UNIVERSITY *of* LIMERICK
OLLSCOIL LUIMNIGH

COLLEGE OF INFORMATICS AND ELECTRONICS

END OF SEMESTER ASSESSMENT PAPER

MODULE CODE: **MA4605** **SEMESTER: Autumn 2003/04**

MODULE TITLE: **Chemometrics** **DURATION: 2½ hours**

LECTURER: **Dr. K. Hayes**

EXTERNAL EXAMINER: **Prof. Philip Boland**

INSTRUCTIONS TO CANDIDATES: **Do any 4 questions.**
University of Limerick approved calculators may be used.
Statistical tables may be used.

Q1 (a) The concentration of lead in the bloodstream was measured for a sample of 41 children from a large school near a busy motorway. The sample mean was 12.12 ng ml^{-1} and the standard deviation was 0.60 ng ml^{-1}. Calculate the 95% confidence interval for the mean lead concentration for all the children in the school.
What sample size is required to reduce the range of the confidence interval to 0.1 ng ml^{-1} (i.e. $\pm$0.05 ng ml^{-1})?

[8 marks]

(b) The solubility product of barium sulphate is 1.3×10^{-10}, with a standard deviation of 0.1×10^{-10}. Calculate the standard deviation of the calculated solubility of barium sulphate in water. HINT:

$$y = b^n \qquad \Rightarrow \qquad \left|\frac{\sigma_y}{E[y]}\right| = \left|\frac{n\sigma_b}{E[b]}\right|$$

[6 marks]

(c) The following data give the recovery of bromide from spiked samples of vegetable matter, measured using a gas-liquid chromatographic method. The same amount of bromide was added to each specimen. Also given is a sample of MINITAB output calculated from these data.

Tomato:	777	790	759	790	770	758	764 mg g^{-1}
Cucumber:	782	773	778	765	789	797	782 mg g^{-1}

Two-Sample T-Test and CI: Tomato, Cucumber

```
Two-sample T for Tomato vs Cucumber

          N      Mean     StDev   SE Mean
Tomato    7     772.6      13.6       5.1
Cucumber  7     780.9      10.4       3.9

Difference = mu Tomato - mu Cucumber
Estimate for difference:  -8.29
95% CI for difference: (-22.37, 5.80)
T-Test of difference = 0 (vs not =):
T-Value = -1.28  P-Value = 0.224  DF = 12
Both use Pooled StDev = 12.1
```

Do the recoveries from the two vegetables have variances which differ significantly?

[4 marks]

Do the mean recovery rates differ significantly?

[4 marks]

(d) What would Grubbs' test be used for?

[3 marks]

Q2 In some eye diseases, eye pressure is disturbed e.g. in glaucoma it is often too high. Eye pressure is determined by two continuous processes (a) the rate of formation of fluid (aqueous humour) in the eye and (b) the resistance to drainage of aqueous humour from the eye. If a dye e.g. fluorescein is added to this fluid its concentration declines as the dye is diluted by newly formed aqueous humour in the eye. This dilution should cause an exponential decline in fluorescence (i.e. fluorescein concentration) of the aqueous humour, the **rate** of **decline** being called the $\mathbf{K_{out}}$ constant. The $\mathbf{K_{out}}$ constant is a measure of the rate of formation of new aqueous humour and a successful drug to treat glaucoma is one which decreases $\mathbf{K_{out}}$ and thus lowers eye pressure.

To study and compare the effect of drugs in suppressing aqueous humour formation, groups of bovine eyes (obtained from the local abattoir) are injected via the ciliary artery with similar volumes of drug solutions (saline *(control)*, timolol, terbutaline, *(standard drugs)* mk 927 *(new drug)*). This injection of drug into the ciliary artery does not directly alter the volume of aqueous humour nor the pressure in the eye. However, enough of the drug will diffuse from the artery perfusate into the eye proper, to alter the rate of aqueous humour formation. After a suitable initial period for thorough mixing of the fluorescein, its concentration is measured at five minute intervals for 2 hours, thus generating a form of repeated measures data.

