EXAMINATIONS

April 2004

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

22 June 2004

General comments

The Examiners are of the view that, overall, the paper was of a comparable standard to those set in recent diets. They note that the responses to questions 1, 3, 8(i), 11(i), and 14 were particularly good, while the responses to questions 4, 6, 7(ii), 8(ii), and 12(iii) were particularly poor.

It was clear that many candidates had difficulty with conditional probabilities and conditional expectations.

1 X = number of policies where a claim is made.
$X \sim bi\,(500, 0.04)$
$X \approx N((500)(0.04), 500(0.04)(0.96)) \approx N(20, 19.2)$

$P(X \leq 30) \equiv P(X \leq 30.5)$ using continuity correction

$$\doteq \Phi\left(\frac{30.5-20}{\sqrt{19.2}}\right) = \Phi\left(\frac{10.5}{\sqrt{19.2}}\right) = \Phi(2.40)$$

$$= 0.9918$$

2 The required values of x_1 and x_2, the lower and upper quartiles, are such that:

$$F(x_1) = 0.25 \quad \text{and} \quad F(x_2) = 0.75,$$

i.e. $$\Phi\left(\frac{x_1-\mu}{\sigma}\right) = 0.25 \quad \text{and} \quad \Phi\left(\frac{x_2-\mu}{\sigma}\right) = 0.75,$$

where Φ is the standard normal distribution function.

From the statistical table of the percentage points of the standard normal distribution, we have

$$\frac{x_1-\mu}{\sigma} = -0.6745 \quad \Rightarrow x_1 = \mu - 0.6745\sigma$$

$$\frac{x_2-\mu}{\sigma} = 0.6745 \quad \Rightarrow x_2 = \mu + 0.6745\sigma$$

where μ is the population mean and σ is the population standard deviation.

$\therefore IQR = x_2 - x_1$
$= 0.6745\sigma + 0.6745\sigma$
$= 1.349\sigma.$

3 $C'(t) = 20(1 - t)^{-11}$, $C''(t) = 220(1 - t)^{-12}$
$E[X] = C'(0) = 20$
$V[X] = C''(0) = 220$

[OR as coefficients of t and of $t^2/2!$ in expansion of $C(t)$]

4 Answer: \$35,000 + 20,000 = \$55,000

Reason: the memoryless property of the exponential distribution (the excess above 35,000 itself has an exponential distribution with mean 20,000).

[*Note:* relatively few candidates were able to exploit the memoryless property of the exponential distribution to advantage.]

5 $MSE(k\hat{\theta}) = E((k\hat{\theta} - \theta)^2) = \text{var}(k\hat{\theta}) + [E(k\hat{\theta})]^2 - 2\theta E(k\hat{\theta}) + \theta^2$

$$= k^2 \frac{\theta^2}{10} + (k\theta)^2 - 2\theta k\theta + \theta^2 = \theta^2 \left(\frac{k^2}{10} + k^2 - 2k + 1 \right)$$

$$\frac{dMSE(k\hat{\theta})}{dk} = \theta^2 \left(2\frac{k}{10} + 2k - 2 \right) = 0 \Rightarrow k = \frac{1}{\frac{1}{10} + 1} = \frac{10}{11}.$$

This is clearly a minimum as the *MSE* is a quadratic in k and the coefficient of k^2 is positive.

OR: Note explicitly in first line that

$$MSE\left(k\hat{\theta}\right) = \text{var}\left(k\hat{\theta}\right) + \left[\text{bias}\left(k\hat{\theta}\right)\right]^2$$

[*Note: Many candidates stated in error that* $k\hat{\theta}$ *was unbiased for* θ*. Some attempted to minimise the function of k by differentiating with respect to* θ *rather than k.*]

6 $E[X] = E[E[X|Y]] = P(Y = 1)\, E[X|Y = 1] + P(Y = 2)\, E[X|Y = 2]$

So $1.2 = 0.6 \times (7/6) + 0.4 \times E[X|Y = 2]$ and so $E[X|Y = 2] = 1.25$

7 (i) Let I_A, I_B, I_C be indicator variables such that

$P(I_A = 1) = 0.01$, $P(I_B = 1) = 0.02$, $P(I_C = 1) = 0.005$ and 0 otherwise.

