Models for Survival Data:
Discrete lifetime

1. Survival, Hazard, Mean Residual Lifetime and Median Lifetime functions
2. Kaplan-Meier Product Limit formula
   2a. Uncensored observations

   Survival Analysis
   Ch.2, Sec. 1-5
Exercises: 2 # 2, 12, 15 (c, d), 16, 17
1. Discrete Models

Let $X$ be the time until some specified event.

*event: death, the appearance of a tumor, the development of some disease, equipment breakdown.

Suppose the r.v. $X$ is discrete positive r.v.

\[ X \in \{ x_1, x_2, \ldots, x_k, \ldots \} \]

with corresponding probability mass function:

\[ p(x_j) \quad : \quad p(x_j) = P(X = x_j) \]

Four functions characterize the distribution of $X$:

(i) Survival function: \[ S(x) = P(X > x) \]

is the probability of an individual surviving beyond time $x$.

(ii) Hazard rate function: \[ h(x_j) = \frac{p(x_j)}{S(x_{j-1})} \]

\[ h(x_j) = P(X = x_j \mid X > x_j) \]

(iii) Probability mass function: \[ p(x_j) = P(X = x_j) \]

(iv) Mean residual life at time $x$: \[ \text{mr}l(x) = e(x) \]
(i) The survival function is expressed as

\[ S(x) = P(X > x) = \sum_{x_j > x} p(x_j) \]

**Example 1.** Let us consider the uniform model:

\[ p(x_j) = P(X = j) = \frac{1}{3}, \quad j = 1, 2, 3 \]

\[ S(x) = P(X > x) = \begin{cases} 
1 & 0 < x < 1 \\
\frac{2}{3} & 1 \leq x < 2 \\
\frac{1}{3} & 2 \leq x < 3 \\
0 & x \geq 3
\end{cases} \]

\[ S(x) - \text{nonincreasing step function} \]

**Example 2.** Suppose that \( X \) has a geometric distribution with pmf:

\[ p(x) = p q^{x-1}, \quad x = 1, 2, \ldots, \quad q = 1 - p \]

Find the survival function \( S(x) \).
\[ S^*(x) = P(X > x) = \sum_{j' = x+1}^{\infty} p(j') = \sum_{j' = x+1}^{\infty} p \cdot q^{j'-1} = \]
\[ = p \sum_{k = x}^{\infty} q^{k} = p \frac{q^x}{1-q} = q^x \]

\[ h_{(j)} = h_{(x_j)} = \frac{p(x_j)}{S(x_{j-1})} = p \cdot q^{j-1} = p \]

\[ h(j) = p = \text{const} \]

as in the exponential model.

(1) \textbf{Hazard function:}

\[ h(x_j) = P(X = x_j \mid X \geq x_j) = \]

\[ = \frac{P(X = x_j \cap X \geq x_j)}{P(X \geq x_j)} = \frac{P(X = x_j)}{P(X > x_{j-1})} = S'(x_{j-1}) \]

\textbf{Remark:} Since

\[ p(x_j) = P(X = x_j) = P(X > x_{j-1}) - P(X > x_j) \]

\[ = S'(x_{j-1}) - S^*(x_j) \]

we can rewrite

\[ h(x_j) = \frac{S'(x_{j-1}) - S'(x_j)}{S'(x_{j-1})} = 1 - \frac{S'(x_j)}{S'(x_{j-1})} \]
Here:

$$
1 - h(x_j) = \frac{S(x_j)}{S(x_{j-1})}, \quad j = 1, 2, \ldots
$$

(iv) $m - l(x) = E(X - x \mid X > x) = e(x) = \int_x^\infty \frac{s(y) dy}{S(y)}$

\[ \frac{\text{area under } S \text{ to the right of } x}{S(x)} \]

\[ \frac{1}{S(x)} \left[ (x_{i+1} - x) S(x_i) + \Delta x_{i+1} S(x_{i+1}) + \Delta x_{i+2} S(x_{i+2}) + \ldots \right] \]

\[ \frac{1}{S(x)} \left[ (x_{i+1} - x) S(x_i) + \sum_{j = i+1}^{\infty} \Delta x_j S(x_j) \right] \]

if $x \in [x_i, x_{i+1})$

**Remark** Mean life $E(X) = \mu = \int_0^\infty S(x) dx = \text{the total area under the survival curve}$
The median lifetime of \( X \): \( m(X) = m \)

\[
F(m) = .5 \quad \text{and} \quad S'(m) = .5
\]

If \( X \) is continuous then \( m \) is defined as

\[
S(m) = .5
\]

**Example 3.** \( X \sim \text{Exp}(\lambda) \)

**Mean lifetime:** \( \mu = \frac{1}{\lambda} \)

**Median lifetime:** \( m = \frac{\ln 2}{\lambda} \)

Indeed:

\[
S'(m) = 2^{-1}
\]

\[
e^{-2m} = 2
\]

\[
-2m = -\ln 2
\]

\[
m = \frac{\ln 2}{2}
\]

**Mean** \( \mu = E(X) = 2 \int_{0}^{\infty} x e^{-x} \, dx = \frac{1}{\lambda} \)
Example 4. Let us consider $X$ from Example 1.

