UNIVERSITY OF CAMBRIDGE

M. PHIL. IN STATISTICAL SCIENCE

Friday 6 June 2008 9.00 to 11.00

BIOSTATISTICS

Attempt no more than ***THREE*** *questions.*

There are ***FIVE*** *questions in total.*

The questions carry equal weight.

STATIONERY REQUIREMENTS
Cover sheet
Treasury Tag
Script paper

SPECIAL REQUIREMENTS
None

You may not start to read the questions printed on the subsequent pages until instructed to do so by the Invigilator.

1 Statistics in Medical Practice

Table 1
Counts of events

Control	5,	4,	7,	2,	6,	5,	9,	8,	6,	4,	4,	7
Treated	11,	10,	12,	14,	6,	7,	8,	10,	8,	18,	11,	18

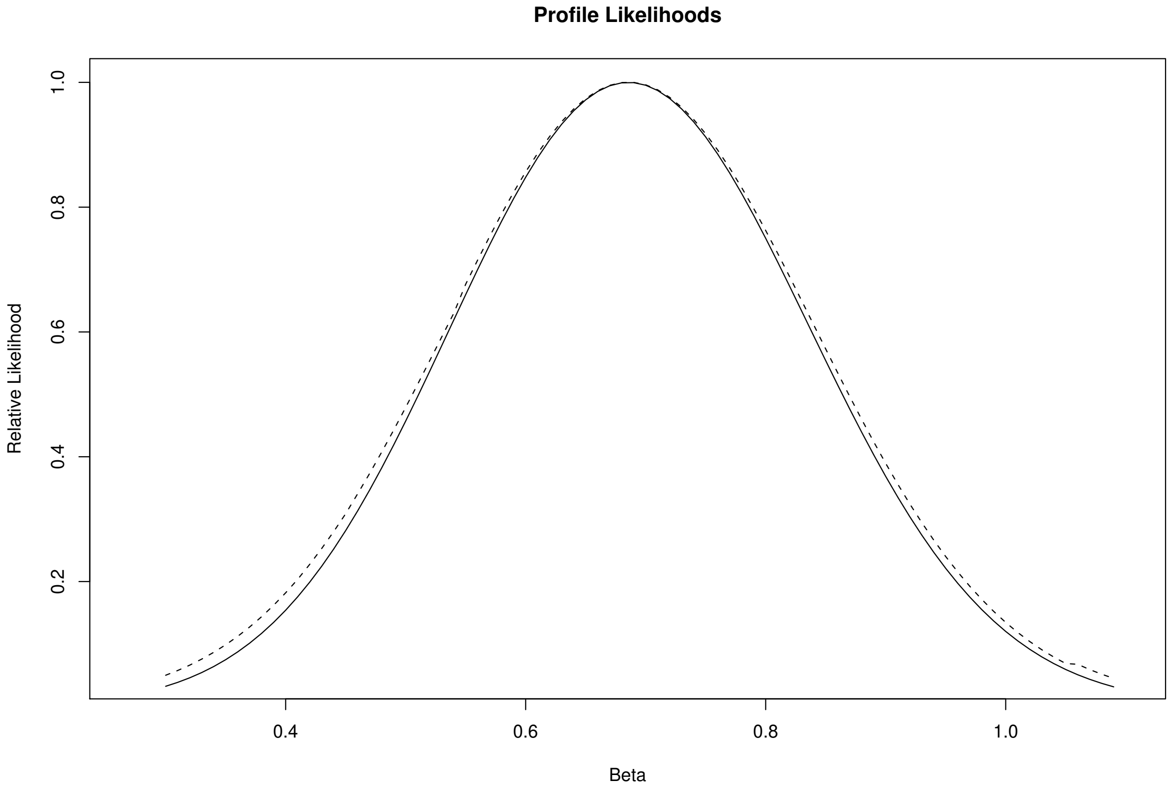

Figure 1

Table 1 presents data on the number of positive communication events initiated by physicians during a filmed patient consultation. Data are presented from 12 'control' physicians with no additional communication training and from 12 'treated' physicians who attended a communication course.

Denote the count for physician i by y_i, $i = 1, \ldots, 24$. Assume that x_i is defined to be a binary covariate taking the value 0 for 'control' physicians ($i = 1, \ldots, 12$) and the value 1 for 'treated' physicians ($i = 13, \ldots, 24$). Assume that y_i is a Poisson variate with mean $\lambda \exp(\beta x_i)$. Further, let $\lambda = \exp(\alpha)$.

(a) The maximum likelihood estimates of α and β are 1.72 and 0.69 respectively, with respective estimated asymptotic standard errors of 0.12 and 0.15. Explain how to use these values to construct a 95% confidence interval for the ratio of the mean number of positive communication events for 'treated' physicians to the mean number for 'control' physicians.

(b) Through an appropriate choice of a conditioning event, E, derive a conditional likelihood for the estimation of β that will eliminate the parameter λ. Express the conditional likelihood as a function of $p = \exp(\beta)/(1 + \exp(\beta))$.

(c) The solid line in Figure 1 presents a profile likelihood for the estimation of β based on unconditional maximum likelihood estimation of α and β in the Poisson model outlined earlier. Also given in Figure 1 is a profile likelihood for the estimation of β based on unconditional maximum likelihood estimation of a negative binomial model for the counts Y , where the mean structure for the model is the same as the Poisson model given earlier and there is an additional shape parameter. Comment briefly on the differences between the two profile likelihoods in Figure 1 and suggest why they differ.

