3. Parametric Survival Models and Likelihood Methods

3.1 Some common parametric survival models

Distribution	$\lambda(t)$	$S(t)$	density $f(t)$	$\mathrm{E}(T)$
Exponential	$\lambda(>0)$	$e^{-\lambda t}$	$\lambda e^{-\lambda t}$	$\frac{1}{\lambda}$
Weibull	$\alpha\lambda t^{\alpha-1}(\alpha,\lambda>0)$	$e^{-\lambda t^{\alpha}}$	$\alpha\lambda t^{\alpha-1}e^{-\lambda t^{\alpha}}$	$\frac{\Gamma(1+1/\alpha)}{\lambda^{1/\alpha}}$
Gamma	$\frac{f(t)}{S(t)}$	$1-I(\lambda t,\beta)$	$\frac{\lambda^{\beta}t^{\beta-1}e^{-\lambda t}}{\Gamma(\beta)}$	$\frac{\beta}{\lambda}$

where $I(t,\beta)=\int_0^t \frac{u^{\beta-1}e^{-u}}{\Gamma(\beta)}du$. See page 38 of Klein and Moeschberger and Chapter 2 of the lecture notes for more distributions.

Exponential distribution: $\lambda(t)=\lambda$, $S(t)=e^{-\lambda t}$ and $f(t)=\lambda e^{-\lambda t}$. So **mean survival time**

$$\mu=\mathrm{E}(T)=\int_0^{\infty} tf(t)dt=\int_0^{\infty} S(t)dt=\int_0^{\infty}\mathrm{e}^{-\lambda t}dt=\frac{1}{\lambda}.$$

Letting $S(t_{0.5})=e^{-\lambda t_{0.5}}=0.5$, then **median survival time** is $t_{0.5}=\frac{\log 2}{\lambda}$.

The **mean residual life time** after t_0 is

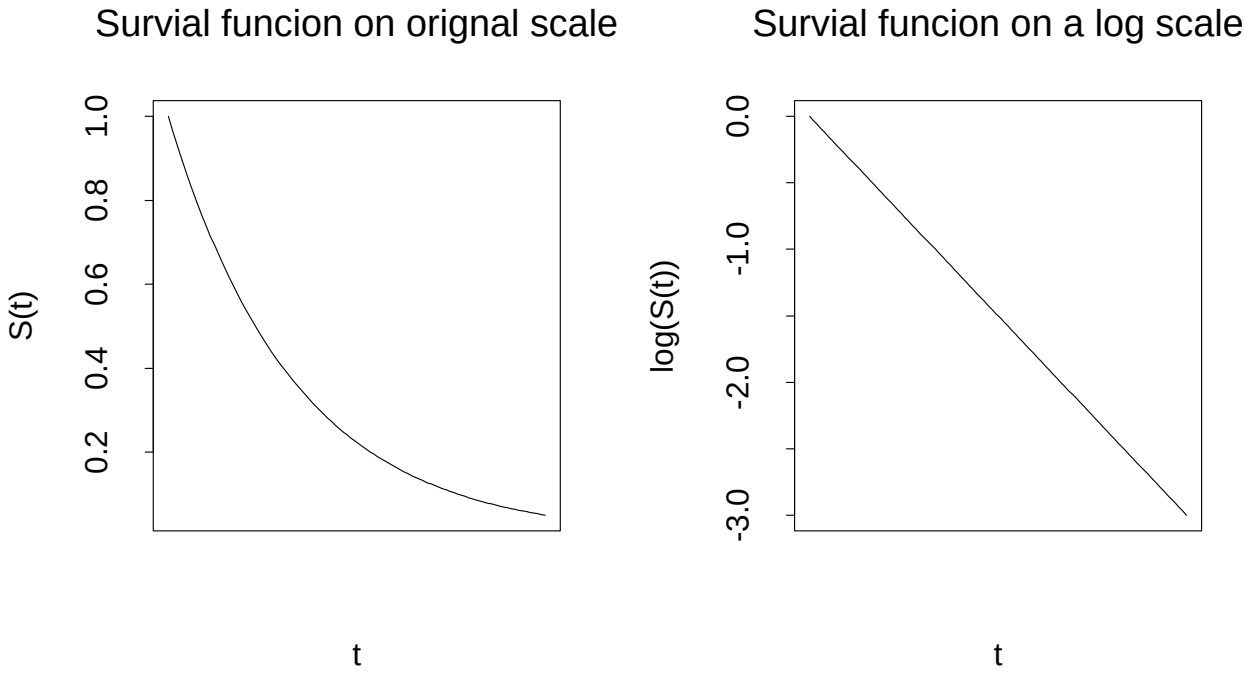
$$mrl(t_0)=\frac{\int_{t_0}^{\infty}S(t)dt}{S(t_0)}=\frac{\int_{t_0}^{\infty}\mathrm{e}^{-\lambda t}}{\mathrm{e}^{-\lambda t_0}}=\frac{1}{\lambda}=\mathrm{E}(T).$$

Sometimes it is useful to plot the survival distribution on a log scale. By so doing, we can identify the hazard rate as minus of the derivative of this function. In particular on a log scale the exponential distribution is a straight line. This is because $S(t)=\exp(-\lambda t)$ for the exponential distribution, so $\log\{S(t)\}=-\lambda t$.

The above equation gives us a way to check if the underlying true distribution of the survival time is exponential or not given a data set. Suppose we can have an estimate $\hat{S}(t)$ of $S(t)$ without assuming any distribution of the survival times (the Kaplan-Meier estimate to be discussed in Chapter 4 is such an estimate). Then we can plot $\log\{\hat{S}(t)\}$ vs t to see if it is approximately a straight line. A (approximate) straight line indicates that the exponential distribution may be a reasonable choice for the data.

Another alternative is to assume the exponential distribution for the data and get the estimate of $S(t) = \exp(-\lambda t)$ (we only need to estimate λ; this kind of estimation may be done by maximizing the likelihood function of the observed data). Denote this estimate by $\hat{S}_1(t)$ and Kaplan-Meier estimate by $\hat{S}_{KM}(t)$. If the exponential distribution assumption is correct, both estimates will be good estimates of the same survival function $S(t)$. Therefore, $\hat{S}_1(t)$ and $\hat{S}_{KM}(t)$ should be close to each other and hence the plot $\hat{S}_1(t)$ vs $\hat{S}_{KM}(t)$ should be approximately a straight line. A non-straight line indicates that the exponential distributional assumption is not appropriate.