Let T = total claim amount = 2.5 I_A + 4.8 I_B + 7.2 I_C

$\therefore E(T)$ = 2.5(0.01) + 4.8(0.02) + 7.2(0.005) = 0.157 (or £15.7m)

Var(T) = $2.5^2(0.01)(0.99) + 4.8^2(0.02)(0.98) + 7.2^2(0.005)(0.995) = 0.7714$

$\therefore$ s.d.(T) = 0.878 (or £87.8m)

(ii) P(1 claim) = P(A not BC) + P(B not AC) + P(C not AB)
= (0.01)(0.98)(0.995) + (0.02)(0.99)(0.995) + (0.005)(0.99)(0.98)
= 0.009751 + 0.019701 + 0.004851 = 0.034303

$P(A|1 \text{ claim}) = 0.009751/0.034303 = 0.2843$
$P(B|1 \text{ claim}) = 0.019701/0.034303 = 0.5743$
$P(C|1 \text{ claim}) = 0.004851/0.034303 = 0.1414$

$\therefore E(T|1 \text{ claim}) = 2.5(0.2843) + 4.8(0.5743) + 7.2(0.1414) = 4.485$
(or £448.5m)

Much larger given that a claim does in fact arise.

8 (i) $M_{X_1}(t) = \exp\{\mu_1(e^t - 1)\}$

$M_{X_2}(t) = \exp\{\mu_2(e^t - 1)\}$

$M_S(t) = M_{X_1}(t)M_{X_2}(t)$
$= \exp\{\mu_1(e^t - 1)\} \cdot \exp\{\mu_2(e^t - 1)\}$
$= \exp\{(\mu_1 + \mu_2)(e^t - 1)\}$

This is the MGF of a $Po(\mu_1 + \mu_2)$.

$\therefore S \sim Po(\mu_1 + \mu_2)$

(ii) $$P(X_1 = x_1 \mid S = X_1 + X_2 = s) = \frac{P(X_1 = x_1, X_1 + X_2 = s)}{P(X_1 + X_2 = s)}$$

$$= \frac{P(X_1 = x_1)P(X_2 = s - x_1)}{P(X_1 + X_2 = s)} = \frac{\dfrac{\mu_1^{x_1}e^{-\mu_1}}{x_1!}\dfrac{\mu_2^{s-x_1}e^{-\mu_2}}{(s-x_1)!}}{\dfrac{(\mu_1+\mu_2)^s e^{-(\mu_1+\mu_2)}}{s!}} = \binom{s}{x_1}\frac{\mu_1^{x_1}\mu_2^{s-x_1}}{(\mu_1+\mu_2)^s}$$

$$= \binom{s}{x_1}\left(\frac{\mu_1}{\mu_1+\mu_2}\right)^{x_1}\left(1-\frac{\mu_1}{\mu_1+\mu_2}\right)^{s-x_1}$$

Binomial distribution with parameters: $n = s, \quad p = \dfrac{\mu_1}{\mu_1 + \mu_2}$.

9 (i) S is approximately normal for large n by Central Limit Theorem.

Using results from the Yellow Book for $X_i \sim U(-\frac{1}{2}, \frac{1}{2})$:

$$E[S] = nE[X_i] = n \times 0 = 0\ ,\ Var[S] = Var[\Sigma X_i] = nVar[X_i] = \frac{n}{12}$$

So $S \sim N(0, n/12)$

(ii) S approx $\sim N(0, 1)$ if $n = 12$.

(iii) The distribution of each X_i is symmetric and so the distribution of the sum $S = \Sigma X_i$ is also symmetric. So skewness is zero, as for any normal distribution.

