Statistics - Final Assignment and Team Project

Announced: 2024/05/22
Due date: 2024/06/20

(有任何問題/不清楚的地方請與我或助教聯絡)

In this final personal assignment and team project, you can use any statistical software (excel, R, SPSS, Python, and others). Summarize and describe/explain the codes/steps and results (step-by-step) in two PDF files, one for personal assignment and one for the team project, and a video for the team project. Submit it to the E3 system by 2024/06/20 23:59 with no extension. Please be as detailed as possible in your response (Mandarin or English).

Important points:

- Please complete the assignment by yourself and work together with your teammate for the team project. assignment是個人作業，請獨立完成，team project則是分組作業。
- If the data you analyze contains missing values or any unreasonable values, please remove them or perform any preprocessing steps. 若資料有空值，或長得很怪(像是該是數字不是數字等)，請記得刪除或是做相對應的資料清理或處理。
- In the report, please clearly indicate the test you perform (like paired t-test, one-way ANOVA, etc.) and explain why you choose this test, the null and alternative hypothesis, one or two-tailed tests applied, the detailed calculation steps, and the conclusion from the test results. You can add descriptive analysis if needed. 在報告中，請寫清楚你要做什麼檢定(檢定名稱)，為什麼？虛無假設與對立假設(若用符號要先定義)，做雙尾還是單尾檢定，為什麼？詳細的計算過程(若從頭到尾用軟體則是要解釋該步驟你怎麼做的)，以及最後的結論。如果敘述性統計有助於解釋，也可做。
- Please check or declare the required assumptions before performing any statistics test if there is no assumption in the question. 如果題目沒講，可以考慮嘗試檢查做該檢定時要符合的假設，或是寫明要用符合什麼假設。
- All the questions fulfill the randomized setting. 以下題目都假設符合隨機性。
- 中文翻譯都是簡單翻題意，英文題會比較詳細

Personal Assignment: 5題 (以小題計)，每題10分 10 points for each question
You don't have to perform sampling for hypothesis tests and other analyses. If the data does not fit the required assumption for the test, please perform the test but state the violated assumption 做假設檢定或其他分析前不用再做抽樣，如果資料測起來不符合假設，還是要繼續做下去，只是請在作業中寫清楚不符合假設。(**2024/5/30** 英翻中)

1. The ship Titanic sank in 1912 with the loss of most of its passengers. Details can be obtained on 1309 passengers and crew onboard the ship Titanic. Dataset, description
    a. Among these passengers, are there age (column name: age) differences in the first-class cabin (pclass=1), second-class cabin (pclass=2), and third-class cabin (pclass=3) group? Which group(s) is(are) different from the others? Assume the population variances of age in each level of class are the same. 請問三種艙等的乘客年紀是否有差？
    b. Among these passengers, are the survival status/proportions (column name: survived) different among passengers in different levels of cabins (column name: pclass)? 乘客存活比例和乘坐的艙等是否有關？
2. In the JHU CSSE COVID-19 Dataset
    a. In the global dataset, is the Fatality Ratio (column name: Case_Fatality_Ratio) on 2022/07/07 (file name 07-07-2022) lower than the Fatality Ratio on 2021/07/07? (file name 07-07-2021) (description) 以全球的資料來看，2022/07/07 (Omicron wave)的疾病死亡比例是否有比2021/07/07 (Delta wave)的死亡比例低呢
3. An international e-commerce company based wants to discover key insights from its customer database. The company sells electronic products. (dataset from Kaggle)
    a. Is the customer rating (column name: Customer_rating) associated with if the products reached on time (Reached.on.Time_Y.N)? That is, please test if the customer rating affected by reached on-time status. 顧客評分跟貨品有沒有準時到是否有關？
    b. Is the cost of the product (column name: Cost_of_the_Product) different among the mode of shipment (column name: Mode_of_Shipment)? 使用不同運送方式的貨品商品價格是否不同？

Team Project: 50 points

In this semester, you've learned about applying statistics to make conclusions from data. Now it's time to apply statistics to the question you really care about. Please raise a question, collect data related to the question, perform statistics analysis and test, and finally make a conclusion based on the results. Please complete this project with your group members. 這學期你們學了如何應用統計學分析資料並下結論，在這個期末專題中，你們必須提出問題、搜集資料，並且應用所學在你有興趣的資料中，最後回答你提出的問題。這個Team project部分是分組作業，請跟組員一起完成。

Please submit **a document in either Mandarin or English** and **a presentation and demo video in either Mandarin or English** (less than 10 mins, including important contents from the document) to the E3 system (1 submission per group. If the video is too large to upload, please upload the video to YouTube and provide a link) by **2024/06/20 23:59 (no late work will be accepted)**, the contents should include:

1. Motivations & your questions
2. The process of data collection

3. Descriptive data analysis (codes & results & descriptions)
4. Statistic test (codes & results & descriptions)
5. The conclusion from the analysis

The scoring criteria, as a group (Will be graded on completeness, correctness, and clarity, by the instructor&TAs 60%, and your peers 25%, the instructions for peer-reviewed will be announced by 2024/6/21):

1. Data collection (30%)
    a. Data collection process and contents are clearly described
    b. The data they collect or download can answer are related to the question
    c. They collect enough data to answer the question
    d. Potential biases are described
2. Descriptive analysis (30%)
    a. Use the correct way to perform descriptive analysis
    b. Appropriate graphs are used to describe the data
3. Statistical test (30%)
    a. The hypotheses are related to their questions
    b. Appropriate statistical tests are used to test their hypotheses
    c. Correct conclusions are made based on the test results
4. Response to the previous peer review comments (HW3) (5%)
5. Creativity and others (5%)

The scoring criteria, as an individual (Will be graded on participation by your group members 15%):

1. Project contributions (5%)
2. Project participation (5%)
3. Easy to communicate/cooperate with as a team member (5%)

Q&A:

1. 如果用Excel而不是python寫題目要怎麼說明?
   比如說t test, 在excel內就是用T.TEST()
   T.TEST(array1,array2,tails,type)
   當然array1, array2要代換成你資料裡面的真實樣貌, 參數則是看題目要求而定, 長相應該像:
   T.TEST(A1:A20,B1:B30,1,1)
   這只是範例, 每份資料會不一樣

2. If the data you analyze contains missing values or any unreasonable values, please remove them. 若資料有空值, 或長得很怪(像是該是數字不是數字等), 請記得刪除或是做相對應的資料清理或處理。這裡所謂的空值是只要填入的資料有任意一行就直接做清理嗎?還是說沒有用到的行沒關係(?)
   只有要用到的資料需要清, 現階段沒有用到的資料是空的也沒關係！

就像蒐集資料的時候，有時候就是找不到或是大家不肯回答，比如說體重好了，只要你整個分析都跟體重無關，其實沒體重也沒關係。