THE ROYAL STATISTICAL SOCIETY

HIGHER CERTIFICATE EXAMINATION

NEW MODULAR SCHEME

introduced from the examinations in 2007

MODULE 4

SPECIMEN PAPER B

AND SOLUTIONS

The time for the examination is 1½ hours. The paper contains four questions, of which candidates are to attempt **three**. 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 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)

2. (i) State and explain a linear model that can be used as the basis for a one-way analysis of variance. Explain clearly what each term in the model represents and state any assumptions required for the analysis to be valid.

(5)

(ii) A psychology researcher has the hypothesis that effective use of leisure time helps to reduce stress. In particular, she suggests that play activity is most effective when the subject feels it is free play, not directed by others.

The researcher recruited 36 college students and divided them randomly into three groups. One group received highly controlled play experience, one received a low level of control and one group performed what they would see as work rather than play.

All subjects first performed a 30-minute stress-producing task, working through mathematics problems while hearing periodic bursts of loud noise through headphones.

Next, each subject had 10 minutes at one of the three play activities described above, "high" or "low" control or "work".

Finally, the subjects attempted to solve two geometric puzzles, one of which was insoluble – but they were not told this. Persistence on the insoluble puzzle (measured in time in total seconds spent on the puzzle before giving up) was the response variable measured and used to assess the effectiveness of the play period in reducing the stress created by the work task. The table below gives the results.

High	*Low*	*Work*
347	504	398
567	420	492
424	583	97
239	183	357
256	279	184
682	381	554
435	118	354
666	317	275
825	359	198
102	77	163
601	336	284
384	197	155

Perform a one-way analysis of variance on these data and, by computing least significant differences, or otherwise, investigate differences between the three means.

Write a brief report for the researcher to use when interpreting the results.

(15)

3. (a) A psychologist wished to examine the degree of association between intelligence and the ability to think laterally. An experiment was conducted in which each member of a random sample of 12 subjects was given both an intelligence test and a test of lateral thinking. The scores obtained by each subject on each test are given in the following table. The intelligence test has a maximum score of 150 while the lateral thinking test has a maximum score of 10.

Subject	1	2	3	4	5	6	7	8	9	10	11	12
Intelligence test score	121	148	108	137	141	124	131	115	118	110	132	127
Lateral thinking test score	3	9	6	7	8	2	5	5	6	4	8	7

The psychologist analyses these data using Spearman's rank correlation coefficient.

(i) Some subjects had the same lateral thinking scores. How, if at all, should this be allowed for in the analysis?

(ii) Calculate Spearman's coefficient between these two test scores, and test it for statistical significance.

(iii) What does your result indicate about the association between intelligence and the ability to think laterally?

(12)

(b) Given a random sample of paired data (x_1, y_1), (x_2, y_2), …, (x_n, y_n), write down the formula for the sample product-moment correlation coefficient, r say, and explain the meaning of this quantity. Also draw scatter diagrams to illustrate

(i) strong positive correlation,

(ii) strong negative correlation,

(iii) independent data,

(iv) data that are uncorrelated but not independent.

(8)

4. (i) Two basic principles of experimental design are *randomisation* and *replication*. Explain why these are important and how they help to validate an analysis of experimental results.

(6)

(ii) Three species of tree were grown in a forestry plantation. Not all the seedlings survived and so the replications, r_i, were not the same for each species. The data shown in the following table are the heights (in metres) of growth made in a fixed time.

Species	*Replicates*	*Observations*	*Total*
Pinus caribea	9	4.20 4.30 3.50 3.90 5.00 4.80 4.60 4.50 4.00	38.80
Pinus kesiya	12	3.95 3.85 4.25 4.70 4.15 3.30 3.65 3.70 3.95 4.00 3.70 4.30	47.50
Eucalyptus deglupta	8	7.95 8.10 8.30 6.60 7.50 7.70 7.25 8.00	61.40
Total	29		147.70

The sum of squares of all 29 observations is 831.8900.

Carry out the usual one-way analysis of variance to examine whether there are overall differences between the species.

Examine in particular whether there are differences

(a) between *pinus* and *eucalyptus*,

(b) between the two *pinus* species.

