THE ROYAL STATISTICAL SOCIETY

GRADUATE DIPLOMA EXAMINATION

NEW MODULAR SCHEME

introduced from the examinations in 2009

MODULE 5

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. (a) Compare the aims of principal component analysis and cluster analysis.

(4)

(b) Data are available from the 1996 Olympic Games, describing the complete results of all 31 competitors in the decathlon event. The variables are defined as follows.

m100	100 metres, time in seconds
longj	long jump, in metres
shot	shot put, in metres
hjump	high jump, in metres
m400	400 metres, in seconds
m110h	110 metres hurdles, in seconds
discus	discus, in metres
polevlt	pole vault, in metres
jav	javelin, in metres
m1500	1500 metres, in seconds

The output **on the next three pages** gives the results of multivariate analyses of these data. Making reference to this output, answer the following questions.

(i) Describe the correlations between the variables, identifying any possible clusters of variables.

(2)

(ii) Using the results of the principal components analysis, draw a scree plot. State how many principal components you would use to summarise the data, justifying your answer.

(4)

(iii) Interpret the first four principal components.

(4)

(iv) A cluster analysis was performed on the 31 observations, with dissimilarities defined as the raw Euclidean distances between the points in the dataset described above. Comment on the validity of such a cluster analysis.

(2)

(v) Compare and contrast the information given in the dendrogram and the labelled plots of the first four principal components.

(4)

Turn over

Output for question 1. This output is printed on this page and the next two pages

Correlations (Pearson)

	m100	longj	shot	hjump	m400	m110h	discus	polevlt	jav
longj	-0.405								
shot	-0.196	0.251							
hjump	0.101	0.285	0.117						
m400	0.685	-0.215	-0.087	0.210					
m110h	0.530	-0.325	-0.155	-0.036	0.351				
discus	-0.280	0.306	0.349	0.216	-0.004	-0.380			
polevlt	-0.337	0.328	0.069	0.043	-0.224	-0.061	0.161		
jav	-0.344	0.281	0.163	0.083	-0.027	-0.429	0.297	0.185	
m1500	-0.225	-0.002	-0.214	0.207	-0.098	-0.205	0.307	-0.008	0.013

Output from principal component analysis

Principal Component Analysis

Eigenanalysis of the Correlation Matrix

Eigenvalue	3.0084	1.5649	1.2961	1.0357	0.8810	0.7503
Proportion	0.301	0.156	0.130	0.104	0.088	0.075
Cumulative	0.301	0.457	0.587	0.691	0.779	0.854

Eigenvalue	0.5154	0.4176	0.3195	0.2112
Proportion	0.052	0.042	0.032	0.021
Cumulative	0.905	0.947	0.979	1.000

Variable	PC1	PC2	PC3	PC4	PC5	PC6
m100	0.472	0.322	-0.065	-0.018	-0.074	-0.027
longj	-0.375	0.169	-0.217	0.326	0.008	0.390
shot	-0.222	0.220	-0.504	-0.288	0.515	-0.160
hjump	-0.088	0.586	0.121	0.345	0.135	0.437
m400	0.315	0.538	-0.043	-0.086	-0.309	-0.183
m110h	0.416	0.039	-0.171	0.371	0.136	-0.251
discus	-0.341	0.378	0.121	-0.184	0.176	-0.507
polevlt	-0.254	-0.083	-0.216	0.653	-0.237	-0.489
jav	-0.323	0.157	-0.101	-0.275	-0.700	0.006
m1500	-0.149	0.105	0.760	0.103	0.141	-0.191

Variable	PC7	PC8	PC9	PC10
m100	0.149	-0.207	0.087	-0.768
longj	0.586	0.371	0.162	-0.132
shot	-0.322	0.097	0.393	-0.083
hjump	-0.428	-0.237	-0.240	0.098
m400	0.211	0.019	0.370	0.539
m110h	-0.178	0.665	-0.321	0.023
discus	0.338	-0.068	-0.533	-0.033
polevlt	-0.112	-0.328	0.192	-0.060
jav	-0.359	0.328	-0.103	-0.220
m1500	-0.128	0.303	0.427	-0.179

Turn over

Output for question 1 (contd)

