The Greenwood and Exponential Greenwood Confidence Intervals in Survival Analysis

S. Sawyer — September 4, 2003

1. Introduction. Let $T_1, T_2, \ldots, T_n$ be the times of either (i) an observed death or failure or (ii) the last time that a living individual was seen. Set $\delta_i = 0$ if $T_i$ is an observed death and $\delta_i = 1$ if the $i^{th}$ individual was last seen alive at time $T_i$, but has not been seen since. If $\delta_i = 1$, the true death time for the $i^{th}$ individual is $X_i > T_i$ but the individual dropped out of the study at that time. In that case, we say that $X_i$ was censored at time $T_i$. If $\delta_i = 0$, $T_i$ is the time of an observed death or failure.

Let $0 < t_1 < t_2 < \cdots < t_r$ be the distinct observed death times in the sample, arranged in increasing order. That is, $t_i$ are the distinct times $t = t_i$ for which $T_j = t$ and $\delta_j = 0$ for some $j$. Let $n_i$ be the size of the risk set at time $t_i$. That is, $n_i$ is the number of individuals in the sample that were alive (or “at risk”) just before time $t_i$. Equivalently, $n_i$ is the number of individuals who are either alive and observed at time $t_i$ or else who died at time $t_i$. For $i < r$, $n_{i+1} = n_i - d_i - c_i$ where $d_i$ is the number who died at time $t = t_i$ and $c_i$ is the number who are censored at times $t$ with $t_i \leq t < t_{i+1}$.

The Kaplan-Meier or product-limit estimator $\hat{S}(t)$ of the survival function $S(t) = \Pr(T > t)$ is

$$\hat{S}(t) = \prod_{t_i \leq t} \left(1 - \frac{d_i}{n_i}\right)$$ (1.1)

The purpose here is to derive two approximate 95% confidence intervals for $S(t)$ for a fixed $t$, or, in general, $(1 - \alpha) \times 100\%$ confidence intervals for $S(t)$. The first is Greenwood’s (1926) confidence interval

$$\hat{S}(t) \pm z_{\alpha/2} \sqrt{\text{Var} [\hat{S}(t)]}$$ where

$$\text{Var} [\hat{S}(t)] = \hat{S}(t)^2 \sum_{t_i \leq t} \frac{d_i}{n_i(n_i - d_i)}$$ (1.2b)

In (1.2b), $z_\alpha$ is the $\alpha$-th quantile of the normal distribution, so that (for example) $z_{0.025} = -1.96$. In particular, $z_{\alpha/2} = -1.96$ for a 95% confidence interval. If $n_i = d_i$, which can only happen if $i = r$ and $t \geq t_r$, the last term in the sum in (1.2b) is omitted.

The second confidence interval is called the “exponential” Greenwood formula. This is attributed by Hosmer and Lemeshow (1999) to the earlier
textbook Kalbfleisch and Prentice (1980). This gives an asymmetric confidence interval
\[
\exp(-\exp(c_+(t))) < S(t) < \exp(-\exp(c_-(t))) \tag{1.3a}
\]
where
\[
c_\pm(t) = \log(-\log \hat{S}(t)) \pm z_{\alpha/2} \sqrt{\hat{V}} \quad \text{and} \quad \hat{V} = \frac{1}{(\log \hat{S}(t))^2} \sum_{t_i \leq t} \frac{d_i}{n_i(n_i - d_i)} 
\tag{1.3b}
\]

Note that \(\exp(-\exp(c_2)) < \exp(-\exp(c_1))\) if \(c_1 < c_2\).

The advantage of (1.3) over the more traditional Greenwood confidence interval (1.2) is that the endpoints of (1.3) are guaranteed to lie in \((0, 1)\), whereas the endpoints of (1.2) can be negative or greater than one. Hosmer and Lemeshow (1999) quote a paper to the effect that the confidence interval (1.3) behaves well for sample sizes as small as 25 with up to 50% of observations being censored.

