Lecture 2

Estimating the Survival Distribution

Statistics 295 - Survival Analysis

Presented January 7, 2016

Dan Gillen
Department of Statistics
University of California, Irvine
Survival Distributions

Functions of the survival distribution

- Probability distributions for a random variable $T$ (failure or survival time)
  - Continuous and Nonnegative

Four functions that characterize the survival distribution

1. **Probability density function** (pdf)
   
   $$f(t) = \lim_{\Delta t \to 0^+} \frac{1}{\Delta t} \Pr[t \leq T < t + \Delta t]$$

2. **Survival function** (1 minus the cumulative distribution function (cdf))
   
   $$S(t) = \Pr[T > t] = 1 - \Pr[T \leq t]$$
   
   $$= 1 - F(t) = 1 - \int_{0}^{t} f(s) \, ds$$
Survival Distributions

Functions of the survival distribution

3. **Hazard function** (or hazard rate)

\[
\lambda(t) = \lim_{\Delta t \to 0^+} \frac{1}{\Delta t} \Pr[t \leq T < t + \Delta t | T \geq t] = \frac{f(t)}{S(t)}
\]

instantaneous death rate at time \(t\) given alive up to time \(t\)

4. **Cumulative hazard function**

\[
\Lambda(t) = \int_0^t \lambda(s) ds = -\log S(t)
\]
Survival Distributions

Notes about the previous functions

1. \( f(t) = \frac{dF(t)}{dt} = -\frac{dS(t)}{dt} \)

2. \( S(t) = \exp\{-\Lambda(t)\} \)

3. When the hazard is high, the cumulative hazard *increases* faster and survival *decreases* faster with time

4. When the hazard is low, the cumulative hazard and survival change very little

5. The hazard function is not a probability (ie. not bounded above by 1)

6. Knowing any one of \( f, S, \lambda, \) or \( \Lambda \) is enough to specify the survival distribution and the remaining three functions

   - This does not mean that they all have the same clinical or scientific importance in a given setting!
Survival Distributions

Characteristics of $S(t)$:

1. $S(0)=1$
   All observations “at risk" at time 0

2. $S(\infty) = \lim_{t \to \infty} S(t) = 0$
   All observations will \textit{eventually} fail (by assumption)

3. $S(t)$ is non-increasing in $t$, ie.
   \[ S(t_1) \geq S(t_2) \text{ if } t_1 < t_2 \]
   Survival in the cohort does not increase with time
Survival Distributions

Hazard Function

- The Hazard function defined as

\[ \lambda(t) = \lim_{\Delta t \to 0^+} \frac{1}{\Delta t} \Pr[t \leq T < t + \Delta t | T \geq t] = \frac{f(t)}{S(t)} \]

- Instantaneous death rate at time \( t \) given alive up to time \( t \)

- Conditions on the **risk set** at time \( t \)
  - A subject is **at risk** at time \( t \) if s/he:
    1. has not yet experienced the failure event
    2. has not yet been censored
  - All subjects at risk at time \( t \) are call the **risk set** (at time \( t \))
    1. Often denoted as \( R_t \) or \( R_j \) if time is \( t_j \)
# Survival Distributions

## Parametric survival distributions

<table>
<thead>
<tr>
<th>Distribution</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Exponential</strong></td>
<td>Constant hazard: $E(T</td>
</tr>
<tr>
<td><strong>Weibull</strong></td>
<td>Hazard proportional to a power of $t$ (monotonic up or down)</td>
</tr>
<tr>
<td><strong>Gamma</strong></td>
<td>Monotonic hazard</td>
</tr>
<tr>
<td><strong>Log-normal</strong></td>
<td>Unimodal, right skewed hazard and density</td>
</tr>
<tr>
<td><strong>Log-logistic</strong></td>
<td>Unimodal, right skewed hazard and density (heavier tail than log-normal)</td>
</tr>
<tr>
<td><strong>Gompertz</strong></td>
<td>Hazard increases exponentially with $t$; used to model survival of human populations</td>
</tr>
<tr>
<td><strong>Generalized gamma</strong></td>
<td>Flexible (3 parameters); used for model checking</td>
</tr>
</tbody>
</table>
## Exponential Distribution

**hazard**: \( \lambda(t) = \lambda \)

**density**: \( f(t) = \lambda e^{-\lambda t} \)

**cumulative hazard**: \( \Lambda(t) = \lambda t \)

**survival function**: \( S(t) = e^{-\lambda t} \)

**mean**: \( \mu = 1/\lambda \)
Survival Distributions

Exponential Distribution

Hazard

Density

Cumulative Hazard

Survival

Survival distributions
Parametric survival distributions
Maximum likelihood estimation
Nonparametric Estimation of $S(t)$
Life Table Estimate of $S(t)$
The Kaplan-Meier estimator
CI for $S(t)$
Estimating percentiles of the survival distribution
### Survival Distributions

#### Weibull Distribution

<table>
<thead>
<tr>
<th>Hazard</th>
<th>Density</th>
<th>Cumulative Hazard</th>
<th>Survival Function</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \lambda(t) = \alpha \lambda t^{\alpha - 1} )</td>
<td>( f(t) = \alpha \lambda t^{\alpha - 1} e^{-\lambda t^{\alpha}} )</td>
<td>( \Lambda(t) = \lambda t^{\alpha} )</td>
<td>( S(t) = e^{-\lambda t^{\alpha}} )</td>
<td>( \mu = \Gamma[1 + 1/\alpha]/\lambda^{1/\alpha} )</td>
</tr>
</tbody>
</table>
Survival Distributions

