Statistics 255
Derivation of the distribution of the logrank statistic

Conditional distribution of \( d_{1k} \):

At the \( k \)-th failure time we observe the following \( 2 \times 2 \) table:

<table>
<thead>
<tr>
<th>Failure</th>
<th>Yes</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group 0</td>
<td>( d_{0k} )</td>
<td>( y_{0k} - d_{0k} )</td>
</tr>
<tr>
<td>Group 1</td>
<td>( d_{1k} )</td>
<td>( y_{1k} - d_{1k} )</td>
</tr>
<tr>
<td>Total</td>
<td>( d_k )</td>
<td>( y_k - d_k )</td>
</tr>
</tbody>
</table>

- Conditional on \( y_{1k} \) and \( y_{0k} \) we have that \( D_{1k} \sim Bin(y_{1k}, p_{1k}) \) and \( D_{0k} \sim Bin(y_{0k}, p_{0k}) \).
- Consider the conditional distribution of \( D_{1k} \) given the margin total \( D_k = D_{1k} + D_{0k} \).
- Thus,

\[
\Pr[D_{1k} = d_{1k}|D_k = d_k] = \frac{\Pr[D_{1k} = d_{1k}, D_{0k} = d_k - d_{1k}]}{\Pr[D_k = d_k]}
\]

\[
\Pr[D_{1k} = d_{1k}, D_{0k} = d_k - d_{1k}] = \binom{y_{1k}}{d_{1k}} p_{1k}^{d_{1k}} (1 - p_{1k})^{y_{1k} - d_{1k}} \binom{y_{0k}}{d_{0k} - d_{1k}} p_{0k}^{d_{0k} - d_{1k}} (1 - p_{0k})^{y_{0k} - (d_k - d_{1k})}
\]

\[
= \binom{y_{1k}}{d_{1k}} \binom{y_{0k}}{d_{0k} - d_{1k}} (1 - p_{1k})^{y_{1k} - d_{1k}} (1 - p_{0k})^{y_{0k} - (d_k - d_{1k})} \exp \left\{ d_{1k} \log \frac{p_{1k}}{1 - p_{1k}} + d_{0k} \log \frac{p_{0k}}{1 - p_{0k}} - d_{1k} \log \frac{p_{0k}}{1 - p_{0k}} \right\}
\]

\[
= \binom{y_{1k}}{d_{1k}} \binom{y_{0k}}{d_{0k} - d_{1k}} (1 - p_{1k})^{y_{1k}} (1 - p_{0k})^{y_{0k}} \exp \left\{ d_{1k} \log \frac{p_{1k}/(1 - p_{1k})}{p_{0k}/(1 - p_{0k})} + d_{0k} \log \frac{p_{0k}}{1 - p_{0k}} \equiv \psi_k \right\}
\]

\[
= \binom{y_{1k}}{d_{1k}} \binom{y_{0k}}{d_{0k} - d_{1k}} (1 - p_{1k})^{y_{1k}} (1 - p_{0k})^{y_{0k}} \exp \{ d_{1k} \log \psi_k + d_{0k} \log \frac{p_{0k}}{1 - p_{0k}} \}
\]
\[
\Pr[D_{1k}|D_k = D_{1k} + D_{0k}] = \frac{\Pr[D_{1k} = d_{1k}, D_{0k} = d_k - d_{1k}]}{\Pr[D_k = d_k]}
\]

\[
\propto \left(\frac{y_{1k}}{d_{1k}}\right)\left(\frac{y_{0k}}{d_k - d_{1k}}\right)\psi_k^{d_{1k}}
\]

\[
\propto \left(\frac{y_{1k}}{d_{1k}}\right)\left(\frac{y_{0k}}{d_k - d_{1k}}\right)\psi_k^{d_{1k}}
\]

Now, consider testing \(H_0: p_{1k} = p_{0k}\) vs. \(H_0: p_{1k} \neq p_{0k}\). Then under \(H_0\),

\[
\psi_k = \frac{p_{1k}/(1-p_{1k})}{p_{0k}/(1-p_{0k})} = 1
\]

So that

\[
\Pr[D_{1k}|D_k = D_{1k} + D_{0k}] = c \left(\frac{y_{1k}}{d_{1k}}\right)\left(\frac{y_{0k}}{d_k - d_{1k}}\right)
\]

Need to find \(c\) such that

\[
\sum_{d_{1k}=1}^{d_k} c \left(\frac{y_{1k}}{d_{1k}}\right)\left(\frac{y_{0k}}{d_k - d_{1k}}\right) = 1
\]

