EXAMINATIONS

September 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.

M Flaherty
Chairman of the Board of Examiners

23 November 2004

1 Let X be claim size in units of £1000: $X \sim N(6,1)$

$P(X > 7.5|X>6) = P(X > 7.5 \text{ and } X > 6) / P(X > 6) = P(X > 7.5) / P(X > 6)$
$= P(Z > 1.5) / P(Z > 0)$ where $Z \sim N(0, 1)$
$= 0.0668 / 0.5 = 0.134$

2 (i) $C_X(t) = \mu(e^t - 1)$

(ii) $C'_X(t) = \mu e^t \quad \therefore C'_X(0) = \mu$ (mean)
$C''_X(t) = \mu e^t \quad \therefore C''_X(0) = \mu$ (variance)

3 CI is $15.6 \pm \{z \times (\sigma / \sqrt{n})\}$

For a symmetrical 90% interval, $z = 1.6449$

i.e. $15.6 \pm \{1.6449 \times (2/5)\}$ i.e. 15.6 ± 0.66 i.e. 14.94 to 16.26

4 $n = 50$ is large, so Central Limit theorem allows the use of normality

$$\textit{P-value} \doteqdot 2 \times P\left(Z > \frac{207 - 200}{\frac{42}{\sqrt{50}}} \right)$$

$= 2\ P(Z > 1.18) = 2(1 - 0.88) = 0.24$

Examiners' Comment: Some candidates assumed that the claim amounts have a normal distribution. This was not justifiable or necessary. What is required is the approximate normality of the distribution of the sample mean, which is justified for large samples by the central limit theorem.

5 $$\frac{s_1^2}{s_2^2} = \frac{139.7}{76.6} = 1.82$$

$F_{24,29}$ critical value at 5% is 2.154 (two-sided test)

$\therefore$ accept H_0 : equal variances at the 5% level.

Examiners' Comments: A test for equality of variances as asked for here is a two-sided test, but since we always use the observed ratio with the larger sample variance on the numerator, we look at the upper tail of the reference F distribution. For a 5% test we look at the upper 2.5% tail, not the upper 5% tail, as some candidates did.

6 Expected frequencies are all 25

$$\therefore \chi^2 = 4 \times \frac{3^2}{25} = 1.44$$

P-value = $P(\chi_1^2 > 1.44) \approx 1 - 0.77 = 0.23$

$\therefore$ there is no evidence to reject the independence of the two criteria.

Examiners' Comments: Although not covered in the Core Reading, the use of "Yates correction" in this situation (in which the χ^2 statistic has only 1 degree of freedom) is acceptable. Using it gives χ^2 = 1, a P-value of 0.32, *and the same conclusion.*

7 (i) We require *k* such that $\int_0^{100} k(100-x)dx = 1$

$$\int_0^{100} k(100-x)dx = k\left[100x - \frac{x^2}{2}\right]_0^{100} = k\left\{10000 - \frac{10000}{2}\right\} = 5000k$$

$$\therefore k = \frac{1}{5000} = 0.0002$$

[or could be argued geometrically]

(ii) Mean = $\int_0^{100} x(0.0002)(100-x)dx = 0.0002\left[50x^2 - \frac{x^3}{3}\right]_0^{100}$

$$= 0.0002\{500000 - \frac{1000000}{3}\} = 33.33$$

(iii) $P(X > 50) = \int_{50}^{100} 0.0002(100-x)dx = 0.0002\left[100x - \frac{x^2}{2}\right]_{50}^{100}$

$$= 0.0002\{100(50) - \frac{100^2 - 50^2}{2}\} = 0.25$$

[or could be argued geometrically]

(iv) $P(X < 60 \mid X > 50) = \frac{P(50 < X < 60)}{P(X > 50)}$

$$P(50 < X < 60) = 0.0002\left[100x - \frac{x^2}{2}\right]_{50}^{60}$$

$$= 0.0002\{100(10) - \frac{60^2 - 50^2}{2}\} = 0.09$$

$$\therefore P(X < 60 \mid X > 50) = \frac{0.09}{0.25} = 0.36$$

[or could be argued geometrically]

8 $S = \sum_{i=1}^{100} X_i$ has mean 400 and variance 400

By CLT, $S \sim N(400,400)$ approximately

$\therefore P(S > 425) \approx P[Z > (425 - 400)/20] = P(Z > 1.25) = 0.106$

9 (i) (a) $k = 4$ using $N(0, 4/2) = N(0, 2)$
5% point = $0 + 1.6449\sqrt{2} = 2.326$

(b) $k = 40$ using $N(0, 40/38) = N(0, 1.0526)$
5% point = $0 + 1.6449\sqrt{1.0526} = 1.688$

(ii) Exact values are: (a) 2.132 and (b) 1.684
for small df approximation is poor, but for large df it is quite good.