Weibull Distribution

Hazard

Density

Cumulative Hazard

Survival

Survival distributions
Parametric survival distributions
Maximum likelihood estimation
Nonparametric Estimation of $S(t)$
Life Table Estimate of $S(t)$
The Kaplan-Meier estimator
CI for $S(t)$
Estimating percentiles of the survival distribution
Log-normal Distribution

\[ \log(T) \sim \mathcal{N}(\mu, \sigma^2) \]

**density**: 
\[ f(t) = \frac{1}{t\sqrt{2\pi\sigma}} \exp \left[ -\frac{1}{2} \left( \frac{\log t - \mu}{\sigma} \right)^2 \right] \]

**survival function**: 
\[ S(t) = 1 - \Phi \left( \frac{\log t - \mu}{\sigma} \right) \]

**mean**: 
\[ \mu = e^{\mu + 0.5\sigma^2} \]
Survival Distributions

Log-normal Distribution

- Hazard
- Density
- Cumulative Hazard
- Survival

Survival distributions
- Parametric survival distributions
  - Maximum likelihood estimation
- Nonparametric
  - Estimation of $S(t)$
    - Life Table Estimate of $S(t)$
    - The Kaplan-Meier estimator
    - CI for $S(t)$
    - Estimating percentiles of the survival distribution
Survival Distributions

Log-normal Distribution

\[ \log(T) \sim \mathcal{N}(\mu, \sigma^2) \]

**density:**
\[
f(t) = [t \sqrt{2\pi}\sigma]^{-1} \exp \left[ -\frac{1}{2} \left( \frac{\log t - \mu}{\sigma} \right)^2 \right]
\]

**survival function:**
\[
S(t) = 1 - \Phi \left( \frac{\log t - \mu}{\sigma} \right)
\]

**mean:**
\[
\mu = e^{\mu + 0.5\sigma^2}
\]
Survival Distributions

Parametric Estimation of the Survival Distribution

▶ Under the assumption of a parametric model, we can use maximum likelihood theory to estimate the survival distribution (and functionals of the distribution)

▶ The only difference is that we must account for censoring in the likelihood function

▶ Example: Suppose that $T_i \sim \text{Exp}(\lambda)$ and $C_i \sim \text{Unif}(a, b)$ where $a$ and $b$ are known constants with $\lambda$ an unknown parameter to estimate. Further assume that the $T_i$’s and $C_i$’s are totally independent.

▶ Goal: Find the MLE of $\lambda$ based on $Y_i = \min(T_i, C_i)$ and $\delta_i = 1[T_i = Y_i]$ and it’s asymptotic distribution. From this, find the MLE of $S_T(t; \lambda) = \Pr[T > t]$ and it’s asymptotic distribution.
Estimating $S(t)$

Nonparametric Estimation of $S(t)$

- A natural target of estimation is the survival function, $S(t)$

- Of course we could assume one of the previous parametric distributions as in our last example and
  - use maximum likelihood to estimate the model parameters
  - use the model parameters to estimate $S(t)$

- However, if our parametric assumption is incorrect this can lead to a biased and inconsistent estimate of $S(t)$

- Because of this, it is of interest to estimate $S(t)$ nonparametrically if possible...
Estimating $S(t)$

Nonparametric Estimation of $S(t)$: Toy Example

<table>
<thead>
<tr>
<th>Subject</th>
<th>Study Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A</td>
</tr>
<tr>
<td>2</td>
<td>A</td>
</tr>
<tr>
<td>3</td>
<td>D</td>
</tr>
<tr>
<td>4</td>
<td>A</td>
</tr>
<tr>
<td>5</td>
<td>D</td>
</tr>
<tr>
<td>6</td>
<td>D</td>
</tr>
<tr>
<td>7</td>
<td>A</td>
</tr>
<tr>
<td>8</td>
<td>A</td>
</tr>
<tr>
<td>9</td>
<td>D</td>
</tr>
<tr>
<td>10</td>
<td>A</td>
</tr>
</tbody>
</table>

Survival distributions:
- Parametric survival distributions
- Maximum likelihood estimation

Nonparametric Estimation of $S(t)$:
- Life Table Estimate of $S(t)$
- The Kaplan-Meier estimator
- CI for $S(t)$
- Estimating percentiles of the survival distribution
Estimating $S(t)$

<table>
<thead>
<tr>
<th></th>
<th>$\Pr[T &gt; 1]$</th>
<th>$\Pr[T &gt; 2]$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Full Sample</strong></td>
<td>8/10 = 0.80</td>
<td>6/10 = 0.60 (too high)</td>
</tr>
<tr>
<td><strong>Reduced Sample</strong></td>
<td>8/10 = 0.80</td>
<td>4/8 = 0.50 (too variable)</td>
</tr>
</tbody>
</table>

- Reduced sample estimate for $T > 2$:

\[
\Pr[T > 2] = \frac{\text{No. alive past 2 yrs}}{\text{No. w/ complete data at 2 yrs}} = \frac{4}{8} = 0.50
\]
Estimating $S(t)$

