EXAMINATIONS OF THE HONG KONG STATISTICAL SOCIETY

GRADUATE DIPLOMA, 2012

MODULE 5 : Topics in applied statistics

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 13 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. Feedback was taken from 1428 students who were taught by a single lecturer. The lecturer's university department is interested in identifying patterns in the responses. The students provided a score for each statement on the scale 1 to 5. The worst score is 1 and the best is 5. The statements were as follows.

- A – 'Lecturer is well prepared.'
- B – 'Lecturer has scholarly grasp.'
- C – 'Lecturer is confident.'
- D – 'Lecturer focuses on examples.'
- E – 'Lecturer uses clear examples.'
- F – 'Lecturer is sensitive to students.'
- G – 'Lecturer allows time for questions.'
- H – 'Lecturer accessible to students outside class.'
- I – 'Lecturer aware of students understanding.'
- J – ‘I am satisfied with my performance.'
- K – 'Compared to other lecturers this one is ...' (The student needed to give a score)
- L – 'Compared to other courses this course was ...' (The student needed to give a score)

(i) Someone has suggested that principal components analysis and discriminant analysis might be appropriate for analysing these data. Discuss the objectives and limitations of these analyses and whether they are relevant to these data.
(7)

(ii) The output displayed **in the following pages** is from an analysis of these data. Name this analysis.
(1)

(iii) What does the correlation matrix (Table 1) reveal about the twelve variables of interest?
(3)

(iv) Looking at the output in Table 2 and the scree plot in Figure 1, how many components would you choose and focus on? Give the reasons behind your choice.
(3)

(v) Table 3 provides the loadings of the first two principal components. Given these loadings, interpret each of the two components.
(6)

Output for Question 1 is on the next 2 pages

Table1: Correlation matrix

	A	*B*	*C*	*D*	*E*	*F*	*G*	*H*	*I*	*J*	*K*	*L*
A	1	0.668	0.610	0.557	0.577	0.408	0.285	0.311	0.477	0.342	0.569	0.462
B	0.668	1	0.653	0.503	0.556	0.435	0.323	0.320	0.451	0.340	0.569	0.456
C	0.610	0.653	1	0.505	0.593	0.460	0.363	0.364	0.512	0.378	0.591	0.453
D	0.557	0.503	0.505	1	0.580	0.395	0.325	0.313	0.438	0.357	0.457	0.426
E	0.577	0.556	0.593	0.580	1	0.551	0.444	0.420	0.587	0.455	0.614	0.525
F	0.408	0.435	0.460	0.395	0.551	1	0.629	0.522	0.553	0.535	0.566	0.472
G	0.285	0.323	0.363	0.325	0.444	0.629	1	0.449	0.500	0.489	0.442	0.374
H	0.311	0.320	0.364	0.313	0.420	0.522	0.449	1	0.427	0.388	0.411	0.365
I	0.477	0.451	0.512	0.438	0.587	0.553	0.500	0.427	1	0.504	0.599	0.503
J	0.342	0.340	0.378	0.357	0.455	0.535	0.489	0.388	0.504	1	0.498	0.451
K	0.569	0.569	0.591	0.457	0.614	0.566	0.442	0.411	0.599	0.498	1	0.706
L	0.462	0.456	0.453	0.426	0.525	0.472	0.374	0.365	0.503	0.451	0.706	1

Table 2: Eigenvalue, variances and cumulative variances

Component	*Eigenvalues*	*% of Variance*	*Cumulative %*
1	6.249	52.076	52.076
2	1.229	10.246	62.322
3	0.719	5.992	68.314
4	0.613	5.109	73.423
5	0.561	4.676	78.099
6	0.503	4.192	82.291
7	0.471	3.927	86.218
8	0.389	3.240	89.458
9	0.368	3.066	92.524
10	0.328	2.735	95.259
11	0.317	2.645	97.904
12	0.252	2.096	100.000

