FACULTY OF SCIENCE AND ENGINEERING

DEPARTMENT OF MATHEMATICS AND STATISTICS

END OF SEMESTER ASSESSMENT

MODULE CODE: MA4605	SEMESTER: Autumn
MODULE TITLE: Chemometrics	DURATION OF EXAM: 2.5 hours
LECTURER: Mr. Kevin O'Brien	GRADING SCHEME: 100 marks 70% of module grade

INSTRUCTIONS TO CANDIDATES

Scientific calculators approved by the University of Limerick can be used.
Formula sheet and statistical tables provided at the end of the exam paper.
Students must attempt any 4 questions from 5.

Question 1. (20 marks) Inference Procedures

(a) (20 Marks) In a clinical trial, volunteer patients were randomly assigned to either the control group or the experimental group. The clinical trial team evaluated both groups to determine whether or not the groups were similar in terms of measures of centrality and dispersion, before the trial commenced.

The following blocks of `R` code (i.e blocks A to F) are based on the data for this assessment.

Write a short report on your conclusion for this assessment, clearly indicating which blocks of `R` code you felt were most relevant, and explain why.

Block A

```
> var.test(EXP,CONT)
        F test to compare two variances
data:  EXP and CONT
F = 1.1669, num df = 11, denom df = 13, p-value = 0.7813
alternative hypothesis: true variance ratio is not 1
95 percent confidence interval:
 0.3649431 3.9578211
sample estimates:   ratio of variances
                              1.166904
```

Block B

```
> shapiro.test(EXP)

        Shapiro-Wilk normality test
data:  EXP
W = 0.9292, p-value = 0.372
>
> shapiro.test(CONT)

        Shapiro-Wilk normality test
data:  CONT
W = 0.9001, p-value = 0.113
```

Block C

```
> t.test(EXP,CONT)

        Welch Two Sample t-test
data:  EXP and CONT
t = -0.5234, df = 22.722, p-value = 0.6058
alternative hypothesis: true difference in means is not 0
95 percent confidence interval:
 -11.562458   6.895791
sample estimates:       mean of x mean of y
                        111.6667  114.0000
```

Block D

```
> t.test(EXP,CONT,var.equal=TRUE)

        Two Sample t-test
data:  EXP and CONT
t = -0.5266, df = 24, p-value = 0.6033
alternative hypothesis: true difference in means is not 0
95 percent confidence interval:
 -11.478353   6.811686
sample estimates:   mean of x mean of y
                    111.6667  114.0000
```

Block E

```
> ks.test(EXP,CONT)

        Two-sample Kolmogorov-Smirnov test
data:  EXP and CONT
D = 0.3571, p-value = 0.382
alternative hypothesis: two-sided
...
```

Block F

```
> wilcox.test(EXP,CONT)

        Wilcoxon rank sum test
data:  EXP and CONT
W = 69.5, p-value = 0.4705
alternative hypothesis: true location shift is not  0
...
```

Question 2. (20 marks) Linear Models

(a) The fluorescence of each of a series of acidic solutions of quinine with concentrations 0,10,20,30,40,50 was determined five times. The mean values and standard deviations of these determinations have been obtained as follows:

Means:	4.0	21.2	44.6	61.8	78.0	105.2
Std Deviations:	0.71	0.84	0.89	1.64	2.24	3.03

Two models have been fitted to the data. These models are described by the following R code output.

Model 1

```
lm(formula = Means ~ Conc)

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.9238 2.1648 1.351 0.248
Conc 1.9817 0.0715 27.715 1.01e-05 ***
---
Residual standard error: 2.991 on 4 degrees of freedom
Multiple R-squared: 0.9948,Adjusted R-squared: 0.9935
F-statistic: 768.1 on 1 and 4 DF, p-value: 1.008e-05
```

Model 2

```
weights=SdInt^(-2)/mean(SdInt^(-2))

lm(formula = Means ~ Conc, weights = weights)
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.48066 1.15736 3.007 0.0397 *
Conc 1.96315 0.06765 29.018 8.4e-06 ***
---
Residual standard error: 2.034 on 4 degrees of freedom
Multiple R-squared: 0.9953,Adjusted R-squared: 0.9941
F-statistic: 842 on 1 and 4 DF, p-value: 8.396e-06
```

i. (4 Marks) What kind of analyses have been performed in each of model 1 and model 2? Write down the linear models fitted by each of the two analyses.

ii. (1 Marks) Explain the term heteroscedascity.

iii. (3 Marks) Describe differences between the two models, making reference to the scatter-plot of the data on the next page. (Also present on the scatter-plot is a regression line fitted using the first analysis).

iiv. (2 Marks) Based on the R code output, which model is the better fit?

