4.2.4 Variance and confidence interval for the CIF estimator

As is often the case in estimation, a researcher might wish to obtain an interval around the CIF calculated at a specific point in time and have some level of confidence that the interval contains the real, but unknown, value. In order to construct a confidence interval for $F_i(t)$ the variance of the CIF estimator, $\hat{F}_i(t)$, is required. To obtain the exact variance of the CIF estimator is a difficult undertaking, since one needs to evaluate

$$V\left(\hat{F}_i(t)\right) = \sum_{t_j \le t} \text{Var}\left(\{d_{ij}/n_j\}\,\hat{S}(t_j)\right) + 2\sum_{t_j<t}\sum_{\substack{t_\nu>t_j\\ t_\nu \le t}} \text{Cov}\left(\{d_{ij}/n_j\}\,\hat{S}(t_j),\{d_{i\nu}/n_\nu\}\,\hat{S}(t_\nu)\right).$$

Using the delta method, $V\left(\hat{F}_i(t)\right)$ can be estimated as:

$$\hat{V}_{\text{DM}}\left(\hat{F}_i(t)\right) = \sum_{t_j \le t}\left\{\left[\hat{F}_i(t)-\hat{F}_i(t_j)\right]^2 \frac{d_j}{n_j\left(n_j-d_j\right)}\right\} + \sum_{t_j\le t}\hat{S}(t_{j-1})^2\frac{d_{ij}\left(n_j-d_{ij}\right)}{n_j^3} - 2\sum_{t_j\le t}\left[\hat{F}_i(t)-\hat{F}_i(t_j)\right]\hat{S}(t_{j-1})\frac{d_{ij}}{n_j^2}. \tag{4.4}$$

The details of how this variance estimator is obtained are not given here, but readers who are interested in its mathematical derivation can find it in Marubini and Valsecchi (1995). Since $\hat{F}_i$ is a step function, terms that include $\hat{F}_i(t)-\hat{F}_i(t_j)$ will be 0 unless an event of type i has occurred between times t_j and t. Likewise, terms involving d_{ij} will be 0 except at those time points where an event of type i has been recorded. Therefore the last term of each sum is non-zero only if at time t there is an event of type i. When there is only one type of event and no competing risks, formula (4.4) for the estimated variance is identical to Greenwood's formula for the usual survival analysis setting (formula (2.7)).

The variance estimator may also be derived using the work of Aalen (1978a) as follows:

$$\hat{V}_{\mathrm{A}}\left(\hat{F}_i(t)\right) = \sum_{t_j \le t} \left\{ \left[\hat{F}_i(t) - \hat{F}_i(t_j)\right]^2 \frac{d_j}{\left(n_j - 1\right)\left(n_j - d_j\right)} \right\}$$
$$+ \sum_{t_j \le t} \hat{S}(t_{j-1})^2 \frac{d_{ij}\left(n_j - d_{ij}\right)}{n_j^2\left(n_j - 1\right)} \tag{4.5}$$
$$- 2\sum_{t_j \le t} \left[\hat{F}_i(t) - \hat{F}_i(t_j)\right] \hat{S}(t_{j-1}) \frac{d_{ij}\left(n_j - d_{ij}\right)}{n_j\left(n_j - d_j\right)\left(n_j - 1\right)}.$$

An approximate $100(1-\alpha)\%$ confidence interval for the CIF for the event of interest, $F(t)$, based on its asymptotic normality, is

$$\hat{F}(t) \pm z_{1-\alpha/2}\sqrt{\hat{V}\left(\hat{F}(t)\right)}, \tag{4.6}$$

where z_α is the α quantile of the standard normal distribution and $\hat{V}$ is the variance estimator of the CIF. While this confidence interval has the advantage of being familiar and simple to construct, it has the unfortunate property that it may result in the bounds being negative or above 1. As shown in Section 2.3, this side effect can be avoided by finding a confidence interval for $\log\left(-\log\left(\hat{F}(t)\right)\right)$ first and transforming the bounds back to their original scale. The resulting confidence interval is of the form

$$\hat{F}(t)^{\exp[\pm A]}, \tag{4.7}$$

where

$$A = \frac{z_{1-\alpha/2}\sqrt{\hat{V}(\hat{F}(t))}}{\hat{F}(t)\log\left(\hat{F}(t)\right)}.$$