REPORT OF THE BOARD OF EXAMINERS

April 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

3 June 2003

General comments

The Examiners were satisfied with the overall performance on this paper, which was of a comparable standard to those set in recent diets. However, while the percentage of candidates passing was relatively high, few candidates managed to score very high marks.

1 x = number of employees absent

f = number of days

$n = \sum f = 91 \quad \sum fx = 106 \quad \sum fx^2 = 318$

Sample mean $\bar{x} = \dfrac{106}{91} = 1.16$

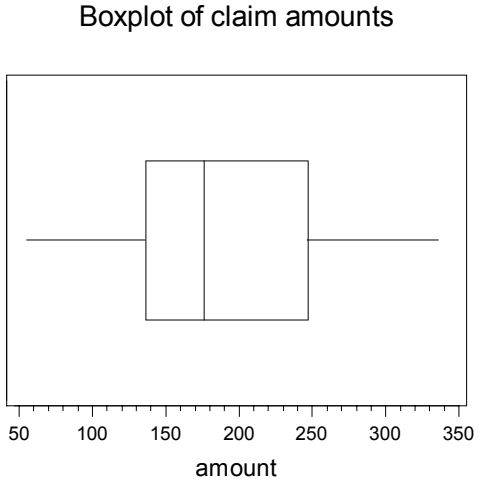
$s^2 = \dfrac{318 - \dfrac{106^2}{91}}{90} = 2.16 \quad \therefore$ Sample s.d. $s = \sqrt{s^2} = 1.47$

Many candidates were unable to makes sense of the frequency distribution of the number of employees absent per day.

2 Ordered: 55 87 112 136 138 159 165 176 192 203 221 253 254 308 336

Median = 8th = 176; Q_1 = 4.25th = 136.5; Q_3 = 11.75th = 245

(*alternatives*: Q_1 = 4th = 136; Q_3 = 12th = 253)



3 P(accident due to faulty brakes | accident attributed to faulty brakes)

$$= \frac{P\left(\text{due to faulty brakes and attributed to faulty brakes}\right)}{P\left(\text{attributed to faulty brakes}\right)}$$

$$= \frac{(0.02)(0.95)}{(0.02)(0.95)+(0.98)(0.01)} = \frac{0.019}{0.0288} = 0.66$$

4 Let X be the number of records with incorrect information.
$X \sim bi(200, 0.13)$
$X \approx N((200)(0.13), 200(0.13)(0.87))$
i.e. $N(26, 22.62)$

$$P(X \le 20)$$

$$\approx P\left(Z < \frac{20.5-26}{\sqrt{22.62}}\right) \text{ where } Z \sim N(0,1), \text{ using continuity correction,}$$

$$= P\left(Z < \frac{-5.5}{\sqrt{22.62}}\right)$$

$$= P\left(Z < -1.16\right) = 1 - 0.88 = 0.12$$

5 (i) $$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$$

$$\approx N\left(n-1, 2\left(n-1\right)\right) \text{ for large } n.$$

(ii) $$P(\frac{100S^2}{\sigma^2} > 110) \approx P(Z > \frac{110-100}{\sqrt{200}} = 0.707) = 1 - 0.76 = 0.24$$

OR: can interpolate in the Yellow Tables (p.169)

6 Approximate CI is "observed proportion ± {1.96 × standard error}"

Maximum value of s.e. is $0.5/\sqrt{1600} = 0.0125$

so maximum width of CI = $2 \times 1.96 \times 0.0125 = 0.049$

7 $\hat{\lambda} = \dfrac{83}{500} = 0.166$

95% confidence interval is: $\hat{\lambda} \pm 1.96\sqrt{\dfrac{\hat{\lambda}}{500}}$

giving $0.166 \pm 1.96\sqrt{\dfrac{0.166}{500}} \quad \Rightarrow \quad 0.166 \pm 0.036 \quad \Rightarrow \quad (0.130, 0.202)$

8 (a) Mean = variance = λ so c.o.v. = $\sqrt{\lambda}/\lambda = 1/\sqrt{\lambda}$
C.o.v. decreases as mean increases

