INSTITUTE AND FACULTY OF ACTUARIES

EXAMINERS' REPORT

April 2013 examinations

Subject CT6 – Statistical Methods
Core Technical

Introduction

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 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. The Examiners have access to the Core Reading, which is designed to interpret the syllabus, and will generally base questions around it but are not required to examine the content of Core Reading specifically or exclusively.

For numerical questions the Examiners' preferred approach to the solution is reproduced in this report; other valid approaches are given appropriate credit. For essay-style questions, particularly the open-ended questions in the later subjects, the report may contain more points than the Examiners will expect from a solution that scores full marks.

The report is written based on the legislative and regulatory context pertaining to the date that the examination was set. Candidates should take into account the possibility that circumstances may have changed if using these reports for revision.

D C Bowie
Chairman of the Board of Examiners

July 2013

General comments on Subject CT6

The examiners for CT6 expect candidates to be familiar with basic statistical concepts from CT3 and so to be 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 April 2013 paper

The examiners had felt that this paper contained a slightly greater proportion of more routine questions than previous papers and this was backed up by some good solutions to most of the questions. There was a marked improvement from previous sessions on topics such as Bayes' Theorem (Q3) and Ruin Theory (Q9).

1 Using the inverse transform method we need to set:

$$u = 1 - e^{-cx^{\gamma}}.$$

i.e.

$$-cx^{\gamma} = \log(1-u).$$

i.e.

$$x = \left(\frac{\log(1-u)}{-c}\right)^{\frac{1}{\gamma}}.$$

Using the equation above with the parameters $c = 0.002$ and $\gamma = 1.1$ we get:

u = 0.238 gives x = 86.96

u = 0.655 gives x = 300.73

This routine question was well answered, although a few candidates struggled with the algebra.

2 We need to solve:

$$1 - e^{-240\lambda} = 0.75$$

$$e^{-240\lambda} = 0.25$$

so

$$\lambda = \frac{\log(0.25)}{-240} = 0.005776$$

and so the mean is $\dfrac{1}{0.005776} = 173.12.$

This straightforward question was well answered.

3 Let L be the state getting up late and let M be the state of getting up on time.

Let Z be the number of minutes late.

According to Bayes' theorem:

$$P(L|Z>20)=\frac{P(Z>20|L)P(L)}{P(Z>20)}$$

but

$$P(Z>20|L)=e^{-\frac{20}{15}}=0.263597138$$

and

$$P(Z>20)=P(Z>20|L)P(L)+P(Z>20|M)P(M)$$

$$=0.263597138\times 1/3+0.2\times 2/3=0.221199046$$

and so

$$P(L|Z>20)=\frac{0.263597138\times 1/3}{0.221199046}=0.3972\,.$$

This question was well answered by most candidates however weaker candidates were unable to apply Bayes' Theorem.

4 (i) A game with 2 players where whatever one player loses in the game the other player wins, and vice versa.

(ii) (a)

Value to Sally		Sally	
		10	**40**
Fiona	**10**	20	–50
	40	–50	80

Sally chooses 10 with probability p.

Then for the expected payoffs to be equal regardless of Fiona's choice we must have:

$$20p - 50(1-p) = -50p + 80(1-p)$$

so $\quad 200p = 130$

so $\quad p = 0.65$

(b) This strategy is optimal for Sally because it produces the same expected payoff regardless of what Fiona does. Under any other randomized strategy Fiona can adopt a strategy that minimizes Sally's expected payoff.

(c) Value = 20 * 0.65 – 50 * 0.35 = –4.5.

This question was generally well answered, although a few candidates were thrown by a less familiar application of decision theory.

5 First accumulate claims:

Cumulative Claims		*Development year* 0	1	2
	2010	2,328	3,812	4,196
Accident year	*2011*	1,749	2,937	
	2012	2,117		

DY1 = (3,812 + 2,937) / (2,328 + 1,749) = 1.655 384
DY2 = 4,196 / 3,812 = 1.100 735

Now complete lower half of table:

Cumulative Claims		*Development year* 0	1	2
	2009	2,328	3,812	4,196
Accident year	*2011*	1,749	2,937	3,232.86
	2012	2,117	3,504.45	3,857.47

So estimated amount of outstanding claims is:

(3,232.86 – 2,937) + (3,857.47 – 2,117) = 2,036.3.

Most candidates scored full marks on this straightforward application of chain ladder theory.

6 (i) We have:

$$M_S(t) = M_N(\log M_X(t)) = e^{\lambda(M_X(t)-1)}$$

Let us work with the cumulant generating function:

$$C_S(t) = \log M_S(t) = \lambda M_X(t) - \lambda.$$

The third central moment is given by $C_S'''(0)$.