Output continued on next page

Figure 1: Scree plot

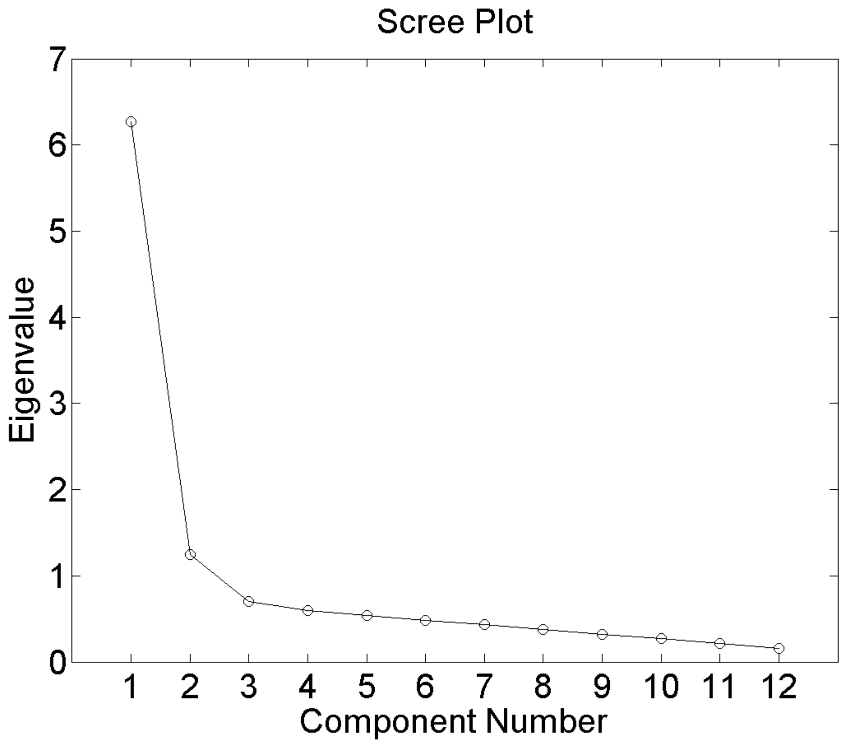


Table 3: Loadings for the first two principal components – PC1 and PC2

	A	*B*	*C*	*D*	*E*	*F*	*G*	*H*	*I*	*J*	*K*	*L*
PC1	0.841	0.811	0.761	0.697	0.670	0.316	0.144	0.187	0.477	0.240	0.636	0.527
PC2	0.153	0.182	0.271	0.250	0.462	0.777	0.798	0.677	0.611	0.707	0.518	0.481

2. (a) Outline the purpose of *cluster analysis*, and give reasons for using it.

(3)

(b) A doctor is interested in dividing patients (cases) who are referred for psychiatric treatment into groups with similar disorders. The patients were evaluated using four questionnaires: Spielberger Trait Anxiety Inventory (STAI), the Beck Depression Inventory (BDI), a measure of Intrusive Thoughts and Rumination (IT) and a measure of Impulsive Thoughts and Actions (Impulse). The rationale behind this analysis is that people with the same disorder should report a similar pattern of numerical scores across the measures (so the profiles of their responses should be similar). The table below shows the range of scores for the four questionnaires.

Questionnaire	*Minimum*	*Maximum*
STAI	0	80
BDI	0	70
IT	0	70
Impulse	0	40

(i) Describe *k-means* and *hierarchical clustering*. What type of cluster analysis would you carry out for these data and why? Would you standardise these variables?

(6)

(ii) Figure 1 is a dendrogram constructed from these data. What clustering method has been used to obtain it? How many possible clusters do you see? Are there any cases that are very dissimilar to the rest? If such cases exist, would you discard them from further analysis or would you investigate them independently? (Justify your answer.)

