UNIVERSITY OF CAMBRIDGE

## M. Phil. in STATISTICAL SCIENCE

Friday, 28 May, 2010   9:00 am to 11:00 am

# BIOSTATISTICS

*Attempt no more than **THREE** questions, with at most **TWO** questions from **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

On January 12th 2010, The Sun had a headline "Caesareans a 'risk' to mums". This study was conducted by the World Health Organisation and was published in a leading medical journal. The study featured 108,000 births from nine Asian countries. The conclusions were based on the following data about the risk to a mother of choosing to have a birth by caesarean without a medical reason, compared to mothers who had a spontaneous 'natural' birth.

| | Mothers who had a spontaneous birth without caesarean or operation (n = 75057) | Mothers who choose before birth to have a caesarean section without a medical indication (n = 1515) | Odds ratios (95% confidence intervals), adjusted for background factors |
|---|---|---|---|
| Maternal death | 53 (0.1%) | 0 | - |
| Maternal adverse event: ie death, admission to intensive care, blood transfusion, etc | 1215 (1.6%) | 9 (0.6%) | 2.7 (1.4 to 5.5) |

(a) Show how to calculate the odds ratio for a mother choosing a caesarean compared to those with a spontaneous birth, for admission for an adverse event based on the unadjusted data, and roughly calculate it. Based on the unadjusted data, does an elective caesarean section increase the risk to the mother [do not do a formal statistical test]? Does this agree with the published odds ratio adjusted for background factors?

(b) The authors report that the risk of maternal death for elective caesarean section "could not be estimated because there were no maternal deaths in this group". Do you think this is a reasonable statement? Would it be possible to calculate an unadjusted and adjusted odds ratio in this situation?

(c) A logistic regression was used to adjust for many background factors, including maternal age, education etc. How would the adjusted odds ratio and its 95% confidence interval be obtained from the logistic regression output?

(d) Is there an apparent contradiction between the unadjusted and adjusted analysis? What could genuinely explain this?

(e) If the elective caesarean group were similar in background factors to the spontaneous birth group, roughly how many adverse maternal events would you expect in the elective caesarean group?

(f) If the adjusted odds ratio is correct, about how many adverse maternal events would be expected if the 1515 elective caesarians had instead had a spontaneous birth? What, in terms of their background risk, would this say about the 1515 mothers who chose a caesarean section?

(g) Briefly, why might the spontaneous birth group not be an appropriate comparison with the elective caesarean group?

(h) The study concluded "The most important finding of the survey is the increased risk of maternal mortality and morbidity in women who undergo caesarean section with no medical indication". Briefly, do you think this conclusion is justified?

(i) What do you think might have gone wrong in the analysis? Note: the authors correctly allowed for clustering for women within clinics, but this meant a very complex program was used.

## 2 Statistics in Medical Practice

Mephedrone is a currently legal drug, attractive to users of ecstasy or cocaine, which may also find a market in young people who hitherto avoided illegal drugs. There is interest in how prevalent, and how dangerous, use of mephedrone is.

(a) The British Army conducts 500 Monday-compulsory-drug-tests (MCDTs) on privates each week. Up to 2010, around 1% of MCDTs were positive for cocaine. Advise the army on how many MCDTs to test additionally for mephedrone to demonstrate using a 5%-level test with 80% power that privates' cocaine positive rate in 2010 is twice their mephedrone positive rate. Explain any assumptions you make.

(b) In the recent past, there have been about 200 cocaine-related deaths per annum in the UK. Would comparison of UK's cocaine-related deaths in 2007+2008 versus 2009+2010 be sufficient to discern a 20% reduction that might be due to displacement of cocaine use in the later period? Explain any assumptions you make.

(c) In the recent past there have been about 30 ecstasy-only deaths per annum in the UK. If mephedrone use is 60% as prevalent as ecstasy use but only half as dangerous in terms of drugs-related deaths, how many mephedrone-only deaths should UK expect in 2010?

