EXAMINATION

April 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

June 2007

Comments

Comments on solutions presented to individual questions for this April 2007 paper are given below.

Question 1
Generally well answered.

Question 2
This bookwork question was poorly answered and only the best prepared candidates scored well.

Question 3
Relatively few candidates correctly identified that the deductible of 100 applied to all claims

Question 4
Most candidates scored full marks on this question.

Question 5
Candidates generally scored well on this straightforward question

Question 6
This was the most difficult question on the paper. Many answered the bookwork in part (i) but were unable to make a start on part (ii).

Question 7
Poorly answered. Most candidates correctly calculated E(S) but failed to work methodically through the calculation of $E(S_I)$.

Question 8
This question was a good differentiator — whilst weaker candidates struggled with (i)(c) and (ii),, Full credit was also given for those candidates who checked the probability that $T>58.5$ *in (i)(c).*

Question 9
This was well answered.

Question 10
Parts (i) and (ii) of this question were well answered on the whole. Only the best candidates answered part (iii) well.

Question 11
Part (i) of this question was well answered. Only the best candidates scored the second mark in (ii)(b). Although many candidates achieved full marks for the calculation of θ^* *relatively few were awarded both marks for their comments in (iii).*

1 (i) The policyholder must have an interest in the risk being insured, to distinguish between insurance and a wager.

The risk must be of a financial and reasonably quantifiable nature.

(ii) Pecuniary loss — protects against bad debts or other failure of a third party.

Fidelity guarantee — protects against losses caused by dishonest actions of employees.

Business interruption cover — protects against losses made as a result of not being able to conduct business.

2 (i) Two time series X, Y are cointegrated if X and Y are $I(1)$ random processes and there exists a non-zero vector (α, β) such that $\alpha X + \beta Y$ is stationary.

$I(1)$ means that ∇X and ∇Y are stationary
(α, β) is called a cointegrating vector

(ii) Examples:

One of the processes is driving the other
Both are being driven by the same underlying process

3 The claim amount, Z, is given by

$$Z = \begin{cases} 0 & Y < 100 \\ Y - 100 & Y \geq 100 \end{cases}$$

$$E[Z] = \int_{100}^{2000} (x-100) f(x)\, dx + 1900 P(X \geq 2000)$$

$$= \int_{100}^{2000} x\lambda e^{-\lambda x} dx - 100\int_{100}^{2000} \lambda e^{-\lambda x} dx + 1900 e^{-2000\lambda}$$

where $\lambda = \dfrac{1}{1000}$

Hence $E[Z] = \left[-xe^{-\lambda x}\right]_{100}^{2000} + \int_{100}^{2000} e^{-\lambda x}dx - 100\left[-e^{-\lambda x}\right]_{100}^{2000} + 1900e^{-2000\lambda}$

$$= 100e^{-100\lambda} - 2000e^{-2000\lambda} + \left[-\frac{1}{\lambda}e^{-\lambda x}\right]_{100}^{2000} - 100\left(e^{-100\lambda} - e^{-2000\lambda}\right) + 1900e^{-2000\lambda}$$

$$= \frac{1}{\lambda}\left(e^{-100\lambda} - e^{-2000\lambda}\right) = 1000\left(e^{-0.1} - e^{-2}\right) = 769.5\,.$$

4 (i) **Revenue (£)**

	Cautious	*Best estimate*	*Optimistic*
Slots	840,000	1,200,000	1,380,000
Dice	1,680,000	2,400,000	2,760,000
Cards	2,240,000	3,200,000	3,680,000

Costs (£)

	Cautious	*Best estimate*	*Optimistic*
Slots	1,550,000	1,550,000	1,550,000
Dice	1,850,000	1,850,000	1,850,000
Cards	2,450,000	2,450,000	2,450,000

Profit (£)

	Cautious	*Best estimate*	*Optimistic*
Slots	-710,000	-350,000	-170,000
Dice	-170,000	550,000	910,000
Cards	-210,000	750,000	1,230,000

Slots is dominated by Dice and Cards, so could be left out from here on.

(ii)

	Maximum Cost
Slots	-710,000
Dice	-170,000
Cards	-210,000

The minimax decision is Dice.

(iii)

	Expected Profit
Slots	-404,000
Dice	442,000
Cards	606,000

The highest expected profit comes from Cards.

