FACULTY OF SCIENCE AND ENGINEERING

DEPARTMENT OF MATHEMATICS AND STATISTICS

REPEAT EXAMINATION PAPER 2017

MODULE CODE: MA4605

SEMESTER: Repeat 2017

MODULE TITLE: Chemometrics

DURATION OF EXAM: 2.5 hours

LECTURER: Mr. Kevin O'Brien

GRADING SCHEME: 100 marks
60% of module grade

EXTERNAL EXAMINER: Prof. A. Marshall

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. (25 marks) Inference Procedures

(a) (i.) (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

(ii.) (2 Marks) Describe any required assumptions for tests, and the limitations of these tests.

(b) Numeric Transformations, such as logarithmic transformation, are often used in statistical analysis as an approach for dealing with non-normal data.

(i) (1 Marks) Discuss the importance of numeric transformations, such as logarithmic transformation, in Statistics.

(ii.) (3 Marks) Give two examples of a transformation for various types of skewed data (i.e. an example for both types of skewness).

(iii.) (1 Mark) Discuss the limitations of numeric transformations.

(c) Consider the following inference procedure performed on data set Z.

```
> shapiro.test(Z)

Shapiro-Wilk normality test

data:  Z
W = 0.8914, p-value = 0.007047
```

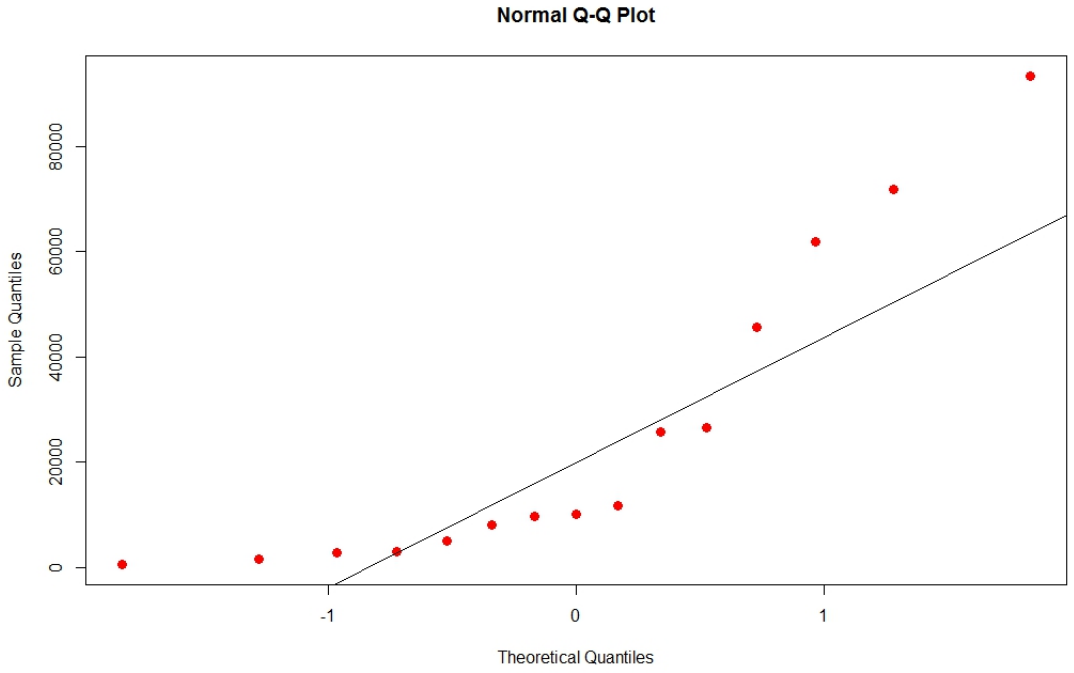
(i.) (1 Mark) Describe what is the purpose of this procedure.

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

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

iv. (1 Mark) Tests for Normality are known to be susceptible to low power. Discuss what is meant by this.

(d) A graphical procedure was carried out to assess whether or not this assumption of normality is valid for data set `Z`. Consider the figure below.



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

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

(e) (5 Marks)

A test of a specific blood factor has been devised such that, for adults in Western Europe, the test score is normally distributed with mean 100 and standard deviation 10. A clinical research organization is carrying out research on the blood factor levels for individuals with a particular disease, with emphasis on the effects of medication on the blood factor level.

For a group of 10 volunteer patients the following test scores were obtained both prior to, and after the medication.

