UNIVERSITY OF CAMBRIDGE

# MATHEMATICAL TRIPOS Part III

Thursday, 7 June, 2012 1:30 pm to 3:30 pm

## PAPER 41

## BIOSTATISTICS

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

.

*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

(a) A randomised controlled trial includes the following stages:

(i) Collect data on potentially eligible individuals and identify those truly eligible

(ii) Ask for consent

(iii) Collect baseline data

(iv) Randomise to experimental intervention or control intervention

Briefly describe the purpose of each of these stages.

(b) The data collected for the $i$th participant in a randomised controlled trial are $x_i$, a vector of baseline covariates; $z_i$, taking values 0/1 if randomised to control/experimental intervention respectively; $d_i$, a measure of the use made of the intervention; and $y_i$, a quantitative outcome measure.

Write down a statistical model that could be fitted to the data to estimate efficiently the effect of allocating a participant to the experimental intervention rather than to the control intervention. Your answer should include which parameter in this model is of primary interest, and any distributional assumption you make.

(c) Unfortunately, some of the values of $y_i$ were not observed. What assumption is made about the missing data if the model you gave in part b is fitted to participants with observed values of $y_i$?

(d) The trial investigators made repeated attempts to collect the outcome $y_i$, continuing until either $y_i$ was collected or four unsuccessful attempts had been made. You are now given data $a_i$ recording the number of *unsuccessful* attempts made for individual $i$, so that individuals who responded at the first attempt have $a_i = 0$, while individuals who did not respond at all have $a_i = 4$. It is proposed that the model you gave in part b be fitted jointly with the response model

$$\text{logit}\{P(a_i \geqslant j | a_i \geqslant j-1, x_i, z_i, d_i, y_i)\} = \eta_j + \theta y_i \ (j = 1, 2, 3, 4).$$

Write down expressions for the likelihood for

(i) an individual with $a_i = 0$;

(ii) an individual with $a_i = 4$.

(e) The trial recruited 300 participants. The outcome was observed for 211 participants with observed mean 8.24 and standard deviation 0.95. The joint model described in part d was fitted to these data, and the estimated parameters for the response model are given in the table.

| Parameter | Estimate | 95% Confidence Interval |
|---|---|---|
| $\eta_1$ | -4.19 | (-7.79 to -0.59) |
| $\eta_2$ | -4.20 | (-7.74 to -0.66) |
| $\eta_3$ | -4.02 | (-7.49 to -0.55) |
| $\eta_4$ | -4.48 | (-7.89 to -1.07) |
| $\theta$ | 0.38 | (-0.06 to 0.82) |

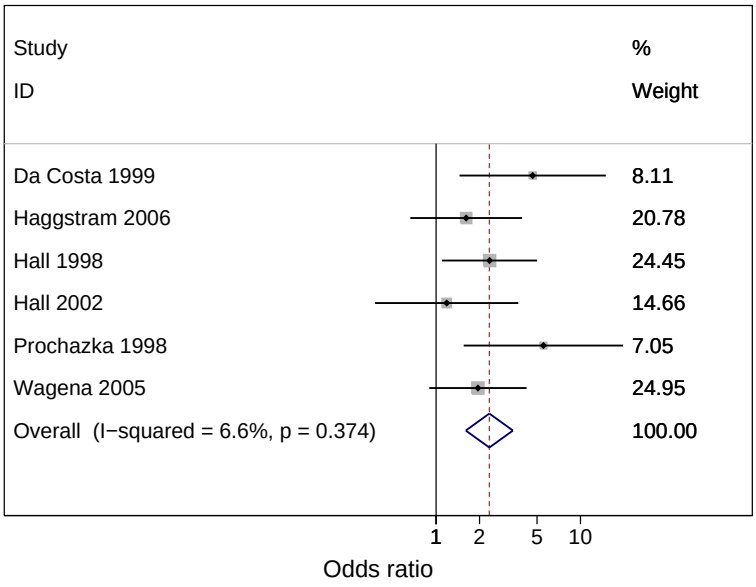
What can you conclude from these results about the truth of the assumption you stated in part c?

