EXAMINATIONS OF THE HONG KONG STATISTICAL SOCIETY

GRADUATE DIPLOMA, 2004

Applied Statistics I

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 silent, cordless, non-programmable electronic calculators.

Where a calculator is used the ***method*** *of calculation should be stated in full.*

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* ${}^nC_r$.

This examination paper consists of 14 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. (i) Explain the terms *stationarity* and *autoregressive model* in the context of time series analysis.

(3)

(ii) Figure 1 shows a plot of a time series x_n. Describe the main features of this time series, including any unusual features.

(4)

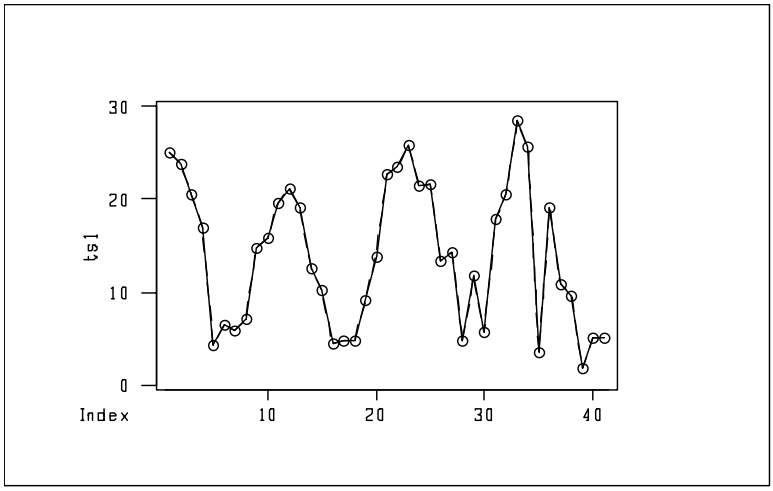


Figure 1. Time series plot

(iii) Figure 2 shows the correlogram associated with this time series. Describe its main features, relating your answer to the features you identified in (ii).

(3)

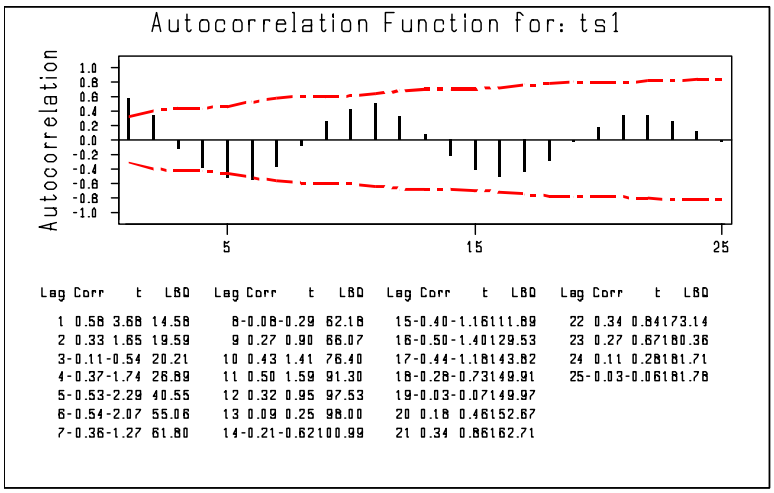


Figure 2. Correlogram of the time series.

Question 1 is continued on the next page

Turn over

(iv) Figure 3 shows the time series plot of the series y_n, given by $y_n = x_n - x_{n-11}$.