.

(b) An ion-selective electrode (ISE) determination of sulphide from sulphate-reducing bacteria was compared with a gravimetric determination. Each pair of determinations were taken from the same sample.

The results obtained by both methods are expressed in milligrams of sulphide, and are tabulated below.

ISE method	108	12	152	3	106	11	128	12	160	128
gravimetry	105	16	113	1	108	11	141	161	182	118

Two simple linear models are fitted to the data. Model C uses the gravimetric determination as an independent variable used to predict the ISE determination. Conversely, Model D uses the ISE determination as an independent variable used to predict the gravimetric determination. The relevant `R` output is presented on the following page.

Model C

```
Call:
lm(formula = ISE ~ grav)
...
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  15.1125    28.8487   0.524    0.615
grav          0.6997     0.2543   2.751    0.025 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
....
```

Model D

```
Call:
lm(formula = grav ~ ISE)
..
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  38.6215    25.8542   1.494    0.174
ISE           0.6949     0.2526   2.751    0.025 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
....
```

i. (4 marks) Write the regression equation for both of the fitted models.

ii. (3 marks) Is a simple linear regression model an suitable approach for this type of analysis? Explain why or why not? What alternatives might you recommend?

iii. (3 marks) Discuss an alternative approach for this analysis, mentioning any disadvantages in using this alternative approach.

Question 3. (20 marks) Statistical Process Control

(a) Answer the following questions.

i (1 marks) Differentiate common causes of variation in the quality of process output from assignable causes.

ii. (1 marks) What is tampering in the context of statistical process control?

iii (4 marks) Other than applying the *Three Sigma* rule for detecting the presence of an assignable cause, what else do we look for when studying a control chart? Support your answer with sketches.

(b) A normally distributed quality characteristic is monitored through the use of control charts. These charts have the following parameters. All charts are in control.

	LCL	Centre Line	UCL
$\bar{X}$-Chart	542	550	558
R-Chart	0	8.236	16.504

i (2 marks) What sample size is being used for this analysis?

ii. (2 marks) Estimate the standard deviation of this process.

iii. (2 marks) Compute the control limits for the process standard deviation chart (i.e. the s-chart).

(c) An automobile assembly plant concerned about quality improvement measured sets of five camshafts on twenty occasions throughout the day. The specifications for the process state that the design specification limits at 600±3mm.

i. (4 marks) Determine the *Process Capability Indices* C_p and C_{pk}, commenting on the respective values. You may use the `R` code output on the following page.

ii. (2 marks) The value of C_{pm} is 1.353. Explain why there would be a discrepancy between C_p and C_{pm}.

iii. (2 marks) Comment on the graphical output of the *Process Capability Analysis*, also presented on the next page.

```
Process Capability Analysis

Call:
process.capability(object = obj, spec.limits = c(597, 603))
Number of obs = 100          Target = 600
       Center = 599.548         LSL = 597
       StdDev = 0.5846948       USL = 603

Capability indices:
      Value   2.5%  97.5%
Cp    ...
Cp_l  ...
Cp_u  ...
Cp_k  ...
Cpm   1.353  1.134  1.572
Exp<LSL 0%   Obs<LSL 0%
```

Question 4. (20 marks) Experimental Design Part 1

(a) Explain the following terms in the context of experimental design

i. (2 marks) levels of a factor.

ii. (2 marks) randomized block design.

(b) Six analysts each made seven determinations of the paracetamol content of the same batch of tablets. The results are shown below. There are 42 determinations in total. The mean determination for each analysts is also tabulated.

Analyst	Content						
A	84.32	84.61	84.64	84.62	84.51	84.63	84.51
B	84.24	84.13	84.00	84.02	84.25	84.41	84.30
C	84.29	84.28	84.40	84.63	84.40	84.68	84.36
D	84.14	84.48	84.27	84.22	84.22	84.02	84.33
E	84.50	83.91	84.11	83.99	83.88	84.49	84.06
F	84.70	84.36	84.61	84.15	84.17	84.11	83.81

The following `R` output has been produced as a result of analysis of these data:

Response: Y	Df	Sum Sq	Mean Sq	F value	$Pr(>F)$
Analyst	?	?	?	?	0.00394 **
Residuals	?	?	0.04065		
Total	?	2.3246			

i. (5 marks) Complete the ANOVA table in your answer sheet, replacing the "?" entries with the correct values.

ii. (2 marks) What hypothesis is being considered by this procedure.

iii. (2 marks) What is the conclusion following from the above analysis? State the null and alternative hypothesis clearly.