(b) Mean = standard deviation = μ so c.o.v. = 1
C.o.v. is unaffected by increasing the mean

(c) Mean = n, variance = $2n$ so c.o.v. = $(2n)^{1/2}/n = (2/n)^{1/2}$
C.o.v. decreases as mean increases

9 (i) $\hat{p} = \dfrac{146}{200} = 0.73$

The usual two-sided 99% confidence interval for the proportion p would be

$$\hat{p} - 2.58 \times s.e.(\hat{p}) \quad \text{to} \quad \hat{p} + 2.58 \times s.e.(\hat{p})$$

Given the nature of the claim, the appropriate one-sided 99% confidence interval for the proportion p is of the form $(0, p_U)$,

where $p_U = 0.73 + 2.326\sqrt{\dfrac{(0.73)(0.27)}{200}} = 0.73 + 0.073$ giving (0, 0.803)

For percentage: (0, 80.3%)

(ii) 99% confidence interval for the mean μ is

$\bar{x} \pm 2.58\dfrac{s}{\sqrt{n}}$ for large n

$\Rightarrow 112.41 \pm 2.58\dfrac{51.62}{\sqrt{200}} \quad \Rightarrow £112.41 \pm 9.42$ or (£102.99, £121.83)

Note: using 2.576 gives £112.41 ± 9.40 or (£103.01, £121.81)

10 (i) Missing entries are:

d.f. = 27 by subtraction
SS = 10713.5 by subtraction
MS = 396.8 being 10713.5/27.

Observed $F = 5.59$ on 2 and 27 d.f.
From tables the 1% critical point is 5.488

Significant at 1%, so there is quite strong evidence of a difference in the mean claim amounts for the three regions.

(ii) 95% confidence interval for $(\mu_A - \mu_B)$, or equivalently $(\tau_A - \tau_B)$, is

$$(\bar{y}_A - \bar{y}_B) \pm t_{0.025,27}\hat{\sigma}\sqrt{\frac{1}{10}+\frac{1}{10}}$$

giving

$$(147.47 - 154.56) \pm 2.052\sqrt{396.8}\sqrt{\frac{1}{10}+\frac{1}{10}}$$

-7.09 ± 18.28 or $(-25.37, 11.19)$

This comfortably contains zero indicating no difference between the underlying means for regions *A* and *B*.

The significant result of the *F*-test clearly comes from region C mean being much lower than the region A and B means.

11 (i) $M_S(t) = E[e^{tS}] = E[E(e^{tS}|N)]$

$$\begin{aligned} E[e^{tS}|N=n] &= E[\exp\{t(X_1 + X_2 + \ldots + X_n\}|N=n] \\ &= E[\exp\{t(X_1 + X_2 + \ldots + X_n\}] \text{ (since the } X_i\text{'s are independent of } N) \\ &= \Pi\, E[\exp(tX_i)] \text{ (since the } X_i\text{'s are iid)} \\ &= \{M_X(t)\}^n \end{aligned}$$

$$\therefore M_S(t) = E[\{M_X(t)\}^N] = E[\exp\{N\log M_X(t)\}] = M_N\{\log M_X(t)\}$$

Here $M_N(t) = \exp\{\lambda(e^t - 1)\}$ and $M_X(t) = (1 - \mu t)^{-1}$ and so

$$M_S(t) = \exp\left[\lambda\left\{(1-\mu t)^{-1} - 1\right\}\right]$$

(ii) $M_S'(t) = M_S(t) \times \lambda\mu(1 - \mu t)^{-2}$

$$M_S''(t) = M_S(t) \times 2\lambda\mu^2 (1 - \mu t)^{-3} + M_S'(t) \times \lambda\mu(1 - \mu t)^{-2}$$

So $E[S] = M_S'(0) = \lambda\mu$

$E[S^2] = M_S''(0) = 2\lambda\mu^2 + (\lambda\mu \times \lambda\mu) = 2\lambda\mu^2 + \lambda^2\mu^2$

