Survival Function

† The probability of an individual surviving beyond time t is given by the survivor function;

$$S(t) = \Pr(T > t) = \int_t^{\infty} f(u)du = 1 - \Pr(T \leq t) = 1 - F(t),$$

where $f(t)$: probability density function of t.

★★ $S(t)$ is a nonincreasing function with $S(0) = 1$ and $S(\infty) = \lim_{t\to\infty} S(t) = 0$.

$t_1 < t_2 \qquad S(t_1) \geq S(t_2)$

† We have

$(1 - F(t))$

$$f(t) = -\frac{dS(t)}{dt}$$

$f(t) = \dfrac{dF(t)}{dt}$

Notation is changed! It was Weibull(λ, α)

♢ eg.: $T \sim \text{Weibull}(\alpha, \lambda)$ ($\alpha > 0$ and $\lambda > 0$). That is,

$$f(t) = \begin{cases} \alpha\lambda t^{\alpha-1} \exp(-\lambda t^{\alpha}), & \text{for } t > 0, \\ 0 & \text{otherwise .} \end{cases}$$

Find the survival function.

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

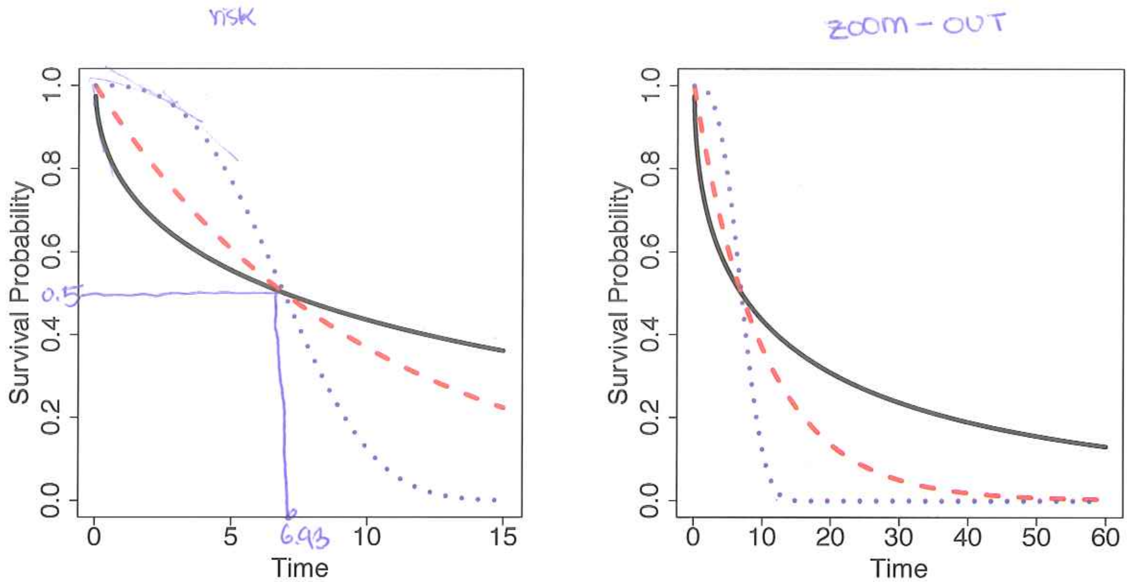
$F(t) = \int_0^t \alpha\lambda u^{\alpha-1} e^{-\lambda u^{\alpha}} du$

$= -\exp(-\lambda u^{\alpha}) \Big|_0^t = 1 - e^{-\lambda t^{\alpha}}, \quad t > 0$

$\Rightarrow \quad S(t) = 1 - (1 - e^{-\lambda t^{\alpha}}) = e^{-\lambda t^{\alpha}}, \quad t > 0$

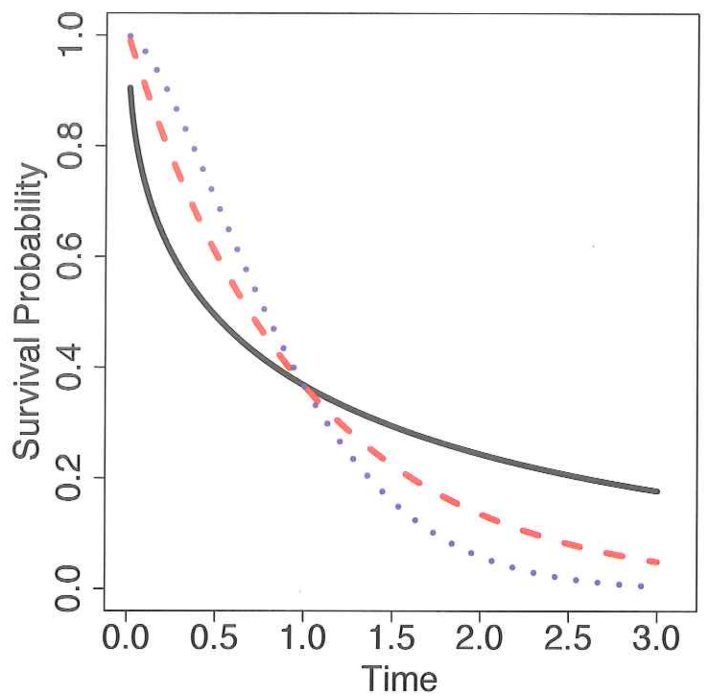
6.93 = F(t)