$p(x_j) = P(X = j) = \frac{1}{3}, \quad j = 1, 2, 3.$

The hazard function

$$h(x_j) = \begin{cases} \frac{1}{3}, & j = 1 \\ \frac{1}{2}, & j = 2 \\ 1, & j = 3 \\ 0 & \text{elsewhere} \end{cases}$$

Indeed, $h(x_1) = \frac{P(X = 1)}{S(0)} = \frac{1/3}{3} = \frac{1}{3}$

$h(x_2) = h(2) = \frac{P(X = 2)}{S(2)} = \frac{1/2}{\frac{1}{3}} = \frac{1}{2}$

$h(x_3) = h(3) = \frac{P(X = 3)}{S(2)} = \frac{1/3}{\frac{1}{3}} = 1$

Remark. Hazard rate function is zero except at points where a failure could occur.
1. **Product limit formula for \( S(x) \).**

**Lemma 1.**

\[
S^*(x) = \prod_{x_j \leq x} \frac{S(x_j)}{S(x_j^{-1})}
\]  

\((5)\)

**Proof:**

Note that

\[
P(X > x_k \mid X > x_{k-1}) = \frac{S'(x_k)}{S(x_{k-1})}
\]

So that

\[
S^*(x) = P(X > x) = P(X > x \mid X > x_{k}) P(X > x_k) = \]

\[
= S(x_k) \times \frac{S'(x)}{S(x_k)} = S(x_{k-1}) \times \frac{S'(x_k)}{S(x_{k-1})} \times \frac{S(x)}{S(x_k)} = \]

\[
= \ldots = \prod_{x_j \leq x} \frac{S'(x_j)}{S(x_j^{-1})}, \quad S(x_0) = 1, \quad x_0 = 0
\]

From (4), we conclude that the survival function is related to the hazard function by

\[
S^*(x) = \prod_{x_j \leq x} (1 - h(x_j))
\]

\((6)\)

**Remark:**

\[
S(x) = \prod_{x_j > x} p(x_j) = \prod_{x_j \leq x} (1 - h(x_j))
\]
Let $X$ be a continuous lifetime

Some useful formulae:

$$S'(x) = e^{-\int_0^x h(u) \, du}$$

$$S'(x) = \frac{e(\theta)}{e(x)} e^{-\int_0^x \frac{du}{e(u)}}$$

$$h(x) = \frac{f(x)}{S(x)} = \frac{\left(\frac{df}{dx} e(x) + 1\right)}{e(x)}$$

Not that if $X$ is a discrete lifetime, then

$$e(x) = \frac{(x_{i+1} - x) S(x_i) + \sum_{j \geq i+1} \Delta(x_{j+1} - x_j) S(x_j)}{S'(x)}$$

$$h(x_i) = \frac{e(x_i)}{S'(x_{i-1})} \quad \text{and} \quad S'(x) = \prod_{x_j < x} \left(1 - h(x_j)\right)$$
Let us prove (8) & (7):

**Lemma 1.** \( h(x) = \frac{e'(x) + 1}{e(x)} \)

**Proof:**

\[
e'(x) = \left[ \frac{\int_{\infty}^{x} S'(u) du}{S(x)} \right]^{1} =
\]

\[
e'(x) = \frac{1}{S^2(x)} \left[ -S(x) \times S'(x) - \int_{x}^{\infty} S'(w) du \right. (-f(x)) \right]
\]

i.e.

\[
e'(x) = -1 + e(x) \frac{f(x)}{S(x)}
\]

Hence

\[
h(x) = \frac{e'(x) + 1}{e(x)}
\]

**Lemma 2.** \( S(x) = \frac{e(0)}{e(x)} \int_{0}^{x} \frac{du}{e(u)} \)

**Proof:** Since

\[
-\int_{0}^{x} h(u) du = -\int_{0}^{x} \frac{e'(u)}{e(u)} \frac{d}{e(u)} = -\int_{0}^{x} \frac{du}{e(u)} = -h \frac{e(0)}{e(x)} - \int_{0}^{x} \frac{du}{e(u)}
\]