Problems with the reduced sample estimate of $S(t)$

- Only uses a portion of the sample to estimate $\Pr[T \geq 2]$
- Throws away information on 1 year survival

Solution: Use conditional probability

\[
\Pr[ \text{Survive 2 yrs } ] = \frac{\text{Pr}[ \text{Surv 2 yrs | Surv past 1 yr } ] \times \text{Pr}[ \text{Surv past 1 yr } ]}{\text{No. alive past 2 yrs}} \times \frac{\text{No. alive at 1 yr}}{\text{No. observed \geq 2 yrs, Alive at 1 yr}} \times \frac{\text{No. observed at 1 yr}}{8} = \frac{4}{6} \times \frac{8}{10} = .53
\]

Assumption: One year survival does not depend on year of entry, ie. future survival is not influenced by censorship
Estimating $S(t)$

**Life table estimate of $S(t)$**

**Example:** Complete data case (6MP data - placebo group)

- The essential data are the $n = 21$ ordered times (in months) to relapse:
  
  $1, 1, 2, 2, 3, 4, 4, 5, 5, 8, 8, 8, 8, 11, 11, 12, 12, 15, 17, 22, 23$

- Organize relapse times into intervals: (include upper boundary in interval)

<table>
<thead>
<tr>
<th>Interval</th>
<th>Beg. Total</th>
<th># Relap’d</th>
</tr>
</thead>
<tbody>
<tr>
<td>0+ - 5</td>
<td>21</td>
<td>9</td>
</tr>
<tr>
<td>5+ - 10</td>
<td>12</td>
<td>4</td>
</tr>
<tr>
<td>10+ - 15</td>
<td>8</td>
<td>5</td>
</tr>
<tr>
<td>15+ - 20</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>20+ - 25</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>

The Kaplan-Meier estimator

CI for $S(t)$

Estimating percentiles of the survival distribution
Estimating $S(t)$

**Life table estimate of $S(t)$**

**Example:** Complete data case (6MP data - placebo group)

- Compute surviving proportion at (after) end of each interval (out of the original 21):

  $0^+ - 5$ : \( \frac{21 - 9}{21} = \frac{12}{21} = 0.57 \)

  \( \Rightarrow \) 57% have not relapsed by end of first interval (i.e. by the 5th month).

  $5^+ - 10$ : \( \frac{12 - 4}{21} = \frac{8}{21} = 0.38 \)

  \( \Rightarrow \) 38% have not yet relapsed by end of second interval (i.e. by the 10th month).
### Life table estimate of $S(t)$

**Example: Complete data case (6MP data - placebo group)**

Therefore:

<table>
<thead>
<tr>
<th>Interval</th>
<th>Beg. Total</th>
<th># Relap’d</th>
<th>S-hat(t)</th>
<th>Survival</th>
<th>t</th>
</tr>
</thead>
<tbody>
<tr>
<td>0+ - 5</td>
<td>21</td>
<td>9</td>
<td>0.5714</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>5+ - 10</td>
<td>12</td>
<td>4</td>
<td>0.3810</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>10+ - 15</td>
<td>8</td>
<td>5</td>
<td>0.1429</td>
<td>15</td>
<td></td>
</tr>
<tr>
<td>15+ - 20</td>
<td>3</td>
<td>1</td>
<td>0.0952</td>
<td>20</td>
<td></td>
</tr>
<tr>
<td>20+ - 25</td>
<td>2</td>
<td>2</td>
<td>0.0000</td>
<td>25</td>
<td></td>
</tr>
</tbody>
</table>
**Estimating $S(t)$**

**Life table estimate of $S(t)$**

- We can also use the proportion surviving each interval:

- $0^+ - 5$: $\frac{21-9}{21} = \frac{12}{21} = 0.57$

- $5^+ - 10$: $\frac{21-9}{21} \times \frac{12-4}{21-9} = \frac{8}{21} = 0.38$

- $10^+ - 15$: $\frac{21-9}{21} \times \frac{12-4}{21-9} \times \frac{8-5}{12-4} = \frac{3}{21} = 0.14$

- This last expansion can be summarized as:

\[
\Pr\{\text{surv. int. 3}\} = \Pr\{\text{surv. int. 1}\} \times \Pr\{\text{surv. int. 2 } | \text{ surv. int. 1}\} \times \Pr\{\text{surv. int. 3 } | \text{ surv. int. 2}\}
\]

\[
= \Pr\{\text{surv. int. 1}\} \times \frac{\Pr\{\text{surv. int. 2}\}}{\Pr\{\text{surv. int. 1}\}} \times \frac{\Pr\{\text{surv. int. 3}\}}{\Pr\{\text{surv. int. 2}\}}
\]

- This form allows for extension to the incomplete data case where we have right censoring.
Estimating $S(t)$

**Life table estimate of $S(t)$**

**Example:** Censored data case (6MP data - 6MP group)

- The essential data are the $n = 21$ ordered times (in months) to relapse or censoring:
  
  $6, 6, 6, 7, 10, 13, 16, 22, 23, 6+, 9+, 10+, 11+, 17+, 19+, 20+, 25+, 32+, 32+, 34+, 35+$
  
  the $+$ indicating that the datum is censored

- Organize relapse times into intervals:

<table>
<thead>
<tr>
<th>Interval</th>
<th>Beg.</th>
<th>Total</th>
<th># Lost</th>
<th># Relap’d</th>
</tr>
</thead>
<tbody>
<tr>
<td>0+ - 5</td>
<td>21</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5+ - 10</td>
<td>21</td>
<td>3</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>10+ - 15</td>
<td>13</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>15+ - 20</td>
<td>11</td>
<td>3</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>20+ - 25</td>
<td>7</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>25+ - 30</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>30+ - 35</td>
<td>4</td>
<td>4</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

---

Survival distributions
Parametric survival distributions
Maximum likelihood estimation
Nonparametric Estimation of $S(t)$
Life Table Estimate of $S(t)$
The Kaplan-Meier estimator
CI for $S(t)$
Estimating percentiles of the survival distribution
Estimating $S(t)$

### Life table estimate of $S(t)$

Problem: Need to account for censoring

- Adjust beginning totals for censoring:
  - Assume that censoring occurs uniformly throughout interval
  - Censored subjects are therefore at risk (of failure) for half of the interval (on average)

- Thus, the *adjusted total at risk* for the interval is:

\[
\text{Adj. Total} = (\text{Uncens. Total}) + (\# \text{ Lost})/2 \\
= (\text{Beg. Total}) - (\# \text{ Lost}) + (\# \text{ Lost})/2 \\
= (\text{Beg. Total}) - (\# \text{ Lost})/2
\]
Estimating $S(t)$

Life table estimate of $S(t)$

- From this we have:

<table>
<thead>
<tr>
<th>Interval</th>
<th>Beg. Total</th>
<th>Lost</th>
<th>Adj. Total</th>
<th>Relap’d</th>
</tr>
</thead>
<tbody>
<tr>
<td>0+ - 5</td>
<td>21</td>
<td>0</td>
<td>21</td>
<td>0</td>
</tr>
<tr>
<td>5+ - 10</td>
<td>21</td>
<td>3</td>
<td>19.5</td>
<td>5</td>
</tr>
<tr>
<td>10+ - 15</td>
<td>13</td>
<td>1</td>
<td>12.5</td>
<td>1</td>
</tr>
<tr>
<td>15+ - 20</td>
<td>11</td>
<td>3</td>
<td>9.5</td>
<td>1</td>
</tr>
<tr>
<td>20+ - 25</td>
<td>7</td>
<td>1</td>
<td>6.5</td>
<td>2</td>
</tr>
<tr>
<td>25+ - 30</td>
<td>4</td>
<td>0</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>30+ - 35</td>
<td>4</td>
<td>4</td>
<td>2</td>
<td>0</td>
</tr>
</tbody>
</table>
Estimating $S(t)$

Life table estimate of $S(t)$

- Now compute surviving proportion for each interval:

  \[
  0^+ - 5: \quad \frac{21-0}{21} = 1.0
  \]
  \[
  5^+ - 10: \quad \frac{19.5-5}{19.5} = \frac{14.5}{19.5} = 0.74
  \]

  \Rightarrow 74\% \ of \ those \ not \ relapsed \ at \ the \ beginning \ of \ the \ second \ interval \ have \ not \ relapsed \ by \ end \ of \ the \ second \ interval \ (i.e. \ by \ the \ 10th \ month).

  \[
  10^+ - 15: \quad \frac{12.5-1}{12.5} = \frac{11.5}{12.5} = 0.92
  \]

  \Rightarrow 92\% \ of \ those \ not \ relapsed \ at \ the \ beginning \ of \ the \ third \ interval \ have \ not \ yet \ relapsed \ by \ the \ end \ of \ third \ interval.
### Life table estimate of $S(t)$

- Now compute surviving proportions using:

  $$\Pr\{\text{surv. int. 2}\} = \Pr\{\text{surv. int. 1}\} \times \frac{\Pr\{\text{surv. int. 2}\}}{\Pr\{\text{surv. int. 1}\}}$$

  i.e., for $5^+ - 10$:

  $$1.00 \times 0.74 = 0.74 \text{ and }$$

  $$\Pr\{\text{surv. int. 3}\}$$

  $$= \Pr\{\text{surv. int. 1}\} \times \frac{\Pr\{\text{surv. int. 2}\}}{\Pr\{\text{surv. int. 1}\}} \times \frac{\Pr\{\text{surv. int. 3}\}}{\Pr\{\text{surv. int. 2}\}}$$

  i.e., for $10^+ - 15$:

  $$1.00 \times 0.74 \times 0.92 = 0.68$$
Estimating $S(t)$

**Life table estimate of $S(t)$**

Therefore, our estimated survival at the end of each interval is:

<table>
<thead>
<tr>
<th>Interval</th>
<th>Adj. Total</th>
<th># Relap’d</th>
<th>(Conditional)</th>
<th>(Marginal)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0+ - 5</td>
<td>21</td>
<td>0</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
<tr>
<td>5+ - 10</td>
<td>19.5</td>
<td>5</td>
<td>0.7436</td>
<td>0.7436</td>
</tr>
<tr>
<td>10+ - 15</td>
<td>12.5</td>
<td>1</td>
<td>0.9200</td>
<td>0.6841</td>
</tr>
<tr>
<td>15+ - 20</td>
<td>9.5</td>
<td>1</td>
<td>0.8947</td>
<td>0.6121</td>
</tr>
<tr>
<td>20+ - 25</td>
<td>6.5</td>
<td>2</td>
<td>0.6923</td>
<td>0.4238</td>
</tr>
<tr>
<td>25+ - 30</td>
<td>4</td>
<td>0</td>
<td>1.0000</td>
<td>0.4238</td>
</tr>
<tr>
<td>30+ - 35</td>
<td>2</td>
<td>0</td>
<td>1.0000</td>
<td>0.4238</td>
</tr>
</tbody>
</table>

