EXAMINATION

September 2007

Subject CT6 — Statistical Methods
Core Technical

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 A Stocker
Chairman of the Board of Examiners

December 2007

Comments

Overall it is clear that candidates found this a tougher paper than in other recent sittings. In particular, candidates who did not have a firm grasp of the basic statistical material covered in subject CT3 struggled with some of the questions.

Comments on individual questions are as follows:

Q1 *The Examiners had intended this to be a straightforward bookwork question as it is taken largely verbatim from the core reading. However, very few candidates could recall this part of the core reading accurately and therefore most candidates struggled to score any marks.*

Q2 *Well answered. There was a minor typographical error in the question, but virtually all candidates seemed to have understood what was intended. The Examiners allowed any alternative interpretation provided it was clearly stated.*

Q3 *In general, this question was very poorly answered. Most candidates appeared unable to apply Bayes' Theorem to the situation given in the question.*

Q4 *Well answered.*

Q5 *Parts (i) and (iii) were generally well answered, a pleasing improvement relative to similar questions from recent sittings. Only the better candidates were able to complete part (ii).*

Q6 *This was the first time in a number of sittings that this material has been tested. Most of the better candidates coped reasonably well with parts (i), (ii) and (v). Weaker candidates struggled badly with (i) and (ii) which required only some calculus. Part (iii) was actually simpler than many candidates appear to have expected and as a result many solutions were over-complicated and scored poorly.*

Q7 *Well answered.*

Q8 *This question was well answered by the better candidates. Many candidates picked up significant follow through marks in parts (ii) and (iii) despite making numerical errors in part (i).*

Q9 *Part (i) was generally well answered. The remaining parts were tougher, with a number of candidates quoting the results in parts (iii) and (iv) rather than deriving them as instructed.*

Q10 *Parts (i) and (iii) were well answered. Most candidates were able to make some progress with (ii) (a) though a number struggled to derive three correct equations. Only the best candidates were able to answer (ii) (b).*

1 An ARCH(p) model is

$$X_t = \mu + e_t \sqrt{\alpha_0 + \sum_{k=1}^{p} \alpha_k (X_{t-k} - \mu)^2}$$

where e_t are independent $N(0,1)$.

Often, this is used to model $\ln(Z_t/Z_{t-1})$ where Z_t is the asset price.

It can be seen that a large departure in X_{t-k} from μ will result in X_t having a larger variance. This will then result in a large volatility for the asset price.

2 (i) θ_1 would dominate θ_2 (and vice versa) if the amount of birds poached under each of a_1, a_2 and a_3 is higher in each case for θ_1 (θ_2)

θ_1 provides a better outcome for a_2 ($90 > 0$) and a_3 ($120 > 75$)

θ_2 provides a better outcome for a_1 ($0 < 75$)

(ii) Minimax solution for gamekeepers. Worst loss under each option would be:

$a_1 = \max(0,75,120) = 120$
$a_2 = \max(90,0,90) = 90$
$a_3 = \max(120,75,0) = 120$

Therefore, gamekeepers would choose a_2 and lose 90 birds.

(iii) $a_1 = 0.25 \times 0 + 0.35 \times 75 + 0.4 \times 120 = 74.25$
$a_2 = 0.25 \times 90 + 0.35 \times 0 + 0.4 \times 90 = 58.5$
$a_3 = 0.25 \times 120 + 0.35 \times 75 + 0.4 \times 0 = 56.25$

Hence the Bayes decision is a_3.

3

$$P(N - r = k \,|\, r \text{ big claims}) = P(N = r + k) \times P(of\; r + k \text{ claims } k \text{ are small}) / P(r \text{ big claims})$$

but

$$P(r \text{ big claims}) = \sum_{j=0}^{\infty} e^{-\lambda} \frac{\lambda^{r+j}}{(r+j)!} \times \frac{(r+j)!}{r!\,j!} p^r (1-p)^j$$

$$= \frac{\lambda^r}{r!} p^r \times \sum_{j=0}^{\infty} e^{-\lambda} \frac{\lambda^j}{j!} (1-p)^j$$

$$= \frac{\lambda^r}{r!} p^r e^{-\lambda} e^{\lambda(1-p)}$$

So

$$P(N - r = k \,|\, r \text{ big claims}) = P(N = r + k) \times P(of\; r + k \text{ claims } k \text{ are small}) / P(r \text{ big claims})$$

$$= \frac{e^{-\lambda} \dfrac{\lambda^{r+k}}{(r+k)!} \times \dfrac{(r+k)!}{r!k!} p^r (1-p)^k}{e^{-\lambda} e^{\lambda(1-p)} p^r \lambda^r / r!}$$

$$= e^{-\lambda(1-p)} \frac{\lambda^k (1-p)^k}{k!}$$

which is a probability from a Poisson distribution with parameter $\lambda(1-p)$. Hence conditional mean of $N - r$ is $\lambda(1-p)$.