◇ eg.: $T \sim \text{Weibull}(\alpha, \lambda)$ ($\alpha > 0$ and $\lambda > 0$) $\Rightarrow$ $S(t) = \exp(-\lambda t^{\alpha})$. Survival curves with a common median of 6.93.



⋆⋆ $(\alpha, \lambda) = (0.5, 0.26328)$ for black solid, (1, 0.1) for red dashed and (3, 0.00208) for blue dotted.

♢ eg.: $T \sim \text{Weibull}(\alpha, \lambda)$ ($\alpha > 0$ and $\lambda > 0$) $\Rightarrow$ $S(t) = \exp(-\lambda t^{\alpha})$.



⋆⋆ $(\alpha, \lambda) = (0.5, 1)$ for black solid, (1, 1) for red dashed and (1.5, 1) for blue dotted.

⋆ Some basic properties of the survival function

⋆⋆ Monotone, nonincreasing functions equal to one at zero and zero as the time approaches infinity. $t=0$

⋆⋆ Their rate of decline varies according to the risk of experiencing the event at time t

Hazard Function

⋆ The hazard rate is defined as;

$$h(t) = \lim_{dt \to 0^+} \frac{\Pr(t \leq T < t + dt \mid T \geq t)}{dt}.$$

⋆ a.k.a.:

⋆⋆ conditional failure rate in reliability

⋆⋆ instantaneous rate of failure at time t

⋆⋆ the force of mortality in demography

⋆⋆ the intensity function in stochastic processes

⋆ The hazard rate is defined as;

$$h(t) = \lim_{dt \to 0^+} \frac{\Pr(t \leq T < t + dt \mid T \geq t)}{dt}.$$

⋆ More observations!

⋆⋆ conditional failure rate ; not a *probability*

⋆⋆ conditional rate (i.e., note division by dt) of failure in $[t, dt)$, given that failure has not occurred as of time t.

⋆⋆ $h(t) \geq 0$ for all $t > 0$ (no upper bound, requirement).

⋆⋆ Useful for modeling association of covariates (will see this later)

Cox - model

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

$\frac{dS(t)}{dt} = -f(t)$

⋆ For continuous T,

$$h(t) = \frac{f(t)}{S(t)} = \frac{-d\log(S(t))}{dt}.$$

⋆ How do we get this?

Assuming ~~coun~~ continuity of $F(t)$

$$f(t) = \lim_{dt \to 0^+} \frac{1}{dt} P(T \in [t, t+dt))$$

$$= \lim_{dt \to 0^+} \frac{1}{dt} P(T \in [t, t+dt), T \geq t)$$

$$= \underbrace{\boxed{\lim_{dt \to 0^+} \frac{1}{dt} P(T \in [t, t+dt) \mid T \geq t)}}_{=h(t)} \underbrace{P(T \geq t)}_{= S(t)}$$

$$= h(t) \cdot S(t)$$

$$\Leftrightarrow f(t) = h(t) S(t)$$

$$\Rightarrow h(t) = \frac{f(t)}{S(t)}$$

⋆ The hazard rate is defined as;

$$h(t) = \lim_{dt \to 0^+} \frac{\Pr(t \leq T < t + dt \mid T \geq t)}{dt}.$$

⋆ Intuition? Informally, using previous ideas

⋆⋆ $f(t)$ is related to P(die at t)

$$\begin{aligned} \text{P(die at } t) &= \text{P(die at } t\text{\& survive until } t) \\ &= \underbrace{\text{P(die at } t \mid \text{ survive until } t)}_{\to h(t)} \underbrace{\text{P(survive until } t)}_{\to S(t)}, \end{aligned}$$

which is related to $h(t)S(t)$.

$\diamondsuit$ eg.: $T \sim \mathrm{Weibull}(\alpha, \lambda)$ ($\alpha > 0$ and $\lambda > 0$). That is,

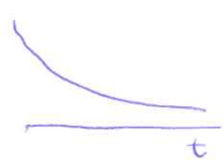
$$f(t) = \begin{cases} \alpha\lambda t^{\alpha-1}\exp(-\lambda t^{\alpha}), & \text{for } t > 0, \\ 0 & \text{otherwise .} \end{cases}$$

The survival function is $S(t) = \exp(-\lambda t^{\alpha})$. Find the hazard function.

