THE ROYAL STATISTICAL SOCIETY

GRADUATE DIPLOMA EXAMINATION

NEW MODULAR SCHEME

introduced from the examinations in 2009

MODULE 3

SPECIMEN PAPER A

SOLUTIONS 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. $\log_{10}$.

1. Let $\{X_n\}$ $(n \geq 0)$ represent a branching process, where X_n denotes the population size in the nth generation. The initial population size is 1, i.e. $X_0 = 1$, and in each generation the number of offspring produced by each individual that survive to the next generation follows the offspring distribution $\{p_i\}$ $(i \geq 0)$ with associated probability generating function $G(z)$. The numbers of surviving offspring produced by different individuals are statistically independent of each other.

Let $G_n(z)$ denote the probability generating function of the number of individuals in the population in the nth generation $(n \geq 1)$.

(i) Define $G(z)$ in terms of the distribution $\{p_i\}$ $(i \geq 0)$.

(1)

(ii) By conditioning on the number of individuals in the first generation, prove that

$$G_n(z) = G(G_{n-1}(z)) \qquad (n \geq 2).$$

(4)

(iii) Let μ denote the mean of the offspring distribution and let μ_n denote the mean population size in the nth generation $(n \geq 1)$. By differentiating the relationship of part (ii), find a recurrence relationship for the μ_n. Using an appropriate initial condition, solve this recurrence relationship to find an expression for μ_n as a function of μ.

(5)

(iv) For $n \geq 1$, let $\theta_n = P(X_n = 0)$, i.e. the probability that the population has become extinct by the nth generation. Using the relationship of part (ii), find a recurrence relationship for the θ_n.

(2)

(v) Let $\theta = \lim_{n\to\infty}\theta_n$, the probability of ultimate extinction of the population. From the result of part (iv), deduce that θ satisfies the equation $\theta = G(\theta)$.

(2)

Consider now the special case where the offspring distribution is a geometric distribution with probability generating function $G(z) = (1 - q)/(1 - qz)$, where the parameter q satisfies $0 < q < 1$.

(vi) Obtain μ_n $(n \geq 1)$ as a function of q.

(3)

(vii) Find an expression as a function of q for the probability θ of ultimate extinction of the population, distinguishing between the cases $q \leq ½$ and $q > ½$.

(3)

[Note. You may assume that θ is given by the smallest positive root of the equation $\theta = G(\theta)$.]

Turn over

2. An insurance firm operates a system of "no claims" bonuses on car insurance. When drivers are first insured, they pay the basic annual premium. Drivers who make no claim on their insurance policy in the current year qualify for a reduction on their premium for the next year. With one claim-free year the reduction is a%, with two consecutive claim-free years it is b ($> a$) %, and with three or more consecutive claim-free years it is c ($> b$) %. Drivers who reach this third level of reduction continue at that level in all subsequent years until they make a claim. A driver who makes at least one claim in the current year loses any "no claims" bonus and returns to the basic premium level for the following year. The insurance firm believes that a proportion θ ($0 < \theta < 1$) of drivers make at least one claim in any given year.

(i) Stating any necessary assumptions, describe how a driver's experience of "no claims" bonuses may be modelled as a Markov chain. Write down the transition matrix for this model.

(4)

(ii) Find the two-step and three-step transition matrices for this model. Suppose that, in the current year, a driver has one year's "no claims" bonus. Find the probability that this driver will be back in the same state of the model in (a) two years' time, (b) three years' time. Comment on your answers.

(8)

(iii) Find the stationary distribution for this model.

(8)

3. A machine in a factory alternates between being in working order and being under repair. The length of any period of time that the machine is in working order (its lifetime) is exponentially distributed with mean $1/\lambda$.

(i) Suppose that the repair time is exponentially distributed with mean $1/\theta$.

(a) Write down the state space and transition rates for an appropriate continuous time Markov chain model for the state of the machine.

(3)

(b) Obtain the corresponding equilibrium distribution and deduce an expression for the long-term proportion of time that the machine is in working order.

(4)

(ii) Suppose now that there are two stages in the repair process and the repair time has a distribution which is equivalent to the sum of two independent exponential distributions each with mean $1/(2\theta)$. The lifetimes and repair times are independent.

(a) What are the mean and variance of the repair time distribution? What is the general name of such a distribution?

(3)

(b) Write down the state space and transition rates for an appropriate continuous time Markov chain model for the state of the machine.

(4)

(c) Obtain the corresponding equilibrium distribution and deduce an expression for the long-term proportion of time that the machine is in working order.

(4)

(iii) Comment briefly on the extent to which the two models give similar results and on how they differ.

(2)

4. Consider a model for an M/M/∞ queue, in which customers arrive according to a Poisson process with rate λ. There is an unlimited number of servers available, and service times are independently and identically distributed, having an exponential distribution with mean $1/\mu$. The quantity λ/μ is denoted by ρ.

(i) For the corresponding continuous time Markov chain, $\{N(t)\}$ $(t \geq 0)$, specify the state space and write down the instantaneous transition rates.