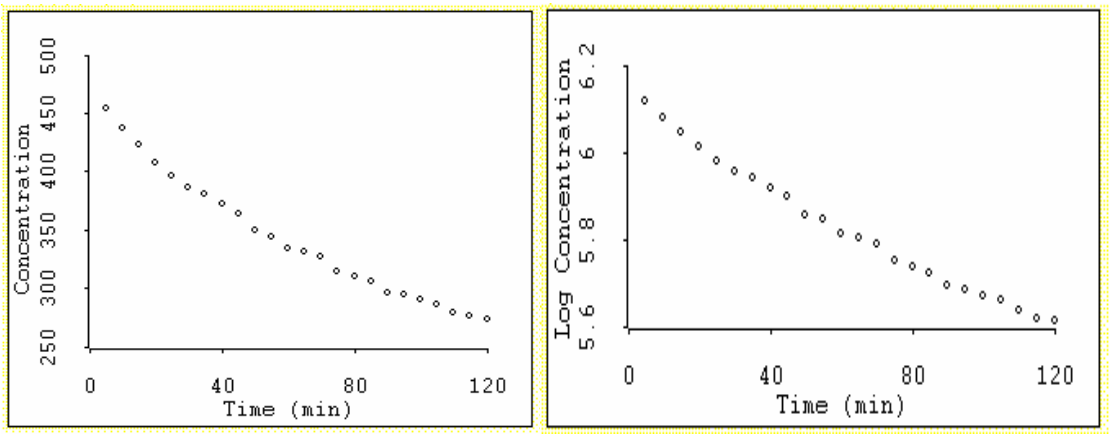
The theory that dye concentration declines exponentially with time suggests the formula :

$$\text{Dye Concentration} = \mathbf{C_0} \times \exp\{-\mathbf{K_{out}} \times \text{Time}\}$$

The $\mathbf{K_{out}}$ parameter should be positive if there is a decline, i.e. dilution in concentration. The parameter $\mathbf{C_0}$ is the concentration at time zero assuming that thorough mixing of the fluorescein has taken place. It follows that

$$\log(\text{concentration}) = \mathbf{L_0} - \mathbf{K_{out}} \times \text{Time}$$

where $\mathbf{L_0} = \log(C_0)$. Plots of concentration and log of concentration against time for a single experiment are shown below. Simple linear regression can be used to determine $\mathbf{K_{out}}$ and $\mathbf{L_0}$ for an individual curve.



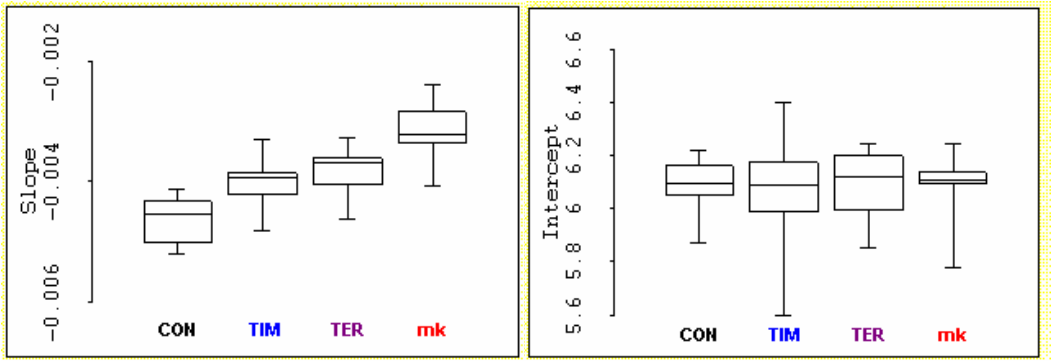
This problem is based on real data. In total 41 eyes were used in the experiment. These were distributed across the drugs as follows:

Drug

Drug	Sample Size (no. of eyes)
Control	11
timolol	14
terbutaline	7
mk 927	9

Of primary interest is comparing the effect of the drugs on the $\mathbf{K_{out}}$ constant; - the drugs used should reduce this **rate** relative to the control and hence reduce eye pressure and, if the new drug is to be better than the others, it should have the numerically smallest **rate**.

