Faculty of Actuaries **Institute of Actuaries**

REPORT OF THE BOARD OF EXAMINERS ON THE EXAMINATIONS HELD IN

April 2002

Subject 101 — Statistical Modelling

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 Forman
Chairman of the Board of Examiners

11 June 2002

1 (i) $P(\text{exactly one}) = (0.95)(0.15)(0.30)+(0.05)(0.85)(0.30)+(0.05)(0.15)(0.70)$
$= 0.04275 + 0.01275 + 0.00525 = 0.06075$

(ii) $P(\text{youngest} \mid \text{exactly one}) = \dfrac{0.04275}{0.06075} = 0.704$

2 (i) $X \sim$ exponential with $\lambda = 1/1000$ and density $\lambda e^{-\lambda x}$.

$P(X > t) = e^{-\lambda t}$ (stated or by integration)

$\therefore P(X > 5000) = e^{-\frac{1}{1000}5000} = e^{-5} = 0.0067$

(ii) $P(X > t_1 \mid X > t_2) = \dfrac{P(X > t_1 \text{ and } X > t_2)}{P(X > t_2)} = \dfrac{e^{-\lambda t_1}}{e^{-\lambda t_2}}$

$\therefore P(X > 5000 \mid X > 1000) = \dfrac{e^{-5}}{e^{-1}} = e^{-4} = 0.0183$

OR: candidates may refer to the memoryless property of the exponential to obtain $P(X > 5000 \mid X > 1000) = P(X > 4000) = e^{-4}$.

3 Let X denote the number of magazine readers in the sample of 200 who are students.

Assuming the magazine's claim is correct, $X \sim$ binomial (200, 0.25).

Using the normal approximation to the binomial distribution,
$X \sim N(50, 150/4)$ approximately:

$$P(X \le 42) \approx P\left(Z < \frac{42.5 - 50}{\sqrt{(150/4)}}\right) \text{ where } Z \sim N(0, 1)$$

$= P(Z < -1.225) = 0.110$ (using continuity correction).

4 Data sum = 82; MLE of μ is $\hat{\mu} = \bar{x} = 82/10 = 8.2$

$\theta = P(X > 9) = P(Z > 9 - \mu)$ where $Z \sim N(0, 1)$

$\therefore$ MLE of $\theta = P(Z > 9 - \hat{\mu}) = P(Z > 9 - 8.2) = P(Z > 0.8) = 0.212$

5 Using binomial model, $P(90 \text{ satisfactory items out of } 100) = \binom{100}{90} p^{90}(1-p)^{10}$

Therefore the log likelihood for p is given by:

$$\log L(p) = 90\log p + 10\log(1-p) + \text{constant}$$

$$\frac{d\log L}{dp} = \frac{90}{p} - \frac{10}{1-p} \Rightarrow \hat{p} = \frac{90}{100} = 0.9 \quad \text{(MLE for } p\text{)}$$

(MLE is sample proportion.)

6 Let n denote the sample size which is determined by the limits of the 99% confidence interval, i.e.

$$2.58 \times \frac{6}{\sqrt{n}} \leq 2$$

$\Rightarrow n \geq (3 \times 2.58)^2 = 59.9$. Therefore n should be at least 60.

7 (i) Using leaves with units of 0.1, the stem and leaf diagrams are:

Group A:	12	1	2	4	4	5	7	7	9	
(Steroid)	13	1	3	6	6	9				
	14	2	2	6						
	15	2	9							
	16									
	17	1	2							

Group B:	12	9	9	9						
(Placebo)	13	0	0	1	1	2	4	5	6	8	9
	14	1	4	4	8					
	15	4	4							
	16									
	17	0								

(ii) The distribution of WBCC is similar for both groups in that the medians are similar (Group A median = (13.3 + 13.6)/2 = 13.45; Group B median = (13.5 + 13.6)/2 = 13.55), and both distributions are slightly positively skewed. There is a slightly greater variability in the steroid group compared to the placebo group. It appears that the steroid treatment has no difference in effect from the placebo.

