Subject CT6 — Statistical Methods. Core Technical

September 2009 examinations

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.

R D Muckart
Chairman of the Board of Examiners

December 2009

Comments for individual questions are given with the solutions that follow.

1

$$\begin{aligned} Y_t &= 2\alpha Y_{t-1} + Z_t \\ Y_t - 2\alpha Y_{t-1} &= Z_t \\ (1-2\alpha B)Y_t &= Z_t \\ Y_t &= (1-2\alpha B)^{-1} Z_t \\ &= (1 + 2\alpha B + (2\alpha B)^2 + (2\alpha B)^3 + \cdots)Z_t \\ &= \sum_{j=0}^{\infty} (2\alpha)^j Z_{t-j} \end{aligned}$$

So $$\begin{aligned} \mathrm{Var}(Y_t) &= \mathrm{Var}\left(\sum_{j=0}^{\infty} (2\alpha)^j Z_{t-j} \right) \\ &= \sum_{j=0}^{\infty} (4\alpha^2)^j \sigma^2 \\ &= \frac{\sigma^2}{(1-4\alpha^2)} \end{aligned}$$

Other valid approaches to deriving the variance were given full credit.

This is a nice, short question involving time series. It requires a little knowledge about series expansions, some algebraic manipulation and the formula for a geometric progression. The question was answered reasonably well, with most candidates recalling the series expansion for $(1-X)^{-1}$. Strong candidates spotted that the condition $|\alpha|<0.5$ was needed to use the formula for an infinite geometric series.

2

Let X_1 denote aggregate claims in year 1, and let X_2 denote aggregate claims in year 2. Then to avoid ruin after the first year, we require X_1<15 and to avoid ruin after 2 years we require $X_1 + X_2$<30.

$$\begin{aligned} P(X_1 < 15 \text{ and } X_1 + X_2 < 30) &= \int_0^{15} f_{X_1}(x) P(X_2 < 30 - x) dx \\ &= \int_0^{15} 0.1e^{-0.1x}(1 - e^{-0.1(30-x)}) dx \\ &= \int_0^{15} 0.1e^{-0.1x} - 0.1e^{-3} dx \end{aligned}$$

$$= \left[-e^{-0.1x} - 0.1xe^{-3} \right]_0^{15}$$

$$= -e^{-1.5} - 1.5e^{-3} + 1$$

$$= 0.702189$$

This was a question involving the concept of ruin in discrete time and requiring candidates to calculate a probability by integrating the pdf of the exponential distribution. This question was not answered well. Many candidates did not recognise the condition for ruin at t=2, ie X_1+X_2<30.

3

(i) Decision d_3 is dominated by d_1 and can be discounted immediately.

(ii) Maximum losses are:

d_1 15

d_2 20

d_4 23

So the minimax solution is to choose d_1

(iii) Expected losses are given by:

$$E(L(d_1)) = 0.4\times10 + 0.25\times15 + 0.35\times5 = 9.5$$
$$E(L(d_2)) = 0.4\times8 + 0.25\times20 + 0.35\times15 = 13.45$$
$$E(L(d_4)) = 0.4\times5 + 0.25\times23 + 0.35\times8 = 10.55$$

So the Bayes solution is also to choose d_1.

A straightforward question involving outcomes of 3 decision functions and requiring the candidates to derive the minimax solution and the Bayes criterion solution. This question was answered very well by most candidates.

4

(i) The likelihood is

$$L = \prod_{i=1}^{k} \prod_{j=1}^{12} \theta_{ij}(1-\theta_{ij})^{y_{ij}}$$

Where y_{ij} is the number of claims on the *i*th policy in the *j*th month.

