Confidence intervals

Louis Dijkstra

August 6, 2014

Suppose we observed k internal segment lengths, $x = (x_1, x_2, \ldots, x_k)$, and l overlapping alignments, $y = (y_1, y_2, \ldots, y_l)$, where y_j is 1 when the alignment does support the presence of the indel of interest and 0 otherwise. Consider the following likelihood function for the parameter $\theta \in [0, 1]$ (representing the variant allele frequency):

$$L(\theta; x, y) = \prod_{i=1}^{k} g_i(x_i; \theta) \times \prod_{j=1}^{l} h_j(y_i; \theta), \tag{1}$$

where

$$g_i(x_i; \theta) = \pi_i^{(X)} \left[\theta f_{\mu+\delta}(x_i) + (1-\theta) f_\mu(x_i)\right] + \left(1 - \pi_i^{(X)}\right) u(x_i) \tag{2}$$

and

$$h_j(y_j; \theta) = \pi_j^{(Y)} \theta^{y_j} (1-\theta)^{1-y_j} + \left(1 - \pi_j^{(Y)}\right) v(y_j). \tag{3}$$

The probability mass functions $f_{\mu+\delta}(\cdot)$ and $f_\mu(\cdot)$ describe the internal segment length distribution when the alignment stems from a chromosome with and without the indel of variant length δ present. It is defined as

$$f_\lambda(x) := \frac{\Phi\left(\frac{x+1-\lambda}{\sigma}\right) - \Phi\left(\frac{x-\lambda}{\sigma}\right)}{1 - \Phi\left(\frac{-\lambda}{\sigma}\right)} \tag{4}$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution function. The functions $u(\cdot)$ and $v(\cdot)$ denote the distributions of the observations when the alignments do not stem from the locus of interest. These distributions are unknown, but need to be chosen such that they do not influence the posterior distribution of θ. A reasonable choice is, therefore:

$$u(z) \propto 1 \qquad \text{and} \qquad v(z) \propto 1.$$

(which we can prove. Do so!).

The loglikelihood function is easily found to be

$$\ell(\theta; x, y) = \sum_{i=1}^{k} \log g_i(x_i; \theta) + \sum_{j=1}^{l} \log h_j(y_j; \theta). \tag{5}$$

(Show that a global maximum exists!). Let $\widehat{\theta}$ be the maximum likelihood estimate of θ, i.e.,

$$L(\widehat{\theta}; x, y) \geq L(\theta; x, y) \qquad \text{for all } \theta \in [0, 1].$$

Likelihood ratio method for constructing CIs

We can construct an approximate $100(1-\alpha)\%$ confidence interval for the variant allele frequency θ by employing the likelihood ratio method, i.e., the CI is defined as

$$\mathrm{CI}_{\widehat{\theta}}^{1-\alpha} := \left\{\theta \in [0,1] : 2\left[\ell(\widehat{\theta}; x, y) - \ell(\theta; x, y)\right] \leq F^{-1}(1-\alpha)\right\}, \tag{6}$$

where $\widehat{\theta}$ is the MLE of the VAF and F^{-1} is the inverse cumulative distribution function of the χ^2 distribution with 1 degree of freedom.

The $100(1-\alpha)\%$ confidence interval contains all θ's in $[0,1]$ for which

$$2\left[\ell(\widehat{\theta}; x, y) - \ell(\theta; x, y)\right] \leq F^{-1}(1-\alpha) \tag{7}$$

where F^{-1} is the inverse cumulative distribution function of the χ^2 distribution with 1 degree of freedom. There are two values that satisfy the equality[1]:

$$2\left[\ell(\widehat{\theta}; x, y) - \ell(\theta; x, y)\right] = F^{-1}(1-\alpha).$$

1. one in the interval $[0, \widehat{\theta}) \to \widehat{\theta}_{low}$,
2. one in the interval $(\widehat{\theta}, 1] \to \widehat{\theta}_{high}$.

We can determine these numerically. The CI is then

$$[\widehat{\theta}_{low}, \widehat{\theta}_{high}].$$

In order to numerically determine the CI (analytically is impossible since it would envolve finding the root of the a higher order polynomial), we first note that the inequality is equal to

$$q(\theta) = \ell(\theta; x, y) - \ell(\widehat{\theta}; x, y) + F^{-1}(1-\alpha)/2 \geq 0.$$

We, thus, need to find the zero points of the function $q(\theta)$ in the intervals $[0, \widehat{\theta})$ and $(\widehat{\theta}, 1]$.

If $\ell(0; x, y) - \ell(\widehat{\theta}; x, y) + F^{-1}(1-\alpha)/2 \geq 0$, then $\widehat{\theta}_{low} = 0$. Otherwise, apply Newton-Rapshon bounded to the interval $[0, \widehat{\theta})$ in order to estimate the CI's lower bound. For determining the upper bound, we do the same.

[1] In case of $\alpha = 5\%$, $F^{-1}(.95) \approx 3.84$.