8 (i) $M(t) = E(e^{tx}) = \sum_{x=0}^{\infty} e^{tx} \binom{k+x-1}{x} p^k q^x$

$= \sum_{x=0}^{\infty} \binom{k+x-1}{x} p^k (qe^t)^x$

$= \left(\frac{p}{1-qe^t}\right)^k \sum_{x=0}^{\infty} \binom{k+x-1}{x} (1-qe^t)^k (qe^t)^x$

$= \left(\frac{p}{1-qe^t}\right)^k$ as the sum equals 1, being the sum of a probability distribution.

Alternative:

$$\sum_{x=0}^{\infty} \binom{k+x-1}{x} p^k (qe^t)^x = p^k \left(1 + k.qe^t + \frac{k(k+1)}{2}.(qe^t)^2 + \ldots\right) = p^k \left(1-qe^t\right)^{-k}$$

(ii) $M(t) = \left(\frac{p}{1-q(1+t+\frac{t^2}{2}+\ldots)}\right)^k = \left(1-\frac{q}{p}(t+\frac{t^2}{2}+\ldots)\right)^{-k}$

$= 1 + k\frac{q}{p}(t+\frac{t^2}{2}+\ldots) + \frac{k(k+1)}{2}(\frac{q}{p})^2(t+\ldots)^2 + \ldots$

coefficient of t: $E(X) = k\frac{q}{p}$

coefficient of $\frac{t^2}{2!}$: $E(X^2) = k\frac{q}{p} + k(k+1)(\frac{q}{p})^2$

mean = $E(X) = \frac{kq}{p}$

variance = $E(X^2) - E^2(X) = k\frac{q}{p} + k(k+1)(\frac{q}{p})^2 - (k\frac{q}{p})^2$

$= k\frac{q}{p} + k(\frac{q}{p})^2 = \frac{kq}{p}(1+\frac{q}{p}) = \frac{kq}{p^2}$

9 (i)

$F_V(t) = P(V \le t) = P(\max\ \{X, Y\} \le t)$

$= P(X \le t \text{ and } Y \le t) = P(X \le t)\ P(Y \le t) = F_X(t)F_Y(t)$ as X and Y are independent

(ii)

$F_W(t) = P(W \le t) = 1 - P(\min\{X, Y\} > t)$
$= 1 - P(X > t \text{ and } Y > t) = 1 - P(X > t)\ P(Y > t)$ as X and Y are independent
$= 1 - [1 - P(X \le t)]\ [1 - P(Y \le t)]$
$= 1 - (1 - F_X(t))\ (1 - F_Y(t))$
$= F_X(t) + F_Y(t) - F_X(t)F_Y(t)$.

(iii)

$F_X(t) = F_Y(t) = 1 - e^{-4t}$, so
$F_W(t) = 1 - e^{-4t} + 1 - e^{-4t} - (1 - e^{-4t})^2$
$= 2 - 2e^{-4t} - 1 - e^{-8t} + 2e^{-4t}$
$= 1 - e^{-8t}$

This is the distribution function of an exponential distribution with parameter 8, and therefore the mean is 1/8.

10 (i) Width of 95% CI $= 2 \times 1.96 \times \{P(1 - P)/200\}^{0.5}$ where P is sample proportion

Max value of $P(1 - P)$ is $0.5^2 = 0.25$

$\therefore$ Max width of CI $= 2 \times 1.96 \times (0.25/200)^{0.5} = 0.139$

In terms of percentages, this is 13.9%.

(ii) $2 \times 1.96 \times (0.25/n)^{0.5} \Leftarrow 0.1 \ \Rightarrow\ n \ge 385$

11 (i) $E(SN) = E[E(SN|N)]$

Now, $E(SN|N = n) = E[(X_1 + \ldots + X_n)n|N = n)] = E[n(X_1 + \ldots + X_n)]$

$= n \times n\mu_X = n^2\ \mu_X$

$\therefore\ E(SN) = E(\mu_X N^2) = \mu_X E(N^2)$

$= \mu_X(\mu_N^2 + \sigma_N^2)$

(ii) $E(S) = E(N\mu_X) = \mu_N\mu_X$

$\therefore \text{Cov}(S,N) = E(SN) - E(S)E(N) = \mu_X(\mu_N^2 + \sigma_N^2) - (\mu_N\mu_X)\mu_N = \mu_X\sigma_N^2$

