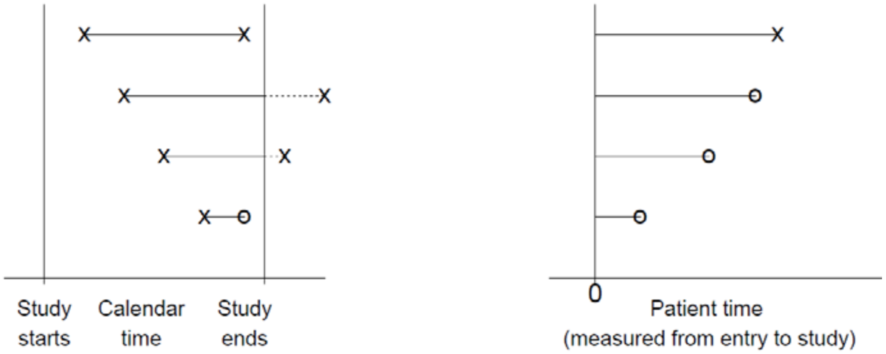
2. Right Censoring and Kaplan-Meier Estimator

Censored Data

Two important issues arise when studying "time to event" data (Event is assumed to be "death" unless specified otherwise):

1. Some individuals are still alive at the end of the study so the event of interest, namely death, has not occurred, which lead to right censored data.
2. Length of follow-up varies due to staggered entry.



Note: It is important to distinguish calendar time and patient time

Figure 2.1: *Illustration of censored data*

In addition to censoring because of insufficient follow-up (*i.e.*, end of study censoring due to staggered entry), other reasons for censoring includes

- loss to follow-up: patients stop coming to clinic or move away.
- deaths from other causes: competing risks.

Censoring from these types of causes may be inherently different from censoring due to staggered entry. We will discuss this in more detail later.

A Myocardial Infarction (MI) Example

Table 2.1: *Data from a clinical trial on MI*

Year since entry to study	Number alive and under observation at the beginning of interval	Number dying during interval	Number censored or withdrawal
[0, 1)	146	27	3
[1, 2)	116	18	10
[2, 3)	88	21	10
[3, 4)	57	9	3
[4, 5)	45	1	3
[5, 6)	41	2	11
[6, 7)	28	3	5
[7, 8)	20	1	8
[8, 9)	11	2	1
[0, 10)	8	2	6

Question: Estimate the 5 year survival rate, *i.e.*, $S(5) = P[T \geq 5]$.

Two Naïve and Incorrect Answers

1. $\hat{F}(5) = P[T < 5] = \frac{76\ deaths\ in\ 5\ years}{146\ individuals} = 52.1\%, \hat{S}(5) = 1 - \hat{F}(5) = 47.9\%.$

2. $\hat{F}(5) = P[T < 5] = \frac{76\ deaths\ in\ 5\ years}{146-29\ (withdrawn\ in\ 5\ years)} = 65\%, \hat{S}(5) = 1 - \hat{F}(5) = 35\%.$

It's obvious that:

1. The first estimate would be correct if all censoring occurred after 5 years. Of cause, this was not the case, leading to overly **optimistic** estimate (*i.e.*, overestimates $S(5)$).
2. The second estimate would be correct if all individuals censored in the 5 years were censored immediately upon entering the study. This was not the case either, leading to overly **pessimistic** estimate (*i.e.*, underestimates $S(5)$)

Note: Both estimates ignore the fact that each one-year interval experienced censoring (or withdrawing). Obviously we need to take this information into account in order to reduce bias.

Life-table Estimate

Idea: If the $S(5)$can be expressed as a function of quantities related to each interval and we get a very good estimate for each quantity, then intuitively, we will get a very good estimate of $S(5)$.

$$\begin{aligned}
S(5) = P[T \geq 5] &= P[T \geq 5 \cap T \geq 4] = P[T \geq 4] \cdot P[T \geq 5 | T \geq 4] \\
&= P[T \geq 4] \cdot \{1 - P[4 \leq T < 5 | T \geq 4]\} \\
&= P[T \geq 4] \cdot q_5 \\
&= P[T \geq 3] \cdot P[T \geq 4 | T \geq 3] \cdot q_5 \\
&= P[T \geq 3] \cdot \{1 - P[3 \leq T < 4 | T \geq 3]\} \cdot q_5 \\
&= P[T \geq 3] \cdot q_4 \cdot q_5 \\
&= \cdots \\
&= q_1 \cdot q_2 \cdot q_3 \cdot q_4 \cdot q_5
\end{aligned}$$

Where, $q_i = 1 - P[i - 1 \leq T < i | T \geq i - 1], i = 1,2, \ldots, 5.$

Note that $1 - q_i$is the mortality rate $m(x)$ at year $x = i - 1.$

Case 1: Assume that anyone censored in an interval of time is censored at the end of that interval.

Each $q_i = 1 - m(i-1)$ can be estimated in the following way:

$$d(0) \sim Bin\big(n(0), m(0)\big) \Rightarrow \hat{m}(0) = \frac{d(0)}{n(0)} = \frac{27}{146} = 0.185, \hat{q}_1 = 1 - \hat{m}(0) = 0.815$$

$$d(1)|H \sim Bin\big(n(1), m(1)\big) \Rightarrow \hat{m}(1) = \frac{d(1)}{n(1)} = \frac{18}{116} = 0.155, \hat{q}_2 = 1 - \hat{m}(1) = 0.845$$

....

Here, H stands for data history (i.e, data before the second interval).

The life table estimate for $S(5)$ would be computed as shown in Table below: **$\hat{S}^R(5)$=0.432**

Duration	$(x =$	$i - 1)$				
$[t_{i-1}, t_i)$	$n(x)$	$d(x)$	$w(x)$	$\hat{m}(x) = \frac{d(x)}{n(x)}$	$1 - \hat{m}(x)$	$\hat{S}^R(t_i) = \prod(1 - \hat{m}(x))$
[0, 1)	146	27	3	0.185	0.815	0.815
[1, 2)	116	18	10	0.155	0.845	0.689
[2, 3)	88	21	10	0.239	0.761	0.524
[3, 4)	57	9	3	0.158	0.842	0.441
[4, 5)	45	1	3	0.022	0.978	0.432

Estimate Variance of $\hat{S}^R(5)$

Notice that: $\log\left(\hat{S}^R(5)\right) = \log(\hat{q}_1) + \log(\hat{q}_2) + \log(\hat{q}_3) + \log(\hat{q}_4) + \log(\hat{q}_5)$

Let $\hat{\phi}_i = \log(\hat{q}_i)$, $\text{var}\{\log\left(\hat{S}^R(5)\right)\} = \underbrace{\sum_{i=1}^{5} \text{var}\left(\hat{\phi}_i\right)}_{(1)} + \underbrace{\sum_{i\neq j} \text{cov}(\hat{\phi}_i, \hat{\phi}_j)}_{(2)}$

