University *of* Limerick
Ollscoil Luimnigh

FACULTY OF SCIENCE AND ENGINEERING

DEPARTMENT OF MATHEMATICS & STATISTICS

END OF SEMESTER ASSESSMENT PAPER

MODULE CODE: MA4605

SEMESTER: Autumn Repeats 2008/2009

MODULE TITLE: Chemometrics

DURATION OF EXAM: 2.5 hours

LECTURER: Dr. N. Coffey

GRADING SCHEME: Examination: 100%

EXTERNAL EXAMINER: Prof. A. Bowman

INSTRUCTIONS TO CANDIDATES

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

Statistical tables are available from the invigilators. Table 25 and a set of formulae is attached to this paper. Calculators may be used.

Q1. (a). Sales personnel for a particular distributor submit weekly reports listing the customer contacts made during the week. A sample of 8 weekly reports yields a sample mean of 19.5 customer contacts per week and a sample standard deviation of 5.2. The 95% confidence interval for the population mean number of customer contacts per week is:

(i) 15.152 to 23.848

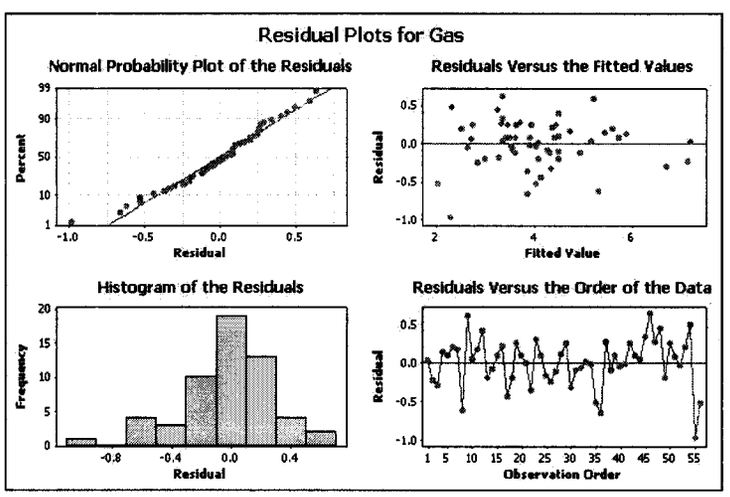
(ii) 15.897 to 23.103

(iii) 16.016 to 22.984

(iv) 15.161 to 23.7395.

[5 marks]

(b). What assumptions are required for simple linear regression analysis? Plots of residuals produced in an analysis using simple linear regression are shown below. Comment on the validity of the assumptions for this analysis based on these plots.



[5 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) $H_0 : \mu_A = \mu_C$

[5 marks]

(d). A Type II error is the error of:

(i) accepting H_0 when it is false

(ii) accepting H_0 when it is true

(iii) rejecting H_0 when it is false

(iv) rejecting H_0 when it is true

(v) both (a) and (c) are correct.

[5 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:	4.266	5.42	6.574
R-Chart:	0	2.0	4.228

(i) What sample size is being used?

(ii) Estimate the standard deviation of the process.

[5 marks]

Q2. (a). A new software package is developed to help reduce the time required to design and develop an information system. To evaluate the benefits of the new software package, 22 analysts are given specifications for a hypothetical information system. 11 of these analysts are instructed to produce the information system using the current technology while the remaining 11 analysts are instructed to produce it using the new technology. The time taken to complete the project is recorded for each analyst. The researcher in charge claims that the new software package reduces the project completion time. The following results were recorded:

New Technology	**Current Technology**
$n_1 = 11$	$n_2 = 11$
$\bar{x}_1 = 288.1$ hours	$\bar{x}_2 = 328.8$ hours
$s_1 = 45.5$	$s_2 = 39.6$

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

[2 marks]

(ii) Do the two software packages have variances that differ significantly? Use $\alpha = 0.05$.

[5 marks]

(iii) Write down the correct null and alternative hypotheses to test the researcher's claim.

[2 marks]

(iv) Test the hypothesis created in part (iii). Use $\alpha = 0.05$ and clearly state your conclusions.

[6 marks]

(v) Construct and interpret a 95% confidence interval for the true mean difference in completion time between the current technology and the new technology.