4 The development factors are:

		Development year			
Accident Year	*0*	*1*	*2*	*3*	*EP*
2003	3,340	3,750	4,270	4,400	4,800
2004	3,670	4,080	4,590		4,900
2005	3,690	4,290			5,050
2006	4,150				5,200
		10,700	7,830	4,270	
		12,120	8,860	4,400	
Development factors		1.1327	1.1315	1.0304	
Cumulative DFs		1.3207	1.1660	1.0304	

Ultimate loss ratio 4,400/4,800 = 0.9167

Estimated ultimate loss for each accident year:

Accident Year	*ULR*	*EP*	*UL*	*Expected claims*	*Claims to be incurred*
2003	0.9167	4,800	4,400	4,400	0
2004	0.9167	4,900	4,492	4,359	133
2005	0.9167	5,050	4,629	3,970	659
2006	0.9167	5,200	4,767	3,609	1,158

Revised ultimate losses are:

Accident Year	*Claims incurred*	*Claims to be incurred*	*Revised UL*
2003	4,400	0	4,400
2004	4,590	133	4,723
2005	4,290	659	4,949
2006	4,150	1,158	5,308
			19,379
Claims to date			15,000
Reserve required			4,379

5 (i) Let X be the size of the first claim, so that X has an exponential distribution with parameter 1. Then for ruin to occur at time t we require $X > U + (1+\alpha)\lambda t$.

$$P(X > U + (1+\alpha)\lambda t) = \int_{U+(1+\alpha)\lambda t}^{\infty} e^{-x}\,dx$$

$$= \left[-e^{-x}\right]_{U+(1+\alpha)\lambda t}^{\infty}$$

$$= e^{-U} e^{-(1+\alpha)\lambda t}.$$

[Note that it would be acceptable to quote the cumulative distribution function for the exponential distribution from the tables rather than calculate the integral]

(ii) Let T denote the time until the first claim. Then T has an exponential distribution with parameter λ and

$$P(\text{Ruin at first claim}) = \int_0^{\infty} P(\text{Ruin at first claim} \mid \text{first claim is at } t) \times f_T(t)\,dt$$

$$= \int_0^{\infty} e^{-U} e^{-(1+\alpha)\lambda t} \lambda e^{-\lambda t}\,dt$$

$$= \int_0^{\infty} e^{-U} \lambda e^{-(2+\alpha)\lambda t}\,dt$$

$$= \left[-e^{-U} \frac{\lambda}{(2+\alpha)\lambda} e^{-(2+\alpha)\lambda t}\right]_0^{\infty}$$

$$= \frac{e^{-U}}{2+\alpha}.$$

(iii) We require

$$\frac{e^{-U}}{2+\alpha} < 0.01$$

i.e. $e^{-U} < 0.01 \times (2 + \alpha)$
i.e. $100e^{-U} < 2 + \alpha$.
i.e. $100e^{-U} - 2 < \alpha$.

6 (i) We find a by solving:

$$\int_0^2 f(x)dx = 1$$

$$\int_0^2 axe^{-x^2} dx = 1$$

$$\left[-\frac{a}{2}e^{-x^2}\right]_0^2 = 1$$

$$\frac{-a}{2}(e^{-4} - 1) = 1$$

$$a = \frac{-2}{e^{-4} - 1} = 2.03731$$

(ii) We find the local maximum value of $f(x)$ by differentiation:

$$f'(x) = ae^{-x^2} - 2ax^2e^{-x^2} = ae^{-x^2}(1 - 2x^2)$$

and this derivative is zero when

$$2x^2 = 1$$
$$x = \pm 0.70711$$

and $f(0.70711) = 0.8738$.

Note that $f(0) = 0$ and $f(2) = 0.07463$ so that the maximum value on [0,2] is 0.8738 and so $|f(x)| < 1$ as required.

(iii) Take as our first point $(2\times 0.7413, 0.4601) = (1.4826, 0.4601)$

Now $f(1.4826) = 0.3353$ which is less than 0.4601 so we reject this point as it lies above the graph of $f(x)$.

Take as our second point $(2\times 0.3210, 0.6316) = (0.6420, 0.6316)$

Now $f(0.6420) = 0.86615 > 0.6316$ so this point lies below the graph of $f(x)$ and is therefore acceptable. Our random sample is therefore the x co-ordinate $= 0.6420$.

