PROJECT ON STATISTICS-FALL 2021

1. Deadline for report submissions: 1 PM, December 27, 2021. Submit a zip file via BB. The zip-file contains all files of Task 1 and Task 2.

Instructions: You can work in a group of 2-3 students (you can work alone if you do want). The project includes two Tasks: **Task 1 of 40 points for an independent submission, Task 2 of 60 points for teamwork. There will be a session for project presentation for Task 2 on December 27, 2021.** If you have any questions or need advice, feel free to ask the instructor.

Task 1 (40 pts): Complete the R-certificate for the open course "Introduction to R" on Datacamp (the link below is free and in Vietnamese):
https://www.datacamp.com/community/open-courses/h%C6%B0%E1%BB%9Bng-d%E1%BA%ABn-c%C6%A1-b%E1%BA%A3n-v%E1%BB%81-r

Task 2 (60pts): Choose one of the following options/projects:

Option 1 for Task 2: Project for Analysing Global Warming in Vietnam
The purpose of this task is to investigate statistically whether there is a significant change in temperature over time in Vietnam. Below are a few ideas for the data collection and the data analysis:

Go to the website: https://climateknowledgeportal.worldbank.org/download-data

1) Select Variable: "Mean-Temperature" (monthly)
2) Country: Vietnam
3) Download the csv file for the two data sets:
    a. 1st data set: from 1931 to 1960;
    b. 2nd data set: from 1991 to 2020;
4) Download the csv file (file 1) and **collect** two data sets from **N**th month (file 2), for example, **N** can be the month of your birthday (or, month of the birthday of a student in the group).
5) For both of your data sets in file 1 and file 2, determine the mean and standard deviation. Plots some appropriate graphs. Compare the two datasets.
6) **Design an appropriate hypothesis test.** Use R/Python to analyse **whether there has been an increase** in the average temperatures over the years at, for example, 10% level of significance and at 2% level of significance, respectively. One can vary the level significance and observe the results.
    i) State the steps of hypothesis testing clearly and justify your statements.
    ii) Discuss what conclusions can be drawn from your findings for each significance level.
7) If it is possible, one can explore/infer some predictions for the future.

Option 2 for Task 2: A team can choose any project that you would like to study!

The most important aspects of any statistical data analysis are stating questions, collecting data, visualizing data, and **analysing the data** to infer the conclusions or predictions. The techniques can be summarizing data, and/or confidence intervals, and/or hypothesis testing, and/or regression models, or a technique that we did not cover in class.

Here are some ideas for the option 2 for Task 2:

2.1 Describe in detail the **logistic regression**: model and a/some applications. and apply the model, for example, to accept the personal loan or to predict the bankruptcy of a company. The probability p of bankruptcy is between 0 and 1 and can be predicted by the logistic regression model.

2.2 Describe in detail the **multi-linear regression**: model and a/some applications.

2.3 Analysing the poverty and equity in Vietnam, and/or Analysing population (urban, rural, largest cities) and making predictions for population in Vietnam, etc.

Data can be taken from the World Bank:

http://povertydata.worldbank.org/poverty/country/VNM

2.4 Analysing the data for a few aspects for Vietnam such as Income, GDP, finance, loan, travel services, Internet users, labour force, employment, education: pupil-teacher ratio, school enrolment, Electricity consumption, Electricity production, tourist, air transport, export, life expectancy, CO2 emission, renewable energy, etc.

Data can be taken from the World Bank:

https://data.worldbank.org/country/vietnam?view=chart

2.5 Study a theoretical model and apply the model for a/some specific applications. For example, study a dimensional reduction problem.

2.6 Study a question/issue by collecting data on your own (e.g., doing a survey, in your classes or taken from other sources), then analyzing the data. Some ideas:

1. What are the factors influencing choice of college major, for example, at IU: what really makes a difference? (Interest in the subject, career aspirations, family influence, or ability in the subject?)

2. Read another research, for example, on the Pew Research Center, a nonpartisan American think tank: https://www.pewresearch.org/

Grading Guidelines:

+ For Task 1 (40 pts): Straightforward (40/40 points for completing the free course “Introduction to R”).

+ For Task 2 (60 pts): Based on the content of the project of 45 pts and the presentation of 15 pts. More specifically, the content of the project will be marked based on the problem, writing/presentation (clarity, fashion), data collection, data analyzing, results, conclusions/discussions. Note that if you are using models/techniques that we did not cover in class, you should explain the models/techniques.

---HAVE FUN! ^-^ ---