## 2 Statistics in Medical Practice

There are reasons to believe that antidepressant drugs might help people to give up smoking. Authors of a systematic review carried out a meta-analysis to assess the effect of the antidepressant Nortriptyline on long-term abstinence from smoking. Their primary conclusions were based on the following data from six randomised controlled trials.

| Study name | Nortriptyline (Number who stopped smoking / Total number randomised) | Placebo (Number who stopped smoking / Total number randomised) | Log Odds ratio | Standard error of log odds ratio |
|---|---|---|---|---|
| Da Costa 1999 | 14/68 | 4/76 | 1.54 | 0.59 |
| Haggstram 2006 | 16/52 | 11/51 | 0.48 | 0.45 |
| Hall 1998 | 24/99 | 12/100 | 0.85 | 0.39 |
| Hall 2002 | 7/73 | 6/73 | 0.17 | 0.58 |
| Prochazka 1998 | 15/108 | 3/106 | 1.71 | 0.65 |
| Wagena 2005 | 20/80 | 13/89 | 0.67 | 0.40 |

(a) Show how to calculate the log odds ratio and its variance for long-term abstinence from smoking, comparing Nortriptyline to placebo, using the data obtained from the Hall 1998 study. Assuming a standard Normal distribution for the log odds ratio statistic, calculate an approximate 95% confidence interval for the log odds ratio. [You may assume that the 97.5% quantile of the standard Normal distribution is approximately equal to 2.]




(b) The above plot depicts a fixed effect meta-analysis of the data in the table. Briefly describe the difference between a fixed effect meta-analysis and a random effects meta-analysis.

(c) Interpret the I-squared value and the p-value from a test for heterogeneity [the p-value is reported in the plot, next to the I-squared value]. Using these results and a

visual inspection of the studies' results to justify your answer, explain whether you believe the results to be heterogeneous across studies.

(d) The researchers obtained the results from a seventh randomised trial comparing Nortriptyline against placebo. Using the published data, they calculated a log odds ratio of 0.64 (corresponding to an odds ratio of 1.90) for long-term abstinence from smoking, with standard error equal to 0.57. If the results from this study were added to the fixed effect meta-analysis displayed in the plot, would you expect the overall conclusions from the meta-analysis to be altered? Explain your answer.

(e) When looking at their results, the authors notice that the three studies which detected a beneficial effect for Nortriptyline compared to placebo at the 5% significance level were carried out before 2000, whereas those showing no evidence of a beneficial effect were carried out after 2000. Which techniques might the authors use to explore this further? Give two reasons why findings from these additional analyses should be viewed with caution.

## 3 Analysis of Survival Data

Explain what is meant by the statements that two survival distributions (a) belong to the same *accelerated life* family and (b) belong to the same *proportional hazards* family.

A time-to-event random variable $T_z$ is related to another random variable $X$ by the relationship:

$$\log T_z = a + bz + cX$$

where $a$, $b$ and $c$ are constants and $z \in \{0, 1\}$ is an explanatory variable.

Let $U$ be a time-to-event random variable given by $U = \exp X$ and let $\mathcal{F}_U(u)$ be the survivor function for $U$. Find the survivor function for $T_z$ in terms of $\mathcal{F}_U$ and show that the distributions of $T_0$ and $T_1$ belong to the same accelerated life family.

Suppose further that $X$ has probability density function $f_X(x)$ given by

$$f_X(x) = \exp\{x - \exp(x)\}.$$

Show that the distributions of $T_0$ and $T_1$ now also below to the same proportional hazards family.

## 4 Analysis of Survival Data

Describe how to calculate the *log-rank* statistic and outline how it can be used to test the null hypothesis that two survival distributions are identical. [There is no need here to give an explicit expression for the variance of the log-rank statistic.]