Figure 1: *The survival function of an exponential distribution on two scales*

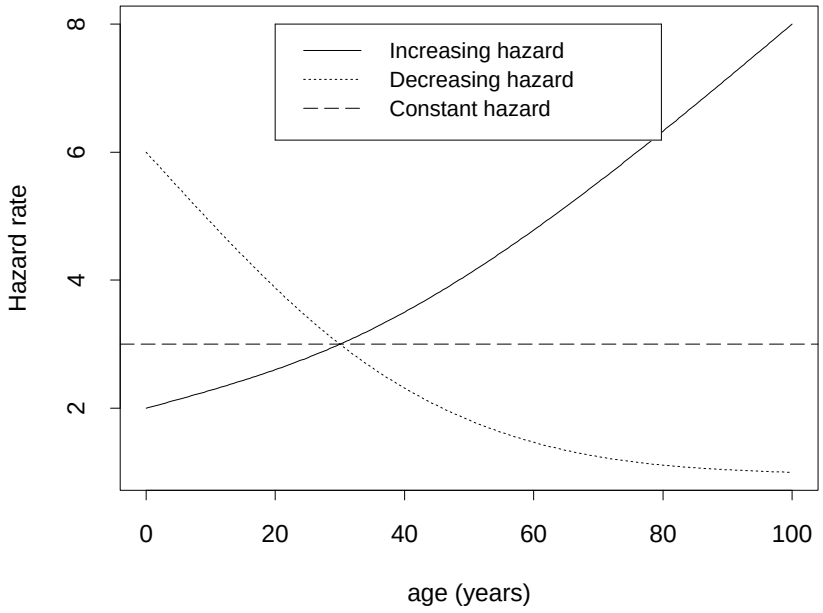


Weibull distribution: $\lambda(t) = \alpha\lambda t^{\alpha-1}, S(t) = \exp(-\lambda t^{\alpha})$. Note this model allows:

- Constant hazard (exponential distribution): $\alpha = 1$
- increasing hazard: $\alpha > 1$
- decreasing hazard: $\alpha < 1$.

and has the hazard patterns shown in Figure 2.

Figure 2: *Three hazard patterns*



The **mean survival time**

$$\mu = \mathrm{E}(T) = \int_0^\infty S(t)dt = \int_0^\infty \mathrm{e}^{-\lambda t^\alpha} dt = \frac{\Gamma(1+1/\alpha)}{\lambda^{1/\alpha}}.$$

The **median survival time** $t_{0.5}$: $\mathrm{e}^{-\lambda t^\alpha} = 0.5 \Longrightarrow$

$$t_{0.5} = \left[\frac{\log 2}{\lambda}\right]^{1/\alpha}.$$

Since $\log\{S(t)\} = -\lambda t^\alpha$, so

$$\log[-\log\{S(t)\}] = \log(\lambda) + \alpha \log(t).$$

A straight line in the plot of $\log[-\log\{S(t)\}]$ vs. $\log(t)$ indicates a Weibull model. We can use the above equation to check if the Weibull model is a reasonable choice for the survival time given a data set. Alternatively, we can assume a Weibull model for the survival time and use the

data to estimate $S(t)$ and plot this estimate against the Kaplan-Meier estimate as we proposed for the exponential distribution. A (approximate) straight line indicates the Weibull model is a reasonable choice for the data.

3.2 Maximum likelihood estimation

3.2.1 Review of parametric likelihood inference

Suppose we have a random sample (*i.i.d.*) $X_1, X_2, ..., X_n$ from distribution $f(x;\theta)$ (here $f(x;\theta)$ is either the density function if the random variable X is continuous or probability mass function is X is discrete; θ can be a scalar parameter or a vector of parameters). The distribution $f(x;\theta)$ is totally determined by the parameter θ. For example, if X_i is known from a log-normal distribution, then

$$f(x;\theta) = \frac{1}{\sqrt{2\pi}x\sigma} e^{-(\log x-\mu)^2/(2\sigma^2)}, \tag{1}$$

and $\theta = (\mu, \sigma)$ are the parameters of interest. Any quantity w.r.t. X can be determined by θ. For example, $\mathrm{E}(X) = e^{\mu+\frac{1}{2}\sigma^2}$. The likelihood function of θ (given data $X_1, ..., X_n$) is

$$\begin{aligned} L(\theta; X_1, ..., X_n) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}X_i\sigma} e^{-(\log X_i-\mu)^2/(2\sigma^2)} &(2)\\ &= (\sqrt{2\pi}\sigma)^{-n} \prod_{i=1}^{n} \frac{e^{-(\log X_i-\mu)^2/(2\sigma^2)}}{X_i}. &(3)\end{aligned}$$

In general, the likelihood function of θ (given data $X_1, ..., X_n$) is given by

$$L(\theta; X_1, ..., X_n) = \prod_{i=1}^{n} f(X_i;\theta) \tag{4}$$

and the log-likelihood function is

$$\ell(\theta; X_1, ..., X_n) = \log\{L(\theta; X_1, ..., X_n)\} = \sum_{i=1}^{n} \log\{f(X_i;\theta)\}. \tag{5}$$

Note that the (log) likelihood function of θ is viewed more as a function of θ than of data $X_1, ..., X_n$. We are interested in making inference on θ: estimating θ, constructing confidence interval (region) for θ, and performing hypothesis testing for (part) of θ.

The maximum likelihood estimate (MLE) $\hat{\theta}$ of θ is defined as the maximizer of $\ell(\theta; X)$, which can be usually obtained by solving the following *likelihood equation* (or *score equation*)

$$U(\theta; X_1, ..., X_n) = \frac{\partial \ell(\theta; X_1, ..., X_n)}{\partial \theta} = \sum_{i=1}^{n} \frac{\partial \log\{f(X_i; \theta)\}}{\partial \theta} = 0,$$

and $U(\theta)$ is often referred to as the *score*. Usually $\hat{\theta}$ does not have a closed form, in which case an iterative algorithm such as Newton-Raphson algorithm can be used to find $\hat{\theta}$.

