University *of* Limerick
Ollscoil Luimnigh

FACULTY OF SCIENCE AND ENGINEERING

DEPARTMENT OF MATHEMATICS & STATISTICS

END OF SEMESTER ASSESSMENT PAPER

MODULE CODE: MA4605 **SEMESTER: Autumn 2008/2009**

MODULE TITLE: Chemometrics **DURATION OF EXAM: 2.5 hours**

LECTURER: Dr. N. Coffey **GRADING SCHEME: Examination: 90%**

EXTERNAL EXAMINER: Prof. A. Bowman

INSTRUCTIONS TO CANDIDATES

Answer Question 1 (15%) and THREE other questions (25% each)

Statistical tables are available from the invigilators. A set of formulae is attached to this paper. Calculators may be used.

Q1. (a). A new technique was developed to extract a contaminant from solids. Ten fish were injected with a particular contaminant and the percentage of contaminant extracted from each fish (recovery percentage) was recorded. The recovery percentages were

$$90, 100, 80, 89, 92, 91, 95, 96, 89, 93.$$

The chemist wants to test whether the data support the conclusion that the mean recovery rate is greater than 90%. The correct null and alternative hypotheses to test this claim is which of the following?

(i) $H_0: \ \mu \leq 90, H_1: \ \mu > 90$

(ii) $H_0: \ \mu \geq 90, H_1: \ \mu < 90$

(iii) $H_0: \ \mu = 90, H_1: \ \mu \neq 90$

(iv) None of the above.

[3 marks]

(b). Explain the term 'multicollinearity'. If multicollinearity exists, what are its implications?

[3 marks]

(c). A balanced ANOVA to determine if 3 different material types (A, B, C) affected tensile strength of a product was carried out. Ten experimental units were examined in each group. The ANOVA table was calculated and is shown below:

```
One-way ANOVA: Strength versus Material

Source    DF      SS      MS      F      P
Material   2  257.07  128.53  49.86  0.000
Error     27   69.60    2.58
Total     29  326.67

Material   N    Mean
   A      10   21.800
   B      10   27.800
   C      10   21.400
```

Use Fisher's LSD to test the following hypotheses (use $\alpha = 0.05$):

(i) $H_0: \mu_A = \mu_B$

(ii) $H_0: \mu_A = \mu_C$

[3 marks]

(d). For the confidence interval estimation of μ when σ is known and the sample size is large, the proper distribution to use is which of the following:

(i) the normal distribution

(ii) the t distribution with n degrees of freedom

(iii) the t distribution with n-1 degrees of freedom

(iv) the t distribution with n-2 degrees of freedom

(v) none of the above

[3 marks]

(e). A normally distributed quality characteristic is monitored through the use of an $\bar{X}$/R chart. These charts have the following parameters. Both charts are in control.

	LCL	Centre Line	UCL
$\bar{X}$-Chart:	614.0	620.0	626.0
R-Chart:	0	8.236	18.795

(i) What sample size is being used?

(ii) Estimate the standard deviation of the process.

[3 marks]

Q2. (a). As part of a class exercise, a first-grade teacher recorded the heights in inches of his 73 students, hoping to determine whether the class was consistent with government standards on height. The government states that the true average height of all first-graders is 50 inches. The following MINITAB output was produced.

```
One-Sample T: Hgt

Test of mu = 50 vs not = 50

Variable   N    Mean  StDev  SE Mean       95% CI           T      P
Hgt       73  43.908  8.333    0.975  (41.963, 45.852)  -6.25  0.000
```

(i) Are the students in this class consistent with government standards? Justify your answer.

[3 marks]

(ii) Interpret the 95% confidence interval produced for the above example.

[3 marks]

(b). Two types of plastic (plastic 1 and plastic 2) are suitable for use by an electronic calculator manufacturer. The breaking strength of this plastic is important. A random sample of size $n_1 = 11$ was taken from plastic 1, with average breaking strength of 166.5 psi and standard deviation of 1.23 psi and a random sample of size $n_2 = 12$ was taken from plastic 2 with average breaking strength of 155.0 psi and standard deviation of 1.0 psi.

(i) Distinguish between matched pairs and independent groups. Into which category do the above data fall? Give reasons.

[2 marks]

(ii) Do the 2 plastics have variances that differ significantly? Use $\alpha = 0.05$.

