INSTITUTE AND FACULTY OF ACTUARIES

EXAMINER'S REPORT

September 2011 examinations

Subject CT6 — Statistical Methods
Core Technical

Purpose of Examiners' Reports

The Examiners' Report is written by the Principal Examiner with the aim of helping candidates, both those who are sitting the examination for the first time and who are using past papers as a revision aid, and also those who have previously failed the subject. The Examiners are charged by Council with examining the published syllabus. Although Examiners have access to the Core Reading, which is designed to interpret the syllabus, the Examiners are not required to examine the content of Core Reading. Notwithstanding that, the questions set, and the following comments, will generally be based on Core Reading.

For numerical questions the Examiners' preferred approach to the solution is reproduced in this report. Other valid approaches are always given appropriate credit; where there is a commonly used alternative approach, this is also noted in the report. For essay-style questions, and particularly the open-ended questions in the later subjects, this report contains all the points for which the Examiners awarded marks. This is much more than a model solution – it would be impossible to write down all the points in the report in the time allowed for the question.

T J Birse
Chairman of the Board of Examiners

December 2011

General comments on Subject CT6

The examiners for CT6 expect candidates to be familiar with basic statistical concepts from CT3 and so to be able to comfortable computing probabilities, means, variances etc for the standard statistical distributions. Candidates are also expected to be familiar with Bayes' Theorem, and be able to apply it to given situations. Many of the weaker candidates are not familiar with this material.

The examiners will accept valid approaches that are different from those shown in this report. In general, slightly different numerical answers can be obtained depending on the rounding of intermediate results, and these will still receive full credit. Numerically incorrect answers will usually still score some marks for method providing candidates set their working out clearly.

Comments on the September 2011 paper

The difficulty of this paper was in line with where the Examiner's seek to set the typical CT6 paper. Well prepared candidates were able to score well. Amongst the questions candidates struggled most with were Q's 2 and 6 reflecting a consistent theme across a number of sittings of candidates struggling with Bayes' Theorem and simulation techniques. The questions on time series and ruin theory were answered well, continuing a trend of better answers on these topics.

1 (i) We can see that D4 is dominated by D1.
D3 is not dominated since it gives the best results under θ_1.

D2 is not dominated since it gives the best results under $\theta_3\theta_3$..

D1 is not dominated by D2 since D1 is better under $\theta_2\theta_2\theta_3$.. Similarly D1 is not dominated by D3 since D1 is better under $\theta_3\theta_3\theta_3$..

(ii) The maximum losses are:

D1	30
D2	25
D3	24

So the minimax solution is D3.

This was a straightforward question and the majority of candidates scored well.

2 By Bayes theorem

$$\Pr(p=0.5\mid X=7)=\frac{\Pr(X=7\mid p=0.5)\times\Pr(p=0.5)}{\Pr(X=7\mid p=0.5)\times\Pr(p=0.5)+\int_{0.5}^{1}f(x)\Pr(X=7\mid p=x)dx}$$

And $\int_{0.5}^{1}f(x)\Pr(X=7\,|p=x)dx=\int_{0.5}^{1}0.4\times8\times x^7(1-x)dx$

$$=3.2\left[\frac{x^8}{8}-\frac{x^9}{9}\right]_{0.5}^{1}$$

$$=3.2\left(\frac{1}{8}-\frac{1}{9}\right)-3.2\left(\frac{0.5^8}{8}-\frac{0.5^9}{9}\right)$$

$$\int_{0.5}^{1}f(x)\Pr(X=7\big|p=x)dx=\int_{0.5}^{1}0.4\times8\times x^7(1-x)\,dx$$

$= 0.043576389$

And so

$$\Pr\left(p=0.5|X\right)=\frac{0.8\times8\times0.5^8}{0.8\times8\times0.5^8+0.043576389}=\frac{0.025}{0.025+0.043576389}=0.364557$$

Many candidates struggled to apply Bayes' theorem, and many of those that did struggled with the mixed prior distribution. Candidates found this one of the harder questions on the paper.