Obviously, the MLE $\hat{\theta}$ is a function of data $\mathbf{X} = (X_1, X_2, ..., X_n)$, and hence a statistic that has a sampling distribution. Let θ_0 be the true value of θ. We have, as n goes to infinity,

$$\sqrt{n}(\hat{\theta} - \theta_0) \to \mathrm{N}\{0, I^{-1}(\theta_0)\},$$

where

$$I(\theta) = -\mathrm{E}\left[\frac{\partial^2 \log\{f(X_1; \theta)\}}{\partial \theta \partial \theta^T}\right]$$

is often referred to as the expected *Fisher information matrix* and

$$I_n(\theta) = -\frac{\partial^2 \ell(\theta; \mathbf{X})}{\partial \theta \partial \theta^T} = -\sum_{i=1}^{n} \frac{\partial^2 \log\{f(X_i; \theta)\}}{\partial \theta \partial \theta^T}$$

is the *observed fisher information matrix*. It can be shown that $I_n(\hat{\theta})/n$ is a consistent estimator for $I(\theta_0)$. These results can be used to construct confidence interval (region) for θ.

Suppose $\theta = (\theta_1, \theta_2)$ and we are interested in testing $H_0 : \theta_1 = \theta_{10}$ v.s. $H_A : \theta_1 \neq \theta_{10}$. Under mild conditions, the following test procedures can be used to test H_0.

Wald test: Suppose the corresponding decompositions of $\hat{\theta}$ and $\hat{I}_n$ are

$$\hat{\theta} = \begin{pmatrix} \hat{\theta}_1 \\ \hat{\theta}_2 \end{pmatrix}, \quad \hat{I}_n \equiv I_n(\hat{\theta}) = \begin{pmatrix} \hat{I}_{n,11} & \hat{I}_{n,12} \\ \hat{I}_{n,21} & \hat{I}_{n,22} \end{pmatrix} \quad \text{and} \quad \hat{I}_n^{-1} = \begin{pmatrix} \hat{I}_n^{11} & \hat{I}_n^{12} \\ \hat{I}_n^{21} & \hat{I}_n^{22} \end{pmatrix}.$$

Then under H_0,

$$\chi^2_{obs} = (\hat{\theta}_1 - \theta_{10})^T (\hat{I}_n^{11})^{-1} (\hat{\theta}_1 - \theta_{10}) \overset{a}{\sim} \chi^2_k,$$

where k is the dimension of θ_1. Therefore, we reject H_0 if $\chi^2_{obs} > \chi^2_{1-\alpha,k}$, where $\chi^2_{1-\alpha,k}$ is the $(1-\alpha)$th percentile of χ^2_k.

Score test: The score test is based on the fact that the score $U(\theta_0; \mathbf{X})/\sqrt{n}$ converges in distribution to a normal random vector with mean zero and covariance matrix $I(\theta_0)$. Decompose $U(\theta; X)$ as $U(\theta; \mathbf{X}) = \{U_1^T(\theta; \mathbf{X}), U_2^T(\theta; \mathbf{X})\}^T$ and let $\tilde{\theta}_2$ be the restricted MLE of θ_2 under $H_0 : \theta_1 = \theta_{10}$, *i.e.*, $\tilde{\theta}_2$ maximizes $\ell(\theta_{10}, \theta_2; \mathbf{X})$. Then under $H_0 : \theta_1 = \theta_{10}$,

$$\chi^2_{obs} = \tilde{U}_1^T \tilde{I}_n^{11} \tilde{U}_1 \sim \chi^2_k,$$

where $\tilde{U}_1 = U_1(\theta_{01}, \tilde{\theta}_2; \mathbf{X})$ and $\tilde{I}_n^{11}$ is the first $k \times k$ submatrix of $I_n^{-1}(\theta_{01}, \tilde{\theta}_2)$. We reject H_0 if $\chi^2_{obs} > \chi^2_{1-\alpha,k}$.

Likelihood ratio test: Under $H_0 : \theta_1 = \theta_{10}$,

$$\chi^2_{obs} = -2\{\ell(\theta_{10}, \tilde{\theta}_2; \mathbf{X}) - \ell(\hat{\theta}; \mathbf{X})\} \sim \chi^2_k.$$

Therefore, we reject H_0 if $\chi^2_{obs} > \chi^2_{1-\alpha,k}$.

An example of score tests: Suppose that the failure times $t_1, t_2, ..., t_n$ are from a Weibull distribution with survival function $s(t) = e^{-\lambda t^\alpha}$ and there is no censoring. We want to construct a score test for testing $H_0 : \alpha = 1$, i.e., the failure times are actually from an exponential distribution.

The likelihood function of (α, λ) is

$$\begin{aligned} L(\alpha, \lambda; \mathbf{t}) &= \prod_{i=1}^n \left[\lambda \alpha t_i^{\alpha-1} e^{-\lambda t_i^\alpha}\right] \\ &= \lambda^n \alpha^n e^{-\lambda \sum_{i=1}^n t_i^\alpha + (\alpha-1)\sum_{i=1}^n \log(t_i)}. \end{aligned}$$

Therefore, the log-likelihood function of (α, λ) is

$$\ell(\alpha, \lambda; \mathbf{t}) = n\log(\lambda) + n\log(\alpha) - \lambda \sum_{i=1}^n t_i^\alpha + (\alpha - 1)\sum_{i=1}^n \log(t_i).$$

So the components of the score are:

$$\begin{aligned} U_1(\alpha, \lambda) = \frac{\partial \ell(\alpha, \lambda; \mathbf{t})}{\partial \alpha} &= \frac{n}{\alpha} - \lambda \sum_{i=1}^n t_i^\alpha \log(t_i) + \sum_{i=1}^n \log(t_i) \\ U_2(\alpha, \lambda) = \frac{\partial \ell(\alpha, \lambda; \mathbf{t})}{\partial \lambda} &= \frac{n}{\lambda} - \sum_{i=1}^n t_i^\alpha, \end{aligned}$$

and the components of the information matrix is

$$\begin{aligned}
\frac{\partial^2 \ell(\alpha,\lambda;\mathbf{t})}{\partial \alpha^2} &= -\frac{n}{\alpha^2} - \lambda \sum_{i=1}^{n} t_i^{\alpha} (\log(t_i))^2 \\
\frac{\partial^2 \ell(\alpha,\lambda;\mathbf{t})}{\partial \alpha \partial \lambda} &= -\sum_{i=1}^{n} t_i^{\alpha} \log(t_i) \\
\frac{\partial^2 \ell(\alpha,\lambda;\mathbf{t})}{\partial \lambda^2} &= -\frac{n}{\lambda^2}.
\end{aligned}$$