Taking the logarithm of *L* we have

$$\log L = \sum_{i=1}^{k} \sum_{j=1}^{12} \left[\log\theta_{ij} + y_{ij}\log(1-\theta_{ij}) \right]$$

and so

$$\frac{\partial \log L}{\partial \theta_{ij}} = \frac{1}{\theta_{ij}} - \frac{y_{ij}}{1-\theta_{ij}}$$

And setting the derivative to zero we find $1-\hat{\theta}_{ij} = y_{ij}\hat{\theta}_{ij}$ so that

$$\hat{\theta}_{ij} = \frac{1}{1+y_{ij}}$$

(ii) $P(Y_{ij} = y) = \theta_{ij}(1-\theta_{ij})^y = e^{\log\theta_{ij} + y\log(1-\theta_{ij})}$

The natural parameter is $\log(1-\theta_{ij})$.

(iii) The range of $\alpha + \beta x_j$ is $(-\infty, +\infty)$ which means it is not suitable for modelling parameters $\theta_{ij} \in [0,1]$.

A possible relationship to consider is $\log\dfrac{\theta_{ij}}{1-\theta_{ij}} = \alpha + \beta x_j$.

Other sensible alternatives should be given credit.

A question testing derivation of the m.l.e. of a non-standard p.d.f, with an example application of the theory in the last part. The question was not answered well, particularly part iii).

5

(i) For the given sample

$$\sum_{i=1}^{8} \frac{x_i}{8} = 128.125$$

$$\sum_{i=1}^{8} \frac{x_i^2}{8} = 18,641.125$$

From the tables:

$$E(X) = e^{\mu + \frac{\sigma^2}{2}}$$

$$E(X^2) = e^{\sigma^2} - 1 \; E(X)^2 + E(X)^2 = E(X)^2 e^{\sigma^2}$$

Substituting into the second of these, we have:

$$18,641.125 = 128.125^2 \times e^{\sigma^2}$$

$$\sigma^2 = \log\left(\frac{18,641.125}{128.125^2}\right) = 0.12711274$$

And substituting back into the first expression

$$128.125 = e^{\mu + \frac{\sigma^2}{2}}$$

$$\mu + \frac{\sigma^2}{2} = \log 128.125$$

$$\mu = \log 128.125 - 0.5 \times 0.12711274 = 4.78945$$

(ii) The lower and upper quartile points in the data set are 95 and 168.5

We need to solve:

$$e^{\mu + 0.6745\sigma} = 168.5 \quad \text{and} \quad e^{\mu - 0.6745\sigma} = 95$$

Dividing the first by the second gives:

$$e^{2 \times 0.6745\sigma} = 1.773684211$$

$$\sigma = \frac{\log 1.773684211}{2 \times 0.6745} = 0.424802711$$

$$\sigma^2 = 0.180457343$$

And substituting back into the first equation:

$$e^{\mu + 0.6745\sigma} = 168.5$$

$$\mu = \log 168.5 - 0.6745 \times 0.424802711$$

$$\mu = 4.840406$$

It is possible to use other definitions of upper and lower quartile. Other sensible choices were given full credit provided the subsequent calculations followed through correctly

A numerical question that tested the theory of fitting a distribution using two different methods to sample data. This question was answered reasonably well, with some candidates scoring very highly indeed. This question was a good differentiator with strong candidates showing they had learnt the theory of distribution fitting thoroughly and accurately calculating the answers. NB Both percentile definitions as per CT3 were given credit.

6

(i) Clearly $\hat{\mu} = \frac{13153.32}{500} = 26.30644.$

Using the known expression of the auto covariance function for AR(1) processes: $\rho_k = a_1^k$, we see that

$$a_1^k, = \hat{\rho}_1 = \frac{\sum_{i=1}^{499}(x_i - \overline{x})(x_{i+1} - \overline{x})}{\sum_{i=1}^{500}(x_i - \overline{x})^2} = \frac{2176.03}{3153.67} = 0.6899993$$

Taking the variance of both sides of

$X_t - \mu = a_1(X_{t-1} - \mu) + \varepsilon_t$

and using the fact that $\gamma_0 = \text{var}(X_t - \mu) = \text{var}(X_{t-1} - \mu)$