(iv) The box $0 < x < 2$, $0 < y < 1$ has area 2, and the area under the curve of $f(x)$ is 1 by definition. Therefore we expect half the points to be rejected as they lie above $f(x)$. Hence it will on average take 4 $U(0,1)$ simulations to determine one point using the acceptance-rejection method.

(v) It would be better to use the same simulated claims to evaluate the re-insurance arrangements

This avoids the possibility that the apparent superiority of one arrangement is in fact due to a favourable series of simulated claims

7 (i) Let $\theta = \exp(-\lambda)$

$\underline{\pi} = \underline{\pi}\, P$

$$P = \begin{bmatrix} 1-\theta & \theta & 0 & 0 \\ 1-\theta & 0 & \theta & 0 \\ 0 & 1-\theta & 0 & \theta \\ 0 & 0 & 1-\theta & \theta \end{bmatrix}$$

$$\begin{aligned}
\pi_0 &= (1-\theta)\pi_0 + (1-\theta)\pi_{25} \\
\pi_{25} &= \theta\pi_0 + (1-\theta)\pi_{40} \\
\pi_{40} &= \theta\pi_{25} + (1-\theta)\pi_{50} \\
\pi_{50} &= \theta\pi_{40} + \theta\pi_{50}
\end{aligned}$$

$\pi_0 + \pi_{25} + \pi_{40} + \pi_{50} = 1$

$$\begin{aligned}
\pi_{25} &= \frac{\theta}{1-\theta}\pi_0 &&= k\pi_0 \\
\pi_{40} &= \frac{\theta}{1-\theta}\pi_{25} &&= k^2\pi_0 \\
\pi_{50} &= \frac{\theta}{1-\theta}\pi_{40} &&= k^3\pi_0
\end{aligned}$$

and hence $\pi_0 + k\,\pi_0 + k^2\,\pi_0 + k^3\,\pi_0 = 1$

hence $\pi_0 = \dfrac{1}{1+k+k^2+k^3}$

hence the average premium paid is

$$500\times\left(\frac{1+0.75k+0.6k^2+0.5k^3}{1+k+k^2+k^3}\right)$$

(ii) (a) $\theta/(1-\theta) = e^{-0.12}/(1-e^{-0.12}) = 7.8433$ Premium = £257.79
(b) $\theta/(1-\theta) = e^{-0.24}/(1-e^{-0.24}) = 3.6866$ Premium = £270.33
(c) $\theta/(1-\theta) = e^{-0.36}/(1-e^{-0.36}) = 2.3077$ Premium = £288.46

(iii) (a) to (b) λ increases by 100% but average premium paid increases only by 4.9%
(b) to (c) λ increases by 50% but average premium paid increases only by 6.7%

The no claims discount system is not effective at discriminating between good and bad drivers.

8 (i) Let the individual loss amounts have distribution X. Then

$$E(X)=\int_0^{100} 0.01333xe^{-0.01333x}dx+100\times P(X>100)$$

$$=\left[-xe^{-0.01333x}\right]_0^{100}+\int_0^{100} e^{-0.01333x}dx+100\int_{100}^{\infty} 0.01333e^{-0.01333x}dx$$

$$=-100e^{-1.333}+\left[-75e^{-0.01333x}\right]_0^{100}+100\left[-e^{-0.01333x}\right]_{100}^{\infty}$$

$$=-100e^{-1.333}-75e^{-1.333}+75+100e^{-1.333}$$

$$=55.2302$$

Hence $E(S)=50\times 55.2302=2761.5$

$$E(X^2) = \int_0^{100} 0.01333x^2 e^{-0.01333x}dx + 100^2 P(X > 100)$$

$$= \left[-x^2 e^{-0.01333x}\right]_0^{100} + \int_0^{100} 2xe^{-0.01333x}dx + 100^2 e^{-1.333}$$

$$= -100^2 e^{-1.333} + \left[-\frac{2x}{0.01333}e^{-0.01333x}\right]_0^{100} + \int_0^{100} \frac{2}{0.01333}e^{-0.01333x}dx + 100^2 e^{-1.333}$$

$$= -\frac{200}{0.01333}e^{-1.333} + \left[-\frac{2}{0.01333^2}e^{-0.01333x}\right]_0^{100}$$

$$= -\frac{200}{0.01333}e^{-1.333} - \frac{2}{0.01333^2}e^{-1.333} + \frac{2}{0.01333^2}$$

$$= 4330.6$$

and so

$$Var(S) = 50 \times 4330.6 = 216529 = (465.33)^2$$

(ii) (a) The normal distribution is $N(2761.5, 465.33^2)$

(b) The Log-Normal distribution has parameters μ and σ with

$$E(S) = e^{\mu+\sigma^2/2}$$
$$Var(S) = e^{2\mu+\sigma^2}(e^{\sigma^2} - 1) = E(S)^2 \times (e^{\sigma^2} - 1)$$

