INSTITUTE AND FACULTY OF ACTUARIES

EXAMINERS' REPORT

September 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

December 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 September 2013 paper

The examiners felt that this paper was slightly less routine than the April paper, but broadly in line with other recent papers. The quality of solutions was often good, with questions 7 and 9 providing the greatest challenge to most students.

1 The posterior distribution of p is given by

$$f(p \mid k \text{ claims}) \propto f(k \text{ claims} \mid p) \times f(p)$$

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

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

which is the pdf of a Beta distribution with parameters $\alpha + k$ and $\beta + n - k$.

Using the fact given in the questions, the mode of the posterior distribution (which is the estimate of p under all or nothing loss is given by:

$$\hat{p} = \frac{\alpha + k - 1}{\alpha + k + \beta + n - k - 2} = \frac{\alpha + k - 1}{\alpha + \beta + n - 2}$$

$$= \frac{\alpha - 1}{\alpha + \beta - 2} \times \frac{\alpha + \beta - 2}{\alpha + \beta + n - 2} + \frac{k}{n} \times \frac{n}{\alpha + \beta + n - 2}$$

$$= (1 - Z) \times \frac{\alpha - 1}{\alpha + \beta - 2} + Z \times \frac{k}{n}$$

where $Z = \dfrac{n}{\alpha + \beta + n - 2}$.

This is in the form of a credibility estimate since $\dfrac{\alpha - 1}{\alpha + \beta - 2}$ is the prior estimate of p under all or nothing loss and k/n is the estimate of p derived from the data.

The first part of this question was answered well. Most candidates didn't recognise the need to base the prior estimate on the mode of the prior distribution and therefore didn't manage to express the posterior estimate as a credibility estimate.

2 The mean amount paid by the reinsurance company is given by:

$$\int_{80}^{160}\left(\frac{1}{2}x-40\right)\times 0.01e^{-0.01x}dx+\int_{160}^{\infty}(x-120)\times 0.01e^{-0.01x}dx\,.$$

The first integral is (using integration by parts):

$$\left[-\left(\frac{1}{2}x-40\right)e^{-0.01x}\right]_{80}^{160}+\int_{80}^{160}\frac{1}{2}e^{-0.01x}dx$$

$$=-40e^{-1.6}-\left[50e^{-0.01x}\right]_{80}^{160}$$

$$=50e^{-0.8}-90e^{-1.6}=4.2957.$$

The second integral is:

$$\left[-(x-120)e^{-0.01x}\right]_{160}^{\infty}+\int_{160}^{\infty}e^{-0.01x}dx$$

$$=40e^{-1.6}+100e^{-1.6}$$

$$=28.26551.$$

So total mean claim = 4.29576 + 28.26551 = 32.56.

This question was generally answered well. Weaker candidates could not integrate by parts accurately.

3 (i) If $U(t)=U+ct-S(t)$ where $U=U(0)=\$0.1\text{m}$ then

$$\Psi(\$0.1m,1)=\Pr\left(U(t)\le 0 \text{ for some } t\in(0,1] \text{ given } U(0)=\$0.1m\right)$$

and

$$\Psi(\$0.1m)=\Pr(U(t)\le 0 \text{ for some } t>0 \text{ given } U(0)=\$0.1m)$$

(ii) The premium charged will be:

$$1.3\times 0.2\times(0.25\times\$1\text{m}+0.75\times\$0.1\text{m})=\$0.0845\text{m}.$$

(iii) The possibilities are tabulated below, where *N* means not injured, *R* means injured but recovered and *X* means injured but career ending:

Year 1	**Year 2**	**Probability**	**Ruin?**
N	*N*	$0.8 \times 0.8 = 0.64$	No
N	*R*	$0.8 \times 0.15 = 0.12$	No
N	*X*	$0.8 \times 0.05 = 0.04$	Yes
R	*N*	$0.15 \times 0.8 = 0.12$	No
R	*R*	$0.15 \times 0.15 = 0.0225$	No
R	*X*	$0.15 \times 0.05 = 0.0075$	Yes
X	N/A	0.05	Yes

Summing the cases where ruin occurs we have:

$$\psi(\$0.1\text{m}, 2) = 0.04 + 0.0075 + 0.05 = 0.0975$$

Many candidates lost marks in part (i) by not giving a sufficiently precise definition to score full marks. For part (iii) candidates who worked through the possibilities methodically generally scored well. A number of candidates unnecessarily used approximate methods in part (iii).

4 (i) The algorithm is as follows:

Step 1 Generate *u* from the uniform distribution on [0,1].

Step 2 If $0 < u < 0.3$ set $X = 1$.

If $0.3 <= u < 0.6$ set $X = 2$.

Otherwise set $X = 3$.

