B.3 Modern Survival Problem sheet 3: Estimating quantiles and excess mortality

To be turned in by noon on 6 November, 2015

(1) Show that Duhamel's equation (6.5) holds at a point s where S_1 or S_2 is discontinuous.

(2) Figure B.1 gives code to calculate confidence for quantiles of the survival curve. Here `SF` is the output of the `survfit(S∼1)` command for a survival object `S`.

```
#Output is a confidence interval for the p-quantile of survival
quantileCI=function(SF,p,alpha=.05){
        sb.fitNA=NAest(SF)
        z=qnorm(1-alpha/2)
        a=-log(p)
        se=sqrt(sb.fitNA$Var)
        xpless=sb.fitNA$Hazard+z*se
        xpmore=sb.fitNA$Hazard-z*se
        upper=max(which(xpmore<a))
        lower=min(which(xpless>a))
        c(sb.fitNA$time[lower],sb.fitNA$time[upper])}
```

Figure B.1: Code to compute confidence intervals for survival quantiles.

The function `NAest` is a homemade function, given in Figure 6.3, to compute Nelson–Aalen estimators.

(a) Explain why this is a reasonable estimator for the quantiles of the survival function. (For more information about quantile estimation, see Section 3.2.3 of Aalen's book. Note that the book is available electronically through the Bodleian website.)

(b) Use this to compute a 95% confidence interval for median survival in the `ovarian` data set (ignore treatment type);

(c) Use this to compute a 95% confidence interval for median survival in a collection of data simulated from an exponential distribution with parameter 1, available on the course web site. Note that this file has been produced with the `save` command, and can be loaded into `R` with the command `load('filename')`. You will then have two vectors of length 1000, called `T` and `ev`. The former is the time of event or right-censoring, the latter is `TRUE` (for event) or `FALSE` (for censoring).

(d) Suppose we ignored the censoring, so only included the uncensored times. What would you estimate for the median survival time?

(3) Look back to the derivation in the lecture notes section 7.2 of the estimator $\hat{\Gamma}(t)$ for cumulative excess mortality in the two-sample setting. Think of $k_c(t)$ as arbitrary predictable random variables.

(a) Construct a martingale to show that the estimator (7.5) for excess mortality in the two-sample case is unbiased for appropriate choice of $k_c(t)$. What conditions must $k_c(t)$ satisfy?

(b) Find an expression for estimating the variance of $\hat{\Gamma}$.

(c) Show that for the particular choice (7.4) the bound (7.6)

$$\mathrm{Var}(\hat{\Gamma}(t)) \approx \sum_{t_i \leq t} \left(\sum_c Y(c, -; t_i) \right)^{-2}$$

is a conservative estimate for the variance of the estimator. That is, it is a good estimate for large samples, and tends not to underestimate the variance.

(d) Since any choice of k_c yields an estimator, we are free to make a convenient choice. Why is the choice (7.4) a good one?

(e) Supposing the groups to be of approximately equal size, what will the relation be between the variance of our estimator for the cumulative excess mortality, and the variance we would estimate for the difference in the cumulative hazards between the groups $\{G_i = 0\}$ and $\{G_i = 1\}$, ignoring the classification c_i.