
# FACULTY OF SCIENCE AND ENGINEERING

## DEPARTMENT OF MATHEMATICS AND STATISTICS

# MID-SEMESTER ASSESSMENT

MODULE CODE: MA4128

SEMESTER: Spring

MODULE TITLE: Advanced Data Modeling

DURATION OF EXAM: 1 hour

LECTURER: Mr. Kevin O'Brien

GRADING SCHEME: 100 marks
20% of module grade

## INSTRUCTIONS TO CANDIDATES

Scientific calculators approved by the University of Limerick can be used.
Formula sheet and statistical tables provided at the end of the exam paper.
Students must attempt any 4 questions from 5.

## Question 1. (10 marks) Distriutional Assumptions

(a) (10 Marks)

The data set `X` and `Y` are both assumed to be normally distributed. The Shapiro-Wilk test was carried out to assess whether or not this assumption is valid for data set `X`.

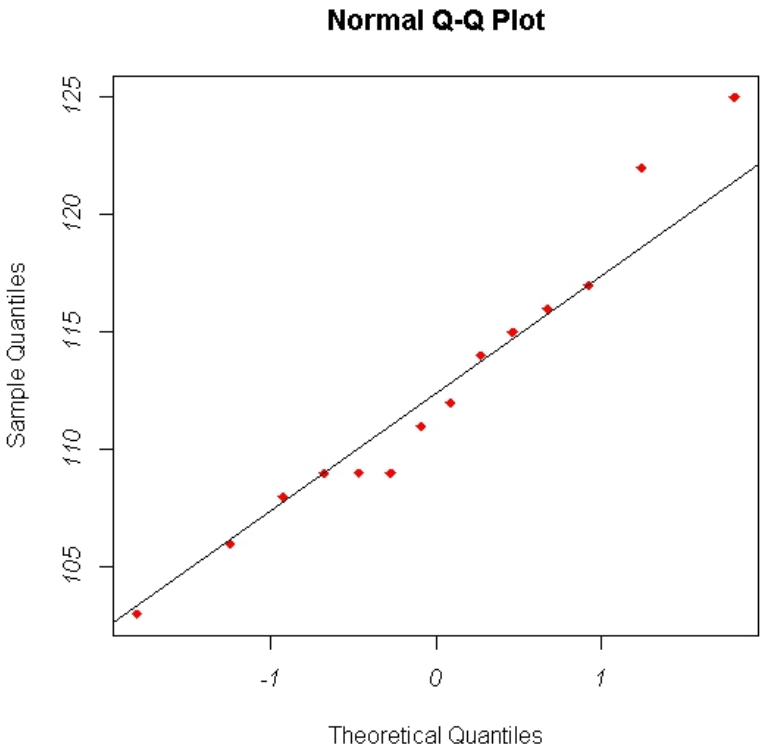
i. (1 marks) Formally state the null and alternative hypothesis.

ii. (2 marks) What is your conclusion for this procedure? Justify your answer.

```
> shapiro.test(X)

Shapiro-Wilk normality test
data:  X
W = 0.9292, p-value = 0.372
```

(c) The data set `X` and `Y` are both assumed to be normally distributed. 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. (2 marks) Provide a brief description on how to interpret this plot.

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

(b) 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:

| | | | | |
|---|---|---|---|---|
| 149 | 146 | 122 | 142 | 153 |
| 137 | 161 | 156 | 170 | 159 |

Use the Dixon Q-test to determine if there is an outlier present in this data. You may assume a significance level of 5%.

(i.) (1 Mark) State the null and alternative Hypothesis for this test.

(ii.) (2 Marks) Compute the test statistic

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

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

## Question 2. (10 marks) Binary Classification

(a) ***Binary Classification (6 Marks)***
For following binary classification outcome table, calculate the following appraisal metrics.

