UNIVERSITY OF CAMBRIDGE

MATHEMATICAL TRIPOS Part III

Friday, 7 June, 2013 1:30 pm to 4:30 pm

PAPER 32

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

A systematic review was carried out to investigate the effectiveness of granulocyte (white blood cell) transfusions for preventing mortality due to infection in patients with neutropenia or disorders of neutrophil function. The following data were extracted from ten randomised trials which compared granulocyte transfusions against a control treatment.

Trial name	Granulocyte transfusion (Deaths / Total)	Control (Deaths / Total)	Log risk ratio	Standard error of log risk ratio
Clift 1978	1/41	5/45	-1.52	1.07
Ford 1982	3/13	2/11	0.24	0.82
Gomez 1984	2/19	6/16	-1.27	0.74
Mannoni 1979	1/20	7/26	-1.68	1.03
Oza 2006	2/53	5/98	-0.30	0.82
Petersen 1987	21/87	11/47	0.03	0.33
Schiffer 1979	1/12	4/10	-1.57	1.03
Strauss 1981	12/54	6/48	0.58	0.46
Sutton 1982	9/29	14/38	-0.17	0.35
Winston 1980	13/19	13/19	0.00	0.22

(a) Write down numerical expressions for the log risk ratio and its variance for the Schiffer 1979 trial, comparing granulocyte transfusions to control. Calculate an approximate 95% confidence interval for the log risk ratio. Interpret the log risk 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.]

(b) Show how to calculate a pooled estimate and variance for the log risk ratio, based on a fixed effect meta-analysis. State the assumptions required for a fixed effect meta-analysis.

(c) Define "publication bias" and describe the likely impact of publication bias on the pooled results from a meta-analysis.

(d) Draw a funnel plot to explore whether publication bias is present in this meta-analysis data set. [The points can be plotted approximately, using just a pencil and paper.] A linear regression model is fitted with log risk ratios as outcome and their standard errors as covariate. The regression coefficient for the covariate is estimated as -1.99 (95% confidence interval -3.40 to -0.59; P=0.01). Interpret this result.

State whether you think there is any evidence of publication bias and justify your answer.

(e) Describe two statistical methods that systematic review authors could use when attempting to adjust for publication bias.

2 Statistics in Medical Practice

Supermarkets require processed meat suppliers to test a random sample of batches for the presence of horse DNA. Suppliers A, B and C each produce six batches per day (Monday to Friday) and propose to organise their random sampling as follows.

(a) With reference to tables of simple random numbers of or random permutations of length 16, advise on how to set up a randomisation scheme which matches each supplier's intention.

Supplier A proposes to rest at random one batch from his six batches daily.

Supplier B proposes to test at random three batches from 15 morning batches per week, and to test at random three batches from his 15 afternoon batches per week.

Supplier C proposes to select every batch on a randomly selected weekday day per working week.

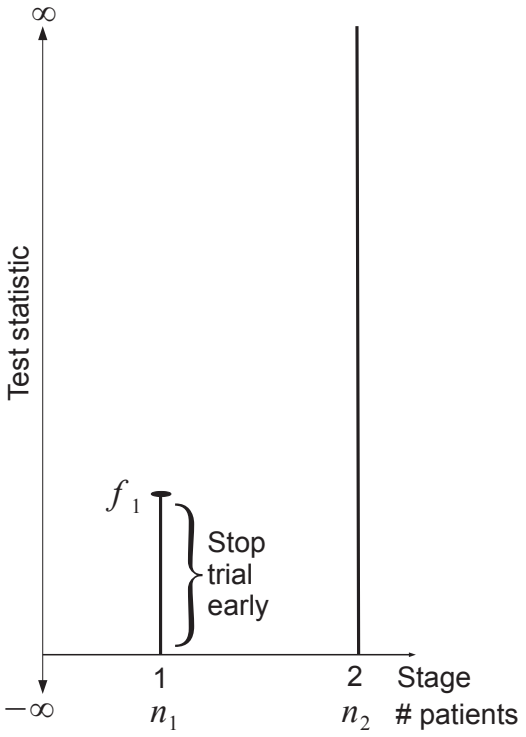
(b) Briefly, compare the merits and demerits of the random sampling schemes proposed by suppliers A, B, C to explain which scheme you prefer, and why.

