R 軟體在統計學教學中的應用

Application of R Software in Teaching Statistics

姜正雄

Cheng-Hsiung Chiang

玄奘大學資訊管理學系
Department of Information Management,
Hsuan Chuang University

摘要

統計學是基礎且重要的學科，統計學技巧可以分析資料、探索資料隱含的意義以及萃取出大數據中有用的資訊。然而部份統計方法所需的計算過程繁瑣，徒手繪製統計圖表也不準確。使用電腦軟體輔助統計學的教學與學習，能提升學生的興趣及學習效果。本論文介紹以 R 軟體輔助統計學教學，R 軟體是免費軟體，目前約有一萬種不同用途的套件可免費下載安裝。其中 Rcmdr 圖形化介面套件常用於統計分析，比較容易使用。依據我們的教學經驗發現，那些對於輸入程式敘述以進行運算會感到困難的學生，比較適合使用 Rcmdr 套件。本論文提出幾個統計學的問題，讓學生練習以 Rcmdr 套件解決這些問題。透過練習學生將能熟練使用 R 軟體以進行統計分析的技巧，並更熟悉統計學的方法及概念。

關鍵字：統計學、R 軟體、R Commander

Abstract

Statistics is a basic and important subject. The Statistical skill can be used for analyzing data, exploring the implied meaning of the data, and extracting useful information for big data. However, the computational process required for some statistical methods is cumbersome, and it is not accurate to draw statistical charts on hand. Assisting statistics teaching and learning by using computer software will enhance students' interest and learning effect. This paper introduces the application of R software for teaching statistics. R software is a free software, and there are about ten thousand kinds of different packages for freely download. The Rcmdr package which is a graphical user interface is commonly used for statistical analysis and it is easy to be used. According to our teaching experience, those students who feel difficult about writing program statements to carry out analysis are suitable use the Rcmdr package. This paper presents several statistical problems for students to practice solving these problems by using Rcmdr. Through practicing, students will be familiar with the skills of statistical analysis by using R software and the methods and concepts of statistics.

Keywords: statistics, R software, R Commander

1. 前言

統計學課程日益重要並已成為高等教育課程中的主要部分。國際商管學院促進協會 (AACSB) 也要求大學生及研究生展現其對於統計方法的熟練度。對於做研究、分析資料甚至是大數據資料分析，統計學的技巧是必要的。由於統計學的知識很有組織，對於定理的推導需要數理基礎。因此很多學生對於理解基礎統計學概念有困難 [3]。在 1990 年代中期，Cobb [4] 指出老師已經開始使用電腦模擬展示重要的統計學概念，並允許學生自己探索這些概念 [7]。

目前常用來輔助統計學教學的主要軟體包括 Excel、R、SAS、SPSS、Matlab 及 Eviews 等軟體。如果沒有軟體輔助教學，無論是老師講解例題，還是學生完成習題都需花費大量的時間進行手工計算，且錯誤率高。使用軟體輔助學習，可以提高學生的學習興趣以及動手操作或程式設計能力；也可以分析和精確地計算巨量資料，並可省略查詢統計學附表 [2]。

R 軟體 (或稱作 R 語言) 是紐西蘭奧克蘭大學的 Ross Ihaka 及 Robert Gentleman 所開發（開發者名字第一個字母都是 R，因此稱為 R 軟體），由於當初沒有合適的統計教學軟體，所以發展出 R 軟體 [9]。R 軟體是一種自由軟體，可以安裝免費的套件 (packages) 以增強功能。根據 CRAN 網站的紀錄，至 2017 年 4 月初為止，目前有 10399 種不同的套件，所涵蓋的領域包括統計學、物理、化學、數學、經濟學、財務、人工智慧、醫學、心理學、社會科學及網路技術等。R 軟體有以下幾個優點 [5]。第一，R 軟體持續地更新，它更新的速度比其它傳統的商業統計軟體更快。第二，由於 R 是免費的，適合於教育用途。R 軟體是一套程式語言，學生可以學習程式設計，並能更有彈性地設計程式解決統計問題。第三，R 可以畫出具出版品質的圖形，它可以畫出複雜的三維圖形。最後，R 很適合用於蒙地卡羅模擬，它可以模擬數千次的運算 (例如：模擬五千次型 II 誤差率)。