So substituting gives

$$216529 = 2761.5^2 \times (e^{\sigma^2} - 1)$$

$$e^{\sigma^2} = \frac{216529}{2761.5^2} + 1 = 1.028394$$

$$\sigma^2 = \log(1.028394) = 0.027998$$

$$\sigma = 0.167327$$

And now we can substitute for σ to give

$$2761.5 = e^{\mu+0.027998/2}$$

$$\mu = \log(2761.5) - 0.027998/2 = 7.90953$$

(iii) Using the Normal distribution:

$$P(N(2761.5, 465.33^2) > 3000) = P\left(N(0,1) > \frac{3000-2761.5}{465.33}\right)$$

$$= P(N(0,1) > 0.51) = 1 - 0.69497 = 0.30503$$

From tables.

Using the log-normal distribution,

$$P(\log N(7.90953, 0.167327^2) > 3000) = P(N(7.90953, 0.167327^2) > \log(3000))$$

$$= P(N(0,1) > \frac{\log 3000 - 7.90953}{0.167327}).$$

$$= P(N(0,1) > 0.58) = 1 - 0.71904 = 0.28096$$

9 (i) The likelihood is $\prod_{i=1}^{n} f(y_i \mid \mu) = \prod_{i=1}^{n} \frac{\mu^{y_i} e^{-\mu}}{y_i!}$

and hence the log-likelihood is

$$\log\mu \sum_{i=1}^{n} y_i - n\mu - \ldots \quad = \theta \sum_{i=1}^{n} y_i - nb(\theta) + \text{ terms not depending on } \theta$$

where $\theta = \log\mu$
$b(\theta) = e^{\theta}$

(ii) (a) The Pearson residual is

$$\frac{y_i - \hat{y}_i}{\sqrt{\hat{y}_i}}$$

(b) The Pearson residuals are skewed.

This makes it difficult to assess the fit of the model by eye.

(iii) The conjugate prior has the same θ dependence as the likelihood, which is proportional to $\exp\left\{y\theta - e^{\theta}\right\}$. Hence the conjugate prior is $\exp\left\{\alpha\theta - \beta e^{\theta}\right\}$.

$$f(\theta \mid y_1, y_2, \ldots, y_n) \propto f(y_1, y_2, \ldots, y_n \mid \theta)\, f(\theta)$$

$$\propto \exp\left\{\theta \sum_{i=1}^{n} y_i - n e^{\theta}\right\} \exp\left\{\alpha\theta - \beta e^{\theta}\right\}$$

$$\propto \exp\left\{\theta\left(\alpha + \sum_{i=1}^{n} y_i\right) - (\beta + n) e^{\theta}\right\}$$

(iv) For the prior $\log f = \alpha\theta - \beta e^{\theta}$ and $\dfrac{\partial \log f}{\partial \theta} = \alpha - \beta e^{\theta}$. Hence

$E\left[\dfrac{\partial \log f}{\partial \theta}\right] = E\left[\alpha - \beta e^{\theta}\right] = 0$, and so $E\left[e^{\theta}\right] = \dfrac{\alpha}{\beta}$.

For the posterior $E\left[\dfrac{\partial \log f}{\partial \theta}\right] = E\left[\alpha + \sum_{i=1}^{n} y_i - (\beta + n) e^{\theta}\right]$, and hence

$$E\left[e^{\theta} \mid y_1, y_2, \ldots, y_n\right] = \frac{\alpha + \sum_{i=1}^{n} y_i}{\beta + n}.$$

Note that $e^{\theta} = \mu$, and the posterior estimate can be written as

$$\frac{\beta}{\beta + n} \times \frac{\alpha}{\beta} + \frac{n}{\beta + n} \times \frac{\sum_{i=1}^{n} y_i}{n} = Z\frac{\alpha}{\beta} + (1 - Z)\frac{\sum_{i=1}^{n} y_i}{n}$$

ie a combination of the prior estimate and the estimate from the data.

10 (i) The characteristic equation is given by:

$$(1 - \frac{8}{15}\lambda + \frac{1}{15}\lambda^2) = (1 - \frac{1}{3}\lambda)(1 - \frac{1}{5}\lambda) = 0$$

which has roots = 3 and 5. They are both greater than 1. Hence, subject to the initial values having appropriate distributions, this implies (weak) stationarity.

