Definition of Survival and hazard functions:

\[ S(t) = \Pr\{T > t\} = 1 - F(t) \]

\[ \lambda(t) = \lim_{u \to 0} \frac{\Pr\{t < T \leq t + u \mid T > t\}}{u} = \frac{f(t)}{S(t)} \]

Relationship between Survival and hazard functions:

\[ \frac{\partial \log S(t)}{\partial t} = \frac{\partial S(t)}{\partial t} \frac{1}{S(t)} = -\frac{f(t)}{S(t)} \]

\[ \lambda(t) = -\frac{\partial \log S(t)}{\partial t} \]

The cumulative hazard function

\[ \Lambda(t) \equiv \int_0^t \lambda(v)dv = -\log S(t) \]

\[ S(t) = \exp[-\Lambda(t)] \]
**Exponential distribution**

\[ \Lambda(t) = \lambda t \]

\[ S(t) = \exp(-\Lambda(t)) = \exp(-\lambda t) \]

\[ \Pr\{T > t_0 + t \mid T > t_0\} = \Pr\{T > t\} \]

**Weibull distribution**

\[ \lambda(t) = \alpha \gamma t^{\gamma-1} \]

\[ \Lambda(t) = \alpha t^\gamma \]

\[ S(t) = \exp(-\Lambda(t)) = \exp(-\alpha t^\gamma) \]

With all deaths observed, can estimate nonparametrically: \( \hat{S}(t) = \text{prop}(T_i > t) \)

Or parametrically (method of moments)

Exponential: \( E(T) = \frac{1}{\lambda} \Rightarrow \hat{\lambda}_m = \frac{1}{\text{mean}(T)} \)

Weibull: \( E(T) = \frac{\gamma}{\lambda}, Var(T) = \frac{\gamma}{\lambda^2} \)

Plug in sample estimates and solve!
Types of Censorship:

Type 1: fixed censoring time (rare in medical applications, more common in engineering)
Type 2: censor after observe r failures (common in engineering)

RANDOM CENSORING: let C be a random censoring time, then for patient i

\[ Y_i = \min(T_i, C_i) \]

Observe

\[ \delta_i = I(T_i < C_i) \]

Convention: assume "death before censoring"!

ASSUME T and C are independent (nearly always false, but usefully so…weaker assumptions usually suffice)
Leads to non-parametric estimation such as KM, or the compactly named: *Altschuler-Nelson-Aalen-Fleming-Harrington* estimator:

\[
\hat{\Lambda}(t) = \sum_{i:t_i<t} \frac{d_i}{n_i}
\]

\(t_1, t_2, t_3, \ldots\) are the ordered unique event times \(d_1, d_2, d_3, \ldots\) corresponding numbers of deaths \(n_1, n_2, n_3, \ldots\) numbers at risk

\[
\hat{S}_\Lambda(t) = \exp[-\Lambda(t)]
\]

Compare to *Kaplan-Meier PL* estimator

\[
\hat{S}_{KM}(t) = \prod_{i:t_i<t} \left(1 - \frac{d_i}{n_i}\right)
\]

\[
\hat{\Lambda}_{KM}(t) = -\log \hat{S}_{KM}(t) = -\sum_{i:t_i<t} \log \left(1 - \frac{d_i}{n_i}\right)
\]
parametric estimation, for random censoring:

Exponential: \( \hat{\lambda}_{ML} = \frac{\sum \delta_i}{\sum Y_i} \)

(note, numerator is count of uncensored observations)

Weibull: not closed form…must iterate!

For now we will concentrate on non-(and semi-) parametric estimation and testing – leave parametrics (especially useful Weibull) to Prof. Olshen.
Mantel-Haenszel log-rank test

At each unique death, 2X2 table of vital status by group

<table>
<thead>
<tr>
<th>group</th>
<th>dead</th>
<th>alive</th>
<th>$n$</th>
</tr>
</thead>
<tbody>
<tr>
<td>group 1</td>
<td>a</td>
<td>b</td>
<td>$n_1$</td>
</tr>
<tr>
<td>group 2</td>
<td>c</td>
<td>d</td>
<td>$n_2$</td>
</tr>
</tbody>
</table>

$m_1$ $m_2$ $n$

For example
Leukemia data, first death time=5

<table>
<thead>
<tr>
<th>Survival Status</th>
<th>dead</th>
<th>alive</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maintained</td>
<td>0</td>
<td>11</td>
</tr>
<tr>
<td>Unmaintained</td>
<td>2</td>
<td>10</td>
</tr>
</tbody>
</table>

