Introduction

About This Course

The goals of this course are to get you in a position to:

1. To learn **how to use different statistical softwares** to analyze data and interpret computer output.
2. To use simulations to solve statistics problems.
3. To solve real life problems through statistics.
4. To appreciate statistics through analyzing data.

LumiNUS

- All announcements will be made through the course web-page on LumiNUS.
- All lecture notes, tutorial sheets and tutorial solutions will be uploaded to the folder "Files" under ST2137 of LumiNUS.
  - The lecture notes will be uploaded before the class using it.
  - Tutorial solution will be uploaded at the end of the week or over the weekend.
  - This course will require the use of R, Python and SAS.

Lectures and Tutorials

Days	Time
Tue & Fri	2 - 4 pm

- Lectures are conducted live via Zoom. The Zoom link is given under "Conferencing" of Luminus.
- Lectures are webcasted also. Do note that it may take few working days for the record to be uploaded onto Luminus.
- Tutorials are recorded by our TA(s) and will be uploaded to LumiNUS for e-learning.
- Tutorial registration will be done online.
- Each tutorial is about 45 minutes to solve the tutorial questions.

Assessment and Final Examination

Assessment in this class will be based on the following elements:

- Two individual assignments will make up 20% of the grade.(Assignment 1 comes around Week 3-5 while Assignment 2 comes around Week 11-12)
- Two online quizzes will make up 30% of the grade. About (Week 7 and Week 10)
- The final examination will make up 50% of the final grade.
- The final examination will be held online, on Sat, 27 Nov 2021, 13:00 - 15:00 (120 Minutes).

There is no makeup for the final examination if student is absent.

Student will need to record the screen of computer during the time taking quizzes and finals. The record must be submitted (to LumiNUS/Multimedia) for checking. Hence, each student should be familiar with computer full screen recording and uploading.

We will need...

- R (RGui or RStudio) and Python (along with Jupyter Notebook): These are free.

- SAS: We'll use the SAS OnDemand for Academics which is also free. SAS will be introduced when students are already familiar with R and Python (Topic 3 from Week 7).

- The quizzes and final exam will test on writing code and analyzing the output. Hence you must be familiar with the routines that we call during the lectures and tutorials.

Some Useful Texts

Some books that may be good for your reference:

- *Statistics: An Introduction Using R, 2nd edition*, Michael J. Crawley.
- *Python for Data Analysis, 2nd edition*, Wes McKinney.
- *Applied Statistics and the SAS Programing Language, 5th edition*, Ronald P. Cody, Jefferey K. Smith.

Asking Questions

- If you have questions about topics in class, I urge you to post it on the Forum on LumiNUS first. And if you know the answer to a question your classmate has asked, I encourage you to answer the post.

  I'll follow the forum closely, so no need to worry that you post a not-correct answer (since I'll correct it).

  All posts on LumiNUS are anonymous.

- You may email me the questions that you have.
  My email: staptkc@nus.edu.sg

  Please put "ST2137" some where in the title of your email. Otherwise I may be delayed in getting to it.

- Face-to-face consultation is not encouraged given the current situation of Covid 19.

Topics

- Statistical software (R, Python, SAS)
- Describing numerical data
- Robust estimators for location and scale
- Analysis of categorical data
- Inferences on one-sample and two-sample data
- One-way ANOVA
- Regression Analysis
- Simulation studies
- Resampling methods: Bootstrap (if time permits)

Types of Statistical Software

There are two major types of statistical software:

- (A) Ready-made type (recipe type)
  - User friendly and easy to use
  - Less flexible
  - Packages: SPSS, Minitab and so on
- (B) Tailor-made type
  - More flexible
  - Need some sort of programming/coding
  - Packages: Python, R, SAS, S-Plus, C++ and so on
- SAS was the statistical dominator software for the past four decades. It is used by FDA (in US) and most pharmaceutical companies, by multinational companies and ministries. It is powerful but expensive for the purchased version.

R

- Used by statistics researchers
- Flexible
- R is a freeware: `http://www.r-project.org`
- R is the main statistical software that we'll use in our course.

Python

- Used by many data scientists
- Flexible and free `https://www.python.org/`
- Now Python is becoming more and more popular and is chosen by many of data scientists. Hence I made a move, that from last semester, ST2137 started to introduce Python to students.
- We'll need some form of editor to write programs. Using integrated development environment (IDE) editor such as PyCharm or Jupyter Notebook (a web-based IDE) is recommended. **For our course, we'll use Jupyter Notebook**.

Data

"Data" is defined as

- things known
- assumed facts or figures

for which conclusions can be inferred.
Examples:

The Use of Data

- To draw conclusions and make decisions
- To confirm statements
- To make predictions on events to come

Data Analysis

What is data analysis?

- The art of examining, summarizing, and draw conclusions from data
- Transform the data into knowledge

Example 1

1. Questions
   ↓
2. Collecting data
   ↓
3. Analyzing data
   ↓
4. Answers/More questions
   ↓
5. Conclusion

Example 2

1. Available data
   ↓
2. Analyzing data
   ↓
3. Any patterns/Relationship/More questions
   ↓
4. Conclusion

Example 3

- A question: Is it easier to get in a pre–university institution (Junior college or pre-university institution) in 2004 than in 2000?
- Getting data:
  - JC/PU enrollment in 2004: 13435
  - JC/PU enrollment in 2000: 12191
  - Is the answer a "Yes"?
- Analysis: The enrollment can be smaller due to a smaller cohort.
  - O-level enrollment(Sec 4(S,E) and Sec 5(N)) in 2003 is 33387
  - O-level enrollment (Sec 4(S,E) and Sec 5 (N) in 1999 is 33163
- Conclusion:

Example 4

A dataset is given with descriptions:

- 97% first year students answered a questionnaire and gave the information.
- Variables: Gender, Height, Weight, Number of siblings

What can we get out of this dataset?

- Are height and weight correlated? If so, is there a relationship between these two variables?
- Are the relationships between height and weight the same for both male and female?
- Are there any differences in the height for different races?
- What is the average height of first year student? Average weight?

How To Do Data Analysis?

1. Get the right data set
   - Clean
   - Merge
   - Transform
   - Aggregate
2. Choose the right kind of statistical tools
   - Identify relevant variables
   - Graphs, charts, tables
   - Numerical summary
   - Hypothesis testing and estimation
   - Building models
3. Write a report: Avoid using technical jargon
4. If possible, monitor the changes in the data and conditions over time and make necessary changes