REPORT OF THE BOARD OF EXAMINERS

September 2003

Subject 101 — Statistical Modelling

EXAMINERS' REPORT

Introduction

The attached subject report has been written by the Principal Examiner with the aim of helping candidates. The questions and comments are based around Core Reading as the interpretation of the syllabus to which the examiners are working. They have however given credit for any alternative approach or interpretation which they consider to be reasonable.

J Curtis
Chairman of the Board of Examiners

11 November 2003

EXAMINATIONS

September 2003

Subject 101 — Statistical Modelling

EXAMINERS' REPORT

General comments

The Examiners are of the view that, overall, the paper was of a comparable standard to those set in recent diets. However they do recognise that most candidates found the last two questions in the paper rather demanding.

1 Let X be the number remaining in the scheme for at least 10 years.

$X \sim$ binomial(50, 0.4)

So, approximately $X \sim N(20, 12)$

We require $P(X > 25)$.

Using a continuity correction,

$$P(X > 25.5) \doteqdot P(Z > \frac{25.5 - 20}{\sqrt{12}}) = P(Z > 1.59) = 1 - 0.944 = 0.056$$

2 $$P\left(\bar{X} > S\right) = P\left(\frac{3\bar{X}}{S} > 3\right) = P\left(t_8 > 3\right)$$

which is between 0.005 and 0.01.

3 $\pm 1.96\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{200}}$ where $\hat{\theta} = 0.16$

$$\Rightarrow \pm 1.96\sqrt{\frac{(0.16)(0.84)}{200}} \Rightarrow \pm 1.96(0.026) \Rightarrow \pm 0.051$$

or, as an interval: $0.16 \pm 0.051 \Rightarrow (0.109, 0.211)$

4 $n = 16,\ \sum_{i=1}^{16} x_i = 51.2,\ \sum_{i=1}^{16} x_i^2 = 243.19;$

$$\bar{x} = 3.20,\ s^2 = \frac{243.19 - \frac{51.2^2}{16}}{15} = \frac{79.35}{15} = 5.29.$$

$s = \sqrt{5.29} = 2.3$

The 95% confidence interval is given by:

$$\bar{x} \pm t_{0.025,n-1}\frac{s}{\sqrt{n}} = 3.2 \pm 2.131\frac{2.3}{4} = 3.2 \pm 1.23$$

i.e. (1.97, 4.43)

5 P(at least one test is significant | each null hypothesis is true)

= 1 – P(no test is significant | each null hypothesis is true)

= $1 - (1 - 0.05)^{10}$ as the 10 tests are independent

= 0.4

Comment: a false-positive is very likely with the 10 multiple tests.

6 $S_{xx} = 5 - 3^2/3 = 2$, $S_{xy} = y + 4 - 3(y+2)/3 = 2$

So fitted slope = 2/2 = 1

7 N is Poisson($k\lambda$) with $M_N(t) = \exp[k\lambda\{\exp(t) - 1\}]$.

S has a compound distribution with mgf $M_S(t) = M_N\{\log M_X(t)\}$

and $M_X(t) = \exp(\mu t + \sigma^2 t^2/2)$.

So mgf of S is $M_N(\mu t + \sigma^2 t^2/2)$ and correct suggestion is A.

OR: by using the result quoted in the Formulae and Tables book

OR: we must have $M_S(0) = 1$, so B is wrong.

8 Let X be the number of forms with incomplete information in a batch of n forms.

Then $X \sim$ Poisson($0.02n$) approximately

With $n = 516$, $X \sim P(10.32)$ and $P(X \leq 16) = 0.965$ approx, by linear interpolation

With $n = 515$, $X \sim P(10.3)$ and $P(X \leq 15) = 0.940$ approx, by linear interpolation

So he requires 516 forms

OR: the distribution of X can be modelled approximately using a normal distribution with mean $0.02n$ and variance $0.0196n$; we require $P(X \leq n - 500)$ to be at least 0.95; the analysis is more awkward, but solving a quadratic in n gives $n \geq 515$

9 (i)

