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: 60 marks

ASSESSORS: Dr. M.F. Brezzo C.F. Ryback

60% 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 all 4 questions

Question 1. Inference Procedures - (Variant A)

Part A : Inference Procedures

The reproducibility of a method for the determination of a pollutant in water was investigated by taking twelve samples from a single batch of water and determining the concentration of pollutant in each. The following results were obtained:

$$5.98, 8.80, 6.89, 8.49, 8.48, 7.47, 7.97, 6.94, 7.32, 6.64, 6.98, 7.94.$$

It is expected that from this sample a 95% confidence interval for the concentration of pollutant will be obtained.

(i) There is a concern that the data may contain an outlier. Thus the following procedure has been performed on the data (which is simply referred to a X in the code output):

```
Dixon test for outliers
data: X
Q = 0.3625, p-value = 0.6326
alternative hypothesis: lowest value 5.98 is an outlier
```

- (2 Marks) Describe what is the purpose of this procedure.
- (1 Mark) Write the conclusion that follows from it.

(ii) After the test for an outlier, another preliminary procedure has been performed on the data

```
Anderson-Darling normality test

data:  X
A = 0.24067, p-value = 0.7132
```

- (2pts) Explain what is the name and purpose of this procedure. State the null and alternative hypothesis used for this procedure.
- (1pts) What is the conclusion and why?

(iii) After these initial verifications, the confidence interval can be obtained from the following computations Based on the obtained values write down the confidence interval for the pollutant.

```
t = 1.9868, df = 11, p-value = 0.07242
alternative hypothesis: true mean is not equal to 7
95 percent confidence interval:
6.946999 8.036334
sample estimates:
mean of x
7.491667
```

Question 1. Inference Procedures - (Variant B)

Part A : Inference Procedures

- The nicotine content in blood can be determined by gas chromatography down to concentrations of 1 ng/ml. The concentration of nicotine was determined in each of two samples of known concentrations 10 ng/ml and 50 ng/ml.

```
Data: Sample (Lo): m = 10 ng/ml, n=14.

 8.40, 9.59, 9.38, 9.10, 10.78, 11.41, 9.94,
10.08, 12.11, 9.10, 9.59, 10.36, 10.41, 10.52.

Data: Sample (Hi): m = 50 ng/ml, n=10.

47.5, 48.4, 48.8, 48.4, 46.8,
46.2, 48.6, 50.6, 45.5, 46.1.
```

A research team evaluated both samples to determine whether or not the samples were similar in terms of measures of centrality and dispersion, before the trial commenced.

The following blocks of `R` code (i.e blocks 1 to 6) are based on the data for this assessment.

(a) (3 Marks) Each of the six blocks of code describes a statistical inference procedure. Provide a brief description for each procedure.

(b) (4 Marks) 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 1

```
            F test to compare two variances

data:  Lo and Hi
F = 0.3945, num df = 13, denom df = 9, p-value = 0.1246
alternative hypothesis:
 true ratio of variances is not equal to 1
95 percent confidence interval:
 0.1029905 1.3066461
sample estimates:
ratio of variances
         0.3945149
```

Block 2

```
> shapiro.test(Lo)

        Shapiro-Wilk normality test

data:  Lo
W = 0.9779, p-value = 0.9609
> shapiro.test(Hi)

        Shapiro-Wilk normality test

data:  Hi
W = 0.9496, p-value = 0.6634
```

Block 3

```
> t.test(Lo,Hi)

        Welch Two Sample t-test

data:  Lo and Hi
t = -67.374, df = 14.016, p-value < 2.2e-16
alternative hypothesis:
   true difference in means is not equal to 0
95 percent confidence interval:
 -38.83294 -36.43706
sample estimates:
mean of x mean of y
   10.055    47.690
```

Block 4 `> t.test(Lo,Hi,var.equal=TRUE)`

```
        Two Sample t-test

data:  Lo and Hi
t = -72.6977, df = 22, p-value < 2.2e-16
alternative hypothesis:
   true difference in means is not equal to 0
95 percent confidence interval:
 -38.70863 -36.56137
sample estimates:
mean of x mean of y
   10.055    47.690
```

Block 5 `> ks.test(Lo,Hi)`

```
        Two-sample Kolmogorov-Smirnov test

data:  Lo and Hi
D = 1, p-value = 1.02e-06
alternative hypothesis: two-sided
```