(1). By CLT, $(\hat{q}_i - q_i) \overset{d}{\rightarrow} N\left(0, \text{var}(\hat{q}_i)\right)$

$$\text{var}(\hat{q}_i) = \text{var}\{\hat{m}(i-1)\} = E\left(\text{var}(\hat{m}(i-1)|H)\right) + \text{var}\left(E(\hat{m}(i-1)|H)\right)$$

$$= E\left(\frac{m(i-1)[1-m(i-1)]}{n(i-1)}\right) + \text{var}\left(m(i-1)\right)$$

$$= m(i-1)[1-m(i-1)]\, E\left(\frac{1}{n(i-1)}\right)$$

So, $\widehat{\text{var}}(\hat{q}_i) \approx \hat{m}(i-1)[1-\hat{m}(i-1)]\frac{1}{n(i-1)}$

By delta methods, $\text{var}\left(\hat{\phi}_i\right) = \left(\frac{1}{q_i}\right)^2 \text{var}(\hat{q}_i)$, so

$$\widehat{\text{var}}\left(\hat{\phi}_i\right) = \left(\frac{1}{\hat{q}_i}\right)^2 \widehat{\text{var}}(\hat{q}_i) = \frac{\hat{m}(i-1)}{[1-\hat{m}(i-1)]n(i-1)} = \frac{d(i-1)}{[n(i-1)-d(i-1)]n(i-1)}$$

(2). It is very amazing that the covariance terms are all approximately equal to zero. For example, let us consider the covariance between $\hat{\phi}_1$ and $\hat{\phi}_2$.

$$
\begin{aligned}
E[\widehat{m}(0)\widehat{m}(1)] &= E\{E[\widehat{m}(0)\widehat{m}(1)|n(0), d(0), w(0)]\} \\
&= E\{\widehat{m}(0)E[\widehat{m}(1)|n(0), d(0), w(0)]\} \\
&= E\{\widehat{m}(0)m(1)\} \\
&= m(1)E\{\widehat{m}(0)\} \\
&= m(1)m(0) \\
&= E[\widehat{m}(0)]\ E[\widehat{m}(1)]
\end{aligned}
$$

So, $\operatorname{cov}\big(\widehat{m}(0), \widehat{m}(1)\big) = 0$ Therefore, $\operatorname{cov}\big(\hat{\phi}_i, \hat{\phi}_j\big) = 0$ (Why?)

Now, let's put everything together:

Let $\hat{\theta} = \log\big(\hat{S}^R(5)\big)$, then $\hat{S}^R(5) = e^{\widehat{\theta}}$, so

$$\operatorname{var}\{\hat{S}^R(5)\} = \left(e^{\theta}\right)^2 \operatorname{var}\{\log\big(\hat{S}^R(5)\big)\} = \big(S(5)\big)^2 \sum_{i=1}^{5} \operatorname{var}\big(\hat{\phi}_i\big)$$

The variance of $\hat{S}^R(5)$ can be estimated by:

$$\widehat{\operatorname{var}}\{\hat{S}^R(5)\} = \big(\hat{S}^R(5)\big)^2 \sum_{i=1}^{5} \widehat{\operatorname{var}}\big(\hat{\phi}_i\big) = \big(\hat{S}^R(5)\big)^2 \sum_{i=1}^{5} \frac{d(i-1)}{[n(i-1) - d(i-1)]n(i-1)}$$

Case 2: Assume that anyone censored in an interval of time is censored at the beginning of that interval.

Similarly, the life table estimate for $S(5)$ would be computed as shown in Table below:
$\hat{\boldsymbol{S}}^{L}(5)$=**0.400**

Duration $[\boldsymbol{t}_{i-1}, \boldsymbol{t}_i)$	$n(x)$	$d(x)$	$w(x)$	$\hat{m}(x) = \frac{d(x)}{n(x) - w(x)}$	$1 - \hat{m}(x)$	$\hat{S}^R(\boldsymbol{t}_i) = \prod (1 - \hat{m}(x))$
[0, 1)	146	27	3	0.189	0.811	0.811
[1, 2)	116	18	10	0.170	0.830	0.673
[2, 3)	88	21	10	0.269	0.731	0.492
[3, 4)	57	9	3	0.167	0.833	0.426
[4, 5)	45	1	3	0.024	0.976	0.417

The variance estimate of $\hat{\boldsymbol{S}}^{L}(5)$ is similar to that of $\hat{\boldsymbol{S}}^{R}(5)$ except that we need to change The “sample size” $n(x)$ for each mortality estimate to $n(x) - w(x)$ in equation

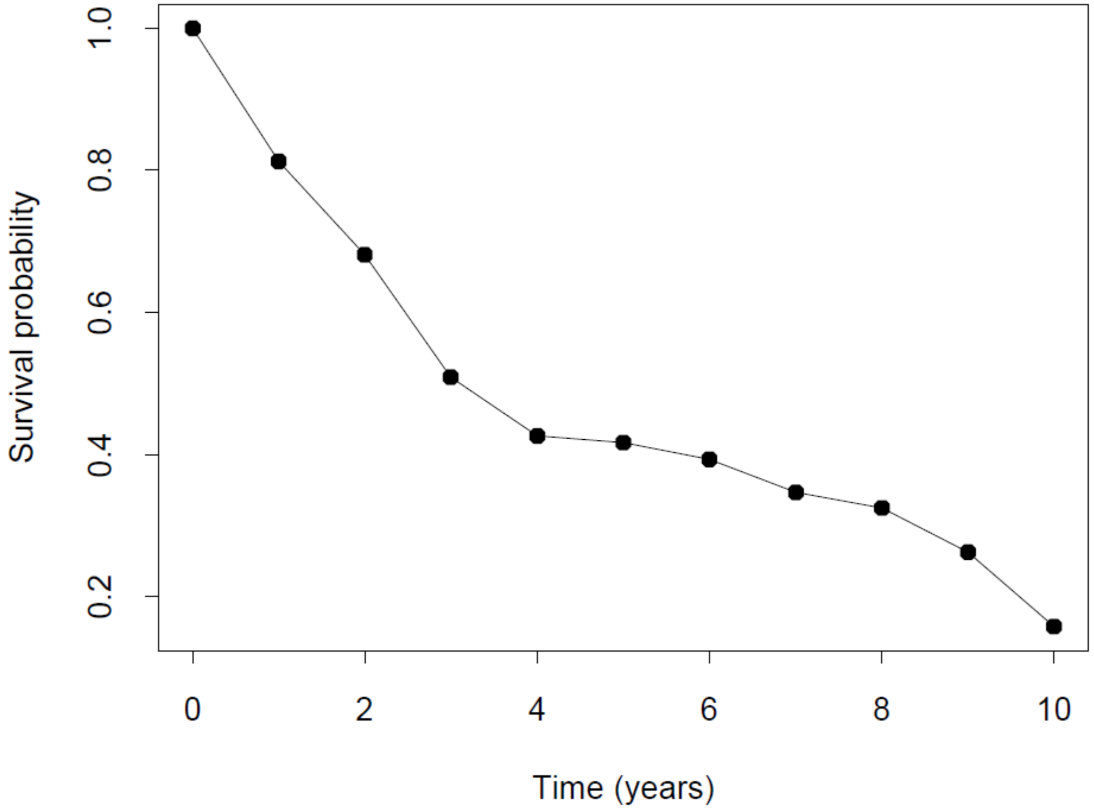
Case 3: Assume that anyone censored in an interval of time is censored during that interval.