5 The cumulative cost of claims paid is (Figures in £000s):

	Development year		
Accident Year	*0*	*1*	*2*
2004	4,144	4,838	5,021
2005	4,767	5,599	
2006	5,903		

The number of accumulated settled claims is as follows:

	Development year			
Accident Year	*0*	*1*	*2*	*Ult*
2004	581	656	684	684
2005	626	697		727
2006	674			788

Grossing up factors for claim numbers

	Development year		
Accident Year	*0*	*1*	*2*
2004	0.849	0.959	1
2005	0.861	0.959	
2006	0.855		

Average cost per settled claim

Accident Year	*Development year* 0	1	2	*Ult*
2004	7.133	7.375	7.341	7.341
2005	7.615	8.033		7.996
2006	8.758			9.104

Grossing up factors for average claim amounts

Accident Year	*Development year* 0	1	2
2004	0.972	1.005	1.000
2005	0.952	1.005	
2006	0.962		

The total ultimate loss is therefore:

Accident Year	*ACPC*	*Claim Numbers*	*Projected Loss*
2004	7.341	684	5,021
2005	7.996	727	5,813
2006	9.104	788	7,174
			18,008

Claims paid to date	16,523
Outstanding claims	1,485

Assumptions:

Claims fully run-off by end of development year 3.
Projections based on simple average of grossing up factors.
Number of claims relating to each development year are a constant proportion of total claim numbers from the origin year.
Similarly for claim amounts i.e. same proportion of total claim amount for origin year.

6 (i) The requirement to calculate cos and sin is time consuming for a computer. The Polar Method avoids this by using the acceptance-rejection method.

(ii) (a) We must transform the values to a log-normal distribution with the appropriate mean and variance by calculating

$$X = e^{5+2Z}$$

The required value is $Y = \sqrt{X} = \sqrt{915.1} = 30.25$

(b) We require $n > \dfrac{z_{\alpha/2}^2 \tau^2}{\varepsilon^2} = \dfrac{1.645^2 \times 26.3}{1} = 71.17$

Hence, 72 simulations are required.

7

$$\begin{aligned} E(S) &= E(N) \times E(X_1) \\ &= 2 \times (1 \times {}^1\!/_3 + 2 \times {}^2\!/_3) \\ &= {}^{10}\!/_3 \end{aligned}$$

and $S = S_I + S_R$.

We will calculate directly the distribution of S_I.

$$P(S_I = 0) = P(N = 0) = e^{-2} = 0.13534$$

$$P(S_I = 1) = P(N = 1)P(X_1 = 1) = e^{-2}\frac{2^1}{1!} \times \frac{1}{3} = 0.09022$$

$$\begin{aligned} P(S_I = 2) &= P(N = 1)P(X_1 = 2) + P(N = 2)P(X_1 = 1)P(X_2 = 1) \\ &= e^{-2}\frac{2^1}{1!} \times \frac{2}{3} + e^{-2}\frac{2^2}{2!} \times \frac{1}{3} \times \frac{1}{3} \\ &= 0.21052 \end{aligned}$$

$$P(S_I = 3) = 1 - 0.13534 - 0.09022 - 0.21052 = 0.56392$$

$$\begin{aligned} E(S_I) &= 0 \times 0.13534 + 1 \times 0.09022 + 2 \times 0.21052 + 3 \times 0.56392 \\ &= 2.20303 \end{aligned}$$

and hence

$$E(S_R) = E(S - S_I) = E(S) - E(S_I) = {}^{10}\!/_3 - 2.20303 = 1.1303.$$

8 (i) (a) Magnitude of the residuals increases over time, suggesting that the variance is increasing over time.

More positive than negative residuals suggesting there is drift in the process.

(b) If y_i ($i = 1, 2, \ldots, n$) is a sequence of numbers, it has a turning point at time k if either $y_{k-1} < y_k$ and $y_k > y_{k+1}$, or $y_{k-1} > y_k$ and $y_k < y_{k+1}$.

(c) Let T represent the number of turning points, and let N=100 be the number of data points. Then

$$E(T) = 2/3(N-2) = 2/3(100-2) = 65.333$$
$$Var(T) = (16N-29)/90 = 17.45556 = 4.178^2$$
$$P(T \geq 59) \approx P(N(65.333, 4.178^2) > 59.5)$$
$$= P(N(0,1) > \frac{65.333-59.5}{4.178}) = P(N(0,1) > -1.396) = 0.919$$

This is a two-sided test so there is approximately a 16% chance of getting such an extreme number of turning points. This value is not significant at the 5% level, and so this test gives no significant evidence to suggest that the residuals are not a white noise process.

Full credit should also be given for calculating a confidence interval and checking if 59 is in this.

