Imperial College
London

UNIVERSITY OF LONDON

BSc and MSci EXAMINATIONS (MATHEMATICS)
MAY–JUNE 2004

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

M3S12 BIOSTATISTICS

Date: Tuesday, 8th June 2004 Time: 2 pm – 4 pm

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.

SPLUS output, including limited statistical tables, is attached.

1. **(a)** The mortality rate in population A is to be assessed with respect to some standardizing population, S. Each population is stratified into $K+1$ age-specific strata, and the numbers of individuals and the numbers of deaths in each stratum are known for each population.

Explain how the mortality rate for population A may be reported after standardization via population S, giving specific details of the following concepts

(i) the *crude mortality rate*

(ii) the *age-specific mortality rates*

(iii) the *directly standardized mortality rate*

(iv) the *indirectly standardized mortality rate*

(v) the *standardized mortality ratio (SMR)*

(b) The following table contains information on the structure and mortality rates stratified by age in a clinically interesting population A (those diagnosed with a particular psychiatric condition) and in the general population (derived from census data) in a city in North America.

	Pop^n A		City Pop^n	
AGE GROUP	No. Deaths	Stratum size	No. Deaths	Stratum size
0-20 years	5	500	240	40000
21-40 years	25	1000	300	40000
41-60 years	20	400	160	32000
61-80 years	10	100	100	16000
TOTAL	60	2000	800	128000

Compute

(i) the directly standardized mortality rate

(ii) the indirectly standardized mortality rate

for population A, using the general city population as the standardizing population. Leave your answers as fractions with denominator 3200.

Comment on the relationship between the psychiatric condition and mortality.

2. **(a)** **(i)** By using a suitable tree diagram and probability notation, outline the structure of a typical observational study designed to discover the relationship between an exposure risk factor and a given clinical outcome. Identify three *measures of effect* of interest.

(ii) Describe the key feature that distinguishes a *case-control* study from a *cohort* study, giving your explanation in terms of assumptions made about certain conditional probabilities.

(iii) In a conventional cohort study (two exposure levels, two outcome categories), the observed data may be summarized in a 2×2 table with entries $(n_{11}, n_{12}, n_{21}, n_{22})$. Show that, in such a study, the estimated standard error of the *log relative risk* (log RR) of outcome 1 in the two exposure categories is approximately

$$\sqrt{\left(\frac{1}{n_{11}}-\frac{1}{n_{11}+n_{21}}\right)+\left(\frac{1}{n_{12}}-\frac{1}{n_{12}+n_{22}}\right)}.$$

Use the following approximation: if

$$U_n \overset{A}{\sim} N\left(\mu, \frac{\sigma^2}{n}\right),$$

g *is some differentiable function with derivative* g', *and* $V_n = g(U_n)$ *then*

$$V_n \overset{A}{\sim} N\left(g(\mu), \frac{\sigma^2\{g'(\mu)\}^2}{n}\right),$$

where $g'(\mu) \neq 0$.

(b) The following tables represent the results of a case-control study into the relationship between a binary exposure factor (smoking in the home) and the incidence of asthma in children from households located in urban or rural locations.

	URBAN		RURAL	
	SMOKE	NON-SMOKE	SMOKE	NON-SMOKE
ASTHMA	5	20	20	100
NO ASTHMA	600	1100	150	1400

On the basis of these data, assess whether there is any evidence of increased risk of asthma due to smoking in the home, and whether there is any confounding or effect modification due to the household location.

3. **(a)** **(i)** Describe how the Analysis of Variance (ANOVA) approach is used to explore the relationship between a continuously varying response variable and a number of factor predictors. In particular, give details of the mathematical modelling assumptions, the sum of squares decomposition, key distributional results, and hypothesis testing.

(ii) Write down the statistical model behind a Two-Way Analysis of Variance with Interaction, where the two factor predictors have K and L levels, for data in a full-factorial design, with an equal number of replicates, m, in each cross-category. Give your answer in the form

$$y_{klj} = \text{MEAN FUNCTION} + \text{RANDOM ERROR}$$

where k and l index the levels of the two factors respectively, and j indexes replicate number.

(iii) Give the sum-of-squares decomposition for this model, explaining the role of each of the four contributions to the total sum of squares.

(b) An investigation into the variability of blood pressure for a cohort of individuals suffering from hypertension was carried out. Each member of the cohort was age-categorized into one of a number of age categories (denoted AGECAT), and then given one of a number of drug treatments (denoted DRUG), one of which was a control group. An equal number of individuals, m, were recruited for each cross-classification.

The SPLUS output below gives details of a Two-Way Analysis of Variance in the form of an ANOVA table. Three entries have been omitted.