More than likely censoring occurs during the interval. Thus $\widehat{\boldsymbol{S}}^{\boldsymbol{L}}$ and b $\widehat{\boldsymbol{S}}^{\boldsymbol{R}}$ are not correct. A compromise is to use $(n(x) - w(x)/2)$ as the “sample size”, which is referred to as the *effective sample size*. Then The life table estimate for $S(5)$ would be computed as shown in Table below: $\widehat{\boldsymbol{S}}^{\boldsymbol{LT}}$**(5)=0.417**.

Duration $[\boldsymbol{t}_{\boldsymbol{i}-\mathbf{1}}, \boldsymbol{t}_{\boldsymbol{i}})$	$n(x)$	$d(x)$	$w(x)$	$\widehat{m}(x) = \frac{d(x)}{n(x) - w(x)/2}$	$1 - \widehat{m}(x)$	$\hat{S}^R(\boldsymbol{t}_{\boldsymbol{i}}) = \prod(1 - \widehat{m}(x))$
[0, 1)	146	27	3	0.187	0.813	0.813
[1, 2)	116	18	10	0.162	0.838	0.681
[2, 3)	88	21	10	0.253	0.747	0.509
[3, 4)	57	9	3	0.162	0.838	0.410
[4, 5)	45	1	3	0.023	0.977	0.400

$$\widehat{\mathrm{var}}\{\hat{S}^{LT}(5)\} = \left(\hat{S}^{LT}(5)\right)^2 \sum_{i=1}^{5} \frac{d(i-1)}{\left[n(i-1) - \frac{w(i-1)}{2} - d(i-1)\right]\left[n(i-1) - \frac{w(i-1)}{2}\right]}$$

Assuming censoring occurred during interval and using the $\hat{S}^{LT}$, the life-table estimates of the survival probability for MI data are shown below:



From this figure, the median survival time is estimated to be about 3 years.

Calculation: Proc Lifetest in SAS **OR** lifetab {KMsurv} in R

R code:

```
> tis <- 0:10
> ninit <- 146
> nlost <- c(3,10,10,3,3,11,5,8,1,6)
> nevent <- c(27,18,21,9,1,2,3,1,2,2)
> lifetab(tis, ninit, nlost, nevent)
```

Output:

```
       nsubs  nlost  nrisk  nevent  surv      pdf       hazard    se.surv   se.pdf    se.hazard
0-1    146    3      144.5  27      1         0.186851  0.206107  0         0.032426  0.039454
1-2    116    10     111    18      0.813149  0.131862  0.176471  0.032426  0.028931  0.041432
2-3    88     10     83     21      0.681287  0.172374  0.289655  0.039337  0.034     0.062542
3-4    57     3      55.5   9       0.508913  0.082526  0.176471  0.043822  0.026163  0.058594
4-5    45     3      43.5   1       0.426387  0.009802  0.023256  0.04452   0.009743  0.023254
5-6    41     11     35.5   2       0.416585  0.02347   0.057971  0.044563  0.016316  0.040974
6-7    28     5      25.5   3       0.393115  0.046249  0.125     0.045037  0.025635  0.072028
7-8    20     8      16     1       0.346866  0.021679  0.064516  0.046992  0.021195  0.064483
8-9    11     1      10.5   2       0.325187  0.06194   0.210526  0.0488    0.040488  0.148038
9-10   8      6      5      2       0.263247  NA        NA        0.055799  NA        NA
```

Note: Here the numbers in the column of hazard are the estimated hazard rates at the midpoint of each interval by assuming the true survival function *S*(*t*) is a straight line in each interval. You can find an explicit expression for this estimator using the relation

$$\lambda(t) = \frac{f(t)}{S(t)},$$

and the assumption that the true survival function *S*(*t*) is a straight line in $[t_{i-1}, t_i)$:

$$S(t) = S(t_{i-1}) + \frac{S(t_i) - S(t_{i-1})}{t_i - t_{i-1}}(t - t_{i-1}), \text{ for } t \in [t_{i-1}, t_i)$$

These estimates are very close to the mortality estimates we obtained before (the column under Conditional Probability of Failure in the SAS output.)

Kaplan-Meier Estimator

The **Kaplan-Meier** or **product limit** estimator is the limit of the life-table estimator when intervals are taken so small that only at most one distinct observation occurs within an interval. Kaplan and Meier demonstrated in a paper in JASA (1958) that this estimator is “maximum likelihood estimate”. By convention, the Kaplan-Meier estimate is a **right continuous** step function which takes jumps only at the death time.

Let $d(x)$ denote the number of deaths at time *x*. Generally $d(x)$ is either zero or one, but we allow the possibility of tied survival times in which case $d(x)$ may be greater than one. Let $n(x)$ denote the number of individuals at risk just prior to time x; *i.e.*, number of individuals in the sample who neither died nor were censored prior to time x. Then Kaplan-Meier estimate can be expressed as