10 Total number of claims $N \sim$ Poisson(50λ)

(a) In the case $\lambda = 0.3$

Power = $P[N \geq 15 | N \sim \text{Poisson}(15)]$

$= 1 - P[N \leq 14 | N \sim \text{Poisson}(15)] = 1 - 0.466 = 0.534$

(b) In the case $\lambda = 0.4$

Power = $P[N \geq 15 | N \sim \text{Poisson}(20)] = 1 - 0.105 = 0.895$

Comment: power higher for $\lambda = 0.4$, which is further away from H_0 value 0.2

11 (i) $n_1 = n_2 = n_3 = n_4 = 7 \qquad n = 28$

$\Sigma y_1 = 227 \quad \Sigma y_2 = 179 \quad \Sigma y_3 = 213 \quad \Sigma y_4 = 218 \quad \Sigma\Sigma y_{ij} = 837$
$\Sigma y_1^2 = 7501 \quad \Sigma y_2^2 = 4703 \quad \Sigma y_3^2 = 7125 \quad \Sigma y_4^2 = 6916 \quad \Sigma\Sigma y_{ij}^2 = 26245$

$$SS_T = 26245 - \frac{837^2}{28} = 26245 - 25020.321 = 1224.7$$

$$SS_B = \frac{1}{7}(227^2 + 179^2 + 213^2 + 218^2) - 25020.321$$
$$= 25209 - 25020.321 = 188.7$$

$\therefore\ SS_R = SS_T - SS_B = 1224.7 - 188.7 = 1036.0$

Source of variation	*d.f.*	*SS*	*MSS*
Companies	3	188.7	62.9
Residual	24	1036.0	43.2
Total	27	1224.7	

$F = \dfrac{62.9}{43.2} = 1.46$ on (3,24) degrees of freedom.

The 10% point of $F_{3,24}$ distribution is 2.327. Therefore, there is insufficient evidence to reject the null hypothesis that the population means for the four companies are equal, i.e., the distributions of the sums insured are the same for the four companies.

(ii) The model used in (i) assumes that the sums insured for each company follow a normal distribution, and the population variances are equal.

The plot of residuals shows:

- normality seems appropriate, but observation £51,000 seems to be an outlier
- companies have similar sample variances, but one could argue that there is a suggestion that the variance for Company 3 is higher than the variances for the other companies.

Therefore, overall the one-way analysis of variance model seems adequate and the conclusions in (i) are valid.

12 (i) $L(p) = \binom{n}{x} p^x (1-p)^{n-x}$

$$\log L(p) = \log \binom{n}{x} + x \log p + (n-x)\log(1-p)$$

$$\frac{\partial}{\partial p} \log L(p) = \frac{x}{p} - \frac{n-x}{1-p}$$

equate to zero $\Rightarrow\ x(1-p) = p(n-x)\ \Rightarrow\ p = \dfrac{x}{n}$

clearly maximises $L(p)$

$\therefore$ MLE is $\hat{p} = \dfrac{X}{n}$

(ii) (a) $\dfrac{\partial^2}{\partial p^2}\log L(p) = -\dfrac{x}{p^2} - \dfrac{n-x}{(1-p)^2}$

$$E\left(\frac{\partial^2}{\partial p^2}\log L(p)\right) = -\frac{np}{p^2} - \frac{n-np}{(1-p)^2} = -\frac{n}{p} - \frac{n}{1-p} = -\frac{n}{p(1-p)}$$

$$\therefore CRlb = \frac{-1}{-\dfrac{n}{p(1-p)}} = \frac{p(1-p)}{n}$$

(b) $\text{Var}(\hat{p}) = \dfrac{np(1-p)}{n^2} = \dfrac{p(1-p)}{n} = CRlb$

(c) $\hat{p} \approx N(p, CRlb) \sim N(p, \dfrac{p(1-p)}{n})$

(iii) $$-1.96 < \frac{\hat{p}-p}{\sqrt{\dfrac{p(1-p)}{n}}} < 1.96 \quad \Rightarrow \quad \frac{(\hat{p}-p)^2}{\dfrac{p(1-p)}{n}} < 1.96^2$$

$$\Rightarrow \quad p^2 - 2\hat{p}p + \hat{p}^2 < \frac{1.96^2}{n}(p - p^2)$$

$$\Rightarrow \quad (1+\frac{1.96^2}{n})p^2 - (2\hat{p} + \frac{1.96^2}{n})p + \hat{p}^2 < 0$$

This is a quadratic and will be negative between its two roots.

