Q1. Theory for Inference Procedures (3 Marks)

Answer the three short questions. Each correct answer will be awarded 1 mark.

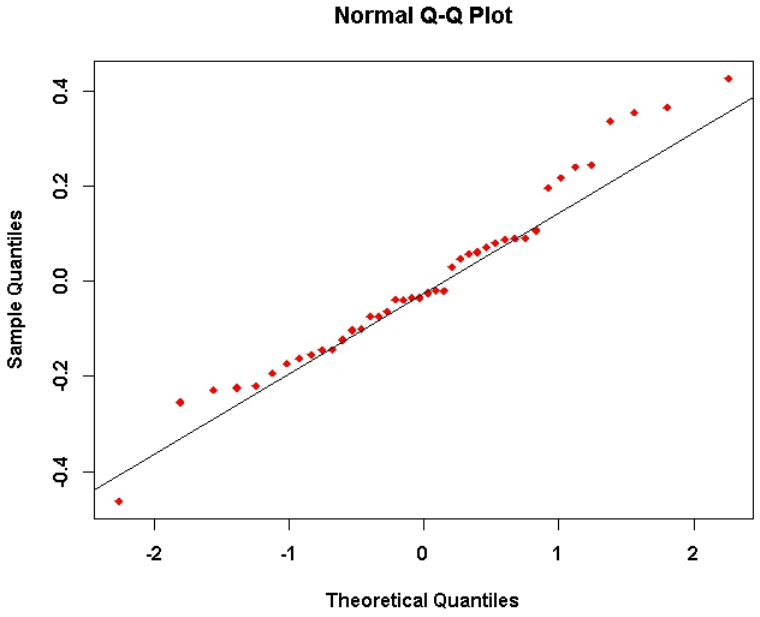
i. (1 Mark) Briefly describe how $p-$value is used in hypothesis testing

ii. (1 Mark) What is meant by a Type I error?

iii. (1 Mark) What is meant by a Type II error? How do Type II errors relate to the "Power" of a hypothesis test?

Q2. Testing Normality (3 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.

Q3. Testing Normality (4 Marks)

Consider the following inference procedure performed on data set X.

```
> shapiro.test(X)

Shapiro-Wilk normality test

data:  X
W = 0.8914, p-value = 0.07047
```

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.

Q4. 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.

Q5. Testing For Outliers (6 Marks)

(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, any required assumptions and the limitations of these tests.

(ii) (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.

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

Q6. 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.

Q7. Regression ANOVA

(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	?	?	?		

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. *(You may assume that the Pearson Correlation Coefficient is a positive number.)*
- (1 Mark) The sample standard deviation of the response variable Y.

Q8. Regression ANOVA

The mercury level of several tests of sea-water from costal areas was determined by atomic-absorption spectrometry. The results obtained are as follows

Concentration in $\mu g\, l^{-1}$	0	10	20	30	40	50	60	70	80	90	100
Absorbance	0.321	0.834	1.254	1.773	2.237	2.741	3.196	3.678	4.217	4.774	5.261

The analysis of the relationship between concentration and absorbance is obtained in R and presented below.

```
x<-seq(0,100,by=10)
y<- c(0.321, 0.834, 1.254, 1.773, 2.237, 2.741, 3.196, 3.678,
 4.217, 4.774, 5.261)
model<- lm(y~x)
summary(model)

Call:
lm(formula = y ~ x)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.2933636  0.0234754   12.50 5.45e-07
x           0.0491982  0.0003968  123.98 7.34e-16
---

Residual standard error: 0.04162 on 9 degrees of freedom
Multiple R-squared: 0.9994,     Adjusted R-squared: 0.9993
F-statistic: 1.537e+04 on 1 and 9 DF,  p-value: 7.337e-16

confint(model)
            2.5 %     97.5 %
(Intercept) 0.24025851 0.34646876
x           0.04830054 0.05009582
```

(i) (2 marks) Determine and interpret the slope and the intercept of the calibration plot.

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

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

(iv) (2 Marks) The following piece of `R` code gives us a statistical metric. What is this metric? What is it used for? How should it be interpreted.

```
> AIC(model)
[1] -34.93389
```

Q9.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

Q10. 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.

Q11.Experimental Design

Explain the following terms in the context of experimental design

i. (2 marks) levels of a factor.

ii. (2 marks) randomized block design.

Q12. One-Way ANOVA

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.

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

Q13. Testing assumptions for ANOVA

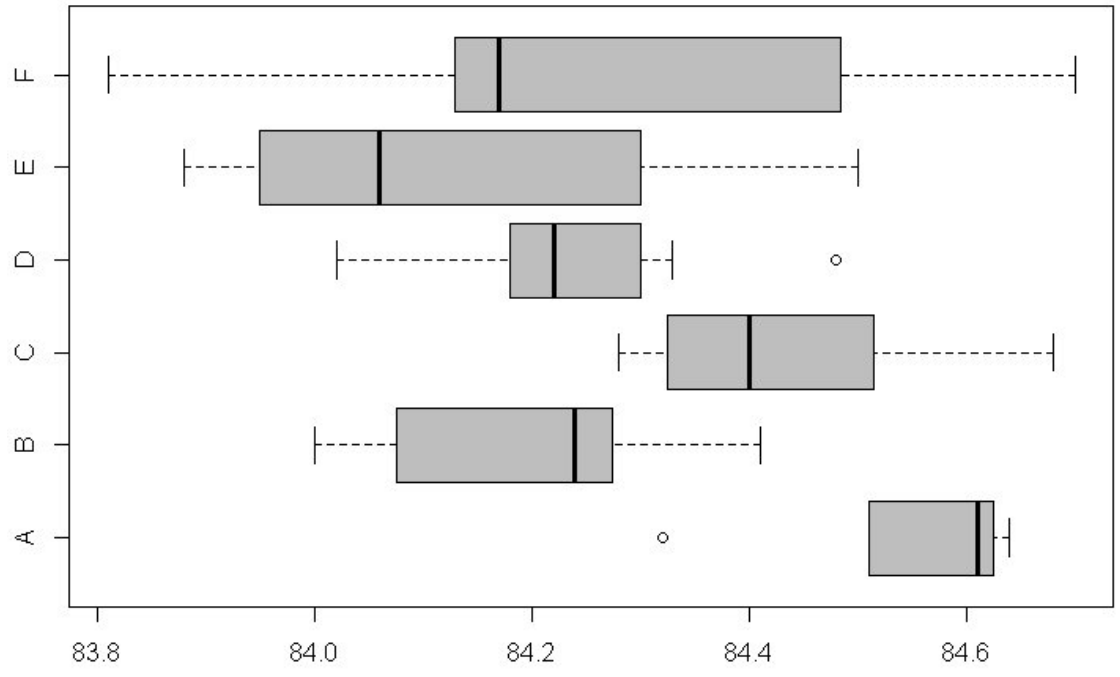
The R code and graphical procedures, below and on the following page, are relevant to checking whether the underlying assumptions are met for the ANOVA model in part (b).