$$KM(t) = \prod_{x \le t} \left(1 - \frac{d(x)}{n(x)}\right)$$

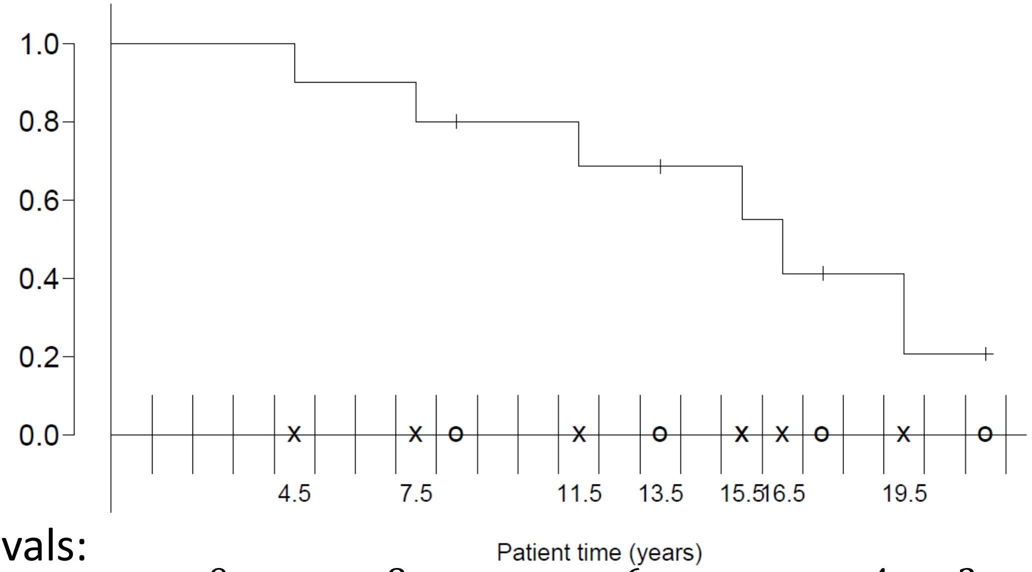
Note: In the notation above, the product changes only at times x where $d(x) \geq 1$, *i.e.*, only at times where we observed deaths.

Calculation: Proc Lifetest in SAS **OR** survfit {survival} in R

EX:

R code:

```
> survtime <- c(4.5, 7.5, 8.5, 11.5, 13.5, 15.5, 16.5, 17.5, 19.5, 21.5)
> status <- c(1, 1, 0, 1, 0, 1, 1, 0, 1, 0)
> fit <- survfit(Surv(survtime, status), conf.type=c("plain"))
> summary(fit)
Call: survfit(formula = Surv(survtime, status), conf.type = c("plain"))
time   n.risk   n.event   survival   std.err   lower 95% CI   upper 95% CI
 4.5      10        1       0.900    0.0949        0.7141          1.000
 7.5       9        1       0.800    0.1265        0.5521          1.000
11.5       7        1       0.686    0.1515        0.3888          0.983
15.5       5        1       0.549    0.1724        0.2106          0.887
16.5       4        1       0.411    0.1756        0.0673          0.756
19.5       2        1       0.206    0.1699        0.0000          0.539
```



At those intervals:

$$1-\widehat{m}(t):\quad 1\quad 1\quad 1\quad 1\quad \frac{9}{10}\quad 1\quad 1\quad \frac{8}{9}\quad 1\quad 1\quad 1\quad \frac{6}{7}\quad 1\quad 1\quad 1\quad \frac{4}{5}\quad \frac{3}{4}\quad 1\quad 1\quad \frac{1}{2}\quad 1\quad 1$$

$$\hat{S}(t):\quad 1\quad 1\quad 1\quad 1\quad \frac{9}{10}\quad .\quad .\quad \frac{8}{10}\quad .\quad .\quad .\quad \frac{48}{70}\quad .\quad .\quad .\quad \frac{192}{350}\quad \frac{144}{350}\quad .\quad .\quad \frac{144}{700}\quad .\quad .$$

Non-informative Censoring

In order that the life-table estimates give unbiased results, there is an important assumption that individuals who are censored are at the same risk of subsequent failure as those who are still alive and uncensored. The risk set at any time point (the individuals still alive and uncensored) should be representative of the entire population alive at the same time. If this is the case, the censoring process is called **non-informative**.

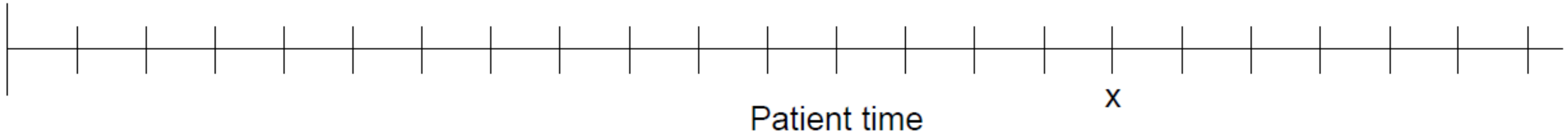
Statistically, if the censoring process is **independent** of the survival time, then we will automatically have non-informative censoring. Actually, we almost always mean independent censoring by non-informative censoring.

If censoring only occurs because of staggered entry, then the assumption of non-informative censoring seems plausible. However, when censoring results from loss to follow-up or death from a competing risk, then this assumption is more suspect. If at all possible censoring from these later situations should be kept to a minimum.

The variance of $\boldsymbol{KM(t)}$

- The derivation given here is heuristic in nature but will try to capture some of the salient features of the more rigorous treatments in more theoretical literatures on survival analysis.

First, let's partition time into many small intervals, say, with interval length equal to Δx, where Δx is small.



Let "x" denote some arbitrary time point on the grid above and define:

- $\{x < t\}$ means $\{\text{all grid potins such that } x + \Delta x \leq t\}$
- $Y(x) =$ number of individuals at risk (*i.e.*, alive and uncensored) at time point *x*.
- $dN(x) =$ number of observed deaths occurring in $[x, x + \Delta x)$.
- $w(x) =$ number of censored individuals in $[x, x + \Delta x)$,

Recall: Previously, $Y(x)$ was denoted by $n(x)$ and $dN(x)$ was denoted by $d(x)$.
It should be straightforward to see that $w(x) = \{[Y(x) - Y(x + \Delta x)] - \Delta x\}$.

<u>Note</u>: In theory, we should be able to choose Δx small enough so that $\{dN(x) > 0 \text{ and } w(x) > 0\}$ should never occur. In practice, however, data may not be collected in that fashion, in which case, approximations such as those given with life-table estimators may be necessary (e.g. effective sample size).

Now, the Kaplan-Meier estimator can be written as

$$KM(t) = \prod_{x<t}\left(1 - \frac{dN(x)}{Y(x)}\right) \quad \text{as } \Delta x \to 0$$

When "Δx" is not small enough, $KM(t)$ can be modified to be

$$LT(t) = \prod_{x<t}\left(1 - \frac{dN(x)}{Y(x) - w(x)/2}\right)$$

Notice that if the sample size is large and Δx is small, then $\frac{dN(x)}{Y(x)}$ is a small number (*i.e.*, close to zero) and as long as x is not close to the right hand tail of the survival distribution (where $Y(x)$ may be very small). If this is the case, then

$$1 - \frac{dN(x)}{Y(x)} \approx e^{-\frac{dN(x)}{Y(x)}},$$

and $KM(t) = \prod_{x<t}\left(1 - \frac{d(x)}{n(x)}\right) \approx \prod_{x<t} e^{-\frac{dN(x)}{Y(x)}} = e^{-\sum_{x<t}\frac{dN(x)}{Y(x)}}$.