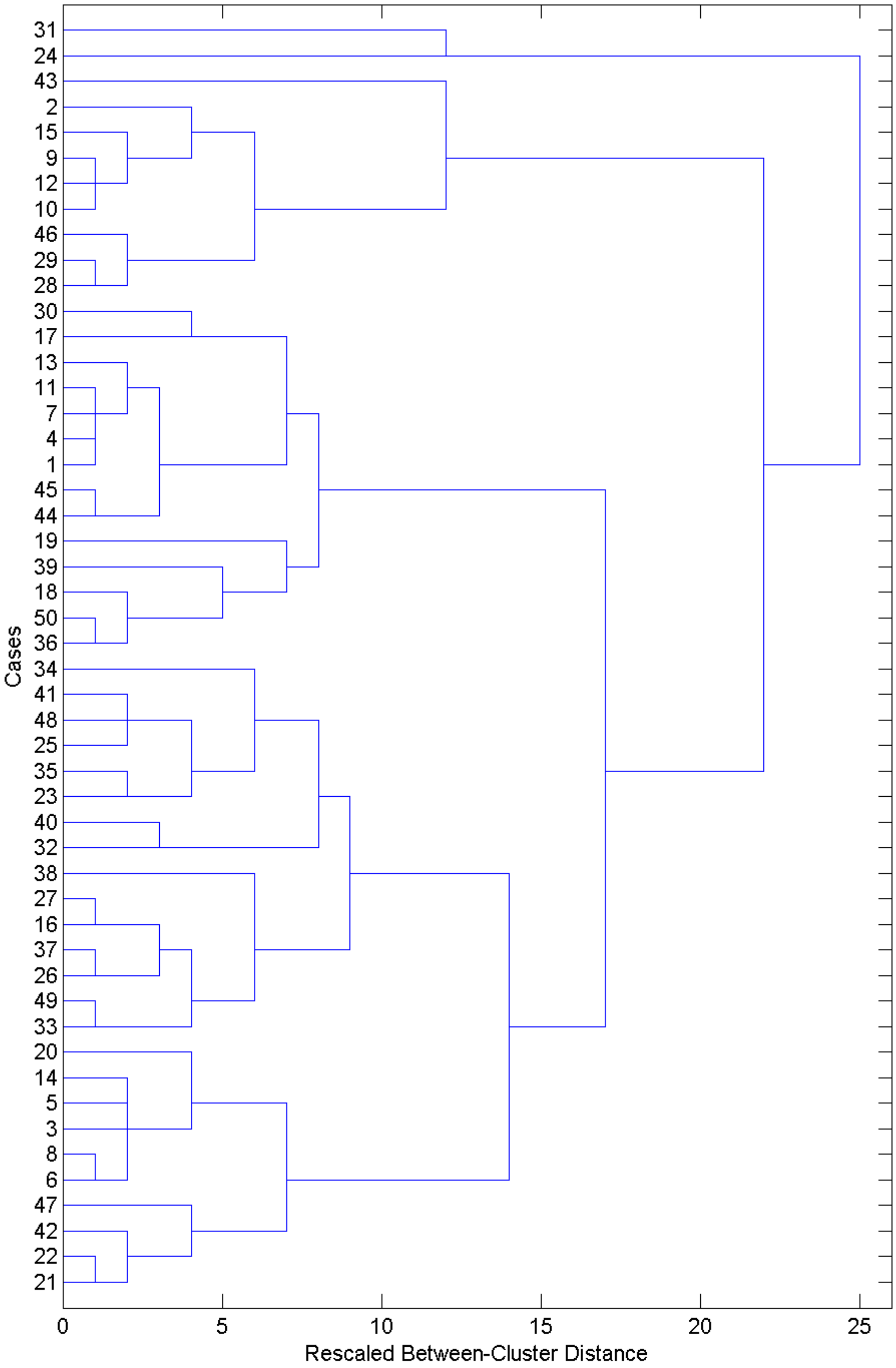
(7)

(iii) Figure 2 is another dendrogram constructed from these data. In what way does it differ from the dendrogram in Figure 1? Do you think it was sensible to carry out the type of cluster analysis that has produced the dendrogram in Figure 2?

(4)

