THE ROYAL STATISTICAL SOCIETY

GRADUATE DIPLOMA EXAMINATION

NEW MODULAR SCHEME

introduced from the examinations in 2009

MODULE 3

SOLUTIONS FOR SPECIMEN PAPER A

THE QUESTIONS ARE CONTAINED IN A SEPARATE FILE

The time for the examination is 3 hours. The paper contains eight questions, of which candidates are to attempt **five**. Each question carries 20 marks. An indicative mark scheme is shown within the questions, by giving an outline of the marks available for each part-question. The pass mark for the paper as a whole is 50%.

The solutions should not be seen as "model answers". Rather, they have been written out in considerable detail and are intended as learning aids. For this reason, they do not carry mark schemes. Please note that in many cases there are valid alternative methods and that, in cases where discussion is called for, there may be other valid points that could be made.

While every care has been taken with the preparation of the questions and solutions, the Society will not be responsible for any errors or omissions.

The Society will not enter into any correspondence in respect of the questions or solutions.

Note. In accordance with the convention used in all the Society's examination papers, the notation log denotes logarithm to base *e*. Logarithms to any other base are explicitly identified, e.g. $\log_{10}$.

Graduate Diploma Module 3, Specimen Paper A. Question 1

(i) $G(z) = \sum_{i=0}^{\infty} p_i z^i$.

(ii) $G_n(z) = E\left(z^{X_n}\right) = \sum_{i=0}^{\infty} p_i E\left(z^{X_n} \middle| X_1 = i\right) = \sum_{i=0}^{\infty} p_i \left(G_{n-1}(z)\right)^i = G\left(G_{n-1}(z)\right)$

(valid for $n - 1 \geq 1$, i.e. for $n \geq 2$).

(iii) Using the function of a function rule, $G_n'(z) = G'\left(G_{n-1}(z)\right).G_{n-1}'(z)$.

Setting $z = 1$, $\mu_n = \mu.\mu_{n-1}$ (for $n \geq 2$). Hence $\mu_n = \mu^{n-1}\mu_1$ (for $n \geq 2$).

Using the initial condition $\mu_1 = \mu$ gives $\mu_n = \mu^n$ (for $n \geq 1$).

(iv) $\theta_n = P(X_n = 0) = G_n(0)$. Setting $z = 0$ in the relationship of part (ii), we obtain

$$\theta_n = G\left(\theta_{n-1}\right) \qquad \text{(for } n \geq 2\text{)}.$$

(v) Letting $n \to \infty$ in the result of part (iv), and noting that G is a continuous function of z so that $G(\theta_{n-1}) \to G(\theta)$ as $n \to \infty$, we obtain the required equation $\theta = G(\theta)$.

(vi) In the special case where the offspring distribution is a geometric distribution with probability generating function $G(z) = (1 - q)/(1 - qz)$, we have $G'(z) = q(1 - q)/(1 - qz)^2$. Hence $\mu = G'(1) = q/(1 - q)$ and $\mu_n = [q/(1 - q)]^n$ $(n \geq 1)$.

(vii) θ is the smallest positive root of the equation $\theta = G(\theta)$. This equation is

$$\theta = \frac{1-q}{1-q\theta} \qquad \text{i.e.} \quad q\theta^2 - \theta + 1 - q = 0\,.$$

This quadratic has roots $\theta = (1 - q)/q$ and $\theta = 1$. So the probability of ultimate extinction is $(1 - q)/q$ (which is < 1) if $q > ½$ and 1 if $q \leq ½$.

Graduate Diploma Module 3, Specimen Paper A. Question 2

(i) Take the states of the model as

0: basic premium
1: a% reduction
2: b% reduction
3: c% reduction.

Assuming that the probability of making a claim in any year is θ, independent of all other years, the conditions for a Markov chain apply. The transition matrix is

$$\mathbf{P} = \begin{pmatrix} \theta & 1-\theta & 0 & 0 \\ \theta & 0 & 1-\theta & 0 \\ \theta & 0 & 0 & 1-\theta \\ \theta & 0 & 0 & 1-\theta \end{pmatrix}$$

(ii) The two-step and three-step transition matrices are $\mathbf{P}^2$ and $\mathbf{P}^3$.