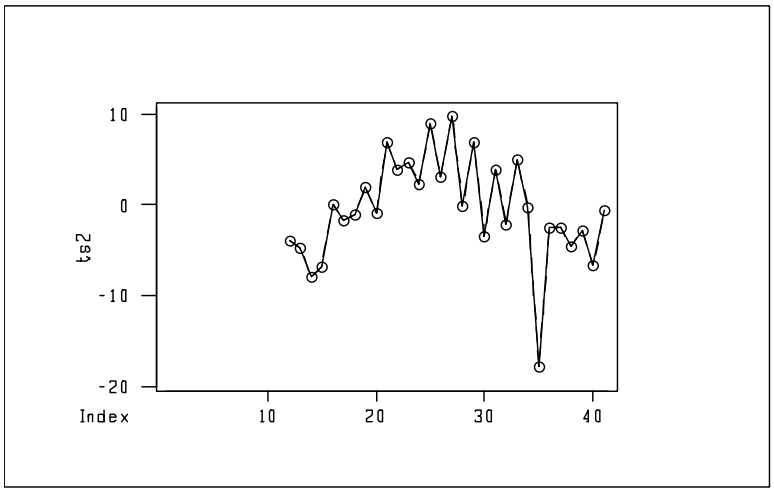


Figure 3. Time series plot of the 11-point differences

Justify the choice of 11-point differences, and describe the plot, contrasting it with the original series and commenting on any unexpected features of figure 3.

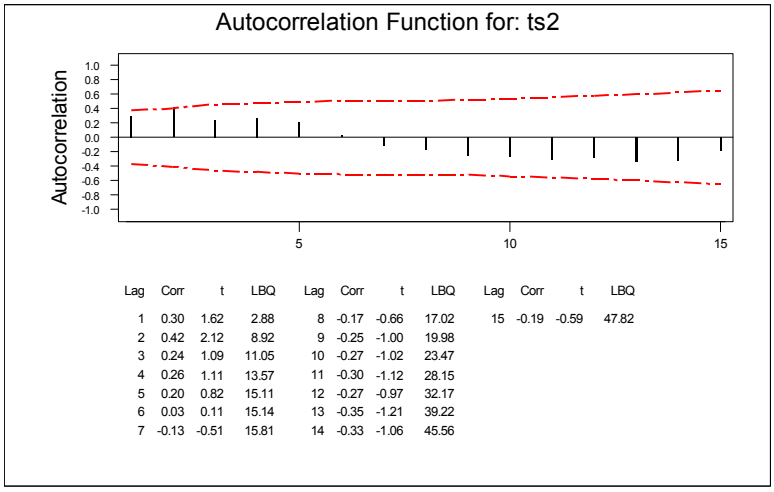
(3)

(v) Briefly describe anything you might do to investigate any unusual features identified in (ii).

(1)

(vi) Figure 4 is the correlogram corresponding to the series y_n. Describe the main features.

(2)



Lag	Corr	t	LBQ	Lag	Corr	t	LBQ	Lag	Corr	t	LBQ
1	0.30	1.62	2.88	8	-0.17	-0.66	17.02	15	-0.19	-0.59	47.82
2	0.42	2.12	8.92	9	-0.25	-1.00	19.98				
3	0.24	1.09	11.05	10	-0.27	-1.02	23.47				
4	0.26	1.11	13.57	11	-0.30	-1.12	28.15				
5	0.20	0.82	15.11	12	-0.27	-0.97	32.17				
6	0.03	0.11	15.14	13	-0.35	-1.21	39.22				
7	-0.13	-0.51	15.81	14	-0.33	-1.06	45.56				

Figure 4. Correlogram of the series y_n

(vii) Suggest suitable time series models to fit to y_n. What would the residuals look like from a well-fitting model?

(4)

Turn over

2. Consider the linear regression model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

where $\mathbf{X}$ is an $n \times (p+1)$ matrix of independent variables with $x_{i1} = 1$ ($i = 1, 2, \ldots, n$), and $\boldsymbol{\varepsilon} \sim \mathrm{N}\left(\mathbf{0},\ \sigma^2 \mathbf{I}_n\right)$. Denote the least squares estimator of $\boldsymbol{\beta}$ by $\hat{\boldsymbol{\beta}}$ and the fitted values by $\hat{\mathbf{Y}}$. Assume that

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}.$$

(i) Show that $\hat{\boldsymbol{\beta}}$ is an unbiased estimator of $\boldsymbol{\beta}$.

