EXAMINATIONS

September 2002

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.

K G Forman
Chairman of the Board of Examiners
12 November 2002

1 $\binom{5}{3} 0.4^3 0.6^2 = 10 \times 0.02304 = 0.2304$

2 With n components the probability that the motor fails is the probability that all of the components fail simultaneously.

$P(\text{motor fails}) = P(n \text{ independent components fail}) = 0.02^n$

This is less than 10^{-9} if $n \log(0.02) \le \log(10^{-9}) \Rightarrow n \ge \log(10^{-9})/\log(0.02) = 5.30$

Therefore the minimum number of components is 6.

[OR: by trial and error]

3 $P(X > 3 \mid X > 1) = P(X > 3 \text{ and } X > 1)/P(X > 1) = P(X > 3)/P(X > 1)$
$= (2/5)^3/(2/3)^3 = 0.216$

4 Using units of £1000:

Total sum assured $S \sim N(100 \times 8, 100 \times 9)$ i.e. $S \sim N(800, 900)$ approximately, by Central Limit Theorem.

$P(S > 845) \approx P[Z > (845 - 800)/30] = P(Z > 1.5)$ where $Z \sim N(0, 1)$
$= 0.067$

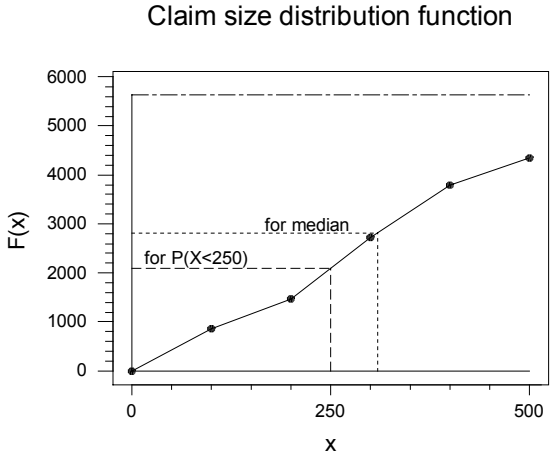
5 $\hat{y}_i = \bar{y} + \hat{\beta}(x_i - \bar{x})$ so $\sum \hat{y}_i = n\bar{y}$ so mean of fitted values is $\bar{y}$

$\therefore$ Correlation coeff$^{t.}$ of observed and fitted values is

$$\frac{\sum (y_i - \bar{y})(\hat{y}_i - \bar{y})}{\left(\sum (y_i - \bar{y})^2 \sum (\hat{y}_i - \bar{y})^2\right)^{1/2}} = \frac{\hat{\beta} S_{xy}}{\left(\hat{\beta}^2 S_{yy} S_{xx}\right)^{1/2}} = r$$

6 (i) Cumulative frequency distribution is

x	*F(x)*
0	0
100	862
200	1470
300	2723
400	3789
500	4347
max	5637



(ii) For $x = 250$ the corresponding $F(x)$ is approximately

$$1470 + \frac{(2723 - 1470)}{100}(250 - 200) \approx 2096.5$$

and the corresponding approximate proportion is $\frac{2096.5}{5637} = 0.37$

(iii) For the median we need x such that $F(x) = 0.5(5637) = 2818.5$

$$\text{median} \approx 300 + \frac{2818.5 - 2723}{1066}(100) = £309$$

The maximum claim size is not given, and candidates were expected to exercise common sense and judgement. One sensible approach is to leave the x-scale open-ended and indicate the presence of a horizontal asymptote at cumulative frequency 5637.

7 (i) $M_Y(t) = E(e^{tY}) = E(e^{-t\log X})$

$= E(X^{-t}) = \int_0^1 x^{-t}dx = \left[\frac{x^{-t+1}}{1-t}\right]_0^1$

$= \frac{1}{1-t}$, $t < 1$.

(ii) This is the MGF of an exponential distribution with parameter 1, and by uniqueness for the MGF this then implies that Y has this distribution.