[5 marks]

(b). In order to determine if men and women differed in the levels of uptake of a particular antioxidant, 20 men and 20 women had blood samples taken and the concentration of the antioxidant was measured. The following MINITAB output was produced.

```
Two-Sample T-Test and CI

Sample     N    Mean  StDev  SE Mean Males     20  10.400  0.520
0.12 Females   20   8.360  0.575     0.13

Difference = mu (Males) - mu (Females) Estimate for difference:
2.040
95% CI for difference:  (1.689, 2.391)
T-Test of difference = 0 (vs not =): T-Value = 11.77  P-Value = 0.000
DF = 38 Both use Pooled StDev = 0.5482
```

(i) From the above output, is there a difference between the average uptake of the antioxidant between men and women? Explain.

[2 marks]

(ii) Interpret the 95% confidence interval produced in the output.

[3 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.

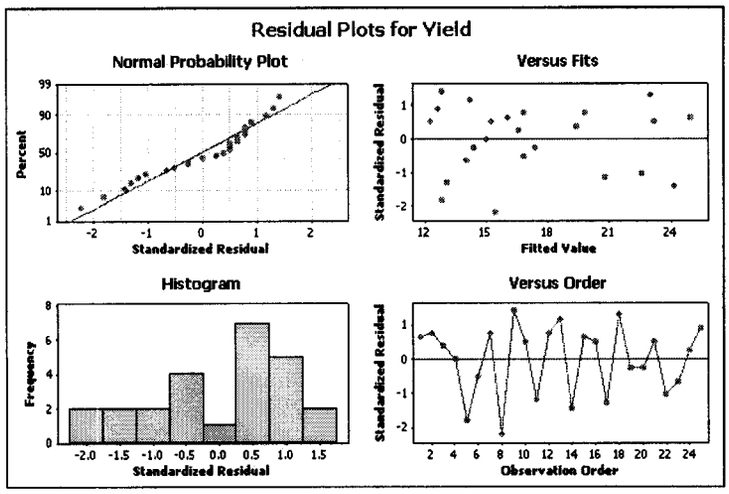
[9 marks]

(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. A computer company that manufactures computer disks takes random samples of five disks on twenty occasions throughout the day and measures the diameter. The table below displays the results.

Sample	Diameters					Means	Ranges
1	3.506	3.509	3.514	3.501	3.503	3.507	0.013
2	3.488	3.509	3.488	3.525	3.503	3.503	0.037
3	3.490	3.490	3.500	3.513	3.497	3.498	0.023
4	3.515	3.512	3.499	3.490	3.484	3.500	0.031
5	3.506	3.511	3.501	3.477	3.480	3.495	0.034
6	3.498	3.496	3.505	3.501	3.506	3.501	0.010
7	3.491	3.491	3.498	3.483	3.504	3.493	0.021
8	3.499	3.485	3.483	3.508	3.509	3.497	0.026
9	3.510	3.516	3.523	3.496	3.500	3.509	0.027
10	3.488	3.502	3.509	3.510	3.515	3.505	0.027
11	3.488	3.489	3.514	3.518	3.486	3.499	0.032
12	3.504	3.487	3.495	3.502	3.478	3.493	0.026
13	3.504	3.477	3.494	3.501	3.490	3.493	0.027
14	3.500	3.503	3.508	3.505	3.523	3.508	0.023
15	3.485	3.494	3.507	3.509	3.501	3.499	0.024
16	3.515	3.483	3.519	3.494	3.499	3.502	0.036
17	3.500	3.504	3.495	3.502	3.489	3.498	0.015
18	3.496	3.482	3.496	3.508	3.487	3.494	0.026
19	3.488	3.486	3.496	3.507	3.498	3.495	0.021
20	3.497	3.514	3.505	3.499	3.489	3.501	0.025
						$\bar{\bar{X}} = 3.499$	$\bar{R} = 0.025$

(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 it is important that the R chart is in control before examining the $\bar{X}$ chart.

[6 marks]

(d). The manufacturer specification limits are 3.5 ± 0.01mm. 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 +0.005mm 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 2^3 factorial design was used to develop a nitride etch process on a single-wafer plasma etching tool. There were three factors in the design, distance between the electrodes, gas flow and power applied to the cathode.

Factor	-1	+1
A (Gap, cm)	0.8	1.2
B (Gas flow)	125	200
C (Power, W)	275	325

Each factor was run at two levels and the design was replicated twice.