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

$$\mathbf{P}^3 = \begin{pmatrix} \theta & \theta(1-\theta) & \theta(1-\theta)^2 & (1-\theta)^3 \\ \theta & \theta(1-\theta) & \theta(1-\theta)^2 & (1-\theta)^3 \\ \theta & \theta(1-\theta) & \theta(1-\theta)^2 & (1-\theta)^3 \\ \theta & \theta(1-\theta) & \theta(1-\theta)^2 & (1-\theta)^3 \end{pmatrix}$$

The required probability is the element of the matrix for transition from state 1 to state 1 (i.e. the element in the second row and the second column), so for both (a) [i.e. $\mathbf{P}^2$] and (b) [i.e. $\mathbf{P}^3$] this is $\theta(1-\theta)$.

It would appear that the probability of going from state 1 back to state 1 is $\theta(1-\theta)$ for any number of steps greater than 1.

Solution continued on next page

(iii) The stationary distribution $\boldsymbol{\pi} = [\pi_0 \quad \pi_1 \quad \pi_2 \quad \pi_3]$ is obtained by solving the equations given by $\boldsymbol{\pi} = \boldsymbol{\pi}\mathbf{P}$, i.e.

$$\pi_0 = \theta\pi_0 + \theta\pi_1 + \theta\pi_2 + \theta\pi_3$$

$$\pi_1 = (1 - \theta)\pi_0$$

$$\pi_2 = (1 - \theta)\pi_1$$

$$\pi_3 = (1 - \theta)\pi_2 + (1 - \theta)\pi_3 ,$$

together with the normalisation condition $\pi_0 + \pi_1 + \pi_2 + \pi_3 = 1$.

The first equation together with the normalisation condition gives $\pi_0 = \theta$.

Then working successively through the equations for the stationary distribution, one by one, we obtain

$$\pi_1 = \theta(1 - \theta), \quad \pi_2 = \theta(1 - \theta)^2, \quad \pi_3 = (1 - \theta)^3.$$

Graduate Diploma Module 3, Specimen Paper A. Question 3

(i) (a) The state space can be defined as follows. 1: in repair; 2: working.

The instantaneous transition rates are θ from 1 to 2 and λ from 2 to 1.

(b) The equilibrium equations are

$$\theta\pi_1 = \lambda\pi_2$$
$$\lambda\pi_2 = \theta\pi_1$$

(where π_1 and π_2 are the respective equilibrium probabilities), which give $\pi_2 = (\theta/\lambda)\pi_1$.

Using the normalisation condition $\pi_1 + \pi_2 = 1$, we get $\pi_2 = \dfrac{\theta/\lambda}{1+(\theta/\lambda)} = \dfrac{\theta}{\lambda+\theta}$, which is the long-term proportion of time that the machine is in working order.

(ii) (a) The repair time distribution has mean $1/(2\theta) + 1/(2\theta) = 1/\theta$ and variance $1/(2\theta)^2 + 1/(2\theta)^2 = 1/(2\theta^2)$.

It is an Erlang distribution which is a gamma distribution with positive integer parameter.

(b) The state space can be defined as follows: 1: in first stage of repair; 2: in second stage of repair; 3: working.

The instantaneous transition rates are 2θ from 1 to 2, 2θ from 2 to 3 and λ from 3 to 1.

(c) The equilibrium equations are

$$2\theta\pi_1 = \lambda\pi_3$$
$$2\theta\pi_2 = 2\theta\pi_1$$
$$\lambda\pi_3 = 2\theta\pi_2$$

which give $\pi_1 = \pi_2$ and $\pi_3 = (2\theta/\lambda)\pi_1$.

Using the normalisation condition $\pi_1 + \pi_2 + \pi_3 = 1$, we get $\pi_1 = \pi_2 = \dfrac{\lambda}{2(\lambda+\theta)}$ and $\pi_3 = \dfrac{2\theta/\lambda}{1+1+(2\theta/\lambda)} = \dfrac{\theta}{\lambda+\theta}$, the last of which is the long-term proportion of time that the machine is in working order.

(iii) The mean repair time and the proportion of time that the machine is working is the same for both models. But in the second model the standard deviation of repair times is less, by a factor of $1/\sqrt{2}$, than in the first – the repair times are less variable.

Graduate Diploma Module 3, Specimen Paper A. Question 4

(i) The state space is the set of all non-negative integers.