A dentist wants to test a new method of filling teeth (Method B) against the standard method (Method A). She chooses 100 patients and for their next two fillings she uses Method A for one tooth and Method B for the other, allocated at random. She records the number of days from treatment until failure for each filling and performs a stratified log-rank test (with each patient in a separate stratum) of the null hypothesis that Method A and Method B are indistinguishable.

(a) Comment briefly on whether the log-rank test captures all the relevant evidence in the study.

(b) The first patient in the dataset has a time from treatment to failure of 7 days for Method A and of 730 days for Method B. What is the contribution of this patient to the log-rank statistic?

(c) Given that a patient makes a non-zero contribution to the log-rank statistic, what are the possible values of that contribution? Under the null hypothesis, and given that the patient makes a non-zero contribution, what is the mean and variance of that contribution?

(d) If $x_{iA}$, $x_{iB}$ are the $i$th patient's time to failure or censoring for Methods A, B respectively and $v_{iA}$, $v_{iB}$ are the corresponding visibility indicators ($v_{iz} = 1$ if $x_{iz}$ is a time to failure and $v_{iz} = 0$ if $x_{iz}$ is a censored observation, for $z \in \{\text{A,B}\}$) then the 100 patients can be divided into eight sets according to the values of $v_{iA}$, $v_{iB}$ and which of $x_{iA}$, $x_{iB}$ is smaller:

| Set | Description | Number of patients |
|---|---|---|
| 1 | $v_{iA} = 0, v_{iB} = 0, x_{iA} < x_{iB}$ | 19 |
| 2 | $v_{iA} = 0, v_{iB} = 0, x_{iA} > x_{iB}$ | 21 |
| 3 | $v_{iA} = 0, v_{iB} = 1, x_{iA} < x_{iB}$ | 13 |
| 4 | $v_{iA} = 0, v_{iB} = 1, x_{iA} > x_{iB}$ | 6 |
| 5 | $v_{iA} = 1, v_{iB} = 0, x_{iA} < x_{iB}$ | 18 |
| 6 | $v_{iA} = 1, v_{iB} = 0, x_{iA} > x_{iB}$ | 11 |
| 7 | $v_{iA} = 1, v_{iB} = 1, x_{iA} < x_{iB}$ | 9 |
| 8 | $v_{iA} = 1, v_{iB} = 1, x_{iA} > x_{iB}$ | 3 |

How many patients provide evidence that Method A is better and how many patients provide evidence that Method B is better? Calculate the log-rank statistic. Decide whether or not this study provides strong evidence that one or the other treatments is better. [You may find the following fact about the standard Normal distribution function useful: $\Phi(3) > 0.99865$.]

**5 Analysis of Survival Data**

(a) What is meant by a *survivor function*?

Show that the function $w(t)$ defined by:

$$w(t) = \frac{1}{4}\left\{\exp(-t) + 3\exp(-2t)\right\}$$

satisfies the necessary conditions to be a survivor function. Derive the corresponding hazard function and show that it is a decreasing function of $t$. Calculate the limit of the hazard function as $t \to \infty$, and show that your answer is consistent with interpreting $w(t)$ as the survivor function of a mixture of two survival distributions.

(b) Explain what is meant by a *frailty* model.

If the frailty random variable $U$ ($U \geqslant 0$) has density $g(u)$ and the integrated hazard conditional on $U = u$ is given by:

$$H(t|U = u) = uH_0(t)\,,$$

show that the unconditional survival function is given by $\tilde{g}\left(H_0(t)\right)$ where $\tilde{g}$ is the Laplace transform of $g$.

Obtain the unconditional survivor function in the case $H_0(t) = \frac{7}{4}t$, $g(u) = \frac{1}{4}\delta\left(u - \frac{4}{7}\right) + \frac{3}{4}\delta\left(u - \frac{8}{7}\right)$ where $\delta$ is the Dirac delta function. Explain why you obtain $w(t)$ as defined above.

**END OF PAPER**