1. (a) [10 marks] Define the following terms. Make sure that all notations are explicitly described.

(i) *Interval censoring*;

(ii) *relative risk model*;

(iii) *predictable* stochastic process;

(iv) *frailty model*;

(v) *Duhamel's equation*.

(b) [5 marks] Let $\lambda : \mathbb{R}_+ \to \mathbb{R}_+$ be any integrable positive function, and $\Lambda(t) = \int_0^t \lambda(s)ds$. Suppose X is a random variable with exponential distribution with parameter 1. Show that $\Lambda^{-1}(X)$ is a random variable with hazard rate $\lambda(t)$.

(c) [10 marks] For a certain population of laboratory mice, we have the null hypothesis H_0, that females have mortality rate $\lambda_f(t) = \theta e^{0.25t}$ at age t (in years), while males have mortality rate $\lambda_m(t) = 2\theta e^{0.2t}$, for some (unknown) value of θ.

We begin with 100 newborn mice of each sex. Suppose that we observe some times of death $t_1 \leq \cdots \leq t_{k_m}$ for males and $s_1 \leq \cdots \leq s_{k_f}$ for females; and right censoring times $c_1 \leq \cdots \leq c_{100-k_m}$ for males and $d_1 \leq \cdots \leq d_{100-k_f}$ for females, where censoring times are assumed to be independent of death times. Let $Y_m(t)$ and $Y_f(t)$ be the count of the number of male and female survivors still under observation at time t.

(i) Explain how to compute $Y_m(t)$ from the event and censoring times.

(ii) Find a function $g(t)$ such that

$$M(t) := \sum_{t_i \leq t} \frac{1}{Y_m(t_i)} - \sum_{s_i \leq t} \frac{g(s_i)}{Y_f(s_i)}$$

is a martingale if the null hypothesis holds, regardless of the value of θ.

(iii) Write down an unbiased estimator for the variance of $M(t)$.

(iv) Explain how we may use these facts to construct an approximate significance test of the hypothesis H_0 at a given level α.

(v) How might we modify this significance test to improve the power?

2. (a) [3 marks] Describe the *multiplicative intensity model.*

(b) [2 marks] Give an example of a data situation that would not satisfy the assumptions of the multiplicative intensity model.

(c) [10 marks] Let $N(t)$ be the counting process associated with a non-homogeneous Poisson process with rate $\lambda(t) = t^2$.

(i) What is the compensator of $N(\cdot)$?

(ii) Compute the expectation and variance of $N(t)$.

(iii) Compute the variance of $\int_0^1 s dN(s)$.

(iv) Suppose the jumps on the interval $[0, 2]$ are at times 1.01, 1.44, 1.82. Sketch $N(t) - A(t)$ for $t \in [0, 2]$, where A is the compensator of N. (For ease of calculation, note that $1.44^3 \approx 3$ and $1.82^3 \approx 6$.)

(d) [10 marks] We are given data for n individuals, including a survival time T_i (possibly right-censored) and a single covariate $x_i > 0$ (for $0 \le t \le T_i$), which are all distinct. (The covariates are constant in time.) As usual, we list the event times in order as $t_1 \le t_2 \le \cdots \le t_k$. We believe that an additive-hazards model should hold, where the hazard for individual i at time t is $\beta_0(t) + \beta_1(t)x_i$.

Let $R(t)$ be the set of individuals at risk at time t, and define

$$S_k(t) = \sum_{i \in R(t)} (x_i)^k \text{ for } k = 0, 1, 2.$$

Explain why

$$\hat{B}_1(t) := \sum_{t_j \le t} \frac{S_0(t_j)x_{i_j} - S_1(t_j)}{S_2(t_j)S_0(t_j) - S_1(t_j)^2}$$

is a reasonable estimator for $B_1(t) = \int_0^t \beta_1(s)ds$ for all t substantially less than $\max T_i$.

3. (a) [9 marks] Archaeologists find 10 skeletons of a certain species of dinosaurs, to which they are able to assign approximate ages at death based on physical evidence. Depending on the biological material available they may be able to estimate the precise age at death, or or they may be able to determine an upper bound on the age at death (that is, to say that the dinosaur died at or before a certain age). The data are as follows:

```
Precise ages: 2 2 3 3 3.5 5
Upper bound ages: 4 4 5
```

(i) What kind of censoring or truncation is this?

(ii) Estimate the survival function for all t, taking account of all the information.

(b) [4 marks] Describe the power variance function (PVF) family of frailty distributions. What is one feature of this family that makes it particularly useful for frailty models?

(c) [12 marks] Suppose we have a data frame `stroke` in `R` that includes results of a study of men at elevated risk for stroke, followed for 6 years, to determine whether body mass index (BMI) were useful for predicting stroke risk. The data frame includes variables `time` (patient's time on test), `delta` (= 1 if the individual had a stroke at that time, = 0 otherwise), and `BMI`. We produce the following output:

```
require(survival)
require(KMsurv)

data(stroke)
lar.cph=coxph(Surv(time,delta)~BMI,data=stroke)

      coef exp(coef) se(coef)    z    p
age 0.0953      1.10   0.04 2.38 0.009

Likelihood ratio test=5.41  on 1 df, p=0.010  n= 828, number of events= 44

lar.fit=survfit(lar.cph,newdata=data.frame(BMI=20))
# baseline estimated hazard curve, taking BMI=20 as baseline
T=lar.fit$times  # Times of events or censoring
S=lar.fit$surv   # Estimate of baseline survival
```

(i) Explain the significance of the number in the column 'z', and how that number was computed.

(ii) How many men in the study were right-censored?

(iii) The threshold for being considered "overweight" is BMI=25, and the threshold for being considered "obese" is BMI=30. Explain how you would compute a 95% confidence interval for the ratio of stroke risk between a man whose BMI is at the obesity threshold and a man whose BMI is at the overweight threshold.

(iv) Suppose that `T[48]=3.0`. Write `R` code that will compute the estimated probability that a man in the sample whose BMI is 30 will survive more than 3 years.

(v) Suppose we have generated a vector `B` that has the same length as `T` and `S`, and `B[i]` gives the BMI of the man who had a stroke event or censoring at time `T[i]`. Write `R` code that will compute the Cox-Snell residuals.

(vi) Explain how the Cox-Snell residuals are used to assess the appropriateness of the Cox proportional hazards model for these data.

End of Last Page