The Chinese University of Hong Kong
Faculty of Science
Science Academy for Young Talent

Summer Courses 2020
Course Outline

CUSA1026 Statistical Modeling and Big Data Analytics
統計模型及大數據分析

Introduction:

Data from various fields, such as climatology, finance and sports, exhibit different properties. This course aims to use the R-package (a statistical software) to visualize the properties of the data, fit the data into various statistical models, assess model performance and predict the data. Topics include exploratory data analysis, time series models, hidden Markov models, classification trees, Poisson process and analysis of big data problems. Students will gain hands-on experience in statistical programming at the computer lab.

各種領域的數據（如氣候學，金融及運動）會展示不同的特質。本課程目標是透過統計軟件 R 去透視數據多方面的特性，從而用適當的統計模型去解釋，評估模型的表現及作出數據預測。本課程涵蓋範圍包括：探索性數據分析，時間序列模型，隱馬爾可夫模型，分類樹，泊松過程和大數據問題的分析。學生將親身體驗統計程式的編寫。

Learning Outcomes:

Upon completion of this course, students should be able to:

1) Understand data from various fields
2) Apply the exploratory data analysis (EDA) to visualize the properties of the data;
3) Understand the theories behind various statistical models, and how the models can be fitted into different data sets;
4) Write computer programs in R to perform various statistical analysis;
5) Develop a systematic approach in solving statistical problems;

Medium of Instruction: Cantonese supplemented with English

Organising Unit:
Department of Statistics, CUHK

Teacher:
Dr. LEE Pak Kuen Philip
Department of Statistics, CUHK
Room 116, Lady Shaw Building, CUHK
E-mail: pklee@sta.cuhk.edu.hk

Course Content:

7 August 2020 (Friday) 9:30am– 1:00pm 2:00pm–5:30pm	Lecture: • Exploratory Data Analysis (EDA): Scatter plot, Box plot, Histogram, Quartile-quartile Plot, Time Series Plot, Autocorrelation (ACF) Plot, PACF Plot • Climate Data: Properties, Seasonal ARIMA Model, Model Prediction • Financial Data: Properties, Hidden Markov Model, GARCH Model Computer Lab Activities: •R programming: The Basics, EDA, Estimation of Time Series Models Assessment: •Data Modeling in R
10 August 2020 (Monday) 9:30am– 1:00pm 2:00pm–5:30pm	Lecture: • Sports Data: Properties, Poisson Process, Implied Probability and Odds • Classification Trees • Big Data Problems and Analysis Computer Lab Activities: • R programming: Advanced Modeling Case Discussion and Assessment: • Data Modeling in R

Duration	2 whole day sessions (total 14 contact hours)
Date	7, 10 August 2020
Time	7 August 2020: 09:30am– 1:00pm; 2:00pm–5:30pm 10 August 2020: 09:30am– 1:00pm; 2:00pm–5:30pm
Teaching Platform	Zoom
Enrollment	30
Expected Applicants	Students who are studying S4-S5with good knowledge in mathematics
Tuition Fee	HKD 2,352.00
Credit	1 Academic Unit Certificates or letters of completion will be awarded to students who attain at least 75% attendance.