EXAMINATIONS OF THE ROYAL STATISTICAL SOCIETY
(formerly the Examinations of the Institute of Statisticians)

HIGHER CERTIFICATE IN STATISTICS, 2006

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 8 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. The duration of a "normal" pregnancy (i.e. one without medical complications) can be modelled by a Normal distribution with mean 266 days and standard deviation 16 days. In an urban hospital in a relatively deprived area of the USA, a random sample of 60 pregnancies was studied and the duration of each pregnancy determined. The relevant statistics were

$$\sum x = 15568, \quad \sum x^2 = 4054484 .$$

(i) Test whether the variance of this sampled population differs from that of "normal" pregnancy durations.

(10)

(ii) Test whether there is evidence that the mean pregnancy duration for this sampled population differs from 266 days. Find an approximate *p*-value for your test, and state your conclusions clearly.

(10)

2. A randomised experiment was carried out to test how well Entonox, a gas made up of 50% oxygen and 50% nitrous oxide, performed in reducing pain during a minor surgical procedure. Twenty patients were randomised to receive air (10 patients) or Entonox (10 patients). Immediately after the procedure, the patients used a 0 – 100 scale to record the level of pain they suffered during the procedure. Their scores are given in the table below.

Air	*Entonox*
37	9
33	0
50	14
8	47
16	15
38	17
55	10
28	22
31	3
39	80

(i) In situations such as this when comparing two independent samples of subjective scores, briefly discuss the advantages and disadvantages of using non-parametric rather than parametric tests.

(6)

(ii) Use an appropriate non-parametric test to determine whether there is evidence that pain scores differ between patients given Entonox and patients given air during this minor surgical procedure. Explain your conclusions in a manner appropriate for a doctor to understand, including a comment on what implications the size of the samples may have for your conclusion.

(14)

3. A sample of 10 sea bass was caught by a fisheries scientist who then measured their length x (in millimetres) and their weight y (in grams). The data are given in the table below.

Length (x)	387	366	329	293	273	268	294	198	185	169
Weight (y)	720	680	480	330	270	220	380	108	89	68

(i) Plot the weights of the 10 sea bass (on the y or vertical axis) against the corresponding lengths (on the x or horizontal axis). Does it appear appropriate to fit a straight line to these data?

(8)

(ii) (a) Calculate the least-squares estimates of the parameters β_0 and β_1 of the regression line $y = \beta_0 + \beta_1 x$.

[Note:

$$n = 10, \qquad \sum_{i=1}^{n} x_i = 2762, \qquad \sum_{i=1}^{n} y_i = 3345,$$

$$\sum_{i=1}^{n} x_i^2 = 812594, \qquad \sum_{i=1}^{n} y_i^2 = 1610009, \qquad \sum_{i=1}^{n} x_i y_i = 1075861.]$$

(b) Comment on the appropriateness of the regression line estimated in part (a) as a model for the relationship between the weights and lengths of sea bass.

(8)

(iii) Calculate and interpret the coefficient of determination.

(4)

4. An experiment was performed to investigate the clotting time (in minutes) of plasma from 8 subjects treated by three different methods. The resulting data are presented in the table below.

Subject	*Method 1*	*Method 2*	*Method 3*
1	6.8	8.3	8.1
2	9.7	10.0	11.1
3	8.3	8.5	10.0
4	9.0	7.9	9.6
5	11.0	10.8	11.1
6	12.4	12.6	14.5
7	8.8	8.4	10.0
8	12.6	12.8	12.5

(i) Treating the subjects as blocks, analyse the data using analysis of variance. State the assumptions necessary for this analysis to be valid and provide a brief interpretation of your findings.

[Note. The sum of all the observations is 244.8 and the sum of their squares is 2585.22.]

