2001/161.111

MTUI DISD

**Student Name: Braden Johnston Student ID: 20005898**

**MASSEY UNIVERSITY**

**MANAWATU AND DISTANCE CAMPUSES**

**EXAMINATION FOR**

**161.111 APPLIED STATISTICS**

**SEMESTER ONE 2020**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Time allowed is **TWO (2) Hours**.

Answer **ALL** questions.

Questions carry different marks, as shown at the bottom of the page.

Type your answers in the relevant section of the Question paper in the area specified.

Type your name and ID number in the spaces provided at the top of this page.

Calculators are permitted with no restrictions.

The exam paper will NOT be made available on the University Library website.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Lecturer Use Only**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Section** | **A** | **B** | **C** | **D** | **Total** |
| **Mark** | 12 | 12 | 16 | 10 | 50 |
|  |  |  |  |  |  |

**PART A [12 marks]**

An experiment was carried out to compare recall ability when reading electronically compared to hard copy. Thirty students from a first-year university English class agreed to participate. Each student was shown a randomly generated list of 40 five-letter words on a tablet and given 2 minutes to memorise them. They were then given 4 minutes to write down as many of the words as they could recall. An hour later, the students repeated the experiment but this time the list of 40 five-letter words was given to them on a piece of paper. In each case, the percentage of correctly recalled words was recorded. The aim is to see if there is a difference in the percentage of words correctly recalled between the electronic list and the paper list.

RStudio was used to do a two-sample t-test and a t-test of differences on the data. The output from both these tests is shown below.

|  |  |
| --- | --- |
| Two-sample t-test: | A screenshot of a cell phone  Description automatically generated |
|  |  |
| t-test of differences: | A picture containing bird  Description automatically generated |

1. Which test is appropriate for this experiment? Explain. [2 marks]

|  |
| --- |
| A t-test of differences would be more appropriate for this experiment. A two-sample t-test is used to compare two groups that are independent to each other, whereas in this experiment, the sample of students remains the same, and the difference in scores are being recorded. |

1. State the null hypothesis for this test. [1 mark]

|  |
| --- |
| The true mean amount of words that university students could recall from paper is equal to the amount of words that they could recall from reading electronically. |

*Part A continues over the page…*

*Part A continued….*

1. What is the value of the test statistic? [1 mark]

|  |
| --- |
| 2.207 |

1. State the decision for the hypothesis test with reason. [1 mark]

|  |
| --- |
| Because the p-value is less than 0.05, we are likely to reason with the alternative hypothesis. |

1. State the conclusion for the hypothesis test. [1 mark]

|  |
| --- |
| We have evidence to suggest that the true mean amount of words that university students can recall if read from paper is not equal to the amount of words that they could recall if read from an electronic device. |

1. Are the conditions met? Explain. [2 marks]

|  |
| --- |
| The sampling distribution must be normal. The sample size in this experiment is 30, so the Central Limit Theorem (CLT) holds. The sample also needs to be representative of the population. It is hard to say with certainty that the students are representative because nowhere does it say that the students were selected using a random sampling method and instead says that they merely ‘agreed’ to take part. |

1. Explain how the confidence interval in the relevant RStudio output adds to your conclusion.

[2 marks]

|  |
| --- |
| The CI adds to the conclusion by telling us that university students can recall more words from paper then they can electronically, as opposed to the same amount. The mean difference being somewhere between 0.65 words and 16.8 words. |

1. The experimental design was not ideal. Explain how it could have been improved. [2 marks]

|  |
| --- |
| The experiment could have been more accurate if two separate groups of students were used, though this would require finding another 30 students to keep the CLT, the students were required to read the same words and recall them twice also, this means that they could remember more of the words on paper than they could on electronic devices because they have read the set of words twice. If I was to conduct this experiment, I would use two samples of students, one set that did the electronic first and then the hard copy, with the other set doing the hard copy first and then the electronic version. |