Of secondary interest is to check that **initial** concentrations of fluorescein are similar across the drugs. If not there might be some differences between the eyes used for different drugs. This could confound conclusions about differences in the effects of the drugs on the **rate** measurements.



One-way analysis of variance was carried out on the slopes of the log concentration curves with the different drugs acting as factor levels (boxplots above). Confidence intervals, with a simultaneous confidence level of 95%, for the mean of the new drug against each of the other drugs were also calculated.

Source	DF	SS	MS	F	P-value
Slopes	3	0.0000102	0.0000034	16.17	<0.001
Error	37	0.0000078	0.0000002		
Total	40	0.0000180			

New – Control	(0.000923,0.00193)
New – Timolol	(0.000345,0.00130)
New – Terbutaline	(0.0000834,0.00121)

One-way analysis of variance was also carried out on the intercepts of the log concentration curves (boxplots above).

Source	DF	SS	MS	F	P-value
Intercepts	3	0.0035	0.0012	0.04	0.988
Error	37	0.9849	0.0266		
Total	40	0.9884			

(a) In the context of statistical experimental design explain what you understand by a balanced design? Outline one disadvantage of using a design that is not balanced. Is the above design balanced ? Give reasons.

[5 marks]

(b) Using the boxplots of the slopes and intercepts only, what conclusions can be reached ? Why are these graphical displays insufficient for reaching formal conclusions about the data ?

[5 marks]

(c) In words, interpret the ANOVA tables in the context of the above problem.

[5 marks]

(d) Comment on the Bonferroni confidence intervals provided.

[5 marks]

(e) If analysis of variance is a method for comparing sample means, why the misleading name?

[5 marks]

Q3 The production of a certain type of red brick requires that the dimensions of the end product fall within specification limits set out by the production engineer. The target width for the bricks is 140 mm $\pm$ 2mm. Four bricks were selected at random every 5 minutes and their widths recorded. Some of these data are displayed below, along with the sample to which each brick belongs.

Sample					means	ranges
1	140.5	139.9	140.0	140.5	140.22	0.6
2	139.7	140.0	139.8	140.2	139.93	0.5
3	139.5	139.8	139.8	140.6	139.93	1.1
:	:	:	:	:	:	:
:	:	:	:	:	:	:
:	:	:	:	:	:	:
98	138.5	140.5	139.0	139.3	139.33	2.0
99	140.5	140.6	140.1	141.4	140.65	1.3
100	140.1	140.3	139.7	139.4	139.88	0.9
					$\overline{\overline{x}} = 140.00$	$\overline{R} = 1.21$

(a) Calculate the control limits for the mean and range charts.

[5 marks]

(b) 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 (i), check for process control. In practice why are two "types" of chart required ?

[5 marks]

(c) Is the process capable? Give reasons.

[5 marks]

(d) Calculate the ARL i.e. average run length for a change of + 1 mm in the average. In words give a practical interpretation of the ARL.

[5 marks]

(e) What is a CUSUM chart ? What type of departures from the production target value is this type of chart useful for detecting ? Explain how it works.

[5 marks]

Q4 The quality of a certain pharmaceutical product is indicated by the percentage contamination of a by-product in the chemical synthesis. This by-product can be removed from the final batch, but only at considerable expense. It is thought that a cheaper way of improving quality would be to add an inhibitor, designed to stop the build-up of the by-product during the synthesis. Thirty two batches of the pharmaceutical were produced at different pH settings, with and without the inhibitor. The percentage contamination of the final product was determined. The data are as reported below.

pH	percentage contamination (no inhibitor used)	percentage contamination (inhibitor added)
6.00	2.22	2.39
6.00	2.26	2.44
6.25	2.24	2.34
6.25	2.24	2.34
6.50	2.24	2.27
6.50	2.20	2.26
6.75	2.24	2.21
6.75	2.19	2.21
7.00	2.21	2.17
7.00	2.13	2.11
7.25	2.15	2.06
7.25	2.15	2.05
7.50	2.14	2.02
7.50	2.10	2.01
7.75	2.14	1.98
7.75	2.08	1.93

(a) The variable "Dummy" = 0 if the inhibitor was not used. The variable "Dummy" = 1 if the inhibitor was used. Explain what you understand by a dummy variable, as used in multiple linear regression. Show how these types of variables can be use to account for different (i) intercepts, (ii) slopes, (iii) slopes and intercepts, for two groups of data in a regression context. Use the regression equations overleaf to illustrate your answer.