	Df	Sum of Sq	Mean Sq	F Value	Pr(F)
DRUG	3	538.74	179.581	10.6162	0.00001810
AGECAT	5	10888.18	2177.637	128.7345	0.00000000
DRUG:AGECAT	**	******	37.580	2.2216	**********
Residuals	48	811.95	16.916		
Total	71	12802.57			

(i) Identify the number of levels of drug treatment (including the control), and the number of age categories.

(ii) Find the number of replicates, m.

(iii) Find the three omitted values, using the $Fisher-F$ distribution tables on page 7.

(iv) Summarize the results of the analysis.

4. **(a)** This question concerns a logistic regression for a binary response data set with a single continuous predictor variable. Denote the observed response and predictor pairs $\{(y_i, x_i), i = 1, ..., n\}$, where the assumed model for the response variable has

$$Y_i \sim Bernoulli\,(\theta_{0i})$$

for naive parameter θ_{0i}.

(i) Write this probability model in *exponential-dispersion family* form.

(ii) Find the *canonical link* function for this GLM.

(iii) Assuming a linear predictor of the form

$$\boldsymbol{x_i}^T\boldsymbol{\beta} = \beta_0 + \beta_1 x_i$$

and the canonical link function, derive the two *score equations* used for the maximum likelihood estimation of β_0 and β_1.

(b) The SPLUS output on page 8 summarizes the analysis of deviance results of a logistic regression analysis of data from an *in vitro* fertilization (IVF) study. In the study, 208 mothers underwent a cycle of IVF, and the outcome (successful/unsuccessful pregancy, $Y = 1/0$) was recorded. The aim of the study was to assess whether the level of a hormone, progesterone, measured in the mother three days before the beginning of the cycle, could be used to predict whether the IVF cycle will be successful. It was also believed that the maternal age was potentially an important predictor of the success of the IVF cycle; this variable was recorded, and discretized into five age categories.

In the SPLUS output, the progesterone level is denoted X and maternal age category is denoted MATAGE.

(i) Summarize the results of the analysis of deviance, and report the most appropriate model.

(ii) Comment on the adequacy of the selected model

(iii) By considering **(a) (i), (ii)** and **(iii)**, briefly describe three variations of the model that may lead to more appropriate inference about the dependence of response on the predictors.

5. **(a)** A two-way, $I \times J$ contingency table of count data is assumed to have entries n_{ij}, that are realizations of independent random variables N_{ij} that are Poisson distributed, that is

$$N_{ij} \sim Poisson\,(\lambda_{ij}) \qquad i = 1, ..., I \text{ and } j = 1, ..., J.$$

Find the likelihood for the entries in the table **conditional** on the row totals $(N_{1.}, N_{2.}, ..., N_{I.})$ taking their observed values $(n_{1.}, n_{2.}, ..., n_{I.})$ where

$$n_{i.} = \sum_{j=1}^{J} n_{ij}$$

(b) The following data are from a small cohort study carried out to discover the relationship between the exposure to radiation of the fathers of children employed at a nuclear power plant, and incidences of leukemia, collated over a ten year period

	EXPOSED	NON-EXPOSED
LEUKEMIA CASE	12	8
HEALTHY	24	100

Is there any statistical evidence of association or dependence between the row and column factors ? Justify your answer, commenting on the validity of the chosen statistical test.

(Note the chi-squared distribution quantiles given at the bottom of page 9)

Explain briefly how and why conditioning might be used in the analysis of the data in this table.

Question 3: $Fisher - F$ tables

The table below contains probabilities in the $Fisher - F$ cumulative distribution function with df_1 and 48 degrees of freedom, where df_1 is the first entry in each row. Specifically, the entries in the table are

$$p = P[X \leq x]$$

when

$$X \sim Fisher(df_1, 48)$$

and x is given by the column heading. For example, if $df_1 = 7$ and $x = 2.60$, then

