A.5 Censoring and truncation, Kaplan-Meier estimator

1. Explain the types of censoring and truncation which are relevant in the following situations:

   (a) Children are tested at monthly intervals from 12 to 18 months of age. On each occasion one of the tests examines the acquisition of the ability to perform a simple task involving tool use. The distribution of the age at which this skill is acquired is studied.

   (b) A study examines the time until the onset of a second episode of clinical depression, in individuals who have suffered a previous episode. An individual can be enrolled in the study if they have suffered precisely one such episode in the past. The study lasts for four years.

   (c) Patients who first experience symptoms of malaria after returning to the UK from a malarial region are observed. The period of their stay in the region and the time of onset of symptoms is obtained, and used to study the incubation time of malaria (i.e. the time from infection to symptoms), which is typically on the order of a few weeks.

2. An example of "Type II censoring" is as follows: failure times of n machines are observed, all started at the same moment. Once k machines have failed, observation ceases. The observed data can be written as $x_1, x_2, \ldots, x_k, x_k+, x_k+, \ldots, x_k+$, where $x_1 \leq x_2 \cdots \leq x_k$ and where $+$ indicates a right-censored observation.

   Write down the likelihood function in terms of the underlying lifetime distribution.

   What form will the Kaplan-Meier estimate of the distribution take in this case?

3. Show that when no censoring or truncation occurs, the Kaplan-Meier estimator corresponds to the empirical distribution function.

4. (a) Adapt the delta-method procedure used in the lectures to justify Greenwood's estimate for the Kaplan-Meier curve, to give the following approximation for the Neslon-Aalen estimator of the survival distribution:

$$\operatorname{Var} \tilde{S}(t) \approx \tilde{S}(t)^2 \sum_{i:t_i \leq t} \frac{d_i(n_i - d_i)}{n_i^3}.$$

   (b) Rather than a normal approximation for the Kaplan-Meier estimator $\hat{S}(t)$ itself, or for $\log \hat{S}(t)$, some authors recommend a normal approximation for $\log\big(-\log \hat{S}(t)\big)$. One advantage is that any value in $(-\infty, \infty)$ is a possible value for this quantity (which is not the case for $\hat{S}(t)$ or $\log \hat{S}(t)$). Starting from the variance of $\log \hat{S}(t)$ as used in the justification of Greenwood's estimate, use the delta method again to obtain an approximation for the variance of $\log\big(-\log \hat{S}(t)\big)$. Find the form of the confidence interval for $\hat{S}(t)$ which results from this approach.

5. Plot (or sketch) the Kaplan-Meier curve for the following sample of size 15:

   $$1.75, 1.89, 1.92, 2.17+, 2.24, 2.27+, 2.60+, 2.88, 3.10+, 3.84+, 4.25, 4.81, 5.05+, 5.11+, 5.25$$

   where $+$ indicates a right-censored observation. Estimate the variance of $\hat{S}(4)$ using Greenwood's formula. Find a 95% confidence interval for $S(4)$ using one (or more, if you like) of the approaches mentioned above.

6. The table below shows data collected by a group of specialist care homes on their residents. For each individual, the following data was available: (curtate) age on entering the home, and age on death or on exit from the home for another reason (or current age for those still living and resident). The following shows the number of arrivals, deaths and other exits recorded at each age:

Age	Arrivals	Deaths	Other Exits
68	1	0	0
69	5	0	2
70	10	2	1
71	24	5	3
72	24	6	3
73	11	9	9
74	11	8	3
75	4	5	4
76	4	6	2

   (a) What sorts of censoring and/or truncation do we have in this data?

   (b) Make a table showing the number of individuals at risk at each age (suitably approximated).

   (c) Find a confidence interval for the probability that an individual on his or her 71st birthday survives a further 5 years.

   (d) What assumptions about the population do we need to make in the calculation above? Are they reasonable?