Block 6 `wilcox.test(Lo,Hi)`

```
        Wilcoxon rank sum test


data:  Lo and Hi
W = 0, p-value = 1.02e-06
alternative hypothesis:
true location shift is not equal to 0
```

Question 1. Inference Procedures - (Variant C)

Part A : Outliers

- (3 Marks) Provide a brief description for three tests from the family of Grubb's Outliers Tests. Include in your description a statement of the null and alternative hypothesis for each test, any required assumptions and the limitations of these tests.
- (3 Marks) Showing your working, use the Dixon Q Test to test the hypothesis that the maximum value of the following data set is an outlier.

$$19, 22, 23, 24, 25, 26, 29, 38$$

Testing for Outliers (3 Marks)

The following statistical procedure is based on this dataset.

6.98	8.49	7.97	6.64
8.80	8.48	5.94	6.94
6.89	7.47	7.32	4.01

```
> grubbs.test(x, two.sided=T)

Grubbs test for one outlier

data:  x
G = 2.4093, U = 0.4243, p-value = 0.05069
alternative hypothesis: lowest value 4.01 is an outlier
```

i. (1 Mark) Describe what is the purpose of this procedure. State the null and alternative hypothesis.

ii. (1 Mark) Write the conclusion that follows from it.

iii. (1 Mark) State any relevant assumptions for this procedure.

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:

121	146	150	149	142	170	153
137	161	156	165	137	178	159

Use the Dixon Q-test to determine if the lowest value (121) is an outlier. You may assume a significance level of 5%.

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

ii. (1 Mark) Compute the Test Statistic.

iii. (2 Mark) By comparing the Test Statistic to the appropriate Critical Value, state your conclusion for this test.

Question 2. Linear Models - (Variant 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

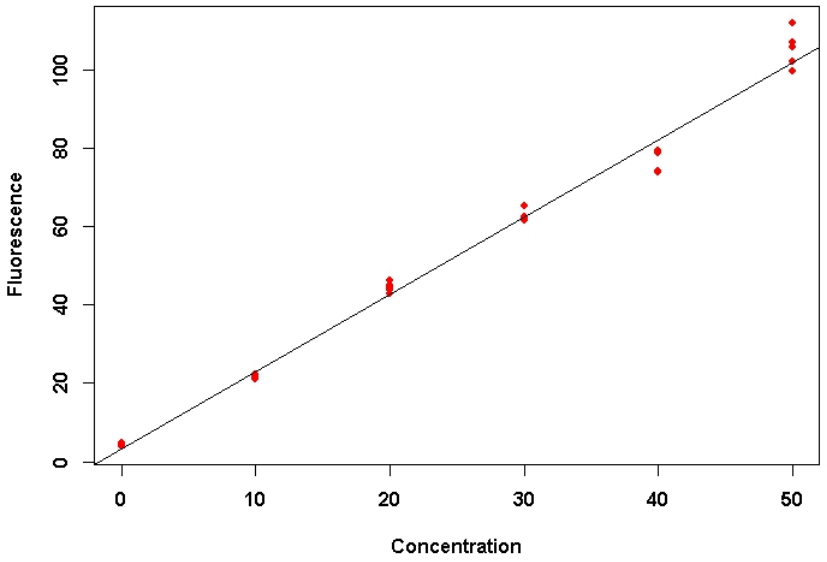
```
weights=SdInt^(-2)/mean(SdInt^(-2))

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 model regression equation fitted by each of the two analyses.

ii. (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).

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



.

Question 2. Linear Models - (Variant B)

(4 Marks) Complete the following *Analysis of Variance* Table for a simple linear regression model based on the data provided. The required values are indicated by question marks.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Regression	?	9160239	?	?	$< 2.2e^{-16}$
Error	50	2134710	?		
Total	?	?	42694		

Once you have completed this table, compute the following

- (1 Mark) The Pearson correlation coefficient for the response variable Y and the predictor variable X.
- (1 Mark) Tthe sample standard deviation of the response variable Y.

Question 2. Linear Models - (Variant C)

The following results were obtained when each of a series of standard silver solutions was analysed by ame atomic-absorption spectrometry. The analysis of these by means of R is also presented below.

Concentration, ng/ml	Absorbance
10	0.251
15	0.390
20	0.498
25	0.625
30	0.763
0	0.003
5	0.127

