Faculty of Engineering and Information Sciences

**School of Mathematics and Applied Statistics**

http://eis.uow.edu.au/smas/index.html

# STAT251 – Fundamentals for Biostatistics

## Spring Session 2020

**Assignment 3**

**Due Date: 6th November, 10pm (Friday, Week 12)**

**This Assignment is worth 15% of the final mark in STAT251.**

*Submission Format: Enter your name and student number on page 2 then enter your answers beneath each question in this Word file. Create a pdf version of your final Answers file (without page 1).   
Submit the pdf file to the Moodle Dropbox by the due date and time.*

**STAT251 Assignment 3 NAME:**

**Student Number:**

1. A new protocol for care of patients following their first myocardial infarction (MI) is proposed.

A study is conducted to directly compare the standard care to the new protocol using a randomised trial. The number of patients under each type of care and the number who were readmitted to hospital within 6 weeks are recorded in the table below.

|  |  |  |
| --- | --- | --- |
| Treatment | Number of patients | Number readmitted to hospital within 6 weeks |
| Standard Care | 125 | 16 |
| New Protocol | 125 | 11 |

1. Based on the given information, is there a significant reduction in the readmission rate under the new protocol?

Carry out an appropriate hypothesis test (by hand) showing all steps.

[*Hint:* Let be the population proportion of patients under the Standard Care who were readmitted to hospital within 6 weeks.

*Let* be the population proportion of patients under the New Protocol who were readmitted to hospital within 6 weeks.]

1a. /6

1. Calculate a 95% confidence interval for given the information in the table above. Write a statement to interpret the interval in context.

1b. /3

1. i. State the margin of error of the 95% confidence interval for in terms of a percentage.

1c.i. /1

1. If a follow-up study required that the estimate of be determined within 1.5% (instead of what was obtained in c.(i)) with 95% confidence, determine the sample   
   size required for the follow-up study.

1c.ii. /2

1. Comment on your result in c(ii) with reference to the sample size of *n*=125 used  
    in the original study.

1c.iii. /1

Question 1 total =13 marks

1. An individual's critical flicker frequency is the highest frequency at which the flicker in a flickering light source can be detected. At frequencies above the critical frequency, the light source appears to be continuous even though it is actually flickering. This investigation recorded critical flicker frequency (in cycles/sec) and iris colour of the eye for 19 subjects.   
   The data is in the SPSS file *Flicker.sav*.

Data source: <http://www.statsci.org/data/general/flicker.html>

We would like to determine if the mean critical flicker frequency is different for different eye colours.

*In your answers to the following questions, you may paste the results from SPSS into your answer but you must refer to and interpret the relevant output in your answer.*

* 1. Create a side-by-side boxplot and a table of the means and standard deviations of the critical flicker frequencies given eye colour. Paste the results here and comment on each.

2a. /4

* 1. Use an appropriate statistical test to test if there is sufficient evidence to believe that the population mean critical flicker frequency is different for different eye colours.

Include all the steps as demonstrated in lectures. Use the notation as given in lectures.

2b. /6

* 1. What are the assumptions and requirements for inference using the technique in Part (b)? Do any of them appear to be violated? Which ones?

[*Hint: you will need to obtain the residuals*.]

2c. /5

* 1. Use the appropriate procedure to produce (experimentwise) 95% confidence intervals for the difference in population mean critical flicker frequency for each pair of eye colours, properly adjusted for multiple comparisons/testing. Which, if any, of the eye colours' mean critical flicker frequencies appear to be significantly different?

2d. /3

* 1. Perform an appropriate non-parametric statistical test to the test performed in (b).

State the hypotheses, paste in the SPSS results, and state test statistic, the degrees of freedom and the decision of the test.

2e. /6

* 1. Compare the p-values produced for the test in (b) and the test in (e). With regard to power, which test has the higher power? Explain you answer.