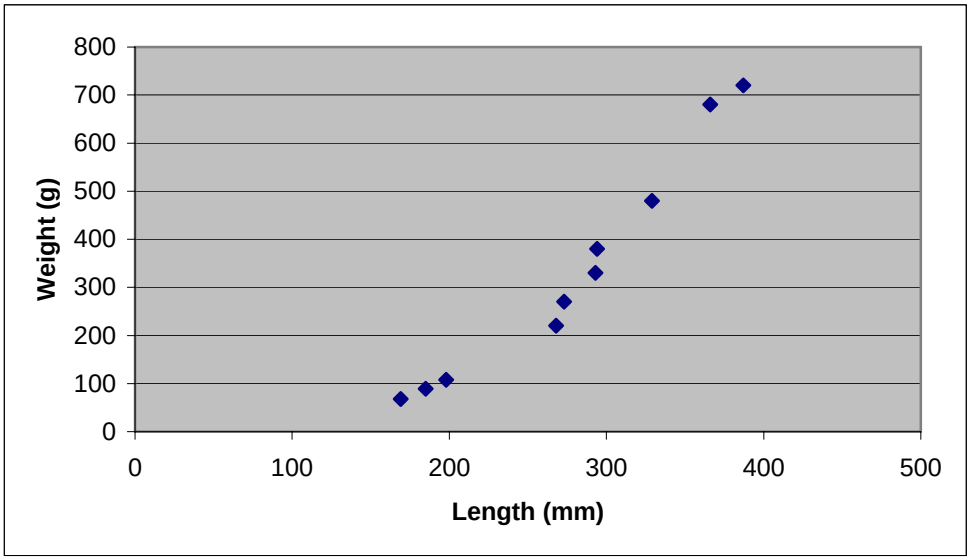
Calculate the residuals for each of the observations, make a dot-plot of these and comment on any information this gives.

(14)

SOLUTIONS

Question 1

(i) The graph suggests that a linear fit will be reasonable, at least as a first approximation. There may be curvature in the relation of weight and length.



(ii) (a) $\hat{\beta}_1 = \dfrac{S_{xy}}{S_{xx}}$.

$$S_{xy} = \sum x_i y_i - \frac{\left(\sum x_i\right)\left(\sum y_i\right)}{n} = 1075861 - \frac{2762 \times 3345}{10} = 151972.0.$$

$$S_{xx} = 812594 - 2762^2/10 = 49729.6.$$

$$\therefore \hat{\beta}_1 = 3.056.$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 334.5 - (3.056 \times 276.2) = -509.56.$$

(b) The first three points are not fitted at all well. Note that the intercept is –509.56, whereas it looks as if it should be much nearer 0 if these are to be fitted. But the remaining points are fitted reasonably well. It would however be sensible to examine also a quadratic relationship.

(iii) The coefficient of determination is given by $R^2 = S_{xy}^2/S_{xx}S_{yy}$.

$$S_{yy} = 1610009 - \frac{3345^2}{10} = 491106.5.$$

$$\therefore R^2 = \frac{151972.0^2}{49729.6 \times 491106.5} = 0.9457.$$

Thus 94.6% of the total variation in the weights of the sea bass is explained by a linear relationship with their lengths.

Question 2

(i) $y_{ij} = \mu + t_i + \varepsilon_{ij}, \qquad i = 1, 2, ..., k, \quad j = 1, 2, ..., r_i, \qquad \{\varepsilon_{ij}\} \sim \text{ind } \mathrm{N}(0, \sigma^2)$.

There are k treatments, indexed by i = 1, 2, …, k. In the experiment or survey, there are r_i units (individuals) in the ith group, i.e. receiving the ith treatment. y_{ij} is the observation (response) for the jth individual in group i.

μ is the overall population general mean. t_i is the population mean effect (departure from μ) due to treatment i, with $\Sigma_i t_i = 0$.

The Normally distributed residual (error) terms ε_{ij} all have variance σ^2 and are uncorrelated (independent).

This is an additive model: the components add together, and together explain all the variation in the responses.

(ii) The "treatments" here are "high", "low" and "work". $r_1 = r_2 = r_3 = 12$.

Totals are:	High	Low	Work
	5528	3754	3511

The grand total is 12793. $\Sigma\Sigma y_{ij}^2 = 5719139$.

"Correction factor" is $\dfrac{12793^2}{36} = 4546134.694$.

