EXAMINATIONS OF THE ROYAL STATISTICAL SOCIETY

GRADUATE DIPLOMA, 2010

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 12 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. Data are available about the employment patterns that existed in 26 European countries in 1979. They consist of the percentages of the working population employed in different types of industry. There are 9 types of industry, and for each of the 26 countries the 9 percentages sum to 100%, subject to rounding error.

The variables defining the percentages employed in the 9 types of industry are as follows.

agr	agriculture
min	mining
man	manufacturing
ps	power supply
con	construction
si	service industries
fin	finance
sps	social and personal services
tc	transport and communications

An analyst has been asked to perform a principal component analysis on the data.

(i) Briefly explain the purpose of principal component analysis for these data.

(2)

(ii) Explain how principal components are derived from the correlation matrix.

(2)

(iii) Someone has claimed that principal component analysis would be invalid because the sum of the 9 variables for each country is 100. Discuss the implications of this collinearity for a principal component analysis.

(3)

(iv) The analyst carries out a principal component analysis on the correlation matrix, and some of the computer output is shown below.

Correlations

	agr	min	man	ps	con	si	fin	sps	tc
agr	1.00								
min	0.04	1.00							
man	-0.67	0.45	1.00						
ps	-0.40	0.41	0.39	1.00					
con	-0.54	-0.03	0.49	0.06	1.00				
si	-0.74	-0.40	0.20	0.20	0.36	1.00			
fin	-0.22	-0.44	-0.16	0.11	0.02	0.37	1.00		
sps	-0.75	-0.28	0.15	0.13	0.16	0.57	0.11	1.00	
tc	-0.56	0.16	0.35	0.38	0.39	0.19	-0.25	0.57	1.00

Question continued on next page

Summary statistics of the variables

Variable	Mean	Std. Dev.	Min	Max
agr	19.131	15.547	2.7	66.8
min	1.254	0.970	0.1	3.1
man	27.008	7.008	7.9	41.2
ps	0.908	0.376	0.1	1.9
con	8.165	1.646	2.8	11.5
si	12.958	4.575	5.2	19.1
fin	4.000	2.807	0.5	11.3
sps	20.023	6.830	5.3	32.4
tc	6.546	1.391	3.2	9.4

Eigenvalues

Component	Eigenvalue	Proportion	Cumulative
Comp1	3.487	0.387	0.387
Comp2	2.130	0.237	0.624
Comp3	1.099	0.122	0.746
Comp4	0.994	0.111	0.857
Comp5	0.543	0.060	0.917

Principal components (eigenvectors)

Variable	Comp1	Comp2	Comp3	Comp4	Comp5
agr	-0.524	0.054	-0.049	0.029	0.213
min	-0.001	0.618	0.201	0.064	-0.164
man	0.348	0.355	0.151	-0.346	-0.385
ps	0.256	0.261	0.561	0.393	0.295
con	0.325	0.051	-0.153	-0.668	0.472
si	0.379	-0.350	0.115	-0.050	-0.284
fin	0.074	-0.454	0.587	-0.052	0.280
sps	0.387	-0.222	-0.312	0.412	-0.220
tc	0.367	0.203	-0.375	0.314	0.513

(a) Describe the correlation structure in the data.

(2)

(b) How many principal components would you choose from the analysis? Justify your answer.

(2)

(c) Someone argues that the analysis cannot be correct because the first principal component is usually the average of the variables, and this is not the case in this output. How would you respond?

(2)

(d) Interpret the first two principal components.

(4)

(e) Someone else suggests that, for these data, it would be useful to do principal component analysis on the covariance matrix. The analyst argues that the results would be very similar. How would you respond?

(3)

2. (i) A biologist is drawing up a taxonomy of a number of species, based on variables defining the shapes and sizes of each species. He knows that he needs to use multivariate methods, but is not sure whether he should choose cluster analysis or linear discriminant analysis.

(a) Explain the difference between the purposes of these two methods, the assumptions that need to be met and how he should decide which method to use.

(6)

(b) The biologist decides that he will use cluster analysis, and has a package that will do agglomerative hierarchical clustering. However, he has also found a webpage which makes the following statements about cluster analysis.

> "Cluster analysis is descriptive and non-inferential. It has no theoretical basis from which to draw inferences about the population. The solutions are not unique because the cluster membership depends on many elements of the procedure and many different solutions can be obtained by varying the elements."