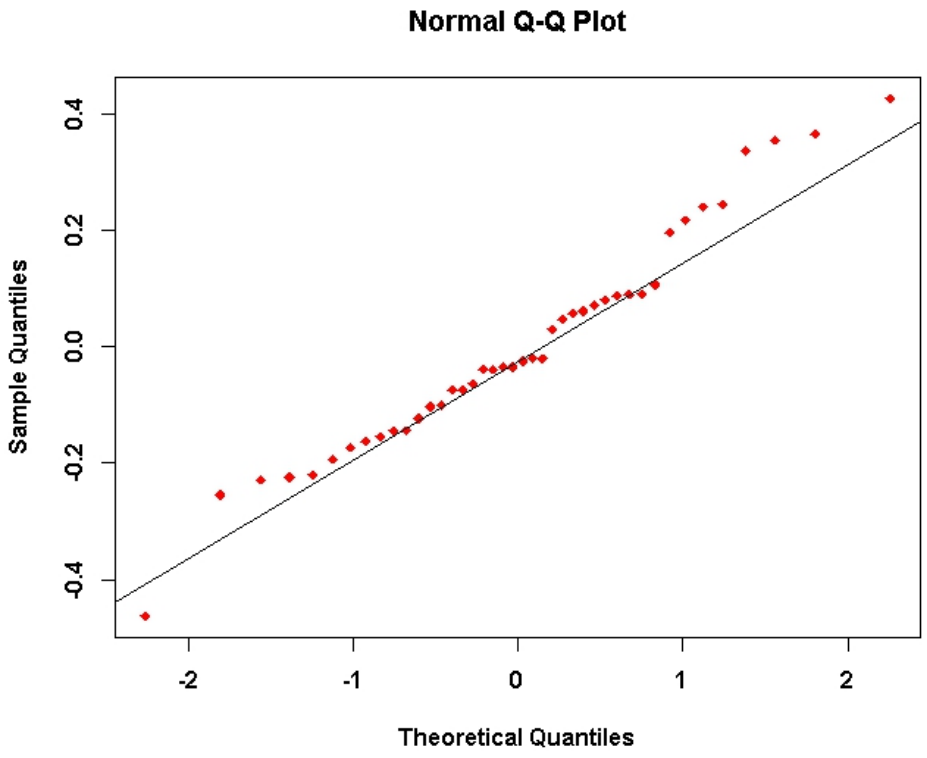
i. (3 marks) What are the assumptions underlying ANOVA?

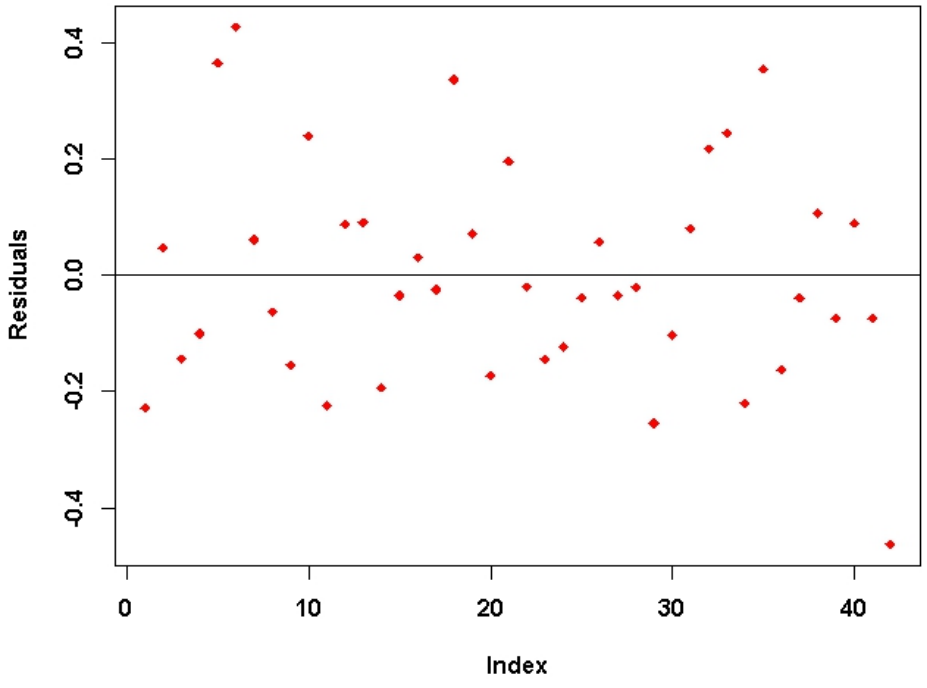
ii. (4 marks) Assess the validity of these assumptions for the ANOVA model in part(b).

```
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
```




Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles
-0.4
-0.2
0.0
0.2
0.4
-2
-1
0
1
2


Residuals
Index
-0.4
-0.2
0.0
0.2
0.4
0
10
20
30
40

Q14. ANOVA Example

Assume that we have three fertilizers to be tested. We wish to determine if there is any difference is the mean yields for the three different types of fertilizer.

