EXAMINATIONS OF THE HONG KONG STATISTICAL SOCIETY

GRADUATE DIPLOMA, 2009
(Modular format)

MODULE 3 : Stochastic Processes and Time Series

Time Allowed: Three Hours

Candidates should answer **FIVE** *questions.*

All questions carry equal marks.
The number of marks allotted for each part-question is shown in brackets.

Graph paper and Official tables are provided.

Candidates may use calculators in accordance with the regulations published in the Society's "Guide to Examinations" (document Ex1).

The notation log *denotes logarithm to base* ***e***.
Logarithms to any other base are explicitly identified, e.g. $\log_{10}$.

Note also that $\binom{n}{r}$ *is the same as* ${}^{n}C_{r}$.

This examination paper consists of 11 printed pages, **each printed on one side only**.
This front cover is page 1.
Question 1 starts on page 2.

There are 8 questions altogether in the paper.

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 $\theta_n = P(X_n = 0)$ $(n \geq 1)$, 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)

(iv) Let $\theta = \lim_{n\to\infty} \theta_n$, the probability of ultimate extinction of the population. From the result of part (iii), deduce that θ satisfies the equation $\theta = G(\theta)$. (2)

Consider now the special case where each individual produces exactly two offspring, both of which survive to the next generation independently of each other with probability p. In this case, the offspring distribution is a binomial distribution with parameters 2 and p.

(v) Find $G(z)$. (1)

(vi) Find θ_1, the probability that the population becomes extinct at the first generation. (1)

(vii) Show that θ_2, the probability that the population has become extinct by the second generation, can be written as

$$\theta_2 = (1-p)^2\left(1+p-p^2\right)^2.$$ (3)

(viii) Find an expression in terms of p for the probability θ of ultimate extinction of the population, distinguishing between the cases $p \leq \frac{1}{2}$ and $p > \frac{1}{2}$. (6)

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

2. Consider a Markov chain with three states, 1, 2 and 3, and transition matrix

$$\begin{pmatrix} \frac{2}{5} & \frac{2}{5} & \frac{1}{5} \\ \frac{1}{5} & \frac{3}{5} & \frac{1}{5} \\ \frac{1}{5} & \frac{2}{5} & \frac{2}{5} \end{pmatrix}.$$

(i) Explain what is meant by the statement that a Markov chain is an *irreducible recurrent chain*, and show, stating any general results that you assume, that this statement is true for the present chain.

(4)

(ii) Find the stationary distribution for this chain.

(8)

Now consider this Markov chain as a simple model for social mobility. People's occupations have been classified into the three classes "Upper" (State 1), "Middle" (State 2) and "Lower" (State 3). The transition probabilities model how the occupational classes of sons depend on the occupational classes of fathers. The transition probabilities as given above are rounded versions of estimates obtained from a social survey.

(iii) If initially, in the first generation, the proportions of males in each class are $\frac{2}{5}$, $\frac{2}{5}$ and $\frac{1}{5}$ respectively, what proportions would you expect to find in each class at the second generation?

(4)

(iv) If, in a large population, this transition matrix remains unchanged over a number of generations, approximately what proportions of males would you expect to find in each of the three occupational classes after several generations?

Explain carefully your reasoning and state any results about stationary distributions that you assume.

(4)

Turn over

3. A factory has two production lines, line 1 and line 2, for manufacturing car seats.

The preferred arrangement is to run both lines at standard speed. When running at standard speed, the lifetime of line 1 has an exponential distribution with a mean of 30 days and the lifetime of line 2 has an exponential distribution with a mean of 15 days.

There is one repair crew to deal with line failures, and repair times are exponentially distributed with mean 2 days.

If one line fails, the other will be run at double speed in order to meet production targets. In this case, the means of the lifetime distributions are then reduced to 10 days and 5 days for lines 1 and 2 respectively.

If both lines fail, the repair crew will repair line 1, because it is the more reliable, even if this means abandoning repair of line 2.

The lifetimes and repair times are statistically independent.

(i) According to this protocol, if the repair crew is repairing line 2 and line 1 fails, the crew immediately moves to line 1, abandoning the repair on line 2 until later. Explain why in setting up a model it is unnecessary to allow for the time spent initially repairing line 2.

(2)

(ii) Set up a continuous time Markov chain model for the state of the factory, defining the state space and writing down the instantaneous transition rates.

(6)

(iii) Find the corresponding equilibrium distribution, expressing the values as fractions and then calculating them correct to 2 decimal places. What, correct to 2 decimal places, is the long-term proportion of time during which neither line is running so that the factory is unable to meet the production target?

(12)

4. Consider a simple M/M/1 queue with arrival rate λ and service rate μ.

(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) Define the *traffic intensity* ρ and state the necessary and sufficient condition for an equilibrium distribution to exist.