(1)

(ii) Find Var($\hat{\boldsymbol{\beta}}$).

(2)

(iii) Interpret each of the three terms shown in brackets {…} in the following identity:

$$\left\{\mathbf{Y}^T\mathbf{Y} - n\bar{Y}^2\right\} = \left\{\hat{\boldsymbol{\beta}}^T\mathbf{X}^T\mathbf{Y} - n\bar{Y}^2\right\} + \left\{(\mathbf{Y} - \hat{\mathbf{Y}})^T(\mathbf{Y} - \hat{\mathbf{Y}})\right\}.$$

Illustrate the use of these terms in an analysis of variance table, explaining the hypothesis of interest and giving the test statistic and its sampling distribution.

(8)

(iv) The Hat matrix $\mathbf{H}$ is defined by $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$. Show that

(a) $\mathbf{H}$ is symmetric, and $\mathbf{H}$ is idempotent (i.e. $\mathbf{HH} = \mathbf{H}$),

(b) $\hat{\mathbf{Y}} = \mathbf{H}\mathbf{Y}$,

(c) $\mathrm{Var}\left(\hat{\mathbf{Y}}\right) = \sigma^2\mathbf{H}$.

(4)

(v) The residuals are defined by $\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}}$. Show that

(a) $\mathbf{e} = (\mathbf{I} - \mathbf{H})\boldsymbol{\varepsilon}$,

(b) $E(\mathbf{e}) = \mathbf{0}$,

(c) $\mathrm{Var}(\mathbf{e}) = \sigma^2(\mathbf{I} - \mathbf{H})$.

(5)

3. (i) Suppose that variable FAC represents a factor with two levels via dummy coding, X is a continuous variable, and PROD is the product of FAC and X. These variables are to be used in selecting a linear model for a response variable Y. Use diagrams to illustrate fitting models with the following predictor variables in a regression command.

(a) X
(b) FAC X
(c) FAC X PROD
(d) X PROD

In each case, write down the form of the model, including any distributional assumptions.

(10)

(ii) A research laboratory is investigating the effect of two variables, the weight of the car and driver experience, on the car performance achieved on a test run under carefully controlled conditions. Weight is measured in kg, drivers are rated as "experienced" or "not experienced" and the performance is measured by km/litre of fuel used.

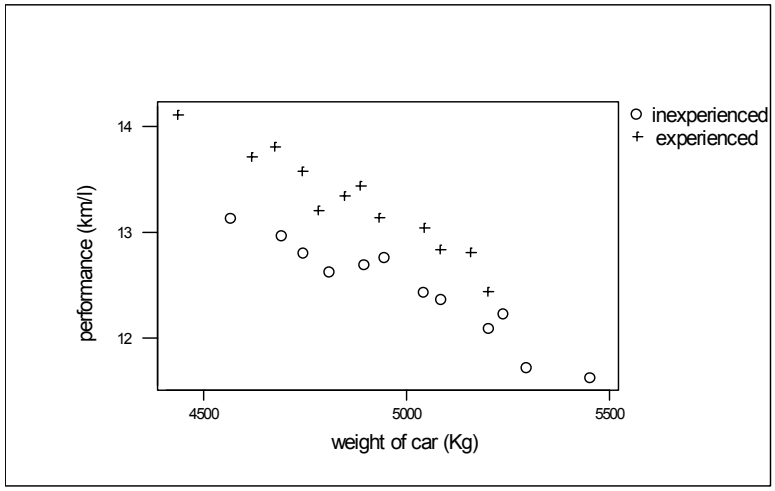
In a computer-based analysis, the weight is held in a variable WTKG, and the performance measurement is held in KMPERL. Two further variables are created in the following way:

EXPER = 0 if the driver is inexperienced
= 1 if the driver is experienced

INTER = WTKG * EXPER

(a) The figure shows a scatter plot of the data. Interpret this scatter plot.