Labelled plots of principal component scores for the 31 athletes

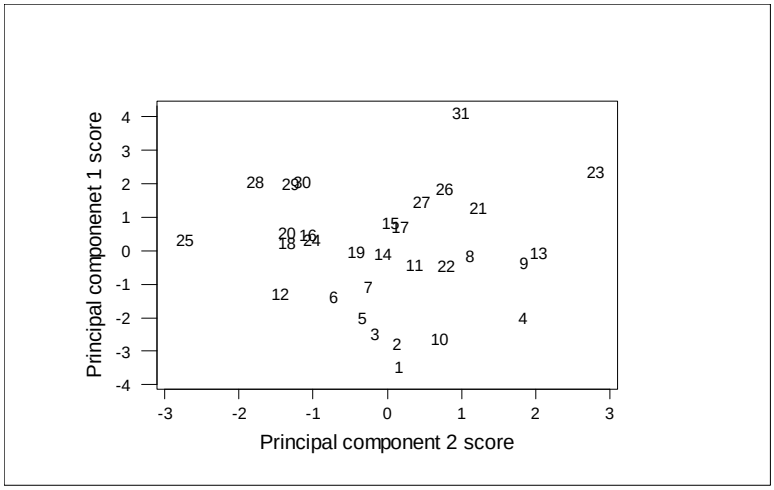


Figure 1.1. Plot of principal component score 1 against principal component score 2

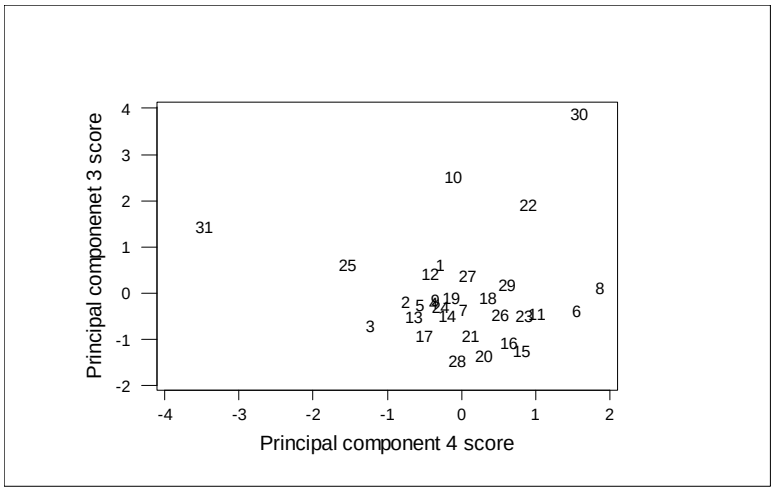


Figure 1.2. Plot of principal component score 3 against principal component score 4

Turn over

Output for question 1 (contd)

Output from cluster analysis of observations, using Euclidean distance and average linkage

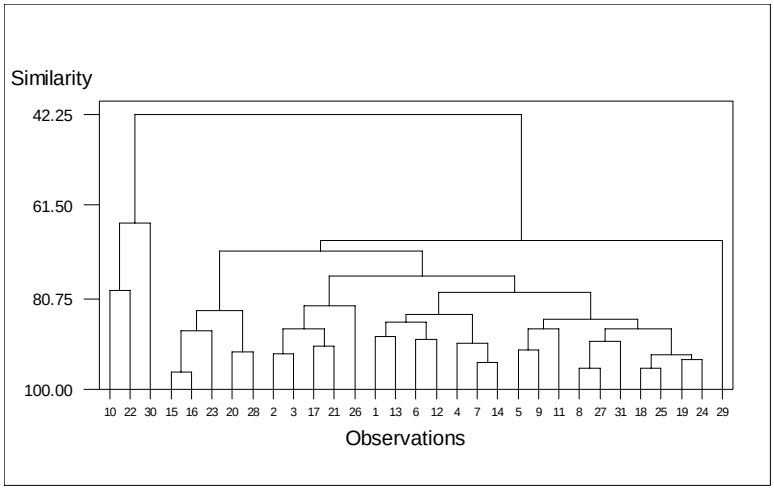


Figure 1.3. Dendrogram from cluster analysis

Turn over

2. Suppose that $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$ are a random sample of p-variate vectors from a multivariate Normal distribution with unknown mean $\boldsymbol{\mu}$ and unknown covariance matrix $\boldsymbol{\Sigma}$.

(i) Write down the form of a test statistic T^2 for testing H_0: $\boldsymbol{\mu} = \boldsymbol{\mu}_0$ against H_1: $\boldsymbol{\mu} \neq \boldsymbol{\mu}_0$ where $\boldsymbol{\mu}_0$ is a specified vector of constants, stating whether rejection of H_0 is indicated by large or small values of the statistic.

