統計應用方法 Applied Methods in Statistics

Spring 2025 課程綱要

授課教師	黃冠華教授 辦公室：綜合一館 423 室 電話：03-5131334 電子郵件：ghuang@nycu.edu.tw
上課時間與地點	每星期三（上午）9:00-12:00 於綜合一館 304 室
課程網頁	E3 數位教學平台：https://e3.nycu.edu.tw/
開課單位	生資碩
永久課號	BTBI30081
學分數	3

課程概述與目標

本課程將以實際的資料為核心，搭配統計軟體 R (http://www.r-project.org/)的使用，引導課程參與者接觸並學習資料探索方法(exploratory data analysis)、不同型態資料的統計檢定(statistical hypothesis testing)方法、迴歸分析(regression analysis)、主成份與因素分析(principal component and factor analysis)、集群分析(clustering analysis)、分類分析(classification analysis)、機器學習(machine learning)等實用的統計分析工具與相關的概念。

上課內容，將廣泛包含所有相關知識，上課時側重講述這些知識的基本觀念與模型解釋（如果需要時）。對於深入的理論與其餘詳細資訊，則僅作重點提示或提供參考文獻。課堂中將以實際的例子來補充上課內容，並討論相關方法的統計軟體 R 的實作。

課程組成部分

課堂講解

每星期三（上午）9:00-12:00，由授課教師講解課程相關的主題。上課內容，將廣泛包含所有相關知識，上課時側重講述這些知識的生成動機、基本觀念與模型解釋（如果需要時）。對於深入的理論與其餘詳細資訊，則僅作重點提示或提供

參考文獻。期盼日後當學生獨立進行統計分析時，這些廣泛的知識，能增廣他們思考問題的角度，並成為眾多他們可選擇的解決方案。若要進行更深入的模型研究與理論推導時，則知道要從何下手與到何處去找尋相關的輔助資訊。

作業

我們將會有 3 份作業，練習主題包含：資料探索方法(exploratory data analysis)、不同型態資料的統計檢定(statistical hypothesis testing)、線性與羅吉斯迴歸分析(linear and logistic regression analysis)。

由於大部份的作業問題，會須要以 R 程式軟體來進行實作、分析，因此要求同學們的作業要以 R Markdown (http://rmarkdown.rstudio.com/)的格式來撰寫。R markdown 能將你的文字說明、數學式子、R 程式、R 執行結果、…等，結合成一個文件，如此將易於他人閱讀與重製(reproduce)你的分析。

你可與其他同學討論作業，以幫助理解所問的問題、釐清課程概念。但是你必須獨立完成所繳交的作業，作業中要求寫的電腦程式、跑的資料分析、解釋的分析結果，都不可與他人共同合作。

考試

本課程將會有一次考試（預計於 4 月 30 日舉行）。考試為 open-book test，範圍將只會包含：資料探索方法(exploratory data analysis)、不同型態資料的統計檢定(statistical hypothesis testing)、線性與羅吉斯迴歸分析(linear and logistic regression analysis)這些內容。考試目的在測試學生對課堂講述之基本觀念、模型、方法的理解程度，還有檢驗學生運用課堂上所學的統計方法與技術來進行資料分析的能力。

課程實作計劃

修課學生須完成一份數據分析的計劃，其目的在讓你能就一個所關心或有興趣的議題，運用課堂上所學的方法與技術，從問題形成、資料來源確認、資料搜集、儲存與整理、模型建立與分析、結果呈現、說明與視覺化，實作整個數據分析計畫，以一窺數據分析的全貌。

每份計劃報告將由**最多 3 位**修課同學共同完成（亦可自己一組），每一報告工作小組，將各自選定一個所關心或有興趣的議題（非模型、方法、理論等技術性探討）。學期中，**每個組員**將先就計畫主題（包含：描述問題、預計如何回答），各自繳交一份書面報告。學期末，**整個工作小組**將就計劃的：問題（目的為何？想預測或估計什麼？）、資料（那裡來的？看起來像什麼？）、分析模型、結果（新發現、與聽眾溝通、視覺化），進行 15 分鐘的口頭報告，與繳交最終書面報告。