[5 marks]

(iii) The company will not adopt plastic 1 unless its breaking strength exceeds that of plastic 2 by more than 10 psi. Set up and test an appropriate hypothesis using $\alpha = 0.05$ to determine if the company should use plastic 1. State clearly the hypothesis being tested and your conclusions.

[8 marks]

(iv) Construct and interpret a 99% confidence interval for the true mean difference in breaking strength between plastic 1 and plastic 2.

[4 marks]

Q3. To determine the effect of standing times on the yield of a chemical process, the yield was measured using five batches of raw material, five acid concentrations and five standing times (A, B, C, D, E). The following results were obtained.

	Acid Conc					
Batch	I	II	III	IV	V	T_j
1	A = 26	B = 16	C = 19	D = 16	E = 13	90
2	B = 18	C = 21	D = 18	E = 11	A = 21	89
3	C = 20	D = 12	E = 16	A = 25	B = 13	86
4	D = 15	E = 15	A = 22	B = 14	C = 17	83
5	E = 10	A = 24	B = 17	C = 17	D = 14	82
T_i	89	88	92	83	78	

$$S = \sum\sum y_{ij}^2 = 7832, \quad T = \sum\sum y_{ij} = 430$$

(a). The above data are an example of a particular experimental design. What is the general name given to this type of design? Write down an appropriate model for the design. Name one advantage and one disadvantage of this design.

[3 marks]

(b). Explain the term 'blocking' in the context of ANOVA. For the above example, distinguish between the treatment and the blocking variables involved.

[3 marks]

(c). Calculate the sum of squares values for batch, acid concentration and standing time.

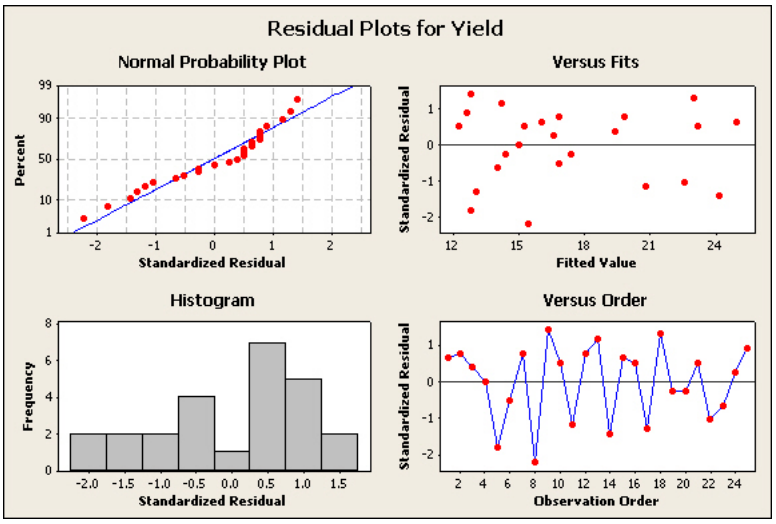
[9 marks]

(d). Set up the ANOVA table and test for significant differences between batches, acid concentrations and standing times. State clearly the hypotheses being tested and your conclusions. Use $\alpha = 0.05$.

[8 marks]

(e). What assumptions are required to carry out the analysis? Based on the output in the figure below, comment on the validity of these assumptions for the above data. Give reasons for your answer.

[2 marks]


Residual Plots for Yield
Normal Probability Plot
Percent
Standardized Residual
Versus Fits
Standardized Residual
Fitted Value
Histogram
Frequency
Standardized Residual
Versus Order
Standardized Residual
Observation Order

Q4. An automobile assembly plant concerned about quality improvement measured sets of five camshafts on twenty occasions throughout the day. The specification limits set out by the production engineer were 600mm ± 3mm.