Patient	A	B	C	D	E	F	G	H	I	J
Before	120	140	112	109	114	116	99	108	109	111
After	104	112	110	107	101	103	101	102	103	102

The organization wishes to determine if there is a significant improvement (lessening of the blood factor level) due to the medication. Using the output, shown below, write a short report discussing your findings. State the null and alternative hypotheses clearly. (You may assume that the case-wise differences are normally distributed.) *(The* `R` *output is presented on the next page)*

```
> t.test(Before,After,paired=TRUE)

Paired t-test

data:  Before and After
t = 3.3881, df = 9, p-value = 0.008023
alternative hypothesis:
  true difference in means is not equal to 0

95 percent confidence interval:
3.090618 15.509382
sample estimates:
mean of the differences
9.3
```

(f) (5 Marks) The research organization wishes to assess the link between the blood factor level and intake of a particular supplement. Using the following output, write a short report discussing your findings. State the null and alternative hypotheses clearly.

```
> cor.test(BloodFactor,Supplement)

Pearson's product-moment correlation

data:  BloodFactor and Supplement
t = -1.384, df = 13, p-value = 0.1896
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
-0.7355908  0.1885016
sample estimates:
cor
-0.3583659

>
```

Question 2. (25 marks) Regression Models

(a) The fluorescence of each of a series of acidic solutions of quinine with concentrations 0,10,20,30,40,50 and 60 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	121.6
Std Deviations:	0.71	0.84	0.89	1.64	2.24	3.03	4.05

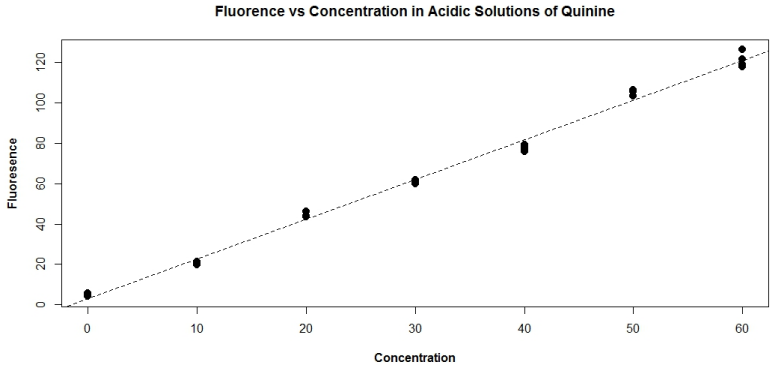


Figure 1:

The model is described by the following R code output. Another code segment in on the top of the next page. Answer the questions on the next page.

```
> summary(lm(Fluo~Conc))

Call:
lm(formula = Fluo ~ Conc)

Residuals:
    Min      1Q  Median      3Q     Max
-5.9645 -2.1473  0.2626  1.9254  5.2553

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.95221    1.01033   2.922  0.00711 **
Conc         1.96844    0.02802  70.247  < 2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 2.966 on 26 degrees of freedom
Multiple R-squared:  0.9948,    Adjusted R-squared:  0.9946
F-statistic:  4935 on 1 and 26 DF,  p-value: < 2.2e-16
```

```
> confint(lm(Fluo~Conc))
                2.5 %   97.5 %
(Intercept) 0.8754463 5.028968
Conc        1.9108368 2.026035
```

(i) (3 marks) What is the regression equation for this fitted model? In your answer, Interpret the slope and the intercept of the regression line.

(ii) (3 Marks) Comment on the significance of the regression estimates.

(iii) (2 Marks) State the 95% confidence interval for the slope and the intercept coefficients. Interpret this intervals with respect to any relevant hypothesis tests

(iv) (2 Marks) Explain in which way is the prediction intervals different from the confidence intervals for fitted values in linear regression?

(b) Model appraisal and validation are importants steps in the statistical modelling process. Answer the following questions.

(i) (2 Marks) In the context of regression models, explain what is meant by Heteroscedascity and Homoscedascity. Support your answers with sketches.

(ii) (1 Mark) Explain how the *Akaike information criterion* would used to compare two models fitted for the same data.

(iii) (1 Mark) Explain why the adjusted R^2 value may differ in value from the corresponding multiple R^2 value for the same fitted model.

(iv) (1 Marks) Explain the term "Influence" in the context of linear regression models. Support your answer with sketches.

(v) (1 Marks) Explain the term "Cook's Distance" in the context of linear regression models.