(ii) We need to solve $P(Y < 2.5) = 0.75$

but $P(Y < 2.5) = 1 - e^{-2.5\lambda}$

so $1 - e^{-2.5\lambda} = 0.75$

so $e^{-2.5\lambda} = 0.25$

so $\lambda = \dfrac{\log(0.25)}{-2.5} = 0.554517744$ and the mean of *Y* is 1.803368801.

(iii) The extended algorithm is:

Step 1 Generate v from the uniform distribution on [0,1].

Step 2 If $v<0.2$ then generate a sample from X as in (i) and finish, otherwise go to step 3.

Step 3 Generate u from the uniform distribution on [0,1].

Step 4 Set $1-e^{-\lambda x}=u$

i.e. $x=\dfrac{\log(1-u)}{-0.55451744}$

This question was answered well.

5 (i) The premium loading θ is given by:

$$74.25=(1+\theta)\times(0.75\times 2000+0.25\times 5000)\times 0.02=55(1+\theta)$$

and so

$$\theta=\frac{74.25}{55}-1=35\%.$$

(ii) Under A expected profit is:

$$2000\times 74.25-2000\times 0.02\times(0.75\times 2000+0.25\times 5000)$$
$$=38,500.$$

Under B expected profit is:

$$2,000\times 74.25-2,000\times 27-0.7\times 2,000\times 0.02\times(0.75\times 2,000+0.25\times 5,000)$$
$$=17,500.$$

Under C expected profit is:

$$2000\times 74.25-2000\times 15-2000\times 0.02\times(0.75\times 2000+0.25\times 3000)$$
$$=28,500$$

so the optimal course under the Bayes criterion is no reinsurance.

(iii) Under the minimax we need to consider the worst case scenario – which is that all 2,000 workers die in industrial accidents.

Under this outcome, the losses are:

Under A: $2000 \times 74.25 - 2000 \times 5000 = -9,851,500$

Under B: $2000 \times (74.25 - 27) - 2000 \times 5000 \times 0.7 = -6,905,500$

Under C: $2000 \times (74.25 - 15) - 2000 \times 3000 = -5,881,500$

so the optimal decision under the minimax criterion is C.

(iv) The approach in (iii) puts all the weight on what is at first seems a pretty unlikely scenario – so that our decision making is driven by something fairly remote.

That said, the workers are all in the same factory, so it is not inconceivable that a single catastrophe could result in a large number of claims all at the same time – i.e. the lives are not independent.

This question was well answered. There are a number of alternative approaches available (for example working on a per policy basis) which all give the same results, and all of which were given full credit. Candidates made a range of comments in part (iv) and all sensible answers were given credit.

6 The average cost per claim is given in the table:

	0	*1*	*2*
2010	497.59	662.35	836.49
2011	588.89	803.14	
2012	750.00		

The grossing up factors for average costs are given in the table below (the underlined figures are the simple averages):

	0	*1*	*2*	*Ult*
2010	497.6	662.3	836.5	836.5
	59.49%	79.18%	<u>100.00%</u>	
2011	588.9	803.1		1014.3
	58.06%	<u>79.18%</u>		
2012	750.0			1276.1
	<u>58.77%</u>			

The grossing up factors for claim numbers are as follows:

	0	*1*	*2*	*Ult*
2010	87.0	132.0	151.0	151.0
	57.62%	87.42%	100.00%	
2011	117.0	156.0		178.5
	65.56%	87.42%		
2012	99.0			160.7
	61.59%			

So the total claims are:

	Average amount	*Number*	*Total*
2010	836.5	151.0	126310
2011	1014.3	178.5	181006
2012	1276.1	160.7	205126
			512442

So the outstanding claims are:

512,442 – 126,310 – 125,290 – 74,250 = 186,592.

This question was well answered.

7 (i) Let X_i be the amount paid on the *i*th claim:

Then

$$E(X_i) = 50\int_0^{50} f(y)dy + \int_{50}^{\infty} yf(y)dy$$

$$= 50\int_0^{50} 0.01e^{-0.01y}dy + I_1$$

Using the notation given in the question.