8 (i) $\hat{\lambda} = \frac{183}{1500} = 0.122$

(ii) (a) Number of claims X for 10 policies in six months is Poisson with parameter 10(0.5)(0.122) = 0.61

P(no claims) = $P(X = 0)$ = exp(–0.61) = 0.543

Alternative:

For a single policy P(no claim) = exp(–0.061) = 0.9408

For 10 policies P(no claims) = $(0.9408)^{10}$ = 0.543

(b) Number of claims X for 250 policies in one year is Poisson with parameter 250(0.122) = 30.5

Outwith scope of the Green tables, so we must use a normal approximation:

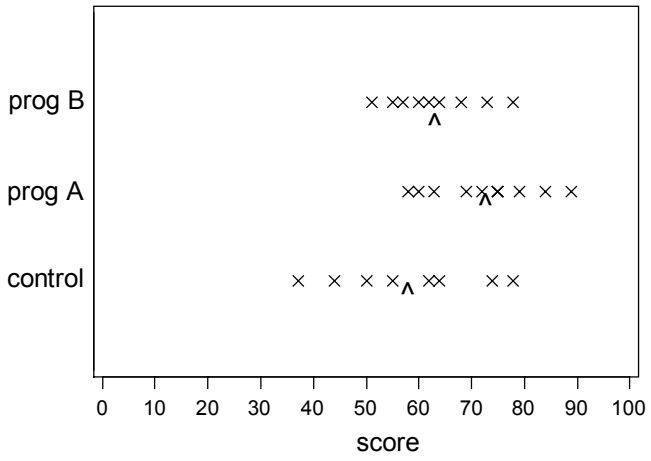
$X \approx N(30.5, 30.5)$

Applying a continuity correction:

$P(X > 40) \to P(X > 40.5) = P(Z > \frac{40.5 - 30.5}{\sqrt{30.5}} = 1.81) = 1 - 0.96485$

$= 0.035$

9 (i)



^ indicates treatment means : helpful but not required for the marks

"Within treatment" variation is the spread within each set of points, "between treatment" variation is the spread among the 3 treatment means.

(ii) $SS_T = 118128 - 1756^2/27 = 3923.0$
$SS_B = (464^2/8 + 724^2/10 + 568^2/9) - 1756^2/27 = 971.67$
$\therefore\ SS_R = 3923.0 - 971.67 = 2951.3$

Source of variation	*df*	*SS*	*MSS*
Between treatments	2	971.67	485.84
Residual	24	2951.3	122.97
Total	26	3923.0	

Under H_0 : no treatment effects $F = 485.84/122.97 = 3.95$ on 2,24 df

P-value is < 0.05, so we have some evidence against H_0.

We conclude that there is evidence of differences among the treatments. The data suggest that training programme A gives a higher mean score than the others.

10 (i) $E(S) = E(N)E(X) = \mu\frac{\alpha}{\lambda}$

$$\text{Var}(S) = E(N)\text{Var}(X) + \text{Var}(N)[E(X)]^2 = \mu\frac{\alpha}{\lambda^2} + \mu(\frac{\alpha}{\lambda})^2 = \frac{\mu\alpha}{\lambda^2}(1+\alpha)$$

$$\therefore sd(S) = \sqrt{\frac{\mu\alpha(1+\alpha)}{\lambda^2}}$$

(ii) (a) $\frac{\alpha}{\lambda} = 100 \quad \& \quad \frac{\alpha}{\lambda^2} = 50^2 \quad \Rightarrow \quad \alpha = 4\,,\, \lambda = 0.04$

$E(S) = 100(100) = £10{,}000$

$$sd(S) = \sqrt{\frac{100(4)(1+4)}{0.04^2}} = £1{,}118$$

(b) Since N is going to be large, the Central Limit Theorem allows normality.

$$P(S > 12500) \approx P(Z > \frac{12500 - 10000}{1118} = 2.236) \text{ where } Z \sim N(0,1)$$
$$= 1 - 0.987 = 0.013$$

11 (i) $P(T < x) = 1 - \exp(-\lambda x)$ so $X \sim \text{bi}(n,\, 1 - \exp(-\lambda t_0))$

(ii) (a) $L(\lambda) \propto (1 - \exp(-\lambda t_0))^x\, [\exp(-\lambda t_0)\,]^{n-x}$

$\therefore \ell(\lambda) = x\log(1 - \exp(-\lambda t_0)) + (n - x)\log[\exp(-\lambda t_0)] + \text{constant}$

$= x\log(1 - \exp(-\lambda t_0)) - (n - x)\lambda t_0 + \text{constant}$

(b) $\frac{\partial \ell}{\partial \lambda} = x t_0 \exp(-\lambda t_0)\,/(1 - \exp(-\lambda t_0)) - (n - x)t_0$