Fertilizer A	5.6	6.4	6.6	5.8
Fertilizer B	5.1	6.2	6.4	5.7
Fertilizer C	5.0	6.1	5.8	5.5

The following R output is a One-Way ANOVA procedure for testing multiple means.

```
Fert <- c("A", "A", "A", "A", "B", "B", "B", "B",
"C", "C", "C", "C")
Yield <- c(5.6, 6.4, 6.6, 5.8, 5.1, 6.2, 6.4, 5.7,
5, 6.1, 5.8, 5.5)

ModelA=aov(Yield~Fert)
```

```
> summary(aov(Yield~Fert))
            Df Sum Sq Mean Sq F value Pr(>F)
Fert         2   0.50  0.2500   0.957   0.42
Residuals    9   2.35  0.2611
```

State you conclusion to this procedure

Q15. Two Way ANOVA

- A standard solution was prepared, containing 16.00% (by weight) of chloride. Three titration methods, each with a different technique of end-point determination, were used to analyse the standard solution.
- The procedure was carried out by four different clinical analysts. The order of the experiments was randomized. The results for the chloride found (% w/w) are shown below:

	Analyst 1	Analyst 2	Analyst 3	Analyst 4
Method A	16.03	16.05	16.02	16.12
Method B	16.13	16.13	15.94	15.97
Method C	16.09	16.15	16.12	16.1

```
> Model=aov(Titr ~ Meth + Anlt)
>
> summary(Model)
Df  Sum Sq   Mean Sq  F value Pr(>F)
Meth           2  0.01202  0.006008 1.279   0.345
```

```
Anlt          3  0.01109  0.003697 0.787   0.543
Residuals     6  0.02818  0.004697

---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

State you conclusion to this procedure

Q16. One Way ANOVA

Four laboratory technicians performed six determination of C of 2,4 dinitrophenol in water, according to the same specified procedure. The results in $C/\mu M$ are as follows

Analyst A	Analyst B	Analyst C	Analyst D
701	550	511	613
677	545	523	623
680	573	540	649
660	532	542	632
654	529	559	614
648	534	554	626

The analysis of variance procedure is used to determine if there is a significiant difference between the mean of the determinations makde by the four investigators.

```
summary(aov(Det~Anlt))
            Df Sum Sq Mean Sq  F value   Pr(>F)
Anlt        ..  99942   ....    .....    4.64e-12 ***
Residuals   ..   6918   ....
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

(a) The value for the degrees of freedom for residuals has been removed from the output. What is this value?

(b) The value for the test statistics (`F value`) has been removed from the output. What is this value?

(c) State the null and alternative hypothesis for this procedure.

(d) Based on the $p-$value, what is your conclusion for this procedure.

Q17. Two Way ANOVA

Four standard solutions of chloride were prepared. Three 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. The

Solution	Method X	Method Y	Method Z
1	10.03	10.13	10.12
2	10.13	10.15	10.12
3	10.09	10.02	9.97
4	10.05	9.94	10.10

following output table is the result of performing a two-way ANOVA (without interactions).

Analysis of Variance Table

```
Response: chloride

          Df   Sum Sq   MeanSq  F value    Pr(>F)
A          ? 0.012017   ?       1.2791    0.3446
B          ? 0.011092   ?       0.7871    0.5435
Residuals  ? 0.028183   ?
Total      ?        ?
```

(a) Identify the two factors A and B, and their corresponding levels. Are they controllable or random? Construct the hypothesis statements. (2 marks)

(b) Fill in the missing values for the mean sum of squares. (2 marks)

(c) Test whether there are significant differences between the concentration of chloride in the different solutions, and whether there are significant differences between the results obtained by the different methods. (2 marks)

(d) Compute the variance of the all results computed for this experiment. (2 marks)

(e) Is it possible to include the interaction term in the model, given the data we have? Why? How can this be changed? (2 marks)

Q18. Residual Analysis for Regression Models

Explain the following terms

(i) Influence

(ii) Leverage

(iii) Cooks Distance

Write a brief explanation of how robust regression differs from linear models computed using the ***Ordinary Least Squares*** method, making reference to one particular weighting method.

Q19. Numeric Transformation of Data

(i) Describe the purpose of transformations

(ii) Describe the process of transformations

(iii) Describe the purpose of Tukey's Ladder (referencing direction and relative strength)

(iv) Give an example of a transformation for various types of skewed data (use Tukey's Ladder, with an example for both directions)

(v) Describe the limitations of transformations

Q20. Inference Procedures with R

The following table gives the concentration of norepinephrine (μmol per gram creatinine) in the urine of healthy volunteers in their early twenties. Male: 0.48 0.36 0.28 0.55 0.45 0.46 0.47 0.25; Female 0.35 0.37 0.27 0.29 0.28 0.31 0.33. The problem is to determine if there is evidence that concentration of norepinephrine differs between genders. The following analyses have been performed on the data:

```
        F test to compare two variances
data:  M and F
F =7.90, num df =7, denom df =6, p-value =0.022
alter. hypothesis:
ratio of variances not equal to 1
95 percent confidence interval:
  1.386947 40.433419
ratio of variances
         7.899317
```

```
        Welch Two Sample t-test
data:  M and F
t = 2.4732, df = 8.953, p-value = 0.03551
alter. hypothesis:
difference in means not equal to 0
95 percent confidence interval:
 0.008309402 0.188119170
mean of x mean of y
0.4125000 0.3142857
```

- (4pts) Explain what is the purpose of each of the two procedure.
- (4pts) Write the conclusions that follow from the given analyses.
- (2pts) How, if at all, does the first procedure affect the second one?

Q21. Inference Procedures with R

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 (A): mean = 10 ng/ml, n=12.