Now

$$I_1 = 50e^{-0.5} + 100\int_{50}^{\infty} 0.01e^{-0.01y}dy$$

$$= 50e^{-0.5} + 100\left[-e^{-0.01y}\right]_{50}^{\infty}$$

$$= 50e^{-0.5} + 100e^{-0.5} = 150e^{-0.5}$$

So

$$E(X_i) = 50\left[-e^{-0.01y}\right]_0^{50} + 150e^{-0.5}$$

$$= -50e^{-0.5} + 50 + 150e^{-0.5} = 50 + 100e^{-0.5} = 110.653066$$

And

$$E(X_i^2) = 50^2\int_0^{50} f(y)dy + \int_{50}^{\infty} y^2 f(y)dy$$

$$= 2,500\int_0^{50} 0.01e^{-0.01y}dy + I_2$$

$$= 2,500\left[-e^{-0.01y}\right]_0^{50} + 50^2 e^{-0.5} + 200I_1$$

$$= -2,500e^{-0.5} + 2,500 + 2,500e^{-0.5} + 200\times 150e^{-0.5}$$

$$= 2,500 + 200\times 150e^{-0.5} = 20,695.91979$$

So finally we have:

$$E(S) = 20\times 110.653066 = 2213.06$$

$$\text{Var}(S) = 20\times 20695.91979 = 413918.40.$$

(ii) We need to solve:

$$e^{\mu+\sigma^2/2} = 2213.06 \quad (1)$$

and

$$e^{2\mu+\sigma^2}\left(e^{\sigma^2}-1\right) = 413918.40. \quad (2)$$

Dividing (2) by the square of (1) we have:

$$e^{\sigma^2}-1 = \frac{413918.40}{2213.06^2} = 0.084514$$

$$\sigma^2 = \log(1.084514) = 0.081132$$

and substituting into (1) we have:

$$\mu = \log(2213.06) - \frac{0.081132}{2} = 7.6615655.$$

Finally:

$$P(S > 4000) = P(N(7.6615655, 0.081132) > \log(4000))$$

$$= P\left(N(0,1) > \frac{8.29404964 - 7.6615655}{\sqrt{0.081132}}\right) = P(N(0,1) > 2.2205)$$

$$= 0.95 \times (1 - 0.98679) + 0.05 \times (1 - 0.98713)$$

$$= 0.01319$$

(iii) The probability will be lower.

This is because the log normal distribution has a "fat tail" and hence gives more weight to extreme outcomes.

Only the best candidates were able to derive the value of the variance in part (i) despite the formula for integration by parts being given in the question paper. The remaining parts were well answered.

8 (i) We have 4 years of observations such that $y_1 + y_2 + y_3 + y_4 = 33$. The likelihood function is then:

$$L = \prod_{i=1}^{4} \frac{\alpha^{y_i - 1}}{(1+\alpha)^{y_i}} = \frac{\alpha^{33-4}}{(1+\alpha)^{33}} = \frac{\alpha^{29}}{(1+\alpha)^{33}}$$

The log-likelihood is then:

$$l = 29\log\alpha - 33\log(1+\alpha)$$

Taking its derivative w.r.t. α and equation it to zero we have:

$$\frac{29}{\alpha} - \frac{33}{1+\alpha} = 0$$

$$29(1+\alpha) = 33\alpha$$

which implies that $29 = 4\alpha$

therefore $\hat{\alpha} = \frac{29}{4} = 7.25.$

Differentiating the log likelihood again gives $-\frac{29}{\alpha^2} + \frac{33}{(1+\alpha)^2}$ which is negative at $\hat{\alpha} = 7.25.$

(ii) We have:

$$p(y) = \frac{\alpha^{y-1}}{(1+\alpha)^{y}} = \exp[y\log\alpha - y\log(1+\alpha) - \log\alpha]$$

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

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

where

$\theta = \log\left(\frac{\alpha}{1+\alpha}\right)$, the natural parameter

$\varphi = 1$

$$a(\varphi) = \varphi$$

$$b(\theta) = \log \alpha = \theta - \log(1 - e^{\theta})$$

$$c(y, \varphi) = 0$$

This question was mostly well answered. Only the best candidates showed that the estimate was a maximum by evaluating the second derivative of the log-likelihood at the value of the estimate. In part (ii) some candidates failed to score full marks as a result of not specifying all the parameters.

9 (i) The three main stages are:

(a) tentative model identification
(b) model fitting
(c) diagnostics

(ii) Since the auto-correlation is non-zero for the first lag only and the partial auto-correlation function decays exponentially it is likely that the observed data comes from an MA(1) (or equivalently a ARMA(0,1) or ARIMA(0,0,1) model).

(iii) First note that for this model:

$$\text{Cov}(X_t, e_t) = \sigma^2$$

and

$$\text{Cov}(X_t, e_{t-1}) = \alpha_1 \text{Cov}(X_{t-1}, e_{t-1}) + \beta_1 \sigma^2 = (\alpha_1 + \beta_1)\sigma^2.$$

Taking the covariance of the defining equation with X_t we get:

$$\gamma_0 = \alpha_1 \gamma_1 + \alpha_2 \gamma_2 + \beta_1(\alpha_1 + \beta_1)\sigma^2 + \sigma^2.$$

Taking the covariance with X_{t-1} we get:

$$\gamma_1 = \alpha_1 \gamma_0 + \alpha_2 \gamma_1 + \beta_1 \sigma^2.$$

Taking the covariance with X_{t-2} we get:

$$\gamma_2 = \alpha_1 \gamma_1 + \alpha_2 \gamma_0$$

and in general

$$\gamma_n = \alpha_1\gamma_{n-1} + \alpha_2\gamma_{n-2} \text{ for } n > 2.$$

(iv) The presence of the term $\beta_1 e_{t-1}$ means that the PACF will decay exponentially to zero, but it will never get there, so that the PACF will always be non-zero.