12 (i) $0 \le \frac{1}{4} - \theta \le 1 \quad \Rightarrow \quad \theta \le \frac{1}{4}, \theta \ge -\frac{3}{4}$

$0 \le \frac{5}{8} + 2\theta \le 1 \quad \Rightarrow \quad \theta \le \frac{3}{16}, \theta \ge -\frac{5}{16}$

$0 \le \frac{1}{8} - \theta \le 1 \quad \Rightarrow \quad \theta \le \frac{1}{8}, \theta \ge -\frac{7}{8}$

combining these $\Rightarrow \quad -\frac{5}{16} \le \theta \le \frac{1}{8}$

(ii)

(a) $\theta = 0.1$: P(down in one period) = $\frac{1}{8} - 0.1 = 0.025$

(b) $\theta = 0$: P(same in two periods) = $[P(\text{same})]^2 = (\frac{5}{8})^2 = 0.391$

(c) $\theta = -0.2$: P(up) = 0.45, P(same) = 0.225

$\therefore p = \frac{4!}{2!2!}(0.45)^2(0.225)^2 = 0.062$

(iii)

(a) (1) $L(\theta) = (\frac{1}{4} - \theta)^{24}(\frac{5}{8} + 2\theta)^{35}(\frac{1}{8} - \theta)^{21}$

$\log L = 24\log(\frac{1}{4} - \theta) + 35\log(\frac{5}{8} + 2\theta) + 21\log(\frac{1}{8} - \theta)$

$\frac{\partial \log L}{\partial \theta} = -\frac{24}{\frac{1}{4} - \theta} + \frac{2(35)}{\frac{5}{8} + 2\theta} - \frac{21}{\frac{1}{8} - \theta}$

equate to zero $\Rightarrow$

$-24(\frac{5}{8} + 2\theta)(\frac{1}{8} - \theta) + 70(\frac{1}{4} - \theta)(\frac{1}{8} - \theta) - 21(\frac{1}{4} - \theta)(\frac{5}{8} + 2\theta) = 0$

$$\Rightarrow \quad 10240\theta^2 - 936\theta - 190 = 0 \quad \Rightarrow \quad 5120\theta^2 - 468\theta - 95 = 0$$

(2) $$\theta = \frac{468 \pm \sqrt{468^2 + 4(5120)(95)}}{2(5120)} = \frac{468 \pm 1471.2661}{10240}$$

$= + 0.189$ or -0.0980

$+ 0.189$ is inadmissible, -0.0980 is admissible

$\therefore MLE \quad \hat{\theta} = -0.0980$

(b) (1) with $\hat{\theta} = -0.0980$, estimated probabilities are

$P(\text{up}) = 0.3480$, $P(\text{same}) = 0.4290$, $P(\text{down}) = 0.2230$

Multiply by 80 for expected frequencies:

up : 27.84, same : 34.32, down : 17.84

(2)

O	e	$(o-e)^2/e$
24	27.84	0.530
35	34.32	0.013
21	17.84	0.560
		$\chi^2 = 1.103$

$\chi^2 = 1.103$ on $(3 - 1 - 1) = 1$ d.f.

this is well below the 5% point for χ^2_1 from tables (3.841)

$\Rightarrow$ large P-value, and so no evidence against the model.

$\therefore$ the model fits these data well.

(c) We need information on the order of the up/same/down's;
and some method of investigating whether the up/same/down's are distributed randomly.

13 (i) $SS_T = 661796 - \dfrac{4426^2}{30} = 8813.47$

$$SS_B = \frac{1}{5}(748^2 + 657^2 + 826^2 + 741^2 + 710^2 + 744^2) - \frac{4426^2}{30} = 3046.67$$