Diagrams for Question 2 are on the next 2 pages

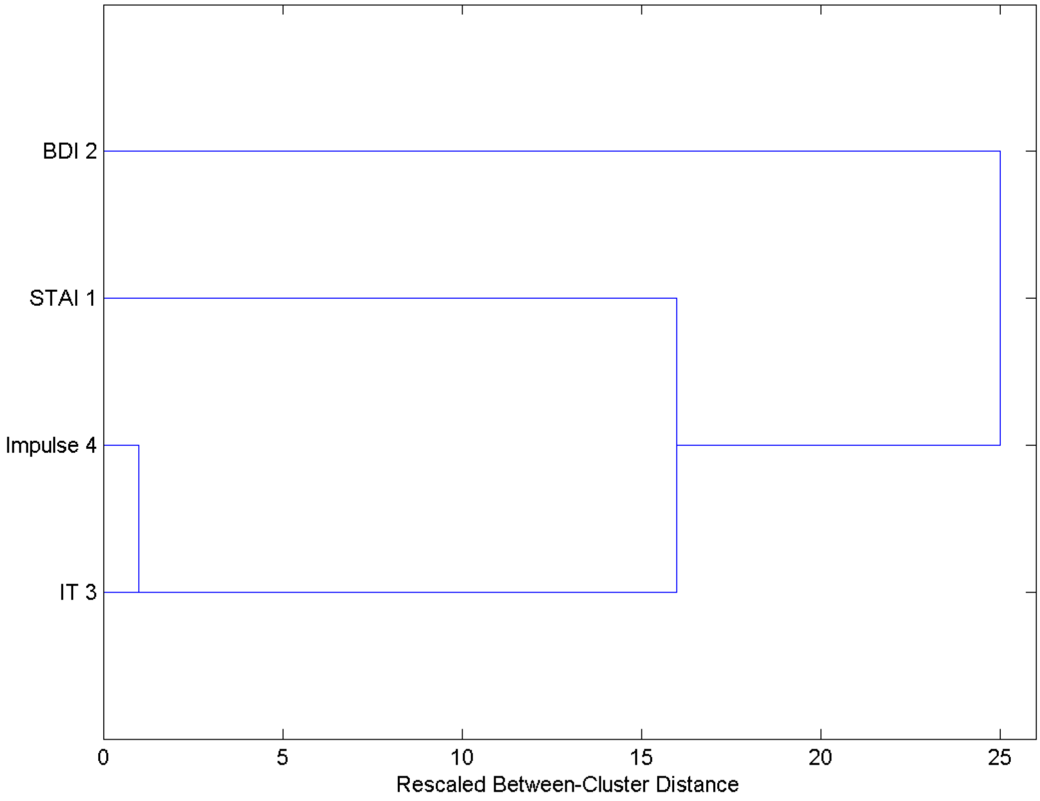
Figure 1



Diagrams continued on next page

Turn over

Figure 2


BDI 2
STAI 1
Impulse 4
IT 3
0
5
10
15
20
25
Rescaled Between-Cluster Distance

3. A study was conducted to consider the effect of two types of graft (allogenic or autologous) for bone marrow transplants. Three other explanatory variables were recorded: the type of disease (Non-Hodgkin's lymphoma or Hodgkin's lymphoma), Karnofsky score and waiting time to transplant in months.

(i) Describe the *Cox proportional hazards model* and explain how it differs from a *parametric proportional hazards model*.

(4)

(ii) The Cox model was fitted to the data. Graft is coded 1 for allogenic and 2 for autologous and disease is coded 1 for Non-Hodgkin's lymphoma and 2 for Hodgkin's lymphoma. Some results are shown below.

```
             coef    se(coef)
Disease    0.981     0.523
Graft     -0.233     0.443
Score     -0.056     0.012
Waiting   -0.008     0.008
```

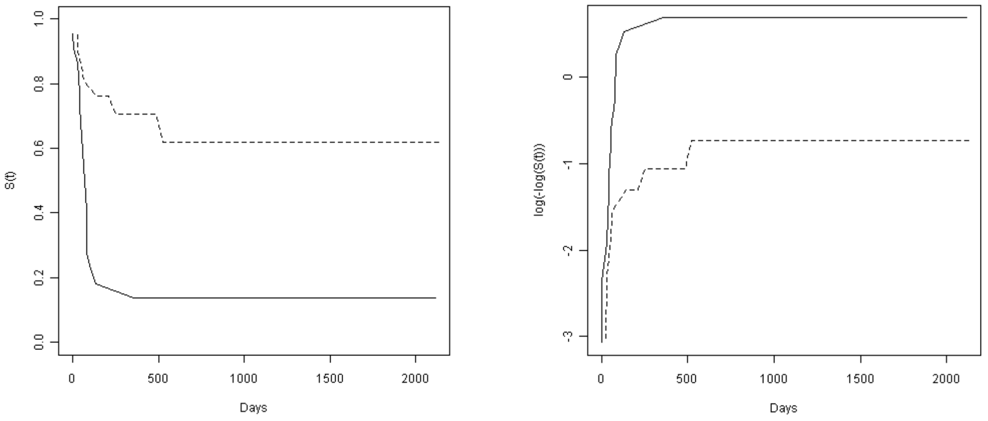
(a) What can be concluded from the results about the effect of these factors on the hazard?

(6)

(b) Construct a 95% confidence interval for the hazard ratio of a patient with a Karnofsky score of 30 compared with a patient with a Karnofsky score of 70 assuming that all other values of the explanatory variables are the same. Give a clear explanation of what the hazard ratio represents.

(6)

(iii) The data were divided into two groups according to the Karnofsky score. The plots below show the Kaplan-Meier estimates of the survivor function (left-hand graph) and the log of the cumulative hazard function (right-hand graph) for the two groups, with the first group shown as the dashed line and the second group shown as the solid line.