For a given data, we can calculate the above quantities under $H_0 : \alpha = 1$ and construct a score test. For example, suppose $n = 25$, $\sum t_i = 6940$, $\sum \log(t_i) = 132.24836$, $\sum t_i \log(t_i) = 40870.268$, $\sum t_i (\log(t_i))^2 = 243502.91$. Then the restricted MLE of λ under H_0 is $\tilde{\lambda} = 1/\bar{t} = 0.0036023$. So the score $U_1(1, \tilde{\lambda}) = 25 - 0.0036023 * 40870.268 + 132.24836 = 10$, the information matrix and its inverse under H_0 are

$$I_n(1,\tilde{\lambda}) = \begin{bmatrix} 902.17186 & 40870.268 \\ 40870.268 & 1926544 \end{bmatrix}, \quad I_n^{-1}(1,\tilde{\lambda}) = \begin{bmatrix} 0.0284592 & -0.000604 \\ -0.000604 & 0.0000133 \end{bmatrix}.$$

Therefore the score statistic is $\chi^2 = 10 * 0.0284592 * 10 = 2.8$ and the p-value is 0.09.

3.2.2 Likelihood construction for censored survival data

Suppose we have a random sample of n individuals from a specific population whose survival times are $T_1, T_2, ..., T_n$. However, due to right censoring such as staggered entry, loss to follow-up, competing risks (death from other causes) or any combination of these, we do not always have the opportunity of observing these survival times. Denote by C the censoring process and by $C_1, C_2, ..., C_n$ the (potential) censoring times. Thus if a subject is not censored we have observed his/her survival time (in this case, we may not observe the censoring time for this individual), otherwise we have observed his/her censoring time (survival time is larger than the censoring time). In other words, the observed data are the minimum of the survival time and censoring time for every subject in the sample and the indication whether or not the subject is censored.

Statistically, we have observed data $(\tilde{T}_i, \Delta_i)$, $i = 1, 2, ..., n$, where

$$\begin{aligned} \tilde{T}_i &= \min(T_i, C_i), \\ \Delta_i &= I(T_i \leq C_i) = \begin{cases} 1 & \text{if } T_i \leq C_i \text{ (observed failure)} \\ 0 & \text{if } T_i > C_i \text{ (observed censoring)} \end{cases} \end{aligned}$$

Namely, the potential data are $\{(T_1, C_1), (T_2, C_2), ..., (T_n, C_n)\}$, but the actual observed data are $\{(\tilde{T}_1, \Delta_1), (\tilde{T}_2, \Delta_2), ..., (\tilde{T}_n, \Delta_n)\}$.

Of course we are interested in making inference on the random variable T, *i.e.*, any one of the following functions

$$\begin{aligned} f(t) &= \text{density function} \\ F(t) &= \text{distribution function} \\ S(t) &= \text{survival function} \\ \lambda(t) &= \text{hazard function} \end{aligned}$$

Since we need to work with our data: $\{(\tilde{T}_1, \Delta_1), (\tilde{T}_2, \Delta_2), ..., (\tilde{T}_n, \Delta_n)\}$, we define the following corresponding functions for the censoring time C:

$$\begin{aligned} g(t) &= \text{density function} \\ G(t) &= \text{distribution function} = P[C \leq t] \\ H(t) &= \text{survival function} = P[C \geq t] = 1 - G(t) \\ \mu(t) &= \text{hazard function} = \frac{g(t)}{H(t)} \end{aligned}$$

Usually, the density function $f(t)$ of T may be governed by some parameters θ and $g(t)$ by some other parameters ϕ. In these cases, we are interested in making inference on θ.

In order to derive the density of $(\tilde{T}, \Delta)$, we assume independent censoring, *i.e.*, random variables T and C are independent. The density function of $(\tilde{T}, \Delta)$ is defined as

$$f(t, \delta) = \lim_{h \to 0} \frac{P[t \leq \tilde{T} < t + h, \Delta = \delta]}{h}, \quad t \geq 0, \delta = \{0, 1\}.$$

Note: Do not mix up the density $f(t)$ of T and $f(t,\delta)$ of $(\tilde{T},\Delta)$. If we want to be more specific, we will use $f_T(t)$ for T and $f_{\tilde{T},\Delta}(t,\delta)$ for $(\tilde{T},\Delta)$. But when there is no ambiguity, we will suppress the subscripts.

1. Case 1: $\delta = 1$, *i.e.*, $T \le C$, $\tilde{T} = \min(T, C) = T$, we have

$$\begin{aligned} & P[t \le \tilde{T} < t+h, \Delta = 1] \\ = \; & P[t \le T < t+h, C \ge T] \\ \approx \; & P[t \le T < t+h, C \ge t] \;\; \text{(Note: } t \text{ is a fixed number)} \\ = \; & P[t \le T < t+h] \cdot P[C \ge t] \;\; \text{(by independence of } T \text{ and } C) \\ = \; & f_T(\xi) h H_C(t), \;\; \xi \in [t, t+h), \;\; \text{(Note: } H(t) \text{ is the survival function of } C). \end{aligned}$$

   Therefore

$$\begin{aligned} f(t, \delta = 1) &= \lim_{h \to 0} \frac{P[t \le \tilde{T} < t+h, \Delta = 1]}{h} \\ &= \lim_{h \to 0} \frac{f_T(\xi) h H_C(t)}{h} \\ &= f_T(t) H_C(t). \end{aligned}$$

2. Case 2: $\delta = 0$, *i.e.*, $T > C$, $\tilde{T} = \min(T, C) = C$, we have

$$\begin{aligned} & P[t \le \tilde{T} < t+h, \Delta = 0] \\ = \; & P[t \le C < t+h, T > C] \\ \approx \; & P[t \le C < t+h, T \ge t] \;\; \text{(Note: } t \text{ is a fixed number)} \\ = \; & P[t \le C < t+h] \cdot P[T \ge t] \;\; \text{(by independence of } T \text{ and } C) \\ = \; & g_C(\xi) h S_T(t), \;\; \xi \in [t, t+h). \end{aligned}$$