		Y			
		2	*4*	*6*	
	1	0.2	0.0	0.2	0.4
X	*2*	0.0	0.2	0.0	0.2
	3	0.2	0.0	0.2	0.4
		0.4	0.2	0.4	

$E[X] = 0.4 \times 1 + 0.2 \times 2 + 0.4 \times 3 = 2$

$E[Y] = 0.4 \times 2 + 0.2 \times 4 + 0.4 \times 6 = 4$

$E[XY] = 1 \times 2 \times 0.2 + 1 \times 4 \times 0.0 + 1 \times 6 \times 0.2$
$+ 2 \times 2 \times 0.0 + 2 \times 4 \times 0.2 + 2 \times 6 \times 0.0$
$+ 3 \times 2 \times 0.2 + 3 \times 4 \times 0.0 + 3 \times 6 \times 0.2 = 8$

$E[XY] - E[X]E[Y] = 0$ Therefore uncorrelated.

X and Y are not independent since

$P(X = x \text{ and } Y = y) \neq P(X = x)\, P(Y = y)$

e.g. $x = 1$ $y = 2$, $0.2 \neq 0.4 \times 0.4 = 0.16$.

(ii) X and Y are independent if joint probability is:

		Y			
		2	*4*	*6*	
	1	0.2	0.0	0.2	0.4
X	*2*	0.1	0.0	0.1	0.2
	3	0.2	0.0	0.2	0.4
		0.5	0.0	0.5	

10 The mean number of claims per day is

$\{(32 \times 1) + (17 \times 2) + (2 \times 3) + (0 \times 4) + (1 \times 5)\}/100 = 0.77.$

Use 0.77 as an estimate of the mean of the Poisson distribution. Thus

$P(X = x) = \dfrac{e^{-\lambda}\lambda^{x}}{x!}$ is estimated by

$$P(X = x) = \frac{e^{-0.77}0.77^{x}}{x!},\ x = 0, 1, 2, \ldots$$

The expected frequencies are given by $100 \times P(X = x)$.

No. of claims (x)	0	1	2	3	4	≥5	Total
Obs. frequency (f_i)	48	32	17	2	0	1	100
Exp. frequency (e_i)	46.3	35.7	13.7	3.5	0.7	0.1	100.0

Categories $x = 3$, 4, and ≥5 are grouped together to ensure that all e_i are greater than 1.

The expected frequency for ≥ 3 is 3.5 + 0.7 + 0.1 = 4.3; the corresponding observed frequency is 3.

$$\chi^2 = \sum (f_i - e_i)^2 / e_i = \frac{1.7^2}{46.3} + \frac{3.7^2}{35.7} + \frac{3.3^2}{13.7} + \frac{1.3^2}{4.3} = 1.63.$$

There are 2 d.f. [4 categories x = 0,1,2, and ≥3, and 1 parameter estimated from the data.]

The probability value =

$$P\left(\chi_2^2 > 1.63\right) \cong 1 - 0.557 = 0.443 \text{ from the Yellow Tables } p164$$

There is insufficient evidence to suggest that the number of claims does not follow a Poisson distribution (i.e. the model provides a good fit to the data).

An alternative solution (in this over-conservative approach some information is thrown away unnecessarily - but it was awarded full marks):

Grouping categories x = 2, 3, 4 and ≥ 5 and using only 3 cells with observed frequencies 48, 32, and 20 and expected frequencies 46.3, 35.7, and 18.0 gives χ^2 = 0.668 on 1 degree of freedom. The probability value is 0.414. Same conclusion.

11 (i) Total sum: 13046 + 12592 + 10965 + 12128 = 48731

Total sum of squares: 17322090 + 16116822 + 12208021 + 14929994 = 60576927

$SS_T = 60576927 - 48731^2/40 = 1209168$

$SS_B = (13046^2 + 12592^2 + 10965^2 + 12128^2)/10 - 48731^2/40$
$= 59607619 - 48731^2/40 = 239860$

$SS_R = SS_T - SS_B = 1209168 - 239860 = 969308$

Source of variation	df	Sums of Squares	Mean Squares
Between regions	3	239860	79953
Residual	36	969308	26925
Total	39	1209168	

$F = 79953/26925 = 2.97$ on 3, 36 d.f.

