Hypothesis Test

Goal of testing: Determine if there is a difference (overall or at a given time point) between two or more groups.

- one-sample tests
- two-sample tests
- K-sample tests
- trend tests
- stratified tests

Some of the "traditional methods" are appropriate for complete survival times but not applicable to censored data.

Complete Data & Fixed Time

Suppose there is no censoring and the data include $t_1, t_2, \ldots, t_n$. For a fixed time point t, we are interested in comparing the survival rates at t between two groups.

		D	$\bar{D}$	
Treatment	A	d_A	$n_A - d_A$	n_A
	B	d_B	$n_B - d_B$	n_B
		m_D	$m_{\bar{D}}$	n

- d_A and d_B are the numbers of subjects who fail before t
- n_A and n_B are the total numbers of subjects in Groups A and B
- $m_{\bar{D}}$ is the total number of subjects who survive beyond t
- $n = m_D + m_{\bar{D}} = n_A + n_B$

- $1 - d_A/n_A$ is the survival rate in Group A at time t
- A typical 2×2 table
- χ^2 test or Fisher exact test

Incomplete Data & Fixed Time

Suppose t-year survival rate is of interest

$$H_0 : S_A(t) = S_B(t) \qquad H_1 : S_A(t) \neq S_B(t).$$

Data could be censored before t. We use the K-M estimate to estimate $S_A(t)$ and $S_B(t)$, and construct a test statistic

$$T = \frac{\widehat{S}_A(t) - \widehat{S}_B(t)}{\widehat{\text{SD}}\left\{\widehat{S}_A(t) - \widehat{S}_B(t)\right\}} \sim N(0,1).$$

Here $\text{SD}\left\{\widehat{S}_A(t) - \widehat{S}_B(t)\right\}$ can be estimated by Greenwood's formula,

$$\begin{aligned}
\text{var}\left\{\widehat{S}_A(t) - \widehat{S}_B(t)\right\} &= \text{var}\left\{\widehat{S}_A(t)\right\} + \text{var}\left\{\widehat{S}_B(t)\right\} \\
\widehat{\text{SD}}\left\{\widehat{S}_A(t) - \widehat{S}_B(t)\right\} &= \sqrt{\widehat{\text{var}}\left\{\widehat{S}_A(t)\right\} + \widehat{\text{var}}\left\{\widehat{S}_B(t)\right\}}
\end{aligned}$$

Incomplete Data & All Time

Consider a two-sample test of the overall difference between two survival functions

$$H_0 : S_A(t) = S_B(t), \quad \text{for all } t$$

(Equivalently, $h_A(t) = h_B(t)$ under H_0)

Log-rank test is used:

1. Create a 2×2 table at each *uncensored* survival time (on the basis of the corresponding risk set).
2. Construct a test statistic based on each 2×2 table.
3. Combine all the test statistics from tables to construct a final test statistic (log-rank test statistic)

Log-Rank Test

Step 1: construct a 2×2 table at each uncensored survival time t_i (from pooled data).

		Event	No Event	
Treatment	A	$d_{A,i}$	$n_{A,i} - d_{A,i}$	$n_{A,i}$
	B	$d_{B,i}$	$n_{B,i} - d_{B,i}$	$n_{B,i}$
		$m_{D,i}$	$m_{\bar{D},i}$	n_i

- $d_{A,i}$: number of failures at t_i from Group A
- $n_{A,i}$: number of individuals at risk at t_i from Group A
- $m_{D,i}$: number of failures at t_i from pooled data
- n_i: number of individuals at risk at time t_i from pooled data

Log-Rank Test

Step 2: construct a test statistic for each 2×2 table.

- Given $n_{A,i}, n_{B,i}, m_{D,i}, m_{\bar{D},i}$, the number of failures $d_{A,i}$ follows a hypergeometric distribution (under H_0)
- Test statistic: $d_{A,i} - m_{D,i} n_{A,i} / n_i$ (i.e., observed $-$ expected)
- $\text{var}(d_{A,i}) = \frac{n_{A,i} n_{B,i} m_{D,i} m_{\bar{D},i}}{n_i^2 (n_i - 1)}$ under H_0

Log-Rank Test

Step 3: combine information from all tables into the final log-rank test statistic.

$$Z = \frac{\sum_{i=1}^{K}\{d_{A,i} - \mathbb{E}(d_{A,i})\}}{\sqrt{\sum_{i=1}^{K} \mathrm{var}(d_{A,i})}} \sim N(0,1)$$

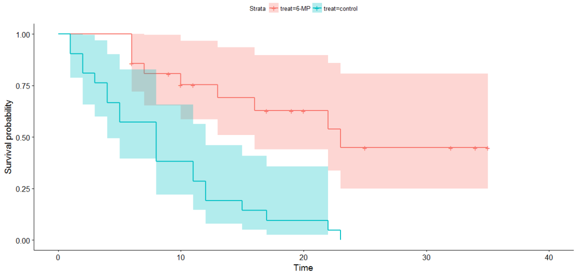
- Replace $\mathbb{E}(d_{A,i})$ and $\mathrm{var}(d_{A,i})$ with the empirical estimates under H_0.
- Normal approximation requires large sample size.
- When Z is positive, treatment B is better.
- When Z is negative, treatment A is better.
- For two-sided test, use $Z^2 \sim \chi^2(1)$.

Example

Acute Leukemia

- 6-MP group: $n_1 = 21$
  $6, 6, 6, 7, 10, 13, 16, 22, 23, 6^+, 9^+, 10^+, 11^+, 17^+,$
  $19^+, 20^+, 25^+, 32^+, 32^+, 34^+, 35^+$ (months)

- Placebo group, $n_2 = 21$
  $1, 1, 2, 2, 3, 4, 4, 5, 5, 8, 8, 8, 8, 11, 11, 12, 12, 15, 17, 22, 23$ (months)



Weighted Log-Rank Test

After constructing a sequence of 2×2 tables at uncensored times, we consider the statistic $T = \sum_{i=1}^{K} w_i \{d_{A,i} - \mathbb{E}(d_{A,i})\}$ where w_i is the "weight" on the table at t_i. The variance of T is $\sum_{i=1}^{K} w_i^2 \text{var}(d_{A,i})$, and the weighted log-rank test statistic is

$$Z = \frac{\sum_{i=1}^{K} w_i \{d_{A,i} - \mathbb{E}(d_{A,i})\}}{\sqrt{\sum_{i=1}^{K} w_i^2 \text{var}(d_{A,i})}}$$

- $w_i = 1$: log-rank test
- $w_i = n_i$: Gehan's test
- $w_i = \sqrt{n_i}$: Tarone and Ware test
- $w_i = [\hat{S}(t_i)]^{\rho}[1 - \hat{S}(t_i)]^{\gamma}$: Fleming and Harrington test

- Gehan's test and Tarone and Ware's test put more weight on earlier times. They are more powerful if the relative hazard is large in the beginning.
- Log-rank test is powerful to detect the survival differences most evident later in time.
- All tests may not be sensitive to the crossing survival curves.
- Log-rank test can be generalized to more than two groups.