2. Proofs. The main idea is the “delta-method” approximation, which assumes
\[
f(X) \approx f(c) + f'(c)(X - c) \tag{2.1}
\]
for a function \(f(X)\) of a random variable \(X\) with \(c\) close to \(E(X)\). This implies
\[
E[f(X)] \approx f(c) + f'(c)(E(X) - c) \tag{2.2}
\]
\[
\text{Var}[f(X)] \approx f'(c)^2 \text{Var}(X)
\]
The traditional Greenwood formula applies (2.1)–(2.2) with \(f(t) = \log t\). The exponential Greenwood formula has essentially \(f(t) = \log(-\log t)\).

The Kaplan-Meier formula (1.1) implies
\[
\log \hat{S}(t) = \sum_{t_i \leq t} \log \left(1 - \frac{d_i}{n_i}\right) \tag{2.3}
\]
We assume that \(d_i\) is binomially distributed with parameters \(p_i\) and \(n_i\) given the size \(n_i\) of the risk set. Thus \(E(d_i) = n_i p_i\) and \(\text{Var}(d_i) = n_i p_i(1 - p_i)\).

Using (2.1) with \(f(t) = \log t\) and \(c = p_i\),
\[
\log \hat{S}(t) \approx \sum_{t_i \leq t} \left(\log(1 - p_i) - \frac{1}{1 - p_i} \left(\frac{d_i}{n_i} - p_i\right)\right)
\]
\[
= C(p) - \sum_{t_i \leq t} \frac{1}{1 - p_i} \left(\frac{d_i}{n_i} - p_i\right) \tag{2.4}
\]
The terms in the sum in (2.4) are not independent, since the $d_i$ in one term affects the risk set counts $n_k$ for $k > i$. However, the $i$th term in the sum in (2.4) has mean zero given $d_1, d_2, \ldots, d_{i-1}$. (The technical term for this is that the partial sums of the sum in (2.4) form a martingale.) The fact each term in the sum has mean zero conditional on the earlier terms can be used to show that the variance of the sum in (2.4) is the sum of the variances, even though the terms are not independent. (Exercise: Prove this.) Since $\text{Var}(d_i/n_i | n_i) = p_i(1-p_i)/n_i$, this implies

$$\text{Var}(\log \hat{S}(t)) \approx \sum_{t_i \leq t} \frac{d_i}{n_i(n_i - d_i)} \approx \sum_{t_i \leq t} \frac{d_i}{n_i}$$

by setting $p_i = \hat{p}_i = d_i/n_i$.

Note $\hat{S}(t) = \exp(Y(t))$ for $Y(t) = \log \hat{S}(t)$. A second application of (2.1)–(2.2) with $f(y) = e^y$ yields $\overline{\text{Var}}(\hat{S}(t)) = \hat{S}^2(t) \overline{\text{Var}}(Y(t))$. This implies Greenwood’s formula (1.2b) for the variance. All that remains is that the distribution of the sum (2.3) is approximately normal, so that we can use normal confidence limits. This can be shown using properties of martingales. This completes the proof of Greenwood’s formula (1.2).

The “exponential” Greenwood formula is based on the random variables $Z(t) = \log(Y(t)) = \log\left(-\log \hat{S}(t)\right)$ for $Y(t) = -\log \hat{S}(t)$

Applying (2.2) with $f(y) = \log(y)$ implies $\text{Var}(Z(t)) \approx \text{Var}(Y(t))/Y(t)^2$. Using (2.5), this implies the confidence interval

$$Z(t) \pm z_{\alpha/2} \sqrt{\overline{\text{Var}}(Z(t))}$$

for

$$\overline{\text{Var}}(Z(t)) = \frac{1}{\left(\log \hat{S}(t)\right)^2} \sum_{t_i \leq t} \frac{d_i}{n_i(n_i - d_i)}$$

Undoing the transformation $Z(t) = \log\left(-\log \hat{S}(t)\right)$ leads to the confidence interval (1.3)

References.