   Therefore

$$\begin{aligned} f(t, \delta = 0) &= \lim_{h \to 0} \frac{P[t \le \tilde{T} < t+h, \Delta = 0]}{h} \\ &= \lim_{h \to 0} \frac{g_C(\xi) h S_T(t)}{h} \\ &= g_C(t) S_T(t). \end{aligned}$$

Combining these two cases, we have the density function of $(\tilde{T}, \Delta)$:

$$\begin{aligned} f(t,\delta) &= [f_T(t)H_C(t)]^\delta [g_C(t)S_T(t)]^{1-\delta} \\ &= \{[f_T(t)]^\delta [S_T(t)]^{1-\delta}\}\{[g_C(t)]^{1-\delta}[H_C(t)]^\delta\}. \end{aligned}$$

Sometimes it may be useful to use hazard functions. Recalling that the hazard function

$$\lambda_T(t) = \frac{f_T(t)}{S_T(t)}, \text{ or } f_T(t) = \lambda_T(t)S_T(t),$$

we can write $[f_T(t)]^\delta [S_T(t)]^{1-\delta}$ as

$$[f_T(t)]^\delta [S_T(t)]^{1-\delta} = [\lambda_T(t)S_T(t)]^\delta [S_T(t)]^{1-\delta} = [\lambda_T(t)]^\delta [S_T(t)].$$

Another useful way of defining the distribution of the random variable $(\tilde{T}, \Delta)$ is through the cause-specific hazard function.

Definition: The cause-specific hazard function is defined as

$$\lambda(t,\delta) = \lim_{h\to 0} \frac{P[t \le \tilde{T} < t+h, \Delta = \delta | \tilde{T} \ge t]}{h}.$$

For example, $\lambda(t, \delta = 1)$ corresponds to the probability rate of observing a failure at time t given an individual is at risk at time t (*i.e.*, neither failed nor was censored prior to time t).

If T and C are statistically independent, then through the following calculations, we obtain

$$\begin{aligned} P[t \le \tilde{T} < t+h, \Delta = \delta | \tilde{T} \ge t] &= \frac{P[(t \le \tilde{T} < t+h, \Delta = \delta) \cap (\tilde{T} \ge t)]}{P[\tilde{T} \ge t]} \\ &= \frac{P[t \le \tilde{T} < t+h, \Delta = \delta]}{P[\tilde{T} \ge t]}. \end{aligned}$$

Hence

$$\lambda(t, \delta = 1) = \frac{\lim_{h\to 0} \frac{P[t \le \tilde{T} < t+h, \Delta = 1]}{h}}{P[\tilde{T} \ge t]} = \frac{f(t, \delta = 1)}{P[\tilde{T} \ge t]}.$$

Since $f(t, \delta = 1) = f_T(t)H_C(t)$ and

$$\begin{aligned} P[\tilde{T} \ge t] &= P[\min(T, C) \ge t] = P[(T \ge t) \cap (C \ge t)] \\ &= P[T \ge t] \cdot P[C \ge t] \text{ (by independence of } T \text{ and } C) \\ &= S_T(t)H_C(t). \end{aligned}$$

Therefore,

$$\lambda(t, \delta = 1) = \frac{f_T(t)H_C(t)}{S_T(t)H_C(t)} = \frac{f_T(t)}{S_T(t)} = \lambda_T(t).$$

Remark:

- This last statement is very important. It says that if T and C are independent then the cause-specific hazard for failing (of the observed data) is the same as the underlying hazard of failing for the variable T we are interested in. This result was used implicitly when constructing the life-table, Kaplan-Meier and Nelson-Aalen estimators in later lectures.

- If the cause-specific hazard of failing is equal to the hazard of underlying failure time, the censoring process is said to be *non-informative*. Except for some pathological examples, non-informative censoring is "equivalent to" independent censoring.

- We assumed independent censoring when we derive the density function for $(\tilde{T}, \Delta)$ and the cause-specific hazard. All results depend on this assumption. If this assumption is violated, all the inferential methods will yield biased results.

- To make matters more complex, we cannot tell whether or not T and C are independent based on the observed data $(\tilde{T}_i, \Delta_i)$, $i = 1, 2, ..., n$. This is an inherent non-identifiability problem; See Tsiatis (1975) in Proceeding of the National Academy of Science.

- To complete, if T and C are independent, then

$$\lambda(t, \delta = 0) = \lambda_C(t).$$

Likelihood for right censoring case

Now we are in a position to write down the likelihood function for a parametric model given our observed data $(\tilde{t}_i, \delta_i)$ (under independence of T and C): $i = 1, 2, ..., n$.

$$L(\theta, \phi) = \prod_{i=1}^{n} \{[f(\tilde{t}_i; \theta)]^{\delta_i}[S(\tilde{t}_i; \theta)]^{1-\delta_i}\}\{[g(\tilde{t}_i; \phi)]^{1-\delta_i}[H(\tilde{t}_i; \phi)]^{\delta_i}\}.$$

Keep in mind that we are mainly interested in making inference on the parameters θ characterizing the distribution of T. So if θ and ϕ have **no** common parameters, we can use the following likelihood function to make inference on θ:

$$L(\theta; \tilde{t}, \delta) = \prod_{i=1}^{n} [f(\tilde{t}_i; \theta)]^{\delta_i} [S(\tilde{t}_i; \theta)]^{1-\delta_i}. \tag{6}$$

Or equivalently,

$$L(\theta; \tilde{t}, \delta) = \prod_{i=1}^{n} [\lambda(\tilde{t}_i; \theta)]^{\delta_i} [S(\tilde{t}_i; \theta)]. \tag{7}$$

Note: Even if θ and ϕ may have common parameters, we can still use (6) or (7) to draw **valid** inference on θ. Of course, we may lose some efficiency in this case.

Likelihood for general censoring case

The likelihood function (6) has the following form

$$L(\theta; \tilde{t}, \delta) = \prod_{d \in D} f(\tilde{t}_d) \prod_{r \in R} S(\tilde{t}_r), \tag{8}$$

where D is the index set of death times, R is the index set of right censored times. For a death time $\tilde{t}_d$, $f(\tilde{t}_d)$ is proportional to the probability of observing a death at time $\tilde{t}_d$. For a right censored observation $\tilde{t}_r$, the only thing we know is that the real survival time T_r is greater than $\tilde{t}_r$. Hence we have $P[T_r > \tilde{t}_r] = S_T(\tilde{t}_r)$, the probability that the real survival time T_r is greater than $\tilde{t}_r$, for a right censored observation.