3 (i) We must solve

$$\int_0^M 0.01e^{-0.01x}dx=0.8$$

$$[-e^{-0.01x}]_0^M=0.8$$

$$1-e^{-0.01M}=0.8$$

$$M=\frac{\log 0.2}{-0.01}=160.9437912$$

(ii) We have

$$E(X_I) = \int_0^M 0.01x e^{-0.01x} dx + MP(X > M)$$

$$= [-xe^{-0.01x}]_0^M + \int_0^M e^{-0.01x} dx + Me^{-0.01M}$$

$$= -Me^{-0.01M} + \left[\frac{-e^{-0.01x}}{0.01}\right]_0^M + Me^{-0.01M}$$

$$= -100e^{-0.01M} + 100$$

$$= -100e^{-1.6094} + 100 = 80$$

And hence $E(X_R) = E(X) - E(X_I) = 100 - 80 = 20$

This standard question was generally well answered. Alternatively, one could calculate $E(X_R)$ first and then apply $E(X_I) = E(X) - E(X_R)$.

4 (i) Let $S(t)$ denote the total claims up to time t and suppose individual claim amounts follow a distribution X.

Then $U(t) = U + \lambda t(1 + \theta) E(X) - S(t)$.

(ii) $\psi(U, t) = \Pr(U(s) < 0 \text{ for some } s \in [0, t])$

$\psi(U) = \Pr(U(t) < 0 \text{ for some } t > 0)$

(iii) The probability of ruin by time t will increase as λ increases. This is because claims and premiums arrive at a faster rate, so that if ruin occurs it will occur earlier, which leads to an increase in $\psi(U, t)$.

The probability of ultimate ruin does not depend on how quickly the claims arrive. We are not interested in the time when ruin occurs as we are looking over an infinite time horizon.

This is another standard theory question. Many candidates lost marks by not specifying the probabilities carefully enough in part (ii) – for example $\psi(U) = Pr(U(t) < 0)$ does not fully specify the probability since no information is given about t.

5 The likelihood is given by

$$l = C \times \prod_{i=1}^{50} e^{-\frac{1}{2}\left(\frac{y_i - \mu_i}{25}\right)^2}$$

So the log-likelihood is given by

$$L = \log l = D - \frac{1}{1250}\sum_{i=1}^{50}(y_i - \alpha - \beta x_i)^2$$

$$= D - \frac{1}{1250}\left(\sum_{i=1}^{50} y_i^2 - 2\alpha\sum_{i=1}^{50} y_i + 2\alpha\beta\sum_{i=1}^{50} x_i + 50\alpha^2 - 2\beta\sum_{i=1}^{50} x_i y_i + \beta^2\sum_{i=1}^{50} x_i^2\right)$$

We can ignore the factor of 1,250.

$$\frac{\partial L}{\partial \alpha} = 2\sum_{i=1}^{50} y_i - 2\beta\sum_{i=1}^{50} x_i - 100\alpha = 10,984 - 1,274\beta - 100\alpha$$

$$\frac{\partial L}{\partial \beta} = -2\alpha\sum_{i=1}^{50} x_i + 2\sum_{i=1}^{50} x_i y_i - 2\beta\sum_{i=1}^{50} x_i^2 = 149,064 - 1,274\alpha - 16,624\beta$$

Setting both partial derivatives to zero and solving:

$100\alpha + 1,274\beta = 10,984$ (AA)
$-1,274\alpha - 16,624\beta = -149,064$ (BB)

(AA) $\times$ 12.74 + (BB) gives $-393.24\beta = -9,127.84$ so that $\beta = 23.212$
And so $\alpha = 0.01(10,984 - 1274 \times 23.212) = -185.88$

This requires some calculations to produce the mle estimates and only the stronger candidates were able to carry the algebra through to the end. Alternatively, solutions for α and β could also be obtained using the least-squares linear regression expressions given in the tables. This approach gave full credit provided it was accompanied by an explanation of why it produces the same estimates.

6 (i) $F_X(x) = P(X \le x) = P(X = X_1 \cap X_1 \le x) + P(X = X_2 \cap X_2 \le x)$
$= pF_1(x) + (1-p)F_2(x)$

and so $f_X(x) = F_X'(x) = pF_1'(x) + (1-p)F_2'(x) = pf_1(x) + (1-p)f_2(x)$

(ii) We need to combine an algorithm for determining whether to sample from X_1 or X_2 with an algorithm for generating a sample from the appropriate exponential distribution.