10 (i) $\hat{p}$ is unbiased with variance $\frac{p(1-p)}{20}$ $\therefore$ MSE $= \frac{p(1-p)}{20}$.

Evaluation gives $\frac{0.5(1-0.5)}{20} = 0.0125$

(ii) $\tilde{p}$ has bias $= \frac{20p+1}{21} - p = \frac{1-p}{21}$ and variance $\frac{20p(1-p)}{21^2}$

$\therefore$ MSE $= \frac{20p(1-p)}{21^2} + \frac{(1-p)^2}{21^2}$

Evaluation gives $\frac{20(0.5)(1-0.5)}{21^2} + \frac{(1-0.5)^2}{21^2} = 0.0113 + 0.0006 = 0.0119$

(iii) Even though $\hat{p}$ is the MLE and is unbiased, $\tilde{p}$ is a more efficient estimate (for $p = 0.5$) having a smaller mean square error.

11 (i) Using results for a uniform distribution given in the Yellow Book we obtain the following

$$E(X \mid Y = y) = \frac{1-y}{2}$$

and

$$\text{var}(X|Y = y) = \frac{(1-y)^2}{12}$$

since the conditional distribution of X given $Y = y$ is a continuous uniform distribution with parameters $a = 0$ and $b = 1–y$.

(ii) $$\text{var}(E(X \mid Y)) = \text{var}\left(\frac{1-Y}{2}\right) = \frac{1}{4}\text{var}(Y) = \frac{1}{4}\left(\frac{1}{18}\right) = \frac{1}{72}$$

since Y has a beta distribution with parameters $\alpha = 1$, $\beta = 2$, and the Yellow Book gives $\text{var}(Y) = \frac{1}{18}$.

$$E(\text{var}(X \mid Y)) = E\left(\frac{(1-Y)^2}{12}\right) = \int_0^1 \frac{(1-y)^2}{12} 2(1-y)dy = \frac{1}{6}\left[-\frac{(1-y)^4}{4}\right]_0^1 = \frac{1}{24}.$$

Therefore,

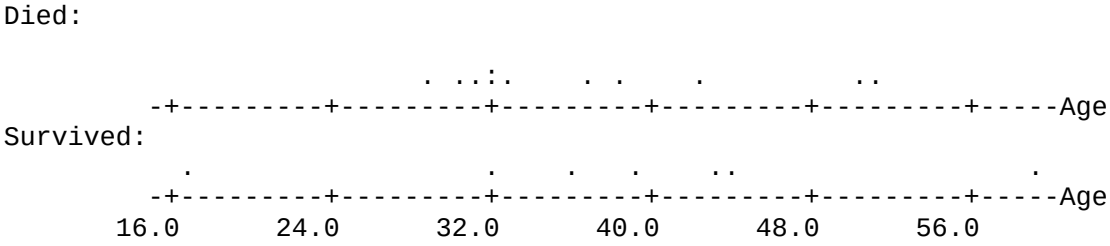
$$\text{var}(E(X \mid Y)) + E(\text{var}(X \mid Y)) = \frac{1}{72} + \frac{1}{24} = \frac{1}{18}.$$

By symmetry with the random variable Y, X has a beta distribution with parameters $\alpha = 1$, $\beta = 2$, and $\text{var}(X) = \frac{1}{18}$.

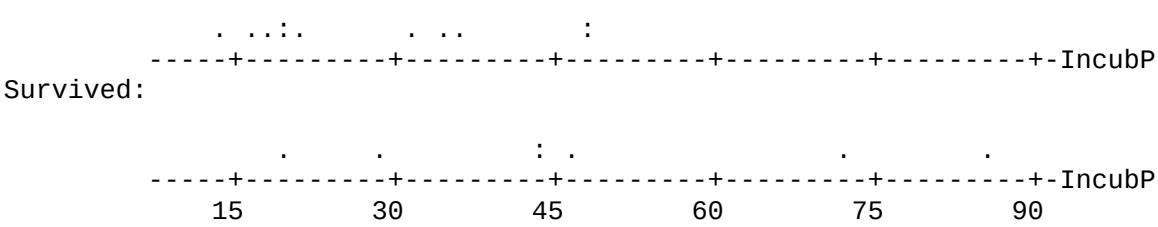
Examiners' Comment: The question helpfully states that the conditional distribution is a uniform distribution and the marginal distribution is a beta distribution. Despite this many candidates failed to quote standard results given in the Yellow Book relating to uniform and beta distributions and instead performed time-consuming integrations.