(i) (1 Mark) accuracy;

(ii) (1 Mark) recall;

(iii) (1 Mark) precision;

(iv) (1 Mark) F-measure.

| | Predict Negative | Predict Positive |
|---|---|---|
| Observed Negative | 9530 | 10 |
| Observed Positive | 300 | 160 |

(v) (2 Marks) Explain why the F-measure is considered a more informative measure of performance than the Accuracy score.

iii. (2 Marks) Define Specificity and Sensitivity. You make reference to previous answers.

iv. (3 Marks) What is a ROC curve? Explain its function, how it is determined, and the means of interpreting the curve. Support your answer with a sketch.

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

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

i. (4 marks) Write the regression equation for both of the fitted models.

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

iii. (3 marks) Discuss an alternative approach for this analysis, mentioning any disadvantages in using this alternative approach.

## Question 3. (10 marks) Hierarchical Clustering

i. (2 Marks) What is the purpose of a cluster analysis?

ii. (2 Marks) A discriminant analysis is similar to a cluster analysis; however, there is one fundamental difference. Explain this difference.

iii. (2 Marks) What is the difference between a linkage method and a distance measure?

iv. (2 Marks) Compare and contrast any two linkage methods.

v. (2 Marks) How do we determine the appropriate number of clusters? Give two different visualization methods that are used to display the outcome of a cluster analysis.

vi. (2 Marks) Standardization

vii. (2 Marks) Explain the difference between Ward's method and k-means clustering.

viii. (2 Marks) In the context of hierarchical cluster analysis, distinguish between agglomerative clustering and divisive clustering.

ix. (2 Marks) What is a vertical icicle plot used for? Give a brief description, supporting your answer with sketches.

x. (2 Marks) Compute the Euclidean distance between the following points.

$$A = (4, 6, 8, 2)$$

$$B = (3, 6, 1, 6)$$

(b) A normally distributed quality characteristic is monitored through the use of control charts. These charts have the following parameters. All charts are in control.

| | LCL | Centre Line | UCL |
|---|---|---|---|
| $\bar{X}$-Chart | 542 | 550 | 558 |
| $R$-Chart | 0 | 8.236 | 16.504 |

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

ii. (2 marks) Estimate the standard deviation of this process.

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

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

i. (4 marks) Determine the *Process Capability Indices* $C_p$ and $C_{pk}$, commenting on the respective values. You may use the `R` code output on the following page.

ii. (2 marks) The value of $C_{pm}$ is 1.353. Explain why there would be a discrepancy between $C_p$ and $C_{pm}$.

iii. (2 marks) Comment on the graphical output of the *Process Capability Analysis*, also presented on the next page.

```
Process Capability Analysis

Call:
process.capability(object = obj, spec.limits = c(597, 603))
Number of obs = 100          Target = 600
       Center = 599.548         LSL = 597
       StdDev = 0.5846948       USL = 603

Capability indices:
      Value   2.5%  97.5%
Cp    ...
Cp_l  ...
Cp_u  ...
Cp_k  ...
Cpm   1.353  1.134  1.572
Exp<LSL 0%   Obs<LSL 0%
```

## Question 4. (10 marks) K-Means Clustering

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

i. (2 marks) levels of a factor.

ii. (2 marks) randomized block design.

**ANOVA**

| | Cluster | | Error | | | |
|---|---|---|---|---|---|---|
| | Mean Square | df | Mean Square | df | F | Sig. |
| Net profit | 495,145 | 3 | 1419,744 | 3 | ,349 | ,795 |
| Own funds | 2878,202 | 3 | 2537,200 | 3 | 1,134 | ,460 |
| Assets | 842788,443 | 3 | 9987,138 | 3 | 84,387 | ,002 |
| Client deposits | 634017,636 | 3 | 35643,498 | 3 | 17,788 | ,021 |
| Loans | 957411,333 | 3 | 37401,709 | 3 | 25,598 | ,012 |

Figure 1:

(b) 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.

