Problem 1. You are working as product manager for a manufacturer of an insulin pump. There are a number of diabetic patients in South East Asia who have been using the insulin pump for a few years. The company decides to run a satisfaction survey among the patients country by country. One question simply asks the patients whether, overall, they are satisfied with the quality of life with the insulin pump (yes/no answer). In Singapore, 220 answered "Yes", while 12 answered "No". In Malaysia, 85 answered "Yes", while 8 answered "No".

a) Determine whether there is any significant difference between Singapore and Malaysia in terms of proportion of satisfied patients.

b) You decide to extend the survey to patients in Thailand by adding data from Thailand to the existing data from Singapore and Malaysia. You are still interested in whether the proportion of satisfied patients is the same across the countries. Your Thai colleague who is in charge of running the survey sends you this email

---

From: Mr Somchai (Thailand branch)
To: JENNIFER
Subject: [CONFIDENTIAL] survey in Thailand

Dear Jennifer,

I hope this email finds you well. I have asked my intern to collect the survey data as you requested. He reported back to me that 245 Thai patients answered "Yes" to the question. Unfortunately, the intern failed to tell me how many answered "No" or to tell me how many he actually asked. I know that at least 14 patients answered "No" because I happened to be there. If it can be of any help, a previous similar study in other countries resulted in a p-value of 0.1. I am terribly sorry for this inconvenience. I will try to contact the intern again and get back to you.

Best regards,

Somchai

Unfortunately, you can't wait for the answer by the intern because the deadline for this report is close. Instead, you decide to estimate the number of those who said "No" by following the clue of the p-value of the previous study and the fact that the number of respondents who said "No" is at least 14. In this situation, what is the number of Thai respondents who would have to say "No" in order for this current study to have a p-value as close as possible to the one suggested by Mr Somchai?

c) As part of the Singapore survey, patients are also asked a series of other questions about their quality of life. Scores for these questions are then combined to form the

"Quality of Life Score" (QLS). QLS scores are reported in the file `qls_data_43.dat`, where the first row is patient's age and the second raw is the QLS score for that patient. Determine whether there is a statistically significant trend of QLS with respect to age in these patients.

d) Based on the data in *c*), what is the predicted QLS score for a 25-years old patient? Provide an interval estimate.

Where necessary, please consider a level of statistical confidence of 93%.

Problem 2. The data in the file `virus_data_43.dat` refer to the number of people infected by a mysterious virus (2^n column, numbers refer to number of new infection cases per million) as a function of time (1^{st} column, in days since the day of the first reported infection) from various districts of a country. Consider the equation

$$y(x) = \beta_0 + \beta_1 * f(x)$$

where x is time and y is the number of new infection cases per million. Select a suitable equation for $f(x)$ to approximate the observations on the number of new infection cases per million over time as closely as possible. For the selected $f(x)$ determine:

a) the values of β_0 and β_1 that best approximate the observations.

b) using a 98% confidence level, the confidence intervals of both parameters β_0 and β_1.

Hint: you may want to try a few[1] *simple functions* $f(x)$*. There is no need for anything fancy.*

Instructions are on the next page

[1]There is no right number as to how many. Please do not ask the lecturer "how many should I try?". Please, use common sense: you should be able to count the number of attempts on the fingers of one hand.

INSTRUCTIONS

You need to submit to LumiNUS two Python files and one pdf file.

- A file called `Jennifer_p1_43.py` containing your commented code to solve problem 1. In your code you may choose the name of the variables. However, you are required to use the following variable names
  - `p_value_a` for the requested answer in a)
  - `n_thai_no` for the requested answer in b)
  - `p_value_c` for the requested answer in c)
  - `upper_bound` and `lower_bound` for the requested bounds of the interval in d)
- A file called `Jennifer_p2_43.py` containing your commented code to solve problem 2. In your code you may choose the name of the variables. However, you are required to use the following variable names
  - `beta_0` and `beta_1` for the parameters (β_0 and β_1) respectively mentioned in a) for the chosen $f(x)$
  - `upper_beta_0` and `lower_beta_0` for the upper and lower bound of β_0 respectively mentioned in b) for the chosen $f(x)$
  - `upper_beta_1` and `lower_beta_1` for the upper and lower bound of β_1 respectively in b) for the chosen $f(x)$
- A pdf file called `Jennifer_43.pdf` containing your answers to the problems. The answers are expected to be the required data analysis reports for each of the studies i.e., the key results as well what you can and can't claim after performing the necessary calculations and your reasons. For problem 2, please explain why you chose that particular $f(x)$. **No more than two pages in total**.

Please do not submit any given data file (I have them) nor any other data file. I will place your code in the same directory as your assigned data files and your code is expected to work in such conditions. You may treat this assignment as a realistic scenario where any questions/considerations/comments on the data and the the analysis should be included in your data analysis report (the pdf file).
This assignment aims at testing your ability to code the correct solution as well as the understanding of the data analysis we covered in this module so far. As such, your code must not contain any built-in Python functionality that was not covered in the tutorials or in the primer document.
The deadline is April 10^{th} at 05:00pm. Late submissions will be penalized according to the amount of delay. Failure to follow the instructions above will also cause a penalty to be applied. Plagiarism will not be tolerated.