(vi) (2 Marks) The Durbin Watson Test was carried out to test for Autocorrelation. Briefly describe autocorrelation. You may support your answer with sketches.

(vii) (1 Mark) State your conclusion to the following procedure.

```
> durbinWatsonTest(myModel)
lag Autocorrelation D-W Statistic p-value
1     -0.08428163      2.143578   0.806
Alternative hypothesis: rho != 0
```

(c) (6 Marks) Suppose we have a regression model, described by the following equation

$$\hat{y} = 18.81 + 6.25x_1 + 7.82x_2 - 1.74x_3$$

We are given the following pieces of information.

* The standard deviation of the response variance y is 10 units.
* There are 76 observations.
* The *Coefficient of Determination* (also known as the *Multiple R-Squared*) is 0.80.

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 3. (25 marks) Experimental Design

(a) Specimens of milk from dairies in four different districts are assayed for their concentrations of the radioactive isotope Strontium-90. The results, in picocuries per litre, are shown in the table below.

District									Mean $\bar{x}_i$	St. Dev s_i
A	28.2	30.8	27.8	32.7	29.6	31.3	32.4	32.8	30.7	1.855
B	30.3	28.5	32.4	31.6	28.6	34.6	31.6	31.2	31.1	1.874
C	32.9	30.6	33.6	37.1	32.6	36.1	35.5	34.4	34.1	1.976
D	32.8	34.8	34	31.6	34.8	35.2	36.4	39.6	34.9	2.251
Overall									32.7	2.660

(i) (8 Marks) Complete the ANOVA table in your answer sheet, replacing the "?" entries with the correct values.

(ii) (1 Mark) What hypothesis is being considered by this procedure.

(iii) (1 Mark) What is the conclusion following from the above analysis? State the null and alternative hypothesis clearly.

(iv) (2 marks) State any two assumptions underlying the ANOVA model?

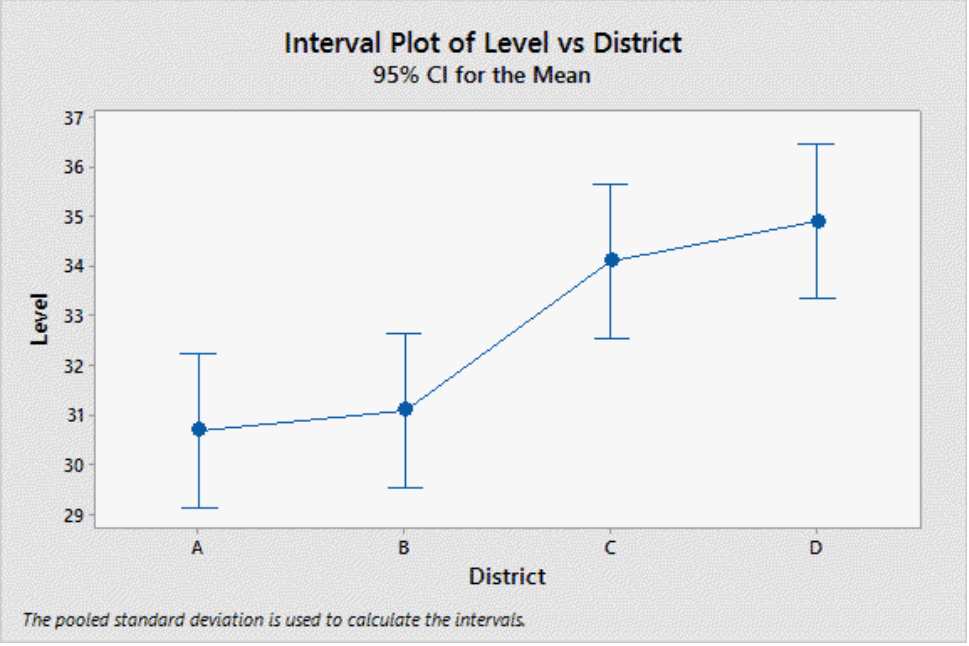
	DF	SS	MS	F
Between	?	?	?	?
Within	?	?	?	
Total	?			

(b) (4 Marks) The following outputs are Post-Hoc Procedures for the example in part (a). Interpret these outputs.