(c) 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
```

## Question 5. (10 marks) Modelling Count Variables

(i)

(ii)

(iii)

(iv)

(v)

# Formulas and Tables

## Critical Values for Dixon Q Test

| N | $\alpha = 0.10$ *Confidence*= 0.90 | $\alpha = 0.05$ *Confidence*= 0.95 | $\alpha = 0.01$ *Confidence*= 0.99 |
|---|---|---|---|
| 3 | 0.941 | 0.97 | 0.994 |
| 4 | 0.765 | 0.829 | 0.926 |
| 5 | 0.642 | 0.71 | 0.821 |
| 6 | 0.56 | 0.625 | 0.74 |
| 7 | 0.507 | 0.568 | 0.68 |
| 8 | 0.468 | 0.526 | 0.634 |
| 9 | 0.437 | 0.493 | 0.598 |
| 10 | 0.412 | 0.466 | 0.568 |
| 11 | 0.392 | 0.444 | 0.542 |
| 12 | 0.376 | 0.426 | 0.522 |
| 13 | 0.361 | 0.41 | 0.503 |
| 14 | 0.349 | 0.396 | 0.488 |
| 15 | 0.338 | 0.384 | 0.475 |
| 16 | 0.329 | 0.374 | 0.463 |

**Factors for Control Charts**

| Sample Size (n) | c4 | c5 | d2 | d3 | D3 | D4 |
|---|---|---|---|---|---|---|
| 2 | 0.7979 | 0.6028 | 1.128 | 0.853 | 0 | 3.267 |
| 3 | 0.8862 | 0.4633 | 1.693 | 0.888 | 0 | 2.574 |
| 4 | 0.9213 | 0.3889 | 2.059 | 0.88 | 0 | 2.282 |
| 5 | 0.9400 | 0.3412 | 2.326 | 0.864 | 0 | 2.114 |
| 6 | 0.9515 | 0.3076 | 2.534 | 0.848 | 0 | 2.004 |
| 7 | 0.9594 | 0.282 | 2.704 | 0.833 | 0.076 | 1.924 |
| 8 | 0.9650 | 0.2622 | 2.847 | 0.82 | 0.136 | 1.864 |
| 9 | 0.9693 | 0.2459 | 2.970 | 0.808 | 0.184 | 1.816 |
| 10 | 0.9727 | 0.2321 | 3.078 | 0.797 | 0.223 | 1.777 |
| 11 | 0.9754 | 0.2204 | 3.173 | 0.787 | 0.256 | 1.744 |
| 12 | 0.9776 | 0.2105 | 3.258 | 0.778 | 0.283 | 1.717 |
| 13 | 0.9794 | 0.2019 | 3.336 | 0.770 | 0.307 | 1.693 |
| 14 | 0.9810 | 0.1940 | 3.407 | 0.763 | 0.328 | 1.672 |
| 15 | 0.9823 | 0.1873 | 3.472 | 0.756 | 0.347 | 1.653 |
| 16 | 0.9835 | 0.1809 | 3.532 | 0.750 | 0.363 | 1.637 |
| 17 | 0.9845 | 0.1754 | 3.588 | 0.744 | 0.378 | 1.622 |
| 18 | 0.9854 | 0.1703 | 3.64 | 0.739 | 0.391 | 1.608 |
| 19 | 0.9862 | 0.1656 | 3.689 | 0.734 | 0.403 | 1.597 |
| 20 | 0.9869 | 0.1613 | 3.735 | 0.729 | 0.415 | 1.585 |
| 21 | 0.9876 | 0.1570 | 3.778 | 0.724 | 0.425 | 1.575 |
| 22 | 0.9882 | 0.1532 | 3.819 | 0.720 | 0.434 | 1.566 |
| 23 | 0.9887 | 0.1499 | 3.858 | 0.716 | 0.443 | 1.557 |
| 24 | 0.9892 | 0.1466 | 3.895 | 0.712 | 0.451 | 1.548 |
| 25 | 0.9896 | 0.1438 | 3.931 | 0.708 | 0.459 | 1.541 |