[10 marks]

(b) Using the various regression outputs overleaf comment on the effectiveness of inhibitor.

[10 marks]

(c) What are differences between multiple linear regression and ANOVA?

[5marks]

Regression 1 :

Dependent variable is: **Percentage**
No Selector
R squared = 75.3% R squared (adjusted) = 74.5%
s = 0.05855 with 32 - 2 = 30 degrees of freedom

Source	Sum of Squares	df	Mean Square	F-ratio
Regression	0.313717	1	0.313717	91.5
Residual	0.102835	30	0.00342785	

Variable	Coefficient	s.e. of Coeff	t-ratio	prob
Constant	3.36764	0.1246	27	≤ 0.0001
pH	-0.172852	0.01807	-9.57	≤ 0.0001

Regression 2 :

Dependent variable is: **Percentage**
No Selector
R squared = 75.5% R squared (adjusted) = 73.8%
s = 0.05933 with 32 - 3 = 29 degrees of freedom

Source	Sum of Squares	df	Mean Square	F-ratio
Regression	0.314482	2	0.157241	44.7
Residual	0.10207	29	0.00351965	

Variable	Coefficient	s.e. of Coeff	t-ratio	prob
Constant	3.37253	0.1267	26.6	≤ 0.0001
Dummy	-0.00978323	0.02098	-0.466	0.6444
pH	-0.172852	0.01831	-9.44	≤ 0.0001

Regression 3 :

Dependent variable is: **Percentage**
No Selector
R squared = 76.0% R squared (adjusted) = 74.3%
s = 0.05877 with 32 - 3 = 29 degrees of freedom

Source	Sum of Squares	df	Mean Square	F-ratio
Regression	0.316402	2	0.158201	45.8
Residual	0.10015	29	0.00345344	

Variable	Coefficient	s.e. of Coeff	t-ratio	prob
Constant	3.36764	0.1251	26.9	≤ 0.0001
pH	-0.171524	0.0182	-9.43	≤ 0.0001
pH*Dmy	-0.00265582	0.003012	-0.882	0.3851

Regression 4 :

Dependent variable is: **Percentage**
No Selector
R squared = 96.0% R squared (adjusted) = 95.5%
s = 0.02449 with 32 - 4 = 28 degrees of freedom

Source	Sum of Squares	df	Mean Square	F-ratio
Regression	0.399757	3	0.133252	222
Residual	0.0167946	28	0.000599808	

Variable	Coefficient	s.e. of Coeff	t-ratio	prob
Constant	2.75296	0.07374	37.3	≤ 0.0001
Dummy	1.22935	0.1043	11.8	≤ 0.0001
pH	-0.0827329	0.01069	-7.74	≤ 0.0001
pH*Dmy	-0.180238	0.01512	-11.9	≤ 0.0001

Q5 A supermarket buys a particular product from four suppliers, A, B, C, D, and regular tasting tests by expert panels are carried out as the product is sold in their food halls. Various characteristics are scored and an analysis of the totals of these scores is made. Four tasters a, b, c, d obtained these results at four sessions 1-4.

Taster		a	b	c	d
Session	1	A:21	B:17	C:18	D:20
	2	B:20	D:22	A:23	C:19
	3	C:20	A:24	D:22	B:19
	4	D:22	C:21	B:22	A:26

Score versus Taster, Session, Supplier

```
Factor     Type Levels Values
Taster    fixed      4 a b c d
Session   fixed      4 1 2 3 4
Supplier  fixed      4 A B C D

Analysis of Variance for Score, using Adjusted SS for Tests

Source     DF     Seq SS        MS       F      P
Taster      ?     0.5000    0.1667    0.20  0.893
Session     ?    28.5000    9.5000   11.40  0.007
Supplier    ?    44.0000   14.6667   17.60  0.002
Error       ?     ?         ?
Total       ?     ?
```

(a) In the context of the above example, distinguish between the treatment and the blocking variables involved. Give reasons.