**PART B [12 marks]**

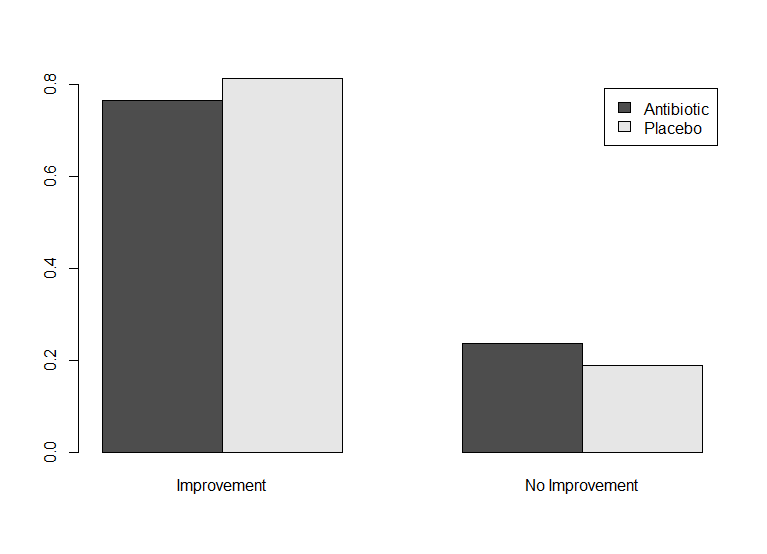
Researchers carried out a study into the effect of antibiotic treatment for sinusitis. They randomly assigned 245 adults diagnosed with sinusitis into two groups. Those in the Antibiotic group received a 10-day course of an antibiotic pill; the rest of the participants received a placebo pill. The placebo pills had the same taste and packaging as the antibiotic. At the end of the 10-day period, patients were asked if their symptoms had improved since they started taking the pills.

1. Explain how this study makes use of blinding. [1 mark]

|  |
| --- |
| Blinding in this case is referring to the concealment of allocation into groups, and not being told if they are being given a placebo pill, which essentially means that the pill does nothing, and as long as the group thinks that they are taking a real antibiotic, their brain might treat the sinusitis on it’s own. |

The table and graph below summarise the results.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | Self-reported improvement in symptoms | |  |
|  |  | Improvement | No Improvement | Total |
| Treatment | Antibiotic | 94 | 29 | 123 |
| Placebo | 99 | 23 | 122 |
|  | Total | 193 | 52 | 245 |



1. What do the table and graph tell you about the effect of antibiotic treatments for sinusitis? [2 marks]

|  |
| --- |
| The table tells us that a very large portion of the sample reported having improved in symptoms. While the graph is a graph of counts, which shows proportions enabling us to compare the amount of improvement in each randomly selected group, with the placebo pill patients having recovered more on average than the patients that were receiving the actual antibiotic, marginally. |

*Part B continues over the page…*

*Part B continued….*

The researchers used RStudio to do a Chi-squared test on the data. Below is the RStudio output for this test along with the expected counts:

A picture containing bird

Description automatically generated

1. State the null hypothesis for the test. [2 marks]

|  |
| --- |
| Sinusitis patients that are given antibiotics will not recover at the same speed as patients who are given placebo pills |

1. State the value of the test statistic. [1 mark]

|  |
| --- |
| 0.8177 |

1. State the decision for the hypothesis test with reason. [1 mark]

|  |
| --- |
| Because the p-value is higher than 0.05, we trust the null hypothesis in this instance. |

1. State the conclusion for the hypothesis test. [1 mark]

|  |
| --- |
| There is evidence to suggest that adults with Sinusitis that are given antibiotics do not recover at the same speed as those that are given placebo pills. |

1. Are the conditions met? Justify your answer. [2 marks]

|  |
| --- |
| The sample size is 245, which is much larger than 30, so the CLT holds.  The sample took those 245 adults and randomly assigned them into groups, which tells me that the sample is representative.  The smallest expected count is 25.89, which is higher than 5, so this condition is also met. |

1. A doctor questions the need for the Chi-squared test. He says the conclusion is obvious from looking at the table of the sample data. Do you agree? Explain. [2 marks]

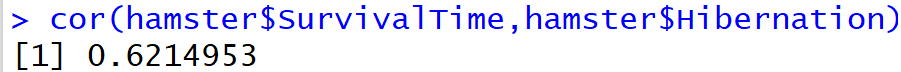
|  |
| --- |
| I agree with the doctor about the need for a chi-squared test in this experiment. While only marginal, the graph shows us that there were more cured sinusitis patients who took the placebo pills compared to those who took antibiotics in this sample. |

**PART C [16 marks]**

A group of biologists wanted to see if the time a hamster spends hibernating is related to its survival time. They recorded the survival time (in days) and the time spent hibernating (in days) for a simple random sample of 25 hamsters.