(2)

(iii) Assuming that the condition for an equilibrium distribution to exist is satisfied, write down the detailed balance equations and deduce that the equilibrium distribution $\{\pi_n\}$ is given by

$$\pi_n = (1-\rho)\rho^n \qquad (n \geq 0).$$

(5)

Define a customer's waiting time to be the length of time from the time that he arrives to the time that he leaves the queue, his service time having been completed.

(iv) If a customer arrives to find the queue empty, write down the probability density function (pdf) of his waiting time.

(2)

(v) More generally, if a customer arrives to find n customers ahead of him in the queue, and if customers are served on a "first come, first served" basis, write down the pdf of his waiting time, explaining the reasons for your answer.

(3)

[**Note.** In answering this part of the question, you may assume that the sum of n independently and identically distributed random variables, each having an exponential distribution with parameter θ, has a gamma distribution with pdf

$$\frac{\theta^n x^{n-1} e^{-\theta x}}{(n-1)!} \qquad (x \geq 0).]$$

(vi) If a customer arrives to join the queue when it is in equilibrium, show that his waiting time distribution is exponential with parameter $\mu - \lambda$.

(5)

5. Consider the AR(2) model

$$Y_t = \tfrac{3}{4}Y_{t-1} - \tfrac{1}{8}Y_{t-2} + \varepsilon_t \qquad (-\infty < t < \infty)$$

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) Consider the infinite moving average representation of Y_t,

$$Y_t = \sum_{i=0}^{\infty} \psi_i \, \varepsilon_{t-i} \,.$$

By substituting this expression into the model equation, find a set of recurrence relations satisfied by the ψ_i together with appropriate initial conditions.

(6)

(iii) Solve these recurrence equations to show that

$$\psi_i = 2\left(\tfrac{1}{2}\right)^i - \left(\tfrac{1}{4}\right)^i \qquad (i \geq 0)\,.$$

(6)

(iv) Deduce that

$$\text{Var}(Y_t) = \tfrac{64}{35}\sigma^2\,,$$

where σ^2 is the variance of the white noise process $\{\varepsilon_t\}$.

(4)

Turn over

6. Consider modelling the series of monthly Australian sales of dry white wine in thousands of litres for the period from 1980 to 1994. In particular, let Y_t denote the natural logarithm of the sales for month t ($1 \leq t \leq 180$). 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) Suggest two different reasons 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 sales of dry white wine for December 1995, together with a 95% prediction interval, giving all your results correct to the nearest thousand litres.
(4)

Question continued on next page

Sample autocorrelation functions for qu 6

Lag	for $\{Y_t\}$	for $\{\Delta\Delta_{12}Y_t\}$
1	0.271	–0.411
2	0.024	–0.083
3	0.154	–0.068
4	0.198	0.080
5	0.003	0.004
6	–0.232	–0.005
7	–0.026	–0.087
8	0.160	0.120
9	0.118	–0.021
10	–0.047	0.029
11	0.163	0.132
12	0.721	–0.403
13	0.188	0.138
14	–0.013	0.025
15	0.094	0.141
16	0.110	–0.104
17	–0.035	0.049
18	–0.270	–0.055
19	–0.075	0.051
20	0.052	–0.089
21	0.074	0.030
22	–0.070	0.020
23	0.145	0.069
24	0.632	–0.095
25	0.162	0.062
26	–0.024	0.067
27	0.042	–0.123
28	0.060	–0.006
29	–0.066	–0.058
30	–0.245	0.115
31	–0.084	0.036
32	0.016	–0.048
33	0.052	–0.011
34	–0.084	–0.053
35	0.130	0.000
36	0.587	0.118
37	0.148	–0.127
38	–0.008	0.030
39	0.062	0.002
40	0.088	0.049

The computer output is on the next page

Computer output for question 6

```
Final Estimates of Parameters

Type        Coef  SE Coef      T      P
MA   1   0.8759   0.0379  23.12  0.000
SMA  12  0.7789   0.0550  14.16  0.000

Differencing: 1 regular, 1 seasonal of order 12
Number of observations:  Original series 180, after differencing 167
Residuals:    SS =  1.73152 (backforecasts excluded)
MS =  0.01049  DF = 165

Modified Box-Pierce (Ljung-Box) Chi-Square statistic
Lag            12     24     36     48
Chi-Square    9.4   20.6   35.1   52.3
DF             10     22     34     46
P-Value     0.492  0.547  0.414  0.243

Forecasts from period 180

                     95% Limits
Period  Forecast    Lower    Upper
   181   7.87442  7.67360  8.07525
   182   8.17948  7.97712  8.38184
   183   8.25273  8.04883  8.45662
   184   8.19244  7.98703  8.39785
   185   8.16790  7.96098  8.37482
   186   8.15206  7.94365  8.36047
   187   8.33409  8.12419  8.54399
   188   8.34188  8.13050  8.55325
   189   8.29073  8.07789  8.50357
   190   8.36048  8.14619  8.57478
   191   8.53378  8.31805  8.74952
   192   8.67702  8.45985  8.89420
```

7. The Holt-Winters forecasting procedure is to be used for a time series with multiplicative seasonal variation of period p.