Therefore, our estimated survival at the end of each interval is:
Estimating $S(t)$

**Life table estimate of $S(t)$**

- **Notes:**
  - Because there are no subjects left at risk after 35 months, the survival function is not estimable for $t > 35$.
  - The “last” estimate available is
    \[
    \hat{\Pr}\{\text{surv. int. 6}\} = \hat{\Pr}\{T > 35 \text{ months}\} = \hat{S}_{LT}(35 \text{ months}) = 0.42
    \]
  - In particular, the *mean* survival time is *not* estimable, since to do so would require an estimate of $S(t)$ over all $t > 0$
  - These *life table estimates* (LT) of the survival function assume that future survival is not influenced by time of entry (censorship)
  - In R, lifetable estimates can be computed with the function `lifetab()` in the `KMsurv` package
**Estimating** $S(t)$

**Kaplan-Meier Estimator**

**Data**

- Possibly right-censored failure times:
  \[(x_1, \delta_1), (x_2, \delta_2), \ldots, (x_n, \delta_n)\]

  where
  \[x_i = \min(t_i, c_i) = \begin{cases} t_i & \text{uncensored case} \\ c_i & \text{censored case} \end{cases}\]
  \[\delta_i = I(x_i = t_i) = \begin{cases} 1 & \text{uncensored case} \\ 0 & \text{censored case} \end{cases}\]

  and $t_i$ is the failure time and $c_i$ is the censoring time for the $i$th subject.

- **Goal**: Estimate $S(t) = \Pr[T > t]$

- **Assume**: $T_i$ and $C_i$ are independent random variables — i.e. independent or uninformative censoring
Estimating $S(t)$

Kaplan-Meier Estimator

- The Kaplan-Meier (KM) estimator considers the probability of surviving a (very small) interval of time, given that a subject is at risk at the beginning of the interval.
- The K-M estimator is the limit of the life table estimator as the intervals shrink to zero.

\[ J \uparrow \infty, \quad l_j = \tau_j - \tau_{j-1} \downarrow 0 \]
## Kaplan-Meier Estimator

### Notation

- Let \( t_1 < t_2 < \ldots < t_D \) denote the **ordered event times** in the sample.
- For \( t_i \), let
  - \( d_i \) denote the total number of failures occurring at time \( t_i \)
  - \( s_i \) denote the total number that haven’t failed by time \( t_i \) (includes censored at \( t_i \))
  - \( n_i \) denote the total number at risk at time \( t_i \) (under observation, not failed)
  - \( d_i = n_i - s_i \)
Estimating $S(t)$

#### Kaplan-Meier Estimator

**Example Data**

<table>
<thead>
<tr>
<th>ID</th>
<th>Entry</th>
<th>Exit</th>
<th>Time</th>
<th>Ind</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.94</td>
<td>3.00</td>
<td>1.06</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0.91</td>
<td>3.00</td>
<td>2.09</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>1.50</td>
<td>2.96</td>
<td>1.46</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>0.66</td>
<td>1.00</td>
<td>0.34</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>0.44</td>
<td>2.23</td>
<td>1.79</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>0.08</td>
<td>2.87</td>
<td>2.78</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>0.90</td>
<td>3.00</td>
<td>2.10</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>0.43</td>
<td>3.00</td>
<td>2.57</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>0.64</td>
<td>1.14</td>
<td>0.50</td>
<td>1</td>
</tr>
<tr>
<td>10</td>
<td>1.71</td>
<td>3.00</td>
<td>1.29</td>
<td>0</td>
</tr>
</tbody>
</table>

- **Ordered event times**: 0.34 < 0.50 < 1.46 < 1.79 < 2.78
- **Tabulate counts**:

\[
\begin{array}{ccccc}
    t_i & n_i & d_i & s_i \\
    0.34 & 10 & 1 & 9 \\
    0.50 & 9 & 1 & 8 \\
    1.46 & 6 & 1 & 5 \\
    1.79 & 5 & 1 & 4 \\
    2.78 & 1 & 1 & 0 \\
\end{array}
\]
### Estimating $S(t)$

#### Kaplan-Meier Estimator

- Choose intervals so small that:
  - Endpoints at $t_i$ and $t_i + \Delta$
  - $\Delta$ so small that no withdrawals (censorings) in the interval
    
    $$(0, 0.34), [0.34, 0.50), [0.50, 1.06), [1.46, 1.79), [1.79, 2.09), [2.78, \infty)$$