The instantaneous transition rates are λ from i to $i + 1$ (for $i \geq 0$) and μi from i to $i - 1$ (for $i \geq 1$).

(ii) The detailed balance equations are $\lambda\pi_{n-1} = \mu n\pi_n$ (for $n \geq 1$). Thus

$$\pi_n = \big(\lambda/(\mu n)\big)\pi_{n-1} = \big(\rho/n\big)\pi_{n-1} \quad \text{(for } n \geq 1\text{), where } \rho = \lambda/\mu,$$

and, using this relation recursively, we obtain $\pi_n = \big(\rho^n/n!\big)\pi_0$ (for $n \geq 0$).

Using the normalisation condition $\Sigma\pi_n = 1$ and the exponential series $\Sigma\rho^n/n! = e^{\rho}$ (summations from $n = 0$ to $n = \infty$), we obtain $\pi_n = e^{-\rho}\rho^n/n!$, i.e. the equilibrium distribution is the Poisson distribution with mean ρ.

(iii) If X denotes the queue size at an arbitrary time in equilibrium, we require to find k such that $P(X \leq k) = 0.99$. Using the Normal approximation we have $X \sim \mathrm{N}(\rho, \rho)$ and, using a continuity correction, we require that (approximately)

$$P(X \leq k + \tfrac{1}{2}) = P[(X - \rho)/\sqrt{\rho} \leq (k + \tfrac{1}{2} - \rho)/\sqrt{\rho}] = 0.99.$$

Therefore $(k + \tfrac{1}{2} - \rho)/\sqrt{\rho} \approx 2.33$, i.e. $k \approx \rho + 2.33\sqrt{\rho} - \tfrac{1}{2}$.

(iv) (a) $\pi_0 = e^{-\rho}$.

(b) The expected length of an idle period is the mean of the exponential distribution with parameter λ, i.e. $1/\lambda$.

(v) The expected length of a cycle of idle plus busy period is $(1/\lambda) + L$. The expected length of time during the cycle that the queue is empty is $1/\lambda$. Hence [it may be proved] the long-run proportion of time that the queue is empty is

$$\frac{1/\lambda}{(1/\lambda) + L} = \frac{1}{1 + \lambda L}.$$

But this is also given by $\pi_0 = e^{-\rho}$. Hence $e^{-\rho} = 1/(1 + \lambda L)$, i.e. $1 + \lambda L = e^{\rho}$, i.e. $L = \dfrac{e^{\rho} - 1}{\lambda}$.

Graduate Diploma Module 3, Specimen Paper A. Question 5

(i) θ is the traffic intensity, the mean number of arrivals in a service time, the (mean) service time divided by the mean inter-arrival time. $\theta < 1$ is a necessary and sufficient condition for the existence of an equilibrium distribution for the queue.

(ii) The distribution of the number of customer arrivals in a service time is a Poisson distribution with mean θ. Hence $k_r = e^{-\theta}\theta^r/r!$ $(r \geq 0)$. The probability generating function is $K(z) = e^{-\theta(1-z)}$.

(iii) Define Q_n to be the number of customers in the queue just after the service of the nth customer is completed. Let R_n denote the number of customers who arrive during the service time of the nth customer.

If $Q_n = 0$ then $Q_{n+1} = R_{n+1}$. Hence $p_{0j} = k_j$ $(j \geq 0)$.

If $Q_n \geq 1$ then $Q_{n+1} = Q_n - 1 + R_{n+1}$, where the -1 arises from the departure of the $(n+1)$th customer. Hence $Q_{n+1} \geq Q_n - 1$ and, for $i \geq 1$, $p_{ij} = k_{j-i+1}$ $(j \geq i-1)$.

Finally, the queue cannot decrease by more than one customer at a time, so $p_{ij} = 0$ for $j < i - 1$.

(iv) The equations $\pi_j = \sum_{i=0}^{\infty} \pi_i p_{ij}$ reduce here to $\pi_j = \pi_0 k_j + \sum_{i=1}^{j+1} \pi_i k_{j-i+1}$ (for $j \geq 0$).