We can rewrite \( S(x) \) in the following way:

\[
S(x) = e^{-\int_{0}^{x} h(u) du} = e^{-\int_{0}^{x} \frac{du}{e(u)}} = \frac{e(0)}{e(x)} e^{-\int_{0}^{x} \frac{du}{e(u)}}
\]
Definition 1. \( X \) is said to follow the log logistic distribution with parameters \( \mu \) and \( \sigma \) if

\[
f_Y(y) = \frac{1}{\sigma} \frac{y - \mu}{1 + \frac{y - \mu}{\sigma}}^{-2}, \quad -\infty < y < \infty
\]

\[M = \mathbb{E}[Y] \text{ and } \sigma^2 \] is the scale parameter of \( Y \).

We can represent \( Y \) as follows:

\[Y = \mu + \sigma W\]

with

\[W \sim \text{log logistic with parameters } \mu = 0 \text{ and } \sigma = 1\]

Hazard function of \( X \) is given by

\[h(x) = \frac{\alpha x^{\alpha-1}}{1 + \alpha x^\alpha} \quad \text{where } \alpha = \frac{1}{\sigma} > 0 \]

\[a = e^{-\frac{M}{\sigma}} \]

and survival function

\[S(x) = \frac{1}{1 + \alpha x^\alpha}.\]
Definition 2: Median residual lifetime $m(x)$:

$$F(m(x) \mid x) = 0.5$$

where

$$F(y \mid x) = P(X-x < y \mid X \geq x) = \frac{F(X+y)-F(x)}{S(x)}$$

is conditional cdf of residual $X-x$ lifetime.

Remark: When $x = 0$, $F(y \mid x) = F(y) - \text{unconditional cdf of } X$. 
Exerc. 2 #1. Assume $X$ - lifetime of light bulbs and
$X \sim \text{Exp}(\lambda)$, i.e. $F(x) = 1 - e^{-\lambda x}$
with $\lambda = 0.001$ failures per hour

$h(x) = \lambda = 0.001$

(a) Mean lifetime of selected light bulb:

$E(X) = \frac{1}{\lambda} = 1000$

(b) Median lifetime: $m$:

$S(m) = \frac{1}{2}$

$e^{-\lambda m} = 2^{-1} \Rightarrow -\lambda m = -\ln 2 \Rightarrow m = \frac{\ln 2}{\lambda} = \frac{\ln 2}{0.001} = 693.15$

(c) $P(X > 2000) = S(2000) = \frac{2e^{-2000\lambda}}{\lambda} = e^{-2 \cdot 2000} = e^{-1} \approx 0.368$
Exercise 2.15.

\[ p(x_1) = S(x_0) - S(x_1) \]
\[ p(6) = 1 - 0.55 = 0.45 \]

\[ p(12) = 0.55 - 0.43 = 0.12 \]
\[ p(18) = 0.43 - 0.34 = 0.09 \]
\[ p(24) = 0.34 - 0.30 = 0.04 \]
\[ p(30) = 0.30 - 0.25 = 0.05 \]
\[ p(36) = 0.25 - 0.18 = 0.07 \]
\[ p(42) = 0.17 - 0.10 = 0.08 \]
\[ p(48) = 0.10 - 0.06 = 0.04 \]
\[ p(54) = 0.06 \]
Exer. 2.18 (continued)

(6) Find the hazard rate function

\[ h(x) = \frac{p(x_j)}{S(x_{j-1})} \]

\[ h(6) = \frac{.45}{1} = .45 \]

\[ h(12) = \frac{.12}{.55} = \frac{218}{55} \]

\[ h(18) = \frac{.09}{.843} = .21 \]

\[ h(24) = \frac{.04}{.34} = .12 \]

\[ h(30) = \frac{.05}{.30} = .167 \]

\[ h(36) = \frac{.07}{.25} = .28 \]

\[ h(42) = \frac{.08}{.18} = .44 \]

\[ h(48) = \frac{.04}{.10} = .40 \]
Exer. 2.17

\( \frac{e(x) = x + 10}{\text{Ch. 2.7}} \)

(a) \( \delta = E(x) = \dot{e}(0) = 10 \)

(b) \( h(x) = \left[ \frac{d}{dx} \frac{e(x) + 1}{e(x)} \right] = \frac{1}{x + 10} \)

(c) \( S(x) = e^{-\int_0^x \frac{du}{e(u)}} \times \frac{e(0)}{e(x)} = \frac{10}{x + 10} - \int_0^x \frac{du}{u + 10} \)