- Estimate conditional survival probabilities:

  \[
  \hat{P}_r \{ T > 0 \} = 1 \\
  \hat{P}_r \{ T > 0.34 \mid T > 0 \} = 1 - \left( \frac{1}{10} \right) = \left( \frac{9}{10} \right) = 0.90 \\
  \hat{P}_r \{ T > 0.50 \mid T > 0.34 \} = \\
  \hat{P}_r \{ T > 1.46 \mid T > 0.50 \} = \\
  \hat{P}_r \{ T > 1.79 \mid T > 1.46 \} = \\
  \hat{P}_r \{ T > 2.78 \mid T > 1.79 \} =
  \]
Estimating $S(t)$

Kaplan-Meier Estimator

- Choose intervals so small that:
  - Endpoints at $t_i$ and $t_i + \Delta$
  - $\Delta$ so small that no withdrawals (censorings) in the interval
    
    $(0, 0.34), [0.34, 0.50), [0.50, 1.06), [1.46, 1.79), [1.79, 2.09), [2.78, \infty)$

- Estimate conditional survival probabilities:

\[
\begin{align*}
\hat{Pr}\{ T > 0 \} &= 1 \\
\hat{Pr}\{ T > 0.34 \mid T > 0 \} &= 1 - \left( \frac{1}{10} \right) = \left( \frac{9}{10} \right) = 0.90 \\
\hat{Pr}\{ T > 0.50 \mid T > 0.34 \} &= 1 - \left( \frac{1}{9} \right) = \left( \frac{8}{9} \right) \\
\hat{Pr}\{ T > 1.46 \mid T > 0.50 \} &= 1 - \left( \frac{1}{6} \right) = \left( \frac{5}{6} \right) \\
\hat{Pr}\{ T > 1.79 \mid T > 1.46 \} &= 1 - \left( \frac{1}{5} \right) = \left( \frac{4}{5} \right) \\
\hat{Pr}\{ T > 2.78 \mid T > 1.79 \} &= 1 - \left( \frac{1}{1} \right) = \left( \frac{0}{10} \right) = 0
\end{align*}
\]
Estimating $S(t)$

Kaplan-Meier Estimator

- In general:

$$\hat{\Pr}\{ T > t_j \mid T > t_{j-1} \} = 1 - \frac{d_j}{n_j} = \frac{s_j}{n_j}$$

- Recall that for $t \in (t_k, t_{k+1})$:

$$S(t) = \Pr[T > t]$$

$$= \Pr[T > t \mid T > t_k] \Pr[T > t_k \mid T > t_{k-1}] \times \ldots \times \Pr[T > t_1]$$
Estimating $S(t)$

**Kaplan-Meier Estimator**

**Definition**

The **Kaplan-Meier (Product Limit) estimate of $S(t)$** is given by:

$$
\hat{S}_{KM}(t) = \prod_{i: t_i \leq t} \left( 1 - \frac{d_i}{n_i} \right) = \prod_{i: t_i \leq t} \left( \frac{s_i}{n_i} \right)
$$
Estimating $S(t)$

**Kaplan-Meier Estimator**

**Question** Why are we able to only focus on intervals where an event occurs?

- Consider dividing time into any set of (very) small intervals: $[0, \tau_1), \ldots, [\tau_k - 1, \tau_k)$
- If the interval contains an observed event time: $\tau_{j-1} < t_i \leq \tau_j$

\[
\hat{\Pr}[T > \tau_j | T > \tau_{j-1}] \approx \frac{s_j}{n_j}
\]

- If the interval contains no event time:

\[
\hat{\Pr}[T > \tau_j | T > \tau_{j-1}] \approx 1
\]

- So that:

\[
\hat{S}(t) = 1 \cdot 1 \cdot 1 \cdot \frac{s_1}{n_1} \cdot \frac{s_2}{n_2} \cdots \frac{s_i}{n_i}
\]

where $t_i$ is the largest observed event time $\leq t$. 
Estimating $S(t)$

Standard error estimate of the Kaplan-Meier Estimator

**Definition** Greenwood’s formula:

$$\hat{SE}_G\{\hat{S}_{KM}(t)\} = \hat{S}_{KM}(t) \sqrt{\sum_{i:t_i \leq t} \frac{d_i}{n_is_i}}$$

<table>
<thead>
<tr>
<th>$t_i$</th>
<th>$n_i$</th>
<th>$d_i$</th>
<th>$s_i$</th>
<th>$\hat{S}_{KM}(t_i)$</th>
<th>$\sum_{t_j \leq t_i} \frac{d_j}{n_js_j}$</th>
<th>$\hat{SE}_G$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.34</td>
<td>10</td>
<td>1</td>
<td>9</td>
<td>0.90</td>
<td>0.011</td>
<td>0.095</td>
</tr>
<tr>
<td>0.50</td>
<td>9</td>
<td>1</td>
<td>8</td>
<td>0.80</td>
<td>0.025</td>
<td>0.127</td>
</tr>
<tr>
<td>1.46</td>
<td>6</td>
<td>1</td>
<td>5</td>
<td>0.67</td>
<td>0.058</td>
<td>0.161</td>
</tr>
<tr>
<td>1.79</td>
<td>5</td>
<td>1</td>
<td>4</td>
<td>0.54</td>
<td>0.108</td>
<td>0.176</td>
</tr>
<tr>
<td>2.78</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0.00</td>
<td>NA</td>
<td>NA</td>
</tr>
</tbody>
</table>

$(1 - \alpha)100\%$ confidence interval for $S(t)$:

$$\left(\hat{S}_{KM}(t) - Z_{1-\alpha/2} \times \hat{SE}_G, \hat{S}_{KM}(t) + Z_{1-\alpha/2} \times \hat{SE}_G\right)$$
Estimating $S(t)$

Confidence Interval for $S(t)$

- **A better CI:**
  - Note that the range of $\log \Lambda(t)$ is $(-\infty, +\infty)$ [range of $\Lambda(t)$ is $[0, -\infty)$]
  - Build a confidence interval for the log cumulative hazard
    $\log \Lambda(t) = \log[- \log S(t)]$
  - Back transform to obtain a confidence interval for $S(t)$.