Describe the shapes of the survivor function $S(t)$ for the two groups. Comment on the validity of the assumptions of the Cox model in the light of these graphs.

(4)

4. (i) Define the *hazard function* of a random variable T measuring lifetime.

(2)

(ii) Suppose that T has the Weibull distribution with density

$$f(t) = \lambda\gamma(\lambda t)^{\gamma-1}\exp\ -(\lambda t)^{\gamma}$$

and cumulative distribution function

$$F(t) = 1 - \exp\ -(\lambda t)^{\gamma}\ .$$

Derive the hazard function and state the ranges of values of γ for which the hazard function is

(a) increasing,

(b) decreasing,

(c) constant.

(3)

(iii) The following analysis concerns a clinical trial for patients with Acute Myelogenous Leukemia. One group of patients was given a standard course of chemotherapy (Nonmaintained) whilst the second group was given a longer course of chemotherapy (Maintained). The following computer output relates to the fitting of a parametric proportional hazards model with a Weibull distribution.

```
                   Value    Std. Error
(Intercept)       -4.109      0.300
Nonmaintained      0.929      0.383
Log(gamma)        -0.235      0.178
```

(a) Construct a 95% confidence interval for the treatment effect and comment on the effectiveness of the maintained treatment relative to the nonmaintained treatment.

(4)

(b) Test whether an exponential distribution would be inappropriate for these data.

(4)

(c) Write down the proportional hazards model, using the notation in part (ii), and describe how you would find an estimate of λ from the information in the computer output. Write down estimates of the hazard function for the two groups.

(4)

(d) Describe how the suitability of this proportional hazards model could be evaluated using an appropriate graphical method.

(3)

5. Survival data, for up to 8 years after diagnosis, are shown in the table below for 2418 males diagnosed to have angina pectoris. Year *i* after diagnosis denotes the time interval [*i*, *i* + 1). For the *i*th time interval: n_i' denotes the number of patients entering, l_i denotes the number who were lost to follow-up for various reasons, w_i denotes the number who withdrew alive from the study, and d_i denotes the number who died.

Year i after diagnosis	[*i*, *i* + 1)	n_i'	l_i	w_i	d_i	n_i	$\hat{q}_i$	$\hat{p}_i$	$\hat{S}(t_i)$	$\hat{h}(t_{mi})$
0	[0, 1)	2418	0	0	456					
1	[1, 2)	1962	9	30	226					
2	[2, 3)	1697	10	12	152					
3	[3, 4)	1523	0	23	171					
4	[4, 5)	1329	9	15	135					
5	[5, 6)	1170	10	97	125					
6	[6, 7)	938	25	108	83					
7	[7, 8)	722	15	87	74					
8	[8, 9)	546	8	60	51					

(i) Explain the notation n_i, $\hat{q}_i$ and $\hat{p}_i$ as used in clinical life tables, and show how each is calculated in terms of n_i', l_i, w_i and d_i. Determine also $\hat{S}(t_i)$, the estimate of the probability of survival to each year after diagnosis. Copy the table and complete all but the last column, i.e. the column headed $\hat{h}(t_{mi})$.

(10)

(ii) The hazard function *h*(*t*) is defined as the probability of death given survival to date and, for the *i*th interval, may be estimated at the mid-point t_{mi} by $\hat{h}(t_{mi}) = \dfrac{2\hat{q}_i}{1+\hat{p}_i}$. Fill in the last column of your table with the appropriate values.

(2)

(iii) Plot graphs of

(a) $\hat{S}(t_i)$ against time t_i,

(b) $\hat{h}(t_{mi})$ against the interval mid-point t_{mi}.

Comment on the inferences that can be made from each graph.

(8)

6. (i) Define the *sensitivity* and *specificity* of a diagnostic test. Derive expressions for positive predictive value and negative predictive value in terms of sensitivity, specificity and prevalence. Explain briefly what information each of these quantities gives about the possibility of detecting a disease, and comment on why sensitivity and specificity may be preferred to positive and negative predictive values in deciding how good a diagnostic test is.

(7)

(ii) A biochemical marker is being used to test whether each of 130 patients suspected of having a particular disease can be diagnosed by a new diagnostic test. A patient is deemed to have the disease if the value of the marker is greater than or equal to a specified cut-off value. The patients have also been classified, by more detailed medical tests, as having, or not having, the disease. Three cut-off values of interest are 1, 2 and 3 units of the marker.