Therefore total SS = 5719139 – 4546134.694 = 1173004.306.

SS for treatments = $\dfrac{5528^2}{12} + \dfrac{3754^2}{12} + \dfrac{3511^2}{12} - 4546134.694 = 202067.056$.

The residual SS is obtained by subtraction.

Hence the analysis of variance table is as follows (SS and MS entries are slightly rounded).

SOURCE	DF	SS	MS	F value
Treatments	2	202067	101034	3.43 Compare $F_{2,33}$
Residual	33	970937	29422	$= \hat{\sigma}^2$
TOTAL	35	1173004		

The upper 5% point of $F_{2,33}$ is about 3.3; the treatments effect is significant. There is evidence to reject the null hypothesis that all treatments have the same effect.

Solution continued on the next page

To investigate treatment differences, first calculate the treatment means, which are (in ascending order, for clarity)

Work : 292.58 Low : 312.83 High : 460.67

The least significant difference between any pair of these means is

$$t_{33}\sqrt{\frac{2\times 29422}{12}} = 70.026\,t_{33} \quad \text{where} \quad t_{33} = \begin{cases} 2.035 & \text{at } 5\% \\ 2.736 & \text{at } 1\% \\ 3.617 & \text{at } 0.1\% \end{cases}$$

so the least significant differences are 142.50 for 5%, 191.59 for 1% and 253.28 for 0.1%. Thus the only apparent difference is that "high" gives a larger mean response than "low" and "work", judged at the 5% level; "low" and "work" do not differ.

Report

After carrying out an analysis which compares group means against internal variability of responses in the groups, we find some evidence that "high" shows more persistence than the other two groups, whose results are quite similar. The within-group variability is very high.

Question 3

Part (a)

(i) Any scores occurring for more than one subject lead to an average rank being given. This is shown for some subjects in part (b).

(ii)

Subject	1	2	3	4	5	6	7	8	9	10	11	12
IT rank	5	12	1	10	11	6	8	3	4	2	9	7
LT rank	2	12	6½	8½	10½	1	4½	4½	6½	3	10½	8½
Difference d_i	3	0	–5½	1½	½	5	3½	–1½	–2½	–1	–1½	–1½

$\Sigma d_i^2 = 9 + 0 + \ldots + 2.25 = 93.$

Spearman's coefficient $r_S = 1 - \dfrac{6\Sigma d^2}{n\left(n^2-1\right)} = 1 - \dfrac{558}{1716} = 0.675$.

From tables, this is significantly different from zero at the 5% level (two-sided).

[Note. Use of this formula for r_s leads to slight inaccuracy where there are tied ranks – but unlikely to make much difference here.]

(iii) There is evidence to reject the null hypothesis of no correlation, so it seems that there is some association between intelligence and the ability to think laterally; but it does not appear that the relationship is a very strong one.

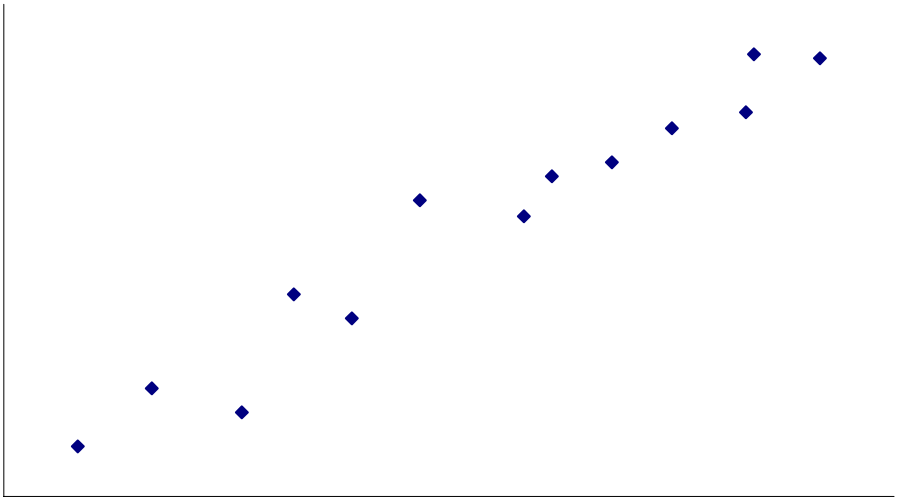
The solution to part (b) starts on the next page