Now:

$$C_S'''(t) = \lambda M_X'''(t)$$

and so

$$C_S'''(0) = \lambda M_X'''(0) = \lambda m_3.$$

Hence the coefficient of skewness is given by:

$$\frac{E((S-E(S))^3}{(Var(S))^{3/2}} = \frac{\lambda m_3}{(\lambda m_2)^{\frac{3}{2}}}.$$

(ii) Since *X* takes only positive values we have $m_3 = E\left(X^3\right) > 0.$

Both λ and $m_2 = E(X^2)$ are also always positive.

This means the coefficient of skewness is always positive.

(iii) Re-writing the equation for the coefficient of skewness we have:

$$\frac{\lambda m_3}{(\lambda m_2)^{\frac{3}{2}}} = \frac{m_3}{\lambda^{0.5} m_2^{1.5}} \to 0 \text{ as } \lambda \to \infty.$$

Hence the distribution of *S* tends to symmetry as $\lambda \to \infty$.

Well prepared candidates who knew their bookwork were able to answer this question well, however weaker candidates struggled with part (i) and gave unconvincing answers to part (ii) & (iii).

7 (i) From the definition of the gamma density given in the question

$$f(y) = \frac{\alpha^{\alpha}}{\mu^{\alpha}\Gamma(\alpha)} y^{\alpha-1} e^{-\frac{\alpha}{\mu}y}$$

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

$$= \exp\left[\frac{(y\theta - b(\theta))}{a(\varphi)} + c(y,\varphi)\right]$$

where:

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

$$\varphi = \alpha$$

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

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

$$c(y,\varphi) = (\varphi-1)\log y + \varphi\log\varphi - \log\Gamma(\varphi).$$

Hence the distribution has the right form for a member of an exponential family.

The natural parameter is $-{}^{1}\!/\!_{\mu}$. The canonical link function is $\frac{1}{\mu}$.

(ii) Using the information given, we can calculate the deviance differences and compare that with the differences of the degrees of freedom for each of the nested models. If the decrease in the deviances is greater than twice the difference in degrees of freedom this suggests an improvement.

Model	*Scaled Deviance*	*Degrees of freedom*	*Difference in scaled deviance*
1	900	12	
Age	789	10	111
Age +location	544	7	245
Age * location	541	1	3

From the table we can see that the interaction model does not indicate any improvement hence the recommended model would be **Age +location**.

Again well prepared candidates were able to score highly on this question, however weaker candidates dropped marks as a result of not specifying a full parameterisation in part (i). In part (ii) full credit was given to candidates who used the chi-squared test rather than the approximation set out above..

8 (i) (a) $E(N) = E[E(N|\mu)]$

$= E[\mu] = 2/8 = 0.25$

(b) $\text{var}(N) = E[\text{var}(N|\mu)] + \text{var}[E(N|\mu)]$

$= E[\mu] + \text{var}[\mu]$

$= 2/8 + 2/8^2 = 0.28125$

(ii) Let Y be aggregate claims from one policy.

Individual claim is gamma with $\alpha = 2$ and $\lambda = 0.001$.

$$E(Y) = E(X)E(N) = 2000 \times 0.25 = 500.$$

$$Var(Y) = E(N)Var(X) + Var(N)E(X)^2$$

$$= 0.25 \times 2000000 + \tfrac{9}{32} \times 2000^2 = 1,625,000.$$

So the mean and variance of total claims are 500,000 and 1,625,000,000 respectively.

(iii) Our approximate distribution for S is $S \sim N(500,000 , 1625000000)$.

$$P(S > 550000) = P\left(Z > \frac{550000 - 500000}{\sqrt{1625000000}}\right) = P(Z > 1.24035) = 0.1074.$$

(iv) The prob three years in a row is $0.1074^3 = 0.00124$.

The probability of this happening is very low. It is more likely that the insurance company's belief about the distribution of claims amounts is incorrect.

The normal approximation tails off quickly and so underestimates the probability of extreme events

Part (i) was straightforward, however some candidates failed to show sufficient working to gain full marks. A surprising number of candidates were unfamiliar with the standard bookwork underlying part (ii). Credit was given for any sensible comments in part (iv).

9 (i) $\psi(U)$ does not depend on λ. This parameter affects the speed with which the process runs, but does not affect the ultimate probability of ruin.

$\psi(U)$ is higher for higher values of μ since the significance of the starting capital falls as μ rises, providing proportionately less of a buffer.
$\psi(U)$ is lower for higher values of θ since the higher θ is the higher the premiums with no change to claim amounts, so that there is a larger buffer against ruin.