- **Standard Error for $\log[- \log(\hat{S}_{KM}(t))]$:**

\[
\hat{SE}_2\{\log[- \log \hat{S}_{KM}(t)]\} = \sqrt{\sum_{i:t_i \leq t} \frac{d_i}{n_i s_i}} / \left( - \log \hat{S}_{KM}(t) \right)
\]
Confidence Interval for $S(t)$

**Confidence Interval:**

- For $S(t)$:

$$\exp \left[ -e^{\log \hat{S}(t) \pm Z_{1-\alpha/2} \times \hat{SE}_2} \right]$$

$$= \left( [\hat{S}_{KM}(t)]^{\exp \left\{ Z_{1-\alpha/2} \times \hat{SE}_2 \right\}}, [\hat{S}_{KM}(t)]^{\exp \left\{ -Z_{1-\alpha/2} \times \hat{SE}_2 \right\}} \right)$$

**Implementation in R:** See the `survfit` function in the R package `survival`
Estimating $S(t)$

Example: Kidney Transplant — K & M

- **Goal**: Estimate time to death from first kidney transplant

- **Covariates**:
  - ptid: Observation number
  - obstime: Time to death or on-study time
  - death: Death indicator (0=alive, 1=dead)
  - gender: Gender (1=male, 2=female)
  - race: Race (1=white, 2=black)
  - age: Age in years
Estimating \( S(t) \)

Example: Kidney Transplant — K & M

- First 5 rows of the data:

```r

> kidney[1:5,]
   ptid obstime death gender race age
 1   1       1     0     1   1  46
 2   2       5     0     1   1  51
 3   3       7     1     1   1  55
 4   4       9     0     1   1  57
 5   5      13     0     1   1  45
```

- Recode/rename gender and race to something meaningful:

```r
## Recode gender and race to something meaningful
kidney$female <- kidney$gender - 1
kidney$black <- kidney$race - 1
```
Estimating $S(t)$

Example: Kidney Transplant — K & M

- Compute the KM estimate of survival for the entire sample
  - Start by “specifying” the outcome variable using the `Surv()` function
  - All other survival functions will rely upon the `Surv` object
  - KM estimates are obtained with `survfit()`

```r
> ## Compute overall KM estimate
> kmEst.all <- survfit(Surv(obstime, death) ~ 1, data=kidney)
> kmEst.all
Call: survfit(formula = Surv(obstime, death) ~ 1, data = kidney)

  records  n.max  n.start  events median 0.95LCL 0.95UCL
          863     863     863   140   NA     NA     NA

> summary( kmEst.all )
Call: survfit(formula = Surv(obstime, death) ~ 1, data = kidney)

          time  n.risk n.event survival std.err lower 95% CI upper 95% CI
          2     861      1   0.999  0.00116     0.997     1.000
          3     860      1   0.998  0.00164     0.994     1.000
          7     857      2   0.995  0.00232     0.991     1.000
         10     853      2   0.993  0.00284     0.987     0.999
         17     848      1   0.992  0.00307     0.986     0.998
```
Estimating $S(t)$

Example: Kidney Transplant — K & M

- Compute and plot the KM estimates of survival for males and females
  - Just apply `plot` to the `survfit()` object

```r
kmEst.gender <- survfit( Surv(obstime, death) ~ female, data=kidney )
plot( kmEst.gender, xscale=365.25, mark.time=FALSE, lty=1:2,
     ylab="Survival", xlab="Years since transplant" )
legend( 1, .6, lty=1:2, legend=c("Male", "Female"), bty="n" )
```
Estimating $S(t)$

Example: Kidney Transplant — K & M

- Estimates of the survival probabilities can also be pulled from the `survfit()` object
  - Easiest to use the `pKM()` function on the course webpage...

- Let’s compute the estimated probability of survival for each gender at 6 months, 1 year, and 2 years post-transplant

```r
> pKM(kmEst.gender, q=c(.5,1,2), xscale=365.25 )
$female=0$
time   km.est  lower   upper
1   0.5 0.94674 0.92739 0.9665
2   1.0 0.91626 0.89199 0.9412
3   2.0 0.87998 0.85076 0.9102

$female=1$
time   km.est  lower   upper
1   0.5 0.94195 0.91694 0.96765
2   1.0 0.92508 0.89663 0.95443
3   2.0 0.91031 0.87900 0.94273
```
### Estimating $S(t)$

#### Additional notes on the Kaplan-Meier Estimator

- The `survfit()` plot could use some improvements in terms of the information it portrays...See HW 1!

- Other notes on the KM estimator:
  - If all remaining subjects fail at time $t_{(J)}$ (the largest failure time), then $\hat{S}_{KM}(t) = 0$ for $t > t_{(J)}$. If not, then $\hat{S}_{KM}(t)$ is generally considered inestimable for $t > t_{(J)}$.
  