$\gamma_0 = a_1^2\gamma_0 + \sigma^2.$

Hence $\hat{\sigma}^2 = \hat{\gamma}_0(1 - a(hat)_1^2) = \dfrac{3153.67}{500}(1 - 0.6899993^2) = 3.304416,$

i.e. $\hat{\sigma} = \sqrt{3.304416} = 1.817805$

(ii) Using the fact that under the white noise assumptions the mean and variance of the number of change points are

$\dfrac{2(N-2)}{3} = 332$ and $\dfrac{(16N-29)}{90} = 88.56667$

respectively where $N = 500$. Therefore since the 95% confidence interval is

$$(332 - 1.96 \times \sqrt{88.56667}, 332 + 1.96 \times \sqrt{88.56667}) = (313.6, 350.4)$$

which does not contain the observed number 280, there is a strong evidence that the errors are not close to those of a white noise.

This was a slightly more complicated parameter fitting question for a time-series model, and with a chi-square significance test to finish. The question was answered poorly, with many candidates finding this one tough. Candidates scoring poorly for part i) usually did not get ii) out as well.

7

(i)

a. Note first that

$$P(X = 0) = 1 - q$$
$$P(X = 1) = q(1 - q)$$
$$P(X \geq 2) = 1 - (1 - q) - q(1 - q) = q^2$$
$$P(X \geq 1) = 1 - (1 - q) = q$$

The transition matrix is

$$P = \begin{pmatrix} q & 1-q & 0 \\ q & 0 & 1-q \\ q^2 & q(1-q) & 1-q \end{pmatrix}$$

b. $(kq^2(2-q), kq(1-q), k(1-q)^2)P = (\pi_1, \pi_2, \pi_3)$

Where

$$\pi_1 = kq^3(2-q) + kq^2(1-q) + kq^2(1-q)^2$$
$$= kq^2 \;\; q(2-q) + (1-q) + (1-q)^2$$
$$= kq^2 \;\; 2q - q^2 + 1 - q + 1 - 2q + q^2$$
$$= kq^2(2-q)$$

$$\pi_2 = kq^2(2-q)(1-q) + k(1-q)^2 q(1-q)$$
$$= kq(1-q) \;\; q(2-q) + (1-q)^2$$
$$= kq(1-q)(2q - q^2 + 1 - 2q + q^2)$$
$$= kq(1-q)$$

$$\pi_3 = kq(1-q)^2 + k(1-q)^3$$
$$= k(1-q)^2(q + 1 - q)$$
$$= k(1-q)^2$$

Since the proportions must sum to 1, we have

$$k = \frac{1}{q^2(1-q) + q(1-q) + (1-q)^2} = \frac{1}{2q^2 - q^3 - q + 1}$$

(ii)

a. Average premium is:

$$\text{L}\,350 \times \frac{1}{2\times 0.1^2 - 0.1^3 - 0.1 + 1} \times \;\; 0.1^2 \times 1.9 + 0.65 \times 0.1 \times 0.9 + 0.5 \times 0.9^2$$

= 183.76

b. Policyholders are twice as likely to claim, but the premium increases only by 3%! Suggests that the NCD system is not effective.

A 3x3 NCD problem with generic P(claim) = q, and P(not claim) = 1-q, and then a numerical application. Despite some fiddly algebra, this question was based on standard NCD theory and answered very well.