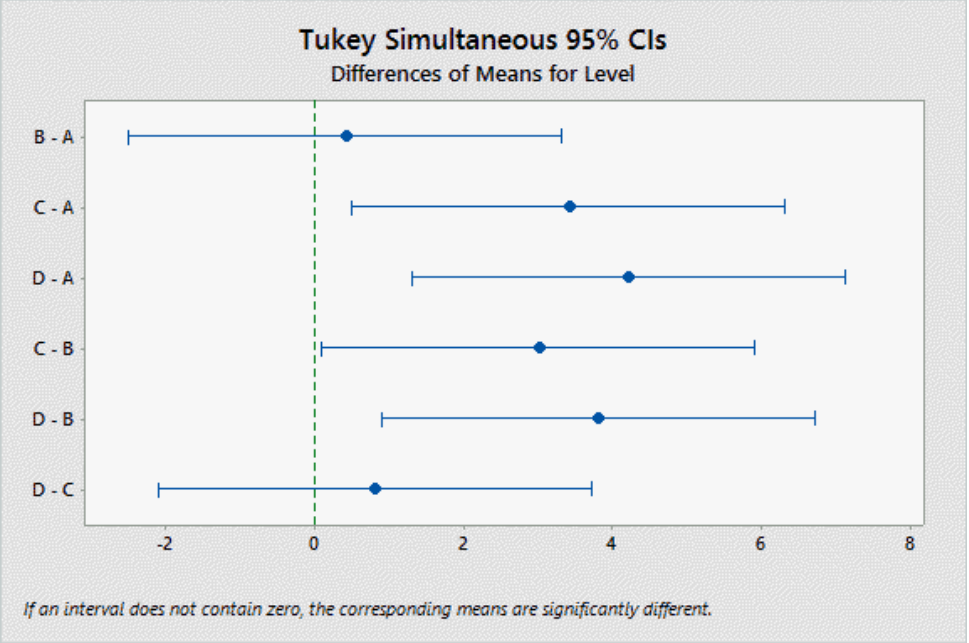
```
Tukey Pairwise Comparisons

Grouping Information Using the Tukey Method and 95% Confidence

District  N    Mean  Grouping
D         8  34.900  A
C         8  34.100  A
B         8  31.100    B
A         8  30.700    B
```



(The computer output continues on the next page.)



(c) Five standard solutions of chloride were prepared. Four titration methods, each with a different technique of end-point determination, were used to analyze each standard solution. The order of the experiments was randomized. The results of chloride found are shown below.

	Method U	Method V	Method W	Method X
Solution 1	10.15	10.64	10.71	10.71
Solution 2	10.13	10.23	10.72	10.01
Solution 3	10.32	10.22	10.60	10.24
Solution 4	10.29	10.72	10.53	10.58
Solution 5	9.81	10.28	10.20	9.99

You are also given the following information:

- $S_r^2 = 0.0394$
- $S_c^2 = 0.0305$
- The variance of concentrations in the table above is $\text{Var}(y) = 0.0771$

The following questions will result in the completion of the ANOVA Table on the next page. The $p-$values for both tests are already provided.

(i) (7 Marks) Complete the ANOVA Table. Show your workings.

(ii) (2 Marks) Based on the p-values, provided, what is your conclusion? Clearly state the null and alternative hypotheses for both tests.

Source	DF	SS	MS	F	p-value
Factor A	?	?	?	?	0.0130 *
Factor B	?	?	?	?	0.0196 *
Error	?	?	?		
Total	?	?			

Question 4. (25 Marks) Experimental Design

(a) (i) (2 Marks) What is a randomised block design?

(ii) (2 Marks) What is an "a x b" factorial experimental design?

(iii) (2 Marks) What is the difference between a between-treatments estimate and a within treatments estimate?

(iv) (2 Marks) What distinguishes a factorial experiment from a completely randomised experiment or a randomised block experiment?

(b) 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
Agitation speed (rpm)	15	30
Concentration of NaOH	40%	50%
Temperature (°F)	150	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.