(i) Let Y_t denote the observed value of the series at time t, L_t the local level, B_t the trend and I_t the seasonal index at time t. If α, γ and δ denote the smoothing constants for L_t, B_t and I_t, respectively, write down the updating equations for L_t, B_t and I_t.

(3)

(ii) Write down an expression for the forecast $\hat{y}_T(h)$ at time T for lead time h.

(2)

Monthly numbers of deaths in road accidents are recorded over a number of years and Holt-Winters forecasting with multiplicative seasonal variation of period 12 is used. In the output below, the numbers of deaths and other quantities that have been calculated month by month, using the smoothing constants α = 0.4, γ = 0.1 and δ = 0.01, are shown for the two years 1992 and 1993.

Year	Month	Deaths	Level	Trend	Index	Fitted	Residual
1992	Jan	234	301.69	-1.36	0.708	199.79	34.21
1992	Feb	232	329.64	1.57	0.622	186.49	45.51
1992	Mar	200	311.51	-0.40	0.709	234.93	-34.93
1992	Apr	229	306.43	-0.87	0.765	237.95	-8.95
1992	May	298	299.90	-1.43	1.022	312.49	-14.49
1992	Jun	314	282.73	-3.01	1.211	361.67	-47.67
1992	Jul	324	264.11	-4.57	1.345	376.54	-52.54
1992	Aug	390	268.77	-3.65	1.381	358.16	31.84
1992	Sep	316	270.88	-3.07	1.131	299.73	16.27
1992	Oct	317	273.61	-2.49	1.123	300.71	16.29
1992	Nov	310	285.79	-1.02	1.008	273.07	36.93
1992	Dec	322	303.84	0.88	0.969	275.82	46.18
1993	Jan	230	312.72	1.68	0.709	215.83	14.17
1993	Feb	225	333.39	3.58	0.622	195.50	29.50
1993	Mar	215	323.54	2.24	0.708	238.79	-23.79
1993	Apr	214	307.41	0.40	0.764	249.11	-35.11
1993	May	288	297.37	-0.64	1.022	314.70	-26.70
1993	Jun	321	284.09	-1.91	1.210	359.26	-38.26
1993	Jul	411	291.54	-0.97	1.346	379.52	31.48
1993	Aug	359	278.35	-2.19	1.380	401.20	-42.20
1993	Sep	323	279.94	-1.81	1.131	312.30	10.70
1993	Oct	323	281.90	-1.44	1.123	312.40	10.60
1993	Nov	308	290.51	-0.43	1.008	282.70	25.30
1993	Dec	333	311.44	1.70	0.970	281.23	51.77

(iii) Given the data as at December 1993, calculate, to the nearest whole number, forecasts of numbers of deaths for (a) January 1994 and (b) December 1994.

(4)

(iv) The number of deaths for January 1994 turned out to be 245. Given this fact, calculate the values of all the remaining figures in the corresponding row of the output.

(8)

(v) Using the whole historical run of the above series of deaths, it was decided to investigate what might be the best set of values to use for the smoothing constants α, γ and δ. Discuss how this might be done.

(3)

8. Let $\{Y_t\}$ be a process that satisfies the equation

$$(1-L)^2 Y_t = (1-\theta L)\varepsilon_t \qquad (-\infty < t < \infty),$$

where L is the backward shift operator and $\{\varepsilon_t\}$ is a white noise process. Suppose that the process $\{Y_t\}$ has been observed up to time T and consider the minimum mean square error forecast $\hat{y}_T(h) = E(Y_{T+h} \mid H_T)$ for lead time $h \geq 1$, where H_T represents the history of the process up to time T.

(i) State to which member of the family of ARIMA processes $\{Y_t\}$ belongs.

(1)

(ii) Expand the model equation to obtain an explicit expression for Y_t in terms of previous process values and white noise terms.

(1)

(iii) By setting $t = T + 1$ in the model equation, find an expression for $\hat{y}_T(1)$.

(2)

(iv) Write down a simple expression for the forecast error $Y_{T+1} - \hat{y}_T(1)$ as a white noise term.

(1)

(v) By setting $t = T + 2$ in the model equation, show that

$$\hat{y}_T(2) = 3Y_T - 2Y_{T-1} - 2\theta\varepsilon_T .$$

(3)

(vi) Show that, for $h \geq 3$,

$$\hat{y}_T(h) = 2\hat{y}_T(h-1) - \hat{y}_T(h-2) .$$

(2)

(vii) Deduce that

$$\hat{y}_T(h) = Y_T + b_T h \qquad (h \geq 1),$$

where

$$b_T = Y_T - Y_{T-1} - \theta\varepsilon_T .$$

(5)

(viii) Show that the $\{b_T\}$ satisfy the recurrence relation

$$b_T = (1-\theta)(Y_T - Y_{T-1}) + \theta b_{T-1} .$$

(5)