Many candidates struggled with this question, with only the best accurately calculating the covariance of X_t with e_{t-1}. The chart on the printed examination paper was not clear, and the examiners took a generous approach to marking part (ii) where candidates had struggled interpreting the chart.

10 (i) Firstly:

$$f(\mu|x_1, x_2, \ldots, x_n) \propto f(x_1, x_2, \ldots, x_n) f(\mu)$$

$$\propto \prod_{i=1}^{n} e^{-\mu} \frac{\mu^{x_i}}{x_i!} \times \frac{\lambda^{\alpha}}{\Gamma(\alpha)} \mu^{\alpha-1} e^{-\lambda\mu}$$

$$\propto \mu^{\sum_{i=1}^{n} x_i + \alpha - 1} e^{-(\lambda+n)\mu}$$

which is the pdf of another gamma distribution. So the posterior distribution is gamma with parameters $\alpha + \sum_{i=1}^{n} x_i$ and $\lambda + n$.

(ii) Under quadratic loss the Bayes estimate is the mean of the posterior distribution, so:

$$\hat{\mu} = \frac{\alpha + \sum_{i=1}^{n} x_i}{\lambda + n}$$

which can be written as

$$\hat{\mu} = \frac{\alpha}{\lambda} \times \frac{\lambda}{\lambda + n} + \frac{\sum_{i=1}^{n} x_i}{n} \times \frac{n}{\lambda + n}$$

$$= (1 - Z)\frac{\alpha}{\lambda} + Z\overline{x}$$

where $Z = \frac{n}{\lambda + n}$. This is in the form of a credibility estimate since the mean of the prior distribution is α/λ and we have written the posterior mean as a weighted average of the prior mean and the mean of the observed data.

(iii) In this case we have:

$$\hat{\mu} = \frac{10 + 42}{2 + 6} = 6.5.$$

(iv) (a) Let S be the time taken to resolve a single query. Then for a simple query:

$$P(S > 30 | \text{simple}) = \left(\frac{20}{20 + 30} \right)^3 = 0.4^3 = 0.064.$$

For a complicated query we have

$$P(S > 30 | \text{complicated}) = e^{-0.4 \times 30^{0.5}} = 0.111817.$$

And finally

$$P(S > 30) = 0.75 \times 0.064 + 0.25 \times 0.111817 = 0.07595.$$

(b) Mean time for simple calls is $\frac{20}{3 - 1} = 10$.

Mean time for complicated calls is

$$\Gamma\left(1 + \frac{1}{0.5}\right) \times 0.4^{-1/0.5} = \Gamma(3) \times 0.4^{-2} = 2 \times 0.4^{-2} = 12.5.$$

Overall mean is $0.75 \times 10 + 0.25 \times 12.5 = 10.625$.

(c) Overall total time is $6.5 \times 10.625 = 69.0625$.

(v) (a) The parameter of the exponential distribution is $\frac{1}{10.625} = 0.094117647$.

(b) The probability of taking more than 30 minutes using this approximation is $P(S > 30) = e^{-\frac{30}{10.625}} = 0.059396$.

(c) This compares to the true value of 0.07595. The exponential distribution underestimates this tail probability since it has less fat tails than the Pareto and Weibull distributions.

(vi) We first find the parameter β of the exponential distribution being used. This is given by:

$$P(S > 30) = e^{-30\beta} = 0.1$$

so

$$\beta = \frac{\log 0.1}{-30} = 0.076752836$$

and the mean of the exponential distribution is 13.0288.

The mean of the given Weibull distribution is $\Gamma\left(1+\frac{1}{0.5}\right)\times c^{-\frac{1}{0.5}} = 2c^{-2}$.

The overall mean is then given by $0.75\times 10 + 0.25\times 2c^{-2} = 7.5 + 0.5c^{-2}$.

Equating this to 13.0288 gives:

$$7.5 + 0.5c^{-2} = 13.0288$$

$$c^{-2} = 11.0576$$

$$c = 0.300725.$$

Parts (i) to (iv) were well answered. Parts (v) and (vi) were attempted by only the more able candidates.

END OF EXAMINERS' REPORT