What would you say to the biologist about the truth or otherwise of these statements, in relation to agglomerative hierarchical clustering?

(3)

(ii) The following 'dissimilarity' matrix is presented for use in cluster analysis.

	obs1	obs2	obs3	obs4	obs5	obs6	obs7	obs8	obs9
obs1	0								
obs2	8.31	0							
obs3	7.50	5.96	0						
obs4	8.89	14.27	9.27	0					
obs5	21.06	14.07	14.14	22.37	0				
obs6	12.60	8.88	5.10	12.34	10.05	0			
obs7	5.87	9.60	5.08	5.41	18.62	9.15	0		
obs8	5.97	2.98	6.76	13.36	17.02	10.88	8.94	0	
obs9	2.71	10.70	8.56	6.88	22.60	13.48	5.32	8.60	0

(a) Use single linkage cluster analysis on the dissimilarity matrix above, and draw the dendrogram for your analysis.

(8)

(b) Do you think there is a clear cluster structure in the data? Justify your answer.

(3)

3. (i) Consider a random variable T measuring the time to failure of machinery and defined by the probability density function

$$f_T(t), \quad t \geq 0 .$$

(a) Define the *survivor function* as used in survival analysis, and show how it is related to the probability density function.

(2)

(b) Derive the survivor function for the Weibull distribution with probability density function

$$f_T(t; \theta, \beta) = \frac{\beta}{\theta^{\beta}} t^{\beta-1} e^{-(t/\theta)^{\beta}} \qquad t \geq 0; \ \beta > 0, \ \theta > 0 .$$

(2)

(c) Show that if time to failure follows a Weibull distribution, a scatter plot of a suitable function of the survivor function plotted against log(time) can be used to estimate the parameters θ and β.

(3)

(ii) A quality control engineer is studying the reliability of a particular type of machine, by measuring the times to failure for eleven randomly selected machines. The times (in thousands of hours) are as follows, where * indicates a censored value.

Machine	1	2	3	4	5	6	7	8	9	10	11
Time (thousands of hours)	7.432	1.537	3.169	9.500	5.993	6.369*	9.400*	4.219	6.683	4.700	6.148

The engineer asks you to estimate the 'average' time to failure.

(a) Explain why the mean time may not be a sensible average for data like these. Compute a preferable alternative measure of location. Justify your choice of measure.

(3)

(b) Compute the Kaplan-Meier survivor function for these data and plot the survival curve.

(4)

(c) Draw a suitable graph to investigate whether these data can be modelled using a Weibull distribution, and interpret the graph.

(4)

(d) Draw a straight line through the points on your graph by eye and use it to estimate the parameters for a Weibull distribution fitted to these data.

(2)

4. An engineer is investigating the time to failure of certain components. He has studied a random sample of 50 of these components of which 18 failed before the end of the study.

The engineer believes that there are three variables that affect time to failure; a factor *F* (with values 0 and 1), and two covariates *X1* and *X2*. The covariates *X1* and *X2* are known to be positively correlated.

(i) Define the *Cox proportional hazards model* and explain how it could be used to model these data.

(5)

(ii) The results of a Cox proportional hazards regression are shown below.

_t	Haz. Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
factor	1.59310	0.92013	0.81	0.420	0.51359	4.94165
x1	0.99581	0.00288	-1.45	0.146	0.99019	1.00146
x2	1.01292	0.04134	0.31	0.753	0.93505	1.09728

(a) The engineer queries why the *p*-values are all greater than 5%, but the confidence intervals for the coefficients do not contain any negative values. Provide a suitable explanation for him.

(2)

(b) The engineer wants to be able to interpret the estimates from the model, despite the fact that they are not statistically significant. Use the following examples of components to explain to him how the coefficients for *F* and *X1* are to be interpreted.