$$p = 0.977$$

df1	2.00	2.10	2.20	2.30	2.40	2.50	2.60	2.70	2.80	2.90	3.00
1	0.836	0.846	0.855	0.864	0.872	0.880	0.887	0.893	0.899	0.905	0.910
2	0.854	0.866	0.878	0.889	0.898	0.907	0.915	0.923	0.929	0.935	0.941
3	0.873	0.887	0.900	0.911	0.921	0.929	0.937	0.944	0.950	0.956	0.960
4	0.890	0.905	0.917	0.928	0.937	0.945	0.952	0.959	0.964	0.969	0.973
5	0.904	0.918	0.930	0.941	0.949	0.957	0.963	0.969	0.973	0.977	0.980
6	0.916	0.929	0.941	0.950	0.959	0.965	0.971	0.976	0.980	0.983	0.986
7	0.925	0.938	0.949	0.958	0.966	0.972	0.977	0.981	0.984	0.987	0.989
8	0.933	0.946	0.956	0.964	0.971	0.977	0.981	0.985	0.987	0.990	0.992
9	0.940	0.952	0.962	0.969	0.975	0.980	0.984	0.987	0.990	0.992	0.994
10	0.946	0.957	0.966	0.973	0.979	0.983	0.987	0.990	0.992	0.994	0.995
11	0.951	0.961	0.970	0.977	0.982	0.986	0.989	0.991	0.993	0.995	0.996
12	0.955	0.965	0.973	0.979	0.984	0.988	0.991	0.993	0.994	0.996	0.997
13	0.959	0.968	0.976	0.982	0.986	0.989	0.992	0.994	0.995	0.996	0.997
14	0.962	0.971	0.978	0.983	0.988	0.991	0.993	0.995	0.996	0.997	0.998
15	0.965	0.973	0.980	0.985	0.989	0.992	0.994	0.995	0.997	0.997	0.998
16	0.967	0.976	0.982	0.987	0.990	0.993	0.995	0.996	0.997	0.998	0.998
17	0.969	0.977	0.983	0.988	0.991	0.993	0.995	0.996	0.997	0.998	0.999
18	0.971	0.979	0.985	0.989	0.992	0.994	0.996	0.997	0.998	0.998	0.999
19	0.973	0.980	0.986	0.990	0.993	0.995	0.996	0.997	0.998	0.999	0.999
20	0.975	0.982	0.987	0.991	0.993	0.995	0.997	0.997	0.998	0.999	0.999

Question 4: SPLUS Output Page 1

```
>options(contrasts=c(``contr.treatment'',``contr.poly''))
>summary(glm(Y~1,data=ivf.data,family=binomial))
Coefficients:
                Value Std. Error  t value
(Intercept) 0.3695331  0.1410489 2.619894

    Null Deviance: 281.3678 on 207 degrees of freedom
Residual Deviance: 281.3678 on 207 degrees of freedom

>summary(glm(Y~X,data=ivf.data,family=binomial))

Coefficients:
                Value    Std. Error  t value
(Intercept) -0.81382457 0.33957134 -2.396623
          X  0.05298365 0.01450141  3.653690

    Null Deviance: 281.3678 on 207 degrees of freedom
Residual Deviance: 264.4305 on 206 degrees of freedom

> summary(glm(Y~MATAGE,data=ivf.data,family=binomial))

Coefficients:
               Value   Std. Error   t value
(Intercept)  0.7777046  0.2929894  2.6543779
    MATAGE2  0.1026541  0.4111418  0.2496805
    MATAGE3 -0.8954876  0.4515510 -1.9831371
    MATAGE4 -0.8730147  0.4257902 -2.0503400
    MATAGE5 -0.9783753  0.5365286 -1.8235285

    Null Deviance: 281.3678 on 207 degrees of freedom
Residual Deviance: 270.1133 on 203 degrees of freedom
```

Question 4: SPLUS Output Page 2

```
>summary(glm(Y~X+MATAGE,data=ivf.data,family=binomial))

Coefficients:
                Value   Std. Error   t value
(Intercept) -0.65804253 0.43965622 -1.4967206
          X  0.07128228 0.01708112  4.1731628
    MATAGE2  0.16393160 0.42926749  0.3818868
    MATAGE3 -1.53313554 0.50938469 -3.0097794
    MATAGE4 -1.17467879 0.45816854 -2.5638574
    MATAGE5 -0.89797446 0.55375000 -1.6216243

    Null Deviance: 281.3678 on 207 degrees of freedom
Residual Deviance: 245.6008 on 202 degrees of freedom

> summary(glm(Y~X*MATAGE,data=ivf.data,family=binomial))

Coefficients:
                    Value Std. Error   t value
(Intercept) -0.191771684 0.72770966 -0.26352774
          X  0.046775026 0.03398600  1.37630282
    MATAGE2 -0.476663278 1.01051003 -0.47170564
    MATAGE3 -1.513788351 1.15459445 -1.31109963
    MATAGE4 -3.388798987 1.39742806 -2.42502572
    MATAGE5  0.021827867 1.49395337  0.01461081
  X:MATAGE2  0.034672757 0.05034709  0.68867449
  X:MATAGE3  0.007105887 0.04458499  0.15937845
  X:MATAGE4  0.101334251 0.06099963  1.66122738
  X:MATAGE5 -0.048402837 0.07326309 -0.66067149

    Null Deviance: 281.3678 on 207 degrees of freedom
Residual Deviance: 240.5079 on 198 degrees of freedom

> round(qchisq(0.95, df = c(1, 2, 3, 4, 5, 6, 7, 8)), 4)
[1] 3.8415 5.9915  7.8147  9.4877 11.0705 12.5916 14.0671 15.5073
```