8

a. The development factors are given by

$R_1 = (136 + 156 + 130) / (96 + 100 + 120) = 1.335443$

$R_2 = (140 + 160) / (136 + 156) = 1.027397$

$R_3 = 168 / 140 = 1.2$

The fully developed table using the chain ladder is below:

Incident year	*0*	*1*	*2*	*3*
2005	96	136	140	168
2006	100	156	160	**192**
2007	120	130	**133.56**	**160.28**
2008	136	**181.62**	**186.60**	**223.92**
R	1.335443	1.027397	1.2	1
f	1.646436	1.232876	1.2	1

Reserve = (168 + 192 + 160.28 + 223.92) – (168 + 160 + 130 + 136) = 150.2

b. B-F method

Estimated loss ratio: 168/175 = 0.96

	2008	*2007*	*2006*	*2005*
F	1.646436	1.232876	1.2	1
1 – 1/f	0.392627	0.188888	0.1666667	0
IUL	188.16	182.4	173.76	168
Emerging liab.IUL(1 – 1/f)	73.87678	34.45325	28.96	0

Reserve is now = 73.87678 + 34.45325 + 28.96 = 137.29

A standard chainladder / Bornhuetter-Ferguson question which candidates answered very well.

9

(i) A distribution is a conjugate prior for an unknown parameter if when used as a prior distribution for that parameter it leads to a posterior distribution which is from the same family.

(ii) $f(p|k) \propto f(k|p) \times f(p)$

$$\propto p^k (1-p)^{n-k} p^{\alpha-1}(1-p)^{\beta-1}$$
$$\propto p^{\alpha+k-1}(1-p)^{\beta+n-k-1}$$

Which is the pdf of a Beta($\alpha+k, \beta+n-k$) distribution.

(iii)

$$E\left(\frac{1}{X}\right) = \int_0^1 \frac{1}{x}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha-1}(1-x)^{\beta-1}dx$$

$$= \int_0^1 \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha-2}(1-x)^{\beta-1}dx$$

$$= \frac{\Gamma(\alpha-1)\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\alpha+\beta-1)}\int_0^1 \frac{\Gamma(\alpha+\beta-1)}{\Gamma(\alpha-1)\Gamma(\beta)}x^{\alpha-2}(1-x)^{\beta-1}dx$$

$$= \frac{\Gamma(\alpha-1)\times(\alpha+\beta-1)\times\Gamma(\alpha+\beta-1)}{(\alpha-1)\times\Gamma(\alpha-1)\times\Gamma(\alpha+\beta-1)}\times 1$$

$$= \frac{\alpha+\beta-1}{\alpha-1}.$$

Other derivations are acceptable.

(iv)Let $h(d) = E(L(d,p))$

$$h(d) = E\left(\frac{(d-p)^2}{p}\right) = E\left(\frac{d^2-2dp+p^2}{p}\right)$$

$$= d^2E(1/p) - 2d + E(p)$$

$$h'(d) = 2dE(1/p) - 2$$

And $h'(d) = 0$

$$2d^*E(1/p) = 2$$

When $d^* = 1/E(1/p) = \dfrac{\alpha+k-1}{\alpha+\beta+n-1}$

Using the result from (iii) applied to the posterior distribution for p.

(v)

$$d^* = \frac{\alpha+k-1}{\alpha+\beta+n-1} = \frac{\alpha-1}{\alpha+\beta-1}\times\frac{\alpha+\beta-1}{\alpha+\beta+n-1} + \frac{k}{n}\times\frac{n}{\alpha+\beta+n-1}$$

$$= \mu(1-Z) + Z\frac{x}{n}$$

Where $Z = \dfrac{n}{\alpha+\beta+n-1}$ and μ is the prior expectation of $1/p$.D

(vi) The estimates are:

Using the given loss function the estimate is

(3 + 2 – 1) / (3 + 3 + 10 – 1) = 4/15 = 0.266666

Using Bayesian loss, we have (3 + 2) / (3 + 3 + 10) = 5/16 = 0.3125.

The mean of the prior is 0.5 and the observed sample mean is 0.2. The loss function in (iv) penalises mis-estimates particularly when the true value of p is lower. This means that the estimate in (iv) is lower than would result from straight quadratic loss.