If u is generated from a $U(0,1)$ distribution then $F_i^{-1}(u)$ is exponentially distributed with mean $1/\lambda_i$. But $F_i(x) = 1 - e^{-\lambda_i x}$ so that $F_i^{-1} = -\frac{\log(1-u)}{\lambda_i}$

So the algorithm is as follows:

(A) Generate u_1 and u_2 from $U(0,1)$

(B) If $u_1 < p$ then set $i = 1$ otherwise set $i = 2$.

(C) Set $x =$ $x = -\frac{\log 1 - u_2}{\lambda_i}$ $-\frac{\log(1-u_2)}{\lambda_i}$

(iii) The algorithm will be as follows:

(A) Generate u_1 and $\mathrm{u}_2 u_2$ from $U(0,1)$

(B) Set $q_0 = 0, q_j = p_1 + ... + p_j$ for $j = 1, 2, \ldots, k$

(C) If $q_{j-1} \le u_1 < q_j$ then set $i = j$.

(D) Set $x = -\frac{\log(1-u_2)}{\lambda_i}$ $x = -\frac{\log 1 - u_2}{\lambda_i}$

A number of candidates struggled to generate the correct algorithm. Some attempted to use the inversion method in parts (ii) and (iii) but the method shown above is much easier.

7 (i) Let the loss amount be X. Then

$E(X) = 0.8 \times 100 + 0.2 \times 115 = 103$

$E(X^2) = 0.8 \times (100^2 + 400) + 0.2\,(115^2 + 900) = 11{,}145$

$\mathrm{Var}(X) = E(X^2) - E(X)^2 = 11{,}145 - 103^2 = 536$

(ii) No, the loss distribution is not Normal. To see this, note that (for example) the pdf of the combined distribution will have local maxima at both 100 and 115. [Consider the case where the variances are very small to see this]

(iii) $\Pr(X > 130) = 0.8 \times \Pr(N(100, 20^2) > 130) + 0.2 \times \Pr(N(115, 30^2) > 130)$

$= 0.8 \times \Pr\left(N(0,1) > \frac{130-100}{20}\right) + 0.2 \times \Pr\left(N(0,1) > \frac{130-115}{30}\right)$ $\Pr(X > 130) = 0.8 \times \Pr(N(100, 20^2) > 130) + 0.2 \times \Pr(N(115, 30^2) > 130)$

$= 0.8 \times \Pr(N(0,1) > 1.5) + 0.2 \times \Pr(N(0,1) > 0.5)$
$= 0.8 \times (1 - 0.93319) + 0.2 \times (1 - 0.69146)$

$= 0.115156$

(iv) The relevant proportion is given by:

$$\frac{0.2 \times (1 - 0.69146)}{0.115156} = 53.6\%$$

Many weaker candidates struggled with this question, with a large number incorrectly asserting the loss distribution was Normal in part (ii).

8 (i) The model is ARIMA(1,0,1) if Y_t is stationary.

(ii) (a) The characteristic polynomial for the AR part is $A(z) = 1 - 0.4z$ the root of which has absolute value greater than 1 so the process is stationary.

(b) The characteristic polynomial for the MA part is $B(z) = 1 + 0.9z$ the root of which has absolute value greater than 1 so the process is invertible.

(iii) Since the process is stationary we know that $E(Y_t)$ is equal to some constant μ independent of t.

Taking expectations on both sides of the equation defining Y_t gives

$E(Y_t) = 0.1 + 0.4E(Y_{t-1})$

$\mu = 0.1 + 0.4\mu$

$\mu = \dfrac{0.1}{1 - 0.4} = 0.1666666$

Note that

$\text{Cov}(Y_t, e_t) = \text{Cov}(0.1 + 0.4Y_{t-1} + 0.9e_{t-1} + e_t, e_t)$

$= 0.4\text{Cov}(Y_{t-1}, e_t) + 0.9\text{Cov}(e_{t-1}, e_t) + \text{Cov}(e_t, e_t) = 0 + 0 + \sigma^2 = \sigma^2$
Similarly