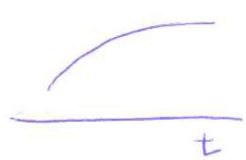
$$h(t) = \frac{f(t)}{S(t)} = \frac{\alpha\lambda t^{\alpha-1} \cancel{e^{-\lambda t^{\alpha}}}}{\cancel{e^{-\lambda t^{\alpha}}}} = \frac{\alpha\lambda t^{\alpha-1}}{>0}, \quad t > 0$$

$\alpha = 1$ $\quad h(t) = \lambda$

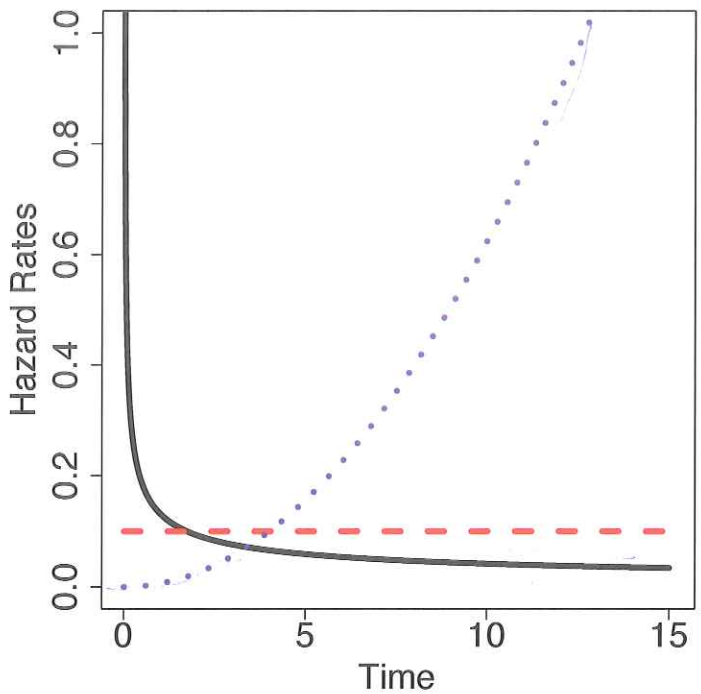
$\alpha < 1$ $\quad h(t)$



$\alpha > 1$ $\quad h(t)$

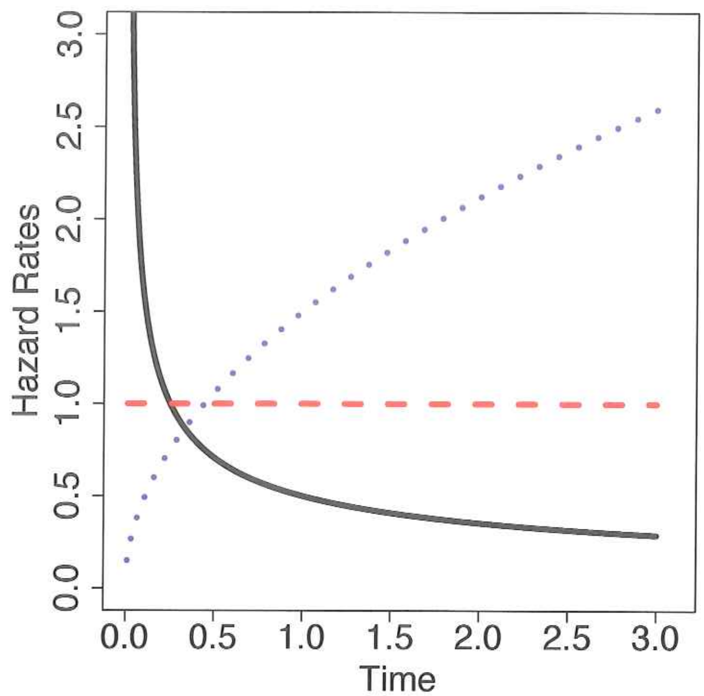


$\diamondsuit$ eg.: $T \sim \mathrm{Weibull}(\alpha, \lambda)$ ($\alpha > 0$ and $\lambda > 0$) $\Rightarrow h(t) = \alpha\lambda t^{\alpha-1}$.



$\star\star$: $(\alpha, \lambda) = (0.5, 0.26328)$ for black solid, (1, 0.1) for red dashed and (3, 0.00208) for blue dotted.

♢ eg.: $T \sim \text{Weibull}(\alpha, \lambda)$ ($\alpha > 0$ and $\lambda > 0$) $\Rightarrow h(t) = \alpha\lambda t^{\alpha-1}$.



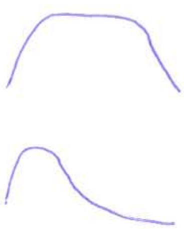
⋆⋆ $(\alpha, \lambda) = (0.5, 1)$ for black solid, (1, 1) for red dashed and (1.5, 1) for blue dotted.