The data is shown in the scatterplot below, along with the correlation coefficient.





1. What does the scatterplot and the correlation coefficient tell us about the relationship between survival time and time spent hibernating? [2 marks]

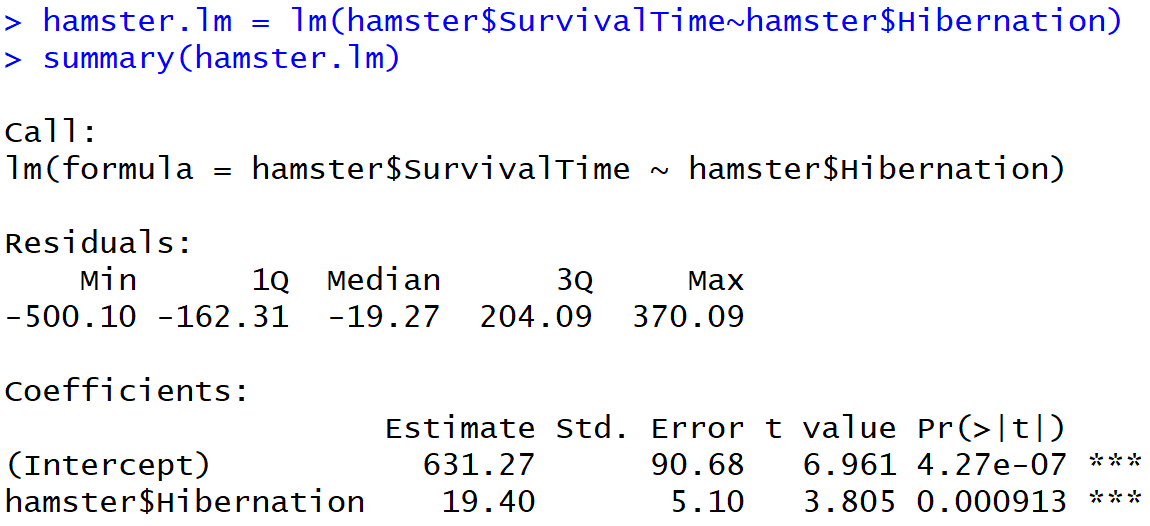
|  |
| --- |
| There appears to be a moderately strong, linear relationship between the amount of time that a hamster spends hibernating and the amount of time which that hamster survives. |

*Part C continues over the page…*

*Part C continued…*

RStudio was used to fit a linear model to the data.

Below are the RStudio output, the residual plot and a histogram of the residuals.





*Part C continues over the page..*

*Part C continued…*

1. Write down the linear model equation. [1 mark]

|  |
| --- |
| Survival time = 631.27 + 19.4 x Hibernation time |

1. What is the slope and what does it tell us? [2 marks]

|  |
| --- |
| The slope of 19.4 tells us that, on average, the survival time of a hamster is expected to increase by ~19 days for each day that that hamster hibernates. |

1. What change in survival time do we expect for an extra half day of hibernating? [1 mark]

|  |
| --- |
| 19.4/2 = 9.7. We would expect a hamster to survive for an extra 9.7 days if there was an extra half day of hibernating |

1. One of the hamsters spent 14 days hibernating and survived for 1132 days. According to the linear model its expected survival time is 902.87 days. Calculate the residual for this hamster.

