ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

EXAMINATIONS OF THE ROYAL STATISTICAL SOCIETY

GRADUATE DIPLOMA, 2017

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* ${}^nC_r$.

This examination paper consists of 12 printed pages.
This front cover is page 1.
Question 1 starts on page 2.

There are 8 questions altogether in the paper.

1. (i) What is a population pyramid?

(2)

(ii) Describe in detail how a population pyramid is constructed.

(4)

(iii) Describe two uses of a population pyramid for a single population.

(2)

(iv) Describe two uses of population pyramids for comparing populations.

(2)

(v) A *life table* presents in tabular layout various probabilities, conditional probabilities, expectations and conditional expectations. Each column has a particular interpretation. For a life table with radix of $l_0 = 1$, where x is age in years, explain the meaning of the following life table quantities.

(a) l_x

(b) d_x

(c) q_x

(d) L_x

(e) T_x

(f) e_x

(6)

(vi) There are two types of life tables: cohort and period life tables. Describe in words each type of life table, and give an example of where it can be used.

(4)

2. (i) Consider a study of the association between a binary outcome (disease versus no disease) and a binary risk factor (exposure to a risk factor versus no exposure). What is *confounding* of the association between disease and exposure? What method is commonly used to detect the presence of confounding?

(4)

(ii) What conditions are necessary for a variable to be a confounder in the association between disease status and a risk factor for the disease?

(6)

(iii) What can a researcher do about confounding in terms of

(a) study design;

(b) study analysis?

In each case, illustrate using brief examples.

(10)

3. (i) In a small agricultural region, a sample of farms is to be selected to estimate the total number of acres used for growing corn. The sampling frame includes the farm size (acres) for each farm. Describe two ways in which this auxiliary information could be used for stratification in sample design. Describe three ways in which the use of stratification would be useful in analysis of the survey. (5)

(ii) The sampling frame includes 511 farms with total acreage of 53 102. A simple random sample of 9 farms was selected, and the number of acres used for growing corn (y) and the farm size (x) in acres were recorded for each farm. The results were as follows.

Table 1: Simple random sample of 9 farms

Farm	*Corn (acres), y*	*Total acres, x*
1	50	60
2	56	72
3	66	68
4	76	94
5	90	90
6	100	102
7	112	116
8	110	130
9	175	200

You are given that $\Sigma y_i = 835$, $\Sigma x_i = 932$, $\Sigma y_i^2 = 89\,137$, $\Sigma x_i^2 = 111104$ and $\Sigma(y_i - \bar{y})^2 = 11\,667.56$.

(a) In fact, stratification was not used and instead both a simple random sample estimator and ratio estimation were used. Estimate the **total** number of acres used for growing corn in the region using

(1) the simple random sample estimator, (2)

(2) the ratio estimator. (2)

(b) Given that the estimated variance of the ratio estimator of this total is 1 672 985, calculate the (estimated) relative efficiency of the ratio estimator of the total compared to the simple random sample estimator of the total. Which estimator is better? (4)

(c) Why are these two estimates very similar in size? (2)

(d) Under what conditions is the ratio estimator useful? (4)

(e) An alternative estimator to the ratio estimator is the regression estimator. When is the regression estimator preferred to the ratio estimator? (1)

4. (i) Explain what is meant by *non-probability sampling*.

(2)

(ii) Explain what is meant by *convenience sampling*. For what type of population does this method of sampling yield accurate results?

(2)

(iii) Explain what is meant by *judgement (or purposive) sampling*. When is this sampling method used by statisticians?

(2)

(iv) Explain what is meant by *quota sampling*. What method of probability-based sampling does it resemble?

(2)

(v) State three reasons why non-probability sampling methods are used.

(3)

(vi) State three limitations/disadvantages of non-probability sampling methods.

(3)

(vii) State three reasons why simple random sampling is rarely used in practice.

(3)

(viii) State three properties of an *ideal sampling frame*.

(3)

5. (i) Explain the purpose of *discriminant analysis*. State two reasons for using it.

(3)

(ii) State three distributional assumptions that are desirable for discriminant analysis to be useful.

(3)

(iii) Data were collected on 189 pregnant women, who gave birth at a medical clinic in the USA. Of the 189, 59 gave birth to a low birth weight baby (< 2.5 kilograms) and 130 gave birth to a normal birth weight baby (≥ 2.5 kilograms). As low birth weight is related to childbirth complications, the identification of risk factors associated with giving birth to a low birth weight baby is important. Commonly collected clinical data thought to be useful in predicting whether a woman would give birth to a low birth weight baby are her age (in years) and weight (in pounds at her last menstrual period).

(a) Summary statistics and scatter plots are given below. Describe these data, and briefly discuss the suitability of discriminant analysis to analyse these data.