If Δx is taken to be small enough so that all distinct times (either death times or withdrawal times) are represented at most once in any time interval, then the estimator $\sum_{x<t}\frac{dN(x)}{Y(x)}$ will be uniquely defined and will not be altered by choosing a *finer* partition for the grid of time points. In such a case the quantity $\sum_{x<t}\frac{dN(x)}{Y(x)}$ is sometimes represented as $\int_0^t \frac{dN(x)}{Y(x)}$.

$\sum_{x<t}\frac{dN(x)}{Y(x)}$ is referred to as the *Nelson-Aalen estimator* for the cumulative hazard function $\Lambda(t) = \int_0^t \lambda(x)\, dx$. That is, $\widehat{\Lambda}(t) = \sum_{x<t}\frac{dN(x)}{Y(x)}$. (Recall $S(t) = e^{-\Lambda(t)}$)

Another Way to Look at the $\sum_{x<t} \frac{dN(x)}{Y(x)}$

- Is basically the sum over all the distinct death times before time *t* of the number of deaths divided by the number at risk at each of those distinct death times.

Notice that $\Lambda(t) = \int_0^t \lambda(x)\, dx \approx \sum_{x<t} \lambda(x)\Delta x$

By the definition of a hazard function,

$$\lambda(x)\Delta x \approx P[x \leq T < x + \Delta x | T \geq x]$$

With independent censoring, it would seem reasonable to estimate $\lambda(x)\Delta x$, *i.e.*, "the conditional probability of dying in $[x, x + \Delta x)$ given being alive at time x" by $\frac{dN(x)}{Y(x)}$.

Therefore we obtain the Nelson-Aalen estimator:

$$\widehat{\Lambda}(t) = \sum_{x<t} \frac{dN(x)}{Y(x)}$$

Goal: show how to estimate the variance of $\widehat{\Lambda}(t)$and then show how it can be used to estimate the variance of the Kaplan-Meier estimator $KM(t)$.

$\widehat{\Lambda}(t)$ is nearly unbiased to $\Lambda(t)$

For a grid point x, denote the history of all deaths and censoring occurring up to time x as

$$H(x) = \{dN(u), w(u); \text{for all value } u \text{ on the grid of points for } u < x\}.$$

Then, we have

$$dN(x)|H(x) \sim \text{Bin}\big(Y(x), \pi(x)\big),$$

Known

which implies:

$$E[dN(x)|H(x)] = Y(x)\pi(x) \Longrightarrow E\left[\frac{dN(x)}{Y(x)}\middle| H(x)\right] = \pi(x)$$

$$Var[dN(x)|H(x)] = Y(x)\pi(x)[1-\pi(x)]$$

$$E\left\{\left[\frac{Y(x)}{Y(x)-1}\right]\left[\frac{dN(x)}{Y(x)}\right]\left[\frac{Y(x)-dN(x)}{Y(x)}\right]\middle| H(x)\right\} = \pi(x)[1-\pi(x)]$$

Consider the Nelson-Aalen estimator $\widehat{\Lambda}(t)$:

$$E\big[\widehat{\Lambda}(t)\big] = E\left[\sum_{x<t}\frac{dN(x)}{Y(x)}\right] = \sum_{x<t} E\left[\frac{dN(x)}{Y(x)}\right] = \sum_{x<t} E\left[E\left(\frac{dN(x)}{Y(x)}\middle| H(x)\right)\right]$$

$$= \sum_{x<t}\pi(x) \approx \sum_{x<t}\lambda(x)\Delta x \approx \int_0^t \lambda(x)\,dx = \Lambda(t)$$

Estimate Variance of $\widehat{\Lambda}(t)$

$$\mathrm{Var}\left(\widehat{\Lambda}(t)\right) = E\left[\widehat{\Lambda}(t) - E\left(\widehat{\Lambda}(t)\right)\right]^2 = E\left[\sum_{x<t} \frac{dN(x)}{Y(x)} - \sum_{x<t} \pi(x)\right]^2 = E\left[\sum_{x<t} \left\{\frac{dN(x)}{Y(x)} - \pi(x)\right\}\right]^2$$

$$= E\left[\sum_{x<t} \left\{\frac{dN(x)}{Y(x)} - \pi(x)\right\}^2 + \sum_{x \neq x' < t} \left\{\frac{dN(x)}{Y(x)} - \pi(x)\right\}\left\{\frac{dN(x')}{Y(x')} - \pi(x')\right\}\right]$$

$$= \sum_{x<t} E\left[\frac{dN(x)}{Y(x)} - \pi(x)\right]^2 + \sum_{x \neq x' < t} E\left[\left\{\frac{dN(x)}{Y(x)} - \pi(x)\right\}\left\{\frac{dN(x')}{Y(x')} - \pi(x')\right\}\right]$$

For the first term:

$$E\left[\frac{dN(x)}{Y(x)} - \pi(x)\right]^2 = E\left[E\left[\left\{\frac{dN(x)}{Y(x)} - \pi(x)\right\}^2 \middle| H(x)\right]\right] = E\left[Var\left[\frac{dN(x)}{Y(x)} \middle| H(x)\right]\right] = E\left[\frac{\pi(x)[1-\pi(x)]}{Y(x)}\right]$$

For the second term, W.L.O.G., suppose $x < x'$,

$$E\left[\left\{\frac{dN(x)}{Y(x)} - \pi(x)\right\}\left\{\frac{dN(x')}{Y(x')} - \pi(x')\right\}\right] = E\left[E\left[\left\{\frac{dN(x)}{Y(x)} - \pi(x)\right\}\left\{\frac{dN(x')}{Y(x')} - \pi(x')\right\} \middle| H(x')\right]\right]$$

$$= E\left[\left\{\frac{dN(x)}{Y(x)} - \pi(x)\right\} E\left[\left\{\frac{dN(x')}{Y(x')} - \pi(x')\right\} \middle| H(x')\right]\right] = 0$$

↖Why?