(3)



Scatter plot from study on cars: the relationship between weight of car, experience of driver and km/l achieved.

Question 3 is continued on the next page

Turn over

(b) Using the computer output below, apply a forward selection technique and a 5% significance level to choose a model for the data. (There is no need to formally specify any hypotheses in terms of the model parameters.) Interpret the form of the selected model, relating it to the figure above.

(7)

MODEL 1

KMPERL = 23.3 – 0.00212 WTKG

Predictor	Coef	Stdev	t-ratio	p
Constant	23.332	1.333	17.51	0.000
WTKG	-0.0021213	0.0002700	-7.86	0.000

MODEL 2

KMPERL = 12.5 + 0.827 EXPER

Predictor	Coef	Stdev	t-ratio	p
Constant	12.4614	0.1366	91.21	0.000
EXPER	0.8268	0.1932	4.28	0.000

MODEL 3

KMPERL = 21.5 + 0.595 EXPER – 0.00181 WTKG

Predictor	Coef	Stdev	t-ratio	p
Constant	21.5115	0.5825	36.93	0.000
EXPER	0.59461	0.05778	10.29	0.000
WTKG	-0.0018123	0.0001164	-15.57	0.000

MODEL 4

KMPERL = 20.9 + 2.07 EXPER – 0.00169 WTKG – 0.000659 INTER

Predictor	Coef	Stdev	t-ratio	p
Constant	20.8767	0.7548	27.66	0.000
EXPER	2.069	1.142	1.81	0.085
WTKG	-0.0016852	0.0001509	-11.16	0.000
INTER	-0.0006595	0.0005099	-1.29	0.211

4. For each of the following scenarios, suggest a potentially suitable method of statistical analysis. Describe any other information you would need before selecting your method. Justify your answers with reference to the theoretical assumptions underlying your chosen method.

(a) A teacher wants to be able to predict pass/failure of students sitting an entrance examination to a university, based on their previous results in tests and homework. He has collected data on last year's group of students.

(5)

(b) A mail order company has collected data about its customers and wants to classify them in groups in order to send suitable advertising letters to them.

(5)

(c) A psychologist has collected data about sex, age, disease severity and treatment of a group of patients with a particular disease, together with their scores on a quality of life scale. She wants to examine relationships between the treatments and reported quality of life.

(5)

(d) A manager of a fire station has collected data on the number of house fires per week in his area over the past 5 years. He wants to predict house fires per week for the next 6 months.

(5)

5. The table below shows data relating to a study of people with cancer of the pancreas, and its possible association with a popular drink. Males and females were classified by the average amount of the drink they consumed each day. The number of people in each category is denoted by n_i, and the number with cancer of the pancreas is r_i. $X1$ is a code for the consumption rate and $X2$ is a dummy variable for sex.

Drink	**Sex**	r_i	n_i	$X1$	$X2$
0 cups per day	male	12	63	0	0
1–2 cups per day	male	65	210	2	0
3–4 cups per day	male	51	134	4	0
5 or more cups per day	male	29	75	5	0
0 cups per day	female	10	40	0	1
1–2 cups per day	female	95	216	2	1
3–4 cups per day	female	52	124	4	1
5 or more cups per day	female	63	148	5	1

(i) Draw a suitable graph to investigate relations between incidence of cancer of the pancreas and the variables, and hence describe the main features of the data.

(3)

(ii) Discuss critically the way the variable $X1$ is coded.

(2)

(iii) (a) Explain what is meant by the exponential family of distributions. State the probability mass function of the binomial distribution and show that it is a member of the exponential family.

(5)

(b) Several generalised linear models are fitted, using the binomial distribution with logit link function $\log\{\pi/(1-\pi)\}$. F1 is modelled as a factor with four levels (level 1 being "no cups per day", level 2 being "1–2 cups per day", etc, as in the data table above).