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

Data: Sample (B): mean = 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=c(8.40, 9.59, 9.38, 9.10, 10.78, 11.41, 9.94, 10.08, 12.11, 9.10, 9.59, 10.36)
> B=c(47.5, 48.4, 48.8, 48.4, 46.8, 46.2, 48.6, 50.6, 45.5, 46.1)
>
> var.test(A,B)

        F test to compare two variances

data:  A and B
F = 0.4514, num df = 11, denom df = 9, p-value = 0.2141
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.115379 1.619474
sample estimates:
ratio of variances
         0.4513712

> t.test(A,B,var.equal=TRUE)

        Two Sample t-test

data:  A and B
t = -67.5402, df = 20, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -38.86779 -36.53887
sample estimates:
mean of x mean of y
 9.986667 47.690000
```

- Do the data sets have variances that differ significantly? Justify your answer. Write down the null and alternative hypotheses being considered.
- Write down the estimate for the difference of means.
- State the 95% confidence interval for the difference of means. Interpret this confidence interval with respect to the null hypothesis.

Q22. Two Way ANOVA - No Replicates

Three varieties of potatoes are being compared for yield. The experiment was carried out by assigning each variety at random to four of twelve equal size plots, one being chosen in each of four locations. The following yields in bushels per plot resulted:

	Potato		
Location	A	B	C
1	18	13	12
2	20	23	21
3	14	12	9
4	11	17	10

Additional Information

- The variance of the Row means is : $S_R^2 = 19.037$.
- the variance of the Column means is : $S_C^2 = 3.0625$.
- Also the overall variance of the 12 observations is $\text{Var}(Y) = 21.6363$.

Exercise Complete the Two Way ANOVA table. You are not required to perform any hypothesis testing.

Q23. One Way ANOVA

A trial is undertaken to investigate the effect on fuel economy of 3 fuel additives A, B and C, where A and B are new and C is the current standard additive. The same driver drives the same car on a fixed test route during 20 working days. The additive used on each day is randomly assigned so that A and B are each used for 5 days and C is used for 10 days. The response variable measured each day is Y, the number of miles per gallon (mpg) achieved.

The results are shown in the following table.

Additive	*y*	*Total*
A	39, 35, 37, 36, 38	$\sum y_A = 185$
B	36, 41, 39, 40, 39	$\sum y_B = 195$
C	37, 33, 30, 34, 36, 34, 31, 36, 34, 35	$\sum y_C = 340$

You are given that the sum of squares of the observations is 26078.

(i) (a) Carry out an analysis of variance to test for differences between the effects on Y of the additives. State clearly your null and alternative hypotheses and present your conclusions.

(11)

You are given the additional piece of information , sufficient to construct ANOVA table

	A	B	C	Overall
Mean	37	39	34	36
Std.Dev	1.5811	1.8708	2.2111	2.8837

The p-value for the Test Statistic is 0.00650

Q24. Two Way ANOVA

Given the following details below, construct the appropriate Two-Way ANOVA Table. *You are not required to do any hypothesis testing.*

- There are 2 factors: A and B. A has 2 levels, while Factor B has 3 levels.
- There are 54 observations in the experiment.
- The variance of the response variable is 174.2075.
- The Sum of Squares for Factors A and B are 451 and 2034 respectively.
- The Sum of Squares for Error is 5745.

Q25. One Way ANOVA

A chemist is trying three different procedures to prepare a solution. Three independent samples were taken. He repeated the first procedure 7 times and recorded the concentration of a certain substance. The average of the observations was 3.1. He repeated the second procedure 8 times and recorded the concentration of the same substance. The average of the observations was 4.9. He repeated the third procedure 5 times and recorded the concentration of the substance. The average of the observations was 4.1. The sum of the squares of all 20 observations $\left(\sum x_i^2\right)$ is 392.102. Let μ_1, μ_2 and μ_3 denote the expected concentrations of the substance under the three procedures.

(a) Test the hypothesis

$$H_0 : \mu_1 = \mu_2 = \mu_3$$

against the alternative that this is not so.

Additional Information:

- The Variance of the Response Variable is 3.2. (You can ignore the part about the sum of squares for all 20 obervations.)
- Ordinarily you would be given the standard deviation for each sample, and from that, directly compute SSwithin. In this question , use this identity : SStotal = SSbetween + SSwithin.
- You have been given enough information to compute the Overall Mean.
- We may not get around to covering the confidence intervals material in MA4605 2015. It will be retained for possible use in future.

Q26. Two Way ANOVA (no replicates)

A taxi company employs four drivers, each one with their own car. The takings of each driver during each one of the seven days of the same week were recorded. The total takings over the week of driver A were €840, of driver B were €858.06, of driver C were €866.88 and of driver D were €921.06. The following is the calculated ANOVA table based on daily takings with some entries missing.

Source	degrees of freedom	sum of squares	mean square	F - value
Day				3.62
Drivers				
Error		1162.26		
Total				

(a) Complete the table using the information provided above.

	Driver A	Driver B	Driver C	Driver D	MEANS
Mon					
Tue					
Wed					
Thu					
Fri					
Sat					
Sun					
MEAN					

For a question like this, you may expect to be given the variance of the row means and column means.

Variance of Row Means : S^2_R = 58.435
Variance of Column Means : S^2_c = 24.8329

Also: The Variance of the Response Variable is 114.3033

Q27. One Way ANOVA

Students from three different schools took the same mathematics test. The average score of students from school A was 57; 7 students took the test from school A. The average score of students from school B was 69.7; 10 students took the test from school B. The average score of students from school C was 63.2; 5 students took the test from school C. The sum of the squares of all 22 test scores $\left(\sum x_i^2\right)$ is 108084. Let μ_A, μ_B and μ_C denote the expected test scores from the three schools.

(a) Test the hypothesis

$$H_0 : \mu_A = \mu_B = \mu_C$$

against the alternative that this is not so.

Additional Information:

- The variance of the response variable is 831.3695. (*You can ignore now some of the information given in the question*)
- In this question , use this identity : SStotal = SSbetween + SSwithin.
- You have been given enough information to compute the Overall Mean.
- We may not get around to covering the confidence intervals material in MA4605 2015. It will be retained for possible use in future.

Q28. Two Way ANOVA (no replicates)

An experiment is conducted to study how long different digital camera batteries last. The aim is to find out whether there is a difference in terms of battery life between four brands of batteries using seven different cameras. Each battery was tried once with each camera. The time the Brand A battery lasted was 43.86 hours. The times for brands B, C and D were 41.28, 40.86 and 40 hours respectively. The following is the calculated ANOVA table with some entries missing.

Source	degrees of freedom	sum of squares	mean square	F - value
Cameras			26	
Batteries				
Error				
Total		343		

(a) Complete the table using the information provided above.

	Battery A	Battery B	Battery C	Battery D
Camera 1				
Camera 2				
Camera 3				
Camera 4				
Camera 5				
Camera 6				
Camera 7				

For a question like this, you may expect to be given the variance of the row means and column means.

Variance of Row Means : $S^2_R = 6.5$
Variance of Column Means : $S^2_c = 2.7585$

Also: The Variance of the Response Variable is 12.7037

Q29. Two Way ANOVA (with replicates)

Consider the following experiment (similar to question 28) where there are 5 measurements per treatment group. Complete the following ANOVA table.

	Battery A	Battery B	Battery C
Camera 1			
Camera 2			
Camera 3			

Source	DF	SS	MS	F
Camera	*	100	*	*
Battery	*	40	*	*
Camera: Battery	*	*	5	*
Error	*	144	*	
Total	*	*		

For the three F test statistics, state the appropriate degrees of freedom for the corresponding critival value. *(You are not required to perform the hypothesis test)*

Question 30 - Residual Diagnostics

Expect a question on Hypothesis Tests from car `R` package, and Model diagnostic plots. This scope of this question covers the following topics.

- `ncvTest()` - Non Constant Error Variance
- `outlierTest()` - Outliers
- `durbinWatsonTest()` - Autocorrelation
- Cook's Distances
- Diagnostic Plot 1 (Fitted Vs Residual)
- Diagnostic Plot 2 (Residual Normality)

Question 31 - Control Limits

Short Question on Calculating Control Limits

Exam Paper Formulas for Control Limits

- Process Mean

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

- Process Standard Deviation

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

- Process Range

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

Question 32 - Method Comparison

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 *
....
```