Multiplying this equation by z^j and summing over j,

$$\Pi(z) = \pi_0 K(z) + \sum_{j=0}^{\infty} z^j \sum_{i=1}^{j+1} \pi_i k_{j-i+1}$$

$$= \pi_0 K(z) + \sum_{i=1}^{\infty} \pi_i z^i \sum_{j=i-1}^{\infty} k_{j-i+1} z^{j-i} = \pi_0 K(z) + \frac{(\Pi(z) - \pi_0) K(z)}{z}.$$

Rearranging this equation gives the required result.

(v) Because of the normalisation condition $\sum_{j=0}^{\infty} \pi_j = 1$, we must have $\Pi(1) = 1$ and $\lim_{h \downarrow 0} \Pi(1-h) = 1$. Similarly, $K(1) = 1$ and $\lim_{h \downarrow 0} K(1-h) = 1$.

Substituting $z = 1 - h$ into the expression of part (iv) gives $\Pi(1-h) =$

$$\frac{\pi_0 h K(1-h)}{K(1-h) - 1 + h} = \frac{\pi_0 h K(1-h)}{K(1) - hK'(1) + o(h) - 1 + h} = \frac{\pi_0 h K(1-h)}{h - hK'(1) + o(h)}.$$

Letting $h \downarrow 0$, $1 = \pi_0/(1 - K'(1)) = \pi_0/(1-\theta)$. Hence $\pi_0 = 1 - \theta$.

Graduate Diploma Module 3, Specimen Paper A. Question 6

(i) The characteristic equation is $1 - (1/3)z - (1/12)z^2 = 0$, which has roots $z = 2$ and $z = -6$. Both the roots are greater than one in modulus, so the stationarity condition is satisfied.

(ii) Multiplying through by $Y_{t-\tau}$ in the model equation and taking expectations, we obtain the equations

$$\gamma_\tau = (1/3)\gamma_{\tau-1} + (1/12)\gamma_{\tau-2} \qquad (\tau \geq 1)$$

where $\gamma_\tau = \text{Cov}(Y_t, Y_{t-\tau}) = E(Y_t Y_{t-\tau})$ since the model mean is zero, and thus $\gamma_0 = \text{Var}(Y_t) = E(Y_t^2)$.

Dividing through by γ_0, we obtain the Yule-Walker equations

$$\rho_\tau = (1/3)\rho_{\tau-1} + (1/12)\rho_{\tau-2} \qquad (\tau \geq 1).$$

(iii) In the special case $\tau = 1$, using the fact that $\rho_0 = 1$ and the symmetry condition $\rho_{-1} = \rho_1$, we obtain $\rho_1 = (1/3) + (1/12)\rho_1$, which gives $\rho_1 = 4/11$.

(iv) The general solution of the difference equation of part (ii) is of the form

$$\rho_\tau = A_1\alpha_1^\tau + A_2\alpha_2^\tau \qquad (\text{for } \tau \geq -1),$$

where A_1 and A_2 are arbitrary constants and α_1 and α_2 are the roots of the auxiliary equation

$$\alpha^2 - (1/3)\alpha - (1/12) = 0.$$

The roots of the auxiliary equation are the reciprocals of the roots of the characteristic equation of part (i), i.e. 1/2 and –1/6. Hence the general solution of the difference equation is of the form

$$\rho_\tau = A_1(1/2)^\tau + A_2(-1/6)^\tau .$$

Using the conditions $\rho_0 = 1$ and $\rho_1 = 4/11$, we obtain the equations

$$A_1 + A_2 = 1$$

and

$$(1/2)A_1 - (1/6)A_2 = 4/11,$$

whose solution is $A_1 = 35/44$ and $A_2 = 9/44$. Hence the solution is as given in the question.

Graduate Diploma Module 3, Specimen Paper A. Question 7

(i) If the underlying trend is an exponential one, taking logarithms will transform the trend to a linear one, in which case an ARIMA model is likely to provide a better fit. If the variability of the series and, in particular, of any seasonal effects, increases with increase in the underlying level of the series, taking logarithms will tend to stabilise the variation, so that, again, an ARIMA model is likely to provide a better fit.

(ii) 95% confidence limits are at $\pm 1.96/\sqrt{240}$, or approximately at $\pm 2/\sqrt{240} = \pm 0.129$. So any autocorrelation outside these limits differs significantly from zero at the 5% level. We see that a large number of autocorrelations lie well outside these limits, with especially large values at lags 1, 2, 11, 12, 13, 23, 24, 25, 35, 36, 37. This clearly indicates the presence of seasonality of period 12 months and also suggests the presence of trend. The purpose of taking differences is to eliminate the trend, and the purpose of taking seasonal differences is to eliminate the seasonality.