Copy the following table into your answer book, adding a column showing the numbers of degrees of freedom. Choose what you consider to be the "best" model on the basis of the evidence given, justifying your choice carefully.

(4)

Variables in model (in addition to the constant term)	Scaled deviance
–	27.373
$X2$	14.434
F1	8.969
F1 $X2$	2.051

(c) Do you think there would be any advantages in using the variable $X1$ instead of the factor F1? Justify your answer.

(2)

Question 5 is continued on the next page

(d) Further computer output from the model with F1 and $X2$ is given below.

_outcome	Coef.	Std. Err.	z	$P > \|z\|$
F1_level_1	0			
F1_level_2	.7579888	.2614022	2.900	0.004
F1_level_3	.8676501	.2728437	3.180	0.001
F1_level_4	.8591938	.2788847	3.081	0.002
sex	.3505732	.1336318	2.623	0.009
constant	–1.447907	.2479709	–5.839	0.000

Derive an approximate confidence interval relating the sex of the person to a relevant odds ratio, and hence discuss the apparent relationship between sex and pancreatic cancer for this population.

(4)

6. (i) Describe the purpose of principal component analysis, and explain why it is usually carried out on the correlation matrix.

(2)

(ii) What quantity is maximised by the first principal component of the correlation matrix?

Show that it follows from this property that the first principal component is an eigenvector of the correlation matrix.

(4)

(iii) Data have been collected from the first 48 runners to complete a 100-kilometre race. The data comprise the ages of the runners and the times in minutes that they took to run each 20-kilometre section of the race.

(a) The covariances and correlations of the times are shown below. (The variable time1 is the time in minutes taken for the first 20km, time2 that for the second 20km, etc.) Describe the patterns in the data.

(4)

Covariances

	time1	time2	time3	time4	time5
time1	113.7059				
time2	77.4105	72.4588			
time3	78.2676	76.9897	116.7553		
time4	85.4905	73.2750	134.2574	266.2151	
time5	95.6801	70.4827	97.1523	181.6348	288.0413

Correlations (Pearson)

	time1	time2	time3	time4
time2	0.853			
time3	0.679	0.837		
time4	0.491	0.528	0.762	
time5	0.529	0.488	0.530	0.656

(b) Explain why for these data it is appropriate to carry out principal component analysis on the covariance matrix. Interpret the first two principal components.

(3)

Eigenanalysis of the Covariance Matrix

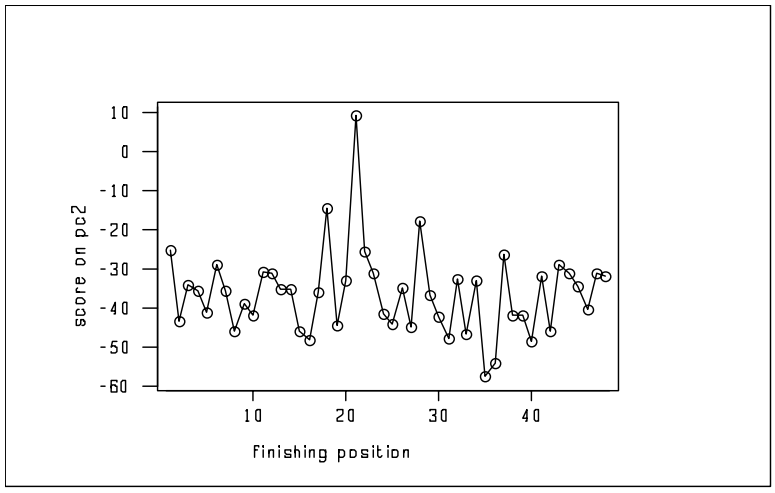
Eigenvalue	609.83	119.66	96.89	24.70	6.09
Proportion	0.711	0.140	0.113	0.029	0.007
Cumulative	0.711	0.851	0.964	0.993	1.000

Variable	PC1	PC2	PC3	PC4	PC5
time1	0.314	-0.368	-0.530	-0.620	0.316
time2	0.256	-0.379	-0.324	0.224	-0.798
time3	0.368	-0.429	0.036	0.661	0.493
time4	0.590	-0.102	0.717	-0.326	-0.144
time5	0.594	0.726	-0.315	0.146	0.015

Question 6 is continued on the next page

(c) The figure shows a plot of the second principal component against finishing position in the race. Interpret the plot.

