UNIVERSITY OF CAMBRIDGE

M. Phil. in STATISTICAL SCIENCE

Tuesday, 2 June, 2009 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 A recently published paper by Dennis et al was entitled *"Effect of peer support on prevention of postnatal depression among high-risk women: multisite randomised trial"*. Details from the paper include:

- 701 women in the first two weeks after birth were identified as high risk for postnatal depression using the Edinburgh postnatal depression scale. They were randomised to either intervention or control with an internet-based randomisation service, stratified by self reported history of depression.
- The intervention was individualised telephone-based peer (mother to mother) support initiated within 48-72 hours of randomisation, provided by a volunteer recruited from the community who had previously experienced and recovered from self-reported postnatal depression and attended a four hour training session.
- Primary outcome measures were the Edinburgh postnatal depression scale (score > 12 considered as indicating 'depression'), and a structured clinical interview to diagnose depression. Secondary outcome measures were an anxiety measure, loneliness scale, and a measure of the use of health services. Their power calculations were based on the proportion with post-natal depression and suggested a sample size of 586 (293 in each of the intervention and control groups). They planned to enrol 700 to allow for losses to follow-up.
- A significance level of 0.05 was used for the primary outcome of postnatal depression and 0.01 for secondary outcomes. The authors used multiple logistic regression analysis to assess the effect of the intervention on postnatal depression at 12 weeks after controlling for baseline characteristics.
- After web-based screening of 21 470 women, 701 eligible mothers were recruited. A blinded research nurse followed up more than 85% by telephone, including 613 at 12 weeks. Out of the 349 women randomised to the intervention group, there was clear documentation of some form of initiation of the intervention in 328 (94%). At 12 weeks, 14% (40/297) of women in the intervention group and 25% (78/316) in the control group had an Edinburgh postnatal depression scale score $> 12 (X^2 = 12.5, P < 0.001)$. The logistic regression gave an odds ratio of 2.1 in favour of the intervention, 95% confidence interval 1.38 to 3.20.
- Only 37 (6%) women in the whole sample were identified with clinical depression using the interview at 12 weeks after birth - 14/297 (5%) in the intervention group and 23/315 (7%) in the control group. This prevalence is significantly lower than the overall 13% reported in a meta-analysis of 59 studies.

(a) How might the stratified randomisation have been conducted, and why?

(b) Define the quantities required in order to carry out their power calculation. [You do not need to provide the formula or calculations.]

(c) Why was a different significance level chosen for the primary and secondary outcomes?

(d) What were the research nurses blinded to, and why? Could there be problems with this blinding?

(e) Describe algebraically how the logistic regression may have been used and how an odds ratio of 2.1 would be obtained from it. What, approximately, is the corresponding raw odds-ratio based on the primary outcome alone? Why is it slightly surprising that they quote an odds ratio greater than 1?

(f) What percentage of women in the intervention arm may not have received the intervention, and why are they included in the analysis?

(g) In the paper the authors state that *"9 women would need to receive the peer support intervention to prevent one case of postnatal depression."* How did they obtain this estimate? Describe in words how would you put a confidence interval around this number using the raw percentage outcomes (you do not need to provide a formula or do calculations)?

(h) The abstract of the paper only discusses one of the primary outcome measures (the Edinburgh postnatal depression scale). Why do you think the other one (the clinical depression interview) does not appear in the abstract, and is this a reasonable omission?

2 Statistics in Medical Practice

(a) Consider the analysis of data which includes a vector-valued response variable of interest, Y, and a matrix of explanatory variable(s) or covariate(s), X. Assume that R is an indicator vector whose elements are coded 1 if the associated element of Y is observed and 0 otherwise. Further define the vectors Y^o and Y^mwhich correspond to the observed and missing values of the vector Y.

Explain in words what is meant by the following patterns of missing data, and give the corresponding specification for the distribution $f(R \mid Y^o, Y^m, X)$.

- Missing Completely at Random (MCAR)
- Covariate Dependent-MCAR (CD-MCAR)
- Covariate Dependent Missing at Random (CD-MAR)
- Missing Not at Random (MNAR)

(b) A survey of sexual lifestyles included a question, with a binary yes/no response, on virginity status. Survey respondents could be divided into three classes:

- Responders: provided answers to all questions
- Item non-responders: refused to answer the virginity question
- Unit non-responders: refused to answer any questions

In addition, the interviewers recorded for all responders and item non-responders an indication of whether the respondent appeared to be embarrassed at answering questions of a sexual nature.

The following table gives a summary of the available information from the survey.

	Responders	Item Non-responders	Unit Non-responders
Embarrassed	200	100	
Not Embarrassed	400	50	
Total	600	150	300

Of the responders who were judged to be embarrassed, 18% indicated that they were virgins whereas of the rest of the responders the percentage was 8%.

(i) Provide an estimate of the level of virginity in responders.

(ii) Carefully explaining any required assumptions, provide an estimate of the level of virginity in the combined sample of responders and item non-responders.

(iii) What additional assumptions would be needed to allow estimation of the level of virginity in the entire population of individuals approached in the survey? Provide an estimate of this, making it clear how all assumptions are used.

(iv) What is the aim of a sensitivity analysis which might be presented along with the estimate from part (iii)?

3 Survival Data

(a) Describe what is meant by (i) a *proportional hazards* family of distributions and (ii) an *accelerated life* family of distributions.

A survival variable with a Weibull(p, λ) distribution has density function

$$f(t; p, \lambda) = p\lambda^p t^{p-1} \exp[-(\lambda t)^p]$$

for $p > 0$ and $\lambda > 0$. Show that the Weibull(p_1, λ_1) and the Weibull(p_2, λ_2) distributions with $p_1 = p_2$ belong both to the same proportional hazards family and the same accelerated life family.

(b) Explain carefully how the partial likelihood is constructed for a proportional hazards model applied to data with no ties.

Describe, using an example, how the partial likelihood needs to be modified if the dataset contains ties. What diffculties occur if the number of ties is large? Explain, using an example, how these difficulties can be avoided.

4 Survival Data

(a) Outline the derivation of the Nelson-Aalen estimator of the integrated hazard for a survival dataset with no tied observations.

(b) Let $\hat{H}_j$ be the Nelson-Aalen estimator of the integrated hazard calculated at the jth event or censoring time and let m be the number of distinct event or censoring times in the dataset. Continuing to assume no ties, show that $\sum_{j=1}^{m} \hat{H}_j$ is equal to the number of individuals with an observed event.

(c) Describe two methods of handling tied event times. For which of these methods does the result in (b) continue to hold?

5 Survival Data

Suppose the hazard function $h^{(i)}(t)$ for the ith individual in a survival dataset can be written as

$$h^{(i)}(t) = h_B^{(i)}(t) + h_E(t)$$

where $h_B^{(i)}(t)$ is known, small and depends on i and $h_E(t)$ is unknown, not necessarily small and does not depend on i.

(a) Describe a context in medical statistics where such models are often used.

(b) Define and interpret the *relative survivor function* $F_E(t)$. Describe how to obtain an estimate $\hat{F}_E(t)$ of $F_E(t)$ paying particular attention to the contribution of the $h_B^{(i)}(t)$.

(c) t_j and t_k are times of consecutive events ($t_j < t_k$) in a relative survival dataset. Describe the behaviour of $\hat{F}_E(t)$ for $t_j < t < t_k$. Describe an example of a situation where $\hat{F}_E(t_K)$ might be greater than $\hat{F}_E(t_j)$.

END OF PAPER