FACULTY OF SCIENCE AND ENGINEERING

DEPARTMENT OF MATHEMATICS AND STATISTICS

END OF SEMESTER EXAMINATION PAPER 2015

MODULE CODE: MA4605 SEMESTER: Autumn 2015

MODULE TITLE: Chemometrics DURATION OF EXAM: 2.5 hours

LECTURER: Mr. Kevin O'Brien GRADING SCHEME: 100 marks
70% of module grade

ASSESSORS: Dr. C.F. Ryback

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.
There are 5 questions in this exam. Students must attempt any 4 questions.

Question 1 Inference Procedures

What is going here?

- Using the Murdoch Barnes Table for Normal Distribution Problems
- Testing that Data is normally distributed (may appear elsewhere)
- Transformation of Data (Tukey's Ladder)
- Outliers and Boxplots (Grubbs Test, Dixon Q-test)
- Non-Parametric Procedures (e.g. Wilcoxon test, Kolmogorov Smirnov Test)

Part A Dixon Q Test For Outliers (4 Marks)

The typing speeds for one group of 12 Engineering students were recorded both at the beginning of year 1 of their studies. The results (in words per minute) are given below:

118	146	149	142	170	153
137	161	156	165	178	159

Use the Dixon Q-test to determine if the lowest value (118) is an outlier. You may assume a significance level of 5%.

i. (1 Mark) State the Null and Alternative Hypothesis for this test.

ii. (1 Marks) Compute the test statistic

iii. (1 Mark) State the appropriate critical value.

iv. (1 Mark) What is your conclusion to this procedure

Part B Normal Distribution (3 Marks)

Assume that the diameter of a critical component is normally distributed with a Mean of 250 mm and a Standard Deviation of 15 mm. You are required to estimate the approximate probability of the following measurements occurring on an individual component.

i. (1 Mark) Greater than 104.1mm

ii. (2 Marks) Less than 95.2 mm

iii. (2 Marks)[$*$] Between 94.2 and 103 mm

Use the normal tables to determine the probabilities for the above exercises. You are required to show all of your workings.

Q
44 36 56 38 63 89 58 37 41 54 71 24 51 49

Question 2 Chi-Squared and One-Way ANOVA F-test

- State the Null and Alternative Hypothesis

A market research survey was carried out to assess preferences for three brands of chocolate bar, A, B, and C. The study group was categorised by gender to determine any difference in preferences.

	X	Y	Z	Total
Children	50	70	80	200
Teenagers	90	50	20	160
Adults	140	120	100	360

i. (1 Mark)Formally state the null and alternative hypotheses.

ii. (2 Marks) Compute the cell values expected under the null hypothesis. Show your workings for two cells.

iii. (3 Marks) Compute the Test Statistic.

iv. (1 Mark) State the appropriate Critical Value for this hypothesis test.

v. (1 Mark) Discuss your conclusion to this test, supporting your statement with reference to appropriate values.

Question 3 Linear Models (25 Marks)

State the regression equation for this model. Use this equation to predict a value for y when the predictor variables take the following values.

Test for Correlation

```
cor.test(X,Y)
```

Multiple Linear Regression

Given the AIC for each candidate model, use ***Backward Selection*** to determine the optimal model for predicting values of y with predictor variables x_1, x_2,x_3 and x_4.

Suppose there were 10 possible predictor variables. How many ways are there to fit a model.

Combinatorial Explosion

Multicollinearity -

Model Selection

- Suppose we have 5 predictor variables.
- Use **Forward Selection** and **Backward Selection** to choose the optimal set of predictor variables, based on the AIC measure.

Variables	AIC	Variables	AIC
$\emptyset$	200	x1, x2, x3	74
		x1, x2, x4	75
x1	150	x1, x2, x5	79
x2	145	x1, x3, x4	72
x3	135	x1, x3, x5	85
x4	136	x1, x4, x5	95
x5	139	x2, x3, x4	83
		x2, x3, x5	82
x1, x2	97	x2, x4, x5	78
x1, x3	81	x3, x4, x5	85
x1, x4	94		
x1, x5	88	x1, x2, x3, x4	93
x2, x3	87	x1, x2, x3, x5	120
x2, x4	108	x1, x2, x4, x5	104
x2, x5	87	x1, x3, x4, x5	101
x3, x4	105	x2, x3, x4, x5	89
x3, x5	82		
x4, x5	86	x1, x2, x3, x4, x5	100

Regression ANOVA

Suppose we have a regression model, described by the following equation

$$\hat{y} = 28.81 + 6.45x_1 + 7.82x_2$$

We are given the following pieces of information.

- The standard deviation of the response variance y is 10 units.
- There are 53 observations.
- The *Coefficient of Determination* (also known as the *Multiple R-Squared*) is 0.75.

Complete the *Analysis of Variance* Table for a linear regression model. The required values are indicated by question marks.

	DF	Sum Sq	Mean Sq	F value	Pr(>F)
Regression	?	?	?	?	$< 2.2e^{-16}$
Error	?	?	?		
Total	?	?			

Question 4 (25 Marks)

ANOVA

Three species of tree were grown in a forestry plantation. Not all the seedlings survived and so the sample size, n_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	n_i	Observations	Total	S_x^2
Pinus Caribea	10	4.9 5.1 4.5 5.0 4.1 4.0 4.4 4.8 3.8 3.4	44.0	0.32
Pinus Kesiya	12	4.2 3.5 4.7 4.1 3.9 4.6 4.3 3.4 4.0 3.3 3.6 4.4	48.0	0.22
Eucalyptus Deglupta	8	5.6 4.6 5.7 6.3 5.4 5.0 5.1 4.7	42.4	0.32

- The overall mean is 4.48
- The overall variance is 0.543
- Carry out the usual one-way analysis of variance to examine whether there are overall differences between the species.

Source	DF	SS	MS	F	$p-$values
Between	?	?	?	?	7.69×10^{-05}
Within	?	?	?		
Total	?	?			

Question 5. (25 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%
```

Control Charts

SPC
    Control Charts

Testing for Normality

Shapiro-Wilk Test
    Tukey ladder
    D'Agostino Test (Multivariate Normality)
    Process Capability Indices

Critical Values for the F distribution

- The significance level is 5%.
- The first degree of freedom ν_1 is arranged along the columns
- The second degree of freedom ν_2 is arranged along the rows.

Critical Values for Dixon Q Test

N	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$
3	0.941	0.97	0.994
4	0.765	0.829	0.926
5	0.642	0.71	0.821
6	0.56	0.625	0.74
7	0.507	0.568	0.68
8	0.468	0.526	0.634
9	0.437	0.493	0.598
10	0.412	0.466	0.568
11	0.392	0.444	0.542
12	0.376	0.426	0.522
13	0.361	0.41	0.503
14	0.349	0.396	0.488
15	0.338	0.384	0.475
16	0.329	0.374	0.463

Critical Values for Chi Square Test

n	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.001$
1	2.705	3.841	6.634	10.827
2	4.605	5.991	7.378	9.21
3	6.251	7.815	9.348	11.345
4	7.779	9.488	11.143	13.277
5	9.236	11.07	12.833	15.086
6	10.645	12.592	14.449	16.812
7	12.017	14.067	16.013	18.475
8	13.362	15.507	17.535	20.09
9	14.684	16.919	19.023	21.666
10	15.987	18.307	20.483	23.209

Process Capability Indices

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

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

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

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