\( = \frac{10}{x + 10} \times e^{-\int_0^x h(u + 10) du} = \frac{10}{x + 10} \cdot e^{-H(x + 10)} \)

\( \approx \frac{10}{x + 10} \times e^{-H(x + 10)} + \ln 10 \)

(d) \( f(x) = h(x) S(x) = \frac{2}{x + 10} \times \frac{10^2}{(x + 10)^2} = \frac{2 \times 10^2}{(x + 10)^3} \)

\( \int_0^\infty \frac{1}{(x + 10)^2} \, dx = 1 \)
The regression coefficients for these models are functions of time so that the effect of a given covariate on survival is allowed to vary over time. The $p$ regression functions may be positive or negative, but their values are constrained because (2.6.5) must be positive.

Estimation for additive models is typically made by nonparametric (weighted) least-squares methods. Additive models are used in section 6.3 to model excess mortality and, in Chapter 10, to model regression effects.

**Practical Notes**

1. From Theoretical Note 1 of section 2.4,

   \[ S(x \mid z) = \exp \left[ - \int_0^x b(t \mid z) dt \right] \]  \hspace{1cm} (2.6.6)

   and, in conjunction with (2.6.4),

   \[ S(x \mid z) = \exp \left[ - \int_0^x b_0(t) \exp(\beta'z) dt \right] \]

   \[ = \left\{ \exp \left[ - \int_0^x b_0(t) dt \right] \right\} \exp(\beta'z) \]

   \[ = \{S_0(x)\} \exp(\beta'z) \]

   which implies that

   \[ \ln[-\ln S(x \mid z)] = \beta'z + \ln[-\ln S_0(x)]. \]  \hspace{1cm} (2.6.7)

So the logarithms of the negative logarithm of the survival functions of $X$, given different regressor variables $z_t$, are parallel. This relationship will serve as a check on the proportional hazards assumption discussed further in Chapter 11.

### 2.7 Exercises

**2.1** The lifetime of light bulbs follows an exponential distribution with a hazard rate of 0.001 failures per hour of use.

(a) Find the mean lifetime of a randomly selected light bulb.
(b) Find the median lifetime of a randomly selected light bulb.
(c) What is the probability a light bulb will still function after 2,000 hours of use?

2.2 The time in days to development of a tumor for rats exposed to a carcinogen follows a Weibull distribution with $\alpha = 2$ and $\lambda = 0.001$.
(a) What is the probability a rat will be tumor free at 30 days? 45 days? 60 days?
(b) What is the mean time to tumor? (Hint $\Gamma(0.5) = \sqrt{\pi}$)
(c) Find the hazard rate of the time to tumor appearance at 30 days, 45 days, and 60 days.
(d) Find the median time to tumor.

2.3 The time to death (in days) following a kidney transplant follows a log logistic distribution with $\alpha = 1.5$ and $\lambda = 0.01$.
(a) Find the 50, 100, and 150 day survival probabilities for kidney transplantation in patients.
(b) Find the median time to death following a kidney transplant.
(c) Show that the hazard rate is initially increasing and, then, decreasing over time. Find the time at which the hazard rate changes from increasing to decreasing.
(d) Find the mean time to death.

2.4 The time to death (in days) after an autologous bone marrow transplant, follows a log normal distribution with $\mu = 3.177$ and $\sigma = 2.084$. Find
(a) the mean and median times to death,
(b) the probability an individual survives 100, 200, and 300 days following a transplant, and
(c) plot the hazard rate of the time to death and interpret the shape of this function.

2.5 A model for lifetimes, with a bathtub-shaped hazard rate, is the exponential power distribution with survival function $S(x) = \exp[1 - \exp((\lambda x)^\alpha)]$.
(a) If $\alpha = 0.5$ show that the hazard rate has a bathtub shape and find the time at which the hazard rate changes from decreasing to increasing.
(b) If $\alpha = 2$, show that the hazard rate of $x$ is monotone increasing.

2.6 The Gompertz distribution is commonly used by biologists who believe that an exponential hazard rate should occur in nature. Suppose that the time to death in months for a mouse exposed to a high dose of radiation follows a Gompertz distribution with $\theta = 0.01$ and $\alpha = 0.25$. Find
2.11 A model used in the construction of life tables is a piecewise, constant hazard rate model. Here the time axis is divided into $k$ intervals, $[\tau_{i-1}, \tau_i]$, $i = 1, \ldots, k$, with $\tau_0 = 0$ and $\tau_k = \infty$. The hazard rate on the $i$th interval is a constant value, $\theta_i$, that is

$$b(x) = \begin{cases} 
\theta_1 & 0 \leq x < \tau_1 \\
\theta_2 & \tau_1 \leq x < \tau_2 \\
& \vdots \\
\theta_{k-1} & \tau_{k-2} \leq x < \tau_{k-1} \\
\theta_k & x \geq \tau_{k-1}
\end{cases}$$

(a) Find the survival function for this model.
(b) Find the mean residual-life function.
(c) Find the median residual-life function.