[5 marks]

(b) The above data are an example of a particular experimental design. What is the general name given to this type of experimental design? Name one serious limitation of this type of experimental design.

[5 marks]

(c) Complete the ANOVA table substituting the symbols **?** with their correct values.

[5 marks]

(d) Interpret the results.

[5 marks]

(e) What is the key property of the experimental design above which allows factor effects to be estimated independently of one another. Show how this property presents itself in the above design.

[5 marks]

Q6. In an investigation into the extraction of nitrate-nitrogen from air dried soil, three quantitative variables were investigated at two levels. These were the amount of oxidised activated charcoal (A) added to the extracting solution to remove organic interferences, the strength of $CaSO_4$ extracting solution (C), and the time the soil was shaken with the solution (T). The aim of the investigation was to optimise the extraction procedure. The levels of the variables are given here:

		-	+
Activated charcoal (g)	A	0.5	1.0
$CaSO_4$ (%)	C	0.1	0.2
Time (minutes)	T	30	60

The concentrations of nitrate-nitrogen were determined by ultra-violet spectrophotometry and compared with concentrations determined by a standard technique. The results are given below and are the amounts recovered (expressed as the percentage of known nitrate concentration).

A	C	T	Amount
-1	-1	-1	45.1
1	-1	-1	44.9
-1	1	-1	44.8
1	1	-1	44.7
-1	-1	1	33.0
1	-1	1	53.8
-1	1	1	32.6
1	1	1	54.2
-1	-1	-1	44.6
1	-1	-1	45.3
-1	1	-1	46.7
1	1	-1	44.8
-1	-1	1	35.0
1	-1	1	51.7
-1	1	1	33.7
1	1	1	53.2
0	0	0	45.0
0	0	0	44.9
0	0	0	44.8

Computer output from MINITAB along with several important graphical summaries of the data are given on the following pages.

(a) Write a suitable response function for these data. Give reasons for any terms you include / exclude.

[5 marks]

(b) Interpret the coefficients of the estimated response function obtained in part (a). How can this function be used to identify the path of steepest ascent. No calculations are necessary.

[5 marks]

(c) Comment on all the graphical displays provided. For each identify what is being plotted, what assumption is being tested, and how the plot should be interpreted in the context of the analysis.

[10 marks]

(d) Explain what advantages a fractional factorial design can confer over a full factorial design. At what stage in a planned approach to experimental design would such designs be used?

[5 marks]

Estimated Effects and Coefficients for Amount

Term	Effect	Coef	StDev Coef	T	P
Constant		44.3579	0.2168	204.59	0.000
A	9.6375	4.8188	0.2363	20.40	0.000
C	0.1625	0.0813	0.2363	0.34	0.737
T	-1.7125	-0.8562	0.2363	-3.62	0.003
A*C	0.1375	0.0688	0.2363	0.29	0.776
A*T	10.0125	5.0063	0.2363	21.19	0.000
C*T	-0.1125	-0.0562	0.2363	-0.24	0.816