(iii) Approximate 95% confidence limits are at $\pm 2/\sqrt{227} = \pm 0.133$. So any autocorrelation outside these limits differs significantly from zero at the 5% level. Here the only significant autocorrelations are at lag 1 and at lags 11, 12, 13. This shows that any trend and seasonality have been removed by the differencing to obtain a stationary series and suggests that the stationary series might be modelled with multiplicative moving average factors at lags 1 and 12. So a seasonal ARIMA$(0, 1, 1)\times(0, 1, 1)_{12}$ is suggested.

(iv) A seasonal ARIMA$(0, 1, 1)\times(0, 1, 1)_{12}$ has been fitted. The equation of the fitted model is

$$(1 - L)(1 - L^{12})Y_t = (1 - 0.7552L)(1 - 0.8102L^{12})\varepsilon_t$$

where L is the backshift operator and $\{\varepsilon_t\}$ is a white noise process, i.e.

$$(1 - L - L^{12} + L^{13})Y_t = (1 - 0.7552L - 0.8102L^{12} + 0.6119L^{13})\varepsilon_t$$

or

$$Y_t = Y_{t-1} + Y_{t-12} - Y_{t-13} + \varepsilon_t - 0.7552\varepsilon_{t-1} - 0.8102\varepsilon_{t-12} + 0.6119\varepsilon_{t-13}\ .$$

(v) None of the p-values of the modified Box-Pierce statistics is significant. So the residuals of the fitted model appear to come from a white noise process – our model appears to give a good fit to the data.

(vi) The forecast and 95% prediction interval for $Y252$ are given by 5.57051 and (5.31135, 5.82968) respectively. The forecast deaths and prediction interval are given by taking exponentials of these values. This, correct to the nearest whole number, gives 263 as the forecast number of deaths for December 1995 and (203, 340) as the 95% prediction interval.

Graduate Diploma Module 3, Specimen Paper A. Question 8

(i) The updating equation is $L_t = \alpha Y_t + (1 - \alpha)L_{t-1}$ and $\hat{y}_T(h) = L_T$.

(ii) The updating equations are

$$L_t = \alpha Y_t + (1 - \alpha)(L_{t-1} + B_{t-1})$$

$$B_t = \gamma(L_t - L_{t-1}) + (1 - \gamma)B_{t-1}$$

and $\hat{y}_T(h) = L_T + hB_T$.

(iii) At any time t, the fitted value is the forecast value of Y_t as made at the previous time point $t - 1$, i.e. $\hat{y}_{t-1}(1) = L_{t-1} + B_{t-1}$. The residual is the difference between the observed and fitted value, i.e. $Y_t - \hat{y}_{t-1}(1)$. For December 2007 we have

"fitted" = 153733.49 + 141.21 = 153874.70,

"residual" = 153866 – 153874.70 = –8.70.

(iv) $\hat{y}_T(1) = 153868.61 + 141.09 = 154009.70$.
$\hat{y}_T(2) = 153868.61 + (2 \times 141.09) = 154150.79$.
$\hat{y}_T(3) = 153868.61 + (3 \times 141.09) = 154291.88$.

(v) For January 2008,

Level: $L_t = (0.7)(153824) + (0.3)(153868.61 + 141.09) = 153879.71$,

Trend: $B_t = (0.02)(153879.71 - 153868.61) + (0.98)(141.09) = 138.49$.

(vi) For any time point t, the "deviation" e_t is the same as the "residual" as described in part (iii), i.e. $e_t = Y_t - \hat{y}_{t-1}(1)$. If the historical series runs from t = 1 to T,

$$\text{MAD} = \sum_{t=1}^{T} \frac{|e_t|}{T} \quad \text{and} \quad \text{MSD} = \sum_{t=1}^{T} \frac{e_t^2}{T} \ .$$

In the present case,

MAD = (162.50 + 337.42 + … + 8.70)/12 = 222.42,

MSD = (26406.36 + 113850.78 + … + 75.69)/12 = 75991.14 (using the squared residual values given to 2 decimal places in the question).