A longer Bayes question with derivation of a posterior Beta distribution, a credibility factor Z and consideration of a non-standard loss function (given) and the quadratic loss. There was a wide range of quality of answers for this 6 part question. Generally i) and ii) were answered well, iii) to v) proved trickier, in particular deriving d in part v).*

10

(i) $M_S(t) = E(e^{tS})$

$= E(E(e^{t(X_1+X_2+\cdots+X_N)} | N)$

$= E(M_X(t)^N)$ since the X_i are independent and identically distributed

$= E(e^{N \log M_X(t)})$

$= M_N(\log M_X(t))$

$= \exp(\lambda(\exp(\log(M_X(t)-1))))$

$= \exp(\lambda(M_X(t)-1))$

(ii) $M'_S(t) = M_S(t) \times \lambda M'_X(t)$

$E(S) = M'_S(0) = M_S(0) \times \lambda \times M'_X(0)$

$= 1 \times \lambda \times \mu$

$= \lambda\mu$

$M''_S(t) = M'_S(t) \times \lambda \times M'_X(t) + M_S(t) \times \lambda \times M''_X(t)$

$E(S^2) = M''_S(0) = M'_S(0) \times \lambda \times M'_X(0) + M_S(0) \times \lambda \times M''_X(0)$

$= \lambda\mu \times \lambda \times \mu + 1 \times \lambda \times (\sigma^2 + \mu^2)$

$= \lambda^2\mu^2 + \lambda\mu^2 + \lambda\sigma^2$

And so

$$\mathrm{Var}(S) = E(S^2) - E(S)^2$$
$$= \lambda^2\mu^2 + \lambda\mu^2 + \lambda\sigma^2 - \lambda^2\mu^2$$
$$= \lambda(\mu^2 + \sigma^2)$$

(iii) First, we must calculate the mean and variance of a single claim, say Y. Let us denote by X the underlying loss. Then

$$
\begin{aligned}
E(Y) &= \int_0^{200} 0.01xe^{-0.01x}dx + \int_{200}^{\infty} (x+50)\times 0.01\times e^{-0.01x}dx \\
&= \int_0^{\infty} 0.01xe^{-0.01x}dx + 50\int_{200}^{\infty} 0.01e^{-0.01x}dx \\
&= E(X) + 50\times P(X>200) \\
&= 100 + 50\times e^{-200\times 0.01} \\
&= 100 + 6.76676 \\
&= 106.76676
\end{aligned}
$$

$$
\begin{aligned}
E(Y^2) &= \int_0^{200} 0.01x^2e^{-0.01x}dx + \int_{200}^{\infty} (x+50)^2\times 0.01e^{-0.01x}dx \\
&= \int_0^{\infty} 0.01x^2e^{-0.01x}dx + \int_{200}^{\infty} xe^{-0.01x}dx + 50^2\int_{200}^{\infty} 0.01e^{-0.01x}dx \\
&= E(X^2) + \left[-100xe^{-0.01x}\right]_{200}^{\infty} + \int_{200}^{\infty} 100e^{-0.01x}d + 50^2P(X>200) \\
&= 100^2 + 100^2 + 20,000e^{-2} + \left[-100^2e^{-0.01x}\right]_{200}^{\infty} + 2,500e^{-2} \\
&= 20,000 + 20,000e^{-2} + 10,000e^{-2} + 2,500e^{-2} \\
&= 20,000 + 32,500e^{-2} \\
&= 24,398.39671
\end{aligned}
$$

And finally, using the results from part (ii)

$$
\begin{aligned}
E(S) &= 500E(X) = 500\times 106.76676 \\
&= 53,383.38
\end{aligned}
$$

and

$$
\text{Var}(S) = 500E(X^2) = 500\times 24,398.39671 = 12,199,198.36
$$

A long question about deriving mgf of a compound Poisson distribution. Mixture of bookwork and proof, and a numerical application to finish. This question was answered reasonably well, in particular part i) and part ii).

END OF EXAMINERS' REPORT