Actuarial Estimates and MLEs in Survival Analysis

Stanley Sawyer — Washington University — October 2, 2005

Assume that we have demographic data for a population over a series of times

$$0 \;=\; t_0 \;<\; t_1 \;<\; t_2 \;<\; \dots \;<\; t_r \tag{1}$$

We assume that the data is *longitudinal*, which means that we follow that same $N = n_1$ individuals over r time intervals, as opposed to observing different individuals in different time intervals. In more detail, let

n_i be the number being followed (or "at risk") just before time t_{i-1} ,
d_i be the number of observed deaths in $[t_{i-1}, t_i)$,
c_i be the number of censored individuals in $[t_{i-1}, t_i)$, and
$\Delta_i = t_i - t_{i-1}$ be the length of the i^{th} time interval

where $[t_{i-1}, t_i)$ means times $t_{i-1} \le t < t_i$. Thus $n_{i+1} = n_i - d_i - c_i$ and c_i is the number of individuals who were last seen at time t_{i-1} but are not observed at times $t \ge t_i$.

The underlying probability model is that individuals die at rate α_i in the interval (t_{i-1}, t_i) (whether they are observed or not) and that they are censored (that is, drop out alive) at rate β_i. The probability that an individual survives until time $t = t_j$ is then

$$S(t_j) \;=\; \prod_{t_i \le t_j} e^{-\alpha_i \Delta_i} \;=\; \prod_{t_i \le t_j} e^{-\mu_i} \qquad \text{for } \mu_i = \alpha_i \Delta_i \tag{2}$$

Similarly, set $\nu_i = \beta_i \Delta_i$ where β_i is the censoring rate. Note that the censoring parameters ν_i and β_i do not enter (2) directly. However, they enter implicitly, since we do not know how many of the initial n_i intervals in any time interval were censored before they had time to die.

By definition, the maximum likelihood estimator (MLE) of $S(t)$ is that function $S(t)$ in the class (2) that maximizes the likelihood or probability of observing all of the data (n_i, d_i, c_i). From (2), the MLE of $S(t)$ depends only on the MLEs $\widehat{\mu}_i$ of the μ_i, where $\widehat{\mu}_i$ depends on the counts (n_i, d_i, c_i).

Theorem 1. *The maximum likelihood estimator (MLE) for $S(t)$ in (2) for data (n_i, d_i, c_i) is*

$$\widehat{S}(t) \;=\; \prod_{t_i \le t} \left(1 \;-\; \frac{(d_i + c_i)}{n_i}\right)^{d_i/(d_i+c_i)} \tag{3}$$

Theorem 2. *Within errors of the form $O(1/n_i^3)$ for large n_i, the estimator $\widehat{S}(t)$ in (3) is the same as*

$$\widehat{S}(t) \;=\; \prod_{t_i \le t} \left(1 \;-\; \frac{d_i}{n_i - (1/2)c_i}\right) \tag{4}$$

Remarks. Equation (4) is usually called the *Actuarial Estimator* of $S(t)$. The notation $O(1/n^3)$ (due to Landau) stands for any expression that is bounded by C/n^3 for $n \ge 1$ for some fixed (but unknown) constant $C < \infty$.

Proof of Theorem 1. Suppressing subcripts for the i^{th} interval, the probability that any one individual out of $n = n_i$ individuals neither died nor was censored in the time interval is $\exp(-\mu-\nu) = \exp\big(-(\alpha+\beta)\Delta\big)$, since $\alpha = \alpha_i$ and $\beta = \beta_i$ are rates and Δ is the length of the time interval. Thus the probability that $m = d + c$ individuals out of the initial $n = n_i$ individuals either died or were censored is

$$\frac{n!}{(n-m)!\,m!}\;\big(e^{-\mu-\nu}\big)^{n-m}\,\big(1 - e^{-\mu-\nu}\big)^m \tag{5a}$$

In general, the probability that a given individual eventually dies before he or she is censored is $\kappa = \mu/(\mu+\nu) = \alpha/(\alpha+\beta)$. One way to see this is to use the fact that if X and Y are independent exponentially-distributed random variables with rates α and β respectively, then $P(X < Y) = \kappa = \alpha/(\alpha+\beta)$.