(d) Outline briefly a statistical model which might be used to analyze the data in Table 1 if the observations in the two groups were paired (say by order of presentation in the table) because each observation in the pair came from the same physician, say before and after a communication course. Assume the effect of the communication course is common across physicians and represented in a similar manner as that in the model outlined earlier. What conditioning would be required to derive a conditional likelihood for the estimation of β in this case? Would conditional likelihood estimation be more or less valuable in this situation than for the model used in part (b)? Is there a possible concern with the use of conditional likelihood in this situation and, if so, can it be alleviated in any way?

2 Statistics in Medical Practice

(a) There is compulsory drugs testing in the armed services. Between 2003 and 2007, the annual number of cocaine positives in Monday tests increased from 250 to 500. In 2007, 30% of the Monday cocaine-positive tests gave a maximum read-out. Street cocaine is believed to be weaker in 2007, because more heavily 'cut' than in the two-year period 2003-2004. Could one construct a 5%-level test with 80% power, to detect statistically whether the maximum read-out percentage had changed from 40% in 2003-2004 to 30% in 2007, if the data on 500 Monday cocaine positives in 2003-2004 were to be retrieved?

(b) One third of male prisoners aged 18-44 years have a history of heroin injection. One in 200 of them will die from heroin overdose within 4 weeks of release from prison. To prevent plausibly 30% of these deaths, a major trial will randomise individually-consented 18-44 year old prisoners with a history of heroin injection to receive, on their release from prison, either a pack containing a syringe pre-loaded with Naloxone, the heroin antidote, to be administered intramuscularly in the event of overdose, or a control pack.

(i) How many consented eligible prisoners need to be randomised for the proposed trial to have 80% power to discern a 30% reduction in heroin overdoses within 4 weeks of release for prisoners randomised to Naloxone?

(ii) Inmates with a history of heroin injection may be detoxified on arrival in prison, or they may receive methadone maintenance, or neither. Methadone maintenance will be increasingly available at more prisons over the course of the trial. Give your reasoning for a recommended method of randomisation that can ensure balanced allocations (between Naloxone and control packs) by prison and also by how the inmate's addiction was managed.

(iii) A referee proposed that prisons, not prisoners, should be randomised as this would save on staff costs in the control prisons. Give one ethical and one statistical reason against the referee's proposal.

3 Survival Data Analysis

(a) A time-to-event distribution has integrated hazard $H(t) = (\lambda t)^p$ for $\lambda > 0$ and $p > 0$.

Obtain the hazard, density and survivor functions for this distribution. Show that the two time-to-event distributions obtained by setting $\lambda = \lambda_1$ and $\lambda = \lambda_2$ respectively (with $p_1 = p_2$) belong both to the same proportional hazards family and the same accelerated life family.

(b) A time-to-event dataset consists of n individuals with x_i being the event time ($v_i = 1$) or the censoring time ($v_i = 0$) of the ith individual. Write down a general expression for the log-likelihood in terms of the hazards $h_i(t)$ and integrated hazards $H_i(t)$.

(i) Obtain the log-likelihood in the case $H_i(t) = (\lambda t)^p (\lambda > 0,\ p > 0)$.

(ii) Obtain the maximum likelihood estimate of λ in the case $H_i(t) = \lambda t$ $(\lambda > 0)$.

(iii) Now assume the hazard functions may differ across individuals and the integrated hazard for the ith individual is given by $H_i(t) = \lambda t + G_i(t)$ $(\lambda > 0,\ G_i(0) = 0)$ where the $G_i(t)$ are known increasing differentiable functions, with differential $G_i'(t)$ small compared with λ. Obtain an approximate maximum likelihood estimator for λ in this case.

4 Survival Data Analysis

Outline a derivation of the *Kaplan-Meier* estimate $\hat{F}(t)$ of the survivor function $F(t) = \mathcal{P}(T > t)$.

(a) A survival dataset is such that no individuals are censored in the interval $a < t < b$. Show that, if d individuals have events in the interval $a < t \leqslant b$, then $\hat{F}(b)$ does not depend on the order of the event times, or on whether or not any event-times are tied.

(b) A researcher enrols 100 subjects into a time-to-death study on 1st January 2006 (cohort A) and 100 subjects into the same study on 1st January 2007 (cohort B).

Of the 100 subjects belonging to cohort A, 28 died in the time period 1st January 2006 to 31 December 2006 inclusive and 24 died in the time period 1st January 2007 to 31st December 2007 inclusive; the remainder were all alive on 1st January 2008.

Of the 100 subjects belonging to cohort B, 31 died in the time period 1st January 2007 to 31st December 2007 inclusive; the remainder were all alive on 1st January 2008.

(i) Calculate the Kaplan-Meier estimate $\hat{F}(t)$ for both $t = 1$ year and $t = 2$ years.

(ii) The researcher is particularly interested in two-year survival. She proposes to estimate $F(2)$ using only cohort A because "they are the only subjects who could have survived that long". Criticise her proposal.

5 Survival Data Analysis

(For the whole of this question you may assume there are no ties in the survival dataset.)

(a) What is the *Cox regression* formulation of proportional hazards modelling in survival analysis? Derive the likelihood for the parameter vector β and obtain the partial derivatives of the log-likelihood with respect to β.

(b) Define and interpret the *Schoenfeld residual.* Show that their sum over the observed event times is zero.

(c) Describe without detailed calculations how Schoenfeld residuals can be used to test the assumptions of a proportional hazards model.

END OF PAPER