```
Conc=c(10,15,20,25,30,0,5)
Abs=c(0.251,0.390,0.498,0.625,0.763,0.003,0.127)
Call:
lm(formula = Abs ~ Conc)
Residuals:
          1           2           3           4           5           6           7
-0.0027500  0.0104286 -0.0073929 -0.0052143  0.0059643  0.0008929 -0.0009286

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.0021071  0.0047874    0.44    0.678
Conc        0.0251643  0.0002656   94.76 2.48e-09 ***
---
Residual standard error: 0.007026 on 5 degrees of freedom
Multiple R-squared: 0.9994,Adjusted R-squared: 0.9993
F-statistic:  8980 on 1 and 5 DF,  p-value: 2.481e-09
```

Based on this information do the following

i. (3pts) Determine the slope and intercept of the calibration plot and make a sketch of the linear fit to the date. Include data points on the graph as well.

ii. (4pts) Based on the above calibration fit find the silver concentration for a sample giving an absorbance of 0.42 in a single determination. Estimate the 95% confidence limits for the silver concentration.

The following formula for the standard deviation of the concentration determination should be used for the purpose

$$s_{x_0} = \frac{s_{y/x}}{b}\sqrt{\frac{1}{m}+\frac{1}{n}+\frac{(y_0-\bar{y})^2}{b^2\sum_{i=1}^{n}(x_i-\bar{x})^2}}.$$

The following computations in `R` allows for effective computation of the above standard deviation