Treatment	A	B	C	y_1	y_2
(1)	-1	-1	-1	28	30
a	+1	-1	-1	22	22
b	-1	+1	-1	29	30
ab	+1	+1	-1	22	20
c	-1	-1	+1	36	37
ac	+1	-1	+1	31	34
bc	-1	+1	+1	36	37
abc	+1	+1	+1	33	34

$$S = \sum\sum y_{ij}^2 = 14,969$$

(a). Construct main effects plots for the effects of A, B, and C and comment.

[5 marks]

(b). Set up the matrix of contrasts for the main effects and interactions. Calculate the contrast, the estimate and the sum of squares for each effect.

[8 marks]

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

[7 marks]

(d). Based on the results of part (c), what settings would you use to minimise the etch rate?

[5 marks]

Q6. A Limerick freight company in the business of making deliveries throughout the mid-west region wishes to build a model which would estimate the total daily travel time (y) for their drivers. Initially the management believed that the total miles travelled (x_1) daily would be an important independent variable. A sample of 10 daily assignments gave the following results:

Assignment	$\mathbf{x_1}$ **= Miles Travelled**	**y = Travel Time (hours)**
1	100	9.3
2	50	4.8
3	100	8.9
4	100	6.5
5	50	4.2
6	80	6.2
7	75	7.4
8	65	6.0
9	90	7.6
10	90	6.1

$$SS_{xx} = \sum(x_i - \bar{x})^2 = 3450 \qquad \sum x_i = 800$$
$$SS_{yy} = \sum(y_i - \bar{y})^2 = 23.9 \qquad \sum y_i = 67$$
$$SS_{xy} = \sum(y_i - \bar{y})(x_i - \bar{x}) = 234$$

(a). (i) Calculate the sample coefficient of linear correlation and comment on the nature of the relationship between travel time and miles travelled.

[2 marks]

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

[4 marks]

(iii) The following MINITAB output displays the ANOVA table generated from fitting this regression model.

```
Analysis of Variance

Source          DF      SS      MS      F
Regression       1  15.871  15.871  15.81
Residual Error   8   8.029   1.004
Total            9  23.900
```

Test the significance of the model using the above ANOVA table. State clearly the hypotheses under consideration and your conclusions.

[4 marks]

(iv) Calculate R^2 and comment.

[2 marks]

(b). In attempting to identify another independent variable, management felt that the number of deliveries (x_2) could also be relevant. The partial print out for fitting the model

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

is as follows:

```
Regression Analysis: Travel Time versus Miles Travelled, Number of Deliveries

Predictor                    Coef  SE Coef
Constant                   -0.869    0.951
Miles Travelled (x1)        0.061    0.010
Number of Deliveries (x2)   0.923    0.221

Analysis of Variance

Source          DF       SS      MS     F
Regression      (i)   21.60     (iv)  (vi)
Residual Error (ii)    2.30     (v)
Total         (iii)   23.90
```

(i) Explain the term multicollinearity? If multicollinearity exists, what are the implications?

[2 marks]

(ii) Complete the ANOVA table by filling in the values for (i)-(vi).

[3 marks]

(iii) Test the hypothesis

$$H_0 : \beta_2 = 0.$$

Use $\alpha = 0.05$. Clearly state your conclusion.

[4 marks]

(iv) Calculate the updated value of R^2. Does including the number of deliveries (x_2) improve the model? Give reasons for your answer.

[2 marks]

(v) Write down the multiple linear regression model.

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

Formulae sheet - 2 of 2

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

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

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

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

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

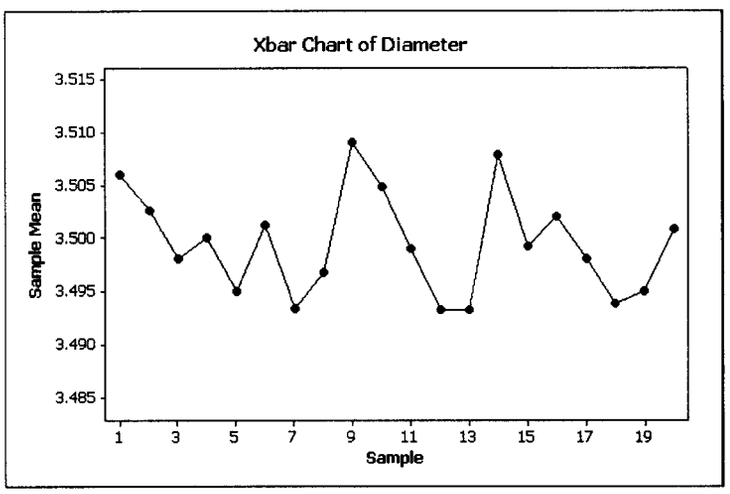
- $$SS_{xy} = \sum(x_i - \bar{x})(y_i - \bar{y})$$
$$SS_{xx} = \sum(x_i - \bar{x})^2$$
$$SS_{yy} = \sum(y_i - \bar{y})^2$$