$\psi(U)$ is lower for higher values of *U* since the higher *U* is the higher the larger the buffer against ruin given by the initial capital.

(ii) The adjustment coefficients are the solutions to:

$$M_X(R) = 1 + 200 \times 1.3 \times R = 1 + 260R$$

for the various choices of the moment generating function.

Our first task is to find the parameters in the gamma distribution in C.

Denoting these by α and β we have:

$$\frac{\alpha}{\beta} = 200 \text{ and } \frac{\alpha}{\beta^2} = 800$$

Dividing the second by the first we get $1/\beta = 4$ so $\beta = 0.25$ and $\alpha = 50$.

Solving for R_B we have:

$$\frac{0.005}{0.005 - R_B} = 1 + 260R_B.$$

$$1 = (1 + 260R_B)(1 - 200R_B).$$

$$1 = 1 + 60R_B - 52{,}000R_B^2$$

$$R_B = \frac{60}{52{,}000} = 0.001153846.$$

Consider the three functions:

$$A = e^{200R} - 1 - 260R$$

$$B = \frac{0.005}{0.005 - R} - 1 - 260R$$

$$C = \left(\frac{0.25}{0.25 - R}\right)^{50} - 1 - 260R.$$

We can tabulate the values of these functions as follows:

R	**A**	**B**	**C**
0.0001	–0.00579	–0.00559	–0.00579
0.0012	–0.04075	0.003789	–0.0400

So the second function has changed sign, but the first and third have not which gives the required result.

(iii) We know that R_C satisfies:

$$\left(\frac{0.25}{0.25 - R_C}\right)^{50} = 1 + 260R_C.$$

We can re-write this as:

$$\left(\frac{0.25 - R_C}{0.25}\right)^{-50} = 1 + 260R_C.$$

$$\frac{1}{(1 - \frac{R_C}{0.25})^{50}} = 1 + 260R_C.$$

$$\frac{1}{\left(1 - \frac{200R_C}{50}\right)^{50}} = 1 + 260R_C.$$

But due to the approximation given in the question, the denominator of the left hand side is approximately e^{-200R_C}.

So we have, approximately:

$$\frac{1}{e^{-200R_C}} = 1 + 260R_C.$$

i.e. $\quad e^{200R_C} = 1 + 260R_C.$

Which is the equation satisfied by R_A. Hence R_C and R_A are approximately equal.

Most candidates scored well in part (i), although many simply stated how the probability of ruin changes without explaining why. Well prepared candidates scored well on part (ii), noting the method in previous examinations for finding the root of the equation; however very few candidates scored well on part (iii) which was stretching.

10 (i) The likelihood function is given by:

$$L \propto \prod_{i=1}^{5} e^{-\lambda}\lambda^{x_{1,i}} \prod_{i=1}^{5} e^{-2\lambda}(2\lambda)^{x_{2,i}} \prod_{i=1}^{5} e^{-5\lambda}(5\lambda)^{x_{3,i}}.$$

Where $x_{j,i}$ is the number of claims on the jth type in the ith year.

The log likelihood is given by:

$$l = \log L = C - 5\lambda - 10\lambda - 25\lambda + (\log\lambda)\sum_{i,j} x_{j,i}.$$

$$= C - 40\lambda + 804(\log\lambda).$$

Differentiating gives:

$$\frac{dl}{d\lambda} = -40 + \frac{804}{\lambda}.$$

and setting this equal to zero gives:

$$\hat{\lambda} = \frac{804}{40} = 20.1.$$

This is a maximum since:

$$\frac{d^2 l}{d\lambda^2} = -\frac{804}{\lambda^2} < 0.$$

(ii) The mean number of claims for the various types are:

$$\bar{X}_1 = 16.4 \text{ and } \bar{X}_2 = 40.6 \text{ and } \bar{X}_3 = 103.8.$$

With overall mean $\bar{X} = 53.6$.