2. R 軟體簡介

R 軟體的免費下載網址為 https://www.r-project.org/，安裝完成後執行的畫面如圖 1 所示。這是原始版本的 R 軟體，透過主控台輸入 R 的程式敘述，以執行對應的動作。我們可以安裝一款常用的整合開發環境 (IDE) 免費軟體 RStudio，讓我們能夠有效率地使用 R 軟體。RStudio 可用於不同的作業系統，例如 Windows、Mac 以及 Linux。

RStudio 的下載網址為 http://www.rstudio.com/，RStudio 將 R 軟體的主控臺、命令稿、工作空間、說明文件及圖表呈現區整合在一起，並分成不同的視窗區域 (如圖 2)，便於進行檔案和套件管理，適合初學者入門使用 [1]。**命令稿**視窗顯示程式原始碼、文字檔、R 的文件、TeX 文件以及 HTML 文件等。右上方**工作空間**視窗的第一個標籤「Environment」用以顯示目前所建立的所有變數，並可匯入資料集，第二個標籤「History」顯示所有執行過的命令、函數及動作。左下方**主控臺**視窗用來輸入程式敘述，執行後產生相對的結果。右下方的**圖表呈現區**有五個標籤，包括 Files、Plots、Packages、Help 及 Viewer [8]。

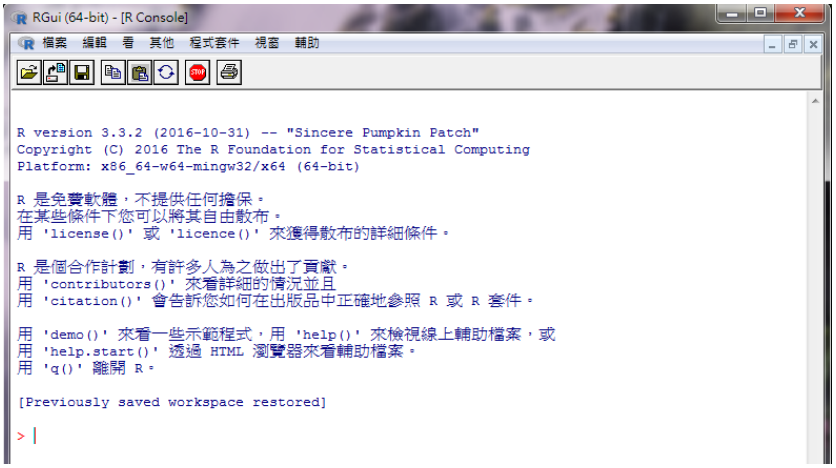


圖1　R軟體的執行畫面 (64位元，3.3.2版)

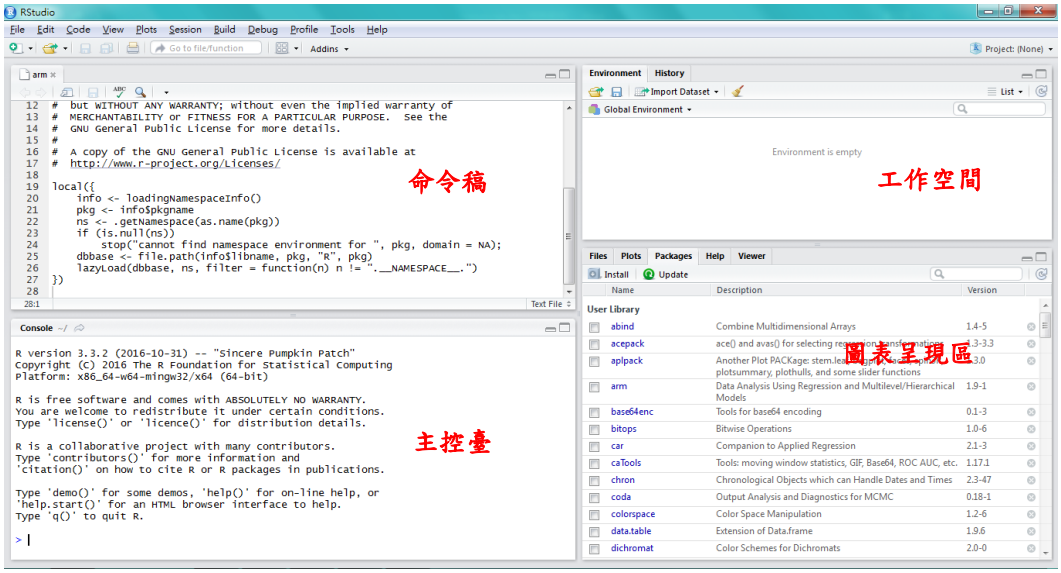


圖2　RStudio軟體的執行畫面 (1.0.136版)