2.12 Let $X$ have a uniform distribution on the interval $0$ to $\theta$ with density function

$$f(x) = \begin{cases} 
1/\theta, & \text{for } 0 \leq x \leq \theta \\
0, & \text{otherwise.}
\end{cases}$$

(a) Find the survival function of $X$.
(b) Find the hazard rate of $X$.
(c) Find the mean residual-life function.

2.13 Suppose that $X$ has a geometric distribution with probability mass function

$$p(x) = p(1-p)^{x-1}, \quad x = 1, 2, \ldots$$

(a) Find the survival function of $X$. (Hint: Recall that for $0 < \theta < 1$, $\sum_{j=0}^{\infty} \theta^j = \theta/(1 - \theta)$.
(b) Find the hazard rate of $X$. Compare this rate to the the hazard rate of an exponential distribution.

2.14 Suppose that a given individual in a population has a survival time which is exponential with a hazard rate $\theta$. Each individual's hazard rate $\theta$ is potentially different and is sampled from a gamma distribution with density function

$$f(\theta) = \frac{\lambda^\beta \theta^{\beta-1} e^{-\lambda \theta}}{\Gamma(\beta)}$$

Let $X$ be the life length of a randomly chosen member of this population.

(a) Find the survival function of $X$.
(b) Find the hazard rate of $X$. What is the shape of the hazard rate?
Based on data reported to the International Bone Marrow Transplant Registry, the survival function for a person given an HLA-identical sibling transplant for refractory multiple myeloma is given by

<table>
<thead>
<tr>
<th>Months Post Transplant</th>
<th>Survival Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>$0 \leq x &lt; 6$</td>
<td>1.00</td>
</tr>
<tr>
<td>$6 \leq x &lt; 12$</td>
<td>0.55</td>
</tr>
<tr>
<td>$12 \leq x &lt; 18$</td>
<td>0.43</td>
</tr>
<tr>
<td>$18 \leq x &lt; 24$</td>
<td>0.34</td>
</tr>
<tr>
<td>$24 \leq x &lt; 30$</td>
<td>0.30</td>
</tr>
<tr>
<td>$30 \leq x &lt; 36$</td>
<td>0.25</td>
</tr>
<tr>
<td>$36 \leq x &lt; 42$</td>
<td>0.18</td>
</tr>
<tr>
<td>$42 \leq x &lt; 48$</td>
<td>0.10</td>
</tr>
<tr>
<td>$48 \leq x &lt; 54$</td>
<td>0.06</td>
</tr>
<tr>
<td>$x \geq 54$</td>
<td>0</td>
</tr>
</tbody>
</table>

(a) Find the probability mass function for the time to death for a refractory multiple myeloma bone marrow transplant patient.
(b) Find the hazard rate of $X$.
(c) Find the mean residual life at 12, 24, and 36 months post transplant.
(d) Find the median residual life at 12, 24, and 36 months.

2.16 Suppose that the hazard rate of $X$ is a linear function $h(x) = \alpha + \beta x$, with $\alpha$ and $\beta > 0$. Find the survival function and density function of $x$.

2.17 Suppose that the mean residual life of a continuous survival time $X$ is given by $\text{MRL}(x) = x + 10$.
(a) Find the mean of $x$.
(b) Find $b(x)$.
(c) Find $S(x)$.

2.18 Given a covariate $Z$, suppose that the log survival time $Y$ follows a linear model with a logistic error distribution, that is,

$$Y = \ln(X) = \mu + \beta Z + \sigma W$$

where the pdf of $W$ is given by

$$f(w) = \frac{e^w}{(1 + e^w)^2}, -\infty < w < \infty.$$

(a) For an individual with covariate $Z$, find the conditional survival function of the survival time $X$, given $Z$, namely, $S(x | Z)$.
(b) The odds that an individual will die prior to time $x$ is expressed by $[1 - S(x | Z)]/S(x | Z)$. Compute the odds of death prior to time $x$ for this model.
(c) Consider two individuals with different covariate values. Show that, for any time $x$, the ratio of their odds of death is independent of $x$. The log logistic regression model is the only model with this property.