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

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

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

Student Name:	
Student ID:	



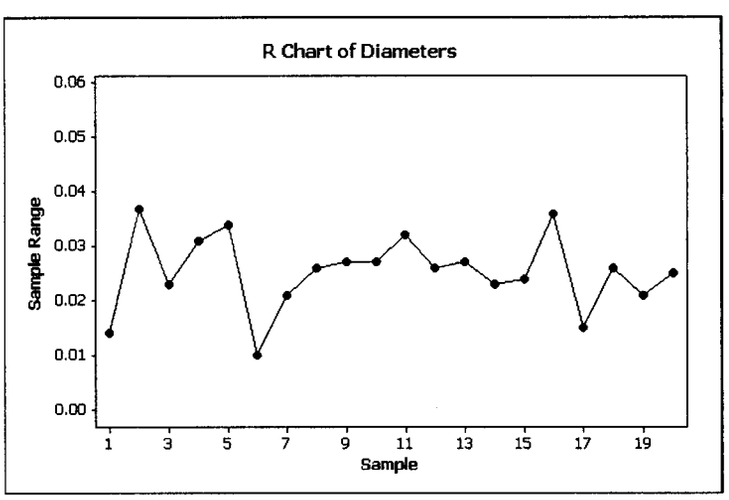


Table 25 Control Chart Factors for Mean and Range (American Usage)

X-bar and R charts

Obser-vations in sample	AVERAGES	RANGES				
	Factors for Control Limits	Factors for Central Line		Factors for Control Limits		
n	A_2	d_2	$1/d_2$	d_3	D_3	D_4
2	1.880	1.128	0.8865	0.853	0	3.267
3	1.023	1.693	0.5907	0.888	0	2.574
4	0.729	2.059	0.4857	0.880	0	2.282
5	0.577	2.326	0.4299	0.864	0	2.114
6	0.483	2.534	0.3946	0.848	0	2.004
7	0.419	2.704	0.3698	0.833	0.076	1.924
8	0.373	2.847	0.3512	0.820	0.136	1.864
9	0.337	2.970	0.3367	0.808	0.184	1.816
10	0.308	3.078	0.3249	0.797	0.223	1.777
11	0.285	3.173	0.3152	0.787	0.256	1.744
12	0.266	3.258	0.3069	0.778	0.283	1.717
13	0.249	3.336	0.2998	0.770	0.307	1.693
14	0.235	3.407	0.2935	0.763	0.328	1.672
15	0.223	3.472	0.2880	0.756	0.347	1.653
16	0.212	3.532	0.2831	0.750	0.363	1.637
17	0.203	3.588	0.2787	0.744	0.378	1.622
18	0.194	3.640	0.2747	0.739	0.391	1.608
19	0.187	3.689	0.2711	0.734	0.403	1.597
20	0.180	3.735	0.2677	0.729	0.415	1.585
21	0.173	3.778	0.2647	0.724	0.425	1.575
22	0.167	3.819	0.2618	0.720	0.434	1.566
23	0.162	3.858	0.2592	0.716	0.443	1.557
24	0.157	3.895	0.2567	0.712	0.451	1.548
25	0.153	3.931	0.2544	0.708	0.459	1.541

To derive control limits

Process Average $\overline{X}$ Chart

Upper Control Limit = Central Line $+A_2\ \overline{R}$
Lower Control Limit = Central Line $-A_2\ \overline{R}$

Central Line = $\overline{X}$ for trial limits
Central Line = Target mean for a controlled process

Range Chart

Upper Control Limit = $D_4\ \overline{R}$
Lower Control Limit = $D_3\ \overline{R}$

Central Line = $\overline{R}$ for trial limits
For achievable known process σ, replace $\overline{R}$ by $d_2\ \sigma$.