Component	*Factor*	*X1*	*X2*
1	1	1000	100
2	1	2000	100
3	0	2000	100

(4)

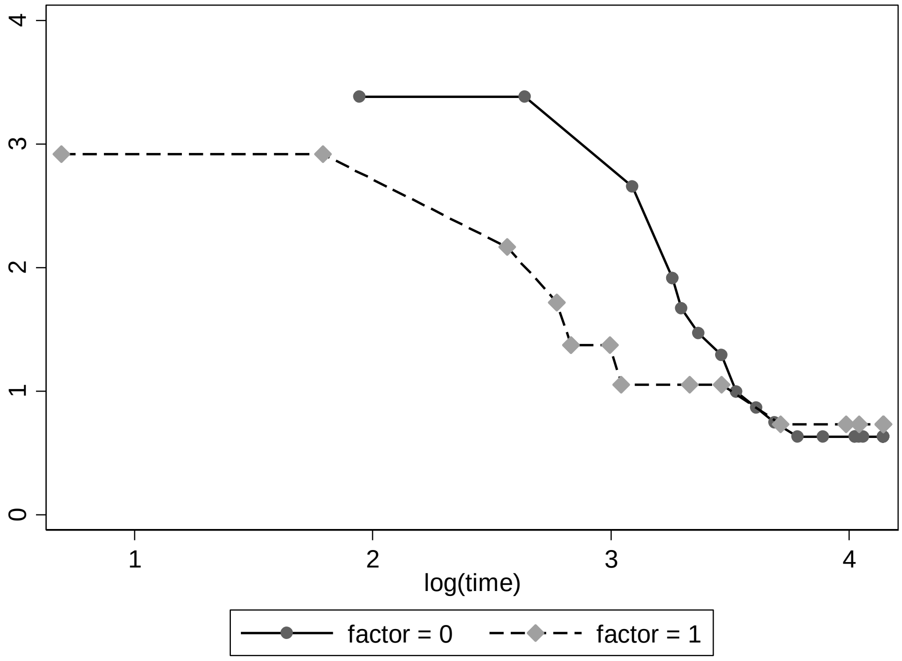
(c) The engineer is worried that previous work has suggested that the proportional hazards assumption may not be valid for the factor *F*, and therefore that the results may be invalid. What would you say to this suggestion?

(2)

Question continued on next page

(d) In order to take the possible problem in part (ii)(c) into account, the engineer produces a plot using *F* as a stratifying variable. This plot is shown below, together with the results from a corresponding Cox model. Based on this information do you think that the results from the original analysis were invalid? Justify your answer.

(3)




```
 Log likelihood = -52.613265                          Prob > chi2  =  0.3971

------------------------------------------------------------------------------
         _t  | Haz. Ratio    Std. Err.       z     P>|z|     [95% Conf. Interval]
-----------+------------------------------------------------------------------
         x1  |   0.99608       0.00292     -1.34    0.180      0.99036    1.00182
         x2  |   1.01126       0.04172      0.27    0.786      0.93271    1.09644
------------------------------------------------------------------------------
                                                                Stratified by F
```

(e) Someone else raises two further objections regarding the original analysis.

Objection 1 the sample size is too small for the number of predictor variables.

Objection 2 the correlation between *X1* and *X2* has not been considered.

What would your advice be?

(4)

Turn over

5. A doctor wants to be able to identify pregnancies where the baby is at risk of shoulder dystocia, a problem caused by a difficult labour during natural birth. He has collected data from a random sample of 100 babies, of which 16 had shoulder dystocia. From logistic regression analysis, one of the variables that appears to be an important predictor is the birthweight of the baby. However, this variable is not easily estimated during pregnancy.

The doctor therefore considers two logistic regression models; model A contains a set of predictor variables plus birthweight, while model B has the same set of predictor variables but omits birthweight. For each baby the response variable is coded as 1 for the presence and 0 for the absence of shoulder dystocia.

(i) The logistic regression produces a 2×2 table showing predicted versus actual values for the response variable. Even though the variable birthweight is highly significant in model A, in the computer output the tables for models A and B are identical and as shown below when a cut-off probability of 0.5 is used to provide the binary predictions.

		Actual		
		0	1	
Predicted	0	84	16	100
	1	0	0	0
		84	16	

Overall accuracy 84%

The overall accuracy is defined to be the percentage of babies that are correctly predicted as 0 or 1 by the model.

Provide an explanation for why the tables are identical for models A and B.

(3)

(ii) Define the *sensitivity* and *specificity* of a diagnostic test and explain why they are more informative than the overall accuracy of the test.

(3)

Question continued on next page

(iii) The doctor obtains data which he has been told can be used to generate sensitivity and specificity values from the two models using different cut-offs for the probabilities.

The data are given below, and are the number of babies predicted to have shoulder dystocia by each model, using different cut-offs.