⋆ The hazard rate for the occurrence of a particular event describes the mechanism of failure (describing the way in which the chance of experiencing the event changes with time).

⋆ The hazard rate can be

- ⋆⋆ **increasing:** when there is natural aging or wear
- ⋆⋆ **decreasing:** much less common but find occasional use when there is a very early likelihood of failure, such as in certain types of electronic devices or in patients experiencing certain types of transplants.
- ⋆⋆ **bathtub-shaped:** appropriate in populations followed from birth. During an early period, deaths result, primarily from infant disease, after which the death rate stabilizes, followed by an increasing hazard rate due to the natural aging process.

⋆ The hazard rate for the occurrence of a particular event can be (contd)



⋆⋆ **hump-shaped**: if hazard rate is increasing early and eventually begins declining. For example, we model survival after successful surgery when there is an initial increase in risk due to infection, hemorrhaging, or other complication just after the procedure, followed by a steady decline in risk as the patient recovers.

⋆ Notes:

⋆⋆ The hazard function is usually more informative about the underlying mechanism of failure than the survival function. For this reason, consideration of the hazard function may be the dominant method for summarizing survival data.

Cumulative Hazard Function

For continuous T_s

⋆ We define the cumulative hazard function $H(t)$,

$$H(t) = \int_0^t h(u)du.$$

$h(t) = - \frac{d \log S(t)}{dt}$

⋆

$$\Rightarrow \quad H(t) = \int_0^t h(u)du = -\log(S(t)).$$

$$\Rightarrow \quad S(t) = \exp(-H(t)) = \exp\left\{-\int_0^t h(u)du\right\}.$$

◇ eg.: $T \sim \text{Weibull}(\alpha, \lambda)$ ($\alpha > 0$ and $\lambda > 0$). That is,

$$f(t) = \begin{cases} \alpha\lambda t^{\alpha-1}\exp(-\lambda t^{\alpha}), & \text{for } t > 0, \\ 0 & \text{otherwise .} \end{cases}$$

The survival function is $S(t) = \exp(-\lambda t^{\alpha})$. The hazard function is $h(t) = \lambda\alpha t^{\alpha-1}$. Find the cumulative hazard function.

$$H(t) = -\log S(t) = -\log(\exp(-\lambda t^{\alpha}))$$

$$= \underline{\underline{\lambda t^{\alpha}}}$$

Discrete Survival Time

⋆ T may be a discrete random variable

⋆⋆ i.e. Suppose T can only take values; $0 \leq t_1 < t_2 \ldots < t_D$

⋆⋆ i.e. If death times are grouped into small intervals (e.g., days, weeks, etc)

⋆ Observed T is almost always discrete; but, models for continuous times are usually used if the underlying time is believed to be continuous.

Discrete Survival Time (contd)

⋆ For discrete T, we define the probability function;

$$p_j = Pr(T = t_j)$$

$\sum p_j = 1,\ p_j \geq 0$

⋆ Survival function: $S(t_j) = Pr(T > t_j) = 1 - \sum_{\ell=1}^{j} p_\ell$

$t_0 = 0$

⋆ Combining the first two, we get $p_j = S(t_{j-1}) - S(t_j)$ & $S(t_0) = S(0) = 1$

⋆ Hazard function:

$$h_j = h(t_j) = Pr(T = t_j \mid T \geq t_j) = \frac{p_j}{S(t_{j-1})}$$

⋆ We also observe

$0 < t_1 < t_2 < \cdots$

$$S(t_j) = \prod_{\ell=1}^{j} (1 - h_\ell).$$

$= Pr(T > t_j) = \prod_{\ell=1}^{j} \underbrace{P(T > t_\ell \mid T \geq t_\ell)}_{= 1 - h_\ell}$

Mean Residual Life Function ~~and Median Life~~

⋆ Mean residual life at time t ($\mathrm{mrl}(t)$) is defined as

· expected remaining lifetime at time t.

$$\mathrm{mrl}(t) = \mathrm{E}(T - t \mid T > t).$$

⋆ For continuous T,

$V = -S(u)$

$$\mathrm{mrl}(t) = \frac{\int_t^\infty \underbrace{(u-t)}_{w}\underbrace{f(u)}_{v'}du}{S(t)} = \frac{\int_t^\infty S(u)du}{S(t)}.$$

$= \int_t^\infty (u - t)\, f(u \mid u > t)\, du$

$= \int_t^\infty (u - t) \dfrac{f(u)}{P(\tilde{u} > t)} du$ $\quad\int_t^\infty f(\tilde{u})\, d\tilde{u}$

$= \int_t^\infty \dfrac{(u-t) f(u)}{S(t)} du$

♣ **Censoring:** some lifetimes are known to have occurred only within certain intervals