Sample	**Lengths**					**Means**	**Ranges**
1	598.0	599.8	600.0	599.8	600.0	599.52	2.00
2	600.0	598.8	598.2	599.4	599.6	599.20	1.80
3	599.4	599.4	600.0	598.8	599.2	599.36	1.20
4	599.4	599.6	599.0	599.2	600.6	599.56	1.60
5	598.8	598.8	599.8	599.2	599.4	599.20	1.00
6	600.0	600.2	600.2	599.6	599.0	599.80	1.20
7	599.0	599.8	600.8	598.8	598.2	599.32	2.60
8	600.0	599.2	599.8	601.2	600.4	600.12	2.00
9	600.2	599.6	599.6	599.6	600.2	599.84	0.60
10	599.2	599.0	599.6	600.4	600.0	599.64	1.40
11	599.0	599.6	599.4	599.2	597.8	599.00	1.80
12	600.4	599.6	600.0	600.8	600.4	600.24	1.20
13	599.4	599.0	598.4	599.0	599.6	599.08	1.20
14	598.8	599.2	599.6	598.6	599.8	599.20	1.20
15	599.6	599.2	599.6	600.2	599.8	599.68	1.00
16	599.6	600.0	599.6	599.2	598.6	599.40	1.40
17	599.6	601.2	599.6	600.2	600.0	600.12	1.60
18	600.0	599.4	599.8	599.2	599.6	599.60	0.80
19	599.4	600.0	600.0	599.2	599.4	599.60	0.80
20	599.6	599.8	599.0	599.6	599.4	599.48	0.80
						$\bar{\bar{X}} = 599.55$	$\bar{R} = 1.34$

(a). Give two different signs that would indicate that a process is out of control.

[2 marks]

(b). Calculate the control limits for the $\bar{X}$ and R charts.

[4 marks]

(c). The data are plotted on the final page of the exam paper, which can be detached and included in your answer sheet. Using the limits calculated in part (ii), check

for process control. Explain why two types of charts are required.

[6 marks]

(d). Is the process capable? Based on your conclusion, draw an appropriate diagram to indicate whether the process is performing within specification limits, exactly on the specification limits or outside of specification limits.

[6 marks]

(e). Calculate the ARL (average run length) for a change of +0.1mm in the average. In words, interpret what this calculated value of the ARL means.

[5 marks]

(f). What is a CUSUM chart? What type of departures from the production target value is this type of chart useful for detecting?

[2 marks]

Q5. A soft drink bottler is interested in obtaining more uniform fill heights in the bottles produced by his manufacturing process. The process engineer can control three variables during the filling process: the percent carbonation (A), the operating pressure in the filler (B) and the line speed (C). Two levels of all three variables are investigated:

Factor	Low (-1)	High (+1)
Percent carbonation, % (A)	10	12
Pressure, psi (B)	25	30
Speed, b/m (C)	200	250

Readings of the deviation from the fill height for each combination of the levels of these factors were recorded, and each combination was tested in duplicate.

	Effect			Deviation	
Run	A	B	C	Rep1	Rep2
(1)	-1	-1	-1	-3	-1
a	+1	-1	-1	0	1
b	-1	+1	-1	-1	0
ab	+1	+1	-1	2	3
c	-1	-1	+1	-1	0
ac	+1	-1	+1	2	1
bc	-1	+1	+1	1	1
abc	+1	+1	+1	6	5

$$S = \sum\sum y_{ij}^2 = 94, \quad T = \sum y_{ij} = 16$$

(a). Figure 1 shows the main effects plot and the interactions plot for these data. Using these plots, which factors have an effect on the deviation from fill height? Is there evidence of any interaction between factors? Give reasons.

[5 marks]

(b). Set up the contrast table for the main effects and interactions. Calculate the contrasts, the main effects and the sum of squares for the effects.

[8 marks]

(c). Set up the associated ANOVA table and test for significant main effects and interactions. State clearly your conclusions.

[7 marks]

(d). Based on the results of part (iii), what settings would you use to minimise the deviations? What is the expected deviation at the optimum settings?

[5 marks]