Of the 130 patients in the trial, 6 were medically classified as having the disease; of these, 3 had a marker value of 3 units or more, 4 had a value of 2 units or more and 5 had a value of 1 unit or more. Of the remaining 124 not medically classified as having the disease, none had a marker level of 3 units or more, 4 had a level of 2 units or more and 7 had a level of 1 unit or more.

(a) Show these results in three tables, one for each marker level.

(4)

(b) Calculate the sensitivity and specificity for each of the three levels of the marker.

Given these calculated values for sensitivity and specificity, determine the positive and negative predictive values of the test, at each of the three marker levels, in a population where prevalence of the disease is 10%.

(5)

(c) Sketch the ROC curve for these three marker levels, and comment on which level might be preferred.

(4)

7. (i) In a population which contains N units, two measurements x, y can be taken on each unit. The ratio $R = \frac{Y}{X}$ is of interest. A simple random sample of n units is taken and the estimate $\hat{R} = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i} = \frac{\bar{y}}{\bar{x}}$ is calculated. Show that the variance $V(\hat{R})$ is approximately

$$\frac{1-f}{n\bar{X}^2} \times \frac{\sum_{i=1}^{N} (y_i - Rx_i)^2}{N-1},$$

where $R = \frac{\bar{Y}}{\bar{X}}$ is the ratio of the population means and $f = \frac{n}{N}$. Explain why this is an approximation. (6)

(ii) National income from manufacturing in a country is being estimated for the current year by taking a simple random sample of n = 6 out of the N = 19 industrial categories that report their results, y, in the early part of the year. (Experience suggests that, by scaling up, a satisfactory estimate for the whole year can be found by this method.) Incomes, x, from all these 19 categories in a recent base year are also available, and the total of these was $674 billion.

	$billion	
Industry	x_i	y_i
Lumber and wood products	21	26
Electric and electronic equipment	63	91
Motor vehicles and equipment	35	47
Food and kindred products	60	70
Textile mill products	16	17
Chemical and allied products	50	76
Total	245	327

$\sum_{i=1}^{6} y_i^2 = 22131$ (current year); $\sum_{i=1}^{6} x_i^2 = 11991$ (base year); $\sum_{i=1}^{6} x_i y_i = 16196$.

Estimate national income from manufacturing in the current year using

(a) the data for y in the simple random sample,

(b) a ratio estimator. (4)

(iii) Estimate the variances of the estimators in part (ii) and compare their efficiencies. Comment on the results and state, with reasons, which method you consider more appropriate. (8)

(iv) Discuss the possible extent of bias there might be in your ratio estimate and the estimate of its variance. (2)

8. (i) Explain briefly how samples are chosen when applying (a) *systematic* and (b) *one-stage cluster* sampling methods. Comment on the use of each of these two methods as alternatives to stratified random sampling.

(7)

(ii) The total weight (kg) of catch landed by fishing boats at a harbour over a 12-hour period is to be estimated by recording the weight landed during a sample of complete hours. In a preliminary trial, two possible suggested sampling methods were

(a) a systematic sample of three separate hours,

(b) a cluster sample of two clusters, chosen from 6 two-hour clusters; for this purpose hours 1 and 2 are treated as cluster 1, hours 3 and 4 as cluster 2, hours 5 and 6 as cluster 3, hours 7 and 8 as cluster 4, hours 9 and 10 as cluster 5 and hours 11 and 12 as cluster 6.

The weights of the hourly catches landed during this trial were as follows.

Hour	1	2	3	4	5	6	7	8	9	10	11	12
Weight (kg)	567	861	231	92	347	1117	946	1301	465	444	96	0

Suppose that the systematic sample chosen consisted of hours 2, 6 and 10, and that the clusters chosen were cluster 2 and cluster 4.

Estimate the total catch predicted for the 12-hour period using the chosen systematic sample. Compare this estimate with those that would be given by using the other three possible systematic samples and with the total for the whole trial.

Comment briefly on these results.

(6)

(iii) Estimate the total catch predicted by the chosen cluster sample. Given that the standard error of this estimate is 3848, calculate an approximate 95% confidence interval for this total.

(3)

(iv) On reviewing the results, the question is asked whether sampling can give reasonable estimates at all, and a simple random sample of 4 hours is taken for comparison. The hours 2, 4, 9 and 10 are used. Estimate the total catch using this sample, together with its standard error.

How would you answer the question asked?

(4)