What is the name of your test statistic?

(4)

(ii) Under H_0, it is known that $\dfrac{n-p}{p(n-1)}T^2 \sim F_{p,n-p}$. Show that for univariate data the test reduces to the usual one-sample t test.

(4)

(iii) Demonstrate how a confidence region for $\boldsymbol{\mu}$ can be obtained from the test procedure in part (i). Give a geometrical description of this region.

(5)

(iv) A medical researcher is interested in two particular fatty acids (A and B) found in human blood. Measurements (micrograms per gram) were taken on 16 new-born babies with Down's syndrome. The sample means were 70 and 50 for fatty acids A and B respectively, giving the sample mean vector

$$\bar{\mathbf{x}} = \begin{pmatrix} 70 \\ 50 \end{pmatrix},$$

and the corresponding (unbiased) sample covariance matrix was

$$\mathbf{S} = \begin{pmatrix} 100 & 80 \\ 80 & 100 \end{pmatrix}.$$

For non-Down's syndrome new-born babies the expected fatty acid levels are 80 and 65 for A and B respectively. Use the multivariate hypothesis test described in part (i) to assess whether the observed data for the Down's syndrome babies are consistent with the expected values for non-Down's syndrome babies.

(7)

Turn over

3. (a) Explain what is meant by the *hazard function* in survival analysis.

The Weibull survival distribution can be characterised by the hazard function $h(t) = \lambda\gamma t^{\gamma-1}$. Define and derive the corresponding survival function, $S(t)$, and the probability density function, $f(t)$, for the Weibull distribution.

(8)

(b) The table below shows the Kaplan-Meier estimate of the survival function $S(t)$ for 13 patients for the time to removal of a catheter following a kidney infection. Sometimes the catheter has to be removed for reasons other than infection, giving rise to right-censored observations.

Survival analysis for "Time", the number of days from insertion of catheter until removal

Time	Status	Cumulative Survival	Standard Error	Cumulative Events	Number Remaining
8.0	Catheter removed	.9231	.0739	1	12
15.0	Catheter removed	.8462	.1001	2	11
22.0	Catheter removed	.7692	.1169	3	10
24.0	Catheter removed	.6923	.1280	4	9
30.0	Catheter removed	.6154	.1349	5	8
54.0	Censored			5	7
119.0	Catheter removed	.5275	.1414	6	6
141.0	Catheter removed	.4396	.1426	7	5
185.0	Catheter removed	.3516	.1385	8	4
292.0	Catheter removed	.2637	.1288	9	3
402.0	Catheter removed	.1758	.1119	10	2
447.0	Catheter removed	.0879	.0836	11	1
536.0	Catheter removed	.0000	.0000	12	0

Number of Cases: 13 Censored: 1 (7.69%) Events: 12

(i) Use a graphical method based on the estimated cumulative hazard function for checking whether the data may reasonably be assumed to come from a Weibull distribution.

(8)

(ii) Use the graph to estimate λ and γ, the parameters of the Weibull distribution.

(4)

4. (i) In a study to compare two treatments for venous leg ulcers, patients were randomised to receive one of two treatments: treatment at a dedicated community leg-ulcer clinic (Intervention) by specially trained nurses, or normal home treatment (Control) by district nurses. They were followed up until either the initial leg ulcer healed or one year from randomisation. Healing times (weeks) for a random sample from the Intervention group were as follows.

3 5 8 10 11* 27 39 52* 52* 52*

(* A star indicates a right-censored observation.)

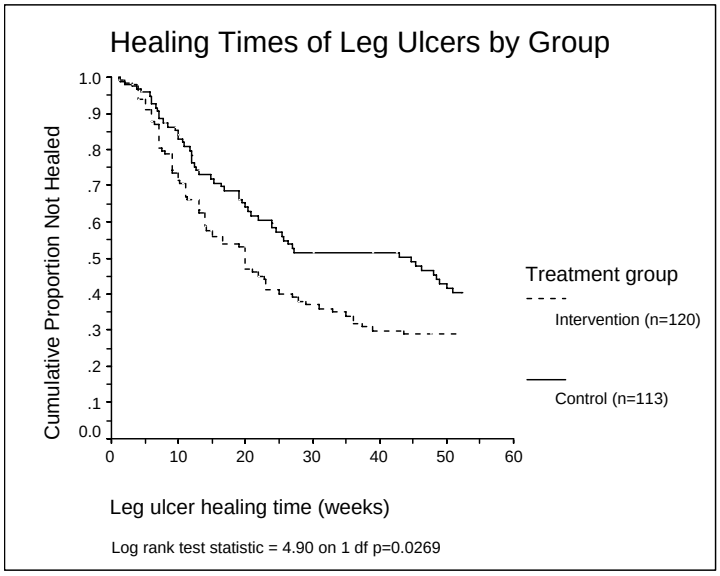
Explain what is meant by a *right-censored* observation. Also explain the meaning of the term *hazard function*.