- **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 *
....
```

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

ii. (2 marks) Provide a brief description of the Bland-Altman plot. Discuss any shortcomings with this approach to method comparison.

Question 33 - 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) (10 Marks) Each of the six blocks of code describes a statistical inference procedure. Provide a brief description for each procedure.

(b) (10 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 34 - Experimental Design

(Remark : This question will not feature in the 2015 Winter Exam)

(i) Give the principal features of a balanced completely randomised design, and explain the role of replication in such a design.

(ii) State the statistical model for this design, define the terms in the model and state the standard assumptions made about the error term.

(iii) Two basic principles of experimental design are **randomisation** and **replication**. Explain why these are important and how they help to validate an analysis of experimental results.

(iv) Briefly explain the principles of randomisation and replication, in the context of a completely randomised experimental design. Write down the model equation for a completely randomised design having equal numbers of replicates in all treatment groups, defining all the symbols that you use.

Question 35 - (NOT USING)

Short Experimental Design Theory Question

- Short Description on Box-Behnken Design
- Short Description on Central Composite Design

- Rationale for Designs like these

Update: This will not feature as a question in the Winter 2015 Exam.

Question 36- Experimental Design Part 2

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

ii. (8 Marks) Using the computed sums of squares values, complete the ANOVA table (see the `R` code below).

iii. (4 Marks) Comment on the tests for significant for the main effects and interactions. State clearly your conclusions.

iv. (4 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
```

Question 37 - ANOVA

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.

Question 38 - Assumptions for ANOVA

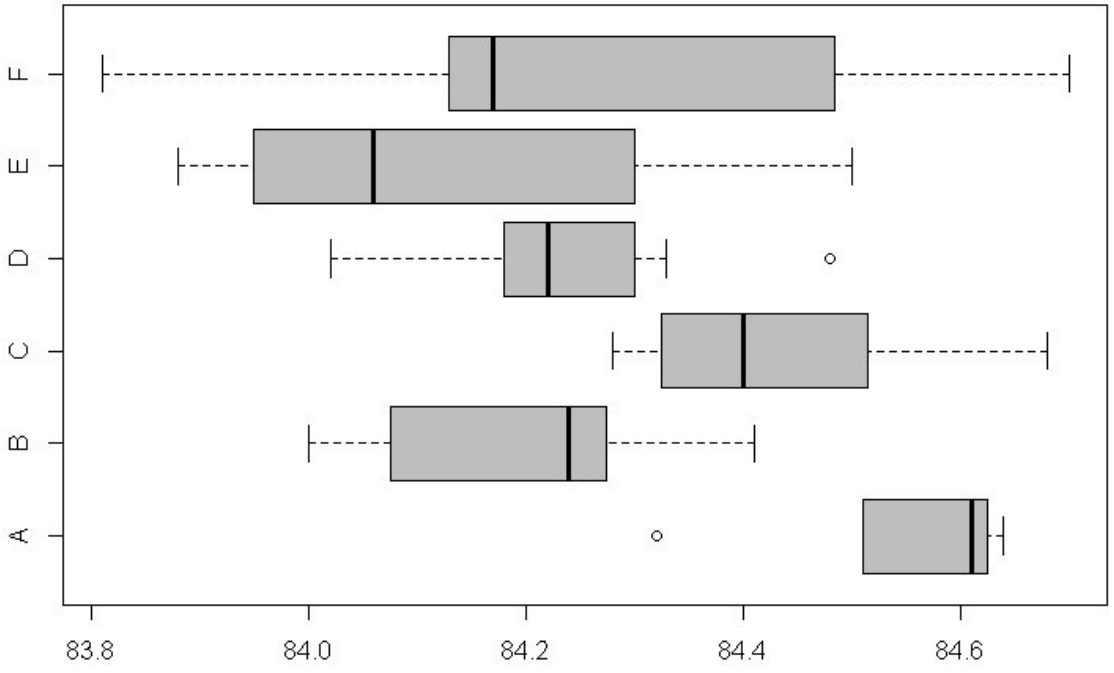
The R code and graphical procedures, below and on the following page, are relevant to checking whether the underlying assumptions are met for the ANOVA model in part (b).