So, p_L, p_U will be the two roots:

$$\frac{(2\hat{p} + \dfrac{1.96^2}{n}) \pm \sqrt{(2\hat{p} + \dfrac{1.96^2}{n})^2 - 4\hat{p}^2(1 + \dfrac{1.96^2}{n})}}{2(1 + \dfrac{1.96^2}{n})}$$

with p_L from the "–" sign, and p_U from the "+" sign.

(iv) $$-1.96 < \frac{\hat{p} - p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} < 1.96$$

$$\Rightarrow \hat{p} - 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < p < \hat{p} + 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \text{ giving } p_L \text{ and } p_U .$$

(v) (a) $x = 4$, $n = 10$

quadratic from (iii) is $1.38416p^2 - 1.18416p + 0.16$

roots give p_L = 0.168 and p_U = 0.687.

from (iv) p_L = 0.096 and p_U = 0.704.

quite a difference, especially in p_L , but n = 10 is small.

(b) $x = 80$, $n = 200$

quadratic from (iii) is $1.019208p^2 - 0.819208p + 0.16$

roots give p_L = 0.335 and p_U = 0.469.

from (iv) p_L = 0.332 and p_U = 0.468.

very similar (and much narrower than (a)) with n = 200 being large.

In (i) many candidates wanted to write the likelihood as a product – this is OK using Bernoulli probabilities as the individual components, as in

$$L(p) = \prod_{i=1}^{n} p^{x_i} (1-p)^{1-x_i}$$

where x_i = 1 or 0, but not as a product of binomial(n,p) components.

13 (i) (a) $\bar{x} = 952.75/100 = 9.53$

$$s = \left\{ \frac{1}{99} \left(13584.5217 - \frac{952.75^2}{100} \right) \right\}^{0.5} = 6.75$$

An exponential distribution has mean = standard deviation so there must be some doubt here as to whether such a distribution will be a good description of these data.

(b) The fitted model is an exponential distribution with mean 9.53 (i.e. with parameter 1/9.5275 = 0.105).

Reason: the maximum likelihood estimate (and the method of moments estimate) of the mean μ of an exponential distribution is the sample mean

(ii) (a) For $X \sim \exp(\text{mean} = \mu)$ we have $P(X > x) = \exp(-x/\mu)$
$\exp(-x/\mu) = p \Rightarrow x = -\mu \log p$

(b) Using $p = 0.8, 0.6, 0.4$, and 0.2 and fitted mean 9.5275
we get $x = 2.13, 4.87, 8.73, 15.33$

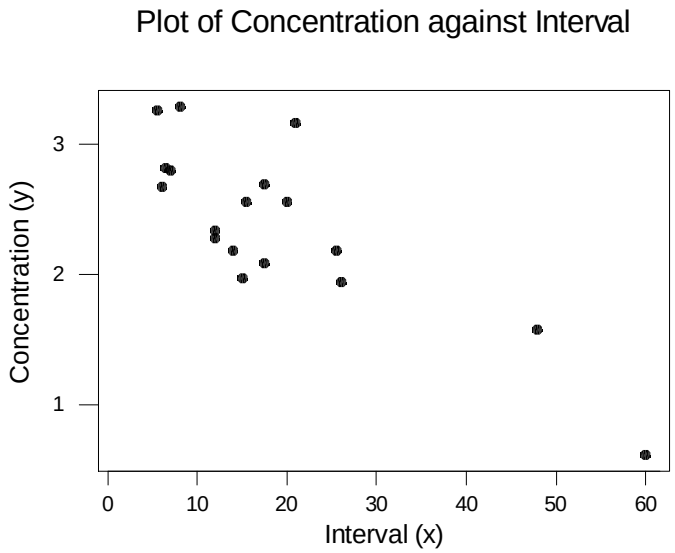
(iii) The numbers of observations in (0,2.13), (2.13,4.87), (4.87,8.73), (8.73,15.33) and (15.33,∞) are, by inspection from the data, respectively 10, 20, 24, 30, and 16.

Interval	0 – 2.13	2.13 – 4.87	4.87 – 8.73	8.73 – 15.33	15.33 – ∞
Observed frequency	10	20	24	30	16
Expected frequency	20	20	20	20	20

$\chi^2 = (10^2 + 0 + 4^2 + 10^2 + 4^2)/20 = 232/20 = 11.6$
df = 3

P-value = $P(\chi^2 > 11.6) = 0.009$ (from Yellow Tables p164), so we reject the exponential distribution as a description – it provides a very poor fit to the data.