(ii) (a)

Firstly, note that $Cov(X_t, Z_t) = 1$ and $Cov(X_t, Z_{t-1}) = 8/15 - 1/7 = 41/105$

We need to generate 3 distinct equations linking γ_0, γ_1 and γ_2

This can be done as follows:

(A)

$$\begin{aligned}\gamma_0 &= Cov(X_t, X_t) = Cov(1 + 8/15X_{t-1} - 1/15X_{t-2} + Z_t - 1/7Z_{t-1}, X_t)\\ &= 8/15\gamma_1 - 1/15\gamma_2 + 1 - 1/7 \times 41/105\\ &= 8/15\gamma_1 - 1/15\gamma_2 + 694/735\end{aligned}$$

(B)

$$\begin{aligned}\gamma_1 &= Cov(X_t, X_{t-1}) = Cov(1 + 8/15X_{t-1} - 1/15X_{t-2} + Z_t - 1/7Z_{t-1}, X_{t-1})\\ &= 8/15\gamma_0 - 1/15\gamma_1 - 1/7\end{aligned}$$

(C)

$$\begin{aligned}\gamma_2 &= Cov(X_t, X_{t-2}) = Cov(1 + 8/15X_{t-1} - 1/15X_{t-2} + Z_t - 1/7Z_{t-1}, X_{t-2})\\ &= 8/15\gamma_1 - 1/15\gamma_0\end{aligned}$$

Next stage is to solve these equations.

Substituting (C) into (A) gives

$$\gamma_0 = (8/15)\gamma_1 - (1/15)((8/15)\gamma_1 - (1/15)\gamma_0) + 694/735$$

so

$$(224/225)\gamma_0 = (112/225)\gamma_1 + 694/735$$

$$\gamma_0 = (1/2)\gamma_1 + 5205/5488$$

Now substituting into (B) gives

$\gamma_1 = 8/15((1/2)\gamma_1 + 5205/5488) - (1/15)\gamma_1 - 1/7$
so
$(4/5)\gamma_1 = 249/686$
$\gamma_1 = 1245/2744 = 0.4537$

And

$\gamma_0 = 1/2 \times 0.4537 + 5205/5488 = 1.1753$
$\gamma_2 = 8/15 \times 0.4537 - 1/15 \times 1.1753 = 0.1636$

Finally,

$$\rho_0 = 1,\ \rho_1 = \frac{\gamma_1}{\gamma_0} = 0.386,\ \rho_2 = \frac{\gamma_2}{\gamma_0} = 0.139$$

(b) $\rho_k = \frac{8}{15}\rho_{k-1} - \frac{1}{15}\rho_{k-2}$ for $k \geq 2$

We will show that the solution has the form:

$$\rho_k = A\,(\frac{1}{3})^k + B\,(\frac{1}{5})^k$$

Substituting the proposed solution into the recurrence relation gives

$$\frac{8}{15}\rho_{k-1} - \frac{1}{15}\rho_{k-2} = \frac{8}{15}[A\,(\frac{1}{2})^{k-1} + B\,(\frac{1}{5})^{k-1}] - \frac{1}{15}[A\,(\frac{1}{3})^{k-2} + B(\frac{1}{5})^{k-2}]$$

$$= A\,(\frac{1}{3})^k(\frac{8}{15} \times 3 - \frac{1}{5} \times 9) + B\,(\frac{1}{5})^k(\frac{8}{15} \times 5 - \frac{1}{15} \times 25)$$

$$= A\,(\frac{1}{3})^k + B\,(\frac{1}{5})^k$$

$$= \rho_k$$

So the solution does have this form.

The values of A and B are fixed by $\rho_0 = 1$, $\rho_1 = 0.386$

$$\therefore A + B = 1$$

$$\frac{1}{3} A + \frac{1}{5} B = 0.386$$

$$\rightarrow \frac{1}{3} A + \frac{1}{5}(1 - A) = 0.386$$

$$A = 1.395$$

$$B = -0.395$$

$$\therefore P_k = 1.395 \left(\frac{1}{3}\right)^k - 0.395 \left(\frac{1}{5}\right)^k$$

(iii) We require mean and variance of X_t which must be normally distributed since Z is normally distributed.

Variance is $\gamma_0 = 1.1753$ from (ii) (a)

$$E(X_t) = 1 + \frac{8}{15} E(X_t) - \frac{1}{15} E(X_t)$$

$$\therefore E(X_t) = \frac{15}{8}$$

END OF EXAMINERS' REPORT