雖然 RStudio 可以更方便地進行以及維護寫程式的工作。但是，輸入程式敘述有時會讓學生困擾。我們可以使用圖形化介面的套件，R Commander 是一套用於統計分析的圖形化介面套件，套件名稱是 Rcmdr。能夠方便地用滑鼠在選單中點選以執行功能。安裝 Rcmdr 套件後，可以在**主控臺**輸入 library(Rcmdr) 載入這個套件， 如圖 3 所示。

在【資料】功能中，可以建立新的資料集或者匯入事先建立好的資料 (可以匯入的檔案格式包括：文字檔、SPSS 檔、SAS 檔、Minitab 檔、STATA 檔及 Excel 檔等)。在【統計量】功能中，能夠對資料進行分析的項目包括：摘要、列聯表、平均數檢定、比率資料分析、變異數檢定、無母數檢定、多變量方法以及迴歸分析等。【繪圖】功能中可以繪製的圖形包括：點圖、直方圖、莖葉圖、盒鬚圖、散佈圖、折線圖及 3D 散佈圖等。使用 R Commander 的優點是不需要輸入程式敘述來進行統計分析，R Commander 會產生相對應的程式碼。

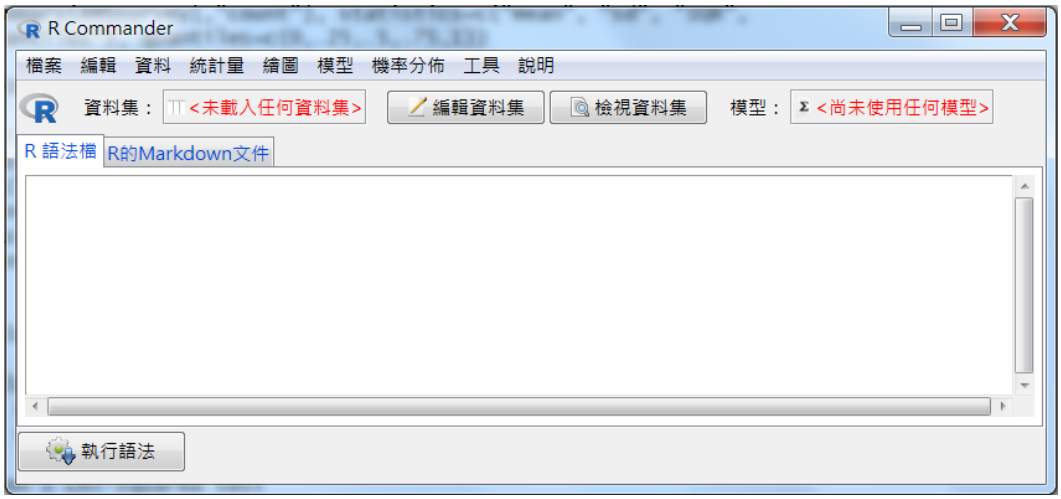


圖3　R Commander 套件的執行畫面

3. 應用 R 軟體於統計學教學

在國內，統計學教科書所採用的輔助學習軟體大多是 Excel、SPSS 以及 R 軟體。前兩種軟體不需要寫程式，透過點選內建的功能選單即可分析資料。R 軟體除了有圖形介面的 R Commander 套件可用，若學生的程度適合，也可以由學生自行寫程式進行資料分析，能夠更有彈性地分析資料。本節將介紹使用 R Commander 套件進行資料分析。以下介紹幾個統計學的應用範例。

3.1. 資料的描述性分析

假設某一座博物館隨機挑選 48 個小時，在一小時內遊客參觀人數，如下所示：

51	59	94	77	50	31	70	96	82	55	45	72	77	87	91	93
53	94	58	74	94	77	84	62	88	53	97	53	83	91	98	86
52	46	57	65	42	32	83	64	40	97	70	98	78	60	52	78

