UNIVERSITY OF CAMBRIDGE

MATHEMATICAL TRIPOS Part III

Monday, 8 June, 2015 9:00 am to 12:00 pm

PAPER 35

BIOSTATISTICS

*Attempt no more than **FOUR** questions, with at most **THREE** questions from **Analysis of Survival Data**.*

*There are **SEVEN** 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

In a survey of sexual attitudes and lifestyles a random sample of individuals was taken from a population and sampled individuals were asked about their sexual behaviours. Researchers were interested in the association between condom use and number of sexual partners in the past five years in the population of individuals who had more than one sexual partner in the past five years. Some individuals declined to provide information about condom use and/or number of sexual partners. The researchers consider it probable that, among individuals with more than one sexual partner, those with many sexual partners and those who did not use condoms were less likely to provide information about these variables.

(a) Consider the data on condom use and number of sexual partners. Define what it means for these data to be missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR).

(b) Say whether you think these data are MCAR, MAR or MNAR, and briefly explain why.

(c) The researchers fit a logistic regression model to the dataset. The outcome (i.e. dependent variable) is the binary variable condom use (1=always use condom, 0=not always use condom). The covariates (i.e. independent variables) are the number of sexual partners in the past five years, age and sex (1=male, 0=female). Age and sex are fully observed. The researchers fit the logistic regression model only to the individuals with more than one sexual partner in the past five years and with no missing data on condom use or number of sexual partners.

Provide an intuitive explanation for why the resulting parameter estimates may be biased.

(d) The researchers realise that this analysis could be inefficient because it ignores available information on individuals with incomplete data. They turn to you for advice on how to use multiple imputation by chained equations (also known as multiple imputation by full-conditional specification).

Provide a detailed, step-by-step description of how multiple imputation by chained equations works in general, and how it could be applied to this dataset. (You may assume that the researchers know about Bayesian statistics and are happy with the use of mathematical notation, but remember to define any notation you use.)

(e) In light of your answer to part (b), briefly discuss whether the assumptions underlying this multiple imputation method are plausible.

(f) Sampled individuals were also asked about their alcohol consumption and their history of sexually transmitted disease. Almost all sampled individuals provided information on these variables, and these variables are moderately correlated with condom use and number of sexual partners.

Explain how you could use the information on alcohol consumption and history of sexually transmitted disease in the multiple imputation procedure you described in part (d), and state what the potential advantages are of doing this.

2 Statistics in Medical Practice

A meta-analysis examining the effectiveness of radiotherapy for treatment of localised prostate cancer includes data from 21 randomised trials comparing radiotherapy against prostatectomy. The table shows the data on mortality at 12 months after randomisation, extracted from 5 of these trials.

Trial name	Radiotherapy (Deaths/Total)	Prostatectomy (Deaths/Total) ratio	log odds ratio	Standard error of log odds
Gallo 2004	20/105	17/103	0.17	0.36
Steineck 2008	30/90	20/60	0.00	0.35
Estwick 2011	20/132	15/131	0.32	0.37
Pollard 2002	183/364	171/367	0.15	0.15
Olsen 2006	89/470	77/466	0.17	0.17

(a) Give three reasons which could motivate researchers to carry out a meta-analysis rather than just presenting a table of results from separate trials.

(b) Show the calculation of the log odds ratio and its variance for the Steineck 2008 trial, comparing radiotherapy against prostatectomy. Calculate an approximate 95% confidence interval for the log odds ratio. Interpret the log odds ratio estimate and confidence interval in words. [You may assume that the 97.5% quantile of the standard Normal distribution is approximately equal to 2.]

(c) The sum of the inverse variances of the log odds ratios from all 21 trials is equal to 400 and the sum of their squares is 16000. Calculate the between-trial heterogeneity estimate . [Hint: the Q statistic has been calculated as 38.] Show how to calculate the I-squared statistic. Interpret the values of the heterogeneity estimate and I-squared statistic in words.

(d) Examine the influence of the Steineck 2008 trial in a fixed effect meta-analysis of all 21 trials, by calculating its percentage weight. Give a formula for the percentage weight given to the same trial in a random effects meta-analysis.

(e) Define "within-study bias", briefly describe its causes and effects, and suggest two approaches that researchers could use to address suspected within-study biases in a meta-analysis.

(f) The type of radiotherapy given to patients differed between the trials included in the meta-analysis. Some trials used conventional radiotherapy techniques (RT), while others used 3-dimensional conformal radiotherapy (3DCRT). The combined log odds ratio estimate obtained for the RT subgroup of trials was 0.45, and its variance was 0.11, while that for the 3DCRT subgroup was 0.16, and its variance was 0.05. Carry out a formal test to compare the subgroups and interpret the result.

3 Statistics in Medical Practice

An infectious disease X has an incubation period T between infection and symptoms that may be described by a distribution with density $f(t;\cdot)$. Assume that new infections occur in a continuous-time non-homogeneous Poisson process with rate $h(t)$ at time t.