$\text{Cov}(Y_t, e_{t-1}) = 0 + 0.4\text{Cov}(Y_{t-1}, e_{t-1}) + 0.9\text{Cov}(e_{t-1}, e_{t-1}) + \text{Cov}(e_t, e_{t-1})$
$= 0.4\sigma^2 + 0.9\sigma^2 + 0 = 1.3\sigma^2$

So

$\gamma_0 = \text{Cov}(Y_t, Y_t) = \text{Cov}(Y_t, 0.1 + 0.4Y_{t-1} + 0.9e_{t-1} + e_t)$

$= 0.4\gamma_1 + 0.9 \times 1.3\sigma^2 + \sigma^2 = 0.4\gamma_1 + 2.17\sigma^2$ (A)

And

$\gamma_1 = \text{Cov}(Y_{t-1}, Y_t) = \text{Cov}(Y_{t-1}, 0.1 + 0.4Y_{t-1} + 0.9e_{t-1} + e_t)$

$= 0.4\gamma_0 + 0.9\sigma^2$ (B)

Substituting for γ_1 in (A) gives

$\gamma_0 = 0.4 \times 0.4\gamma_0 + 0.4 \times 0.9\sigma^2 + 2.17\sigma^2 = 0.16\gamma_0 + 2.53\sigma^2$

$$\gamma_0 = \frac{2.53}{0.84}\sigma^2 = 3.011905\sigma^2$$

Substituting into (B) gives

$\gamma_1 = 0.4 \times 3.011905\sigma^2 + 0.9\sigma^2 = 2.104762\sigma^2$

And in general

$\gamma_s = 0.4\gamma_{s-1}$ for $s \geq 2$

So $\gamma_s = 0.4^{s-1} \times 2.104762\sigma^2$.

(iv) We have $(1 - 0.4B)\, Y_t = 0.1 + 0.9e_{t-1} + e_t$

so $Y_t = (1 - 0.4B)^{-1}\,(0.1 + 0.9e_{t-1} + e_t)$

$$= \sum_{i=0}^{\infty} 0.4^i B^i (0.1 + 0.9e_{t-1} + e_t)$$

$$= \frac{0.1}{1-0.4} + 0.9\sum_{i=0}^{\infty} 0.4^i e_{t-i-1} + \sum_{i=0}^{\infty} 0.4^i e_{t-i}$$

$$= 0.16667 + e_t + 1.3\sum_{i=1}^{\infty} 0.4^{i-1} e_{t-i}$$

Overall, this time series question was reasonably well answered, consistent with the improvement in the standard of answers to this type of question in recent sittings. Weaker candidates could not generate the correct auto-covariance function here.

9 (i) $M_X(t) = E(e^{tX}) = \int_0^\infty e^{tx} f(x)dx$

$= \int_0^\infty 0.01^2 xe^{(t-0.01)x} dx$

$= \left[\dfrac{0.01^2 xe^{(t-0.01)x}}{t-0.01}\right]_0^\infty - \int_0^\infty \dfrac{0.01^2 e^{(t-0.01)x}}{t-0.01} dx$

$= 0 - 0 - \left[\dfrac{0.01^2 e^{(t-0.01)x}}{(t-0.01)^2}\right]_0^\infty$ provided that $t < 0.01$

$= \dfrac{0.01^2}{(t-0.01)^2}$ again provided that $t < 0.01$

(ii) The adjustment coefficient is the unique positive solution of

$M_X(R) = 1 + 1.45E(X)R$

But $E(X) = M'_X(0) = \dfrac{d}{dt}\left[\dfrac{0.01^2}{(t-0.01)^2}\right]\Bigg|_{t=0}$