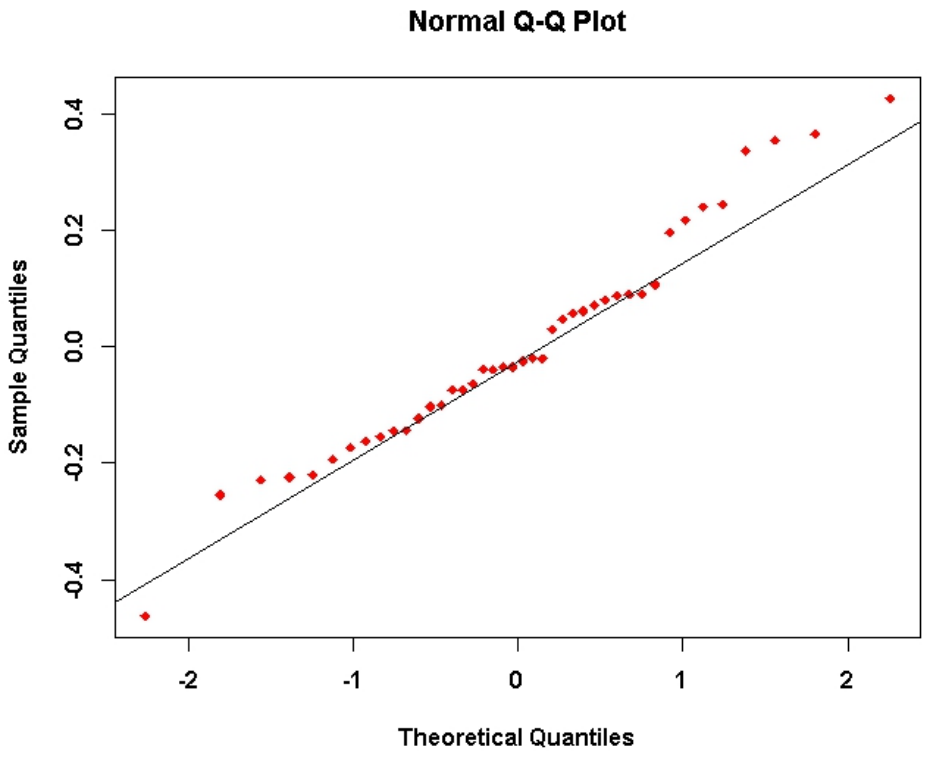
i. (3 marks) What are the assumptions underlying ANOVA?

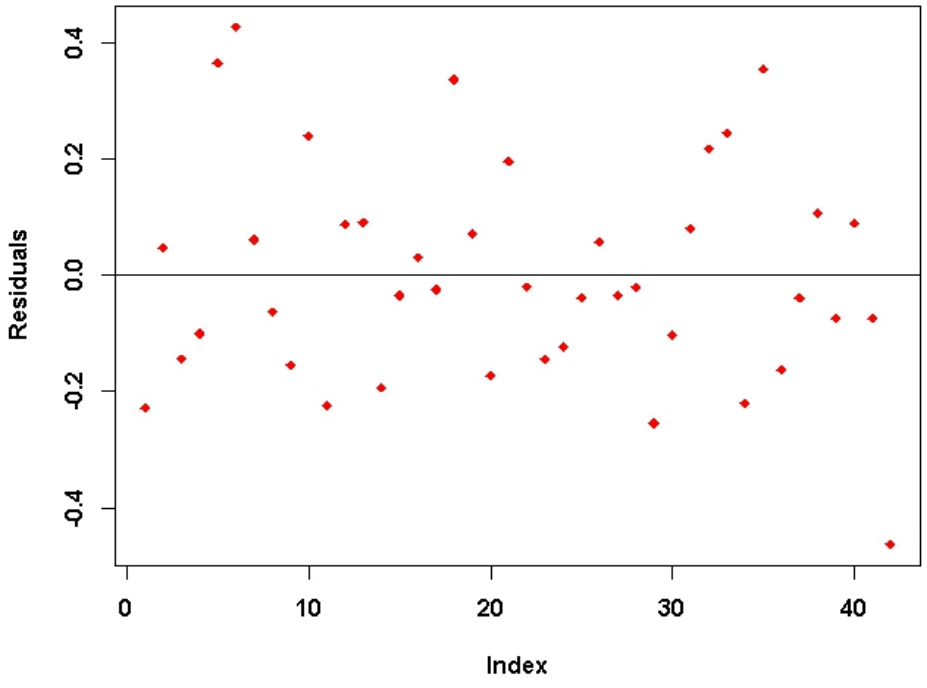
ii. (4 marks) Assess the validity of these assumptions for the ANOVA model in the previous question (Question 37).

```
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
```




Normal Q-Q Plot
Sample Quantiles
Theoretical Quantiles
-0.4
-0.2
0.0
0.2
0.4
-2
-1
0
1
2


Residuals
Index
-0.4
-0.2
0.0
0.2
0.4
0
10
20
30
40

Question 39 - Control Charts Arithmetic

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
X-Chart	614	620	626
R-Chart	0	8.236	18.795

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

ii. (2 marks) Estimate the mean of the standard deviations $\bar{s}$ for this process.

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

Question 40 - Process Capability Indices

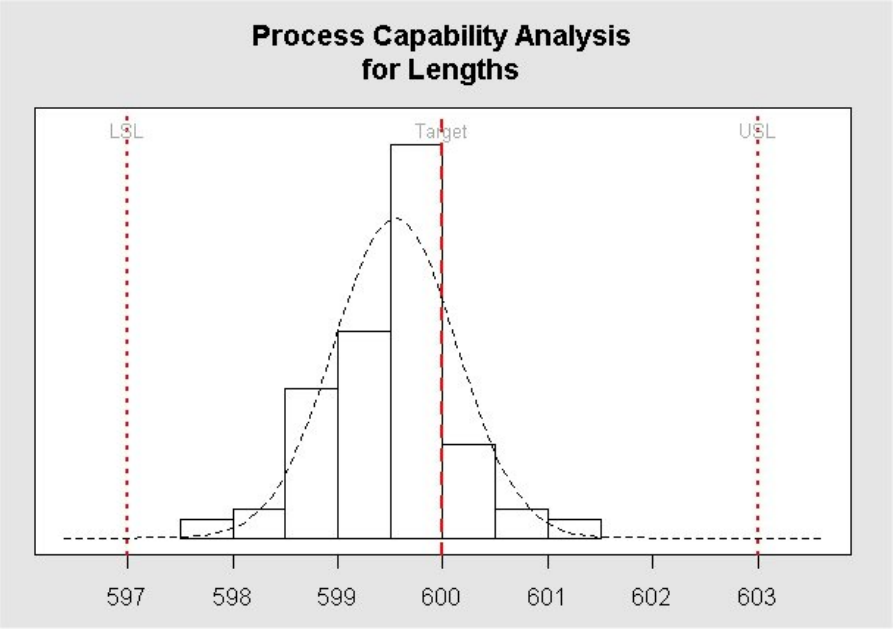
An automobile assembly plant concerned about quality improvement measured sets of five camshafts on twenty occasions throughout the day. The specifications for the process state that the design specification limits at 600±3mm.

i. (2 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 41 - Statistical Process Control

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.

(Remark - Part (iii) of this question is similar to Question 47. Question 47 prompts you discuss the probabilities more.)

Question 42 - Control Charts Arithmetic

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 mean of the standard deviations $\bar{s}$ for this process.

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

Question 43 - Process Capability Indices

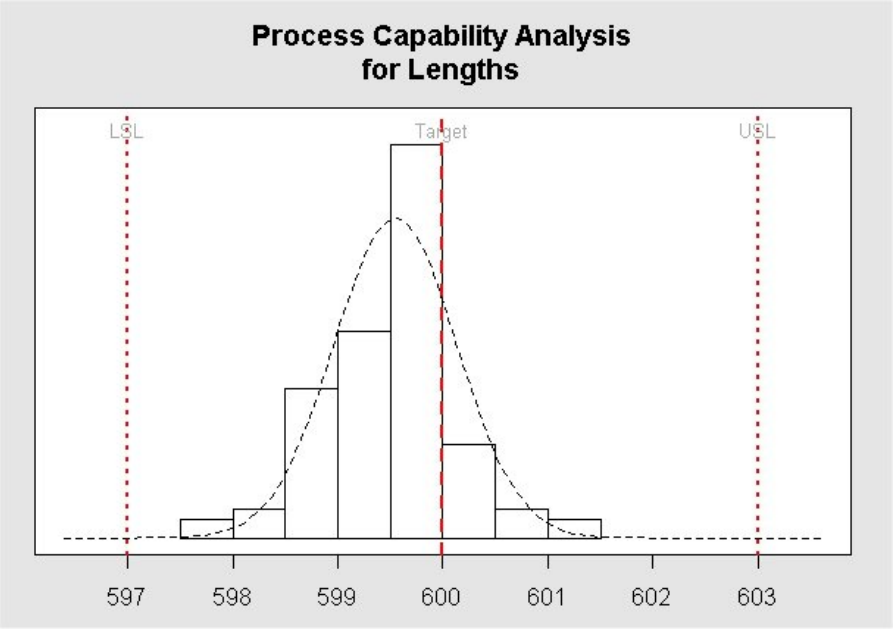
An automobile assembly plant concerned about quality improvement measured sets of five camshafts on twenty occasions throughout the day. The specifications for the process state that the design specification limits at 600±3mm.

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 44 - Factorial Design

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
-	-	-	90,70
+	-	-	100,120
-	+	-	90,110
+	+	-	120,150
-	-	+	110,100
+	-	+	100,130
-	+	+	100,80
+	+	+	160,140

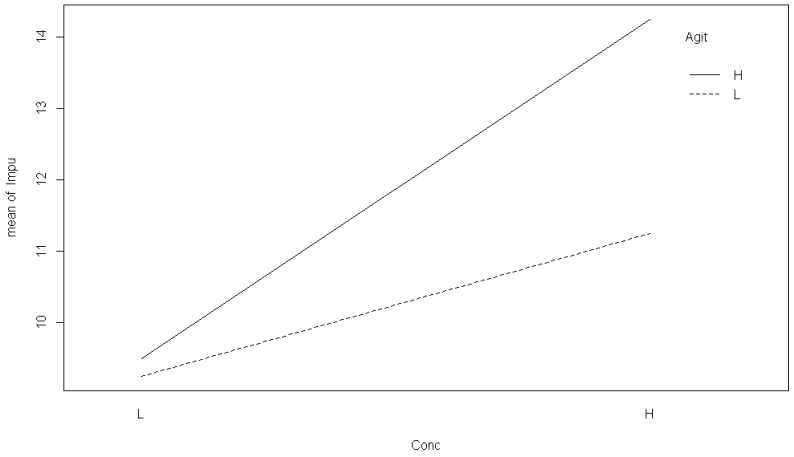
i. (8 Marks) Calculate the contrasts, the effects and the sum of squares for the effects.

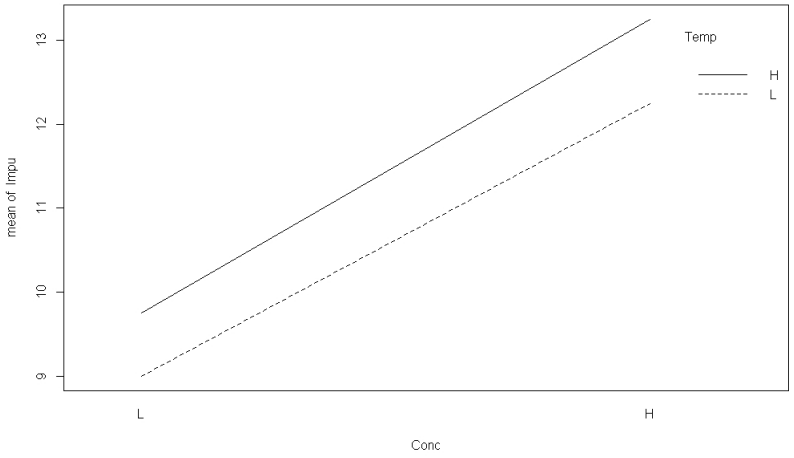
ii. (8 Marks) Using the computed sums of squares values, complete the ANOVA table (see the `R` code below).

