

A.5 Censoring and truncation, Kaplan-Meier estimator

Please hand in your work by Monday 2 March 2009, 4pm, at the Department of Statistics.

- Explain what is meant by right censoring, left censoring, right truncation, left truncation.
 - In a study of the elderly, individuals were enrolled in the study, at varying times, if they had already had one episode of depression. The event of interest was the onset of a second episode. An individual could be enrolled if at some previous time an episode of depression had been diagnosed. Which of the above mechanisms are relevant if it is also known that the study finished after four years?
 - In 1988 a study was published of the incubation time (waiting time from infection until symptoms develop) of AIDS. The sample was of 258 adults who were known to have contracted AIDS from blood transfusion. The data reported were the date of the transfusion, and the time from infection until the disease was diagnosed. Which of the above mechanisms are relevant for analysing these data?
- Prove that in a single sample with no censoring, the Kaplan-Meier estimator for the survival function satisfies

$$\hat{S}(t) = 1 - \hat{F}(t),$$

where $\hat{F}(t)$ is the empirical distribution function (right-continuous).

Below are remission times (in weeks) of leukaemia patients in the control group (Gehan used these in a 1965 paper):

1, 1, 2, 2, 3, 4, 4, 5, 5, 8, 8, 8, 8, 11, 11, 12, 12, 15, 17, 22, 23.

There are no censored subjects. Find and plot the Kaplan-Meier estimator for this set of data. Compare with the estimator based on the Nelson-Aalen estimator for the same dataset. Derive a 95% confidence interval for $S(t)$ at time $t = 4$ using Greenwood's formula.

- If x is the observed value of a random variable $X \sim \text{Binom}(n, p)$, with known n , find the maximum-likelihood estimator \hat{p} , and deduce that

$$\text{Var}(\hat{p}) \approx \frac{x(n-x)}{n^3}.$$

If $\hat{S}(t)$ is the Kaplan-Meier estimator, an alternative estimator for the variance is

$$\text{Var}(\hat{S}(t)) = \frac{\hat{S}(t)^2(1 - \hat{S}(t))}{n(t)}$$

where $n(t)$ is the number at risk at time $t+$. If $d(t)$ is the number of failures up to and including time t , justify the estimation

$$\hat{S}(t) \approx \frac{n(t)}{n(t) + d(t)} = \frac{n(t)}{n(0)},$$

making the conservative assumption that all the censoring in the interval $[0, t)$ takes place at $t = 0$. What is the distribution of $d(t)$ given this assumption? Explain how this can be used to justify the expression for $\text{Var} \hat{S}(t)$ in terms of a binomial proportion estimator (as \hat{p} above). In the special case of no censoring, what is the connection between this estimator and Greenwood's estimator for the variance?

4. We are carrying out a hypothetical study of the survival of Alzheimer patients. We enrol 30 subjects in a clinic, and follow them over five years. We record their age at being enrolled in the study and the age at which they left, and the cause of exit, whether death (1) or something else (0).

Entry Age	Exit Age	Death Indicator	Entry Age	Exit Age	Death Indicator
67	72	0	69	74	1
70	71	0	69	71	0
70	73	1	66	68	0
65	70	0	73	76	1
65	68	1	67	68	0
73	78	1	66	70	1
69	74	1	69	73	1
76	78	1	66	70	1
66	67	0	78	81	1
72	76	1	66	70	1
65	70	1	68	73	1
71	75	1	70	74	1
69	71	0	66	68	0
71	74	1	89	92	1
68	73	0	68	72	1

- (a) What sorts of censoring and/or truncation do we have in this study?
- (b) Make a table indicating the number of subjects at risk at ages from 65 to 75.
- (c) Estimate the survival curve over this age range.
- (d) Compute a 95% confidence interval for the survival probability from age 70 to 75.