(3)

Low birth weight	*Age*		*Weight*	
	Mean	*(SD)*	*Mean*	*(SD)*
Yes	22.31	(4.51)	122.14	(26.56)
No	23.66	(5.58)	133.30	(31.72)

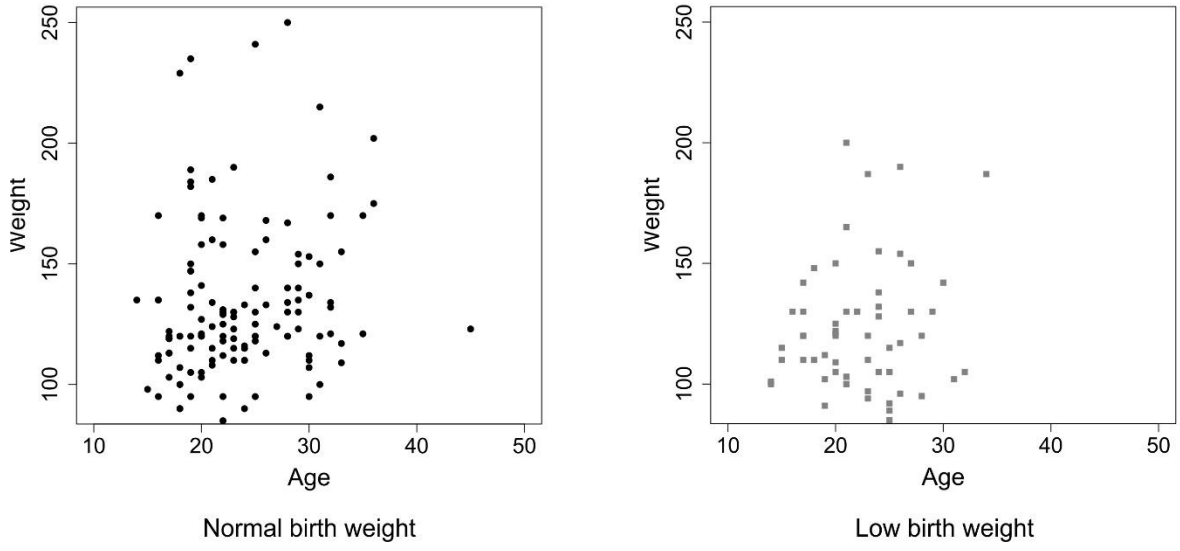

Figure 1: Plots of weight by age for each of the two groups (normal and low birth weight)

Question 5 continued on the next page

(b) Consider the following edited computer output.

```
lda(low ~ age + lwt, data = birthwt, prior = c(0.5, 0.5))

Prior probabilities of groups:      No   Yes
                                   0.5   0.5

Group means:                  age         lwt
                 No     23.66154    133.3000
                 Yes    22.30508    122.1356

Coefficients of linear discriminants:                LD1
                                              age  -0.09127569
                                              lwt  -0.02651833
```

Assume that the prior probabilities of being in the two groups (normal birth weight and low birth weight) are both 0.5. Consider two pregnant women, A and B, where A is aged 30 and her last weight was 150 pounds, while B is aged 15 and her last weight was 150 pounds. Use the above computer output to predict whether woman A will have a low birth weight baby; and to predict whether woman B will have a low birth weight baby.

(6)

(c) The discriminant analysis model was applied to each of the 189 women in the sample; the table below compares the prediction and the actual outcome. What is the misclassification rate? Assess whether this discriminant analysis was suitable.

(2)

		Group	
		No	Yes
Prediction	No	65	19
	Yes	65	40

(iv) Briefly compare and contrast the statistical methods of discriminant analysis (DA) and principal component analysis (PCA).

(3)

6. A doctor investigates whether a drug affects blood pressure on 15 patients. The responses of interest are the changes, positive or negative, in blood pressure. Pearson's product-moment correlation between the changes in systolic blood pressure and diastolic blood pressure is 0.660, with 95 percent confidence interval (0.223, 0.876).

(i) Briefly comment on the change data in Table 2.

(3)

(ii) Explain why Hotelling's T^2 test could be used to test the effect of this drug on blood pressure and state the null and alternative hypotheses.

(5)

Table 2: Blood pressure (mm Hg) before and after administration of a drug.

Systolic blood pressure			*Diastolic blood pressure*		
Before	*After*	*Change*	*Before*	*After*	*Change*
210	201	–9	130	125	–5
169	165	–4	122	121	–1
187	166	–21	124	121	–3
160	157	–3	104	106	2
167	147	–20	112	101	–11
176	145	–31	101	85	–16
185	168	–17	121	98	–23
206	180	–26	124	105	–19
173	147	–26	115	103	–12
146	136	–10	102	98	–4
174	151	–23	98	90	–8
201	168	–33	119	98	–21
198	179	–19	106	110	4
148	129	–19	107	103	–4
154	131	–23	100	82	–18
mean change –18.93			mean change –9.27		

(iii) State the distributional assumptions that are desirable for the method to be useful.