- ⋆⋆ Right censoring (most common)
- ⋆⋆ Left censoring
- ⋆⋆ Interval censoring

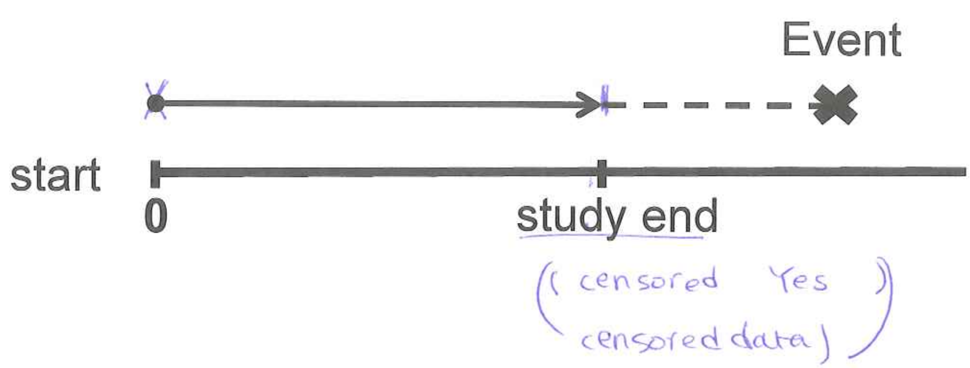
†† Assume that censoring is *noninformative.* In other words,

- ⋆⋆ potential censoring time is unrelated to the potential event time
- ⋆⋆ Inferences on survival do not depend on the censoring process.

5. Survival data usually contain censored observations (Incomplete observation of failure time).

   e.g.: right censoring, left censoring, interval censoring. (⇒ will be discussed in detail later)

   - Leukemia patients in remission



♣ **Right Censoring:**

⋆⋆ Right Censored at C: survival times are known for only a portion of the individuals under study, and the remainder of the survival times are known only to exceed certain values.

- If T is less than or equal to C, T is known
- If T is greater than C, T is censored at C
  ($T > C$ is known)

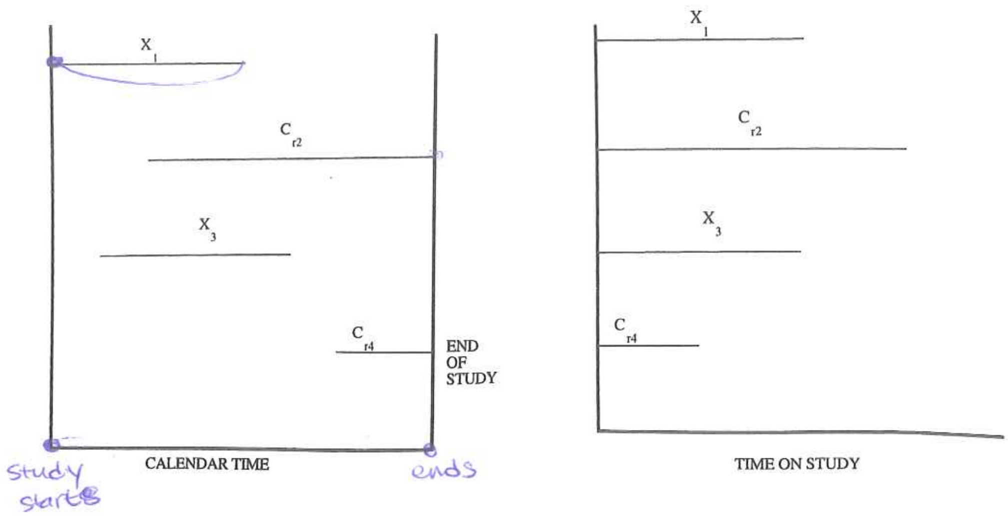
† To represent data, use (Y, ν),

time

$$\nu = \begin{cases} 1 & \text{if } T \text{ corresponds to an event}(T \le C), \\ 0 & \text{if } T \text{ is censored}(T > C). \end{cases}$$

and $Y = \min(T, C)$

♣ **Right Censoring:** (contd)



From K-M, page 66

- Write down the likelihood for right censored data.

We have n subjects. Assume independence among subjects. Data is $\mathcal{D} = (n, \mathbf{y}, \boldsymbol{\nu})$.

$$\mathcal{L} = \prod_{i=1}^{n} (f(y_i))^{\nu_i} (\underbrace{P(t_i > y_i)}_{=1-F(y_i) = S(y_i)})^{1-\nu_i} \qquad \int_{y_i}^{\infty} f(t)\,dt$$

$$= \prod_{i=1}^{n} (f(y_i))^{\nu_i} (S(y_i))^{1-\nu_i} \qquad S(y_i) = \exp(-H(y_i))$$

