Imperial College London

Department of Mathematics

Course:	M3S12
Setter:	Walden
Checker:	Heard
Editor:	McCoy
External:	Kent
Date:	June 6, 2005

BSc and MSci EXAMINATIONS (MATHEMATICS) MAY–JUNE 2005

This paper is also taken for the relevant examination for the Associateship.

M3S12 Biostatistics

DATE: Someday, May 2005 TIME: 10 am – 12 noon

Credit will be given for all questions attempted but extra credit will be given for complete or nearly complete answers.

Calculators may not be used. Statistical tables will be available.

Setter's signature ....

Checker's signature ....

1.

(a) The mortality rate in a population A is to be assessed with respect to some standardizing population S.

(i) Explain why standardization is important when comparing mortality rates between different geographical areas.

(ii) After carefully defining your notation, write down the expression for the directly standardized rate ratio $SRR(A; S)$ with the numerator and denominator expressed as weighted sums. Do likewise for the indirectly standardized mortality ratio $SMR(A; S)$. Explain the forms of the weights in both cases.

(iii) If the subpopulation death rates $\{R_{Ak}\}$for population A are unknown, what single piece of information about population A will enable $SMR(A; S)$ to be calculated? [Assume full information on population S.]

(b) The table below shows the age distribution of the population of Southampton during the 1991 Census and the age-specific death rates for England and Wales during the same year.

Age	Southampton population (1991)	Death rates England and Wales (per 1000)
0-4	25000	0.8
5-14	40000	0.4
15-24	55000	0.9
25-34	50000	1.0
35-44	42000	2.3
45-54	27000	7.1
55-64	17000	20
65-74	10000	52
75-84	5000	120
85+	1000	240

The observed number of deaths in Southampton between 1990 and 1992 inclusive was 6900.

(i) Calculate the standardized mortality ratio (SMR) for these data and interpret your result.

(ii) Give two possible explanations for your result.

2.

(a) Describe the main features of (i) a cohort study and (ii) a case-control study. Give an advantage and disadvantage of each type of study.

Let events E, F and S denote exposure to a risk factor, incidence outcome, and inclusion in the study, respectively.

(b) For a prospective case-control study, show, with full justification, that

$$\frac{\mathbf{P}(E \cap F \cap S)}{\mathbf{P}(E' \cap F \cap S)} = \frac{\mathbf{P}(E|F)}{\mathbf{P}(E'|F)}.$$

(c) Let R be a binary *potential* risk factor, and consider the following pair of 2×2 tables from a cohort study.

$R = 0$

	E	E'
F	90	60
F'	10	40

$R = 1$

	E	E'
F	80	20
F'	120	180

(i) Given $\pi_1 = P(F|E)$ and $\pi_0 = P(F|E')$ use the ratios $\pi_1^{(R=0)}/\pi_1^{(R=1)}$ and $\pi_0^{(R=0)}/\pi_0^{(R=1)}$ to decide whether R is a risk factor for the disease, and if so, in what sense. Fully explain your reasoning

(ii) By pooling the tables appropriately determine whether R and E are related.

(iii) Calculate an approximate 95% confidence interval for $\log \widehat{\psi}$ from this pooled table, where $\widehat{\psi}$ is the maximum likelihood estimator for the odds ratio.

3.

(a) Describe what is meant by a balanced 2-factor design.

(b) An investigation into the variability of survival times for subjects suffering from poisoning was carried out. There were three categories of poisoning and four treatments for poisoning, with m observations per cross-classification.

The following partially completed two-way ANOVA table was obtained.

	D.F.	Sum of squares	Mean square	F
Poison	*	*	50	*
Treatment	*	92	30.66	*
Interaction	6	25	4.16	*
Residual	*	72	*	
Total	47	*		

(i) Find the common number of replicates m for each cross-classification. What would be the consequences for such a study if there was only one replicate for each cross-classification?

(ii) Complete the table entries.

(iii) Suppose the following model of an observation is adopted,

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk},$$

where μ is an overall mean, α_i is the effect of the ith level of poison, β_j is the effect of the jth level of treatment, $(\alpha\beta)_{ij}$ are the interactions, $k = 1, \dots, m$, and $\epsilon_{ijk} \sim N(0, \sigma^2)$.

What constraints must the parameters satisfy, and why are they necessary?

(iv) Using a significance level of $\alpha = 0.05$ for each test, derive the form of model suggested by the ANOVA table.