The above likelihood can be generalized to the case where there might be any kind of censoring:

$$L(\theta; \tilde{t}, \delta) = \prod_{d \in D} f(\tilde{t}_d) \prod_{r \in R} S(\tilde{t}_r) \prod_{l \in L} [1 - S(\tilde{t}_l)] \prod_{i \in I} [S(u_i) - S(v_i)], \tag{9}$$

where L is the index set of left censored observations, I is the index set of interval censored observations with the only knowledge that the real survival time T_i is in the interval $[u_i, v_i]$. Note that $S(u_i) - S(v_i) = P[u_i \leq T_i \leq v_i]$ is the probability that the real survival time T_i is in $[u_i, v_i]$.

Likelihood for left truncated observations

Suppose now that the real survival time T_i is left truncated at Y_i. The conditional density of T_i given that $T_i \geq Y_i$ is given by

$$f(t|T_i \geq Y_i) = \frac{f(t)}{P[T_i \geq Y_i]} = \frac{f(t)}{S(Y_i)}. \tag{10}$$

Therefore, the probability to observe a death at $\tilde{t}_d$ is proportional to $f(\tilde{t}_d)/S(y_d)$. The probability that the survival time T_r is right censored at $\tilde{t}_r$ ($\tilde{t}_r \geq y_r$) is

$$P[T_r \geq \tilde{t}_r|T_r \geq y_r] = S(\tilde{t}_r)/S(y_r).$$

The probability that the survival time T_l is left censored at $\tilde{t}_l$ ($\tilde{t}_l \geq y_l$) is

$$P[T_l \leq \tilde{t}_l|T_l \geq y_l] = [S(y_l) - S(\tilde{t}_l)]/S(y_l).$$

And the probability that the real survival time T_i is in $[u_i, v_i]$ ($u_i \geq y_i$) is

$$P(u_i \leq T_i \leq v_i|T_i \geq y_i) = P(T_i \geq u_i|T_i \geq y_i) - P(T_i \geq v_i|T_i \geq y_i) = [S(u_i) - S(v_i)]/S(y_i).$$

In this case, the likelihood function is given by

$$L(\theta; \tilde{t}, y, \delta) = \prod_{d\in D} \frac{f(\tilde{t}_d)}{S(y_d)} \prod_{r\in R} \frac{S(\tilde{t}_r)}{S(y_r)} \prod_{l\in L} \frac{[S(y_l) - S(\tilde{t}_l)]}{S(y_l)} \prod_{i\in I} \frac{[S(u_i) - S(v_i)]}{S(y_i)} \tag{11}$$

$$= \left[\prod_{d\in D} f(\tilde{t}_d) \prod_{r\in R} S(\tilde{t}_r) \prod_{l\in L} (S(y_l) - S(\tilde{t}_l)) \prod_{i\in I} (S(u_i) - S(v_i))\right] \Big/ \prod_{i=1}^{n} S(y_i) \tag{12}$$

Likelihood for right truncated observations

We consider the special case of right truncation, that is, only deaths are observed. In this case the probability to observe a death at time t given that T_i is less than or equal to Y_i ($t \leq Y_i$) is proportional to $f(t)/(1 - S(Y_i))$. So the likelihood function is

$$L(\theta; T, Y, \delta) = \prod_{i=1}^{n} \frac{f(T_i)}{1 - S(Y_i)}. \tag{13}$$

An Example of right censored data: Suppose the underlying survival time T is from an exponential distribution with parameter λ (here the parameter θ is λ) and we have observed

data: $(\tilde{t}_i, \delta_i)$, $i = 1, 2, ..., n$. The likelihood function of the observed data is

$$L(\lambda; \tilde{t}, \delta) = \prod_{i=1}^{n} \lambda^{\delta_i} \mathrm{e}^{-\lambda \tilde{t}_i} = \lambda^{\sum_{i=1}^n \delta_i} \mathrm{e}^{-\lambda \sum_{i=1}^n \tilde{t}_i}.$$

So the log-likelihood is

$$\ell(\lambda; \tilde{t}, \delta) = \log(\lambda) \sum_{i=1}^{n} \delta_i - \lambda \sum_{i=1}^{n} \tilde{t}_i.$$

Obviously, the likelihood equation is

$$U(\lambda; \tilde{t}, \delta) = \frac{d\ell(\lambda; \tilde{t}, \delta)}{d\lambda} = \frac{\sum_{i=1}^n \delta_i}{\lambda} - \sum_{i=1}^{n} \tilde{t}_i = 0.$$

So the MLE of λ is given by

$$\hat{\lambda} = \frac{\sum_{i=1}^n \delta_i}{\sum_{i=1}^n \tilde{t}_i} = \frac{\text{\# of failures}}{\text{person time at risk}} = \frac{D}{PT},$$

where D is the number of observed deaths and PT is the total patient time. Since

$$\frac{d^2\ell(\lambda; \tilde{t}, \delta)}{d\lambda^2} = -\frac{\sum_{i=1}^n \delta_i}{\lambda^2},$$

the estimated variance for $\hat{\lambda}$ is

$$\widehat{\mathrm{Var}}(\hat{\lambda}) = -\left[\frac{d^2\ell(\lambda; \tilde{t}, \delta)}{d\lambda^2}\bigg|_{\lambda=\hat{\lambda}} \right]^{-1} = \frac{\sum_{i=1}^n \delta_i}{\left[\sum_{i=1}^n \tilde{t}_i\right]^2} = \frac{\hat{\lambda}^2}{D},$$

and asymptotically, we have

$$\hat{\lambda} \overset{a}{\sim} \mathrm{N}\left(\lambda, \frac{\sum_{i=1}^n \delta_i}{\left[\sum_{i=1}^n \tilde{t}_i\right]^2}\right) = \mathrm{N}\left(\lambda, \frac{\hat{\lambda}^2}{D}\right).$$

This result can be used to construct confidence interval for λ or perform hypothesis testing on λ. For example, a $(1-\alpha)$ confidence interval for λ is given by

$$\hat{\lambda} \pm z_{\alpha/2} \frac{\hat{\lambda}}{\sqrt{D}}.$$

Note:

- Sometimes the exponential distribution is parameterized in terms of the mean parameter $\theta = 1/\lambda$. In this case the MLE of θ is given by

$$\hat{\theta} = \frac{\sum_{i=1}^n \tilde{t}_i}{\sum_{i=1}^n \delta_i} = \frac{\text{total person time at risk}}{\text{\# of failures}} = \frac{PT}{D},$$

and asymptotically,

$$\hat{\theta} \overset{a}{\sim} \mathrm{N}\left(\theta, \frac{\hat{\theta}^2}{D}\right).$$

(The estimated variance of $\hat{\theta}$ can be obtained by inverting the observed information or using delta-method.)