先修科目或先備能力

1. 有寫電腦程式的經驗
   - 像：C, C++, Java, Python, R,…
2. 修過基礎統計學
   - 知道：隨機變數、信賴區間、假設檢定、…
3. 願意學習新的軟體、工具
   - 常會非常花時間
   - 要大量閱讀網路上的文件
   - 閱讀許多英文文件

課程實作軟體與教科書

本門課將會以 R 程式軟體(http://www.r-project.org/)，來當作資料分析實作的工具。因此不論演習課助教講解與作業問題，皆會以 R 程式軟體的操作與撰寫為基礎。

本門課雖無指定、必須購買的教科書，然相關的自行閱讀、補充教材內容，將出自以下幾本參考書籍：

1. Irizarry RA, Love MI (2015): Data Analysis for the Life Sciences. 這本書的相關訊息，可從以下連結獲得：https://leanpub.com/dataanalysisforthelifesciences
2. Montgomery DC, Peck EA, Vining GG (2012): Introduction to Linear Regression Analysis (5th Edition). Wiley. 這本書是「迴歸分析」的主要參考書目。
3. Johnson RA, Wichern DW (2007): Applied Multivariate Statistical Analysis (6th Edition). Prentice Hall, Upper Saddle River, NJ. 這本書是「多變量分析」的主要參考書目。

本課程所有上課投影片與相關補充資料，還有用以執行演習課實際例子與上課講議圖形的 R 程式，都將會公佈於課程網頁。

學期成績評分方式

學期成績的計算方式為：

1. 作業（3 次）：30%

2. 考試（1 次）：30%
3. 實作計劃期中報告：10%（根據個人繳交之書面報告）
4. 實作計劃期末報告：30%（根據整個工作小組的報告）

課程大綱

- Exploratory data analysis
  - Measurement scales, data types
  - R graphic package: ggplot2
  - Displaying distribution of univariate data: stem-and-leaf plot, q-q plot, histogram, box plot, bar chart, pie chart
  - Displaying correlation for bivariate data: scatterplot, box plots, stacked bar chart, faceting bar charts, stacked area chart, time series plot
  - Displaying association for multivariate data: 3d scatterplot, lattice in the 3rd dim, map the 3rd dim to colors, lay out panels in the 3rd dim, scatterplot matrices, heatmap
- Statistical decision making: hypothesis testing
  - Basic concepts: null versus alternative hypothesis, type I type II errors, significance level, test statistic, power, p-values
  - Hypothesis testing for continuous random variables: one-sample t-test, two-sample t-test, F-test for equal variance, ANOVA, paired t-test,
  - Hypothesis testing for categorical data: binomial test, x^2 test / Fisher's exact test, McNemar's test, Cohen's kappa test, Mantel-Haenszel test
  - Nonparametric statistical methods: sign test, Wilcoxon signed-rank test, Wilcoxon rank-sum test, Kruskal-Wallis test
  - Computational methods: permutation test, bootstrap
- Regression analysis
  - Simple and multiple linear regressions for continuous data
  - Interpretation and estimation of regression coefficients
  - Confounding and interaction
  - Regression diagnostics
  - Variable selection
  - Logistic regressions for binary data
- Principal component and factor analysis
  - Population principal components
  - Summarizing sample variation by principal components
  - Orthogonal factor model

- Factor rotation
- Factor scores

- Clustering analysis
  - Similarity and distance measures
  - Hierarchical clustering methods
  - K-means clustering methods
  - Multidimensional scaling
- Classification and discrimination analysis
  - Linear discrimination analysis
  - Quadratic discrimination analysis
  - Evaluation of classification: cross-validation
- Machine learning
  - K-nearest neighbor (KNN)
  - Classification and regression trees (CART)
  - Artificial neural network (ANN)
  - Support vector machine (SVM)