(ii) $X_{100}(1) = 5 + 0.9(X_{100} - 5) + 0 + 0.5e_{99} = 5 + 0.9(7-5) - 0.5 \times 0.7 = 6.45$

$X_{100}(2) = 5 + 0.9(X_{100}(1) - 5) + 0.5e_{100} = 5 + 0.9(6.45 - 5) + 0.5 \times 1.4 = 7.005$

$X_{100}(3) = 5 + 0.9(X_{100}(2) - 5) + 0.5e_{101} = 5 + 0.9(7.005 - 5) = 6.8045$

9 (i)

	£	Starting level (Year 0) 0%	25%	40%
Year 1	Premium if claim in year 0	1,000	1,000	750
	Premium if no claim in year 0	750	600	600
	Saving	250	400	150
Year 2	Premium if claim in year 0	750	750	600
	Premium if no claim in year 0	600	600	600
	Saving	150	150	0
	Claim threshold	400	550	150

(ii) $P(X > 400) = P(\log X > \log 400) = P(Z > (\log 400 - 6)/3.33) = P(Z > -0.0026)$
$= 1 - 0.4990 = 0.5010$

$P(X > 550) = P(\log X > \log 550) = P(Z > (\log 550 - 6)/3.33)$
$= P(Z > 0.0931) = 1 - 0.5371 = 0.4629$

$P(X > 150) = P(\log X > \log 150) = P(Z > (\log 150 - 6)/3.33)$
$= P(Z > -0.2971) = 0.6168$

$0.12 \times 0.5010 = 0.0601$
$0.12 \times 0.4629 = 0.0556$
$0.12 \times 0.6168 = 0.0740$

Hence the transition matrix is

$$\begin{bmatrix} 0.0601 & 0.9399 & 0 \\ 0.0556 & 0 & 0.9444 \\ 0 & 0.0740 & 0.9260 \end{bmatrix}$$

(iii) $\underline{\pi} = \underline{\pi}\, P$

$\pi_0 = 0.0601\,\pi_0 + 0.0556\,\pi_{25}$
$\pi_{25} = 0.9399\,\pi_0 + 0.0740\,\pi_{40}$
$\pi_{40} = 0.9444\,\pi_{25} + 0.9260\,\pi_{40}$
$\pi_{25} = 16.905\,\pi_0$
$\pi_{40} = 215.740\,\pi_0$
$\pi_0 + \pi_{25} + \pi_{40} = 1$

Hence

$\underline{\pi} = (0.0043, 0.0724, 0.9234)$

(iv) Stationary distribution if policyholder claims after a loss

$$P = \begin{bmatrix} 0.12 & 0.88 & 0 \\ 0.12 & 0 & 0.88 \\ 0 & 0.12 & 0.88 \end{bmatrix}$$

$$\begin{aligned} \pi_0 &= 0.12\,\pi_0 + 0.12\,\pi_{25} \\ \pi_{25} &= 0.88\,\pi_0 + 0.12\,\pi_{40} \\ \pi_{40} &= 0.88\,\pi_{25} + 0.88\,\pi_{40} \end{aligned}$$

$$\begin{aligned} \pi_{25} &= 7.333\,\pi_0 \\ \pi_{40} &= 7.333\,\pi_{25} = 53.778\,\pi_0 \\ \pi_0 + \pi_{25} &+ \pi_{40} = 1 \end{aligned}$$

Hence:

$\underline{\pi} = (0.0161, 0.1181, 0.8658)$

(v) Award 1 mark for any sensible comment on the reduction in the number of policyholders in the lower discount categories (or increase in the higher discount categories)

10 (i) (a) $$f(y) = \exp\left[-\frac{y\alpha}{\mu} - \alpha\log\mu + (\alpha-1)\log y + \alpha\log\alpha - \log\Gamma(\alpha)\right]$$

$$= \exp\left[\alpha\left(-\frac{y}{\mu} - \log\mu\right) + (\alpha-1)\log y + \alpha\log\alpha - \log\Gamma(\alpha)\right]$$

which is in the form of an exponential family.

$$\theta = -\frac{1}{\mu}$$

$$b(\theta) = \log\mu$$

$$= \log\left(-\frac{1}{\theta}\right) = -\log(-\theta)$$

(b) The mean and variance of the distribution are $b'(\theta)$ and $a(\varphi)b''(\theta)$

$$b'(\theta) = -\frac{1}{\theta} = \mu$$

which confirms that the mean is μ.

$$b''(\theta) = \frac{1}{\theta^2} = \mu^2$$

$$a(\varphi) = \frac{1}{\alpha}$$

hence the variance is $\frac{1}{\alpha}\mu^2$, as required

(ii) A factor is categorical e.g. male/female.