- If we ignored censoring and treated the data $\tilde{t}_1, \tilde{t}_2, ..., \tilde{t}_n$ from the exponential distribution, then the "MLE" of θ would be

$$\tilde{\theta} = \frac{\sum_{i=1}^{n} \tilde{t}_i}{n},$$

which, depending on the percentage of censoring, would severely underestimate the true mean (note that the sample size n is always larger than D, the number of deaths).

A Data Example: The data below show survival times (in months) of patients with certain disease

$$3, 5, 6^*, 8, 10^*, 11^*, 15, 20^*, 22, 23, 27^*, 29, 32, 35, 40, 26, 28, 33^*, 21, 24^*,$$

where * indicates right censored data. If we fit exponential model to this data set, we have $D = 13$ and $PT = \sum x_i = 418$, so

$$\hat{\lambda} = \frac{D}{PT} = \frac{13}{418} = 0.0311/\text{month},$$

and the estimated standard error of $\hat{\lambda}$ is

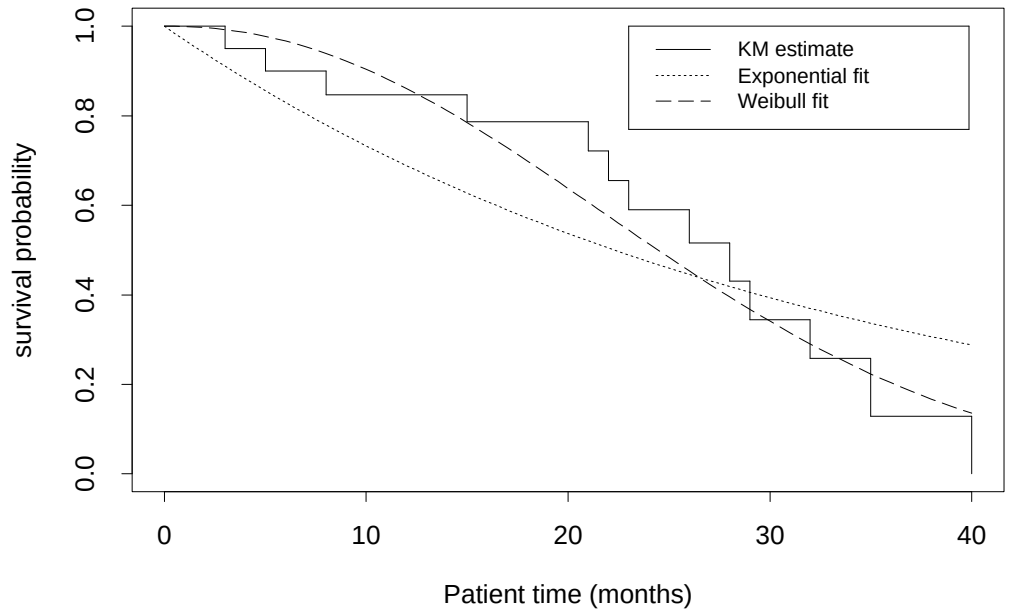
$$\text{se}(\hat{\lambda}) = \frac{\hat{\lambda}}{\sqrt{D}} = \frac{0.0311}{\sqrt{13}} = 0.0086,$$

and a 95% confidence interval of λ is

$$\hat{\lambda} \pm z_{0.025} \cdot \text{se}(\hat{\lambda}) = 0.0311 \pm 1.96 \cdot 0.0086 = [0.0142, 0.0480].$$

To see how well the exponential model fits the data, the fitted exponential survival function is superimposed to the Kaplan-Meier estimate as shown in Figure 3 using the following R functions:

Figure 3: *Three fits to the survival data*




```
> example <- read.table(file="tempsurv.dat", header=T)

> fit <- survfit(Surv(survtime, status), conf.type=c("plain"), example)
> plot(0,0, xlim=c(0,40), ylim=c(0,1),
    xlab="Patient time (months)", ylab="survival probability", pch=" ")
> lines(fit, lty=1)
> x <- seq(0,40, by=0.5)
> sx <- exp(-0.0311*x)
> lines(x, sx, lty=2)
```

where the data file `tempsurv.dat` looks like the following

```
survtime status
3  1
5  1
6  0
8  1
10 0
11 0
15 1
20 0
22 1
23 1
27 0
29 1
32 1
35 1
```

```
40 1
26 1
28 1
33 0
21 1
24 0
```

Obviously, the exponential distribution is a poor fit. In this case, we can choose one of the following options

- Choose a more flexible model, such as the Weibull model.
- Be content with the Kaplan-Meier estimator which makes no assumption regarding the shape of the distribution. In most biomedical applications, the default is to go with the Kaplan-Meier estimator.

To complete, we fit a Weibull model to the data set. Recall that Weibull model has the following survival function $S(t) = \mathrm{e}^{-\lambda t^\alpha}$ and the hazard function $\lambda(t) = \alpha\lambda t^{\alpha-1}$. So the likelihood function of $\theta = (\lambda, \alpha)$ is given by

$$L(\lambda, \alpha; x, \delta) = \prod_{i=1}^{n} \left[\alpha\lambda \tilde{t}_i^{\alpha-1}\right]^{\delta_i} \mathrm{e}^{-\lambda \tilde{t}_i^{\alpha}}.$$

However, there is no closed form for the MLEs of $\theta = (\lambda, \alpha)$. So we used `Proc Lifereg` in `SAS` to fit Weibull model implemented using the following `SAS` program

```
options ls=80 ps=200;

Data tempsurv;
  infile "tempsurv.dat" firstobs=2;
  input survtime status;
run;

Proc lifereg data=tempsurv;
  model survtime*status(0)= / dist=weibull;
run;
```

The above program produced the following output:

```
                                        10:41 Friday, January 28, 2005
```

The LIFEREG Procedure

Model Information

Data Set	WORK.TEMPSURV
Dependent Variable	Log(survtime)
Censoring Variable	status
Censoring Value(s)	0
Number of Observations	20
Noncensored Values	13
Right Censored Values	7
Left Censored Values	0
Interval Censored Values	0
Name of Distribution	Weibull
Log Likelihood	-16.67769141

Algorithm converged.