(4)

(ii) Construct the Kaplan-Meier survival curve for the Intervention group and show it on a suitable graph.

(6)

(iii) The full trial involved 233 patients, with 120 randomised to the Intervention group and 113 to the Control group. The figure below shows Kaplan-Meier estimates of survival functions for healing times for the two treatment groups, and the results of a log-rank test; the test statistic was 4.90 on 1 d.f. (p = 0.0269).



Use the diagram to estimate the median leg ulcer healing times for the two treatment groups. Is there a difference between the survival patterns of the two treatment groups? Comment on the results of the log-rank test.

(4)

Question continued on next page

(iv) Previous studies have suggested that age, initial ulcer size, ulcer duration and history of deep vein involvement are important factors in predicting leg ulcer healing time. Cox proportional hazards regression analysis was used to adjust healing times for these prognostic variables.

The following tables show abbreviated computer output from two Cox regression models. In Analysis 1, a simple regression of ulcer healing time on treatment group alone was performed. Analysis 2 involved a multiple regression of ulcer healing time on age (years), ulcer duration (months), base area (cm^2), deep vein involvement (DVT) and treatment group. (Note that DVT was coded as "No deep vein involvement" = 0 and "Deep vein involvement" = 1. Similarly, treatment group was coded as 0 = Control and 1 = Intervention.)

Analysis 1: Cox regression – model: group

```
No. of subjects =    213            Number of obs   =       213
No. of failures =    129
LR chi2(1)      =   4.85            Prob > chi2     =    0.0276
```

	Haz.Ratio	z	P>|z|	[95% Conf.	Interval]
group	1.480299	2.18	0.029	1.041186	2.104605

Analysis 2: Cox regression – model: age basearea duration dvt group

```
No. of subjects =    213              Number of obs   =       213
No. of failures =    129
LR chi2(5)      =  48.25              Prob > chi2     =    0.0000
```

	Haz.Ratio	z	P>|z|	[95% Conf.	Interval]
age	1.006661	0.82	0.413	.9907812	1.022794
basearea	.9642388	-4.04	0.000	.9473715	.9814064
duration	.9967765	-1.39	0.165	.9922462	1.001327
dvt	.9172132	-0.39	0.695	.5958852	1.411816
group	1.651079	2.75	0.006	1.15515	2.35992

(a) How well can an individual's leg ulcer healing time be predicted from a multiple Cox regression model including age, base area, DVT, ulcer duration and treatment group? Comment on the regression coefficients from this model.

(4)

(b) Compare the results obtained from the multiple regression model with those derived from the regression on treatment group alone.

(2)

5. The mortality rates for a certain stationary population, *A*, with 1000 births per year are given in the following table, where ${}_{10}q_x$ is the probability that a person aged x years dies within the next 10 years.

Age	${}_{10}q_x$
0	0.029
10	0.009
20	0.015
30	0.020
40	0.047
50	0.125
60	0.291
70	0.551
80	0.846
90	0.979
100	1.000

Using only this information, estimate:

(i) the age distribution in 10 year class intervals,

(ii) the expected ages at death of groups of people now aged 20, 40, 60, 90 and 100,

(iii) the life-expectancy of people in this population.

(15)

Using the unabridged life table, the life-expectancy of people in this population is 67.92 years, and the expected ages at death of people now aged 20, 40, 60, 90 and 100 are 70.40, 71.84, 75.61, 93.05 and 101.80 respectively. Explain why these figures differ from those obtained in parts (ii) and (iii).

(2)

A different stable population, *B*, experiences the same mortality rates as population *A* and an annual growth rate of 1%. Without doing any additional calculations, explain how you would find the age distribution of population *B* in a form which is suitable for comparison with the distribution obtained for population *A*. How would you expect these distributions to differ?