Therefore, $\mathrm{Var}\left(\widehat{\Lambda}(t)\right) = \sum_{x<t} E\left[\frac{\pi(x)[1-\pi(x)]}{Y(x)}\right]$

It is reasonable to use $\frac{\frac{dN(x)}{Y(x)}\left[\frac{Y(x)-dN(x)}{Y(x)}\right]}{Y(x)-1}$ to estimate $\frac{\pi(x)[1-\pi(x)]}{Y(x)}$

Therefore, an estimator for $\mathrm{Var}\left(\widehat{\Lambda}(t)\right)$ is

$$\widehat{\mathrm{Var}}\left(\widehat{\Lambda}(t)\right) = \sum_{x<t}\left[\frac{\frac{dN(x)}{Y(x)}\left[\frac{Y(x)-dN(x)}{Y(x)}\right]}{Y(x)-1}\right]$$

In fact, $\widehat{\mathrm{Var}}\left(\widehat{\Lambda}(t)\right)$ is unbiased:

$$E\left[\widehat{\mathrm{Var}}\left(\widehat{\Lambda}(t)\right)\right] = E\left[\sum_{x<t}\left[\frac{\frac{dN(x)}{Y(x)}\left[\frac{Y(x)-dN(x)}{Y(x)}\right]}{Y(x)-1}\right]\right] = \sum_{x<t} E\left[\frac{\frac{dN(x)}{Y(x)}\left[\frac{Y(x)-dN(x)}{Y(x)}\right]}{Y(x)-1}\right]$$

$$= \sum_{x<t} E\left[E\left[\frac{\frac{dN(x)}{Y(x)}\left[\frac{Y(x)-dN(x)}{Y(x)}\right]}{Y(x)-1}\middle| H(x)\right]\right] = \sum_{x<t} E\left[\frac{\pi(x)[1-\pi(x)]}{Y(x)}\right] = \mathrm{Var}\left(\widehat{\Lambda}(t)\right)$$

Note: If the survival data are continuous (*i.e.*, no ties) and Δx is taken small enough, then $dN(x)$ would take on the values 0 or 1 only. In this case:

$$\frac{dN(x)}{Y(x)}\left[\frac{Y(x)-dN(x)}{Y(x)}\right] = \frac{dN(x)}{Y^2(x)} \quad \text{and} \quad \widehat{\mathrm{Var}}\left(\widehat{\Lambda}(t)\right) = \sum_{x<t}\frac{dN(x)}{Y^2(x)}$$

Remark:

Since $\widehat{\Lambda}(t) = \sum_{x<t} \frac{dN(x)}{Y(x)}$ is made up of a sum of random variables that are conditionally uncorrelated, they have a “martingale” structure for which there exists a body of theory that enables us to show that $\widehat{\Lambda}(t)$ is asymptotically normal with mean $\Lambda(t)$ and variance $\mathrm{Var}\left(\widehat{\Lambda}(t)\right)$, which can be estimated by $\widehat{\mathrm{Var}}\left(\widehat{\Lambda}(t)\right)$.

Refer to the estimated standard error $\widehat{\Lambda}(t)$ of by

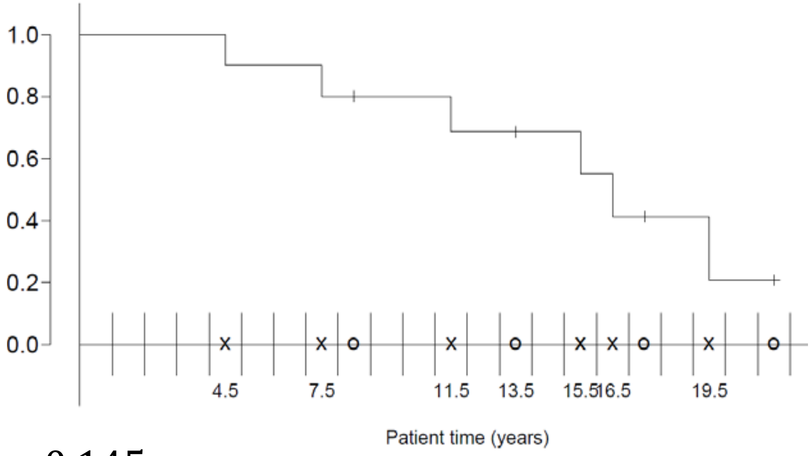
$$se\left(\widehat{\Lambda}(t)\right) = \left[\sum_{x<t}\left[\frac{\frac{dN(x)}{Y(x)}\left[\frac{Y(x)-dN(x)}{Y(x)}\right]}{Y(x)-1}\right]\right]^{1/2}$$

The unbiasedness and asymptotic normality of $\widehat{\Lambda}(t)$ about $\Lambda(t)$ allow us to form confidence intervals for $\Lambda(t)$ (at time t). Specifically, the $(1-\alpha)$th confidence interval for $\Lambda(t)$ is given by $\widehat{\Lambda}(t) \pm z_{\alpha/2} * se\left(\widehat{\Lambda}(t)\right)$

Since $S(t) = e^{-\Lambda(t)}$, this result could also be used to construct $(1-\alpha)$th confidence intervals for the survival function $S(t)$, which is given by

$$\left[e^{-\widehat{\Lambda}(t)-z_{\alpha/2}*se\left(\widehat{\Lambda}(t)\right)}, e^{-\widehat{\Lambda}(t)+z_{\alpha/2}*se\left(\widehat{\Lambda}(t)\right)}\right].$$

An Example



For illustration, let us take $t = 17$. Note that there are no ties in this example.
So,

$$\widehat{\Lambda}(t) = \sum_{x<t} \frac{dN(x)}{Y(x)} = \frac{1}{10} + \frac{1}{9} + \frac{1}{7} + \frac{1}{5} + \frac{1}{4} = 0.804$$

$$\widehat{\text{Var}}\left(\widehat{\Lambda}(t)\right) = \sum_{x<t} \frac{dN(x)}{Y^2(x)} = \frac{1}{10^2} + \frac{1}{9^2} + \frac{1}{7^2} + \frac{1}{5^2} + \frac{1}{4^2} = 0.145$$

95% C.I. for $\Lambda(t)$ is: $0.804 \pm 1.96 * 0.381 = [0.0572, 1.551]$

The Nelson-Aalen estimate of $S(t)$ is: $\hat{S}(t) = e^{-\widehat{\Lambda}(t)} = e^{-0.804} = 0.448$

95% C.I. for $S(t)$ is: $[e^{-1.551}, e^{-0.0572}] = [0.212, 0.944]$

Note:

- The above Nelson-Aalen estimate $\hat{S}(t) = 0.448$ is different from (but close to) the Kaplan-Meier estimate $\text{KM}(t) = 0.411$.
- The above confidence interval for the survival probability $S(t)$ is not symmetric about the estimator $\hat{S}(t)$.