$= \dfrac{-2\times 0.01^2}{(t-0.01)^3}\Bigg|_{t=0} = \dfrac{-2}{-0.01} = 200$

So we need to solve $\dfrac{0.01^2}{(R-0.01)^2} = 1 + 290R$

i.e. $0.01^2 = (1 + 290R)(R - 0.01)^2 = (1 + 290R)(0.01^2 - 0.02R + R^2)$

i.e. $0.012 = 0.01^2 + 0.029R - 0.02R - 5.8R^2 + R^2 + 290R^3$

i.e. $290R^2 - 4.8R + 0.009 = 0$

$R = \dfrac{4.8 \pm \sqrt{4.8^2 - 4\times 290\times 0.009}}{2\times 290}$

i.e. $R = 0.00215578$ or $R = 0.0143959$

So taking the smaller root we have $R = 0.00215578$ since that is less than 0.01

The upper bound for the probability of ruin is given by Lundberg's inequality as

$\psi(U) \le e^{-RU} = e^{-0.00215578U}$

(iii) We want $\psi(U) \le e^{-0.00215578U} \le 0.01$

i.e. $-0.00215578U \le \log 0.01$
i.e. $U \ge \dfrac{\log 0.01}{-0.00215578} = 2136.20$

(iv) This time the adjustment coefficient is the solution to:

$e^{200R} = 1 + 290\mathrm{R}$

So the question is whether $y = e^{200R}$ crosses the line $y = 1 + 290R$ before or after $y = 0.01^2(0.01 - R)^{-2}$ crosses the same line
But when $R = 0.00215578$ we have
$e^{200R} = e^{200\times0.00215578} = 1.539 < 1 + 290R = 1.625.$

So $y = e^{200R}$ has not yet crossed the given line, and the second scenario has a larger adjustment coefficient that the first.

This means the second risk has a lower probability of ruin, which is to be expected since although the mean claim amounts are the same in each scenario, the claim amounts in the first scenario are more variable suggesting a greater risk.

This was found one of the more challenging questions on the paper. In part (i), the final expression could be quoted from the tables but for full marks candidates had to show it from the definitions. Special care is needed here in calculations as decimal places of R can affect the final figures.

10 (i) The development factors are:

$r_{0,1} = \dfrac{134+180+149}{120+140+135} = \dfrac{463}{395} = 1.172151899$

$r_{1,2} = \dfrac{146+185}{134+180} = \dfrac{331}{314} = 1.054140127$

$r_{2,3} = \dfrac{148}{146} = 1.01369863$

The ultimate claims are therefore:

For AY2008: $185 \times 1.01369863 = 187.53$

For AY2009: 149 × 1.05414027 × 1.01369863 = 159.22
For AY2010: 138 × 1.172151899 ×1.054140127 × 1.01369863 = 172.85

So the outstanding claim reserve is

187.53 + 159.22 + 172.85 – 185 – 149 – 138 = 47.60

(ii) The individual development factors are as follows:

	Development Factor		
Accident Year	*0 to 1*	*1 to 2*	*2 to 3*
2007	1.1167	1.0896	1.0137
2008	1.2857	1.0278	
2009	1.1037		
Max	1.2857	1.0896	1.0137

The ultimate claims are therefore:

For AY2008: 185 × 1.0137 = 187.53
For AY2009: 149 × 1.0896 × 1.0137 = 164.57
For AY2010: 138 × 1.2857 × 1.0896 × 1.0137 = 195.97

So the outstanding claim reserve is

187.53 + 164.57 + 195.97 – 185 – 149 – 138 = 76.07

This is an increase of 28.47 which is 59.8% higher.

(iii) Selecting the maximum DF in each column increases the reserves by 60%. Better to take a weighted average of each column as per usual chain ladder approach, UNLESS we know something in particular why we should give full credence to the 1.286 factor (which is much larger than the other two factors in column 2/1) and the 1.09 factor (which is much larger than the 1.028 factor in column 3/2)

This question was well answered. Some candidates dropped marks in part (iii).

11 (i) We need to find the parameters of the Gamma distribution, say α and λ. Then

$$\frac{E(X)}{Var(X)} = \frac{\alpha/\lambda}{\alpha/\lambda^2} = \lambda = \frac{50}{25} = 2$$

And hence $\alpha = E(X) \times \lambda = 50 \times 2 = 100$

The posterior distribution is given by:

$f(\theta_1|x) \propto f(x|\theta_1)\, f(\theta_1)$

$$\propto \left(\prod_{j=1}^{5} e^{-\theta_1}\theta_1^{n_{1j}}\right) \times \theta_1^{\alpha-1} e^{-\lambda\theta_1}$$

$$\propto e^{-(\lambda+5)\theta_1}\theta_1^{\alpha+\sum_{j=1}^{5} n_{1j}-1}$$

Which is the pdf of a gamma distribution with parameters $\alpha + \sum_{j=1}^{5} n_{1j} = 100 + 240 = 340$ and $\lambda + 5 = 7$.