Model A

		Number predicted to have shoulder dystocia				
		Cut-off 1	Cut-off 2	Cut-off 3	Cut-off 4	Cut-off 5
Actual	1	2	5	8	12	13
	0	1	12	14	16	19

Model B

		Number predicted to have shoulder dystocia	
		Cut-off 1	Cut-off 2
Actual	1	5	12
	0	34	64

Use the data to compute the sensitivity and specificity values for each model and plot the two corresponding ROC curves on the same graph.

(5)

(iv) Interpret the ROC curves, with special reference to the importance of the variable birthweight.

(3)

(v) Using the results from model B, obtain the positive and negative predictive values for the test for each of the two cut-offs. Explain why in practice the sensitivity and specificity are preferred.

(4)

(vi) Comment on the advisability of using results from model B in clinical practice to inform mothers of the chance of their delivering a baby with shoulder dystocia.

(2)

6. (i) Explain the role of *life tables* in population studies. Distinguish between *current* and *cohort* life tables, and state when each would be used. Explain briefly the relationship between life tables and *age-specific death rates*.

(4)

(ii) 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	0.250
10	0.024
20	0.040
30	0.051
40	0.062
50	0.091
60	0.172
70	0.335
80	0.624
90	1.000

Using only this information construct a life table and estimate

(a) the age distribution in 10-year class intervals,

(6)

(b) the expected age at death of groups now aged 20,

(3)

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

(2)

Define any notation used in parts (a) to (c).

(d) State any assumptions required for the validity of these calculations and comment on whether they are appropriate.

(3)

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

(2)

Turn over

7. In order to estimate the total cattle population in a district consisting of $N = 1238$ villages, a simple random sample of $n = 16$ villages was selected. The number of cattle, y, recorded in the survey, together with the most recent census figures for cattle, x, are given below.

	Number of cattle			*Number of cattle*	
Village	*Survey, y*	*Census, x*	*Village*	*Survey, y*	*Census, x*
1	654	623	9	292	371
2	696	690	10	555	298
3	530	534	11	2110	2045
4	315	293	12	592	1069
5	78	69	13	707	706
6	640	842	14	1890	1795
7	692	475	15	1123	1406
8	210	161	16	115	118
Total	3815	3687	Total	7384	7808

The census showed that there were 680 900 cattle in the 1238 villages.

$\Sigma x^2 = 13\,462\,957$, $\Sigma y^2 = 12\,773\,061$, $\Sigma xy = 12\,875\,489$.

(i) Estimate the total cattle population from the survey data, using

(a) the ratio estimator,

(b) the regression estimator.

(12)

(ii) Explain briefly why these estimates are lower than that obtained using the survey information alone.

Given that the variances of the ratio and regression estimators are respectively 2 922 981 150 and 2 852 776 962, estimate and compare the efficiencies of these estimators relative to an estimator based on the survey information alone.

(4)

(iii) Giving your reasons, select one of these estimators and construct an approximate 95% confidence interval for the number of cattle in the 1238 villages.

(4)

Turn over

8. There are 2026 households in a city, divided into 4 strata on the basis of household income. It is required to estimate the proportion of households living in rented houses. In each stratum (h = 1, 2, 3, 4) a simple random sample of households is selected, and interviews are carried out to find the number of households living in rented houses. The following table shows the results of this survey. (Income is coded in suitable units.)

Stratum based on income	*Stratum population size* N_h	*Stratum sample size* n_h	*Number renting*
< 50	1190	40	30
50 – 100	523	35	18
100 – 200	215	35	7
> 200	98	40	5

Let N, n denote the total population and sample size respectively, and p_h the sample proportion of those renting in stratum h.

(i) Show that $p_{st} = \frac{1}{N}\sum_{h=1}^{4} N_h p_h$ is an unbiased estimator for the proportion of households living in rented houses.

(5)

(ii) Find the variance of p_{st}, and show that an unbiased estimator of this variance is $\frac{1}{N^2}\sum_{h=1}^{4} N_h (N_h - n_h)\frac{p_h(1-p_h)}{n_h - 1}$.

(7)

(iii) Estimate the proportion of households living in rented houses and the standard error of this estimate, for the data above.

(5)

(iv) Find the stratum sample sizes that would be needed for optimal allocation.

Do the allocations in the above survey seem reasonable? Give reasons for your answer.

(3)

[Note. Results for sampling continuous variables Y may be assumed without proof.]