(c) In the first two weeks of a supermarket's surveillance, 60 (2%) out of 3,000 randomly sampled batches of processed meat from suppliers B and C were discovered to contain horse-DNA. Advise how many randomly selected batches should be tested for horse DNA per supplier if a supermarket wishes to determine if the horse-contamination rate, which is expected to be around 2%, is twice as high in batches from supplier A versus supplier B.

(d) It is important to differentiate between traces of a veterinary medicine, VB, being found in 2 per 10,000 sampled batches from supplier B versus in 6 per 10,000 batches from supplier C. How many ten-thousands of batches need to be VB-checked to give 80% power that, if supplier B's claim to have a VB-rate that is one-third of supplier's C were true, then the difference would be apparent by the yardstick of statistical significance at the 5% level?

3 Statistics in Medical Practice

A randomised controlled trial is to be conducted to test the effectiveness of an experimental chemotherapy treatment for cancer. The trial will be sequential, only continuing to full enrollment if the new treatment looks sufficiently promising compared to standard care at the interim analysis. The trial protocol stipulates that n_1 patients will be randomised to arm 0 (standard care) and to arm 1 (treatment group) at stage 1. If needed, a further $n_2^* = n_2 - n_1$ patients will be randomised to each arm at stage 2. This gives a cumulative total of n_1 and n_2 patients per-arm at stages 1 and 2 respectively. An illustration of the trial is shown in Figure 1 below, further technical details follow.



Let the mean treatment effect for all patients in group i at stage j ($i = 0, 1$, $j = 1, 2$) be measured by an outcome Y_{ij}, which follows a $N(\theta_i, \frac{\sigma^2}{n_i})$ distribution. Outcomes Y_{0j} and Y_{1j} are independent for j=1,2. Assume that the θ_is are unknown but the common variance σ^2 is known. At the end of stage 1 the trialists wish to test the one-sided null hypothesis $H_0 : \delta \leqslant 0$, where $\delta = \theta_1 - \theta_0$.

(a) (i) Describe an advantage and a disadvantage of a group sequential trial compared to a traditional, fixed sample size trial.

(ii) Write down the maximum likelihood estimator (MLE) $\hat{\delta}_1$ for δ at stage 1. Show that $\hat{\delta}_1$ follows a $N(\delta, I_1^{-1})$ distribution, where I_1 is a quantity you should define.

(iii) The trial will continue to stage 2 if the stage 1 Wald statistic, $Z_1 = \hat{\delta}_1\sqrt{I_1}$, is greater than or equal to the futility threshold f_1. Using the cumulative

distribution function of Z_1, derive an expression for the expected sample size (per-arm) of the two-stage trial as a function of n_1, n_2^*, δ, I_1 and f_1.

(b) At the end of stage 1, the observed test statistic z_1 is above the futility threshold f_1, hence a further n_2^* patients are recruited to each arm and a test statistic z_2 is observed. The MLE for δ at the end of the trial is $\hat{\delta}_2$. This can be written as

$$\hat{\delta}_2 = \frac{n_1\hat{\delta}_1 + n_2^*\hat{\delta}_2^*}{n_2}$$

where $\hat{\delta}_2^*$ is the unbiased estimate for δ based on the n_2^* patients recruited at stage 2.

(i) Describe why, conditional on reaching full enrollment, the MLE $\hat{\delta}_2$ is a biased estimate for δ.

(ii) For a generic normal random variable $u \sim N(\mu, \tau^2)$ the following expressions are given:

$$E[u|u \geqslant T] = \frac{\int_T^\infty \frac{u}{\tau}\phi\left(\frac{u-\mu}{\tau}\right) du}{P(u \geqslant T)}$$

and

$$\int_{-\infty}^{T} \frac{u}{\tau}\phi\left(\frac{u-\mu}{\tau}\right) du = -\tau\phi\left(\frac{T-\mu}{\tau}\right) + \mu\Phi\left(\frac{T-\mu}{\tau}\right)$$

where $\phi(.)$ and $\Phi(.)$ are the standard normal probability density function and cumulative distribution function respectively. Furthermore, $\Phi(u) = 1-\Phi(-u)$ and $\phi(u) = \phi(-u)$.