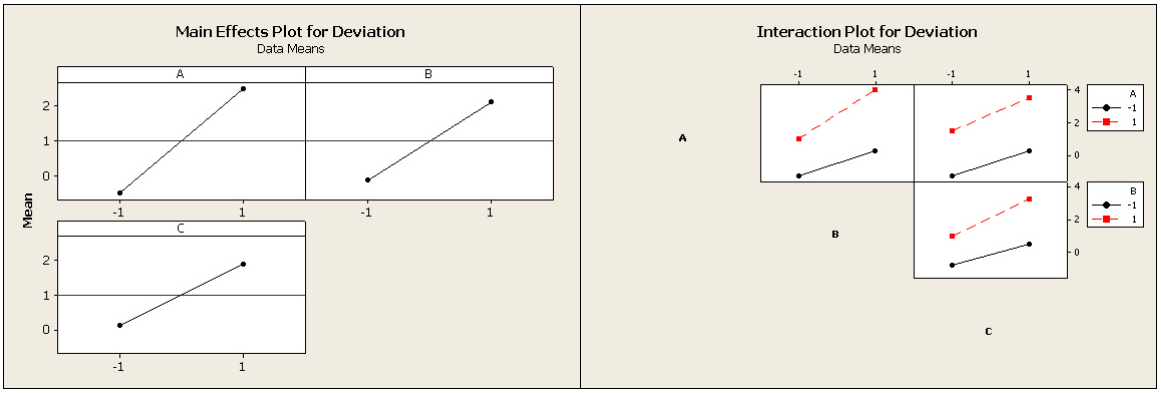


Figure 1: Main effects plot and Interactions plot.

Q6. A hospital administrator wished to study the relation between patient satisfaction (Y) and the patient's age (X_1, in years). She randomly selected 22 patients and collected the data some of which is presented below, where larger values of Y indicated more satisfaction.

i	1	2	3	4	5	6	7	8	$\cdots$	22
y_i	48	57	66	70	89	36	46	54	$\cdots$	52
x_{i1}	50	36	40	41	28	49	42	45	$\cdots$	44

$$SS_{xx} = \sum (x_i - \bar{x})^2 = 1565.45 \qquad \sum x_i = 868$$
$$SS_{yy} = \sum (y_i - \bar{y})^2 = 6143.32 \qquad \sum y_i = 1351$$
$$SS_{xy} = \sum (y_i - \bar{y})(x_i - \bar{x}) = -2404.09$$

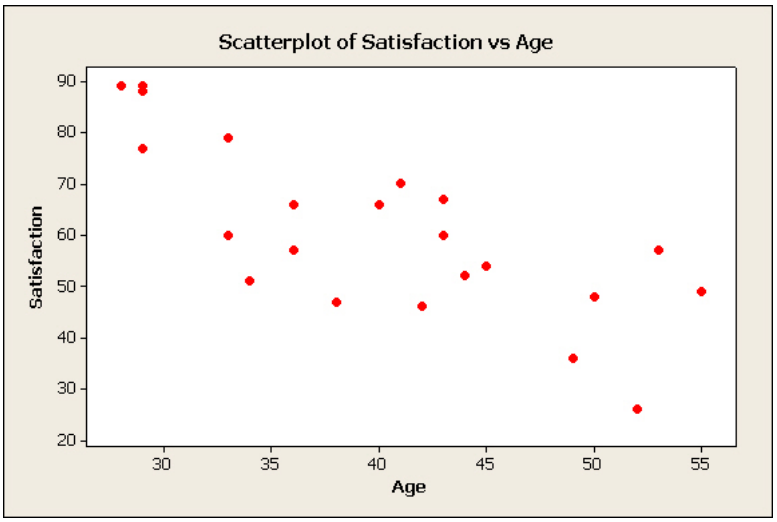


Figure 2: Scatterplot

(a). Figure 2 displays a scatterplot for the above data. Comment on the relationship between satisfaction levels and patient's age. Is the use of simple linear regression appropriate for these data? Justify your answer.

[2 marks]

(b). Fit a linear regression model and interpret the meaning of the coefficients of the intercept and slope.

[4 marks]

(c). The following MINITAB output displays the standard errors of the slope and intercept estimates obtained in part (b).

```
Predictor     Coef  SE Coef   T
Constant             11.29
Age                 0.2798
```

Test (using $\alpha = 0.05$) whether there is a significant linear relationship between satisfaction levels and age. Clearly state the null and alternative hypothesis and your conclusion.

[4 marks]

(d). The administrator also believes that severity of illness (X_2) and anxiety level (X_3) may also affect patient satisfaction levels. (Note: larger values of X_2 and X_3 are associated with increased severity of illness and more anxiety respectively.)