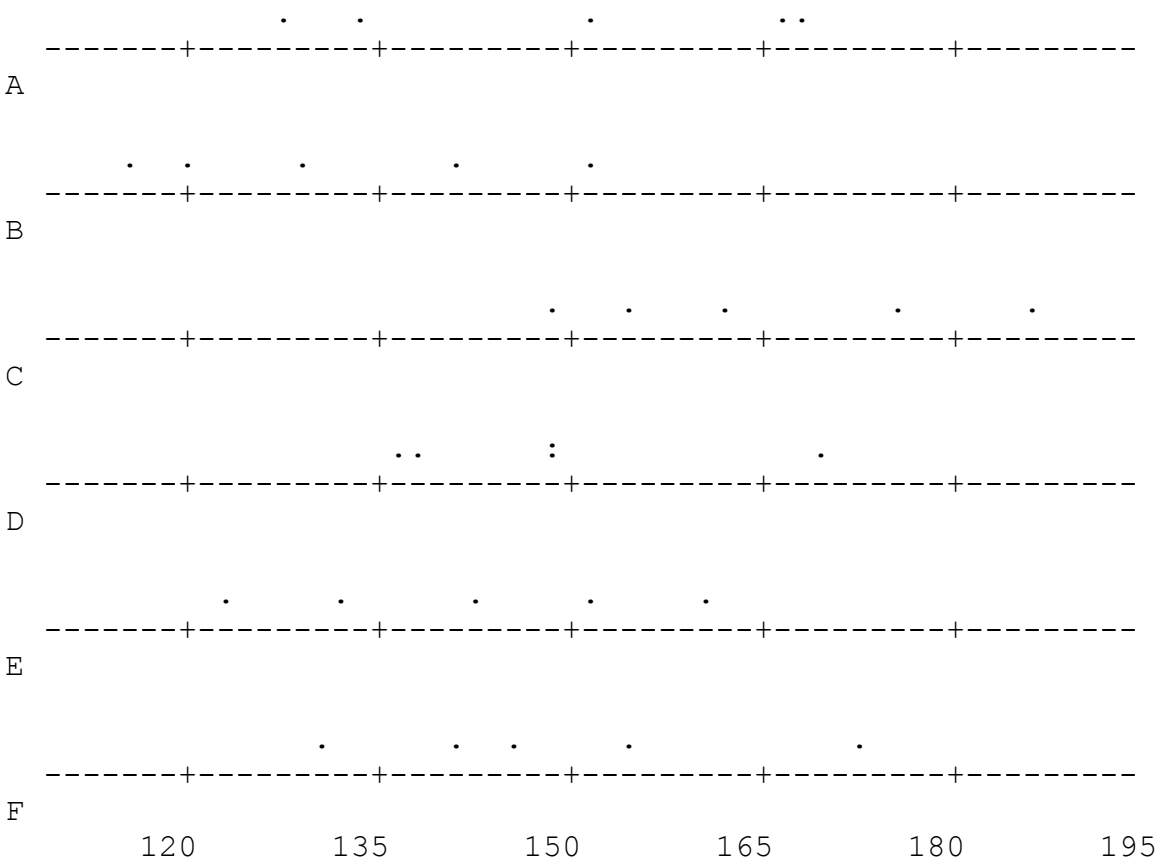
$\therefore\ SS_R = 8813.47 - 3046.67 = 5766.80$

Source	*df*	*SS*	*MS*	*F*
Treatments	5	3046.7	609.3	2.54
Residual	24	5766.8	240.3	
Total	29	8813.5		

From tables $F_{5,24}(5\%) = 2.621$

Observed $F < 2.621$. Therefore <u>not</u> significant at 5% level.

(ii) Assumptions are that, for each company, such premiums are normally distributed with the same variance.



normality and equality of variance both seem reasonable.

(iii) Here $\frac{SS_R}{\sigma^2} \sim \chi^2_{30-6=24}$

$\therefore P(12.40 < \frac{SS_R}{\sigma^2} < 39.36) = 0.95$

$\Rightarrow$ 95% CI for σ^2 is $\frac{SS_R}{39.36} < \sigma^2 < \frac{SS_R}{12.40}$

Data gives $SS_R = 5766.8$

$\Rightarrow$ $146.51 < \sigma^2 < 465.06$

and so 95% CI for σ is $12.1 < \sigma < 21.6$

(iv) (a) Using $\hat{\sigma}^2 = 240.3$ from the ANOVA

For comparing B and C: $t = \frac{165.2 - 131.4}{\sqrt{240.3(\frac{1}{5} + \frac{1}{5})}} = \frac{33.8}{9.80} = 3.45$

$t_{24}(0.5\%) = 2.797$ from tables.