(3)

6. (i) Define the sensitivity, specificity, positive predictive value and negative predictive value of a diagnostic test. Comment on why sensitivity and specificity are often preferred to positive and negative predictive values in deciding how good a diagnostic test is.

(9)

(ii) In respiratory medicine, clinicians need a simple diagnostic test to detect those patients with the coalworkers' disease pneumoconiosis. The data in the table show the forced expiratory volume (FEV1), expressed as a percentage of normal values, for a random sample of 40 non-smoking subjects.

FEV1 values (% normal)
for subjects with and without coal-workers' pneumoconiosis

Men with pneumoconiosis $n = 27$								
40	43	47	49	50	50	53	57	58
58	58	62	65	69	71	73	74	75
75	77	78	79	80	87	90	100	105

Men without pneumoconiosis $n = 13$								
60	67	73	75	79	80	83	87	89
100	105	109	115					

Four possible cut-off values for a diagnostic test of pneumoconiosis are FEV1 values less than 60% of normal, less than 70%, less than 80% and less than 90%. Estimate the corresponding test sensitivities and specificities.

(6)

Sketch the ROC curve using the four cut-off values above. Giving your reasons, suggest a suitable cut-off value which gives an appropriate balance between sensitivity and specificity.

(5)

7. A wholesale food distributor in a large city wants to assess the demand for a new product based on mean monthly sales. He plans to sell this product in a sample of stores he services. He only services four large chains in the city. Hence, for administrative convenience, he decides to use stratified random sampling with each chain as a stratum. A stratified random sample of 20 stores yields the following sales figures after a month:

Stratum (Chain)			
1	2	3	4
$N_1 = 24$	$N_2 = 36$	$N_3 = 30$	$N_4 = 30$
$n_1 = 4$	$n_2 = 6$	$n_3 = 5$	$n_4 = 5$
$\bar{y}_1 = 99.25$	$\bar{y}_2 = 100.0$	$\bar{y}_3 = 98.0$	$\bar{y}_4 = 100.0$
$s_1 = 9.00$	$s_2 = 7.46$	$s_3 = 6.28$	$s_4 = 10.61$
94	91	108	92
90	99	96	110
103	93	100	94
110	105	93	91
	111	93	113
	101		

(i) Explain what is meant by *stratification with proportional allocation*, and verify that this has been used to construct the above stratum sample sizes.

(4)

(ii) Write down an expression for the unbiased estimator of the population mean in terms of the stratum means and derive its standard error.

(5)

(iii) Estimate the mean monthly sales and obtain an estimate of the standard error of your estimator. Construct an approximate 95% confidence interval for the population mean.

(3)

(iv) Suppose instead that a simple random sample of 20 stores from the population of 120 stores had been selected, with the same responses as given in the table. Had this been done, the estimate for the population standard deviation would have been 7.75. Construct an approximate 95% confidence interval for the population mean.

(2)

(v) Compare the efficiencies of your estimators in parts (iii) and (iv). Suggest why stratified random sampling gives a less precise estimate of the population mean than simple random sampling in this case.

(3)

(vi) Advise the wholesaler on how to select appropriate strata for a stratified random sample when the objective of stratification is to produce estimators with small variance.

(3)

8. (i) Explain the difference between *random* and *non-random* methods of sampling, discussing both the construction of samples and the methods of analysing data collected by them. Suggest reasons why non-random samples may sometimes be preferred. Include an explanation of *systematic sampling*, and whether it should be treated as random or non-random.

(6)

(ii) Write down the formula for calculating an unbiased estimate, s^2, of the variance of a large (but finite) population, based on a simple random sample of n items. Define any symbols you use. Show also that, for a binary variable, $s^2 = np(1 - p)/(n - 1)$, where n and p are to be defined.

A pilot survey has given rough estimates of the mean and variance of a measurement x, and of the proportion p of a special type of member, in the population being studied. The main sample survey will be required to estimate, at the 95% confidence level, the population mean of x within ± 1.5 units, and also the proportion p within ± 0.04. If the pilot survey value of the variance of x was 168.33 and the value of the required proportion was 0.36, find the minimum sample size that should be used to meet requirements.

(7)

(iii) Define *one-stage* and *two-stage cluster sampling*. How do cluster sampling and *stratified sampling* differ, both in their construction and in their use? Give an example of a survey in a country of your choice that uses both stratification and clustering in the sample design.

(7)