14 (i)



The concentration of 3-MT in the brain decreases as the post-mortem interval increases from 5.5 hours to 60 hours. There are two points with a much higher post-mortem interval than the other observations.

The data seem to be appropriate for linear regression, but care should be taken when evaluating the effect for the higher interval values as there are only 2 points in the higher x-range.

(ii) $n = 18$

$S_{xx} = \Sigma x^2 - (\Sigma x)^2/n$
$= 9854.5 - (337)^2/18$
$= 3545.1111$

$S_{yy} = \Sigma y^2 - (\Sigma y)^2/n$
$= 109.7936 - (42.98)^2/18$
$= 7.1669111$

$S_{xy} = \Sigma xy - (\Sigma x)(\Sigma y)/n$
$= 672.8 - (337)(42.98)/18$
$= -131.88111$

Correlation coefficient:

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{-131.88111}{\sqrt{(3545.1111)(7.1669111)}}$$

$= -0.827$

Test H_0: $\rho = 0$ vs H_1: $\rho \neq 0$

$$t = r\sqrt{\frac{n-2}{1-r^2}} \sim t_{n-2}$$

$$= (-0.827)\sqrt{\frac{16}{1-(-0.827)^2}}$$

$= -5.89$

$t_{16}(2.5\%) = 2.120$
$t_{16}(0.5\%) = 2.921$

$|t| = 5.89$ is larger than the critical values at the 5% and 1% significance levels. Therefore reject the null hypothesis that the population correlation coefficient is equal to zero, and conclude that there is a linear relationship between interval and concentration.

(iii) Model: $y = \alpha + \beta x$

Slope:

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}} = \frac{-131.88111}{3545.1111}$$

$= -0.0372008$ or -0.0372 (to 4 d.p.)

Intercept:

$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$
= 42.98/18 – (–0.0372008)(337/18)
= 3.084259 or 3.0843 (to 4 d.p.)

$y = \hat{\alpha} + \hat{\beta}x$
= 3.0843 – 0.0372 x
i.e. Concentration = 3.0843 – 0.0372 × Interval

(a) At 1 day = 24 hours:

y = 3.0843 – 0.0372 x
= 3.0843 – 0.0372 (24)
= 2.19

(b) At 2 days = 48 hours:

y = 3.0843 – 0.0372 x
= 3.0843 – 0.0372 (48)
= 1.30

This data set contains accurate data up to 26 hours, as for observations 17 and 18 (at 48 hours and 60 hours respectively) there was no eye-witness testimony directly available. Predicting 3-MT concentration after 26 hours may not be advisable, even though x = 48 is within the range of the x-values (5.5 hours to 60 hours).

(iv) $$\hat{\sigma}^2 = \frac{1}{n-2}\left(S_{yy} - \frac{S_{xy}^2}{S_{xx}}\right)$$

$$= \frac{1}{16}\left(7.1669111 - \frac{(-131.88111)^2}{3545.1111}\right)$$

$$= \frac{2.2608231}{16}$$

= 0.1413014

$\hat{\sigma}$ = 0.3759

$$\text{s.e.}(\hat{\beta}) = \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} = \frac{0.3759}{\sqrt{3545.111}}$$

= 0.00631331

99% confidence interval for slope is $\hat{\beta} \pm t_{n-2}$ s.e.$(\hat{\beta})$ df = $n - 2 = 16$

$= -0.0372008 \pm (2.921)\ (0.00631331)$
$= -0.0372 \pm 0.0184$
or $(-0.0556, -0.0188)$

$\beta = 0$ is not within this 99% confidence interval, therefore we would reject the null hypothesis $\beta = 0$, at the 1% significance level.

Therefore there does appear to be a (negative) linear relationship between Interval and Concentration.

This confirms the result in (ii) where the correlation coefficient was shown to not equal zero at the 1% significance level.

Note: Plots of data need not be works of graphic art, but should at least be neat and clear enough (with adequate title and labels) to convey the information.

END OF REPORT