(d) Cocaine and ecstasy are both illegal class A drugs. Cocaine is twice as lethal as ecstasy per 100,000 users. If in 2010, increased use of mephedrone was associated with at 20% reduction in cocaine-related deaths but itself caused around 10 deaths, comment on the issues to consider in advising on mephedrone's legal status.

**3 Survival Data**

A survival dataset comprises $n$ individuals with survival times $x_j$ and censoring indicators $v_j$, with $j = 1, \ldots, n$, where $v_j = 0$ when $x_j$ corresponds to a right-censored observation and $v_j = 1$ when $x_j$ corresponds to an observed event. It is assumed that the survival distribution for the $j$th individual is exponential with rate parameter $\theta_j$.

(a) Derive the log-likelihood as a function of the $\theta_j$.

(b) The explanatory variable $g_j$ takes values in $\{0, 1\}$ and $0 < \sum_{j=1}^{n} g_j < n$. If $\theta_j = \beta^{(k)}$ when $g_j = k$ find the maximum likelihood estimates of $\beta^{(0)}$ and $\beta^{(1)}$ and obtain the likelihood-ratio test of the hypothesis $\beta^{(0)} = \beta^{(1)}$.

(c) It is now assumed that the hazard for the $j$th individual has form $m(t)\beta^{(k)}$ when $g_j = k$ and $m(t)$ is an unknown function which does not depend on $i$. Describe how you would test the hypothesis that $\beta^{(0)} = \beta^{(1)}$.

## 4 Survival Data

Derive the Nelson-Aalen estimator of the integrated hazard in the case where the integrated hazard does not depend on the individual and when there are no ties in the dataset.

(a) The hazard of the $i$th individual in a survival dataset is given by

$$h^i(t) \;=\; h_0^i(t) + h_1(t)$$

where $h_0^i(t)$ is known and $h_1(t)$ does not depend on $i$. Obtain an estimator for:

$$H_1(t) \;=\; \int_0^t h_1(u)\,\mathrm{d}u\,.$$

(b) It is required to test the hypothesis that the survival distribution of treated individuals is different from that of untreated individuals in the same experiment. At time $a_j$ $(1 \leqslant j \leqslant g)$ there are $r_j^A$ and $r_j^B$ individuals at risk in the treated (A) and untreated (B) groups respectively, with $d_j^A$ and $d_j^B$ individuals experiencing events at that time. No individuals experience events before $a_1$ or after $a_g$. For all $j$: $d_j^A + d_j^B = 1$.

Explain why

$$\sum_{j=1}^{g} \omega_j \left( \frac{d_j^A}{r_j^A} - \frac{d_j^B}{r_j^B} \right)$$

where $\omega_j > 0$ could be an appropriate form for a statistic testing that there is no difference between the survival distributions.

Show that the log-rank statistic can be written in this form with:

$$\omega_j \;=\; \frac{r_j^A\, r_j^B}{r_j^A + r_j^B}\,.$$

**5 Survival Data**

The hazard $h^i(t)$ of the $i$th individual belonging to a survival dataset is given by:

$$h^i(t) = h_0(t)\exp(\beta z^i)$$

where $h_0(t)$ is a baseline hazard, $z^i$ is the value of a constant scalar explanatory variable for the $i$th individual and $\beta$ is a parameter.

Conditional both on (i) the previous history of the process and on (ii) exactly one individual having an event at time $t^*$:

(a) find the probability that it is the $j$th individual who has the event at time $t^*$;

(b) find $\bar{z}(t^*, \beta)$, the expectation of the scalar explanatory variable for the individual having an event at $t^*$;

(c) interpret $s(t^*, \beta) = z^{\pi(t^*)} - \bar{z}(t^*, \beta)$, where $\pi(t^*)$ represents the individual that did have the event at $t^*$, and argue that the expectation of $s(t^*, \beta)$ is zero.

Let the distinct event times be $a_1, \ldots a_k, \ldots a_d$. Assuming no tied event times, show that:

$$\sum_{k=1}^{d} s(a_k, \hat{\beta}) = 0$$

where $\hat{\beta}$ is the proportional hazards estimate for $\beta$.

## END OF PAPER