(3)

(ii) Write down the detailed balance equations and deduce that the equilibrium distribution $\{\pi_n\}$ is a Poisson distribution.

(6)

(iii) In designing the facilities for the management of such a queue, it is required to find a number k such that, in the long run, the queue size is less than or equal to k for 99% of the time. Assume that ρ $(= \lambda/\mu)$ is considerably greater than 1. By using the Normal approximation to the Poisson distribution, show that the required value of k is given approximately by

$$k \approx \rho + 2.33\sqrt{\rho} - \frac{1}{2} .$$

(5)

(iv) Denote periods during which the queue is empty as "idle periods" and periods during which the queue is not empty as "busy periods". The queue alternates between such idle and busy periods. Assuming that the queue is in equilibrium, write down expressions in terms of ρ and λ for (a) the probability that at an arbitrary time point the queue is empty, (b) the expected length of an idle period.

(2)

(v) Outline an argument that shows why the expected length L of a busy period is given by

$$L = \frac{e^{\rho} - 1}{\lambda} .$$

(4)

[Note. In examination questions for this module, the word "queue" refers to all units in a system, i.e. those being served as well as those still waiting to be served.]

Turn over

5. Consider a particular model for an M/G/1, namely a single-server queue in which customers arrive according to a Poisson process and the service time is a fixed constant which is the same for all customers. Let λ be the rate of the Poisson process for customer arrivals, let d be the length of the service time for all customers, and define $\theta = \lambda d$.

(i) Comment on the meaning of the parameter θ, and explain how it relates to the existence of an equilibrium distribution for the model.

(2)

(ii) Write down an expression for k_r ($r \geq 0$), the probability that r customers arrive during any given service time, and an expression for the corresponding probability generating function $K(z)$.

(2)

Consider the imbedded Markov chain $\{X_n\}$ for this model, where X_n denotes the number of customers left behind in the queue by the nth customer when he leaves. This Markov chain has transition probabilities p_{ij} given by

$$p_{0j} = k_j \qquad (j \geq 0)$$

and, for $i \geq 1$,

$$p_{ij} = 0 \qquad (j < i - 1)$$
$$p_{ij} = k_{j-i+1} \qquad (j \geq i - 1).$$

(iii) Explain the reasoning behind the specification of the above transition probabilities.

(4)

(iv) Assume that a stationary distribution $\{\pi_j\}$ exists for the imbedded Markov chain and that the probability generating function of this distribution is $\Pi(z)$. Write down the equations satisfied by the π_j and deduce that

$$\Pi(z) = \frac{\pi_0 (1-z) K(z)}{K(z) - z}.$$

(7)

(v) Find an expression for π_0 in terms of θ.

(5)

6. Consider the AR(2) model

$$Y_t = \frac{1}{3}Y_{t-1} + \frac{1}{12}Y_{t-2} + \varepsilon_t \qquad (-\infty < t < \infty)$$

for a process $\{Y_t\}$, where $\{\varepsilon_t\}$ is a white noise process.

(i) Find the roots of the autoregressive characteristic equation and check that the stationarity condition is satisfied.

(4)

(ii) Find the Yule-Walker equations that are satisfied by the autocorrelation function ρ_t.

(4)

(iii) Obtain the value of ρ_1.

(3)

(iv) Show that a general expression for the autocorrelation function is given by

$$\rho_\tau = \frac{35}{44}\left(\frac{1}{2}\right)^\tau + \frac{9}{44}\left(-\frac{1}{6}\right)^\tau \qquad (\tau \geq 0).$$

(9)

7. Consider modelling the series of monthly Canadian road fatalities for the period from 1975 to 1994. In particular, let Y_t denote the natural logarithm of the number of deaths for month t ($1 \leq t \leq 240$). The sample autocorrelation functions (acfs) for the series $\{Y_t\}$ and for the series $\{\Delta\Delta_{12}Y_t\}$ obtained after differencing and seasonally differencing at lag 12 are tabulated **on the next page**, followed by some computer output.

(i) Explain why logarithms of the data have been analysed.

(2)

(ii) Calculate approximate limits beyond which sample autocorrelations differ significantly from zero at the 5% level. Comment on what you can learn about the series from inspection of the acf for $\{Y_t\}$ and explain the purpose of taking differences and seasonal differences.

(5)

(iii) Comment on the acf for $\{\Delta\Delta_{12}Y_t\}$ and how it is of help in identifying a possible ARIMA model to fit to the data.

(4)

(iv) State which of the family of ARIMA models has been fitted to the data in the output at the end of the question and write down explicitly the equation of the fitted model for the series $\{Y_t\}$.

(3)

(v) What can you deduce from the output about how well the model fits the data?

(2)

(vi) Obtain from the output the forecast number of deaths for December 1995, together with a 95% prediction interval, giving all your results correct to the nearest whole number.