Part (b)

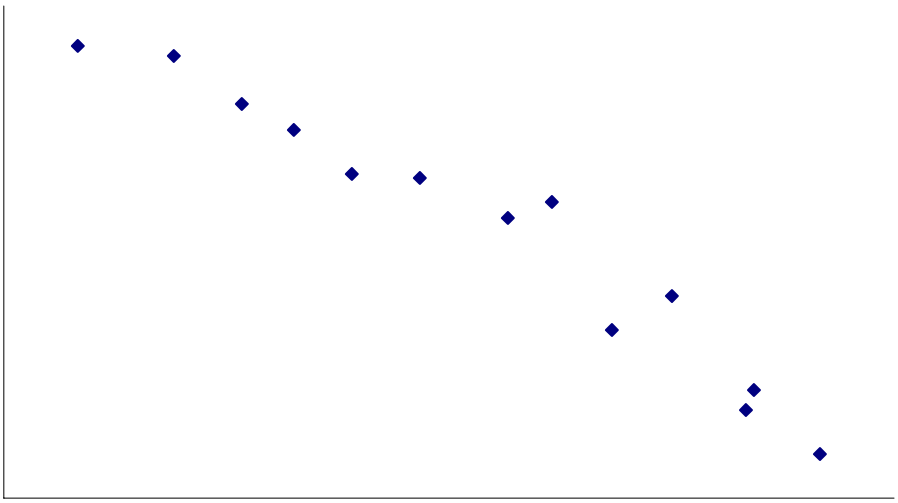
$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}.$$

This explains the strength of linear relationship between the x_i and y_i, with $r = \pm 1$ showing linearity and $r = 0$ showing no linear relationship. The underlying X and Y are both random variables.

(i) r near to +1, small amount of scatter about an (increasing) linear relationship

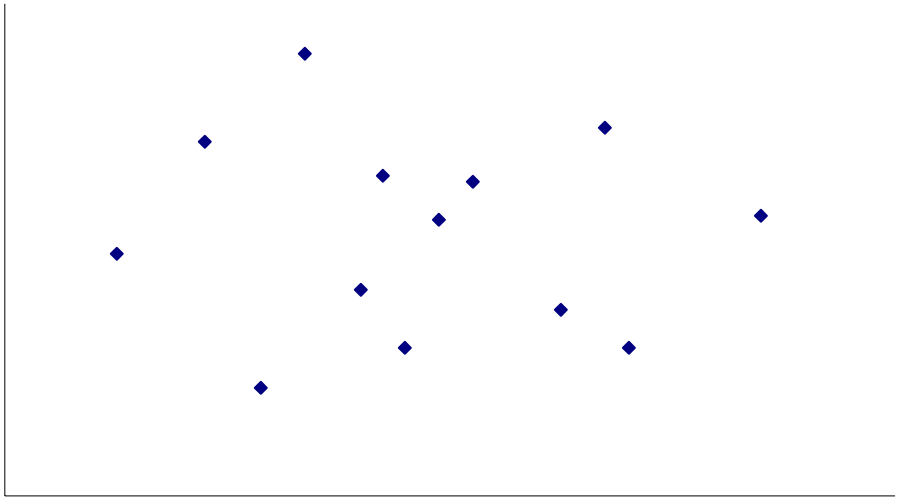


(ii) r near to –1, y decreases as x increases, otherwise as in (a)