```
> mean(Abs)
[1] 0.3795714
> mean(Conc)
[1] 15
> sum((Conc-mean(Conc))^2)
[1] 700
```

- (6pts) Find the silver concentration for a sample giving absorbance values of 0.30, 0.31, 0.29 in three separate analyses of the same sample. Estimate the confidence limits for the concentration in this case. How the accuracy improved comparing to the one in the previous question? (In your computations you should use the information from the previous question.)

- (5pts) Estimate the limit of detection of the silver analysis.

Robust Regression

In certain circumstances, Robust Regression may be used in preference to Ordinary Least Squares Regression.

Answer the following questions relating to Robust Regression.

(i) (1 Mark) Describe what these circumstances might be.

(ii) (1 Mark) State one difference between OLS and Robust regression techniques, in terms of computing regression equations.

(iii) (2 Marks) Explain the process of Huber Weighting, stating the algorithm used to compute weightings.

(iv) (2 Marks) Suppose that Huber Weigthing, with a tuning constanct of $k = 13.45$ was applied to the observations tabulated below. What would be the outcome of the procedure for each case.

Observation i	Residual e_i
11	-9.07
14	14.54
18	22.91

Method Comparison

(i) (1 Mark) Write a brief note on the topic of method comparison studies.

(ii) (1 Mark)Why are OLS regression models not suitable for Method Comparison.

(iii) (1 Mark) Describe an alternative regression technique. Include in your answer any variants of the technique and any limitations of using those technique.

(iv) (2 Marks) A Bland Altman Plot is a graphical technique used in Method Comparison. Sketch a Bland-Altman plot and discuss how the various components are calculated.

Question 3. ANOVA and Experimental Design (Variant A)

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

i. (1 Mark) levels of a factor.

ii. (1 Mark) randomized block design.

(b) Six analysts each made seven determinations of the paracetamol content of the same batch of tablets. The results are shown in the table below. In the last two columns are the sample means and standard deviations for each sample.

Group								$\bar{X}$	S_X
A	84.32	84.51	84.63	84.61	84.64	84.51	84.62	84.5486	
B	84.24	84.25	84.41	84.13	84.00	84.30	84.02	84.1928	
C	84.29	84.40	84.68	84.28	84.40	84.36	84.63	84.4342	
D	84.14	84.22	84.02	84.48	84.27	84.33	84.22	84.2400	
E	84.50	83.88	84.49	83.91	84.11	84.06	83.99	84.1343	
F	84.70	84.17	84.11	84.36	84.61	83.81	84.15	84.2729	

For the aggregrate sample (all 42 observations) the standard deviation is 0.2381.

(i) (5 Marks) Complete the following One Way Analysis of Variance Table.

(ii) (1 Marks) Describe what is the purpose of this procedure. include a statement of the null and alternative hypothesis in your answer.

(iii) (2 Marks) (5 Marks)

Source	DF	Sum Squares	Mean Square	F	p-value
Between-Groups					0.003941
Within-Groups					
Total					

(c) Complete the following ANOVA table:

Source	DF	SS	MS	F
A	1		3088	
B		3400		
AB	3	49000		
Error				
Total	23	63000		

How many replicates were used?

Question 3. ANOVA and Experimental Design (Variant B)

(a) 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 CaSO4 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
CaSO4 (%)	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	Amounts	(2 Replicates)
-1	-1	-1	45.1	44.6
1	-1	-1	44.9	45.3
-1	1	-1	44.8	46.7
1	1	-1	44.7	44.8
-1	-1	1	33	35
1	-1	1	53.8	51.7
-1	1	1	32.6	33.7
1	1	1	54.2	53.2

i. (6 Marks) Calculate the contrasts, the effects and the sum of squares for the effects.

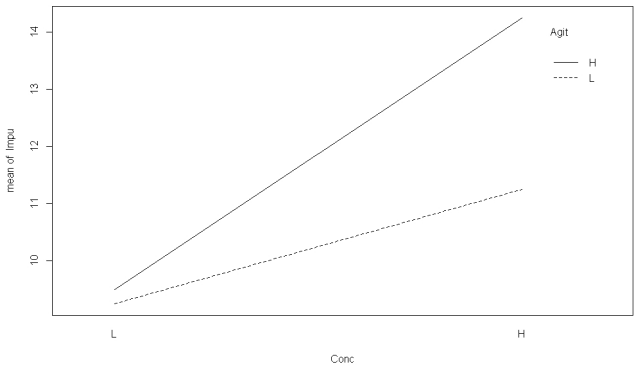
ii. (4 Marks) Using the computed sums of squares values, complete the ANOVA table (see the `R` code below).

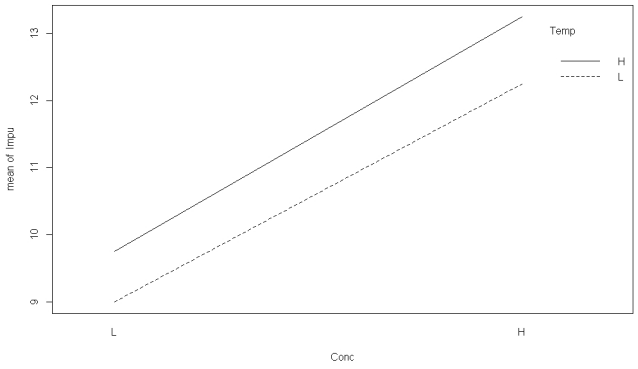
iii. (2 Marks) Comment on the tests for significant for the main effects and interactions. State clearly your conclusions.

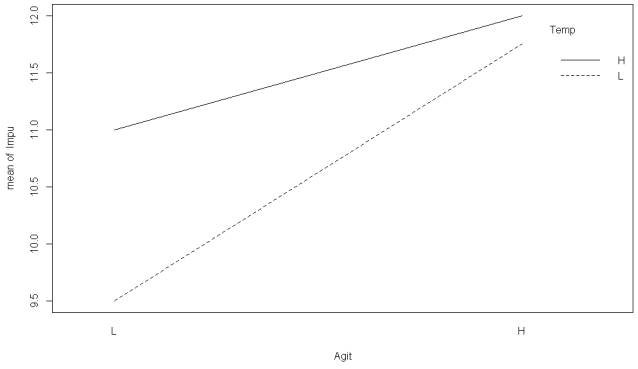
iv. (3 Marks) Write down a regression equation that can be used predicting amounts based on the results of this experiment.

