HT 2007: Statistical Lifetime Models, Sheet 4

1. 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 occur if it is also known that the study finished after four years?

2. Prove the assertion that in a single sample with no censorship then the Kaplan-Meier estimator for the survival function satisfies
$$\widehat{S}(t) = 1 - \widehat{F}(t)$$
   where $\widehat{F}(t)$ 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.

3. If x is the observed value of a random variable $X \sim B(n, p)$, find the maximum likelihood estimator $\widehat{p}$ and deduce that
$$\text{var } \widehat{p} \approx \frac{x(n-x)}{n^3} \; .$$
   If $\widehat{S}(t)$ is the Kaplan-Meier estimator, an alternative estimator for the variance is
$$\text{var } \widehat{S}(t) = \frac{\widehat{S}(t)^2(1-\widehat{S}(t))}{n(t)}$$
   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
$$\widehat{S}(t) \approx \frac{n(t)}{n(t)+d(t)} = \frac{n(t)}{n(0)}$$
   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 var $\widehat{S}(t)$ in terms of a binomial proportion estimator (as $\widehat{p}$ above). In the special case of no censoring what is the connection between this estimator and Greenwood's estimator for the variance?

4. The simplest parametric model used in lifetime analysis is exponential with parameter λ. Write down the hazard, the integrated hazard and the survival function in this case. Given data with failure times $t_{1,}t_2, \cdots t_r$ and right-censored times $c_1, c_2, \cdots c_s$, assuming that censoring is non-informative, find the maximum likelihood estimator of λ. Give an approximate estimator for its variance.

   A more complicated model would be to assume a piecewise continuous hazard rate, λ_i in the ith interval $[a_{i-1}, a_i)$, where $a_0 = 0$, $a_n = \infty$. Show that the maximum likelihood estimator $\widehat{\lambda_i}$ is essentially the number of failures in the interval divided by the total time at risk in the interval.

   Suppose we have the following model

$$\begin{aligned} \lambda &= 0.1, 0 < t < 2 \\ &= 0.05, 2 \leqslant t \end{aligned}$$

   representing the hazard rates for failure of a spotlight bulb in time t (months). A more expensive bulb has hazard rate 0.05 throughout. Calculate the expected lifetime of each bulb. Viewed over the long term what would be the ratio of costs of expensive to less expensive bulb if the long term expenditure is the same?