(15)

(ii) Suppose that both the "method" sum of squares and the total sum of squares you calculated in part (i) had resulted from a similar experiment, which used 24 different subjects, eight being randomised to each method, rather than blocking on 8 subjects. Re-analyse the data on this basis and compare your conclusion with that from part (i). What impact has blocking had on the power to detect a difference between methods?

(5)

5. The data given below are the numbers (in thousands) of farms in each of the 50 states of the USA in 1987, given in order of increasing numbers of farms.

1	1	2	3	3	4	4	6	7	8	8	9	9
14	14	17	21	24	24	27	28	33	36	36	37	38
39	42	46	49	50	50	52	57	57	61	70	71	73
78	79	82	87	88	93	96	99	109	115	160		

(i) Draw a stem and leaf diagram to display these data. Briefly discuss whether the data appear to be symmetrical or skew.

(6)

(ii) Compute the median, mean, standard deviation and inter-quartile range.

[Note: $\sum x = 2217, \quad \sum x^2 = 164441.$]

(7)

(iii) State, with reasons, which measure of central location you consider more appropriate for summarising these data. Interpret the results for this measure and for the corresponding measure of variability (or spread).

(3)

(iv) Suppose you were given the corresponding areas of the 50 states and were asked to investigate the strength of the relationship between the number of farms and the area of the state. Suggest two possible measures that could be used for the strength of this relationship and briefly discuss which you would prefer.

(4)

Turn over

6. A coin is supposed to weigh 5 gm. A random sample of 100 such coins was taken and weighed, yielding the results tabulated below.

Weight (gm)	*Frequency*
< 4.895	4
≥ 4.895 but < 4.925	4
≥ 4.925 but < 4.955	11
≥ 4.955 but < 4.985	13
≥ 4.985 but < 5.015	30
≥ 5.015 but < 5.045	18
≥ 5.045 but < 5.075	11
≥ 5.075 but < 5.105	7
≥ 5.105	2

(i) Construct a histogram of these data, stating any assumptions you make.

(6)

(ii) Perform a χ^2 goodness-of-fit test of the null hypothesis that the data are from a Normal distribution with mean 5 gm. [Note: to save you calculation, the sample standard deviation of these data is 0.055 gm.]

(14)

7. (i) Compare the uses of the independent samples and paired samples *t* tests. Explain clearly in which circumstances each method should be used in preference to the other, illustrating your answer with appropriate examples. State briefly the appropriate assumptions made in each case.

(7)

(ii) In a particular population, it was of interest whether married men were, on average, younger or older than their respective wives. A random sample of 15 couples was taken, their ages being given in the table below.

Couple number	*Husband's age*	*Wife's age*
1	39	32
2	38	31
3	73	68
4	54	58
5	24	26
6	57	53
7	49	48
8	63	69
9	48	47
10	44	46
11	26	25
12	64	62
13	42	40
14	45	48
15	61	57

(a) Use the above data to calculate the mean and standard deviation of the differences between the ages of the husbands and their respective wives.

(4)

(b) Is there evidence that the mean difference in ages between husbands and wives is non-zero?

(6)

(c) Obtain a 95% confidence interval for the mean difference in ages of the husbands and their respective wives.

(3)

8. In the UK, all patients surviving a stroke are supposed to have their cholesterol levels measured soon after their stroke and regularly thereafter. A sample of medical records of men and women who had suffered a stroke was examined to determine whether there was a difference between the sexes in the proportions of stroke survivors who had had a recently recorded cholesterol measurement. The following data were obtained.

		Cholesterol level recorded	
		No	*Yes*
Sex	*Female*	109	22
	Male	97	77

(i) Perform a χ^2 test of the null hypothesis that there is no association between an individual's sex and the chance of he or she having a recently recorded cholesterol measurement.

(9)

(ii) Compute and interpret an approximate 95% confidence interval for the difference between the proportions of females and males having a recently recorded cholesterol measurement.

(8)

(iii) Comment on your conclusions to parts (i) and (ii), relating one to the other.

(3)