Exercises I (for slides 1-49 of notes)

1. For the dataset $\left\{t_1 = 5, t_2 = 10, t_3^* = 15, t_4 = 20, t_5^* = 25, t_6 = 30\right\}$, where the asterisks denote censoring times, compute the K-M and N-A estimators. From each of these compute by hand the corresponding discrete density (probability mass function), showing that these are identical.

2. The sampling distribution of the estimated survival function is more nearly normal on the scale $\hat{V}(t) = \log\left[-\log\left\{\hat{S}(t)\right\}\right]$. Use the "delta method" and Greenwood's formula (notes) to obtain an approximate expression for this, and evaluate this at $t = 20$ for the above example.

3. For a sample $\left\{t_i,\, f_i = 0,1\right\}$ where the latter coordinate indicates failure, with a parametric model $\lambda_i(t, \theta)$ for the hazard, show that (under usual assumptions regarding "independent" [or "uninformative"] censoring) the loglikelihood can be expressed as

$$l = \sum_{i=1}^{n} f_i \log \lambda_i(t_i, \theta) - \int_0^{t_i} \lambda_i(s, \theta) ds$$

4. Show that the corresponding score can be expressed as

$$\partial l / \partial \theta = \sum_{i=1}^{n} \int_0^{t_i} \left\{ \frac{\partial \log \lambda_i(s, \theta)}{\partial \theta} \right\} \left\{ dN_i(s) - \lambda_i(s, \theta) ds \right\}$$

5. Fully explore using Cox regression for the cervical cancer data *cervical_cancer.dta* . The main interest is on the effect of screening, but you must get some understanding of the effects of stage and diagnosis-age. Using a few categories of diagnosis-age will help with this. In the final analyses you should stratify on stage (see slide 36), including some effect of diagnosis-age as well as screening.

6. The dataset *gastric.dta* (using time scale *survtime*) has extremely non-proportional hazards for the two treatment levels. Carry out analysis similar to that in the lectures (around slides 27-32) in regard to this.