Symmetric C.I. for $\hat{S}(t)$

$$\left(\hat{\Lambda}(t) - \Lambda(t)\right) \overset{d}{\rightarrow} N\left(0, \mathrm{Var}\left(\hat{\Lambda}(t)\right)\right)$$

By delta method,

$$\left(e^{-\hat{\Lambda}(t)} - e^{-\Lambda(t)}\right) \overset{d}{\rightarrow} N\left(0, \left(e^{-\Lambda(t)}\right)^2 \mathrm{Var}\left(\hat{\Lambda}(t)\right)\right)$$

i.e.

$$\left(\hat{S}(t) - S(t)\right) \overset{d}{\rightarrow} N\left(0, [S(t)]^2 \mathrm{Var}\left(\hat{\Lambda}(t)\right)\right)$$

Then a reasonable estimate for $[S(t)]^2 \mathrm{Var}\left(\hat{\Lambda}(t)\right)$ is $\left[\hat{S}(t)\right]^2 \widehat{Var}\left(\hat{\Lambda}(t)\right)$

The $(1-\alpha)$th confidence intervals for the survival function $S(t)$: $\hat{S}(t) \pm \hat{S}(t)^{*} se\left(\hat{\Lambda}(t)\right)$

Remark:

Note that $\left[\hat{S}(t)\right]^2 \mathrm{Var}\left(\hat{\Lambda}(t)\right)$ is an estimate of Var $\left(\hat{S}(t)\right)$, where $\hat{S}(t) = e^{-\hat{\Lambda}(t)}$. Previously, we showed that the Kaplan-Meier estimator $KM(t) = \prod_{x<t}\left(1 - \frac{dN(x)}{Y(x)}\right)$ was well approximated by $\hat{S}(t) = e^{-\hat{\Lambda}(t)}$. Thus a reasonable estimator of Var$\left(KM(t)\right)$would be to use the estimator of Var$\left(e^{-\hat{\Lambda}(t)}\right)$, or $\left[\hat{S}(t)\right]^2 \widehat{Var}\left(\hat{\Lambda}(t)\right) = \left[\hat{S}(t)\right]^2 \sum_{x<t} \frac{dN(x)}{Y^2(x)}$.

This is very close (asymptotically the same) as the estimator for the variance of the Kaplan-Meier estimator given by Greenwood. Namely

$$\widehat{Var}\left(KM(t)\right) = \prod_{x<t} \frac{dNx}{[Y(x) - w(x)/2][Y(x) - dNx - w(x)/2]}$$

(Used by SAS)

Ex (continued)

- using the delta-method approximation for getting a confidence interval with the Nelson-Aalen estimator, we get that a 95% C.I. for $S(t)$ at $t = 17$ is

$$e^{-\widehat{\Lambda}(t)} \pm 1.96 * e^{-\widehat{\Lambda}(t)} * se\left[\widehat{\Lambda}(t)\right] = e^{-0.084} \pm 1.96 * e^{-0.084} * 0.381 = [0.114, 0.784]$$

- If we use the Kaplan-Meier estimator, together with Greenwood's formula for estimating the variance,

$$KM(t) = \left[1 - \frac{1}{10}\right]\left[1 - \frac{1}{9}\right]\left[1 - \frac{1}{7}\right]\left[1 - \frac{1}{5}\right]\left[1 - \frac{1}{4}\right] = 0.411$$

$$\widehat{Var}\big(KM(t)\big) = 0.411^2\left[\frac{1}{10 * 9} + \frac{1}{9 * 8} + \frac{1}{7 * 6} + \frac{1}{5 * 4} + \frac{1}{4 * 3}\right] = 0.03077$$

$$\widehat{se}\big(KM(t)\big) = \sqrt{0.03077} = 0.175$$

$$\mathrm{Var}\left(\widehat{\Lambda}(t)\right) = \frac{1}{10 * 9} + \frac{1}{9 * 8} + \frac{1}{7 * 6} + \frac{1}{5 * 4} + \frac{1}{4 * 3} = 0.182$$

$$\widehat{se}\left(\widehat{\Lambda}(t)\right) = \sqrt{0.182} = 0.427$$

a 95% confidence interval for $S(t)$ is

$$KM(t) \pm 1.96 * \widehat{se}\big(KM(t)\big) = [0.068, 0.754]$$

Note:

If we want to use *R* function survfit() to construct a confidence interval for $S(t)$ with the form $KM(t) \pm 1.96 * \widehat{se}\big(KM(t)\big)$, we have to specify the argument conf.type=c("plain") in survfit(). The default constructs the confidence interval for $S(t)$ by exponentiating the confidence interval for the cumulative hazard using the Kaplan-Meier estimator. For example, a 95% CI for $S(t)$ is

$$KM(t) * \left[e^{-1.96*se\left(\hat{\Lambda}_{KM}(t)\right)}, e^{1.96*se\left(\hat{\Lambda}_{KM}(t)\right)}\right] = 0.411 * [e^{-1.96*0.427}, e^{1.96*0.427}]$$
$$= [0.178, 0.949] .$$

Comparison of confidence intervals for $S(t)$

1. Exponentiating the 95% CI for cumulative hazard using Nelson-Aalen estimator: $[0.212, 0.944]$.
2. Delta-method using Nelson-Aalen estimator: $= [0.114, 0.784]$.
3. Exponentiating the 95% CI for cumulative hazard using Kaplan-Meier estimator: $[0.178, 0.949]$.
4. Kaplan-Meier estimator together with Greenwood's formula for variance: $[0.068, 0.754]$.

Of the different methods for constructing confidence intervals, “usually” the most accurate is based on exponentiating the confidence intervals for the cumulative hazard function based on Nelson-Aalen estimator. We don't feel that symmetry is necessarily an important feature that confidence interval need have.

Estimator of Quantiles

Suppose we want to estimate the median $S^{-1}(0.5)$ or any other quantile $\varphi = S^{-1}(\theta)$, $0 < \theta < 1$. Then the point estimate of φ is obtained (using the Kaplan-Meier estimator of $S(t)$):

$$\hat{\varphi} = KM^{-1}(\theta), \text{ i.e. } KM(\hat{\varphi}) = \theta$$

An approximate $(1-\alpha)$th confidence intervals for φ is given by $[\hat{\varphi}_L, \hat{\varphi}_U]$, where

$$KM(\hat{\varphi}_L) - Z_{\frac{\alpha}{2}} * se[KM(\hat{\varphi}_L)] = \theta \quad \text{and} \quad KM(\hat{\varphi}_U) + Z_{\frac{\alpha}{2}} * se[KM(\hat{\varphi}_U)] = \theta$$

Proof: Here the argument is provided for a general estimator $\hat{S}(t)$. So if we use the Kaplan-Meier estimator, then $\hat{S}(t)$ is $KM(t)$. It can also be the Nelson-Aalen estimator. Then

$$P[\hat{\varphi}_L < \varphi < \hat{\varphi}_U] = P[S(\hat{\varphi}_U) < \theta < S(\hat{\varphi}_L)] = 1 - \{P[S(\hat{\varphi}_U) > \theta] + P[\theta > S(\hat{\varphi}_L)]\}$$