$$= \prod_{i=1}^{n} (f(y_i))^{\nu_i} \{\exp(-H(y_i))\}^{1-\nu_i}$$

$$= \prod_{i=1}^{n} (f(y_i) \cdot \exp(H(y_i)))^{\nu_i} \exp(-H(y_i))$$

† We have $f(t) = h(t)S(t)$ so the likelihood becomes

$$\mathcal{L} \propto \prod_{i=1}^{n} \{h(y_i)\}^{\nu_i} \exp\{-H(y_i)\}.$$

$$h(y_i) = \frac{f(y_i)}{S(y_i)}$$

$$S(y_i) = (\exp(-H(y_i)))^{-1} = \frac{1}{\exp(H(y_i))}$$

$$= \prod_{i=1}^{n} \left(\frac{f(y_i)}{S(y_i)}\right)^{\nu_i} \exp(-H(y_i))$$

$$= \prod_{i=1}^{n} (h(y_i))^{\nu_i} \exp(-H(y_i))$$

♣ **Left Censoring:**

⋆⋆ Left Censored at C for a portion of the individuals under study, and the remainder of the survival times are known only to exceed certain values.

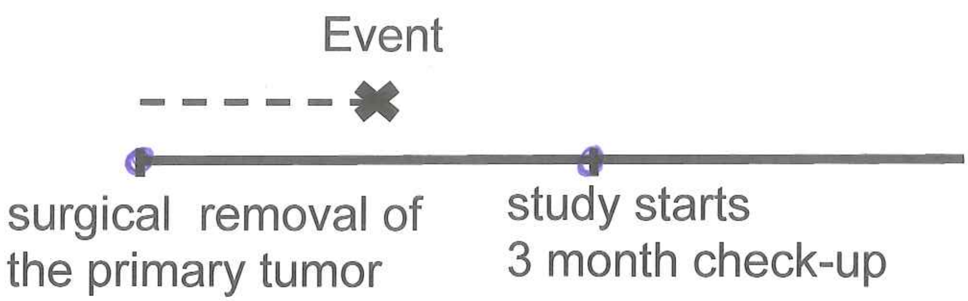
- If T is greater than or equal to C, T is known
- If T is less than C, T is left-censored at C
  ($T < C$ is known)

† To represent data, use (Y, ν),

$$\nu = \begin{cases} 1 & \text{if } T \text{ is observed}(T \leq C), \\ 0 & \text{if } T \text{ is censored}(T > C). \end{cases}$$

and $Y = \max(T, C)$

• e.g. Suppose the event of interest is recurrence of cancer. At the first follow-up visit, cancer already recurred.



- Write down the likelihood for left censored data.

We have n subjects. Assume independence among subjects. Data is $\mathcal{D} = (n, \mathbf{y}, \boldsymbol{\nu})$.

$$\mathcal{L} = \prod_{i=1}^{n} (f(y_i))^{\nu_i} (P(t < y_i))^{1-\nu_i}$$

$$= \prod_{i=1}^{n} (f(y_i))^{\nu_i} (F(y_i))^{1-\nu_i}$$

♣ **Interval Censoring:**

⋆⋆ We only observe $T \in (A_i, B_i)$

• e.g. In HI~~B~~V incidence studies (e.g., vaccine trials), endpoint of interest is time until HIV infection. Tested for HIV periodically, e.g, every 6 months. So T only known to be between last negative and first positive test.

- Write down the likelihood for interval censored data.

We have n subjects. Assume independence among subjects. Data is $\mathcal{D} = (n, \mathbf{y}, \boldsymbol{\nu})$.

$y_i = (A_i, B_i)$

$$\mathcal{L} = \prod_{i=1}^{n} P(A_i < t_i < B_i)$$

$$= \prod_{i=1}^{n} (F(B_i) - F(A_i))$$

$$= \prod_{i=1}^{n} (S(A_i) - S(B_i))^{1-\nu_i} \cdot f($$

♣ Nonparametric Estimation of Basic Quantities:

- [Setting]: Suppose
  - ⋆⋆ The events occur at D distinct times, $t_1 < t_2 < \ldots < t_D$.
  - ⋆⋆ At time t_j, there are d_j events ($j = 1, \ldots, D$).
  - ⋆⋆ N_j: the number of individuals who are at risk at time t_j (the number of individuals who are alive at time t_j or experience the event of interest at t_j)
- See KM Chapters 4.1–4.3 and ICS Chapter 3.7.2

- How does this work? What does d_j/N_j mean?

- $S(t)$ can be expressed...

$$S(t) = P(T > t) = P(T > t \mid T > t_5)\, P(T > t_5 \mid T > t_4)$$

$$\cdots \quad P(T > t_1)$$

$$P(T > t_j \mid T > t_{j-1}) = 1 - P(T \leq t_j \mid T > t_j)$$

$$\approx 1 - \frac{d_j}{N_j}$$

♣ The Product Limit Estimator (Kaplan-Meier!): For all values of t in the range where there is data,

$$\hat{S}(t) = \begin{cases} 1 & \text{if } t < t_1, \\ \prod_{t_j < t} \{1 - \frac{d_j}{N_j}\} & \text{if } t_1 \leq t. \end{cases}$$

⋆⋆ How does this work? What does d_j/N_j estimate?