$= x t_0[1 + \exp(-\lambda t_0)\,/(1 - \exp(-\lambda t_0))] - n t_0$

$= x t_0/[(1 - \exp(-\lambda t_0)] - n t_0$

Setting $\frac{\partial \ell}{\partial \lambda} = 0 \Rightarrow \exp(-\lambda t_0) = 1 - x/n \;\; \therefore \hat{\lambda} = -\frac{1}{t_0}\log(1-\frac{x}{n})$

[OR: direct from MLE of P(survive) = $\exp(-\lambda t_0)$ is observed proportion which survive, namely $1 - x/n$; hence MLE of λ.*]*

(iii) From (ii) (b) $\frac{\partial^2 \ell}{\partial \lambda^2} = -x t_0 (1 - \exp(-\lambda t_0))^{-2} (t_0 \exp(-\lambda t_0))$

$$= -x t_0^2 \exp(-\lambda t_0)\,(1 - \exp(-\lambda t_0))^{-2}$$

$$\therefore E(-\frac{\partial^2 \ell}{\partial \lambda^2}) = t_0^2 \exp(-\lambda t_0)(1 - \exp(-\lambda t_0))^{-2} E(X)$$

$$= t_0^2 \exp(-\lambda t_0)(1 - \exp(-\lambda t_0))^{-2}\, n(1 - \exp(-\lambda t_0))$$

[using mean of binomial distribution from part (i)]

$$= n\, t_0^2 \exp(-\lambda t_0) / [1 - \exp(-\lambda t_0)]$$

(iv) (a) $\hat{\lambda} = -(1/20)\log(0.68) = 0.019283$

Estimate of $\exp(-\lambda t_0)$ is 0.68,

so $E(-\frac{\partial^2 \ell}{\partial \lambda^2}) \approx 1000 \times 20^2 \times 0.68/0.32 = 850000$

$\therefore$ s.e.$(\hat{\lambda}) \cong 850000^{-1/2} \cong 0.0010847 \left(\cong 0.00108\right)$

(b) $0.019283 \pm (1.96 \times 0.0010847)$

i.e. 0.019283 ± 0.002126 i.e. 0.01716 to 0.02141

$\therefore$ 95% CI for mean lifetime $1/\lambda$ is $(1/0.02141, 1/0.01716)$

i.e. 46.7 years to 58.3 years

There was a slight error in line 6 of Question 11: the last word should have appeared as "years" instead of "hours". The examiners took this into account, and gave full credit to any candidates who provided an alternative answer because of this.

12 (i) (a) Males: $n_1 = 10$ $\bar{x}_1 = 7.8$ $s_1 = 6.8605$

95% CI for male data:

$$\bar{x}_1 \pm t_9(0.025)\frac{s_1}{\sqrt{n_1}}$$

$$= 7.8 \pm 2.262\ \frac{6.8605}{\sqrt{10}}\ = 7.8 \pm 4.907$$

$$= (2.89, 12.71)$$

(b) Females: $n_2 = 10$ $\bar{x}_2 = 9.4$ $s_2 = 5.37897$

95% CI for female data:

$$\bar{x}_2 \pm t_9(0.025)\frac{s_2}{\sqrt{n_2}}$$

$$= 9.4 \pm 2.262\frac{5.37897}{\sqrt{10}} = 9.4 \pm 3.848$$

$$= (5.55, 13.25)$$

(c) Neither of the intervals include zero, and therefore there is evidence that the alcohol has an effect on reaction times, i.e. it increases the reaction time.

(ii) (a) Two sample t-test

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

$$= \frac{9(6.8605)^2 + 9(5.37897)^2}{18} = 37.9999$$

$$s_p = 6.1644$$

$$t = \frac{7.8 - 9.4}{6.1644\sqrt{\frac{2}{10}}} = \frac{-1.6}{2.7568} = -0.58$$

$t_{18}(0.25) = 0.688$, and probability value is > 0.5.

$\therefore$ There is no evidence to reject the null hypothesis that the means for males and females are the same.

We can conclude that alcohol has a similar effect.