For a continuous covariate, the value is included. For example, if x is a continuous covariate, a main effect would be $\alpha + \beta x$.

(iii) (a) The linear predictor has the form

$$\alpha_i + \beta_j + \gamma x$$

where α_i is the factor for policyholder gender (i = 1,2)
β_j is the factor for vehicle rating group
x is the policyholder age

$(\alpha_1 = 0, \beta_1 = 0)$

(b) The linear predictor becomes

$$\alpha_i + \beta_j + \gamma_i x$$

11 (i) (a) The likelihood is given by

$$L \propto \theta^{x_1}(1-\theta)\cdots\theta^{x_n}(1-\theta)$$
$$= \theta^{x_1+\cdots x_n}(1-\theta)^n$$

and so the log-likelihood is given by

$$l = LogL = (x_1+\cdots+x_n)\log\theta + n\log(1-\theta)$$

and differentiating gives

$$\frac{dl}{d\theta} = \frac{x_1+\cdots+x_n}{\theta} - \frac{n}{1-\theta}$$

setting this expression to zero to find the maximum gives

$$\frac{x_1+\cdots+x_n}{\hat\theta} - \frac{n}{1-\hat\theta} = 0$$
$$(x_1+\cdots+x_n)(1-\hat\theta) = n\hat\theta$$
$$(x_1+\cdots+x_n) = (n+x_1+\cdots+x_n)\hat\theta$$
$$\hat\theta = \frac{x_1+\cdots+x_n}{n+x_1+\cdots+x_n}$$

to check this is a maximum, note that

$$\frac{d^2l}{d\theta^2} = -\frac{x_1+\cdots+x_n}{\theta^2} - \frac{n}{(1-\theta)^2} < 0$$

(b) The posterior distribution is given by

$$f(\theta|\underline{x}) \propto g(\underline{x}|\theta)f(\theta)$$
$$= \theta^{x_1+\cdots+x_n}(1-\theta)^n \times \theta^{\alpha-1}(1-\theta)^{\alpha-1}$$
$$= \theta^{\alpha+x_1+\cdots+x_n-1}(1-\theta)^{n+\alpha-1}$$

which is the pdf of a beta distribution with parameters $\alpha+x_1+\cdots+x_n$ and $\alpha+n$.

(c) First note that the prior distribution of θ is Beta with parameters α and α. Hence its mean is

$$\frac{\alpha}{\alpha+\alpha} = 1/2.$$

Under Bayesian loss, the estimator is given by the mean of the posterior distribution, which is

$$\theta^* = \frac{\alpha + \sum x_i}{2\alpha + \sum x_i + n} = \frac{\sum x_i}{\sum x_i + n} \times \frac{\sum x_i + n}{2\alpha + \sum x_i + n} + \frac{2\alpha}{2\alpha + \sum x_i + n} \times \frac{1}{2}$$
$$= \hat{\theta} \times Z + (1 - Z)\mu$$

Where $Z = \dfrac{\sum x_i + n}{2\alpha + \sum x_i + n}$ and $\mu = 1/2$ is the prior mean of θ.

(d) As n increases, Z tends towards 1, and the Bayes estimate approaches the maximum likelihood estimate, as more credibility is put on the data, and less on the prior estimate.

(ii) (a) The variance of the prior distribution is given by:

$$\frac{\alpha^2}{(2\alpha)^2(2\alpha+1)} = \frac{1}{4(2\alpha+1)}.$$

(b) Higher values of α result in a lower variance and hence imply greater certainty over the prior value of θ.

In the special case where $\alpha = 1$ the prior distribution is Uniform on [0,1] implying that we have no particular reason to believe that any prior value of θ is more or less likely than any other.

(iii) $\sum x_i = 3 + 3 + 5 = 11$

(a) $\theta^* = \dfrac{5 + 11}{2 \times 5 + 11 + 3} = \dfrac{16}{24} = \dfrac{2}{3} = 0.6667$

(b) $\theta^* = \dfrac{2 + 11}{2 \times 2 + 11 + 3} = \dfrac{13}{18} = 0.7222$

The first set of parameters has greater certainty attached to the prior estimate (i.e. a higher value of α), and therefore the posterior estimate is closer to the mean of the prior distribution (which is 0.5) than in the second case.

END OF MARKING SCHEDULE