Agitation A	Conc NaOH C	Temperature T	Impurity Replicate 1	Impurity Replicate 2
-1	-1	-1	39	34
1	-1	-1	40	47
-1	1	-1	23	34
1	1	-1	25	36
-1	-1	1	75	89
1	-1	1	61	75
-1	1	1	59	43
1	1	1	21	20

The data was analysed with statistical computing software, creating the output presented on the next page. Some numbers have been removed.

```
> summary(aov(y~A*C*T, Fact))

            Df Sum Sq Mean Sq F value  Pr(>F)
A            1  .....   .....   6.694 0.03225 *
C            1  .....   .....  16.594 0.00357 **
T            1  .....   .....  15.276 0.00449 **
A:C          1  .....   .....   0.245 0.63361
A:T          1  .....   .....  18.436 0.00264 **
C:T          1  462.2   462.2   5.603 0.04545 *
A:C:T        1  .....   .....   0.048 0.83124
Residuals    8  660.0    82.5
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

You are not required to complete this table. Read the questions carefully.

(i) (2 Marks) Calculate the contrasts, the effects and the sum of squares for the main effect A.

(ii) (2 Marks) Calculate the contrasts, the effects and the sum of squares for the main effect T.

(iii) (2 Marks) Calculate the contrasts, the effects and the sum of squares for the interaction effect between A and T.

(iv) (2 Marks) Calculate the contrasts, the effects and the sum of squares for the interaction effect between A, C and T.

(v) (3 Marks) Comment on the tests for significant for the main effects and interactions. State your conclusions clearly.

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

(vii) (3 Marks) Comment on the interaction plots on the next page. What does each one indicate?

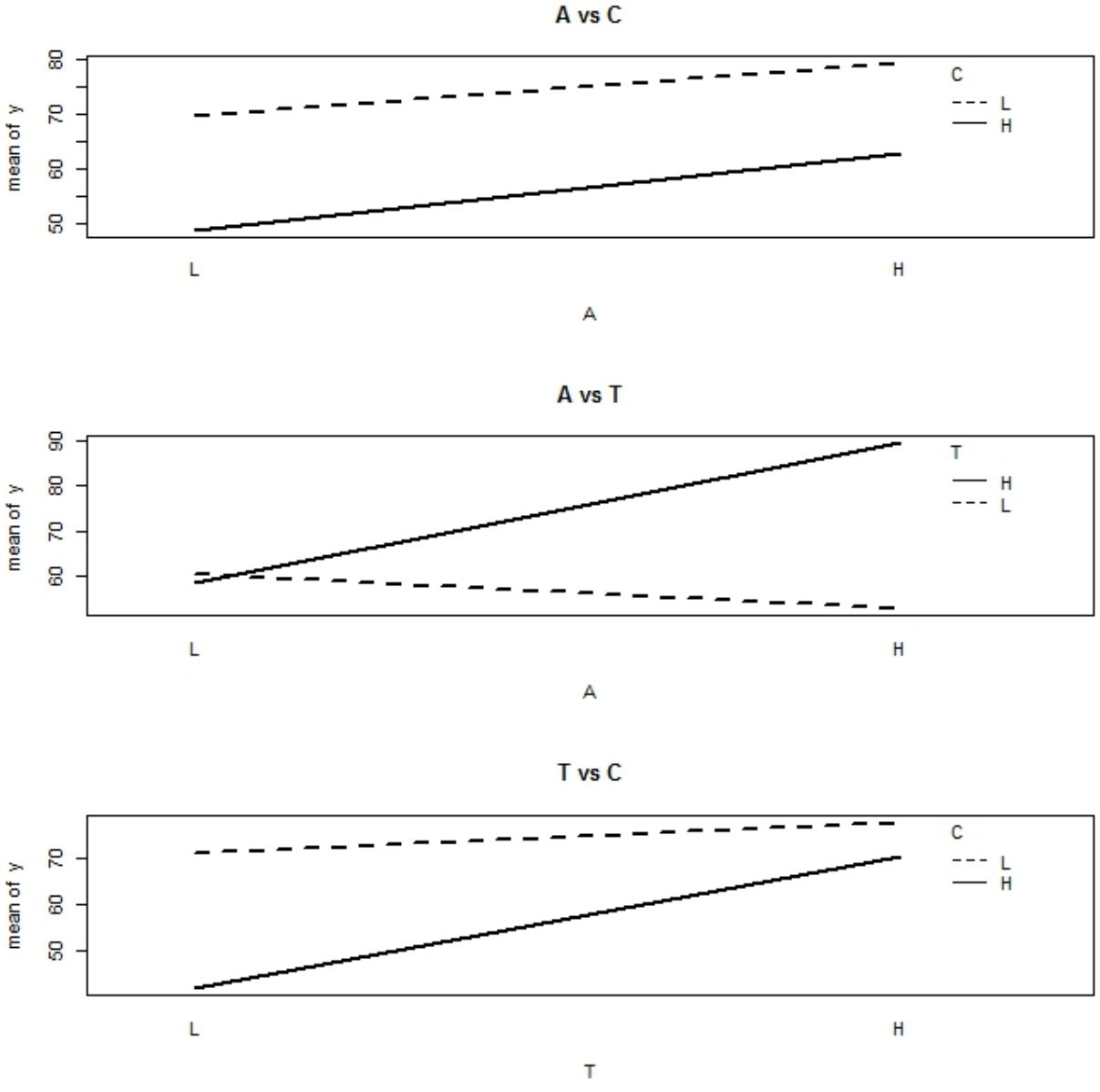


Figure 2: Question 4 Part B : Interaction Plots

Question 5. (25 marks) Statistical Process Control

(a) (3 Marks) Write a brief description of the Mahalanobis Distance. Illustrate your answer with a sketch.

(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 are 600±3mm.

(i) (4 Marks) Determine the *Process Capability Indices* C_p and C_{pk}, commenting on the respective values. Use the `R` code output shown below.

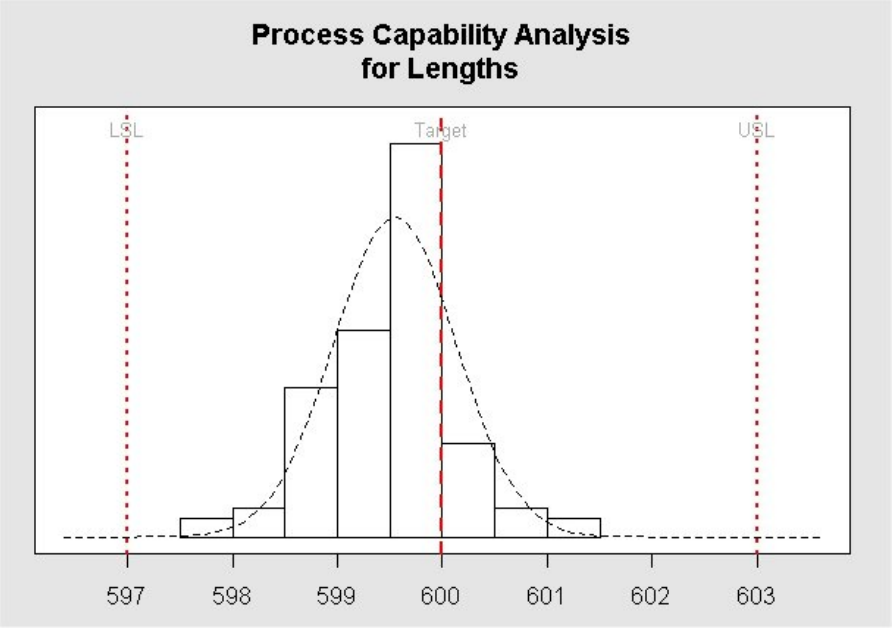
(ii) (2 Mark) Explain why there would be a discrepancy between C_p and C_{pk}. Illustrate your answer with sketches.

(iii) (1 Mark) 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%
```



