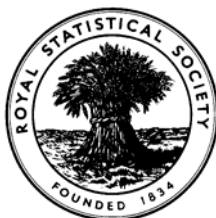**EXAMINATIONS OF THE ROYAL STATISTICAL SOCIETY**

(formerly the Examinations of the Institute of Statisticians)

HIGHER CERTIFICATE IN STATISTICS, 2007

Paper II : Statistical Methods

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 9 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)   State the *Central Limit Theorem*.

(2)

(ii)   A random sample is drawn from a distribution with known variance. Explain how, and under what circumstances, the Central Limit Theorem enables a statistic based on the sample mean to be used to test hypotheses about the mean of the distribution.

(4)

(iii)   A die is rolled 100 times. Use a suitable approximation to find the probability that the sum of the scores on the uppermost face is more than 365, assuming the die to be fair. [You are given that the mean and variance of the score on the uppermost face for each rolling of a fair die are respectively 3.5 and $\frac{35}{12}$.]

(8)

(iv)   A test of whether the die is fair is to be constructed using the sum of the scores on 100 rolls of the die. The alternative hypothesis is that the die is not fair and a 5% significance level is to be used. Use a suitable approximation to find the critical region.

(6)

2. In a random sample of 53 patients attending a clinic with the disease ankylosing spondylitis, their ages, in years, when they were diagnosed with the disease were as follows.

40	27	58	46	33	22	14	18	44	20	36	28
19	22	32	48	40	42	54	26	17	22	26	41
50	38	39	42	32	31	37	24	25	38	22	42
46	52	47	48	48	27	12	33	40	33	13	59
39	32	43	40	34							

(i) Form a grouped frequency distribution table of the ages at diagnosis, using class widths of 5 years.

(5)

(ii) Using your frequency distribution table, construct a histogram of the ages at diagnosis.

(5)

(iii) Describe the distribution of ages at diagnosis. Do you consider that the distribution is symmetric? If you consider it to be skewed, describe the degree and direction of the skewness.

(3)

(iv) Use

(a) your grouped frequency distribution table,

(b) the raw data,

to estimate the median age at diagnosis.

(4)

(v) Compare the values you computed in (iv)(a) and (iv)(b). If you were asked to estimate the median age at diagnosis of ankylosing spondylitis, which value should you quote? Why?

(3)

3

Turn over

3. (i) Explain the rationale behind one-way analysis of variance. State the assumptions necessary for the method to be valid.

(6)

(ii) In a study of the strength of pencil leads, standard length pieces of pencil lead were supported on two ends and loaded at their mid-points. Samples of five pieces each of 4H, H and B lead were used and the breaking stresses (load, in grams) at which each piece failed were recorded. The resulting data are given below.

4H lead:	56.7	63.8	56.7	63.8	49.6
H lead:	99.2	99.2	92.1	106.0	99.2
B lead:	56.7	63.8	70.9	63.8	70.9

(a) Perform an analysis of variance on the above data and determine whether there are any differences between the population mean breaking stresses of 4H, H and B pencil lead. You are given that the assumptions hold for the analysis to be valid.

(10)

(b) Calculate the standard error of the difference between two mean breaking stresses. Without further calculation, briefly explain how you would use this to obtain confidence intervals for the pairwise differences between the population mean breaking stresses of the three types of pencil lead.

(4)

Turn over

4.   In a factory manufacturing electronic components, there are four machine operators and five machines producing similar items. The quality of the components having been variable, an experiment was conducted to determine whether the variability was caused by differences between machines, differences between operators or both. The production was observed for each shift, with each operator using each of the five machines for a whole shift. The quality of the components produced during any shift was assessed by quality control inspectors on a scale of 0 to 100, with 100 corresponding to perfect quality and 0 to useless material. The values of the quality scores were as given in the following table.

		Machine				
		A	B	C	D	E
Operator	1	56	92	53	93	68
	2	64	83	55	95	62
	3	62	80	56	96	62
	4	51	78	44	88	69

Sums for the five machines are 233, 333, 208, 372, 261; sums for the four operators are 362, 359, 356, 330; the sum of squares of the 20 values is 104127.

(i)   Write down a linear model which could be used as the basis for an analysis of a set of data such as that shown above. Explain clearly what each term in the model represents, and state any assumptions required for the analysis to be valid.

(5)

(ii) Carry out an appropriate analysis of variance, and test for differences between operators and between machines.

(9)

(iii) Explain briefly how you might attempt to check whether the assumptions required for an analysis of variance are satisfied.

(3)

(iv) Write a short report on your findings, using non-technical language as far as possible, to assist a manager to choose combinations of machine and operator with good chances of a high score for quality.

(3)

5. The hormone leptin is thought to be important in energy expenditure in humans. In a study to investigate the effect of diet on leptin levels, the leptin levels (in ng/ml) of 19 healthy volunteers were measured before and after a 72-hour controlled diet. The data are given in the table below.