我們可以利用 R Commander 套件或者再修改 R Commander 所產生的語法以調整圖表格式，來繪製統計圖表了解資料的分佈情形。以下有幾個問題讓學生練習回答：

1. 計算這些資料的重要統計量，包括：平均數、標準差以及四分位數。
2. 判斷資料中是否有離群值。
3. 繪圖說明這些資料所呈現的分佈情形，並判斷資料分佈是否對稱。

以下是這些問題的提示：

1. 利用 R Commander【統計量】功能之下的【摘要】項目，可以匯整算出資料的平均數、標準差以及四分位數等統計量。
2. 當數據小於 $Q_1 - 1.5\,(\mathrm{IQR})$ 或者大於 $Q_3 + 1.5\,(\mathrm{IQR})$，則此為離群值。其中 Q_1/ Q_3 為第一/第三四分位數，$\mathrm{IQR} = Q_3 - Q_1$ 為四分位距。

3. 可以利用 R Commander【繪圖】功能來繪製直方圖、密度估計圖、莖葉圖或者盒鬚圖。圖 4 (a) 是密度估計圖，圖 4 (b) 是盒鬚圖，由圖 4 (a) 可看出資料有兩個高峰，分別位於 [40, 60] 以及 [80, 100]。

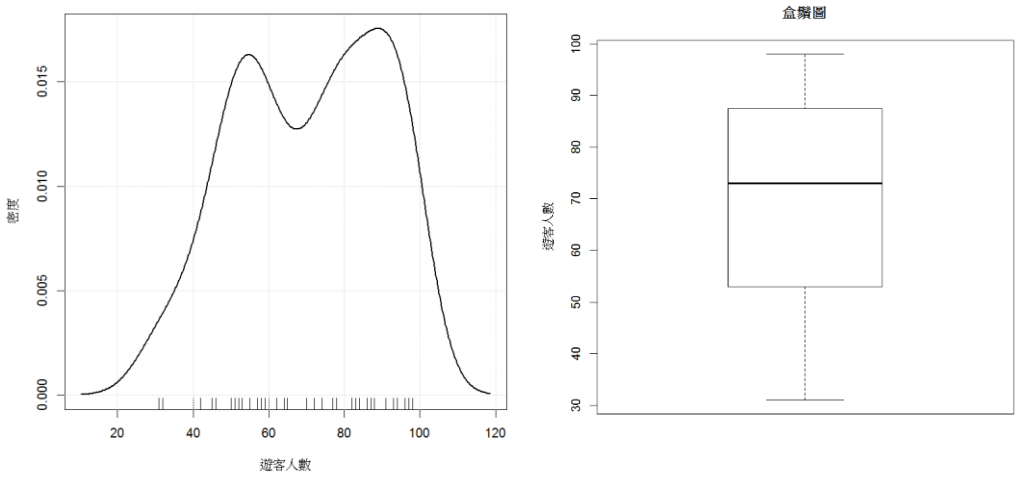


圖4　參觀博物館的人數 (a) 密度估計圖 (b) 盒鬚圖

3.2.　變異數分析

本節使用一個資料集，引導學生練習回答變異數分析的問題。以下的資料 (表 1) 是美國在 2008-09 年及 2011-12 年獲得數學科學博士學位的統計人數 (AMSsurvey 資料集，[6])。**type** 變數的 I(Pu)及 I(Pr)分別代表公立及私立大學授予博士學位的學系，其評比分數為 3.00-5.00 分；II 表示其評比分數為 2.00-2.99 分。III 表示該學系未參與 1995 年的評比；IV 包含授予統計學、生物統計學及生物識別技術的博士學位的學系；Va 包含授予應用數學及應用科學博士學位的學系。**count** 及 **count11** 分別代表 2008-09 年及 2011-12 年獲得博士學位的人數。