Therefore, since the value of $F_{3,36}$ (0.05) is 2.866, the observed F value (2.97) exceeds it and so the null hypothesis that the population means are equal is rejected at the 5% level of significance. However, as $F_{3,36}$ (0.01) is 4.377, the null hypothesis is not rejected at the 1% level.

(ii) Means:

A: $\bar{y}_{1.} = 1304.6$ B: $\bar{y}_{2.} = 1259.2$

C: $\bar{y}_{3.} = 1096.5$ D: $\bar{y}_{4.} = 1212.8$

Least significant difference, for each pair of regions, is (5% level):

$t_{0.025,36}\ \hat{\sigma}\ (1/10 + 1/10)^{1/2} = 2.028\ \sqrt{26925}\ (2/10)^{1/2} = 149$

Differences between pairs of means:

$\bar{y}_{1.} - \bar{y}_{2.} = 45.4$, $\bar{y}_{1.} - \bar{y}_{3.} = 208.1$, $\bar{y}_{1.} - \bar{y}_{4.} = 91.8$

$\bar{y}_{2.} - \bar{y}_{3.} = 162.7$, $\bar{y}_{2.} - \bar{y}_{4.} = 46.4$, $\bar{y}_{3.} - \bar{y}_{4.} = -116.3$

Region C	Region D	Region B	Region A
$\bar{y}_{3.}$	$\bar{y}_{4.}$	$\bar{y}_{2.}$	$\bar{y}_{1.}$

(Alternative answers which have the following conclusion are acceptable: The population mean claim amount for region C appears to be less than the population mean of region A and the population mean of region B. However, the population mean for region C and the population mean for region D do not appear to differ.)

12 (i) $E[X] = E[E(X|U)] = E[U] = \alpha/\lambda$

$V[X] = E[V(X|U)] + V[E(X|U)] = E[U] + V[U] = \alpha/\lambda + \alpha/\lambda^2$

(ii) Using the method of moments, α and λ may be estimated by solving the equations

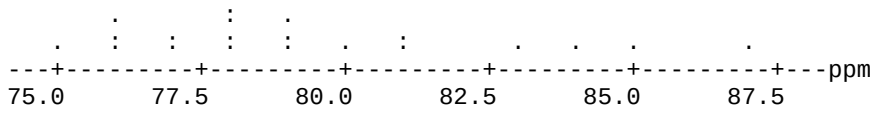
$$\bar{x} = \frac{\alpha}{\lambda} \text{ and } s^2 = \frac{\alpha}{\lambda} + \frac{\alpha}{\lambda^2}$$

which gives

$$\hat{\alpha} = \frac{\bar{x}^2}{s^2 - \bar{x}} \text{ and } \hat{\lambda} = \frac{\bar{x}}{s^2 - \bar{x}}.$$

(iii) If $s^2 \le \bar{x}$, then the method of moments produces inadmissible estimates as the parameters α and λ must be positive and finite.

13 (i) (a)



Dotplot shows moderate positive skewness

(b) $\bar{x} = \frac{1587}{20} = 79.35, \quad s^2 = \frac{126131 - \frac{1587^2}{20}}{19} = 10.66$

95% confidence interval is $\bar{x} \pm t_{0.025,19}\sqrt{\frac{s^2}{20}}$

giving $79.35 \pm 2.093\sqrt{\frac{10.66}{20}} \Rightarrow 79.35 \pm 1.53 \Rightarrow (77.82, 80.88)$

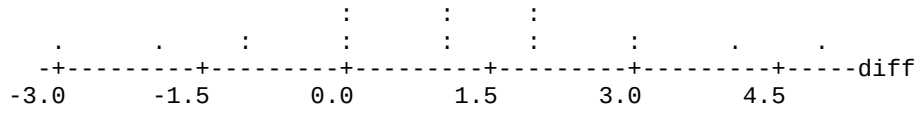
(c) This *t* confidence interval requires normality of the observations.

This may be doubtful in view of the skewness shown in part (a), but the sample size of 20 is perhaps large enough to justify the validity due to the robustness of the *t* analysis.