12 (i) There does not seem to be a clear relationship between age and incubation period either for those subjects who died or for those subjects who survived.

(ii) (a) Dotplots of age for died and survived subjects.



(b) Dotplots of incubation period for died and survived subjects.



The dotplots suggest an association between survival and incubation period (the people who survived tended to have longer incubation periods), but do not suggest an association between survival and age.

(iii) Survived: $n_1 = 7, \bar{y}_1 = 339/7 = 48.429, s_1 = \sqrt{3247.71/6} = 23.266$

Died: $n_2 = 11, \bar{y}_2 = 305/11 = 27.727, s_2 = \sqrt{1578.18/10} = 12.563$

Pooled variance and standard deviation:

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = \frac{6(23.266)^2 + 10(12.563)^2}{16} = 301.633$$

$$s_p = 17.37$$

95% confidence interval:

$$\bar{y}_1 - \bar{y}_2 \pm t_{0.025, n_1+n_2-2} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$= 48.429 - 27.727 \pm (2.120)(17.37)\sqrt{\frac{1}{7} + \frac{1}{11}} = 20.702 \pm 17.804$$

i.e. (2.9, 38.5) hours.

99% confidence interval:

$$\bar{y}_1 - \bar{y}_2 \pm t_{0.005, n_1+n_2-2} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$= 48.429 - 27.727 \pm (2.921)(17.37)\sqrt{\frac{1}{7} + \frac{1}{11}} = 20.702 \pm 24.531$$

i.e. (−3.8, 45.2) hours.

The 95% confidence interval does not contain zero, therefore a two-sided, two sample *t*-test conducted at the 5% significance level would conclude that there is a difference between the means of the two populations.

However, as the 99% confidence interval does contain zero, conducting a more stringent 1% level *t*-test would not reject the null hypothesis that the population means are the same.

(iv) (a) Assuming that the variances of the two populations are equal, $S_1^2 / S_2^2 \sim F_{n_1-1, n_2-1}$.

$s_1^2 / s_2^2 = 23.266^2 / 12.563^2 = 3.430$

The value of the test statistic is below the 5% significance level critical value of $F_{6,10}(2.5\%) = 4.072$. This indicates that there is insufficient evidence to reject the null hypothesis that the two populations have equal variances.

(b) In part (iii) an assumption of normality of each sample was required. The dotplots suggest that this assumption is valid.

Also, the assumption of equal variances of the two groups seems valid from the result of the test conducted in part (iv)(a).

13 (i) $E(Y) = 0 \,.\, P(Y = 0) + \sum_{r=1}^{\infty} r.P(Y = r)$

$= (1 - \alpha) \sum_{r=1}^{\infty} r.P(X = r) = (1 - \alpha)E(X) = (1 - \alpha)\mu$

[Note: there is an alternative solution using conditional expectation]

$E(Y^2) = 0 + \sum_{r=1}^{\infty} r^2.P(Y = r)$

$= (1 - \alpha) \sum_{r=1}^{\infty} r^2.P(X = r) = (1 - \alpha)E(X^2) = (1 - \alpha)(\mu + \mu^2)$

$V(Y) = (1 - \alpha)(\mu + \mu^2) - (1 - \alpha)^2 \mu^2 = (1 - \alpha)\,\mu\,\{1 + \mu - (1 - \alpha)\,\mu\,\}$

$= (1 - \alpha)\,\mu\,(1 + \alpha\mu)$

$E(Y) < E(X)$ as expected since there are more values equal to zero in the adjusted distribution.

The original Poisson has $V(X) = E(X)$. Here, with the extra zeros, we get greater variability relative to the mean, as we see in the fact that $V(Y) > E(Y)$.