(a) Write down the back-calculation equation relating the new infection rate $h(\cdot)$ to the rate of new symptomatic infections $\mu(\cdot)$.

(b) Assume now that time is discretised. Write down the discrete-time approximation to the back-calculation equation, defining any notation you use.

(c) Assume now that $f(t;\cdot)$ is the density of a Weibull distribution with scale parameter λ and shape parameter κ, i.e. that

$$f(t;\lambda,\kappa) = \frac{\kappa}{\lambda}\left(\frac{t}{\lambda}\right)^{\kappa-1}\exp\left\{-\left(\frac{t}{\lambda}\right)^{\kappa}\right\}.$$

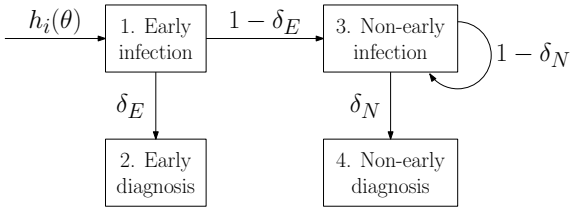
Write down the back-calculation equation in terms of the Weibull distribution, in both continuous-time and discrete-time representations, defining any further notation you use.

The remainder of the question assumes the discrete-time representation of back-calculation.

(d) Denote by h_i the expected number of new infections occurring at the end of a time interval $[t_{i-1}, t_i)$, and assume that h_i is parameterised in terms of a parameter θ, i.e. that $h_i := h_i(\theta)$. Assume a time series $y_{1:N} := (y_1, y_2, \ldots, y_N)$ of counts of new symptomatic infections is observed over the time interval $[t_0, t_N)$, where y_k is the number of symptomatic cases occurring in the interval $[t_{k-1}, t_k)$, for each $k = 1, 2, \ldots, N$. Denote by $Y_k, k = 1, 2, \ldots, N$ the random variables of which the observations $y_{1:N}$ are realisations. Show that for $j \neq k$, Y_j and Y_k are independent Poisson random variables.

(e) Hence write down the likelihood of observing the data $y_{1:N}$ given the parameters θ, λ and κ.

(f) Assume now that disease X is in general diagnosed when symptoms occur after the incubation period, but that for a small proportion of infections, diagnosis occurs early in the progression of the disease. The observations $y_{1:N}$ now represent diagnosis counts, the sum of diagnoses occurring either early or "late". The disease progression and diagnosis process may be represented by the multi-state model below.



Assume that the "early infection" stage lasts 1 time interval, after which individuals are either diagnosed with probability δ_E or progress to "non-early" infection with

probability $1-\delta_E$. Denote the random variables representing the number of individuals entering states 1-4 in the k^{th} interval by $S_k = (S_{1k}, S_{2k}, S_{3k}, S_{4k})$. Assuming no-one is infected prior to time t_0, for the four time points $t_1, \ldots, t_4$, write down expressions for the state incidences $S_k, k = 1, \ldots, 4$, in terms of θ and the diagnosis probabilities δ_E and δ_N.

(g) Hence derive a general expression for p_{Ek}, the expected proportion of new diagnoses that are early at time t_k.

4 Analysis of Survival Data

(a) A time-to-event dataset comprises n individuals: x_i being either the time of the observed event ($v_i = 1$) or the time of censoring ($v_i = 0$) for the ith individual. The common density and survivor function are $f(t)$ and $F(t)$ respectively. Write down the contribution to the likelihood of

(i) an individual censored at x_i;

(ii) an individual with an observed event at x_i.

In the case that $F(t) = \exp(-\theta t)$, $\theta > 0$ derive the log-likelihood function and obtain the maximum likelihood estimator of θ, checking that it is indeed a maximum.

(b) A student attends 13 one-hour lectures given by a lecturer who uses an overhead projector (OHP). During one of those lectures the OHP bulb failed after 30 minutes and was not replaced; during the remaining lectures the bulb did not fail. OHP bulbs have a time-to-failure distribution which is exponential with rate parameter λ. Failed bulbs are replaced between lectures, so - although it is known that a bulb has survived to the start of the lecture - it is not known how long that bulb has survived for.

By considering survival probabilities conditional on having already survived to a particular time:

(i) write down the contribution to the likelihood for λ of a bulb first used at time $t = 0$, which is working at $t = \tau$ and is still working at $t = \tau + 1$.

(ii) write down the contribution to the likelihood for λ of a bulb first used at time $t = 0$, which is working at $t = \tau$ and fails at $t = \tau + \frac{1}{2}$.

Comment on the dependence of your answers on τ. Does it matter that it is not known whether the bulb used in a particular lecture is the same as the bulb used in a previous lecture? Find the maximum likelihood estimate of λ.