Under quadratic loss the Bayes estimate is the mean of the posterior distribution. So we have an estimate of $\frac{340}{7} = 48.57$.

(ii) We have $\bar{n}_1 = \frac{240}{5} = 48$ and $\bar{n}_2 = \frac{260}{5} = 52$ and $\bar{n}_3 = \frac{144}{5} = 28.8$.

This gives $\bar{n} = \frac{48+52+28.8}{3} = 42.9333$

$$\sum_{j=1}^{5}(n_{1j}-\bar{n}_1)^2 = \sum_{j=1}^{5} n_{1j} - 2\sum_{j=1}^{5} n_{1j} \times \bar{n}_1 + 5 \times \bar{n}_1^2$$

$$= 12{,}144 - 2 \times 240 \times 48 + 5 \times 48^2 = 624$$

Similarly

$$\sum_{j=1}^{5}(n_{2j}-\bar{n}_2)^2 = 13{,}934 - 2 \times 260 \times 52 + 5 \times 52^2 = 414$$

$$\sum_{j=1}^{5}(n_{3j}-\bar{n}_3)^2 = 4{,}354 - 2 \times 144 \times 28.8 + 5 \times 28.8^2 = 206.8$$

So

$E(s^2(\theta)) = \frac{1}{3} \times \frac{1}{4}(624 + 414 + 206.8) = 103.733$

and

$$\text{Var}(m(\theta)) = \tfrac{1}{2}((48 - 42.9333)^2 + (52 - 42.9333)^2 + (28.8 - 42.9333)^2) - \tfrac{1}{5} \times 103.7333 = 133.06667$$

So $Z = \dfrac{5}{5 + \dfrac{103.733}{133.06667}} = 0.86512$

So expected claims for next year are:

Cat 1 $0.13488 \times 42.9333 + 0.86512 \times 48 = 47.32$
Cat 2 $0.13488 \times 42.9333 + 0.86512 \times 52 = 50.78$
Cat 3 $0.13488 \times 42.9333 + 0.86512 \times 28.8 = 30.71$

This question contained a minor typographical error in the summary statistics. Based on the figures given in the question a direct calculation of $\sum_{j=1}^{5} n_{1j}^2$ *gives the correct figure 12,114 and not 12,144 which is given in the question. Candidates who used 12114 will have produced slightly different results as follows:*

$$\sum_{j=1}^{5}\left(n_{1j} - \overline{n_1}\right)^2 = 594$$
$$E(s^2(\theta)) = 101.2333$$
$$Var\left(S\left(\theta\right)\right) = 133.5666$$
$$Z = 0.86837$$

And the final three figures will change from 47.32, 50.78, 30.71 to 47.33, 50.81 and 30.66 respectively. Candidates producing these figures scored full marks.

(iii) The main differences are that:

- The approach under (i) makes use of prior information about the distribution of θ_1 whereas the approach in (ii) does not.

- The approach under (i) uses only the information from the first category to produce a posterior estimate, whereas the approach under (ii) assumes that information from the other categories can give some information about category 1.

- The approach under (i) makes precise distributional assumptions about the number of claims (i.e. that they are Poisson distributed) whereas the approach under (ii) makes no such assumptions.

(iv) The insurance policies were newly introduced 5 years ago, and it is therefore likely that the volume of policies written has increased (or at least not been constant) over time. The assumption that the number of claims has a Poisson distribution with a fixed mean is therefore unlikely to be accurate, as one would expect the mean number of claims to be proportional to the number of policies.

Let P_{ij} be the number of policies in force for risk i in year j. Then the models can be amended as follows:

The approach in (i) can be taken assuming that that the mean number of claims in the Poisson distribution is $P_{ij}\theta_i$.

The approach in (ii) can be generalised by using EBCT Model 2 which explicitly incorporates an adjustment for the volume of risk.

This long question was answered well generally. A bit of care was needed in the final two parts where only the better candidates were able to give a full discussion of the assumptions underlying the models and how the models could be amended.

END OF EXAMINERS' REPORT