$\therefore V[S] = 2\lambda\mu^2 + \lambda^2\mu^2 - \lambda^2\mu^2 = 2\lambda\mu^2$

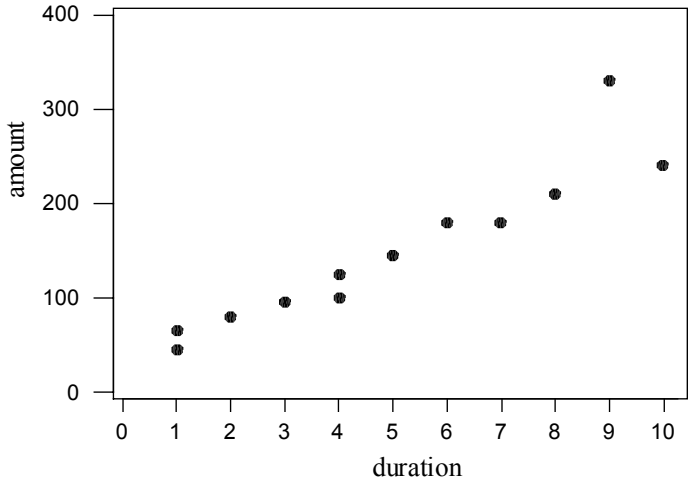
Alternative solution using the cumulant generating function

Let $C_S(t) = \log M_S(t) = \lambda\{(1 - \mu t)^{-1} - 1\}$

$C_S'(t) = \lambda\mu(1 - \mu t)^{-2}$, $C_S''(t) = 2\lambda\mu^2(1 - \mu t)^{-3}$

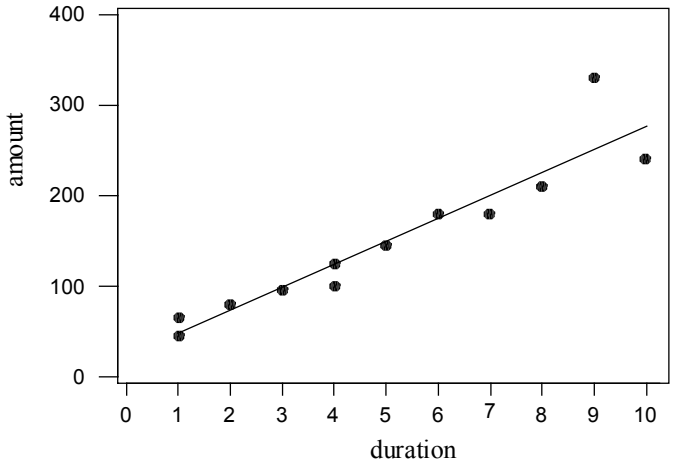
So $E[S] = C_S'(0) = \lambda\mu$ and $V[S] = C_S''(0) = 2\lambda\mu^2$

12 (i) (a)



The point (9,330) is an "outlier" from the general pattern, which is strongly linear.

(b)



(ii) (a) Now $\Sigma x = 51$, $\Sigma x^2 = 321$, $\Sigma y = 1{,}465$, $\Sigma y^2 = 234{,}825$, $\Sigma xy = 8{,}600$

$S_{xx} = 84.545$, $S_{yy} = 39{,}713.636$, $S_{xy} = 1807.727$

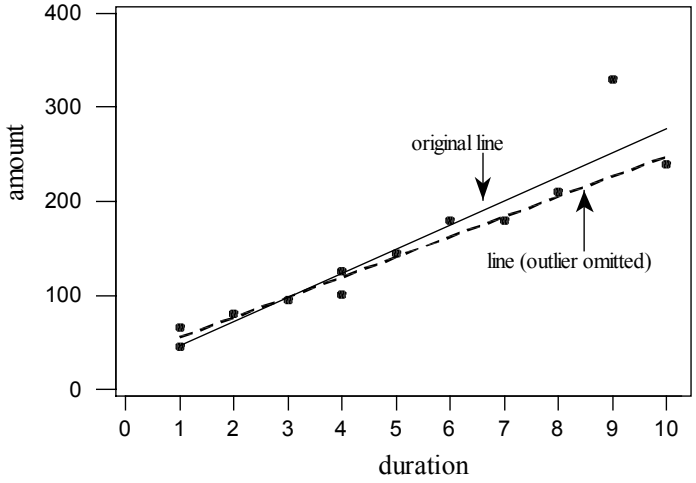
$\Rightarrow \hat{\beta} = 1807.727 / 84.545 = 21.382,\ \hat{\alpha} = 1465/11 - \hat{\beta}(51/11) = 34.048$