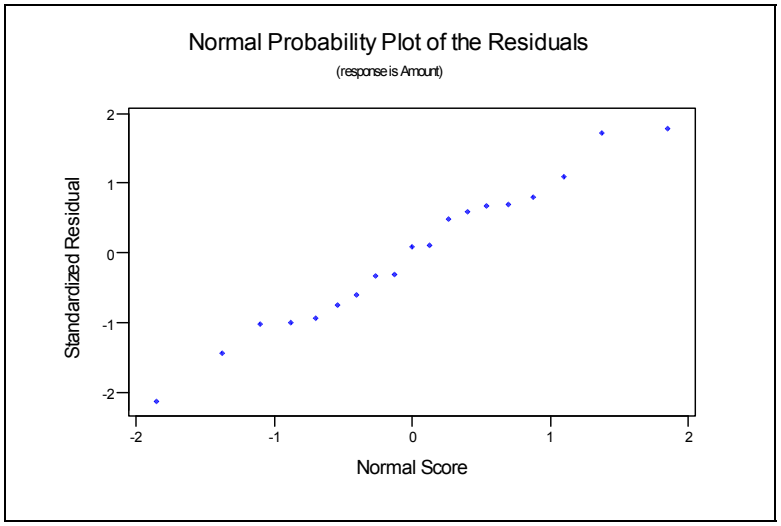
Analysis of Variance for Amount

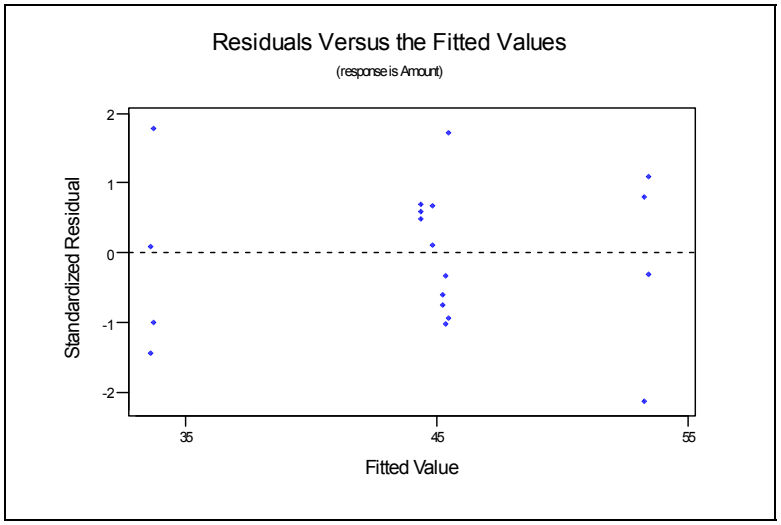
Source	DF	Seq SS	Adj MS	F	P
Main Effects	3	383.362	127.787	143.08	0.000
2-Way Interactions	3	401.127	133.709	149.71	0.000
Residual Error	12	10.718	0.893		
Curvature	1	1.047	1.047	1.19	0.298
Lack of Fit	1	2.326	2.326	3.17	0.106
Pure Error	10	7.345	0.735		
Total	18	795.206			

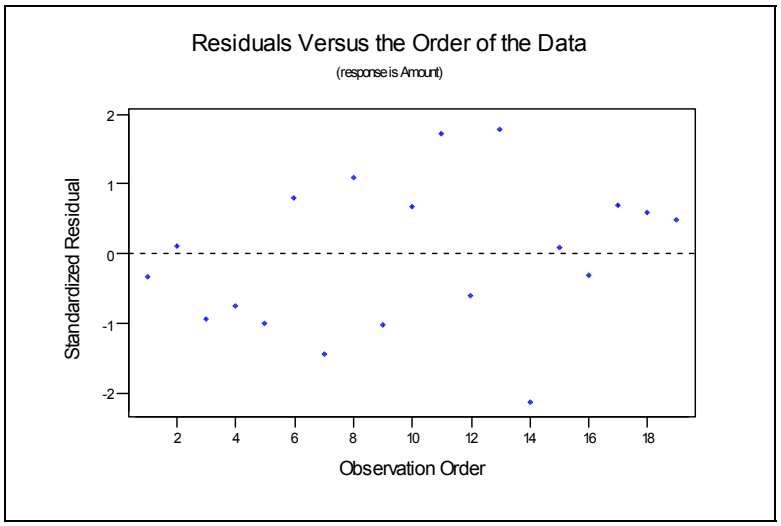
Unusual Observations for Amount

Obs	Amount	Fit	StDev Fit	Residual	St Resid
14	51.7000	53.2329	0.6180	-1.5329	-2.14R

R denotes an observation with a large standardized residual


Normal Probability Plot of the Residuals
(response is Amount)
Standardized Residual
Normal Score
2
1
0
-1
-2
-2
-1
0
1
2


Residuals Versus the Fitted Values
(response is Amount)
Standardized Residual
Fitted Value
2
1
0
-1
-2
35
45
55


Residuals Versus the Order of the Data
(response is Amount)
Standardized Residual
Observation Order
2
1
0
-1
-2
2
4
6
8
10
12
14
16
18

Plots for question 1. Detach and hand up with answer book.

Student Name :	
Student ID number :