(4)

Question continued on next page

Sample autocorrelation functions for qu 7

Lag	for $\{Y_t\}$	for $\{\Delta\Delta_{12}Y_t\}$
1	0.797	–0.510
2	0.543	0.046
3	0.221	0.012
4	–0.057	–0.057
5	–0.224	0.058
6	–0.311	–0.020
7	–0.237	–0.042
8	–0.076	0.076
9	0.176	–0.042
10	0.467	–0.055
11	0.711	0.264
12	0.833	–0.415
13	0.706	0.153
14	0.472	0.052
15	0.171	0.028
16	–0.095	–0.040
17	–0.259	–0.045
18	–0.337	0.061
19	–0.266	0.010
20	–0.115	–0.050
21	0.132	0.009
22	0.408	0.040
23	0.642	0.009
24	0.751	–0.050
25	0.639	0.050
26	0.419	–0.024
27	0.130	–0.032
28	–0.118	0.037
29	–0.284	0.034
30	–0.369	–0.039
31	–0.302	–0.023
32	–0.156	0.029
33	0.074	0.043
34	0.329	–0.070
35	0.551	0.014
36	0.663	0.071
37	0.564	–0.107
38	0.354	0.054
39	0.081	–0.050
40	–0.154	0.014

The computer output is on the next page

Turn over

Computer output for question 7

```
Final Estimates of Parameters

Type       Coef  SE Coef      T      P
MA   1   0.7552   0.0433  17.45  0.000
SMA  12  0.8102   0.0464  17.45  0.000

Differencing: 1 regular, 1 seasonal of order 12
Number of observations:  Original series 240, after differencing 227
Residuals:    SS =  2.37009 (backforecasts excluded)
              MS =  0.01053  DF = 225

Modified Box-Pierce (Ljung-Box) Chi-Square statistic

Lag            12     24     36     48
Chi-Square   14.7   26.6   31.3   44.9
DF             10     22     34     46
P-Value     0.142  0.225  0.602  0.518

Forecasts from period 240

                     95% Limits
Period  Forecast    Lower    Upper
   241   5.31689  5.11568  5.51809
   242   5.18851  4.98137  5.39565
   243   5.21676  5.00384  5.42968
   244   5.27027  5.05173  5.48881
   245   5.55732  5.33329  5.78134
   246   5.68882  5.45945  5.91819
   247   5.81867  5.58407  6.05327
   248   5.85931  5.61960  6.09903
   249   5.65930  5.41458  5.90402
   250   5.67653  5.42690  5.92616
   251   5.53600  5.28156  5.79044
   252   5.57051  5.31135  5.82968
```


Turn over

8. Let Y_t denote the observed value of a time series at time t and $\hat{y}_T(h)$ the forecast at time T for lead time h.

(i) Let L_t denote the smoothed value (the level) of the series at time t obtained using simple exponential smoothing. If α denotes the smoothing constant, write down (a) the updating equation for L_t and (b) an expression for $\hat{y}_T(h)$.

(2)

(ii) If, instead, Holt's two-parameter smoothing method is to be used for forecasting, let L_t denote the local level and B_t the trend at time t. If α and γ denote the smoothing constants for L_t and B_t respectively, write down (a) the updating equations for L_t and B_t and (b) an expression for $\hat{y}_T(h)$.

(3)

The time series of monthly figures for the US civilian labour force in thousands over the period from 1948 to 2007 is available from US government sources. There appears to be no seasonal effect, and Holt's method is to be used for forecasting. The following output is obtained for the final section of the data covering the twelve months of the year 2007, using the smoothing constants $\alpha = 0.7$ and $\gamma = 0.02$.

Month	Labour Force	Level	Trend	Fitted	Residual	Squared Residual
Jan	152958	152909.25	153.17	152795.50	162.50	26406.36
Feb	152725	152826.23	148.44	153062.42	-337.42	113850.78
Mar	152884	152911.20	147.17	152974.67	-90.67	8220.94
Apr	152542	152696.91	139.95	153058.38	-516.38	266643.66
May	152776	152794.26	139.09	152836.86	-60.86	3703.71
Jun	153085	153039.51	141.22	152933.35	151.65	22997.47
Jul	153182	153181.62	141.23	153180.72	1.28	1.63
Aug	152886	153017.06	135.12	153322.85	-436.85	190838.73
Sep	153506	153399.85	140.07	153152.17	353.83	125193.00
Oct	153306	153376.18	136.80	153539.92	-233.92	54720.52
Nov	153828	153733.49	141.21	153512.97	315.03	99241.15
Dec	153866	153868.61	141.09	153874.70	-8.70	75.69

(iii) Explain what the "fitted" and "residual" values are in the output, illustrating your explanation by showing how the values for December 2007 have been calculated. (4)

(iv) Given the data up to December 2007, obtain the forecast labour force in thousands for the next three months. (3)

(v) The labour force for January 2008 turned out to be 153824. Given this fact, calculate the values of "level" and "trend" in the corresponding row of the output. (4)

(vi) Given a historical run of a time series, to which Holt's method has been applied, define the *mean absolute deviation* (MAD) and the *mean square deviation* (MSD). Illustrate the definitions by calculating the MAD and the MSD using the data above for the twelve months of 2007. (4)