Leptin level before diet (ng/ml)	Leptin level after diet (ng/ml)
11.58	12.54
13.47	12.96
13.62	9.84
8.94	9.84
10.80	12.45
17.91	12.33
11.22	11.88
16.62	18.24
9.12	14.67
11.31	14.01
12.12	15.45
15.12	15.60
9.45	11.67
11.34	13.08
12.12	13.80
15.78	16.56
9.78	9.54
12.60	13.83
13.23	14.61

(i) Use an appropriate parametric test to determine whether there is a difference (at the 5% significance level) between the mean leptin levels before and after the diet.

(11)

(ii) (a) Suggest two non-parametric tests, either of which might be used here if the distribution of within-pair differences could not reasonably be assumed to be Normal.

(2)

(b) Why should you prefer to perform an appropriate parametric test rather than either of the two non-parametric tests if the distributional assumption required for the parametric test holds?

(1)

(c) Explain how you would decide whether to perform the parametric test or one of the two non-parametric tests.

(3)

(d) Calculate the sample mean and sample median of the differences. How do these results affect your choice of test?

(3)

6. The data in the table below are from a study of HIV-infected men from Rio de Janeiro, Brazil. They are from the 19 individuals with semen and saliva viral load means above detection limits (the levels cannot be detected if they are below 400 copies/ml, equivalent to 2.60 $\log_{10}$ copies/ml).

Log$_{10}$ serum viral load (log$_{10}$ copies/ml) (y_i)	Log$_{10}$ saliva viral load (log$_{10}$ copies/ml) (x_i)
2.98	3.28
3.04	4.89
3.15	3.89
3.26	3.18
3.60	4.15
3.74	3.96
3.77	4.67
3.78	4.36
3.86	4.18
4.04	5.15
4.04	4.60
4.20	4.61
4.41	4.87
4.49	4.32
4.54	3.28
5.04	5.20
5.26	5.52
5.54	6.08
5.97	4.11

(i) Plot the $\log_{10}$ serum viral load (on the y or vertical axis) against the corresponding $\log_{10}$ saliva viral load (on the x or horizontal axis). Do these log-transformed variables appear to be linearly related?

(8)

(ii) (a) State the necessary conditions for the valid and appropriate use of Pearson's product-moment correlation coefficient in the assessment of a relationship between two variables.

(3)

(b) Calculate and briefly interpret the sample product-moment correlation coefficient for the data presented in the table. You are given the following summary statistics.

$$\sum_{i=1}^{19} x_i = 84.30 \quad \sum_{i=1}^{19} y_i = 78.71 \quad \sum_{i=1}^{19} x_i^2 = 384.55 \quad \sum_{i=1}^{19} y_i^2 = 338.93 \quad \sum_{i=1}^{19} x_i y_i = 355.21$$

(7)

(iii) How is the coefficient of determination related to the correlation coefficient?

(2)

7

Turn over

7. A motorist is trying to determine the quicker of two possible routes, A and B, to work. Over a period of 24 days, she randomly chooses 12 on which to use one route and 12 on which to use the other. The table below gives the journey times from home to work, in minutes.

Route A	Route B
20.2	18.7
27.1	31.2
22.4	22.3
23.2	28.1
28.8	22.1
21.3	19.3
22.5	21.0
23.9	25.7
24.0	21.2
26.3	21.4
34.3	25.0
34.1	30.3

(i) Perform an appropriate non-parametric test to determine whether the median journey time for one of these routes is lower than the median journey time for the other. State the null and alternative hypotheses used in your test and state the assumptions necessary for the test to be valid.

(11)

(ii) Using the data given in the table, compute the sample median and the sample mean of the journey times for each route. Compare the median and mean journey times for the two routes. Briefly discuss the advantages and disadvantages of performing a non-parametric test, rather than a t test for two independent samples, to determine whether one of these routes is quicker than the other.

(9)

Turn over

8. In the British Household Panel survey of 1996, data were collected on family structure during childhood. One question asked "Did you live with both your biological mother and biological father from the time you were born until you were 16?" Those who responded 'yes' are classed as living in an intact family until the age of 16; those who responded 'no' are classed as living in a non-intact family before the age of 16. The responses from those born between 1974 and 1980 are tabulated below.

		Lived in an intact family until age 16	
		No	*Yes*
Sex	*Female*	133	460
	Male	155	418

(i) Perform an appropriate test to determine whether there is an association between an individual's sex and whether he or she lived in an intact family until the age of 16.

(9)

(ii) Compute and briefly interpret an approximate 95% confidence interval for the difference between the proportions of males and females born between 1974 and 1980 who lived in a non-intact family before the age of 16.

(7)

(iii) The confidence interval found in (ii) is of width approximately 0.1. How large a sample size would you need to reduce the width to approximately 0.05?

(4)