(ii) (a) Differences (before – after) are:

2	4	0	-1	1	3	3	0	1	0
-3	2	1	0	-2	2	1	5	2	-1

Dotplot of differences:



Seems quite symmetrical and normal

(b) Paired *t* test is appropriate.

$\Sigma d = 20$ and $\Sigma d^2 = 94$

$$\bar{d} = \frac{20}{20} = 1.0 \quad s^2 = \frac{94 - \frac{20^2}{20}}{19} = 3.895$$

Observed $t = \dfrac{1.0}{\sqrt{\frac{3.895}{20}}} = 2.27$ on 19 d.f.

For one-sided test: 5% point = 1.729, 2.5% point = 2.093 and 1% point = 2.539

P-value is approx. 0.020

So there is some evidence that the modifications have reduced the contaminant content.

(c) This *t* analysis requires normality of the differences and this seems reasonable from part (a).

14 (i) (a) The least squares estimate of β minimises

$$q = \sum_{i=1}^{n} (y_i - \beta x_i)^2 = \sum_{i=1}^{n} y_i^2 - 2\beta \sum_{i=1}^{n} x_i y_i + \beta^2 \sum_{i=1}^{n} x_i^2.$$

Differentiating with respect to β gives

$$\frac{dq}{d\beta} = 2(\beta\sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i\, y_i).$$

Equating to zero gives the least squares estimator as

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i Y_i}{\sum_{i=1}^{n} x_i^2} \text{ as required.}$$

(b) Mean and variance of $\hat{\beta}_1$:

$$E(\hat{\beta}_1) = \sum_{i=1}^{n} x_i E(Y_i \mid x_i) / \sum_{i=1}^{n} x_i^2$$

$$= \sum_{i=1}^{n} x_i \beta x_i / \sum_{i=1}^{n} x_i^2 = \beta$$

$$V(\hat{\beta}_1) = \sigma^2 \sum_{i=1}^{n} x_i^2 / (\sum_{i=1}^{n} x_i^2)^2 = \sigma^2 / \sum_{i=1}^{n} x_i^2.$$

(ii) (a) The alternative estimator $\hat{\beta}_2 = \sum_{i=1}^{n} Y_i / \sum_{i=1}^{n} x_i$ has expectation and variance

$$E(\hat{\beta}_2) = \sum_{i=1}^{n} E(Y_i \mid x_i) / \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} \beta x_i / \sum_{i=1}^{n} x_i = \beta,$$

$$V(\hat{\beta}_2) = n\sigma^2 / \left(\sum_{i=1}^{n} x_i\right)^2 = \sigma^2 / (n\bar{x}^2).$$

(b) $V(\hat{\beta}_2) \geq V(\hat{\beta}_1)$

$$\Leftrightarrow \sigma^2 / n\bar{x}^2 \geq \sigma^2 / \sum_{i=1}^{n} x_i^2$$

$$\Leftrightarrow \sum_{i=1}^{n} x_i^2 - n\bar{x}^2 \geq 0$$

$$\Leftrightarrow \sum_{i=1}^{n} (x_i - \bar{x})^2 \geq 0$$

$\therefore$ The variance of $\hat{\beta}_2$ is at least as large as the variance of the least squares estimator $\hat{\beta}_1$.

(iii) (a) $$E(\hat{\beta}_3) = E\left(\sum_{i=1}^{n} a_i Y_i\right) = \sum_{i=1}^{n} a_i E(Y_i \mid x_i) = \sum_{i=1}^{n} a_i \beta x_i = \beta \sum_{i=1}^{n} a_i x_i$$

$\therefore E(\hat{\beta}_3) = \beta$, i.e. unbiased, if $\sum_{i=1}^{n} a_i x_i = 1$

$$V(\hat{\beta}_3) = V\left(\sum_{i=1}^{n} a_i Y_i\right) = \sum_{i=1}^{n} a_i^2 \sigma^2$$