The MINITAB printout for fitting the model:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i$$

is as follows:

```
Regression Analysis: Satisfaction versus Age, Severity of , Anxiety Leve

The regression equation is
Satisfaction = 163 - 1.21 Age - 0.666 Severity of Illness - 8.6 Anxiety Level

Predictor               Coef  SE Coef
Constant              162.47    26.51
Age                  -1.2179   0.3112
Severity of Illness  -0.6505   0.8452
Anxiety Level          -8.69    12.56

S = 10.2895

Analysis of Variance

Source          DF       SS      MS     F
Regression      (i)  4137.2  1379.1   (vi)
Residual Error (ii)     (iv)     (v)
Total         (iii)  6143.3
```

(i) Complete the ANOVA table by filling in the values for (i)-(vi). Conduct a hypothesis test to determine the significance of the linear regression model.

State clearly the hypothesis being tested and your conclusion. Use $\alpha = 0.05$.

[7 marks]

(ii) Test each variable X_1, X_2 and X_3 for inclusion in the model. State clearly the hypotheses being tested and your conclusions. Use $\alpha = 0.05$.

[6 marks]

(iii) Calculate R^2 and comment.

[2 marks]

Formulae Sheet - 1 of 2

Parameter (Population Value)	Statistic (Sample Value)	Standard Error
μ	$\bar{x}$	$SE(\bar{x}) = \frac{s}{\sqrt{n}}$
μ_d	$\bar{x}_d$	$SE(\bar{x}_d) = \frac{s_d}{\sqrt{n}}$ **where** $s_d = \sqrt{\frac{\sum d_i^2 - \frac{(\sum d_i)^2}{n}}{n-1}}$
$\mu_1 - \mu_2$	$\bar{x}_1 - \bar{x}_2$	$SE(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

Small sample - n < 30:

Parameter (Population Value)	Statistic (Sample Value)	Standard Error
$\mu_1 - \mu_2$	$\bar{x}_1 - \bar{x}_2$	$SE(\bar{x}_1 - \bar{x}_2) = \sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$ **where** $s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1 + n_2 - 2}$

Control Charts:

$\bar{X}$ chart: Process mean and standard deviation known	$LCL = \mu - 3\frac{\sigma}{\sqrt{n}} \quad UCL = \mu + 3\frac{\sigma}{\sqrt{n}}$
$\bar{X}$ chart: Process mean and standard deviation unknown	$LCL = \bar{\bar{x}} - A_2\bar{R} \quad UCL = \bar{\bar{x}} + A_2\bar{R}$
Relationship between R and σ	$\bar{R} = d_2\hat{\sigma}$
R chart	$LCL = D_3\bar{R} \quad UCL = D_4\bar{R}$

Formulae sheet - 2 of 2

- $$C_P = \frac{USL - LSL}{6\hat{\sigma}}$$
$$C_{PU} = \frac{USL - \bar{\bar{x}}}{3\hat{\sigma}} \qquad C_{PL} = \frac{\bar{\bar{x}} - LSL}{3\hat{\sigma}} \qquad C_{PK} = \min(C_{PU}, C_{PL})$$

- $$P[> UCL]: \ Z = \frac{UCL - \mu_{shift}}{\hat{\sigma}/\sqrt{n}} \qquad P[< LCL]: \ Z = \frac{LCL - \mu_{shift}}{\hat{\sigma}/\sqrt{n}}$$

- $$SS_{Total} = S - T^2/n \qquad SS_T = \sum_k T_k^2/n_k - T^2/n$$
$$SS_{B1} = \sum_i T_i^2/n_i - T^2/n \qquad SS_{B2} = \sum_j T_j^2/n_j - T^2/n$$

- $$LSD = t_{(\alpha/2,\textbf{df error})}\sqrt{MS_{Error}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

- $$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

- $$SS_{xy} = \sum(x_i - \bar{x})(y_i - \bar{y})$$
$$SS_{xx} = \sum(x_i - \bar{x})^2$$
$$SS_{yy} = \sum(y_i - \bar{y})^2$$

- $$\hat{\beta}_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{SS_{xy}}{SS_{xx}}$$
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$$

- $$SS_{Reg} = \sum(\hat{y}_i - \bar{y})^2 = (SS_{xy})^2/SS_{xx}$$

- $$Estimate = \frac{1}{2^{k-1}r}[Contrast] \qquad \frac{1}{2^k r}[Contrast]^2$$

Student Name:	
Student ID: