Q1. 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.

Q2. 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	0.1209
B	84.24	84.25	84.41	84.13	84.00	84.30	84.02	84.1928	0.1416
C	84.29	84.40	84.68	84.28	84.40	84.36	84.63	84.4342	0.1459
D	84.14	84.22	84.02	84.48	84.27	84.33	84.22	84.2400	0.1583
E	84.50	83.88	84.49	83.91	84.11	84.06	83.99	84.1343	0.2749
F	84.70	84.17	84.11	84.36	84.61	83.81	84.15	84.2729	0.3324

For the aggregrate sample (all 42 observations) the standard deviation is 0.2334.

(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					

Q3. 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)

(Important: For MA4505, Disregard the comment about the **Sum of Square of the Observations**.)
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

Q4. 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.

Q5. 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.

Q6. 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.

Q7. 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.

Q8. 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

Q9. 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

Q10. 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			

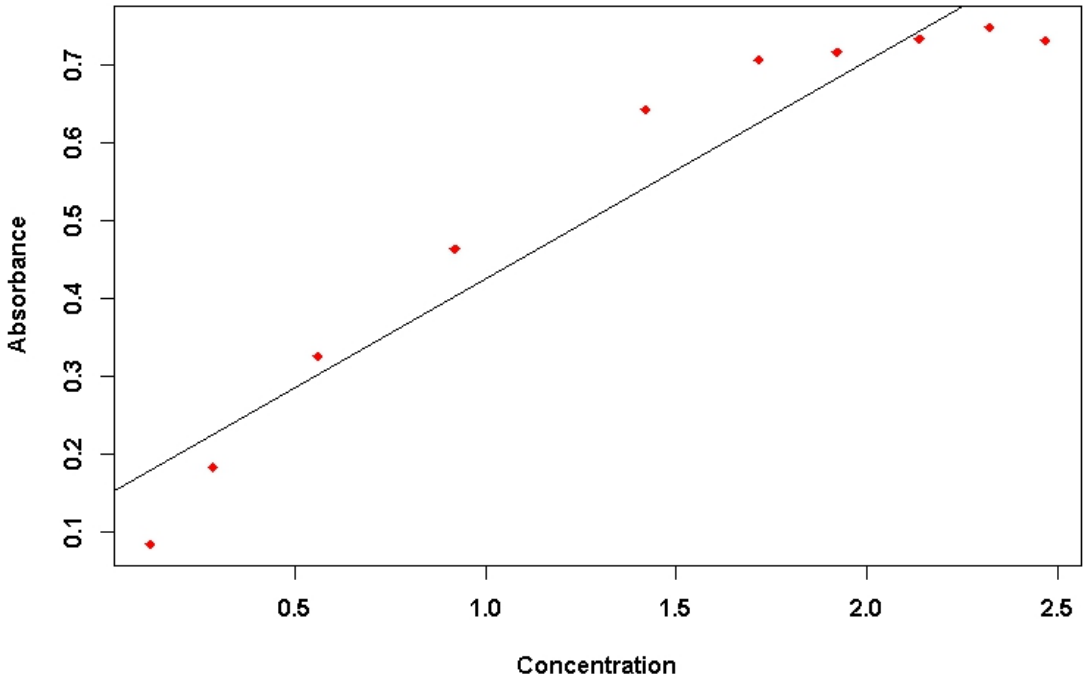
Source	DF	SS	MS	F
Camera	*	100	*	*
Battery	*	40	*	*
Camera: Battery	*	*	5	*
Error	*	144	*	
Total	*	*		

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)*

Q11. Polynomial Regression / Model Selection

In an experiment to determine hydrolysable tannins in plants by absorption spectroscopy, the following results from ten samples were obtained and are tabulated below. A simple linear regression model, predicting absorbance values using concentration as the independent variable, was fitted to the data.

Sample	1	2	3	4	5
Absorbance	0.084	0.183	0.326	0.464	0.643
Concentration	0.123	0.288	0.562	0.921	1.420
Sample	6	7	8	9	10
Absorbance	0.707	0.717	0.734	0.749	0.732
Concentration	1.717	1.921	2.137	2.321	2.467



A Simple Linear regression model and two polynomial models were fitted to the data. Description of all three fitted models are found in the three blocks of `R` code below. The *Akaike Information Criterion* is listed, for each of the three fitted models.

i. (1 marks) Is the simple linear regression model approach suitable for this study? Explain your answer with reference to the scatter-plot.

ii (3 marks) Write down the regression equations of each of the three models.

iv. (2 marks) Specify which one of the models you would use. Justify your answer with appropriate statistical values.

v. (2 marks) Using the best fit model, predict a value for absorbance when the concentration level is 1.2 mg/ml.

- **Model 1**

```
> summary(Model1)
Call:
lm(formula = Absorb ~ Conc)
...
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)    0.14412    0.04721   3.053   0.0158 *
Concentration  0.28088    0.02930   9.586 1.16e-05 ***
...

Residual standard error: 0.07584 on 8 degrees of freedom
Multiple R-squared: 0.9199,     Adjusted R-squared: 0.9099
F-statistic: 91.89 on 1 and 8 DF,  p-value: 1.163e-05
>
>AIC(Model1)
[1] -19.4343
```

- **Model 2**

```
> summary(Model2)
Call:
lm(formula = Absorb ~ Conc + Conc.Squared)
...
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)    0.006582   0.008013   0.821    0.439
Concentration  0.642935   0.015568  41.299 1.27e-09 ***
Conc.Squared  -0.140573   0.005894 -23.851 5.79e-08 ***
...
Residual standard error: 0.008939 on 7 degrees of freedom
Multiple R-squared: 0.999,      Adjusted R-squared: 0.9987
F-statistic:  3592 on 2 and 7 DF,  p-value: 2.879e-11
>
> AIC(Model2)
[1] -61.5338
```

- **Model 3**

```
> summary(Model3)

Call:
lm(formula = Absorb ~ Conc+ Conc.Squared + Conc.Cubed)
...
...
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)    0.013712   0.011629   1.179   0.2830
Concentration  0.608682   0.042825  14.213 7.58e-06 ***
Conc.Squared  -0.108186   0.038088  -2.840   0.0296 *
Conc.Cubed    -0.008196   0.009518  -0.861   0.4223
...

Residual standard error: 0.009109 on 6 degrees of freedom
Multiple R-squared: 0.9991,     Adjusted R-squared: 0.9987
F-statistic:  2306 on 3 and 6 DF,  p-value: 1.422e-09
>
> AIC(Model3)
[1] -60.69903
```

Q12. Model Selection

- Suppose we have 4 predictor variables :$\{X_1, X_2, X_3, X_4\}$
- Use Forward and Backward Selection to Choose the Optimal set of Predictor Variables, based on the AIC metrics listed for each combination of predictor variables listed below.

Variables	AIC	Variables	AIC
$\{\emptyset\}$	208.75	$\{X_2, X_4\}$	166.97
		$\{X_3, X_4\}$	164.16
$\{X_1\}$	169.30		
$\{X_2\}$	190.79	$\{X_1, X_2, X_3\}$	168.40
$\{X_3\}$	170.20	$\{X_1, X_2, X_4\}$	158.01
$\{X_4\}$	166.02	$\{X_1, X_3, X_4\}$	157.55
		$\{X_2, X_3, X_4\}$	165.78
$\{X_1, X_2\}$	169.61		
$\{X_1, X_3\}$	167.14	$\{X_1, X_2, X_3, X_4\}$	159.55
$\{X_1, X_4\}$	156.01		
$\{X_2, X_3\}$	166.97		

Q13. Model Selection

Use Forward and Backward Selection to Choose the Optimal set of Predictor Variables

(None)	AIC(Fit0)	255.4247	Mult R^2	0	adj. R^2	0
Acetic	AIC(FitA)	246.6389	Mult R^2	0.3019934	adj. R^2	0.2770646
H2S	AIC(FitB)	232.0245	Mult R^2	0.5711615	adj. R^2	0.5558458
Lactic	AIC(FitC)	236.8724	Mult R^2	0.4959486	adj. R^2	0.4779468
Acetic, H2S	AIC(Fit1)	233.2438	Mult R^2	0.5821773	adj. R^2	0.5512274
Acetic,Lactic	AIC(Fit2)	237.3884	Mult R^2	0.5202762	adj. R^2	0.4847411
H2S, Lactic	AIC(Fit3)	227.7838	Mult R^2	0.6517024	adj. R^2	0.6259025
All Three	AIC(FitAll)	229.7775	Mult R^2	0.6517747	adj. R^2	0.6115948

Q14. 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 metric.

$\emptyset$	200	x1,x2,x3	74
		x1,x2,x4	75
x1	150	x1,x2,x5	78
x2	170	x1,x3,x4	72
x3	135	x1,x3,x5	82
x4	130	x1,x4,x5	70
x5	140	x2,x3,x4	80
		x2,x3,x5	82
x1,x2	90	x2,x4,x5	78
x1,x3	81	x3,x4,x5	75
x1,x4	84		
x1,x5	78	x1,x2,x3,x4	83
x2,x3	87	x1,x2,x3,x5	130
x2,x4	78	x1,x2,x4,x5	104
x2,x5	87	x1,x3,x4,x5	101
x3,x4	85	x2,x3,x4,x5	89
x3,x5	88		
x4,x5	86	x1,x2,x3,x4,x5	100

Q15. 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.

Q16. 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 regression line.

(ii) (2 marks) 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
```

Q17. Regression Models

```
> summary(CheesesModel)

Call:
lm(formula = Taste ~ Acetic + H2S + Lactic)

Residuals:
Min      1Q  Median      3Q     Max
-17.390  -6.612  -1.009   4.908  25.449

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -28.8768    19.7354  -1.463  0.15540
Acetic        0.3277     4.4598   0.073  0.94198
H2S           3.9118     1.2484   3.133  0.00425 **
Lactic       19.6705     8.6291   2.280  0.03108 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 10.13 on 26 degrees of freedom
Multiple R-squared:  0.6518,    Adjusted R-squared:  0.6116
F-statistic: 16.22 on 3 and 26 DF,  p-value: 3.81e-06
```

(i) (2 marks) State the Regression Equation for this model

(ii) (8 marks) For each of the regression coefficients, interpret the test for significance. State the null and laternative hypothesis in each case. State your conclusion to each test.

Q18. Regression Analysis

(MA4505 2015 - just parts 2,3, and 4)

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%

Q19. Assumptions for ANOVA

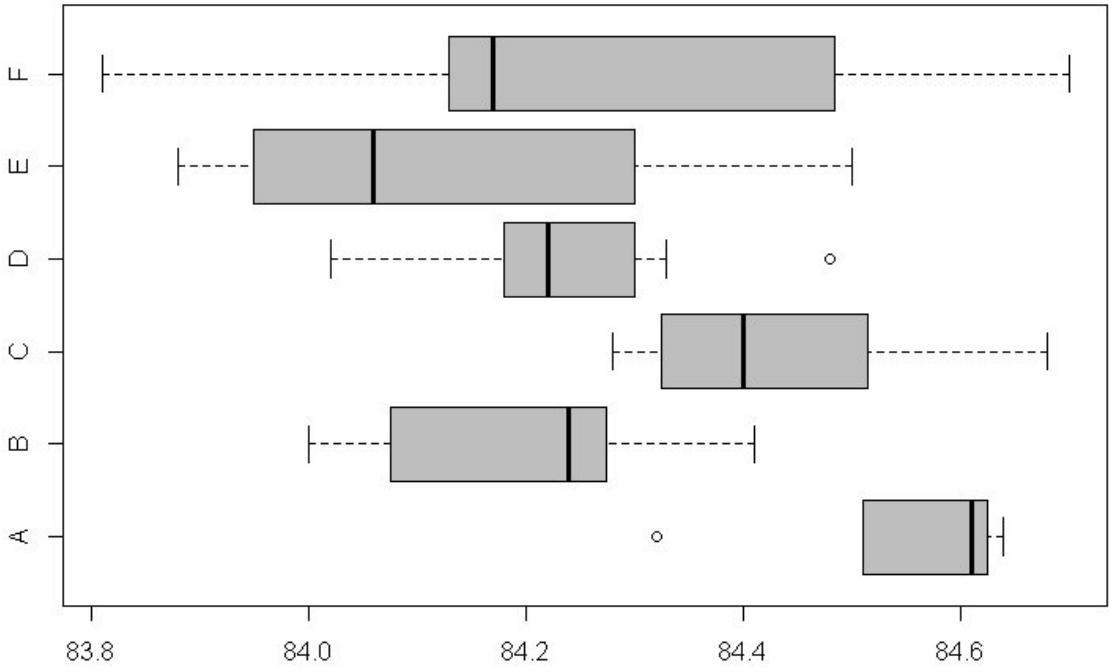
The R code and graphical procedures, below and on the following page, are relevant to checking whether the underlying assumptions are met for the ANOVA model in part (b).

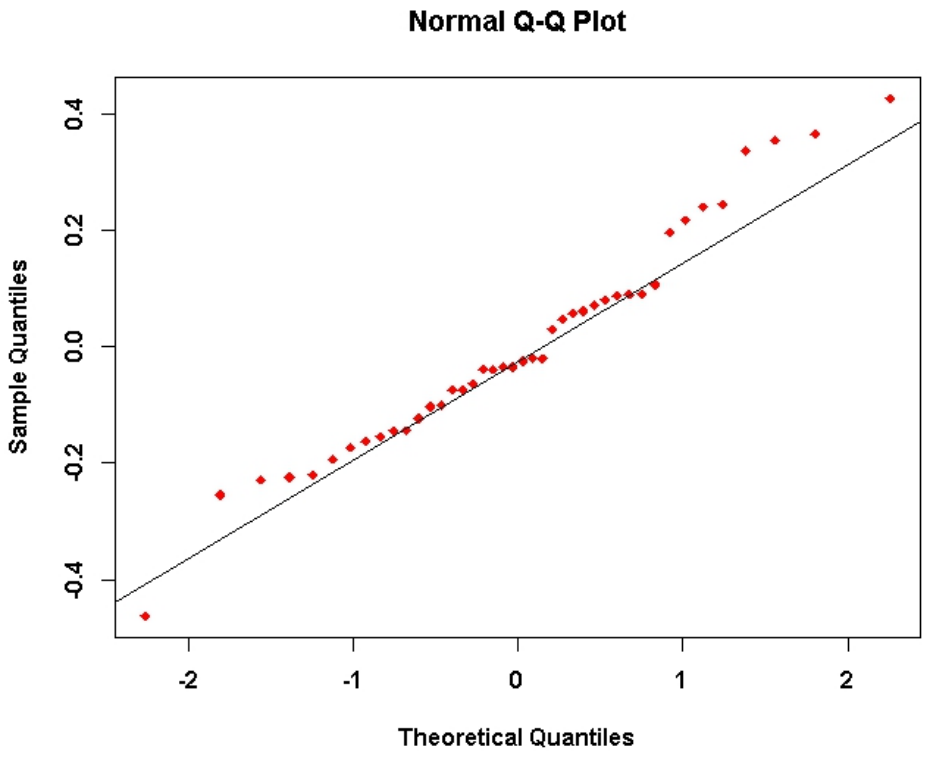
i. (3 marks) What are the assumptions underlying ANOVA?

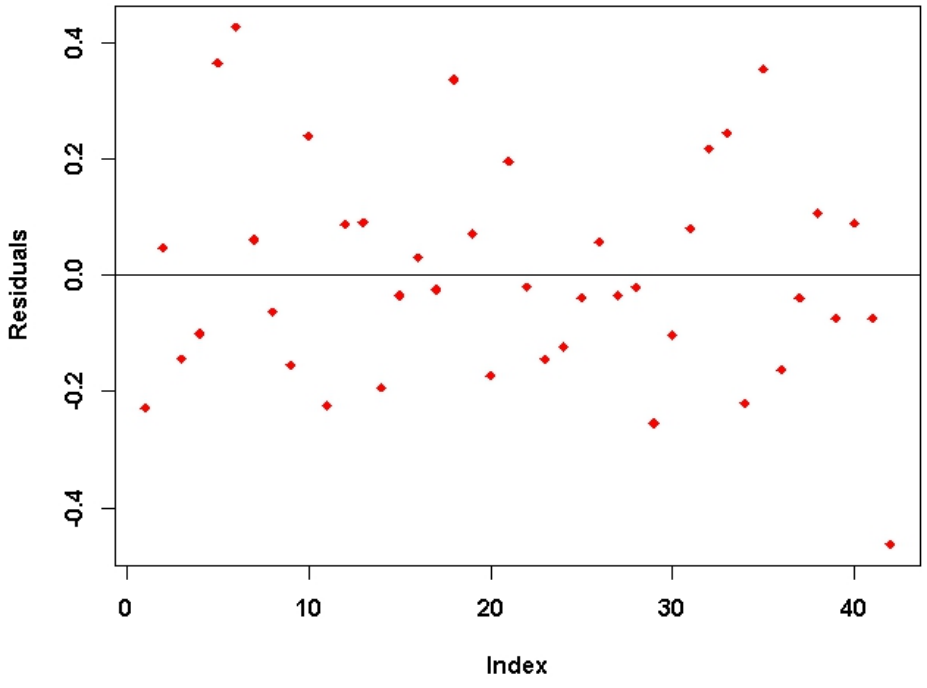
ii. (4 marks) Assess the validity of these assumptions for the ANOVA model in the previous question (Question 37).

```
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

Q20. 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?

Q21. 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

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

Q22. 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).

Q23. 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).

Q24. 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.

Q25. Process Capability Indices

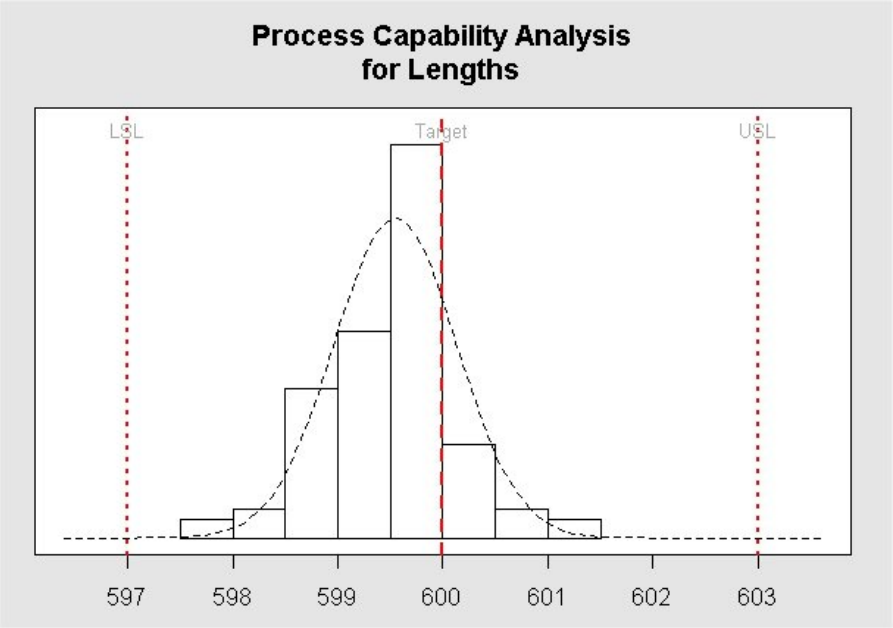
An automobile assembly plant concerned about quality improvement measured sets of five camshafts on twenty occasions throughout the day. The specifications for the process state that the design specification limits at 600±3mm.

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%
```



Q26. Process Capability Indices

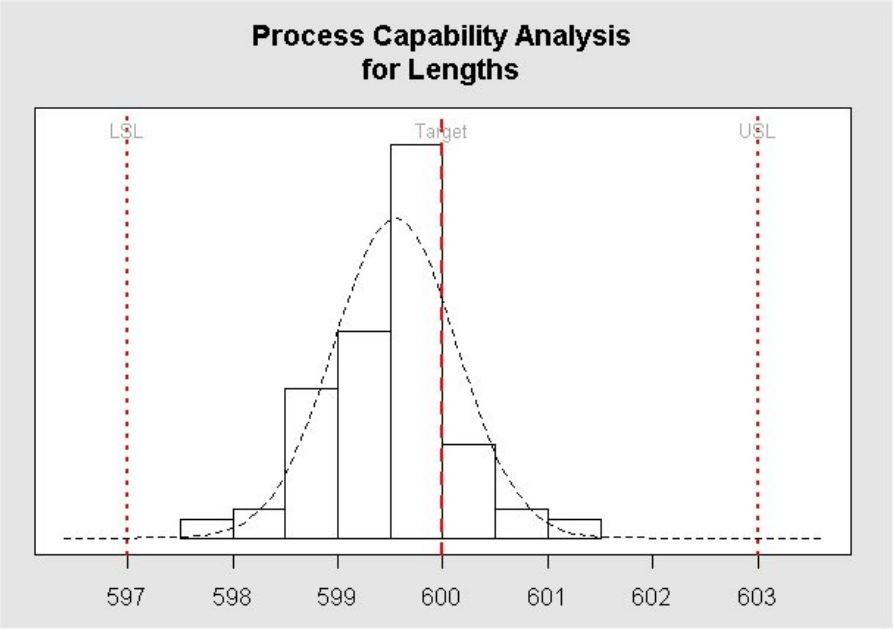
An automobile assembly plant concerned about quality improvement measured sets of five camshafts on twenty occasions throughout the day. The specifications for the process state that the design specification limits at 600±3mm.

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%
```



Q27. 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	?	?			

Q28. 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

Q29- 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) (4×3 Marks) Discuss any four of these rules, and how they would be used to detect "out of control" processes. Support your answer with sketches.

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.

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

Q30. 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.

Q31. Theory for Inference Procedures (3 Marks)

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

i. Briefly describe how $p-$value is used in hypothesis testing

ii. What is meant by a Type I error?

iii. What is meant by a Type II error?

Q32. Normal Distribution (6 Marks)

Assume that the diameter of a critical component is normally distributed with a Mean of 50mm and a Standard Deviation of 2mm. You are required to estimate the approximate probability of the following measurements occurring on an individual component.

i. (2 Marks) Between 50 and 51.2mm

ii. (2 Marks) Less than 48.5 mm

iii. (2 Marks) Between 48.2 and 51.6 mm

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

Q33. 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%.

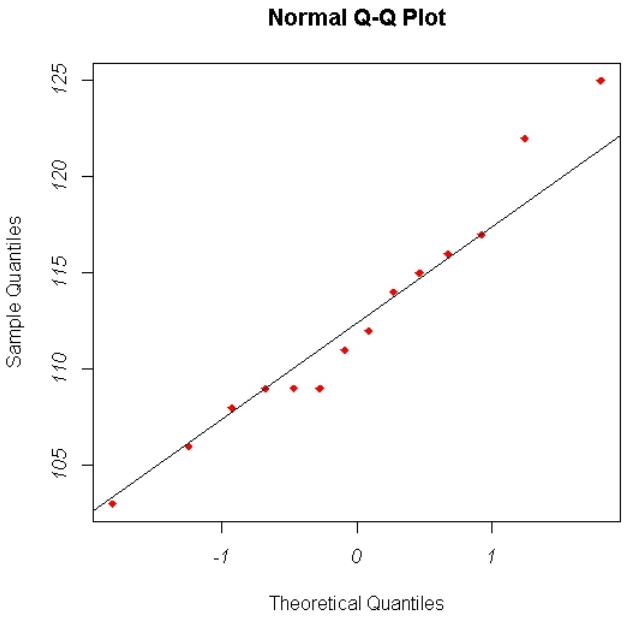
i. (1 Mark) Formally state the null hypothesis and the alternative hypothesis.

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.

Q34. 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.

Q35. 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.

Q36. 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.

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

Q37. 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.

Q38. Chi-Square Test (9 Marks)

Suppose you conducted a drug trial on a group of animals and you hypothesized that the animals receiving the drug would show increased heart rates compared to those that did not receive the drug. You conduct the study and collect the following data:

	Heart Rate Increased	No Heart Rate Increase	Increase
Treated	36	14	50
Not Treated	30	25	55
	66	39	105

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. (2 Marks) Discuss your conclusion to this test, supporting your statement with reference to appropriate values.