Fitted line is $y = 34.0 + 21.4x$

(b) Coefficient of determination $R^2 = 1807.727^2/(84.545 \times 39713.636)$ $= 0.973$ i.e. 97.3%

(or find these by first calculating the three sums of squares SSTOT = 39,713.636 as above, SSREG = $1807.727^2/84.545 = 38652.515$, and so SSRES = 1061.121)

(c)



(d) Removing the influence of the single point (9,330) results in a fitted line with a lower slope and a much better fit for the remaining data (R^2 has increased from 87.8% to 97.3%).

(e) The hourly rate corresponds to the slope in the model

H_0: slope = 25 v H_1: slope $\neq$ 25

Estimate of error variance = 1061.121/9

$\Rightarrow$ standard error of slope estimate = $[(1061.121/9)/84.545]^{1/2} = 1.181$

$t = (21.382 - 25)/1.181 = -3.06$ on 9 df

Upper tail probability is between 0.005 and 0.01 so P-value is between 0.01 and 0.02, so we have quite strong evidence against H_0. We conclude that the data are not consistent with an hourly rate of £25.

13 $f(x) = \dfrac{x^{m-1}\exp(-x/\beta)}{\beta^m \Gamma(m)} \quad (x > 0)$

$$L = \prod_{i=1}^{n} f(x_i) = \frac{\prod_{i=1}^{n} x_i^{m-1} \exp\left(\dfrac{-\sum_{i=1}^{n} x_i}{\beta}\right)}{\beta^{mn}\Gamma^n(m)}$$

(i) m is known case.

(a) $$l = \log L = (m-1)\sum_{i=1}^{n} \log x_i - \frac{\sum_{i=1}^{n} x_i}{\beta} - mn \log \beta - n \log \Gamma(m)$$

$$\frac{\partial l}{\partial \beta} = \frac{\sum_{i=1}^{n} x_i}{\beta^2} - \frac{mn}{\beta} \text{ Then } \frac{\partial l}{\partial \beta} = 0 \Rightarrow \hat{\beta} = \frac{\sum_{i=1}^{n} x_i}{mn} \text{ (MLE)}$$

(b) $$E(\hat{\beta}) = \frac{nE(X_i)}{mn} = \frac{nm\beta}{mn} = \beta \Rightarrow \hat{\beta} \text{ is unbiased}$$

(c) $$\frac{\partial^2 l}{\partial \beta^2} = -2\frac{\sum_{i=1}^{n} x_i}{\beta^3} + \frac{mn}{\beta^2}$$

$$E\left(\frac{\partial^2 l}{\partial \beta^2}\right) = -2\frac{nE(X)}{\beta^3} + \frac{mn}{\beta^2} = -\frac{mn}{\beta^2}$$

$$\therefore \text{ CRlb} = \left[-E\left(\frac{\partial^2 l}{\partial \beta^2}\right)\right]^{-1} = \frac{\beta^2}{mn}$$

(d) Variance of $\hat{\beta}$:

$$V(\hat{\beta}) = \frac{1}{m^2 n^2} V\left(\sum_{i=1}^{n} X_i\right) = \frac{n}{m^2 n^2} V(X) = \frac{nm\beta^2}{m^2 n^2} = \frac{\beta^2}{mn}$$

Cramer-Rao lb is attained.

(ii) m is unknown case.