⋆⋆ ⇒ d_j/N_j: estimate of the conditional probability that an individual who survives to just prior to time t_j experiences that event at time t_j

⋆⋆ ⇒ Use this to construct estimators of the survival function and the cumulative hazard rate

• Example: the following data is observed, with $^+$ denoting censoring.

$D = 6$

i	1	2	3	4	5	6	7	8	9	10
Y_i	2	5^+	8	12^+	15	21^+	25	29	30^+	34

j	t_j	N_j	d_j	d_j/N_j
0	0	10	0	
1	2	10	1	
2	8	8	1	
3	15	6	1	
4	25	4	1	
5	29	3	1	
6	34	1	1	

- For all values of t in the range where there is data,

$$\hat{S}(t) = \begin{cases} 1 & \text{if } t < t_1, \\ \prod_{t_j < t}\{1 - \frac{d_j}{N_j}\} & \text{if } t_1 \leq t. \end{cases}$$

★★ The estimator is a step function with jumps at the observed event times.

★★ $\hat{S}(t)$ has a total of D jumps; one at each unique failure time.

★★ The jump sizes depend on (1) the number of events observed d_j at each event time t_j and (2) the pattern of the censored observations prior to t_j. N_j

★★ For values of t beyond the largest observation t_D time, this estimator is not well defined.

★★ If the last subject at risk is a death, $\hat{S}(t)$ will drop to 0; if not, $\hat{S}(t)$ will not reach 0.

★★ Provides an efficient means of estimating the survival function for right-censored data

- For all values of t in the range where there is data,

$$\hat{S}(t) = \begin{cases} 1 & \text{if } t < t_1, \\ \prod_{t_j < t} \{1 - \frac{d_j}{N_j}\} & \text{if } t_1 \leq t. \end{cases}$$

★★ Kaplan and Meier show that the Kaplan-Meier estimator of $S(t)$ as a nonparametric MLE.

★★ Susarla and Van Ryzin show that $\hat{S}(t)$ is a limiting case of the estimator under Dirichlet process (assume the Dirichlet process for $S(t)$).

- CI for the Product Limit estimator (Kaplan-Meier!):

For all values of t in the range where there is data, the variance of the KM estimator is given by Greenwood's formula;

$$\hat{V}(\hat{S}(t)) = \hat{S}^2(t) \sum_{t_j \leq t} \frac{d_j}{N_j(N_j - d_j)}.$$

$\Rightarrow$ We can draw pointwise confidence intervals using the normal approximation!

- Now estimate the cumulative hazard function,
  $H(t) = -\log(S(t))$
- We can use KM estimator so that $\hat{H}(t) = -\log(\hat{S}(t))$.
- The Nelson-Aalen estimator of the cumulative hazard is defined up to the largest observed time on study:

$$\tilde{H}(t) = \begin{cases} 0 & \text{if } t \leq t_1, \\ \prod_{t_j < t} \frac{d_j}{N_j} & \text{if } t_1 \leq t. \end{cases}$$

★★ Better small sample size performance!

- The estimated variance is given by

$$\sigma_H^2(t) = \sum_{t_j < t} \frac{d_j}{N_j^2}.$$

∗ Example 4.2 of KM:

Please read Section 1.3 of KM for a full description of the data. In short,

- Bone marrow transplants are a standard treatment for acute leukemia.
- A total of 137 patients are treated.
- Several risk factors were measured at the time of transplantation.
- One of them is the disease group; patients were grouped into three risk categories based on their status at the time of transplantation;
  - ALL (38 patients)
  - AML low-risk first remission (54 patients)
  - AML high-risk second remission or untreated first relapse (15 patients) or second or greater relapse or never in remission (30 patients).

* Example 4.2 of KM (contd):

  - An individual is said to be disease-free at a given time after transplant f that individual is alive without the recurrence of leukemia.
  - The event indicator for disease-free survival is $\nu = 1$ if the individual has died or has relapsed.
  - The days on study for a patient is the smaller of their relapse or death time.

* Example 4.2 of KM:

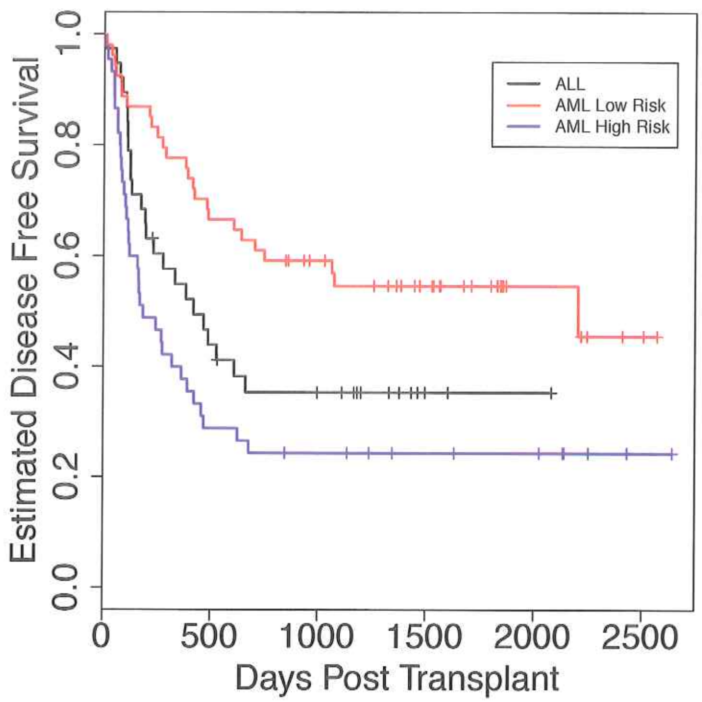
```
> rm(list=ls(all=TRUE))
> library(KMsurv)
> library(survival)
>
> data(bmt)
> mfit <- survfit(Surv(bmt$t2, bmt$d3)~bmt$group)
>
> pdf("BMT-K-M.pdf")
> par(mar=c(4.5, 4.5, 2.1, 2.1), mfrow=c(1,1))
> plot(mfit, conf.int=FALSE, col=c(1, 2, 4), cex.axis=2, cex.lab=2, lwd=2,
xlab="Days Post Transplant", ylab="Estimated Disease Free Survival")
> legend(1800, 0.95, c("ALL","AML Low Risk", "AML High Risk"),
col = c(1, 2, 4), lwd=2)
> dev.off()
quartz
     2
>
> pdf("BMT-K-M-1.pdf")
> par(mar=c(4.5, 4.5, 2.1, 2.1), mfrow=c(1,1))
> plot(mfit, conf.int=TRUE, col=c(1, 2, 4), cex.axis=2, cex.lab=2, lwd=2,
xlab="Days Post Transplant", ylab="Estimated Disease Free Survival")
> legend(1800, 0.95, c("ALL","AML Low Risk", "AML High Risk"),
col = c(1, 2, 4), lwd=2)
> dev.off()
quartz
```

∗ Example 4.2 of KM:

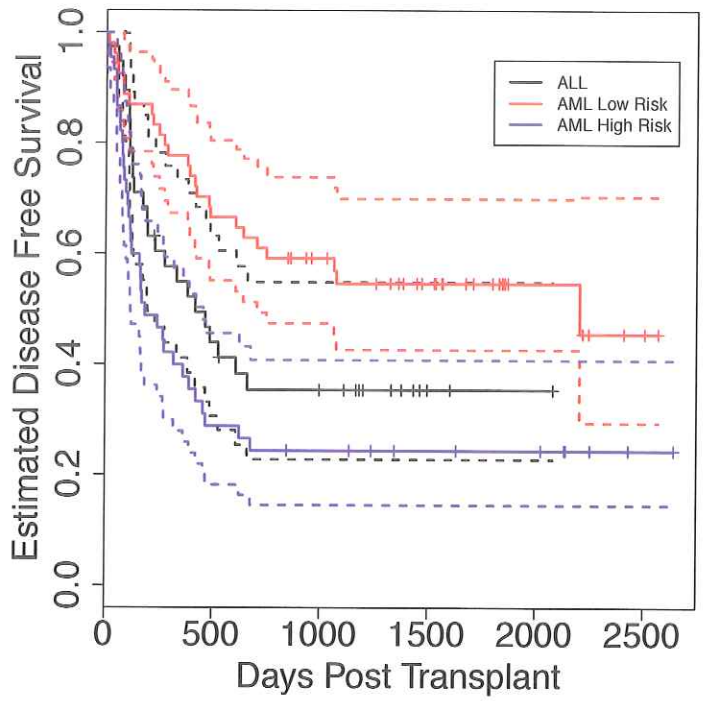
```
> summary(mfit)
Call: survfit(formula = Surv(bmt$t2, bmt$d3) ~ bmt$group)

                bmt$group=1
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    1     38       1    0.974  0.0260        0.924        1.000
   55     37       1    0.947  0.0362        0.879        1.000
   74     36       1    0.921  0.0437        0.839        1.000
   86     35       1    0.895  0.0498        0.802        0.998
  104     34       1    0.868  0.0548        0.767        0.983
  107     33       1    0.842  0.0592        0.734        0.966
...
```

* Example 4.2 of KM:



* Example 4.2 of KM:



* Want to play with KM estimator and R functions?

  check

- https://www.openintro.org/download.php?file=survival_analysis_in_R&referrer=/stat/surv.php
- https://cran.r-project.org/web/packages/survival/survival.pdf
- For data, https://cran.r-project.org/web/packages/KMsurv/KMsurv.pdf