University of Stirling MATPMD1

Computing Science & Mathematics 2021

**MATPMD1 : STATISTICS FOR DATA SCIENCE**

**Specimen Class Test**

**This specimen contains 10 multiple-choice questions of the type you can expect to meet in the MATPMD1 Class Test.**

**NOTE, in the actual test there will be between 20 and 25 questions. You will be expected to complete the test in 50 minutes.**

**BELOW ARE THE INSTRUCTIONS THAT YOU WILL BE GIVEN WITH THE ACTUAL CLASS TEST**

**This is the MATPMD1 Class Test which is to be taken as a Quiz on Canvas.**

**This is an OPEN-BOOK class test and you may consult your notes, but you may not consult any online resource, except for CANVAS.**

**You may use an installation of R on your personal computer.**

**You should complete this test under exam conditions.**

**YOU SHOULD TALK TO NO ONE AND SHOULD WORK ALONE.**

**You should attempt ALL questions.**

**Statistical tables are provided.**

Q1 What is the median value of 23.7, 22.1, 56.8, 31.3, 54.6, and 23.2?

|  |  |  |
| --- | --- | --- |
| A | 27.5 | 1 |
| B | 23.7 |  |
| C | 31.3 |  |
| D | 35.3 |  |

Q2 Which description fits the following data?

Chart, box and whisker chart

Description automatically generated

|  |  |  |
| --- | --- | --- |
| A | Bimodal, positively skewed, no gaps, two outliers. |  |
| B | Unimodal, negatively skewed, with gaps, one outlier. |  |
| C | Unimodal, positively skewed, with gaps, two outliers. | 1 |
| D | Unimodal, negatively skewed, no gaps, two outliers. |  |

Q3 For data that is normally distributed with mean 50, variance 36 which of the following is true?

|  |  |  |
| --- | --- | --- |
| A | 95% of the data lies between 44 and 56. |  |
| B | More than 5% of the data lies outside the interval between 14 and 86. |  |
| C | 95% of the data lies between 38 and 62 | 1 |
| D | Over 95% of the data lies between 50 and 62 |  |

Q4 Data is collected on the exam results of school leavers from 2021. This data is stored in a table indexed by student, with the data held in a separate column logging the alphabetical grade attained, graded from A through to F. No star grades are attained under this system. How can the variable that classifies the grade for the students be described?

|  |  |  |
| --- | --- | --- |
| A | Quantitative variable taking the values A to F. |  |
| B | Nominal Categorical variable taking the values A to F |  |
| C | Ordinal Categorical variable taking the values A to F | 1 |
| D | Quantitative variable taking the values 1 to 6. |  |

Q5 What is the rejection region for a one-sample two-sided t-test performed on a sample of size 19 with and at 1% significance level?

|  |  |  |
| --- | --- | --- |
| A |  |  |
| B |  |  |
| C |  |  |
| D |  | 1 |

Q6 A city GP practice undertook a study to see whether their patients would prefer a first-come first served system i.e. no appointments. As part of the study they examined whether the patient’s view and their gender were associated. A random sample of patients was questioned. The results are given below:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | View of a ‘No Appointments’ system | | |
|  |  | For | Against | No Preference |
| Gender | Male | 106 | 27 | 20 |
|  | Female | 97 | 166 | 34 |

Which of the following describes the conclusion of a chi-squared test for association between gender and view?

|  |  |  |
| --- | --- | --- |
| A | Observed test statistic (16.94) is greater than 3.84 so there is sufficient evidence that there is an association between gender and view |  |
| B | Observed test statistic (64.7) is greater than 5.99 so there is sufficient evidence that there is an association between gender and view. | 1 |
| C | Observed test statistic (16.94) is greater than 5.99 so there is insufficient evidence to suggest that there is an association between gender and view. |  |
| D | Observed test statistic (64.7) is greater than 3.84 so there is sufficient evidence that there is an association between gender and view. |  |

Q7 While preparing the end of year financial statements a chain of stores examined the relationship between profits and the size of ten of their stores. The data are presented below.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Store, i | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| Size  (000s m2) | 6.0 | 1.4 | 5.0 | 1.0 | 1.2 | 1.5 | 3.5 | 1.3 | 2.1 | 1.3 |
| Profit  (£000s) | 42.1 | 6.3 | 38.5 | -0.3 | 3.6 | 7.8 | 24.4 | 2.1 | 17.5 | 7.2 |

If a new store of size 2000m2 is developed, using linear regression to model the expected profit by store size which of the following statements holds true?

|  |  |  |
| --- | --- | --- |
| A | There is a significant linear regression relationship between store size and profit. The predicted profit for the new store is £11.30. |  |
| B | There is a significant linear regression relationship between store size and profit. The predicted profit for the new store is £16,846. |  |
| C | There is a significant linear regression relationship between store size and profit. The predicted profit for the new store is £11,297. | 1 |
| D | There is not a significant linear regression so no conclusion can be inferred. |  |

Q8 Given the R output from the following linear regression, what is the value of the response variable when the independent variable is zero.

Residuals:

Min 1Q Median 3Q Max

-35.207 -0.051 3.996 8.346 14.225

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -20.618 8.755 -2.355 0.0463 \*

Q8var1[1:10] 4.489 1.478 3.037 0.0161 \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.82 on 8 degrees of freedom

Multiple R-squared: 0.5355, Adjusted R-squared: 0.4775

F-statistic: 9.223 on 1 and 8 DF, p-value: 0.01614

|  |  |  |
| --- | --- | --- |
| A | There is a significant linear regression relationship, and the value of the response variable is -2.355 when the independent variable is zero. |  |
| B | There is a significant linear regression relationship, and the value of the response variable is -20.618 when the independent variable is zero. | 1 |
| C | There is a significant linear regression relationship, and the value of the response variable is 4.489 when the independent variable is zero. |  |
| D | There is not a significant linear regression relationship so no conclusion can be inferred. |  |

Q9 What is the null Hypothesis tested by the following R output, and what is the conclusion of the test?

Anderson-Darling normality test

data: specd$Q7y

A = 0.67902, p-value = 0.05232

|  |  |  |
| --- | --- | --- |
| A | H0: the data is normally distributed. As p-value>0.05 there is sufficient evidence to say that the data is not normally distributed. |  |
| B | H0: the data is normally distributed. As p-value>0.05 there is sufficient evidence to say that the data is not normally distributed. |  |
| C | H0: the data is not normally distributed. As p-value>0.05 there is not sufficient evidence to say that the data is not normally distributed. |  |
| D | H0: the data is normally distributed. As p-value>0.05 there is not sufficient evidence to say that the data is not normally distributed. | 1 |

Q10 What is a valid conclusion to draw from the following Multiple Linear Regression output?

Residuals:

Min 1Q Median 3Q Max

-3.2986 -2.4135 0.2623 1.5277 5.3648

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -5.560039 2.444727 -2.274 0.0571 .

specd$Q10x1[1:10] 8.425998 0.579600 14.538 1.74e-06 \*\*\*

specd$Q10x2[1:10] 0.002284 0.755407 0.003 0.9977

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.086 on 7 degrees of freedom

Multiple R-squared: 0.9682, Adjusted R-squared: 0.9592

F-statistic: 106.7 on 2 and 7 DF, p-value: 5.705e-06

|  |  |  |
| --- | --- | --- |
| A | There is a significant multiple linear regression relationship which can be modelled as  . |  |
| B | There is a significant multiple linear regression relationship which can be modelled as  . |  |
| C | There is not a significant multiple linear regression relationship so no conclusion can be inferred. |  |
| D | There is not a significant relationship between the response variable and the independent variable We should remove this variable from the regression before extracting model coefficients. | 1 |