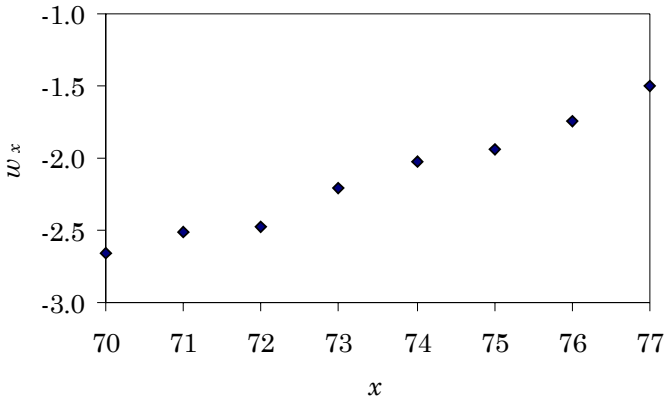
Observed $t > 2.797$. Therefore significant at 1% level (two-sided).

(b) There is no contradiction.

It is wrong to pick out the largest and the smallest of a set of treatment means, test for significance, and then draw conclusions about the set.

Even if H_0 : "μ_i's all equal" is true, the largest and smallest sample means would, of course, differ.

14 (i) Plot



Comments: y_x *v* x relationship is not linear

w_x *v* x relationship appears to be linear, strong

(ii) (a) $S_{xx} = 43260 - 588^2/8 = 42$

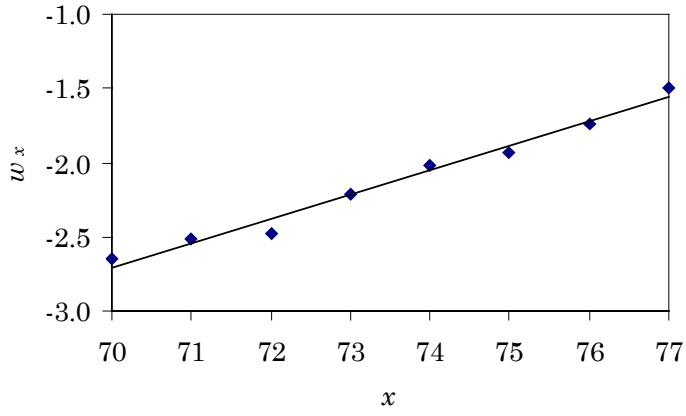
$S_{xw} = -1246.7879 - (588 \times -17.0568)/8 = 6.8869$

$\therefore$ slope $= 6.8869/42 = 0.1640$

so intercept $= -17.0568/8 - (6.8869/42) \times (588/8) = -14.1842$

Fitted line is $w_x = -14.184 + 0.1640x$

(b) Line on plot



(c) $S_{ww} = 37.5173 - ((-17.0568)^2/8) = 1.1505$

$\therefore$ estimate of error variance = $[1.1505 - 6.8869^2/42]/6 = 0.003538$

$\therefore$ standard error of slope coefficient = $(0.003538/42)^{0.5} = 0.00918$

$t_6\,(0.975) = 2.447$

$\therefore$ 95% CI for slope coefficient is $0.1640 \pm (2.447 \times 0.00918)$

i.e. 0.1640 ± 0.0225 or $(0.141, 0.187)$

(d) Fitted value of $w_{71} = -2.5421 \Rightarrow$ fitted value of $y_{71} = \exp(-2.5421) = 0.0787$

$\therefore$ fitted value of $n_{71} = 471 \times 0.0787 = 37.1$

Fitted value of $w_{76} = -1.7222 \Rightarrow$ fitted value of $y_{71} = \exp(-1.7222) = 0.1787$

$\therefore$ fitted value of $n_{76} = 468 \times 0.1787 = 83.6$

(iii) $E(N_x) = E_x bc^x \Rightarrow E(N_x/E_x) = bc^x \Rightarrow E(Y_x) = bc^x$

$\Rightarrow \log E(Y_x) = \alpha + \beta x$ where $\alpha = \log b$, $\beta = \log c$.

The procedure above is a linear regression analysis of $w_x = \log y_x$ on x , which is a simple and approximate approach to fitting the stated model .

Note: The method used in (ii) is based on $E(W_x) = E(\log Y_x) = c + d\,x$, whereas the model stated in (iii) is based on $\log E(Y_x) = c + dx$.