2 21 23
leukemia data in full

<table>
<thead>
<tr>
<th>time</th>
<th>status</th>
<th>group</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>9</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>13</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>13</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>18</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>23</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>28</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>31</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>34</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>45</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>48</td>
<td>1</td>
</tr>
<tr>
<td>11</td>
<td>161</td>
<td>0</td>
</tr>
<tr>
<td>12</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>13</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>14</td>
<td>8</td>
<td>1</td>
</tr>
<tr>
<td>15</td>
<td>8</td>
<td>1</td>
</tr>
<tr>
<td>16</td>
<td>12</td>
<td>1</td>
</tr>
<tr>
<td>17</td>
<td>16</td>
<td>0</td>
</tr>
<tr>
<td>18</td>
<td>23</td>
<td>1</td>
</tr>
<tr>
<td>19</td>
<td>27</td>
<td>1</td>
</tr>
<tr>
<td>20</td>
<td>30</td>
<td>1</td>
</tr>
<tr>
<td>21</td>
<td>33</td>
<td>1</td>
</tr>
<tr>
<td>22</td>
<td>43</td>
<td>1</td>
</tr>
<tr>
<td>23</td>
<td>45</td>
<td>1</td>
</tr>
</tbody>
</table>
Given fixed margins, null hypothesis (equal death rates)
A=upper left hand entry has hypergeometric distribution:

\[
\Pr\{A = a\} = \frac{\binom{n_1}{a} \binom{n_2}{m_1 - a}}{\binom{n}{m_1}}
\]

\[
E_0(A) = \frac{n_1 m_1}{n}
\]

\[
Var_0(A) = \frac{n_1 n_2 m_1 m_2}{n^2 (n - 1)}
\]

Accumulate the differences between observed and expected over the set of 2X2 tables at each death time, and divide by the std error of the sum:

\[
MH = \frac{\sum [a_i - E_0(A_i)]}{\sqrt{\sum Var_0(A_i)}} \quad \text{…….refer to N}(0,1)
\]
Gehan modification of Wilcoxon
Let $T$ and $R$ be the true (possibly unobserved) times in Maintained and Unmaintained subjects, $X$ and $Y$ the corresponding censored values

$$U(X_i, Y_j) = U_{ij} = \begin{cases} 
+1 & \text{if we know } T_i > R_j \\
0 & \text{otherwise} \\
-1 & \text{if we know } T_i < R_j 
\end{cases}$$

$$U = \sum_{i,j} U_{ij}$$

Reject null if $U$ is large.
If no censorship:

$$Var_{0,P}(U) = \frac{mn(m+n+1)}{3}$$

but with censorship more complex (and larger)

……for leukemia data
MH = 1.8429, p= 0.0653
GW = 1.6671, p= 0.0955
Tarone-Ware class of tests:

\[ MH = \frac{\sum w_i [a_i - E_0(A_i)]}{\sqrt{\sum w_i^2 Var_0(A_i)}} \]

- \( w_i = 1 \) gives MH
- \( w_i = n_i \) gives Gehan
- \( w_i = \sqrt{n_i} \) is TW suggestion
COX PH Model:

\[ \lambda(t; X) = \lambda_0(t) \exp(X\beta) \]

single binary covariate \( X \):

\[ \frac{\lambda(t; X = 1)}{\lambda(t; X = 0)} = \frac{\lambda_0(t) \exp(1\beta)}{\lambda_0(t) \exp(0\beta)} = \exp(\beta) \]

Lehman Alternatives:

\[ S(t; X) = \exp \left[ -\int_0^t \lambda_0(s) \exp(X\beta) ds \right] = S_0(t)^\gamma \]

\[ S_0(t) = \exp \left[ -\int_0^t \lambda_0(s) ds \right] \]

\[ \gamma = \gamma(X) = \exp(X\beta) \]
suppressing times, and taking logs...

\[
\log(S) = \gamma \log(S_0) = -\gamma \Lambda_0
\]
\[
\log(-\log(S)) = \log(\gamma) + \log(\Lambda_0) = X\beta + \log(\Lambda_0)
\]

So estimates of survival for various subgroups should look parallel on the "log-minus-log" scale.

And – if the hazard is constant:

\[
\log(\Lambda_0(t)) = \log(\lambda_0 t) = \log(\lambda_0) + \log(t)
\]

so the survival estimates are all straight lines on the log-minus-log (survival) against log (time) plot.
How many subjects to enroll?

To detect a true log hazard ratio of \( \theta = \log\left(\frac{\lambda_1}{\lambda_2}\right) \)
(power \(1 - \beta\) using a 1-sided test at level \(\alpha\))

require \(D\) observed **death**, where:

\[
D = \frac{4(z_{1-\alpha} + z_{1-\beta})^2}{\theta^2}
\]

(for equal group sizes- if unequal replace 4 with \(\frac{1}{P(1-P)}\) where \(P\) is proportion assigned to group 1)

*The censored observations contribute nothing to the power of the test!*
Sample size required for non-binary covariate X:

Deaths:

\[ D = \frac{\left(z_{1-\alpha} + z_{1-\beta}\right)^2}{\sigma_X^2 \theta^2} \]

where \( \sigma_X^2 \) is the variance of X and \( \theta \) is the log hazard ratio for a unit change in X.

Note that "wider" X gives more power, as it should!

Epidemiology: non-binary exposure X (say, amount of smoking)

Adjust for confounders Z (age, sex, etc.), in the Cox model.

Adjust D above by "Variance Inflation Factor"

\[ VIF = \frac{1}{1 - R^2} \]

where \( R^2 \) = variance of X explained by Z.