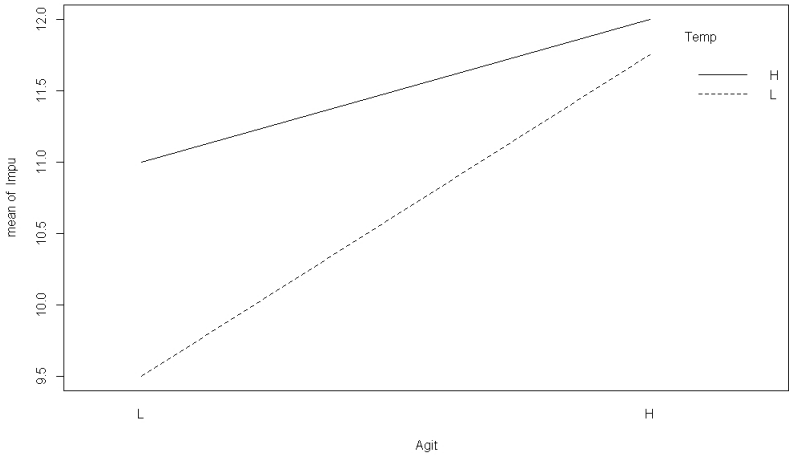
iii. (4 Marks) Comment on the tests for significant for the main effects and interactions. State clearly your conclusions.

iv. (4 Marks) Write down a regression equation that can be used predicting impurity based on the results of this experiment.

```
Df Sum Sq Mean Sq F value  Pr(>F)
Conc            1    ...     ...   ...   0.00253 **
Agit            1    ...     ...   ...   0.07093 .
Temp            1    ...     ...   ...   0.29485
Conc:Agit       1    ...     ...   ...   0.48239
Conc:Temp       1    ...     ...   ...   0.87675
Agit:Temp       1    ...     ...   ...   0.44646
Conc:Agit:Temp  1    ...     ...   ...   0.18751
Residuals       8   1950     244
```


mean of Impu
14
13
12
11
10
Agit
H
L
L
H
Conc


mean of Impu
13
12
11
10
9
Temp
H
L
L
H
Conc


mean of Impu
12.0
11.5
11.0
10.5
10.0
9.5
Temp
H
L
L
H
Agit

Question 45 - Statistical Process Control

Answer the following questions.

i. (1 marks) What is the purpose of maintaining control charts?

ii. (1 marks) What is the ***Three Sigma*** rule in the context of statistical process control?

iii. (2 Marks) What is a CUSUM chart? What type of departures from the production target value is this type of chart useful for detecting?

Question 46 - Regression Analysis

For a study into the density of population around a large city, a random sample of 10 residential areas was selected, and for each area the distance from the city centre and the population density in hundreds per square kilometre were recorded. The following table shows the data and also the log of each measurement.

distance, x (km)	*population density, y*	*log x*	*log y*
0.4	149	–0.916	5.004
1.0	141	0.000	4.949
3.1	102	1.131	4.625
4.5	46	1.504	3.829
4.7	72	1.548	4.277
6.5	40	1.872	3.689
7.3	23	1.988	3.135
8.2	15	2.104	2.708
9.7	7	2.272	1.946
11.7	5	2.460	1.609

(i) By plotting three separate graphs, decide which of the following regressions is best represented by a straight line.

(a) y on x (b) y on $\log x$ (c) $\log y$ on x

(7)

(ii) On the basis of the regression results **on the next page**, which regression do you consider to be best? Justify your answer by reference to the diagnostic criteria given in the output and relating these to your plots in (i). Would you consider regressing $\log y$ on $\log x$? If not, why not?

(5)

(iii) For the model you consider to be best in (ii), obtain an expression for y in terms of x.

(3)

(iv) Using your chosen model, estimate the density of the population at a distance of 5 km from the city centre.

(2)

(v) State any reservations you have about using the model to predict population density.

(3)

Regression Analysis: y versus x

The regression equation is y = 140 - 14.0x

Predictor	Coef	SE Coef	T	P
Constant	139.70	11.12	12.56	0.000
x	-13.958	1.663	-8.39	0.000

S = 18.2834 R-Sq = 89.8% R-Sq(adj) = 88.5%

Observation 10 has an unusually large positive residual

Regression Analysis: y versus logx

The regression equation is y = 127 - 48.0logx

Predictor	Coef	SE Coef	T	P
Constant	126.990	9.147	13.88	0.000
logx	-47.980	5.293	-9.07	0.000

S = 17.0492 R-Sq = 91.1% R-Sq(adj) = 90.0%

Observation 1 has an unusually large negative residual

Regression Analysis: logy versus x

The regression equation is logy = 5.41 - 0.322x

Predictor	Coef	SE Coef	T	P
Constant	5.4133	0.1621	33.40	0.000
x	-0.32157	0.02425	-13.26	0.000

S = 0.266544 R-Sq = 95.6% R-Sq(adj) = 95.1%

Question 47 - Nelson Rules for Control Charts

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.

(i) (4×2 Marks) Discuss any four of these rules, 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 an random variable X.

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

Question 48 - Tukey's Ladder For Data Transformation

(i) (1 Mark) Describe the purpose of Tukey's Ladder (referencing direction and relative strength)

(ii) (2 Marks) Give an example of a transformation for various types of skewed data (use Tukey's Ladder, with an example for both directions)

(iii) (2 Marks) Describe the limitations of transformations

Formulae 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

ANOVA Pocedures

$$\mathrm{Var}(Y) = \frac{SS_{tot}}{n-1}$$

Two Way ANOVA

$$\mathrm{MS}_A = c \times S_R^2 \qquad (= \mathrm{MS}_{Trt})$$
$$\mathrm{MS}_B = r \times S_C^2 \qquad (= \mathrm{MS}_{Block})$$