Denote φ_U the solution to the equation

$$S(\varphi_U) + Z_{\frac{\alpha}{2}} * se\left[\hat{S}(\varphi_U)\right] = \theta,$$

then $\hat{\varphi}_U$ will be close to φ_U. So,

$$P[S(\hat{\varphi}_U) > \theta] = P\left[S(\hat{\varphi}_U) > \hat{S}(\hat{\varphi}_U) + Z_{\frac{\alpha}{2}} * se\left[\hat{S}(\varphi_U)\right]\right] = P\left[\frac{\hat{S}(\hat{\varphi}_U) - S(\hat{\varphi}_U)}{se\left[\hat{S}(\hat{\varphi}_U)\right]} < -Z_{\frac{\alpha}{2}}\right]$$

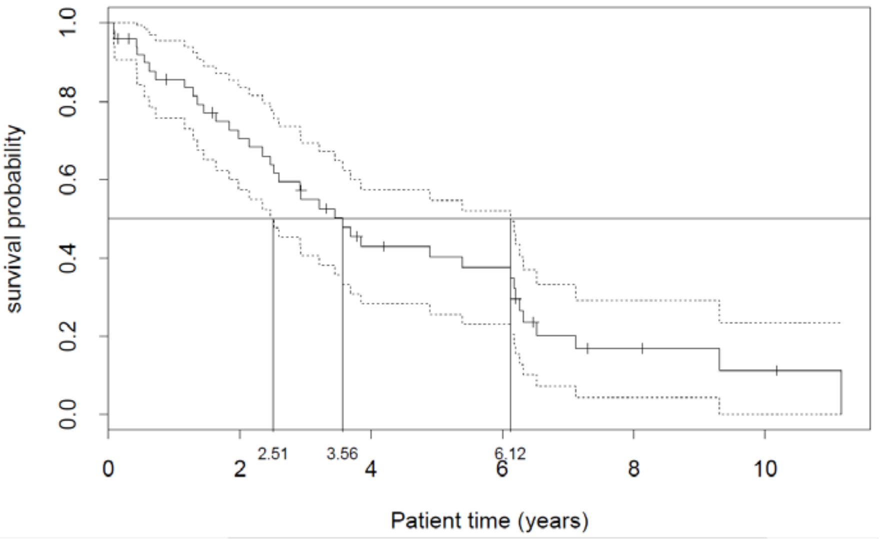
$$\approx P\left[\frac{\hat{S}(\varphi_U) - S(\varphi_U)}{se[\hat{S}(\varphi_U)]} < -Z_{\frac{\alpha}{2}}\right] \approx P\left[Z < -Z_{\frac{\alpha}{2}}\right] = \frac{\alpha}{2}.$$

Similarly, can show that $P[\theta > S(\hat{\varphi}_L)] \approx \frac{\alpha}{2}$.

Example

```
survtime <- rexp(50, 0.2)
censtime <- rexp(50, 0.1)
status <- (survtime <= censtime)
obstime <- survtime*status + censtime*(1-status)
fit <- survfit(Surv(time=obstime, event=status)~1)
summary(fit)
plot(fit, xlab="Patient time (years)", ylab="survival
probability")
```

(Note: You might not have exactly the same number as below since data are simulated.)



The true survival time has an exponential distribution with ¸ $\lambda = 0.2$/year (so the true mean is 5 years and median is $5 * \log(2) \approx 3.5$ years). The (potential) censoring time is independent from the survival time and has an exponential distribution with ¸ $\lambda = 0.1$/year (so it is stochastically larger than the survival time). The Kaplan estimate (solid line) and its 95% confidence intervals (dotted lines) are shown in the Figure. Note that these CIs are constructed by exponentiating the CIs for ¤(t). From this figure, the median survival time is estimated to be 3.56 years, with its 95% confidence interval [2.51, 6.20].

For symmetric confidence intervals of $S(t)$;

```
fit2 <- survfit(Surv(time=obstime, event=status)~1, conf.type=c("plain") )
summary(fit2)
plot(fit2, xlab="Patient time (years)", ylab="survival probability")
```

(Note: use conf.int=0.9 in survfit() for 90% C.I.)

Other Types of Censoring and Truncation

- *Left censoring*: occurs when the event of interest is only known to happen before a specific time point. For example, in a study of *time to first marijuana use* (example 1.17, page 17 of Klein & Moeschberger), 191 high school boys were asked “when did you first use marijuana?”. Some answers were “I have used it but cannot recall when the first time was” . For these boys, their *time to first marijuana use* is left censored at their current age. For the boys who never used marijuana, their *time to first marijuana use* is right censored at their current age. Of course, we got their exact *time to first marijuana use* for those boys who remembered when they first used it.

- *Interval censoring:* occurs when the event of interest is only known to take place in an interval. For example, in a study to compare time to cosmetic deterioration of breasts for breast cancer patients treated with radiotherapy and radiotherapy + chemo, patients were examined at each clinical visit for breast retraction and the breast retraction is only known to take place between two clinical visits or right censored at the end of the study. (Example 1.18 on page 18 of Klein & Moeschberger.)

- *Left truncation:* occurs when the *time to event* of interest in the study sample is greater than a (left) truncation variable. For example, in a study of life expectancy (survival time measured from *birth* to *death*) using elderly residents in a retirement community (example 1.16, page 15 of Klein & Moeschberger), the individuals must survive to a sufficient age to enter the retirement community. Therefore, their survival time is left truncated by their age entering the community. Ignoring the truncation will lead to a biased sample and the survival time from the sample will over estimate the underlying life expectancy.

- *Right truncation:* occurs when the *time to event* of interest in the study sample is less than a (right) truncation variable. For example, to study the induction period between infection with HIV and the onset of clinical AIDS, the ideal approach will be to collect a sample of patients infected with HIV and then follow them for some period of time until some of them develop clinical AIDS. However, this approach may be too lengthy and costly, or even infeasible. An alternative approach is to study those patients who may be infected with HIV from potentially contaminated blood transfusion and later developed clinical AIDS. However, in this case, it is unknown which patients were infected with HIV. So one strategy is to enroll only those individuals who have already developed clinical AIDS. (Example 1.19 on page 19 of Klein & Moeschberger for more description and the data.)

Note:

The K-M survival estimation approach cannot be directly applied to the data with the above censorings and truncations. Modified K-M approach or others have to be used. Similar to right censoring case, the censoring time and truncation time are often assumed to be independent of the time to event of interest (survival time). Since right censoring is the most common censoring scheme, we will focus on this special case most of the time in this course. Nonparametric estimation of the survival function (or the cumulative distribution function) for the data with other censoring or truncation schemes can be found in Chapters 4 and 5 of Klein & Moeschberger.