If m has also to be estimated, ML equations are (approx):

$$\frac{\partial l}{\partial \beta} = \frac{\partial l}{\partial m} = 0$$

$$l = \log L = (m-1)\log\left(\prod_{i=1}^{n} x_i\right) - \frac{\sum_{i=1}^{n} x_i}{\beta} - mn\log\beta$$
$$-n\left[-m + (m-\tfrac{1}{2})\log m + \tfrac{1}{2}\log 2\pi\right]$$

$$\frac{\partial l}{\partial \beta} = \frac{\sum_{i=1}^{n} x_i}{\hat{\beta}^2} - \frac{\hat{m}n}{\hat{\beta}} = 0 \Rightarrow \hat{\beta} = \sum_{i=1}^{n} x_i / n\hat{m} \quad \text{(MLE of } \beta\text{)}$$

$$\frac{\partial l}{\partial m} = \log\left(\prod_{i=1}^{n} x_i\right) - n\log\hat{\beta} + n - n\left(\frac{\hat{m} - \frac{1}{2}}{\hat{m}}\right) - n\log\hat{m} = 0$$

Substituting for $\hat{\beta}$ gives

$$n\log\left(\prod_{i=1}^{n} x_i\right)^{\frac{1}{n}} - n\log\left(\frac{\bar{x}}{\hat{m}}\right) + \frac{n}{2\hat{m}} - n\log\hat{m} = 0$$

$$\therefore\ 2\hat{m} = 1/\log(\bar{x}/\tilde{x}) \quad \text{where} \quad \tilde{x} = \left(\prod_{i=1}^{n} x_i\right)^{\frac{1}{n}},\ \bar{x} = \sum_{i=1}^{n} x_i / n$$

$$\therefore\ \hat{m} = 1/\log\left[\left(\bar{x}/\tilde{x}\right)^2\right] \quad \text{(MLE of } m\text{)}$$

There are alternative versions, for example, $\hat{m} = \dfrac{0.5}{\log\bar{x} - \dfrac{1}{n}\sum\log x_i}$

Most candidates found Question 13(ii) difficult to complete successfully (the Examiners recognise that this part was on the "hard side").

(iii) $\bar{x} = 68.5,\ \tilde{x} = 59.147 \Rightarrow \hat{m} = 3.406,\ \hat{\beta} = 20.114$

14 (i) (a) Suitable table is gender × family size.

Family size 3: no. of girls = 36 + 43(2) + 12(3) = 158
so no. of boys = 300 – 158 = 142, etc.

	family size				
	1	*2*	*3*	*4*	*total*
no. of girls	27	94	158	92	371
no. of boys	23	106	142	108	379
total	50	200	300	200	750

Overall proportion of girls = 371/750 = 0.495

(b) H_0: gender and family size are independent
H_1: gender and family size are not independent

(c) Expected frequency (under H_0) in brackets

27	94	158	92
(24.73)	(98.93)	(148.40)	(98.93)
23	106	142	108
(25.27)	(101.07)	(151.60)	(101.07)

$\chi^2 = (27\text{-}24.73)^2/24.73 + \ldots..$

$= 0.208 + 0.246 + 0.621 + 0.486 +$
$0.203 + 0.241 + 0.608 + 0.476 = 3.088$

df = 3, upper 5% point is 7.815 so *P*-value exceeds 0.05.

(d) We have no evidence against H_0, which can therefore stand, and so we conclude that the proportion of girls is independent of family size.

(ii) (a) Models: no. of girls ~ binomial(n, 0.495) for n = 1, 2, 3, 4

(b) The model assumes that the "trials are independent" i.e. that the gender of each child is independent of that of all other children in the family. We would interpret the lack of fit as evidence that the genders of children in a family are not independently determined.

In addition, the gender of the first child (or the genders of the first and second children) may have an influence on – or even decide - the family size.

Another possible reason is variation in *P*(girl) across families.

The Examiners did not anticipate that so many candidates would be unable to construct the basic 2×4 contingency table appropriate for investigating the matter in question.