(3)



Plot of score on second principal component against finishing position in the race

(d) What would you expect the plot of scores on the *first* principal component against finishing position to look like? Justify your answer.

(2)

(e) A researcher wants to investigate the relationship between runners' ages and their times for the sections of the race. Suggest analyses that she could do.

(2)

Turn over

7. (i) Explain briefly the purpose of cluster analysis.

(2)

(ii) When carrying out a cluster analysis of observations, why is it usually sensible to standardise the variables?

(2)

(iii) The tables below show 5 variables $X1$ – $X5$ measured on 6 items (the "raw data"). Standardised values, and the Euclidean distance matrix derived from the standardised values, are also given.

Raw Data

	*X*1	*X*2	*X*3	*X*4	*X*5
Observation 1	0.95	0.10	0.85	0.85	3.02
2	0.97	0.36	0.10	0.65	0.84
3	1.98	2.54	2.03	0.65	3.06
4	1.20	2.69	2.35	0.36	4.25
5	1.38	2.86	2.87	1.09	4.69
6	1.89	1.63	1.36	1.56	2.05
Mean	1.395	1.697	1.593	0.860	2.985
Sd	0.448	1.216	1.023	0.420	1.413

Standardised Data

	*X*1*S*	*X*2*S*	*X*3*S*	*X*4*S*	*X*5*S*
Observation 1	-0.99	-1.31	-0.73	-0.02	0.02
2	-0.95	-1.10	-1.46	-0.50	-1.52
3	1.31	0.69	0.43	-0.50	0.05
4	-0.44	0.82	0.74	-1.19	0.89
5	-0.03	0.96	1.25	0.55	1.21
6	1.10	-0.05	-0.23	1.67	-0.66

Distance Matrix

	2	**3**	**4**	**5**	**6**
1	1.79	3.30	3.02	3.42	3.09
2		3.78	3.88	4.57	3.50
3			2.08	2.23	2.50
4				1.89	3.83
5					3.04

(a) Describe how the standardised values and the distance matrix are obtained. In particular, reproduce the calculations leading to the standardised value (–0.99) for variable $X1$ for observation 1, and to the distance (1.79) between observations 1 and 2.

(3)

(b) Carry out a cluster analysis on the observations, using the standardised variables, and using single linkage and Euclidean distance. Show all your working and produce a dendrogram.

(5)

Question 7 is continued on the next page

(c) The figures below show dendrograms from various cluster analyses on these data. They all used Euclidean distance. Compare and contrast these dendrograms, explaining possible reasons for any differences in the apparent structure shown in them.

(8)