```
            Df Sum Sq Mean Sq F value     Pr(>F)
A            1    ...     ...     ...   0.000979 ***
C            1    ...     ...     ...   0.934131
T            1    ...     ...     ...   0.395554
A:C          1    ...     ...     ...   0.944243
A:T          1    ...     ...     ...   0.017582 *
C:T          1    ...     ...     ...   0.072101
A:C:T        1    ...     ...     ...   0.028522 *
Residuals    8  116.2    14.5
```

(Find the NEW plots)



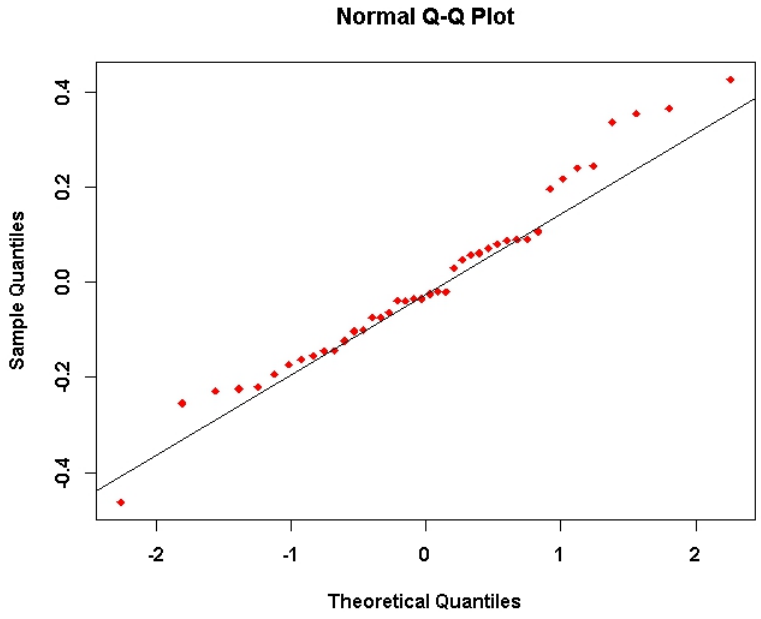




Question 4. Statistical Process Control (Variant A)

Part A: Testing Normality (2 Marks)

A graphical procedure was carried out to assess whether or not this assumption of normality is valid for data set `Y`. Consider the Q-Q plot in the figure below.



i. (1 Mark) Provide a brief description on how to interpret this plot.

ii. (1 Mark) What is your conclusion for this procedure? Justify your answer.

Part B: Testing Normality (3 Marks)

Consider the following inference procedure performed on data set X.

```
> shapiro.test(X)

Shapiro-Wilk normality test

data:  X
W = 0.9619, p-value = 0.6671
```

i. (1 Mark) Describe what is the purpose of this procedure.

ii. (1 Mark) What is the null and alternative hypothesis?

iii. (1 Mark) Write the conclusion that follows from it.

Question 4. Statistical Process Control (Variant B)

(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 (2 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. (3 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.

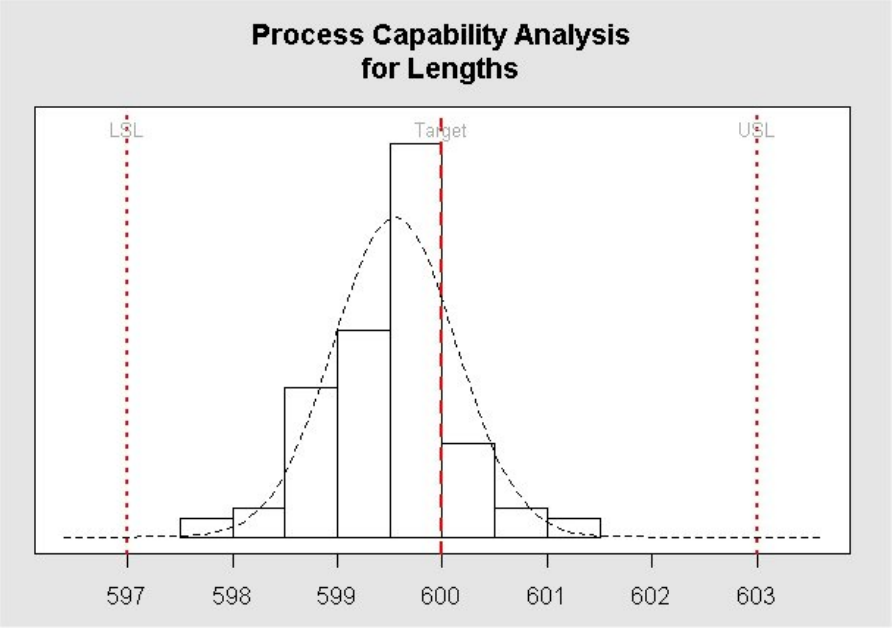
ii. (1 marks) The value of C_{pm} is 1.353. Explain why there would be a discrepancy between C_p and C_{pm}.

iii. (1 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%
```



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

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

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

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

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