STAT6160 DATA ANALYTICS FOR BUSINESS

Assignment 2 Questions

Number of questions: 4, Total marks: 35, Weight: 25%
Due time and date: 11:59pm AEST, Sunday end of Week 11

Submission instructions and general marking criteria

- Prepare your assignment in WORD, Latex, R Markdown, or any other appropriate software system for document preparation.
- Submit a copy in **PDF** format via Canvas.
- Assignments submitted by any other means (e.g., email, scanned copy) will attract no marks.
- Late submission penalty is detailed in the Course Outline.
- It is expected that R is used to assist with calculations and preparation of appropriate graphs; however, using Python is also acceptable if you prefer. All relevant scripts and output must be included with your assignment. However, raw computer output without explanatory text is not acceptable. Answers must be written in clear English sentences linked to appropriate supporting computer output.
- The assessment requires you to apply concepts from Modules 5-6 to various scenarios/data sets and to write up the results of a statistical analysis.

Question 1 (Total 9 Marks)

Mean entry-level salaries for college graduates with electrical engineering degrees and mechanical engineering degrees are believed to be approximately the same. A recruiting office thinks that the mean electrical engineering salary is greater than the mean mechanical engineering salary. The recruiting office randomly surveys 80 entry-level electrical engineers and 100 entry-level mechanical engineers. The data is provided in the file "**Question1.csv**".

Conduct a hypothesis test to determine if you agree that the mean entry-level electrical engineering salary is greater than the mean entry-level mechanical engineering salary through the following six steps:

(i) What are the hypotheses to test whether the mean entry-level electrical engineering salary is greater than the mean entry-level mechanical engineering salary? Define the appropriate parameters. **[2 Marks]**

(ii) What are the assumptions that you should make in using this test? Are they reasonable in this case? **[2 Marks]**

(iii) What is the test statistic? **[1 Mark]**

(iv) What is the null distribution? **[1 Mark]**

(v) What is the p-value? **[1 Mark]**

(vi) Write a conclusion. **[2 Marks]**

Question 2 (Total 9 Marks)

A cheese manufacturer is seeking to determine whether two different kinds of packaging of a certain kind of cheese differ in their ability to protect cheese from mould. To do so, 80 equal-sized pieces of cheese are cut from one randomly selected block, 40 are packaged by Method A and 40 are packaged by Method B. At 40 different locations, one Method A pack and one Method B pack are left until mould appears. The results (in days) before mould appears are provided in the file "**Question2.csv**".

Implement the following six steps to test whether there is a difference, at a 5% significance level, between the abilities of the two different kinds of packaging to protect the cheese from mould.

(i) Define the parameters and state the null and alternative hypotheses. **[2 Marks]**

(ii) Check the assumptions for this hypothesis test. **[2 Marks]**

(iii) Find the test statistic. **[1 Mark]**

(iv) State the null distribution. **[1 Mark]**

(v) Calculate the p-value. **[1 Mark]**

(vi) Write the conclusion in plain language. **[2 Marks]**

Question 3 (Total 7 Marks)

A video game developer is testing a new game on three different groups. Each group represents a different target market for the game. The developer collects scores from a random sample from each group. The results are provided in the file "**Question3.csv**".

The developer would like to test if there is a difference between the population mean scores of the three different groups, at a 5% significance level. In retrospect, we assume that the scores of each group are approximately normally distributed.

(i) Define the appropriate parameters and write down the null and alternative hypotheses. **[2 Marks]**

(ii) What is the test statistic? **[1 Mark]**

(iii) What is the null distribution? **[1 Mark]**

(iv) What is the p-value? **[1 Mark]**

(v) Express your conclusion to the experiment. **[2 Marks]**

Question 4 (Total 10 Marks)

Imagine you are a Data Analyst at a retail company. The marketing team has experimented to understand the effectiveness of three different types of marketing campaigns for a new product launch. These campaigns are: a) Social Media Campaign; b) Email Marketing Campaign; and c) In-store Promotions.

The company randomly selected 45 stores nationwide, dividing them equally among the three campaign types (15 stores per campaign type). After a month of running these campaigns, the marketing team collected data on the increase in sales (percentage) for the new product in each store. Your task is to analyze the data to determine if there are statistically significant differences in the effectiveness of these marketing campaigns on sales increase percentages.

i) What assumptions about the data should be verified for ANOVA? How do you verify these assumptions with R? **[3 Marks]**

ii) Use R to calculate the p-value based on the data provided in "**Question4.csv**". What does this indicate about the effectiveness of the marketing campaigns? **[2 Marks]**

iii) *If* the ANOVA indicates significant differences, perform a post-hoc analysis to find out which campaign(s) differ. **[1 Marks]**

iv) Provide a brief explanation of how to interpret the results of this post-hoc test, and provide your conclusions regarding the effectiveness of the three marketing campaigns?. **[2 Marks]**

v) Discuss what is the potential pitfall of conducting a post-hoc test without first performing the ANOVA test. (Hint: think why we need ANOVA instead of perform multiple pairwise t-tests) **[2 Marks]**