(c) The R code and graphical procedures, below and on the following page, are relevant to checking whether the underlying assumptions are met for the ANOVA model in part (b).

i. (3 marks) What are the assumptions underlying ANOVA?

ii. (4 marks) Assess the validity of these assumptions for the ANOVA model in part(b).

```
        Shapiro-Wilk normality test

data:  Residuals
W = 0.9719, p-value = 0.3819
```

```
        Bartlett test of homogeneity of variances

data:  Experiment
Bartlett's K-squared = 105.9585, df = 1, p-value < 2.2e-16
```

Question 5. (20 marks) Experimental Design Part 2

(a) An experiment is run on an operating chemical process in which the aim is to reduce the amount of impurity produced. Three continuous variables are thought to affect impurity, these are concentration of NaOH, agitation speed and temperature. As an initial investigation two settings are selected for each variable these are

Factor:	low level	highlevel
Concentration of NaOH	40%	45%
Agitation speed (rpm)	15	25
Temperature (°F)	170	200

Readings were recorded of the impurity produced from the chemical process for each combination of the levels of these factors, and each combination was tested twice.

Conc NaOH	Agitation	Temperature	Impurity
-	-	-	39,34
+	-	-	40,47
-	+	-	23,34
+	+	-	25,36
-	-	+	75,89
+	-	+	61,75
-	+	+	59,43
+	+	+	21,20

Control Limits for Control Charts

$$\bar{\bar{x}} \pm 3\frac{\bar{s}}{c_4\sqrt{n}}$$

$$\bar{s} \pm 3\frac{c_5\bar{s}}{c_4}$$

$$\left[\bar{R}D_3, \bar{R}D_4\right]$$

Process Capability Indices

$$\hat{C}_p = \frac{\text{USL} - \text{LSL}}{6s}$$

$$\hat{C}_{pk} = \min\left[\frac{\text{USL} - \bar{x}}{3s}, \frac{\bar{x} - \text{LSL}}{3s}\right]$$

$$\hat{C}_{pm} = \frac{\text{USL} - \text{LSL}}{6\sqrt{s^2 + (\bar{x} - T)^2}}$$

2^3 Design: Interaction Effects

$$AB = \frac{1}{4n}\left[abc - bc + ab - b - ac + c - a + (1)\right]$$

$$AC = \frac{1}{4n}\left[(1) - a + b - ab - c + ac - bc + abc\right]$$

$$BC = \frac{1}{4n}\left[(1) + a - b - ab - c - ac + bc + abc\right]$$

$$ABC = \frac{1}{4n}\left[abc - bc - ac + c - ab + b + a - (1)\right]$$

Factorial Design: Sums of Squares

$$\text{Effect} = \frac{(\text{Contrast})}{n2^{k-1}}$$

$$\text{Sums of Squares} = \frac{(\text{Contrast})^2}{n2^k}$$

Factors for Control Charts

Sample Size (n)	c4	c5	d2	d3	D3	D4
2	0.7979	0.6028	1.128	0.853	0	3.267
3	0.8862	0.4633	1.693	0.888	0	2.574
4	0.9213	0.3889	2.059	0.88	0	2.282
5	0.9400	0.3412	2.326	0.864	0	2.114
6	0.9515	0.3076	2.534	0.848	0	2.004
7	0.9594	0.282	2.704	0.833	0.076	1.924
8	0.9650	0.2622	2.847	0.82	0.136	1.864
9	0.9693	0.2459	2.970	0.808	0.184	1.816
10	0.9727	0.2321	3.078	0.797	0.223	1.777
11	0.9754	0.2204	3.173	0.787	0.256	1.744
12	0.9776	0.2105	3.258	0.778	0.283	1.717
13	0.9794	0.2019	3.336	0.770	0.307	1.693
14	0.9810	0.1940	3.407	0.763	0.328	1.672
15	0.9823	0.1873	3.472	0.756	0.347	1.653
16	0.9835	0.1809	3.532	0.750	0.363	1.637
17	0.9845	0.1754	3.588	0.744	0.378	1.622
18	0.9854	0.1703	3.64	0.739	0.391	1.608
19	0.9862	0.1656	3.689	0.734	0.403	1.597
20	0.9869	0.1613	3.735	0.729	0.415	1.585
21	0.9876	0.1570	3.778	0.724	0.425	1.575
22	0.9882	0.1532	3.819	0.720	0.434	1.566
23	0.9887	0.1499	3.858	0.716	0.443	1.557
24	0.9892	0.1466	3.895	0.712	0.451	1.548
25	0.9896	0.1438	3.931	0.708	0.459	1.541