[1 mark]

|  |
| --- |
| Expected survival time = 631.27 + 19.4 x 14 = 902.87  Residual = 1132 – 902.87 = 229.13 |

1. Are the conditions of the linear model met? Explain. [2 marks]

|  |
| --- |
| Linearity is met, as the residual plot shows no trend.  Equal spead is met, as the residuals are equally spead above and below the line through zero on the residual plot  The histogram looks to have a left skew, instead of a normal distribution. So this condition is not met.  Indepenence is met, as the hamsters were randomly selected. |

1. What is the null hypothesis of the test for the slope? [1 mark]

|  |
| --- |
| True slope = 0 |

1. What is the p-value for this test? [1 mark]

|  |
| --- |
| 0.000913 |

*Part C continues over the page…*

*Part C continued…*

1. What is the conclusion for this test? [1 mark]

|  |
| --- |
| There is evidence of a linear relationship between the amount of time that a hamster spends hibernating and the amount of time in which that hamster survives in total. |

1. The study above did not take account of the fact that there are two common species of hamsters: golden hamster and dwarf hamsters. These two species differ in lifespan as well as in their hibernation behaviour. Golden hamsters are the most common, occurring in twice the numbers that dwarf hamsters do. The biologists want to repeat the study on 100 hamsters selected by stratified sampling. How many of the golden hamsters should they select?

[2 marks]

|  |
| --- |
| It would be appropriate to have a total sample of 99, with 66 golden hamsters and 33 dwarf hamsters. |

1. Why would a cluster sampling method not be appropriate here? Explain. [2 marks]

|  |
| --- |
| Using cluster sampling in this experiment would be biased because the clusters are of different sizes, with the golden hamsters being twice as common as the dwarf hamsters. |

**PART D [10 marks]**

A study was carried out to investigate if poets die young. Data was gathered on the age at death of North American female writers. Three categories of writers were included: poets, novelists and nonfiction writers. The data are summarised in the plot below.



1. What does the graph tell you? [2 marks]

|  |
| --- |
| North American female non-fiction writers have a higher mean lifespan compared to novelists from the same group. While poets appear to have the lowest mean lifespans. Though, poets and novelists have a much higher spread of lifespans than nonfiction writers. |

*Part D continues over the page..*

*Part D continued…*

A hypothesis test was carried out to compare the three population means. The hypotheses for this test are:

**The null:** The mean age at death is the same for each type of writer.

**The alternative:** The mean age at death for at least one of the types of writer is different to the others.

The test gave a p-value of 0.00197.

95% confidence intervals for the differences between each pair of writers are given in the table below.

|  |  |  |
| --- | --- | --- |
| Comparison | Lower limit (years) | Upper limit (years) |
| Novelist – Nonfiction writer | -13.59 | 2.74 |
| Poet – Nonfiction writer | -22.95 | -4.42 |
| Poet - Novelist | -15.63 | -0.89 |

1. State the decision for the hypothesis test with reason. [1 mark]

|  |
| --- |
| Because the p-value is much lower than 0.05 which means we side with the alternative hypothesis. |

1. State the conclusion for the hypothesis test. [1 mark]

|  |
| --- |
| There is evidence to suggest that the true mean age at death for at least one of the types of writer is different to the others. |

1. Explain how the confidence intervals add to this conclusion. [2 marks]

|  |
| --- |
| The confidence interval confirms the conclusion of there being a difference  in the mean age at death for different writers based on the genre they write, because the interval doesn’t include 0 in all cases, only in the case of the nonfiction novelist. |

1. This study was reported with the headline ‘Female poets die young!’. Is this headline justified by the study? Explain. [2 marks]

|  |
| --- |
| This headline is not justified by the study. While it can be said that North American female poets die younger than other North American female writers on average, they still have a fairly long life, and would not be considered to die young. Also, this study is only taking North American writers into account, and not female writers from all backgrounds, and as such this headline is only being used to gain attention for the study, and doesn’t necessarily have any traction. |

1. A young male Australian writer reads the study and decides not to become a poet because he wants to live to a ripe old age. Discuss two problems with his thinking. [2 marks]

|  |
| --- |
| There are a few things to take into account. While this report could seem daunting to writers, we must remember that there are other lurking variables that contribute to an earlier death for North American female poets. There is a great chance that poets have a different lifestyle to novelists and it is for this reason that they could die younger. This young Australian writer should feel free to pursue his dreams, as he is also not a North American female, infact nearly the opposite. The statistics recorded in this survey should not be applied to anyone except North American female writers. |

+ + + + + + + +