Inference for Kaplan-Meier Estimator

Outline

- Greenwood's formula
- Confidence interval for survival function

Introduction

- We have ued the Kaplan-Meier method to estimate the survival function non-parametrically.
- How reliable is the estimate?
- Introduce a Greenwood's formula to estimate the variance of the Kaplan-Meier estimator.
- Use i to construct confidence intervals.

Variance of Kaplan-Meier estimator

- Recall that K-M estimator, for $t \geq t_1$,

  $\hat{S}(t) = \prod_{t_i \leq t} \left[1 - \frac{d_i}{Y_i}\right].$
- Greenwood's formula provides an estimator of variance of K-M estimator:

  $\widehat{Var}\left(\hat{S}(t)\right) = \hat{S}^2(t) \sum_{t_i \leq t} \frac{d_i}{Y_i(Y_i - d_i)}$

  $= \hat{S}^2(t) \hat{\sigma}_S^2(t),$

  where $\hat{\sigma}_S^2(t) = \sum_{t_i \leq t} \frac{d_i}{Y_i(Y_i - d_i)}.$

An example

A total of 18 quails were radio tagged with survival times in weeks,

3,3,6,8,8+,9,9+,9+,10,10+,12+,13+,13+,13+,13+,13+,13+,13+

t	d_i	Y_i	$\hat{S}$(t)= $\prod_{t_i \leq t}\left[1 - \frac{d_i}{Y_i}\right]$	$\hat{\sigma}_S^2(t)$ $\sum_{t_i \leq t} \frac{d_i}{Y_i(Y_i - d_i)}$	V($\hat{S}(t)$)=$\hat{S}^2(t)\hat{\sigma}_S^2(t)$
3	2	18	0.89	$\frac{2}{18\times16}$ =**0.0069**	**0.0055**
6	1	16	0.83	**0.0069+**$\frac{1}{16\times15}$ =**0.0111**	**0.0076**
8	1	15	0.78	**0.0111+**$\frac{1}{15\times14}$=**0.0159**	**0.0096**
9	1	13	0.72	**0.0159+**$\frac{1}{12\times11}$=**0.0234**	**0.0121**
10	1	10	0.65	**0.0234+**$\frac{1}{11\times10}$=**0.0325**	**0.0137**

Confidence intervals

- Greenwood's formula for the estimation of variance of K-M estimator:

  $\hat{\sigma}_{KM}^2(t) = \hat{S}^2(t) \sum_{t_i \leq t} \frac{d_i}{Y_i(Y_i - d_i)}.$
- Then a naïve $100(1-\alpha)\%$ confidence interval (CI) for $S(t)$ is

  $\hat{S}(t) \pm Z_{\alpha/2} \hat{\sigma}_{KM}(t).$

Data example

t	$\hat{S}$(t)	V($\hat{S}(t)$)	95% naïve CI $\hat{S}(t) \pm 1.96\sqrt{V(\hat{S}(t))}$
3	0.89	0.0055	(0.74,1.03)=0.89$\pm$1.96$\sqrt{0.0055}$
6	0.83	0.0076	(0.66,1.01)
8	0.78	0.0096	(0.59,0.97)
9	0.72	0.0121	(0.51,0.93)
10	0.65	0.0137	(0.41,0.88)

An important criticism of the naïve CI is that the bounds of the interval can lie outside [0, 1] and can contain insensible values.

Confidence intervals

- The naïve CIs may include impossible values outside of [0, 1].
- Better CIs can be obtained by applying transformations on S(t).
- Kalbeisch & Prentice (2002) suggested

$$T(t) = \log[-\log S(t)]$$

  - $S(t) \in (0,1)$
  - $\log(S(t)) \in (-\infty, 0)$
  - $\log(-\log(S(t))) \in (-\infty, +\infty)$

Confidence intervals

- Build CI for $T(t) = \log[-\log S(t)]$.
- Back transform to obtain a CI for $S(t)$.
- Specifically, $100(1-\alpha)\%$ CI for $T(t)$ is

$$c_1 = \log[-\log \hat{S}(t)] + Z_{\frac{\alpha}{2}}\sqrt{V\left(\log[-\log \hat{S}(t)]\right)}$$

$$c_2 = \log[-\log \hat{S}(t)] - Z_{\frac{\alpha}{2}}\sqrt{V\left(\log[-\log \hat{S}(t)]\right)}$$

Confidence intervals

- Then the asy. $100(1-\alpha)\%$ CI for $S(t)$ is

  $(\exp(-e^{c_2}), \exp(-e^{c_1})) = \left[\hat{S}(t)^{\theta}, \hat{S}(t)^{\frac{1}{\theta}}\right]$,

  where $\theta = \exp\left\{\frac{Z_{\alpha/2}\hat{\sigma}_S(t)}{\log[\hat{S}(t)]}\right\}$.
- Always yields proper bounds.
- It is not defined for $\hat{S}(t) = 0$ or 1. In these cases, use $(0,0)$, or $(1,1)$ for CI.
- Default in SAS.

Interpretation

- For a given time t, a 95% CI for $S(t)$ is

  $$\left[\hat{S}(t)^{\theta}, \hat{S}(t)^{\frac{1}{\theta}}\right]$$
- How to interpret this CI?
  - If repeated samples were taken and the 95% CI is calculated for each sample, then about 95% of these CIs would contain the true value of $S(t)$.
  - But for any particular sample, the confidence bounds are fixed. It either covers $S(t)$ or not.

Interpretation

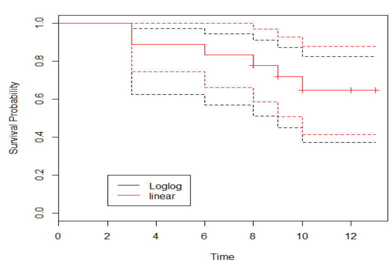
- This CI is point-wise: for any fixed point t, its coverage probability is $100(1-\alpha)\%$.
- It doesn't tell us the coverage probability of the entire curve.
- To make inference about the entire survival curve, we need confidence bands with $100(1-\alpha)\%$ coverage for the entire curve.
- Hall & Wellner (1980) proposed a confidence band, but the formula is very complicated.

Data example

t	$\hat{S}$(t)	V($\hat{S}(t)$)	95% naïve CI	95% loglog CI
3	0.89	0.0055	(0.74,1.03)	(0.62 , 0.97)
6	0.83	0.0076	(0.66,1.01)	(0.57 , 0.94)
8	0.78	0.0096	(0.59,0.97)	(0.51, 0.91)
9	0.72	0.0121	(0.51,0.93)	(0.45, 0.87)
10	0.65	0.0137	(0.41,0.88)	(0.37, 0.83)

Interpretation(loglog CI): we are 95% confident that the probability of surviving more than 3 weeks is between 0.62 and 0.97.

Data example



Summary

- Greenwood's formula
- Confidence intervals for S(t)
  - Naïve CI
  - Loglog transformed CI
- Confidence bands