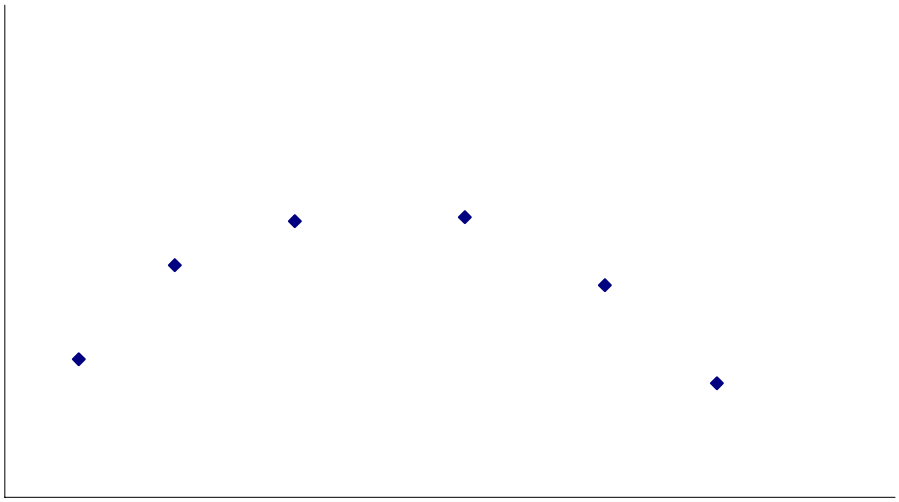


Solution continued on the next page

(iii) Independent data ($r \approx 0$)



(iv) Non-linear relationship, e.g. $y = x^2$



Question 4

(i) The existence of random natural variation (plot-to-plot, unit-to-unit) among experimental material must be recognised and catered for. The results from several unit plots must be used in assessing the response to a treatment, to obtain satisfactory estimates of treatment mean responses and to estimate the natural variability among plots treated alike. This is why there must be replication.

Randomisation of allocation of treatments to individual units helps to guard against (possibly unsuspected) sources of bias, by ensuring that there is no conscious allocation of particular treatments to better units or poorer units. Each unit has the same chance of receiving any one of the treatments being studied. The usual analysis of variance model [see question 2 on this paper for details] requires each observation to have the same underlying variance and that there are no correlations among the observations; randomisation should help achieve this.

[Note. Other devices such as blocking may be needed also. These are explored in module 6 of the Higher Certificate and, in greater depth, in parts of the Graduate Diploma.]

(ii) The grand total is 147.70. The sum of squares of all 29 observations is 831.8900. (These are given in the question.)

"Correction factor" is $\dfrac{147.70^2}{29} = 752.2514$.

Therefore total SS = 831.8900 – 752.2514 = 79.6386.

SS for species = $\dfrac{38.80^2}{9} + \dfrac{47.50^2}{12} + \dfrac{61.40^2}{8} - 752.2514 = 74.2855$.

The residual SS is obtained by subtraction.

Hence the analysis of variance table is as follows.

SOURCE	DF	SS	MS	*F* value
Species	2	74.2855	37.1428	180.4 Compare $F_{2,26}$
Residual	26	5.3531	0.2059	$= \hat{\sigma}^2$
TOTAL	28	79.6386		

The upper 0.1% point of $F_{2,26}$ is 9.12; the species effect is very highly significant. There is very strong evidence to reject the null hypothesis that all the species perform similarly.

Solution continued on the next page

We use *t* tests to examine the particular differences in (a) and (b).

(a) Overall *pinus* mean = $\frac{38.8+47.5}{21} = 4.110$; *eucalyptus* mean = 7.675.

Variance of difference is estimated by $\frac{\hat{\sigma}^2}{21} + \frac{\hat{\sigma}^2}{8} = 0.035542$.

So the *t* statistic for examining the difference is

$$\frac{7.675 - 4.110}{\sqrt{0.035542}} = 18.91,$$

which is clearly extremely highly significant as an observation from t_{26}. There is overwhelming evidence of a difference in mean heights of growth between *pinus* and *eucalyptus*.

(b) Means of the two *pinus* species are $\frac{38.8}{9} = 4.311$ and $\frac{47.5}{12} = 3.958$.

Variance of difference is estimated by $\frac{\hat{\sigma}^2}{9} + \frac{\hat{\sigma}^2}{12} = 0.040036$.

So the *t* statistic for examining the difference is

$$\frac{4.311 - 3.958}{\sqrt{0.040036}} = 1.76,$$

which is not significant as an observation from t_{26}. (the double-tailed 5% point is 2.056) There is no evidence of a difference in mean heights of growth between the two *pinus* species.

The residuals are (obs – fitted), i.e. here (obs – treatment mean):

PC	–0.11	–0.01	–0.81	–0.41	0.69	0.49	0.29	0.19	–0.31			
PK	–0.01	–0.11	0.29	0.74	0.19	–0.66	–0.31	–0.26	–0.01	0.04	–0.26	0.34
ED	0.27	0.42	0.62	–1.08	–0.18	0.02	–0.43	0.32				

See dotplots below. PK looks the most "Normal". ED has more scatter and a possible outlier.