=.5)

lines( exp( - t.seq / lambda.hat) ~t.seq,lty=3)
lines( exp( - t.seq / lambda.P.hat) ~t.seq,lty=3)
```



References

[1] Emil J. Freireich, Edmund Gehan, Emil Frei, Leslie R. Schroeder, Irving J. Wolman, Rachad Anbari, E. Omar Burgert, Stephen D. Mills, Donald Pinkel, Oleg S. Selawry, et al. The effect of 6-mercaptopurine on the duration of steroid-induced remissions in acute leukemia: A model for evaluation of other potentially useful therapy. *Blood*, 21(6):699–716, 1963.