2f. /2

Question 2 total =26 marks

1. A new café owner adjacent to a large IT company is planning the café’s opening and would like to get an idea of milk choices for potential customers. A random sample of 250 IT company employees were emailed and asked about their milk preferences in coffee as well the brand of coffee they preferred (Vittoria or Campos). For those who drink milk coffee, the results are as follows:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Full Cream** | **Low Fat** | **Non-dairy (Soy / Almond)** | **Total** |
| **Campos** | 63 | 43 | 17 | 123 |
| **Vittoria** | 57 | 50 | 20 | 127 |
| **Total** | 120 | 93 | 37 | 250 |

1. Enter the data into SPSS. [*Hint: you will need to weight the cases].*Use SPSS todetermine the row percentages and paste the table here.

3a. /1

* + 1. Use the results in (a), to produce a stacked bar chart for these data and paste in below. [*See SPSS Hints*].

3b.i. /2

* + 1. What does it suggest about the independence of the preferences   
       for milk and coffee brand?

3b.ii. /2

1. Assuming these data represent a random sample of ‘coffee with milk’ drinkers, carry out the following.
   * 1. Produce the expected frequencies for each cell and paste the table of output here.

3c.i. /1

* + 1. Confirm, by hand, the value of the expected frequency for the Campos and   
       Soy milk cell, showing all working.

3c.ii. /1

* + 1. Test the hypothesis that preferences for milk and coffee brand are statistically independent, showing all steps.   
       As part of your answer, paste in your SPSS output into the relevant step.   
       You may use the *p*-value from the SPSS output to make your decision, given a level of significance of 5%.

3c.iii. /6

Question 3 total =13 marks

1. A researcher is interested in the relationship between birth weight (in kg) of newborns and gestational age at birth (weeks). Perform a regression analysis for these data to predict birth weight (in kg) from gestational age at birth (weeks).

Download the dataset *Birthweight.sav* on moodle and use the variables named *Birthweightkg* and *Gestation*.

Dataset is adapted from <https://www.sheffield.ac.uk/mash/statistics/datasets> [www.statstutor.ac.uk](http://www.statstutor.ac.uk) by Ellen Marshall https://licensebuttons.net/l/by-nc-sa/3.0/88x31.png

In your answer, produce and interpret the following:

* 1. the scatterplot;

4a. /2

* 1. the correlation coefficient;

4b. /2

* 1. the regression table containing the proportion of variance explained;

4c. /2

* 1. the regression ANOVA table;

4d. /4

* 1. the coefficient table with 95% confidence intervals;   
     [*See SPSS Hints on last page*.]

4e. /3

* 1. the predicted versus residual plot;

4f. /2

* 1. the histogram and boxplot of residuals.

[*Hint:* *PLEASE MAKE SURE YOU WATCH THE LECTURE TO SEE HOW TO SAVE*

4g. /2

*THE RESIDUALS*].

* 1. Does the data appear to meet the assumptions of linear regression (discuss   
     your choice of answer).

4h. /4

Question 4 total =21 marks

**Assignment 3 Total = 13 + 26 + 13 +21 = 73 marks**

**SPSS Hints.**

**Q2** You will need to obtain and save the residuals.

SPSS: Analyze -> General Linear Model -> Univariate. And request the “Standardized” residuals.

**Q3b.** To create a stacked bar chart using the weighted data, carry out the following.

Yourfile should have three variables already: *Coffeecode*, *Milkcode* and *Frequency*.   
Create a new variable called *ProbCondCoffee* which will be the conditional probabilities of milk preference given coffee preference (these will be the row percentages determined in (a) entered into SPSS as proportions). You will need to enter the 6 values in by hand.

Then do Graphs -> Legacy dialogs -> Bar. Then choose the Stacked bar with Summaries for groups of cases. Enter the *Coffeecode* as the Category Axis and the *Milkcode* into the ‘Define stacks by’. Then put the Bars represent ‘Other statistic – mean’ as *ProbCondCoffee*.

This should produce the stacked bar chart. To change the order to have Full cream at bottom and Soy at top, double click to open the Chart editor, click on the bar and open the Properties box and find the Categories tab. You can rearrange the order to be Full cream, Low fat then Soy.

Change other properties such as the number of decimal places, and the scale increment as desired.

**Q4e.** To produce the CI make sure you click on the Statistics button when you are doing the regression and then click on the CI under regression coefficients.