Notice that \(\left(\frac{y_{1k}}{d_{1k}}\right)\left(\frac{y_{0k}}{d_k - d_{1k}}\right)\) is the kernel of a hypergeometric distribution with parameters \(y_{1k}, y_{0k}, d_k\). Therefore we have that

\[
c = \frac{1}{\left(\frac{y_{1k} + y_{0k}}{d_k}\right)}
\]

and

\[
\Pr[D_{1k}|D_k = D_{1k} + D_{0k}] = \frac{\left(\frac{y_{1k}}{d_{1k}}\right)\left(\frac{y_{0k}}{d_k - d_{1k}}\right)}{\left(\frac{y_{1k} + y_{0k}}{d_k}\right)}
\]

\[
= \frac{\left(\frac{y_{1k}}{d_{1k}}\right)\left(\frac{y_{0k}}{d_k}\right)}{\left(\frac{y_{1k}}{d_k}\right)}
\]

Therefore \(D_{1k} \sim \text{Hypergeometric}(y_k, y_{1k}, d_k)\).

Mean and variance of the LR statistic:
Claim 1: \( E[U_k] = E[O_k - E_k] = E \left[ D_{1k} - \frac{y_{1k}d_k}{y_k} \right] = 0 \)

Proof: Let \( \mathcal{F}(x) \) denote the filtration at time \( x \) so that
\[
\mathcal{F}(x) = \{ D_{0k}, D_{1k}, Y_{0k}, Y_{1k}, w_{0k}, w_{1k}, D_k, \text{for all } k < x \}.
\]

That is, knowing the filtration \( \mathcal{F}(x) \), we know all the failure and censoring that has occurred prior to time \( x \), and the number of individuals at risk at time \( x \) and the total number of deaths. What we don’t know is what group the deaths at time \( x \) occurred in. Then
\[
E[U_k] = E \left\{ E \left[ D_{1k} - \frac{y_{1k}d_k}{y_k} \left| \mathcal{F}(k) \right. \right] \right\} = E \left\{ E[D_{1k}\mid \mathcal{F}(k)] - \frac{y_{1k}d_k}{y_k} \right\} = E \left\{ \frac{y_{1k}d_k}{y_k} - \frac{y_{1k}d_k}{y_k} \right\} = E \{ 0 \}.
\]

Thus, by linearity of expectation we have that \( E[T_{L,R}] = 0 \).

Claim 2: Define \( U = \sum_k U_k \), where \( U_k = O_k - E_k = D_{1k} - \frac{y_{1k}d_k}{y_k} \). Then an unbiased estimate of \( \text{Var}[U] \) is given by
\[
\sum_k V_k = \sum_k \frac{y_{1k}y_k(y_k - d_k)d_k}{y_k^2(y_k - 1)}.
\]

Proof: First, the variance of \( U \) is
\[
\text{Var}[U] = \text{Var} \left[ \sum_k U_k \right] = \sum_k \text{Var}[U_k] + \sum_{j \neq k} \text{Cov}[U_k, U_j].
\]

Consider an arbitrary \( k \) and \( j \), where WLOG \( j < k \). Then from Claim 1, we know that \( E[U_k] = E[U_j] = 0 \). Therefore,
\[
\text{Cov}[U_k, U_j] = E[U_kU_j] = E \{ E[U_kU_j\mid \mathcal{F}(k)] \} = E \{ U_jE[U_k\mid \mathcal{F}(k)] \} = E \{ 0 \} = 0.
\]
Thus,

\[
\text{Var}[U] = \sum_k \text{Var}[U_k] = \sum_k \text{E}[U_k^2]
\]

\[
= \sum_k \text{E} \left\{ \text{E}[U_k^2 | F(k)] \right\}
\]

\[
= \sum_k \text{E} \left\{ \text{E} \left[ \left( D_{1k} - \frac{y_{1k}d_k}{y_k} \right)^2 \bigg| F(k) \right] \right\}
\]

\[
= \sum_k \text{E} \left\{ \text{E} \left[ (D_{1k} - \text{E}[D_{1k}|F(k)])^2 \bigg| F(k) \right] \right\}
\]

\[
= \sum_k \text{E} \left\{ \text{Var}[D_{1k}] \right\}
\]

\[
= \sum_k \text{E} \left\{ \frac{y_{1k}y_{0k}(y_k - d_k)d_k}{y_k^2(y_k - 1)} \right\},
\]

so that an unbiased estimate of \text{Var}[U] is given by

\[
\sum_k V_k = \sum_k \frac{y_{1k}y_{0k}(y_k - d_k)d_k}{y_k^2(y_k - 1)}.
\]