(4)

(iv) Hotelling's T^2 statistic for this null hypothesis is

$$T^2 = n\bar{y}'S_Y^{-1}\bar{y} \, .$$

Define each of these symbols in the context of the above study.

(3)

(v) Recall that Hotelling's T^2 is proportional to F; here T^2 = 68.499. Test the null hypothesis you introduced in part (ii) above, and state your conclusions about the effect of this drug.

(5)

7. (i) Define the *survivor function* $S(t)$ and the *hazard function* $h(t)$ for a continuous non-negative random variable *T* measuring lifetime.

(2)

(ii) For a continuous non-negative random variable *T*, use integration by parts to show that

$$E(T) = \int_0^\infty S(t)\, dt$$

and hence, or otherwise, prove that

$$E\left(\frac{1}{h(T)}\right) = E(T).$$

(5)

After chemotherapy, 42 patients with cancer of the white blood cells were in remission. Remission means that after treatment there is no sign of the cancer. These patients were then randomised to receive either a further drug treatment (treatment group) or no further treatment (control group). Remission times in months were recorded and are given in the table below. A right-censored observation is denoted by +, so 6+ denotes a right-censored observation at 6 months.

Control	1 8	1 8	2 11	2 11	3 12	4 12	4 15	5 17	5 22	8 23	8
Treatment	6 10+	6 11+	6 17+	7 19+	10 20+	13 25+	16 32+	22 32+	23 34+	6+ 35+	9+

(iii) For the control group, compute the Kaplan-Meier estimate of the survivor function at months 1, 2, 3 and 4. Calculate the standard error for $\hat{S}(4)$.

(5)

(iv) For the treatment group, compute the Kaplan-Meier estimate of the survivor function at months 6, 7 and 10. Apply Greenwood's formula to calculate a 95% confidence interval for $\hat{S}(10)$.

(6)

(v) Is using Greenwood's formula to calculate a 95% confidence interval for $\hat{S}(10)$ accurate? Justify your answer.

(2)

8. A proportional hazards regression model of the factors associated with the risk of cycling-related injuries resulting in hospitalisation among members of a University's Cycling Club was estimated. The two risk factors, which were statistically significant, were gender and whether a Club member was an undergraduate student, a postgraduate student, or one of the University's staff.

The model estimated was

$$h(t) = h_0(t)e^{\beta_M M + \beta_U U + \beta_P P}$$

where $h(t)$ is the hazard of injury at duration *t* since joining the Club, and $h_0(t)$ is the baseline hazard at duration *t* since joining the Club.

Indicator variables are:

M takes the value 1 if the member is male, and 0 otherwise,

U takes the value 1 if the member is an undergraduate student, and 0 otherwise,

P takes the value 1 if the member is a postgraduate student, and 0 otherwise;

and β_M, β_U and β_P are parameters.

(i) To what category of member does the baseline hazard relate?

(1)

(ii) According to the results:

male undergraduates had six times the hazard of injury of female staff,

male postgraduates had twice the risk of injury of female undergraduates,

male staff had the same risk of injury as female postgraduates.

Calculate the values of the parameter estimates $\hat{\beta}_M$, $\hat{\beta}_U$ and $\hat{\beta}_P$.

(7)

(iii) Give three reasons why the proportional hazards regression model is an attractive approach to this kind of problem.

(3)

(iv) Describe a situation in which the proportional hazards regression model would be an inappropriate model for analysis of data of this kind.

(2)

(v) Define the cumulative hazard function, $H(t)$. State how it is related to the survivor function, $S(t)$.

(2)

Question 8 continued on the next page

(vi) A colleague suggests to you that, instead of using a semi-parametric form of the proportional hazards model in which the baseline hazard is non-parametric, you might model the baseline hazard using the exponential or Weibull distributions. If *t* is duration, the exponential distribution has a baseline hazard $h_0(t) = \lambda$ and the Weibull distribution has a baseline hazard function

$$h_0(t) = \lambda\gamma t^{\gamma-1}, \qquad t \geq 0$$

where $\lambda > 0$ and $\gamma > 0$ are parameters.

(a) Show that, if the exponential distribution is good fit to the data, a plot of $-\log S(t)$ against duration *t* should be roughly linear and pass through the origin.

(2)

(b) Suggest a similar plot, i.e., some function of $S(t)$, which you could use to test whether the Weibull distribution would fit the data.

(3)

BLANK PAGE