The overall level of significance α for a family of n tests with individual significance levels $\alpha_1, \dots, \alpha_n$ satisfies the Bonferroni condition $\alpha \leq \sum_{i=1}^{n} \alpha_i$. If your tests are repeated at the largest common significance level which ensures $\alpha \leq 0.05$, does your chosen model change?

4.

A discrete random variable Y with single parameter λ belongs to the exponential-dispersion family in canonical form if its probability mass function can be written in the form

$$f_{Y|\theta,\phi}(y;\theta,\phi) = \exp\left\{\frac{\theta b(y) + c(\theta)}{r(\phi)} + d(y,\phi)\right\}.$$

(a) Use the properties of the score function to show that for a member of the exponential-dispersion family for which $b(y) = y$,

$$E_{f_{Y|\theta,\phi}}[Y] = -c'(\theta) \qquad \text{and} \qquad \text{var}_{f_{Y|\theta,\phi}}[Y] = -c''(\theta)r(\phi).$$

(b) Suppose the random variable Y has a Poisson distribution with probability mass function

$$f_{Y|\lambda,\phi}(y;\lambda,\phi) = \frac{e^{-\lambda}\lambda^y}{y!}.$$

(i) Show that Y belongs to the exponential-dispersion family in canonical form, identify the canonical link function for the parameter λ, and the forms of $b(y), c(\theta), d(y,\phi)$ and $r(\phi)$.

(ii) Assuming a generalized linear model, show that for a single predictor X with values $x_i, i = 1,\ldots,n$, for n observations $y_1,\ldots,y_n$, the canonical link function gives

$$\lambda_i = \exp\{\beta_0 + \beta_1 x_i\}$$

where β_0 and β_1 are the usual intercept and slope parameters.

(iii) Describe what is meant by a *saturated model* in the context of generalized linear modelling.

(iv) If $\widehat{\beta}_M$ represents the maximum likelihood parameter estimates under the model in (ii), and $\widehat{\beta}_S$ likewise for the saturated model, and l_M and l_S denote the corresponding likelihood functions, show that the deviance takes the form

$$D = -2\sum_{i=1}^{n}\left[y_i\left(\widehat{\beta}_0 + \widehat{\beta}_1 x_i - \log y_i\right)\right].$$

(v) The number of deaths from AIDS in Australia were recorded for 14 consecutive three-month periods, and were modelled as observations of independent Poisson random variables. The following model was fitted to the data, $\lambda_i = \exp\{\beta_0 + \beta_1 x_i\}$, where $x_i = \log i$ and $i = 1,\ldots,14$. The deviance of the form in (iv) was found to be 17.09. Comment on the fit of the model.

5.

(a) Consider regression for a binary response data set.

(i) Define the three standard link functions, along with the identity link.

(ii) Give two properties which apply to one or more of these functions, stating which one(s) you are referring to.

(b) To examine the effect of smoking upon the onset of menopause, the numbers of women who had, and had not, reached menopause were recorded for a cohort of women aged 45-54, divided into 5 two-year age groups and into smokers and non-smokers. A model proposed for the number of women having reached menopause is that it follows a Binomial distribution with parameter p such that

$$\log(p/[1-p]) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2,$$

with x_1 denoting age group with levels $1, \ldots, 5$, and $x_2 = 0$ for non-smokers and $x_2 = 1$ for smokers.

(i) When this model was fitted to the data, the following results were obtained, $D = 2.29, \widehat{\beta}_0 = -4.55, \widehat{\beta}_1 = 1.10, \widehat{\beta}_2 = 0.23, \widehat{\beta}_3 = 0.14$, where D denotes deviance. Assess the quality of the model fit.

(ii) Under the constraint $\beta_3 = 0$ the fit produced the results $D = 2.64, \widehat{\beta}_0 = -4.75, \widehat{\beta}_1 = 1.15, \widehat{\beta}_2 = 0.71$, and when smoking was also dropped from the model, the results were $D = 11.03, \widehat{\beta}_0 = -4.39, \widehat{\beta}_1 = 1.12$. Use the analysis of deviance to decide on a preferred model for this data.

(iii) Suggest the rationale for not considering setting β_1 to zero.

(iv) For women in age group 4, use your preferred model to estimate the odds ratio ψ, defined in the usual way.

[To calculate the estimate you can make use of: $e^{0.35} \approx \sqrt{2}$.]