So we have parameter estimates:

$$E\left(m(\theta)\right) = \bar{X} = 53.6.$$

$$E\left(s^2(\theta)\right) = \frac{1}{3}\sum_{i=1}^{3}\left[\frac{1}{4}\sum_{j=1}^{5}\left(X_{ij} - \bar{X}_i\right)^2\right]$$

$$= \frac{1}{12}\left(139.2 + 417.2 + 2322.8\right) = 239.9333333.$$

$$Var\left(m(\theta)\right) = \frac{1}{2}\sum_{i=1}^{3}\left(\bar{X}_i - \bar{X}\right)^2 - \frac{1}{5}E(s^2(\theta))$$

$$= 0.5\left[(16.4-53.6)^2 + (40.6-53.6)^2 + (103.8-53.6)^2\right] - 0.2\times 239.93333333$$

$$=1988.4533333.$$

And so:

$$Z = \frac{5}{5+\dfrac{E(s^2(\theta))}{Var(m(\theta))}} = \frac{5}{5+\dfrac{239.9333333}{1988.4533333}} = 0.976436003.$$

and the expected claims from the three types are:

Type	*Credibility Premium*
1	0.976436002 × 16.4 + 0.023563998 × 53.6 = 17.3
2	0.976436002 × 40.6 + 0.023563998 × 53.6 = 40.9
3	0.976436002 × 103.8 + 0.023563998 × 53.6 = 102.6

(iii) The corresponding estimates based on our computed $\hat{\lambda}$ are 20.1, 40.2 and 100.5.

The estimates are remarkably similar. The biggest difference is for type 1 buildings, where the maximum likelihood estimate gives a lower weight to the data from that risk, but the credibility estimate gives greater weight.

(iv) The main limitation is that the model in (ii) does not take account of the volume of buildings covered, which will probably vary from year to year.

Again well prepared candidates found this question relatively straightforward. Weaker candidates were unable to construct the likelihood function in part (i). A disappointing number of candidates were unable to accurately render the standard formulae in part (ii).

11 (i) $X_t - \alpha X_{t-1} = e_t \sim N(0,\sigma^2)$.

So $X_t \,|\, X_{t-1} \sim N(\alpha X_{t-1}, \sigma^2)$

and so the likelihood is given by:

$$L \propto \prod_{i=1}^{n} P(X_i = x_i \,|\, x_{i-1}) \times P(x_0)$$

$$L \propto \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i - \alpha x_{i-1})^2}{2\sigma^2}} \times 1$$

(ii) We can see that maximising the likelihood with respect to α is the same as minimising the expression:

$$L \propto \sigma^{-n} e^{-\frac{\sum_{i=1}^{n}(x_i - \alpha x_{i-1})^2}{2\sigma^2}}$$

$$\sum_{i=1}^{n}(x_i - \alpha x_{i-1})^2 .$$

(iii) The log-likelihood is given by:

$$l = logL = -n \log \sigma - \frac{1}{2\sigma^2} \sum_{i=1}^{n}(x_i - \alpha x_{i-1})^2 + \text{Constant}$$

Differentiating with respect to α gives:

$$\frac{\partial l}{\partial \alpha} = \frac{1}{2\sigma^2} \sum_{i=1}^{n} 2x_{i-1}(x_i - \alpha x_{i-1}) \;\; = \frac{1}{\sigma^2} \sum_{i=1}^{n} x_i x_{i-1} - \frac{\alpha}{\sigma^2} \sum_{i=1}^{n} x_{i-1}^2$$

and setting $\frac{\partial l}{\partial \alpha} = 0$ we have:

$$\hat{\alpha} = \frac{\sum_{i=1}^{n} x_i x_{i-1}}{\sum_{i=1}^{n} x_{i-1}^2} .$$

Differentiating with respect to σ we have:

$$\frac{\partial l}{\partial \sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(x_i - \alpha x_{i-1})^2.$$

Setting this expression equal to zero we have:

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\alpha} x_{i-1})^2.$$

(iv) The Yule Walker equations are:

$$\gamma_0 = \text{cov}\,(\alpha X_{t-1} + e_t, X_t) = \alpha\,\text{cov}(X_{t-1}, X_t) + \text{cov}\,(e_t, X_t) = \alpha\gamma_1 + \sigma^2$$
$$\gamma_1 = \text{cov}\,(\alpha X_{t-1} + e_t, X_{t-1}) = \alpha\,\text{cov}(X_{t-1}, X_{t-1}) + \text{cov}\,(e_t, X_{t-1}) = \alpha\gamma_0$$

Using these to estimate the parameters we get:

$$\hat{\sigma}^2 = \hat{\gamma}_0 - \hat{\alpha}\hat{\gamma}_1.$$

(v) $$\hat{\alpha} = \frac{\hat{\gamma}_1}{\hat{\gamma}_0} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(x_{i-1} - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

The difference between them is that in the second approach we need to centralise the data around the mean $\bar{x}$.

This question was relatively well answered for a time series question. It was clear that some candidates had learnt the bookwork, but struggled with this more unfamiliar application of time series. In particular only the best candidates accurately completed the differentiation needed in part (iii).

END OF EXAMINERS' REPORT