(b) $\dfrac{s_1^2}{s_2^2} = 1.626$

The $F_{9,9}$ distribution has upper 5% critical point at 3.179. Our observed value (1.626) is well within the main body of the distribution and is not significant at the 10% level of testing (and therefore not at the 5% level). There is no evidence to suggest that the variances differ. The assumption of common variance was made when conducting the test in (ii)(a).

13 (i) n_i = number in group i , r_i = number with coronary heart disease in group i.

(a) $\sum r_i = 43 \quad \sum n_i = 100$

Estimate of the probability of having coronary heart disease is given by

$$\hat{\theta} = \frac{\sum r_i}{\sum n_i} = \frac{43}{100} = 0.43.$$

(Assuming constant probability of having coronary heart disease over age groups.)

(b)

		Coronary heart disease		
		Yes	*No*	*Total*
Age groups	20–29	1	9	10
		(4.30)	(5.70)	
	30–34	2	13	15
		(6.45)	(8.55)	
	35–39	3	9	12
		(5.16)	(6.84)	
	40–44	5	10	15
		(6.45)	(8.55)	
	45–49	6	7	13
		(5.59)	(7.41)	
	50–54	5	3	8
		(3.44)	(4.56)	
	55–59	13	4	17
		(7.31)	(9.69)	
	60–69	8	2	10
		(4.30)	(5.70)	
	Total	43	57	100

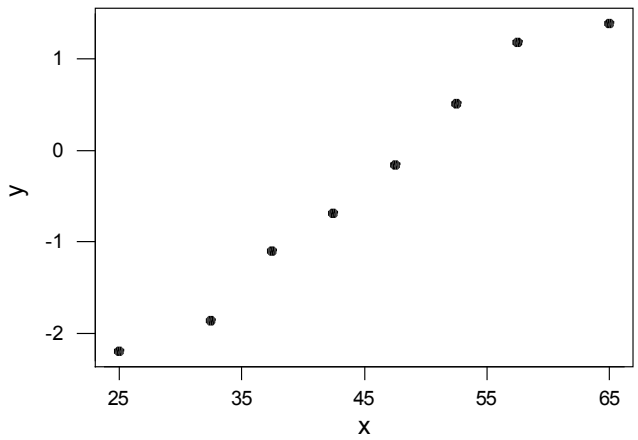
The expected values assuming a constant probability of having coronary heart disease are given in parentheses (= row total $\times\hat{\theta}$).

$$\chi^2 = \sum \frac{(f_i - e_i)^2}{e_i} = \frac{(1-4.30)^2}{4.30} + \ldots + \frac{(2-5.7)^2}{5.7} = 26.6 \text{ on } 7 \text{ d.f.}$$

$\chi^2_7(0.01) = 18.48.$

Strongly reject null hypothesis of constant probability over the different age groups. [Note: If one decides to combine cells to safeguard against very low expected frequencies, one should combine *adjacent* cells.]

(ii) (a)



Linear model seems appropriate, but extremes ($x = 25$ and $x = 65$) are not as good as 32.5–57.5 age range.

(b) $S_{xx} = 17437.5 - \frac{360^2}{8} = 1237.5$

$S_{yy} = 13.615 - \frac{(-2.9392)^2}{8} = 12.535$

$S_{xy} = -9.0429 - \frac{(360)(-2.9392)}{8} = 123.22$

Least squares estimates:

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}} = \frac{123.22}{1237.5} = 0.09957$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = -0.3674 - 0.09957(45) = -4.85$$

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

$$= \left(12.535 - \frac{(123.22)^2}{1237.5} \right) / 6 = 0.04430$$

$\therefore \quad \hat{\sigma} = 0.210$ on 6 d.f.

$$\text{s.e.}(\hat{\beta}) = \frac{\hat{\sigma}}{\sqrt{S_{xx}}} = \frac{0.210}{\sqrt{1237.5}} = 0.0060$$

99% CI for $\hat{\beta} : \hat{\beta} \pm t_6(0.005)$ s.e. $(\hat{\beta})$

$= 0.0996 \pm 3.707\ (0.0060)$

$= 0.0996 \pm 0.0222$ i.e. (0.0774, 0.1218)

(d) In (i)(b), the probability of having coronary heart disease was found to vary with age. The 99% confidence interval for the slope parameter in the regression obtained in (ii)(c) also shows that the probability of having coronary heart disease depends (linearly) on age as zero is not within the interval.