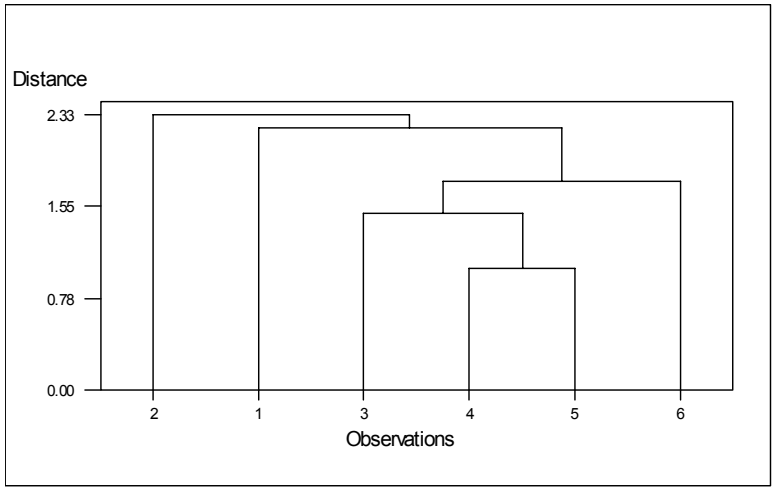


Figure (i). Cluster analysis with single linkage on the raw data

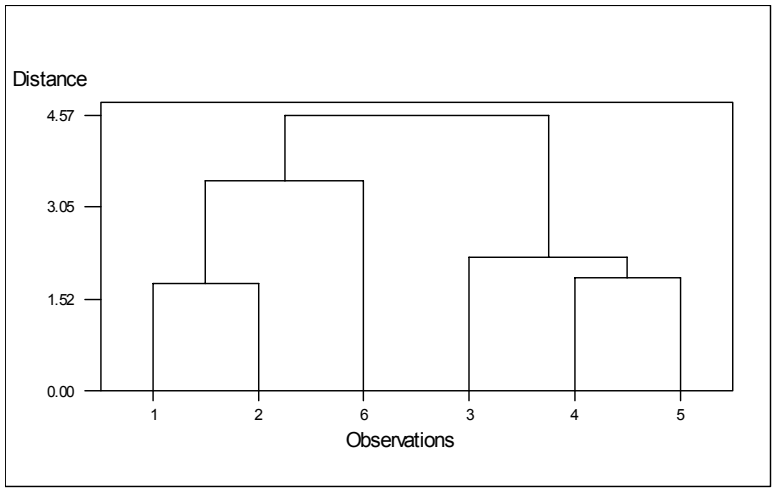


Figure (ii). Cluster analysis with complete linkage on standardised data

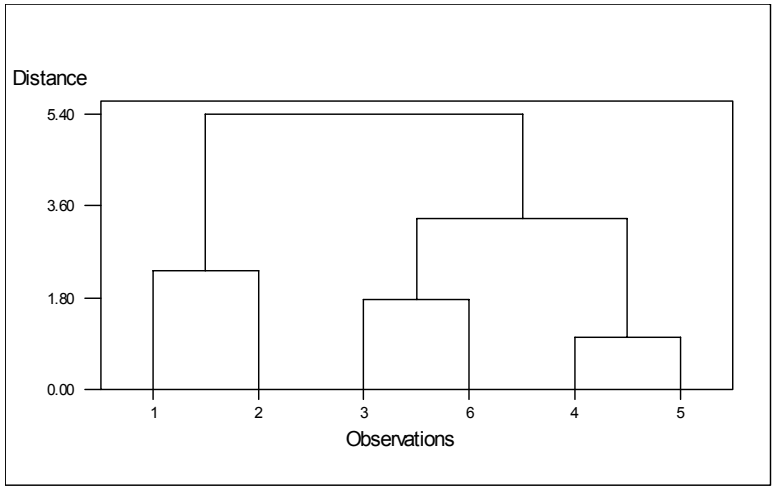


Figure (iii). Cluster analysis with complete linkage on raw data

8. A study was carried out to investigate the exposure to chemicals of men working in petroleum refineries. Three refineries were chosen at random from a large set of refineries and then within each refinery two days were chosen at random. The response was the exposure to o-cyclene (in parts per 10,000) on the 12 randomly chosen men studied.

The data are given below.

Refinery	1	1	1	1	2	2	2	2	3	3	3	3
Day	A	A	B	B	C	C	D	D	E	E	F	F
Response (parts per 10,000)	15	14	16	17	11	13	12	11	14	16	24	19

(i) Describe the data, pointing out any unusual features.

(3)

(ii) Write down a suitable model for these data, explaining all the terms in it and stating the necessary assumptions for it to be valid.

(5)

(iii) Construct an analysis of variance for the given data, complete your analysis and state your conclusions carefully.

(9)

(iv) Comment on the likely validity of the model assumptions for these data, and the implications for your result in (iii).

(3)