Analysis of Parameter Estimates

Parameter	DF	Estimate	Standard Error	95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	3.3672	0.1291	3.1141	3.6203	679.81	<.0001
Scale	1	0.4653	0.1087	0.2943	0.7355		
Weibull Scale	1	28.9964	3.7447	22.5121	37.3483		
Weibull Shape	1	2.1494	0.5023	1.3596	3.3979		

This SAS program fits a Weibull model with two parameters: intercept β_0 and a scale parameter σ. Two parameters we use λ and α are related to β_0 and σ by (the detail will be discussed in later lectures)

$$\lambda = \mathrm{e}^{-\beta_0/\sigma} \text{ and } \alpha = \frac{1}{\sigma}.$$

Since the MLE of β_0 and σ are $\hat{\beta}_0 = 3.36717004$ and $\hat{\sigma} = 0.46525153$, the MLEs $\hat{\lambda}$ and $\hat{\alpha}$ are

$$\begin{aligned} \hat{\lambda} &= \mathrm{e}^{-\hat{\beta}_0/\hat{\sigma}} = \mathrm{e}^{-3.36717004/0.46525153} = 0.00072, \\ \hat{\alpha} &= \frac{1}{\hat{\sigma}} = \frac{1}{0.46525153} = 2.149. \end{aligned}$$

So $\hat{\alpha}$ is the `Weibull Shape` parameter in the SAS output. However, SAS uses the parameterization $S(t) = e^{-(t/\tau)^{\alpha}}$ for Weibull distribution so that τ is the `Weibull scale` parameter.

Comparing this to our parameterization, we see that

$$\left(\frac{1}{\tau}\right)^{\alpha} = \lambda, \implies \tau = \left(\frac{1}{\lambda}\right)^{1/\alpha}.$$

The estimate of this `Weibull scale` parameter is

$$\widehat{\tau} = \left(\frac{1}{0.00072}\right)^{1/2.149} = 28.996.$$

The fitted Weibull survival function was superimposed to the Kaplan-Meier estimator in Figure 3 using the the following R functions

```
> alpha <- 1/0.46525153
> lambda <- exp(-3.36717004/0.46525153)
> sx <- exp(-lambda * x^alpha)
# the object "x" was created before
> lines(x, sx, lty=4)
> legend(25,1, c("KM estimate", "Exponential fit", "Weibull fit"),
lty=c(1,2,4), cex=0.8)
```

Compared to the exponential fit, the Weibull model fits the data much better (since its estimated survival function tracks the Kaplan-Meier estimator much better than the estimated exponential survival function). In fact, since the exponential model is a special case of the Weibull model (when $\alpha = 1$), we can test $H_0 : \alpha = 1$ using the Weibull fit. Note that $H_0 : \alpha = 1$ is equivalent to $H_0 : \sigma = 1$. Since

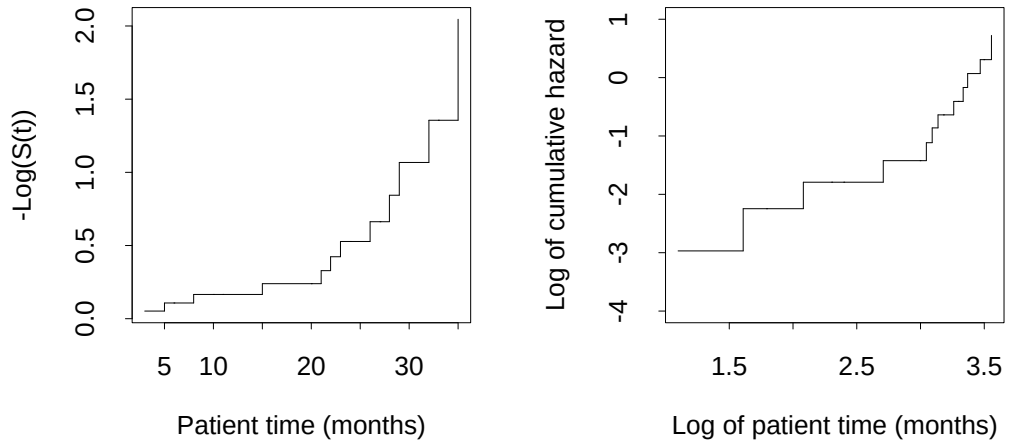
$$\left(\frac{\hat{\sigma} - 1}{\text{se}(\hat{\sigma})}\right)^2 = \left(\frac{0.46525153 - 1}{0.108717}\right)^2 = 24.194,$$

and $P[\chi^2 > 24.194] = 0.0000$, we reject $H_0 : \alpha = 1$, *i.e.*, we reject the exponential model. Note also that $\hat{\alpha} = 2.149 > 1$, so the estimated Weibull model has an increasing hazard function.

The inadequacy of the exponential fit is also demonstrated in the first plot of Figure 4. If the exponential model were a good fit to the data, we would see a straight line. On the other hand, plot 2 in Figure 4 shows the adequacy of the Weibull model, since a straight line of the plot of $\log\{-\log(\hat{S}(t))\}$ vs. $\log(t)$ indicates a Weibull model. Here $\hat{S}(t)$ is the KM estimate.

This graph was plotted using the following R codes:

Figure 4: *Two empirical plots*




```
postscript(file="fig4.2.ps", horizontal = F,
  height=6, width=8.5, font=3, pointsize=14)
  par(mfrow=c(1,2), pty="s")

example <- read.table(file="tempsurv.dat", header=T)

fit <- survfit(Surv(survtime, status), conf.type=c("plain"),
example)

plot(fit$time, -log(fit$surv), type="s", xlab=c("Patient time
(months)"), ylab=c("-Log(S(t))"))

plot(log(fit$time), log(-log(fit$surv)), type="s", ylim=c(-4,1),
xlab=c("Log of patient time (months)"), ylab=c("Log of cumulative
hazard")) dev.off()
```