(d) (3×3 Marks) The **Nelson Rules** are a set of eight decision rules for detecting "out-of-control" or non-random conditions on control charts. These rules are applied to a control chart on which the magnitude of some variable is plotted against time. The rules are based on the mean value and the standard deviation of the samples.

Discuss any three of these rules, stating their mathematical basis and how they would be used to detect "out of control" processes. Support your answer with sketch.

In your answer, you may make reference to the following properties of the Normal Distribution. Consider the random variable X distributed as

$$X \sim \mathcal{N}(\mu, \sigma^2)$$

where μ is the mean and σ^2 is the variance of X.

* $\Pr(\mu - 1\sigma \leq X \leq \mu + 1\sigma) = 0.6827$
* $\Pr(\mu - 2\sigma \leq X \leq \mu + 2\sigma) = 0.9545$
* $\Pr(\mu - 3\sigma \leq X \leq \mu + 3\sigma) = 0.9973$

Formulas and Tables

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

Two Way ANOVA

$$MS_A = c \times S_r^2 \qquad MS_B = r \times S_c^2$$

Control Limits for Control Charts

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

$$\bar{\bar{x}} \pm 3\frac{\bar{s}}{c_4\sqrt{n}} \qquad \left[\bar{R}D_3, \bar{R}D_4\right]$$

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

$$\text{Effect} = \frac{\text{Contrast}}{4n}$$

$$\text{Sums of Squares} = \frac{(\text{Contrast})^2}{8n}$$

Process Capability Indices

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

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

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