How would your answer be different if the failing bulb had been replaced immediately and the replacement bulb did not fail before the end of that lecture?

Suppose instead the time-to-failure followed a Weibull distribution with $F(t) = \exp\left[-(\lambda t)^p)\right]$, $\lambda > 0$, $p > 0$. Could you use this dataset to estimate the parameters? Would it help if it were a larger dataset with more bulb failures?

5 Analysis of Survival Data

A time-to-event dataset comprises n individuals: x_i being either the time of the observed event or the time of censoring for the ith individual. The times of the observed events are a_j, $j = 1, \ldots, d$ with exactly one event occurring at each a_j and $a_j < a_{j+1}$. The individuals are divided into two groups 0 and 1 with g_i indicating the group membership for the ith individual ($g_i \in \{0, 1\}$).

Interpret $Y_{j,k}$, defined by:

$$Y_{j,k} = \#\{i : x_i \geqslant a_j \ \& \ g_i = k\}.$$

Denoting the individual that has the event at a_j by $\pi(j)$, write down in terms of $Y_{j,k}$ the probability $p(j, k)$ that an individual of group k had the event at a_j given the history of the process up to just before time a_j and assuming all individuals are subject to the same hazard $h_0(t)$.

Interpret z_j given by:

$$z_j = g_{\pi(j)} - p(j, 1)$$

and indicate how z given by

$$z = \sum_{j=1}^{d} z_j$$

can be used to test the hypothesis that the hazard function for group 1 is the same as that for group 0.

Suppose now that the hazard function for group k is $\exp(\beta k)h_0(t)$. What then is the probability that the individual with the event at a_j is from group k given the history of the process up to just before a_j? Find the expectation of z_j, conditional on the history up to just before a_j, as a function of β. Verify that the expectation when $\beta = 0$ is as you expect.

6 Analysis of Survival Data

What is meant by the *history*, $\mathcal{H}_{t-}$, of a time-to-event process up to but not including time t? What is meant by the *at-risk* function?

Write down an expression for the probability of the ith individual of a set of n having an event in the time interval $[t, t+\mathrm{d}t)$, conditional on $\mathcal{H}_{t-}$, in terms of the common hazard function $h(t)$ and the ith individual's at-risk function.

Let $\mathrm{d}N_+(t)$ be the total number of events in the interval $[t, t+\mathrm{d}t)$, with $\mathrm{d}N_+(t) \in \{0, 1\}$. What, conditional on $\mathcal{H}_{t-}$, is:

1. the expectation of $\mathrm{d}N_+(t)$?
2. the expectation of the square of $\mathrm{d}N_+(t)$?
3. the variance of $\mathrm{d}N_+(t)$?

Use your answer to (1) to obtain the *Nelson-Aalen* estimator $\hat{H}(t)$ for the integrated hazard. How does $\mathrm{d}\hat{H}(t)$ relate to $\mathrm{d}N_+(t)$?

From your answer to (3), obtain the variance of $\mathrm{d}\hat{H}(t)$, conditional on $\mathcal{H}_{t-}$, in terms of $\mathrm{d}H(t)$ and the at-risk functions. Substitute $\mathrm{d}\hat{H}(t)$ for $\mathrm{d}H(t)$ to obtain an estimated variance of $\mathrm{d}\hat{H}(t)$, conditional on $\mathcal{H}_{t-}$. Hence obtain an estimator for the variance of the Nelson-Aalen estimator of $H(t)$.

7 Analysis of Survival Data

Individuals are at risk of events of two distinct types: A and B; the two events have independent continuous survival distributions. What is meant by the *cause-specific hazard* for A?

An individual's cause-specific hazards for events A and B are given by $h_A(t) = \alpha$ and $h_B(t) = \beta$ respectively.

1. What is the probability that no event of either type has occurred before time t?
2. What is the probability that no event of either type has occurred before t and that an event of type A occurs at or later than t but before $t + \mathrm{d}t$?
3. What is the probability that an event occurs before or at time t and that event is of type A?
4. What is the probability that an event of type A occurs at some time?
5. Show that the density function for the time to an event of type A, given that an event of type A occurs, has an exponential distribution with rate parameter $\alpha + \beta$.

You are reviewing a research paper in which an author reports a time-to-death analysis. He states that out of 150 subjects he observed 100 deaths, the remaining 50 subjects being censored. He assumes an exponential distribution of time-to-death with common rate parameter λ and obtains a maximum likelihood estimate for λ of 0.3/year.

Unfortunately, he estimated the parameter λ using only the subjects who died as "the censored subjects provided no useful information".

6. What is wrong with his reasoning?

Assuming the time-to-censoring distribution is also exponential:

7. What is 0.3/year really an estimate of?
8. What proportion of the subjects died? How can you use this information?
9. Obtain a proper estimate of the cause-specific hazard for death.

END OF PAPER