  - In the absence of censoring (ie. completely observed survival times), the KM estimator reduces to the empirical survival function, $1 - F$, where
    
    $$F(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{(-\infty, x]}(X_i)$$
    
  - Kaplan and Meier (1958) showed that the KM estimator is the unique non-parametric mle over the class of probability distributions assigning probability to (and only to) the observed event times.
Estimating $S(t)$

### Estimating percentiles

- **The median** failure time $t_{.50}$ is that time such that half of all failure times are larger than $t_{.50}$ and half are smaller. i.e.

  $$F(t_{.50}) = S(t_{.50}) = 0.5$$

- **The $p$th percentile** of failure times $t_p$ is that time such that a fraction $p$ of failure times are less than $t_p$ and the other remaining fraction $1 - p$ of times are larger than $t_p$. i.e.

  $$F(t_p) = p \iff S(t_p) = 1 - p$$

- The estimated $p$th percentile of failure times is then the $p$th percentile obtained from the estimate $\hat{S}_{KM}(t)$

- To deal with the discreteness in $\hat{S}_{KM}(t)$, define:

  $$\hat{t}_p = \min \left\{ t : \hat{S}_{KM}(t) \leq 1 - p \right\}$$
Estimating $S(t)$

**Example: Kidney Transplant — K & M**

- Estimate the 25th percentile of failure times for each gender (ie. the 75th percentile of survival)

```r
# First separate the estimates by strata
> male.est <- as.data.frame( cbind( kmEst.gender$time, 
                                kmEst.gender$surv )[ 1:kmEst.gender$strata[1], ] )
> female.est <- as.data.frame( cbind( kmEst.gender$time, 
                                     kmEst.gender$surv )[ (kmEst.gender$strata[1]+1): 
                                     sum(kmEst.gender$strata), ] )
> names( male.est ) <- names( female.est ) <- c("time", "surv")

# Now pull out the appropriate for the 75th percentile of survival
> male.est[ min( which(male.est$surv <=.75 ) ), ]
  time  surv
347 2557 0.74798
> female.est[ min( which(female.est$surv <=.75 ) ), ]
  time  surv
256 2795 0.74027
```
Estimating $S(t)$

Confidence intervals for percentiles

- What about a confidence interval...consider the inversion of a hypothesis test

- A $100(1 - \alpha)\%$ confidence interval for an unknown parameter $\theta$ is the set of values $\theta^*$ that would not be rejected by a (two-sided) $\alpha$-level hypothesis test:

  \[
  H_0 : \quad \theta = \theta^* \quad \text{versus} \quad H_A : \quad \theta \neq \theta^*
  \]

- This definition is quite useful for constructing confidence intervals for the median or other percentiles...
### Confidence intervals for percentiles

- Recall: $p$th percentile of failure times is $t_p$ such that
  \[ S(t_p) = 1 - p \]

- Suppose we wished to test
  \[
  H_0 : \quad t_p = t^* \quad \text{versus} \\
  H_A : \quad t_p \neq t^*
  \]
  at the $(1 - \alpha)$ level

- This is equivalent to testing
  \[
  H_0 : \quad S(t^*) = 1 - p \quad \text{versus} \\
  H_A : \quad S(t^*) \neq 1 - p
  \]
**Estimating** $S(t)$

<table>
<thead>
<tr>
<th>Confidence intervals for percentiles</th>
</tr>
</thead>
<tbody>
<tr>
<td>► The above hypothesis test can be conducting by determining if the $100(1 - \alpha)%$ CI for $S(t^*)$ contained $(1 - p)$:</td>
</tr>
<tr>
<td>► If <em>so</em>, do not reject $H_0$</td>
</tr>
<tr>
<td>► If <em>not</em>, reject $H_0$ for $H_A$</td>
</tr>
<tr>
<td>► From this, we have that all the times $t^*$ for which we do not reject $H_0$ form a $100(1 - \alpha)%$ CI for $t_p$.</td>
</tr>
<tr>
<td>► Messy to do by hand, so see the function $qKM()$ on the course webpage...</td>
</tr>
</tbody>
</table>
Estimating $S(t)$

Example: Kidney Transplant — K & M

- Use `qKM()` to estimate the 15th and 25th percentile of failure times for each gender (ie. the 85th and 75th percentile of survival)

```r
##
##### qKM() for estimating quantiles of the survival distribution along with 95% CIs
##

> qKM(kmEst.gender, p=c(.85,.75), xscale=365.25 )

$'female=0'$

<table>
<thead>
<tr>
<th>perc.surv</th>
<th>km.est</th>
<th>lower</th>
<th>upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.85</td>
<td>3.0253</td>
<td>2.0726</td>
</tr>
<tr>
<td>2</td>
<td>0.75</td>
<td>7.0007</td>
<td>6.2724</td>
</tr>
</tbody>
</table>

$'female=1'$

<table>
<thead>
<tr>
<th>perc.surv</th>
<th>km.est</th>
<th>lower</th>
<th>upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.85</td>
<td>4.8652</td>
<td>2.5435</td>
</tr>
<tr>
<td>2</td>
<td>0.75</td>
<td>7.6523</td>
<td>5.8289</td>
</tr>
</tbody>
</table>
```