Show that the expected value of $\hat{\delta}_1$, conditional on reaching full enrollment, is equal to $\delta + \frac{1}{\sqrt{I_1}}\frac{\phi(\delta\sqrt{I_1}-f_1)}{\Phi(\delta\sqrt{I_1}-f_1)}$.

(iii) Describe and illustrate either the procedure for finding the bias-corrected MLE, or the uniform minimum variance conditionally unbiased (UMVCUE), in order to provide a better estimate for the parameter δ.

4 Analysis of Survival Data

(a) A survival dataset comprises n individuals with survival times x_i and censoring indicators v_i, with $i = 1, \ldots, n$, where $v_i = 0$ when x_i corresponds to a right-censored observation and $v_i = 1$ when x_i corresponds to an observed event. Derive carefully the maximum likelihood estimator for the rate parameter θ of an exponential distribution fitted to that dataset.

(b) A piecewise-constant hazard model is being fitted to a survival dataset: the hazard experienced in the kth month (that is: the time period $k - 1 \leqslant t < k$) is θ_k.

There are 112 individuals in the dataset, 104 of whom were at risk at the start of the 2nd month.

Six of those 104 individuals had events in the 2nd month, at times 1.08, 1.16. 1.22, 1.50, 1.63 and 1.72 months respectively. One individual was censored in the 2nd month, at time 1.69 months. The remainder were at risk at the start of the 3rd month.

Calculate the maximum likelihood estimate for θ_2.

5 Analysis of Survival Data

Explain the principles of the standard log-rank test, describing carefully how the log-rank statistic is calculated. (You should indicate how the variance of the log-rank statistic is obtained but you need not give any formulae.)

A clinical time-to-event study randomised subjects to one of two treatments A and B.

(i) The time-to-event observations for individuals at risk after 160 days are:

Treatment A: 165, 173+

Treatment B: 180, 191

where the times are in days and '+' indicates a right-censored observation.

Calculate the contribution to the log-rank statistic from these event times. Interpret the contribution in terms of whether treatment A or treatment B shows the higher event rate.

(ii) Suppose that additional information becomes available about the individual who was censored at 173 days: it is now known that the individual did not have an event before 193 days. Re-calculate the contribution of events after 160 days to the log-rank statistic. Interpret the re-calculated contribution and explain why the change in censoring time has led to a change in interpretation.

6 Analysis of Survival Data

A time-to-event analysis of data from a recurrent headache study assumes that the hazard function for the next headache is $e^{\beta z}h_0(t)$ where $h_0(t)$ is a baseline hazard function, t is the number of days since treatment started, $z \in \{0, 1\}$ is a treatment indicator and β is the corresponding parameter. The objective of the time-to-event analysis is to estimate β parametrically and $h_0(t)$ non-parametrically.

(a) Describe how to set out the analysis dataset. Illustrate your answer by presenting the rows for patient number 001, whose treatment indicator z equalled zero and who had headaches at $t = 24.8$, 33.1, 40.2 and 51.9, with the study being closed at $t = 60.0$.

(b) Explain how you have ensured that patient 001 cannot have their second headache before their first headache.

(c) How would you modify your example dataset if it were known for this disease that headaches were followed by a recovery period: so that, if a patient had a headache at time t^*, the next headache could not be earlier than time $t^* + 3$? (You may assume that treatment started after at least a seven day period free of headaches.)

(d) How would you further modify your example dataset if the argument of the hazard function was, for second and subsequent headaches, the time since the start of the previous headache?

(e) How would you further modify your example dataset to permit the baseline hazard function for the second and subsequent headaches to be different to the baseline hazard function for the time to the first headache?

Explain why it is inappropriate to analyse the observations from the whole dataset as if they were independent. Briefly describe in general terms the principles of two methods for coping with the lack of independence.

7 Analysis of Survival Data

You are asked to analyse a large time-to-event dataset with many time-independent explanatory variables (both continuous and categorical). Write an essay on how you would construct, fit and assess the adequacy of a proportional hazards model with time-independent coefficients. Give careful explanations of:

(i) how you would choose which explanatory variables to include;

(ii) how you would check that you are using an appropriate functional form of the explanatory variables (for example: the model contains explanatory variable z but should it contain $\log z$ instead, or in addition?);

(iii) how you would check the assumption that the model coefficients are independent of time.

END OF PAPER