(b) $$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i Y_i}{\sum_{i=1}^{n} x_i^2} = \sum_{i=1}^{n} a_i Y_i, \quad \text{where} \quad a_i = \frac{x_i}{\sum_{i=1}^{n} x_i^2}, \; i = 1, \ldots, n.$$

$$\therefore \sum_{i=1}^{n} a_i x_i = \sum_{i=1}^{n} \frac{x_i}{\sum_{i=1}^{n} x_i^2} \; x_i = \frac{\sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n} x_i^2} = 1$$

i.e. the condition $\sum_{i=1}^{n} a_i x_i = 1$ is satisfied.

$$\hat{\beta}_2 = \frac{\sum_{i=1}^{n} Y_i}{\sum_{i=1}^{n} x_i} = \sum_{i=1}^{n} a_i Y_i, \quad \text{where} \quad a_i = \frac{1}{n\bar{x}}, \; i = 1, \ldots, n.$$

$$\therefore \sum_{i=1}^{n} a_i x_i = \sum_{i=1}^{n} \frac{1}{n\bar{x}} \; x_i = \frac{n\bar{x}}{n\bar{x}} = 1$$

i.e. the condition $\sum_{i=1}^{n} a_i x_i = 1$ is satisfied.

(c) Among estimators of the form $\sum_{i=1}^{n} a_i Y_i$, the minimum variance unbiased estimator of β is

$$\hat{\beta}_3 = \sum_{i=1}^{n} a_i Y_i = \frac{\sum_{i=1}^{n} x_i Y_i}{\sum_{i=1}^{n} x_i^2} = \hat{\beta}_1$$

i.e. the least squares estimator.

15 (i) (a) Let Y = ratio for one week, and $Y = e^X$

P(decrease in one week) = $P(Y < 1)$

$= P(e^X < 1) = P(X < 0)$

$$= P(Z < \frac{0 - 0.0125}{0.055} = -0.227) = 1 - 0.5898 = 0.41$$

(b) P(decrease in next two weeks) = $(0.41)^2 = 0.17$

(c) We require $P(\frac{S(2)}{S(0)} > 1) = P(\frac{S(2)}{S(1)} . \frac{S(1)}{S(0)} > 1)$

$= P(Y_2 . Y_1 > 1) = P(X_2 + X_1 > 0)$

where X_2, X_1 are independent $N(\mu,\sigma^2)$

$\therefore X_2 + X_1 \sim N(2\mu, 2\sigma^2)$

$$\therefore P = P(Z > \frac{0 - 2(0.0125)}{\sqrt{2}(0.055)} = -0.321) = 0.63 \quad \text{from tables.}$$

(d) Extending the method of part (c):

$$\sum_{i=1}^{20} X_i \sim N(20\mu, 20\sigma^2)$$

$$\therefore P = P(Z < \frac{0 - 20(0.0125)}{\sqrt{20}(0.055)} = -1.016) = 0.155$$

(ii) (a) The ratios are <u>independent</u> and <u>identically distributed</u> lognormal r.v.'s.

This defines a random sample from a lognormal distribution.

(b) For the 10 observed ratios $y_1, \ldots, y_{10}$:

$\Sigma y = 10.192 \quad \Rightarrow \bar{y} = 1.0192$

$\Sigma y^2 = 10.441562 \quad \Rightarrow s^2 = 0.005986 \Rightarrow s = 0.0774$

(c) For the method of moments:

solve the following equations for μ and σ^2

$$e^{\mu+\frac{1}{2}\sigma^2} = 1.0192 \qquad (1)$$

$$e^{2\mu+\sigma^2}(e^{\sigma^2} - 1) = 0.005986 \qquad (2)$$

$$(2) \div (1)^2 \quad \Rightarrow e^{\sigma^2} - 1 = 0.0057625$$

$$\therefore \sigma^2 = 0.005746 \quad \therefore \sigma = 0.0758$$

$$(1) \quad \Rightarrow \mu = \log(1.0192) - \frac{1}{2}\sigma^2 = 0.0161$$

[*Note*: in MME candidates could use $\hat{\sigma}^2 = 0.005388$ with divisor n not $(n-1)$ to obtain $\sigma = 0.0719$ and $\mu = 0.0164$]