Thus, conditional on $m = d+c$ individuals having died or been censored in the time interval (t_{i-1}, t_i), the probability that we observed d died and c censored is

$$\frac{m!}{d!\,c!}\left(\frac{\mu}{\mu+\nu}\right)^d \left(\frac{\nu}{\mu+\nu}\right)^c \tag{5b}$$

Since (5b) is the probability of observing (d, c) conditional on $m = d + c$, and (5a) is the probability of observing $m = d+c$ out of n, the probability of observing $(d, c, n-d-c)$ for observed deaths, censoring events, and neither is the product of (5a) and (5b), which is the trinomial probability

$$\frac{n!}{(n-d-c)!\,d!\,c!}\;\big(e^{-\mu-\nu}\big)^{n-d-c}\left[\big(1 - e^{-\mu-\nu}\big)\,\frac{\mu}{\mu+\nu}\right]^d \left[\big(1 - e^{-\mu-\nu}\big)\,\frac{\nu}{\mu+\nu}\right]^c$$

In terms of $\lambda = e^{-\mu-\nu}$ and $\kappa = \mu/(\mu+\nu)$, this is

$$\frac{n!}{(n-d-c)!\,d!\,c!}\;\lambda^{n-d-c}\,(1-\lambda)^{d+c}\,\kappa^d(1-\kappa)^c \tag{6}$$

For given values (d, c, n), the probability in (6) is maximized when $\lambda = \widehat{\lambda} = (n - d - c)/n$ and $\kappa = \widehat{\kappa} = d/(d + c)$. (*Exercise*: Prove this.)

In particular, the trinomial likelihood after (5b) is maximized for those values $\widehat{\mu} = \mu$ and $\widehat{\nu} = \nu$ that are equivalent to $\lambda = \widehat{\lambda} = (n - d - c)/n$ and $\kappa = \widehat{\kappa} = d/(d + c)$. Thus

$$\widehat{\mu} = (-\log \widehat{\lambda})\,\widehat{\kappa} = -\log\left(1 - \frac{d+c}{n}\right)\frac{d}{d+c} \tag{7}$$
$$e^{-\widehat{\mu}} = \widehat{\lambda}^{\widehat{\kappa}} = \left(1 - \frac{d+c}{n}\right)^{d/d+c}$$

This completes the proof of Theorem 1, or equivalently of equation (3).

Proof of Theorem 2. Expanding the logarithm in (7) in a power series,

$$\begin{aligned}
\widehat{\mu} &= \left(\left(\frac{d+c}{n}\right) + \frac{1}{2}\left(\frac{d+c}{n}\right)^2 + O\left(\frac{1}{n^3}\right)\right)\frac{d}{d+c} \\
&= \frac{d}{n}\left(1 + \frac{1}{2}\frac{d+c}{n}\right) + O\left(\frac{1}{n^3}\right) \\
&= \frac{d}{n}\left(\frac{1}{1 - (1/2)(d+c)/n}\right) + O\left(\frac{1}{n^3}\right) \\
&= \frac{d}{n - (d+c)/2} + O\left(\frac{1}{n^3}\right)
\end{aligned} \tag{8}$$

The first term on the right in (8) is the intuitive estimate for the *hazard rate* $\mu = \mu_i$, but not for the *survival probability* $e^{-\mu_i}$ in (2). For the latter, we need $e^{-\widehat{\mu_i}}$ from $\widehat{\mu}_i$ in (8). Thus

$$\begin{aligned}
1 - e^{-\widehat{\mu}} &= \widehat{\mu} - \frac{1}{2}\widehat{\mu}^2 + O(\widehat{\mu}^3) \\
&= \frac{d}{n}\left(1 + \frac{1}{2}\frac{(d+c)}{n} - \frac{1}{2}\frac{d}{n} + O\left(\frac{1}{n^2}\right)\right) \\
&= \frac{d}{n}\left(1 + \frac{1}{2}\frac{c}{n}\right) + O\left(\frac{1}{n^3}\right) \\
&= \frac{d/n}{1 - (1/2)c/n} + O\left(\frac{1}{n^3}\right) \\
&= \frac{d}{n - c/2} + O\left(\frac{1}{n^3}\right)
\end{aligned}$$

This implies (4), which completes the proof of Theorem 2.