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

(i) *martingale*;

(ii) *relative risk regression model*;

(iii) *interval censoring*;

(iv) *hazard rate* for a positive continuous random variable T;

(v) *predictable* stochastic process.

(b) [8 marks] We observe 10 identical machines, which are known to each have a hazard rate of failure at time t (in weeks) of $\lambda(t) = 0.1 + 0.02t$. The machines work independently, and they are run until they fail, or until they reach their prescheduled maintenance time. The times recorded are

```
5.6  2.0  0.9+  8.8  7.8+ 17.6 11.9  2.8 19.3 20.5+,
```

where + denotes a maintenance time, rather than an event (failure) time. Let $N(t)$ be the counting process associated with the failure times; let A be the compensator of N; let $M(t) = N(t) - A(t)$.

(i) Compute $A(5)$.

(ii) Compute the predictable variation $\langle M\rangle(5)$.

(iii) Compute the optional variation $[M](5)$.

(c) [7 marks] A study is performed to measure the time required for rats to get through a maze to find the food reward. There are 50 subject animals, each of whom runs the maze once. The data have been recorded in three vectors of length 50: `T` is the vector of times that a trial was stopped (either because the rat found the target or seemed too tired to continue; `delta` records 1 if the rat found the target, 0 if the trial was stopped for some other reason; `type` records 1,2, or 3, depending on whether this rat received enhanced diet (1); enhanced exercise (2); or no treatment. A researcher uses the following command to fit a Cox proportional hazards model: `rat.cph=coxph(Surv(T,delta)∼type).`

(i) Why is this not an appropriate model?

(ii) Write one or more lines of `R` code that will calculate an appropriate model.

(iii) Describe two different plots that you could make to test whether the data fit the assumptions of the model you defined. Describe the quantities to be plotted, and explain what features one would look for to find evidence of a bad fit.

2. (a) [8 marks] Let $N(t)$ be the counting process associated with a Poisson process with parameter λ, and let

$$f(t) = \begin{cases} 1 & \text{if } t \leq 10, \\ 2 & \text{if } t > 10. \end{cases}$$

We have a realisation in which $N(20) = 4$, and the jumps on $[0, 20]$ are at times `4.7 9.2 11.0 17.9`.

(i) Compute $\int_0^{20} f(s)N(s)ds$.

(ii) Compute the variance of $\int_0^{20} f(s)N(s)ds$.

(iii) Compute $\int_0^{20} sN(s)ds$.

(iv) Compute the variance of $\int_0^{20} sN(s)ds$.

(b) [10 marks] Suppose we have a relative risk model where individual i with covariate x_i has hazard rate $r(\beta, x_i)\alpha_0(t)$ at time t. We observe right-censored data (T_i, x_i, δ_i), where δ_i is the indicator of an uncensored observation. There are no ties among the times T_i. We have computed the maximum partial likelihood estimator $\hat{\beta}$.

(i) Write down Breslow's estimator for the baseline cumulative hazard A_0. Define all variables used in your expression.

(ii) Explain why Breslow's estimator is approximately unbiased estimator for the baseline cumulative hazard.

(iii) Suppose you wish to test the appropriateness of the relative risk model by using Cox-Snell residuals. Explain how they would be calculated.

(iv) Describe, with proof, the distribution of the Cox-Snell residuals, under the hypothesis that the observed times were sampled from the given model.

(v) Explain what plot you would make to test whether the Cox-Snell residuals have the appropriate distribution. State what features of the plot would show a good or bad fit.

(c) [7 marks] Death times for a zoo population of canaries are recorded:

```
4 6 11 6 <7 9 3 <7 5 <4
```

where a number is recorded with < to denote the canary in question is known to have died before that age, but it is not known exactly when. Estimate the survival function for all t, taking account of the censoring.

3. (a) [7 marks] A population is composed of two genetic types, in equal proportions at birth. Type A has mortality rate (hazard rate of death) $2 + t$ at age t; type B has mortality rate $4 + t$ at age t. No one enters or leaves the population.

(i) Compute an expression for the population mortality rate as a function of age;

(ii) What fraction of the population aged 2 is type A?

(b) [11 marks] In a study, patients at risk for stroke are randomly assigned to an exercise regime ($G_i = 1$) or no special exercise ($G_i = 0$). The goal is to determine how much effect (if any) the exercise had on the time T_i until a subsequent stroke. T_i is a censored observation: if $\delta_i = 1$ then T_i is the time of a stroke; otherwise, it is the time when the patient left the study for independent reasons. The times recorded are all distinct.

Suppose that individuals in group 0 have hazard rate $\lambda_0(t)$ at time t, and those in group 1 have hazard rate $\lambda_1(t)$ at time t. We define, as usual, $Y_g(t)$ to be the number of subjects still under observation up to time t in group g. We let $\tau := \inf\{t : Y_0(t)Y_1(t) = 0\}$. An estimator for the difference in cumulative hazards

$$\Gamma(t) = \int_0^{t\wedge\tau} \lambda_1(s)ds - \int_0^{t\wedge\tau} \lambda_0(s)ds$$

may be presented as

$$\hat{\Gamma}(t) = \sum_{t_j \le t\wedge\tau} k(t_j)\left(\frac{G_j}{Y_1(t_j)} - \frac{1-G_j}{Y_0(t_j)}\right)$$

[Recall that $t \wedge \tau := \min\{t, \tau\}$.]

(i) Define all the terms in the above estimator. (That is, what are t_j, $k(t)$, G_j, $Y_1(t)$ and $Y_0(t)$?)

(ii) Show that $\hat{\Gamma}(t)$ is an unbiased estimator for $\Gamma(t)$.

(iii) Derive an unbiased estimator for the variance of $\hat{\Gamma}(t)$.

(iv) How could we test the hypothesis that the exercise has no effect on stroke risk?

(c) [7 marks]

(i) State one advantage of additive hazards regression over relative risk regression. (This should be a general advantage, not the fact that in a particular case additive hazards may be better suited to the data.)

(ii) We have survival times that fits the additive hazards model with a single covariate X, so the hazard at time t is $\beta_0(t) + x\beta_1(t)$ for a subject with $X = x$. Suppose now that X is normally distributed, but that the observed covariate is $Y = X + \epsilon$, where ϵ is an independent normal error. Show that the model for survival time as a function of the observed covariate Y is also an additive hazards model, and compute the new regression functions $\beta_0'(t)$ and $\beta_1'(t)$.