(ii) For method of moments we seek α and μ being solutions of

$$\bar{y} = (1-\alpha)\mu$$
$$s^2 = (1-\alpha)\mu(1+\alpha\mu)$$

First equation gives: $\mu = \dfrac{\bar{y}}{(1-\alpha)}$

Substituting into second equation gives: $\dfrac{s^2}{\bar{y}} = 1+\alpha\mu = 1+\alpha\dfrac{\bar{y}}{(1-\alpha)}$

Solving for α gives: $\alpha = \dfrac{s^2 - \bar{y}}{\bar{y}^2 + s^2 - \bar{y}}$

Substituting into expression for μ gives: $\mu = \dfrac{\bar{y}^2 + s^2 - \bar{y}}{\bar{y}}$

(iii) (a) $n = 200$, $\Sigma y = 187$, $\Sigma y^2 = 401$

$$\therefore \bar{y} = \frac{187}{200} = 0.935$$

$$s^2 = \frac{401 - 187^2/200}{199} = 1.1365$$

$$\tilde{\alpha} = \frac{1.1365 - 0.935}{0.935^2 + 1.1365 - 0.935} = 0.1873$$

$$\tilde{\mu} = \frac{0.935^2 + 1.1365 - 0.935}{0.935} = 1.1505$$

(b) $$P(X = 4) = \frac{1.1505^4 e^{-1.1505}}{4!} = 0.0231$$

$$\therefore P(Y = 4) = 0.8127(0.0231) = 0.0188$$

$\therefore$ Exp. freq. = 200(0.0188) = 3.8

Similarly $P(X = 5) = 0.0053$, $P(Y = 5) = 0.0043$,
and exp. freq. = 0.9

By subtraction exp. freq. for $y > 5$ is 200 - 199.9 = 0.1

obs.	90	56	37	12	4	1	0
exp.	88.9	59.2	34.0	13.1	3.8	0.9	0.1

which show good agreement with the observed frequencies so that the model seems to fit the data well.

14 (i) (a) $\Sigma y = 29.12$, $\Sigma y^2 = 70.8744$
$\Rightarrow SS_{TOT} = 70.8744 - 29.12^2/16 = 17.8760$
$\Sigma x = 4 \times 10 = 40$, $\Sigma x^2 = 4 \times 30 = 120 \Rightarrow S_{xx} = 120 - 40^2/16 = 20$
$\Sigma xy = 1 \times 2.73 + 2 \times 6.26 + 3 \times 9.22 + 4 \times 10.91 = 86.55$
$\Rightarrow S_{xy} = 86.55 - 40 \times 29.12/16 = 13.75$
$\Rightarrow$ Regression sum of squares $SS_{REG} = 13.75^2/20 = 9.4531$
$\Rightarrow$ Residual sum of squares $SS_{RES} = 17.8760 - 9.4531 = 8.4229$

(b) $$\hat{\beta} = 13.75/20 = 0.6875$$
$$\hat{\alpha} = 29.12/16 - 0.6875 \times (40/16) = 0.1012$$
Fitted model is $\hat{y} = 0.1012 + 0.6875x$
$y = 0.11$, $x = 1 \Rightarrow$ fitted value = 0.7887

$y = 4.08$, $x = 4 \Rightarrow$ fitted value = 2.8512

(c) $s.e.\left(\hat{\beta}\right) = \left(\dfrac{8.4229/14}{20}\right)^{0.5} = 0.1734$

Under H_0, $P\left(\hat{\beta} > 0.6875\right) = P\left(t_{14} > 0.6875/0.1734\right) = P\left(t_{14} > 3.965\right)$

which is very much lower than 0.005, so *P-value* of test statistic is very much lower than 0.01.

We have strong evidence against the "no linear relationship" hypothesis ($p << 0.01$)

(ii) (a) $SS_{TOT} = 17.8760$

Between companies sum of squares

$SS_B = (2.73^2 + 6.26^2 + 9.22^2 + 10.91^2)/4 - 29.12^2/16 = 9.6709$

$\Rightarrow$ Residual sum of squares $SS_{RES} = 17.8760 - 9.6709 = 8.2051$

(b) $\hat{\mu} = 29.12/16 = 1.82$
$\hat{\tau}_1 = 2.73/4 - 1.82 = -1.1375$, $\hat{\tau}_2 = 6.26/4 - 1.82 = -0.255$
$\hat{\tau}_3 = 9.22/4 - 1.82 = 0.485$, $\hat{\tau}_4 = 10.91/4 - 1.82 = 0.9075$

(c) $y = 0.11$, company A $\Rightarrow$ fitted value = 2.73/4 = 0.6825
$y = 4.08$, company D $\Rightarrow$ fitted value = 10.91/4 = 2.7275

(d) Observed F statistic is (9.6709/3) / (8.2051/12) = 4.715 on 3,12 df

P-value of test statistic is lower than 0.05 (but higher than 0.01)

We have some evidence against the "no company effects" hypothesis ($0.01 < p < 0.05$)

END OF REPORT