表 1　美國在 2008-09 年及 2011-12 年獲得數學科學博士學位的統計人數

No.	type	sex	citizen	count	count11	No.	type	sex	citizen	count	count11
1	I(Pu)	Male	US	132	148	**13**	I(Pu)	Male	Non-US	130	136
2	I(Pu)	Female	US	35	40	**14**	I(Pu)	Female	Non-US	29	32
3	I(Pr)	Male	US	87	63	**15**	I(Pr)	Male	Non-US	79	82
4	I(Pr)	Female	US	20	22	**16**	I(Pr)	Female	Non-US	25	26
5	II	Male	US	96	161	**17**	II	Male	Non-US	89	116
6	II	Female	US	47	53	**18**	II	Female	Non-US	50	56
7	III	Male	US	47	71	**19**	III	Male	Non-US	53	61
8	III	Female	US	32	28	**20**	III	Female	Non-US	39	30
9	IV	Male	US	71	89	**21**	IV	Male	Non-US	122	153
10	IV	Female	US	54	55	**22**	IV	Female	Non-US	105	115
11	Va	Male	US	34	42	**23**	Va	Male	Non-US	28	27
12	Va	Female	US	14	21	**24**	Va	Female	Non-US	12	17

依據以上的資料，請學生練習回答以下問題：

1. 不同的性別分別對於在 2008-09 年及 2011-12 年獲得博士學位的平均人數是否有顯著不同？男生還是女生的人數較多呢？
2. 在 2008-09 年及 2011-12 年，不同的學系授予的博士學位平均人數是否有顯著差異？
3. 在 2008-09 年及 2011-12 年獲得博士學位的平均人數是否有顯著增加？

以下提供這些問題的提示：

1. 問題 1 可以執行 R Commander【統計量】功能的【平均數檢定】➔【單因子變異數分析】。從計算結果可以觀察到性別對於獲得博士學位人數的影響。
2. 問題 2 和問題 1 相似，只是要分析的自變數改成 **type** 變數。
3. 問題 3 可以執行【統計量】功能的【平均數檢定】➔【成對樣本 T 檢定】，以判斷這兩個時期的博士學位平均人數是否顯著不同。

3.3. 迴歸分析

本節使用 2011-12 年，美國紐約市公寓評價資料做為迴歸分析的範例資料。請學生先至網路下載整理過的資料，可以在 RStudio 的主控臺輸入以下敘述：

house <- read.table("http://www.jaredlander.com/data/housing.csv", header=TRUE, sep=",", na.strings="NA", dec=".", strip.white=TRUE)

下載的資料集會儲存在 house。這個資料集有 13 個變數，總共有 2,626 筆記錄。我們想透過多元迴歸分析找出影響每平方呎價格的顯著因素。圖 5 左方是每平方呎價格的直方圖，右邊列出資料集內 13 個變數名稱。

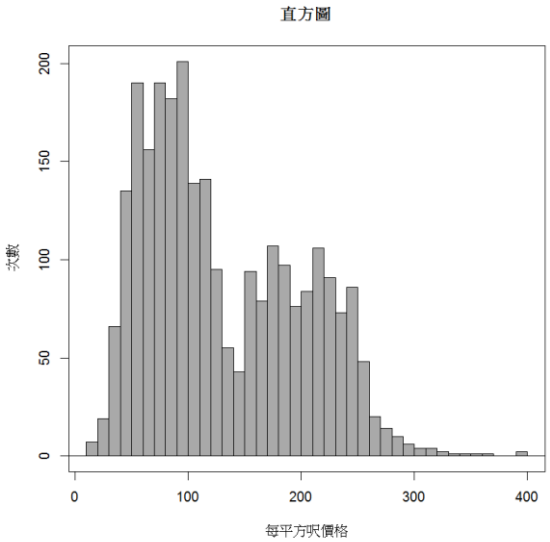
若要使用 R Commander 分析資料，則須先匯入資料。點選【資料】功能，再選擇【匯入資料】➔【匯入文字檔, 剪貼簿或 URL 檔】。以下幾個問題讓學生練習回答：

1. 繪製公寓每平方呎價格 (Market.Value.per.SqFt 變數) 的密度估計圖或其它類型圖表，觀察資料分佈情形。
2. 繪製 Market.Value.per.SqFt 變數的盒鬚圖，找出離群值並剔除。
3. 以多元迴歸分析方法找出影響公寓每平方呎價格的影響因素。

以下提供上述問題的提示：

1. 點選【繪圖】功能的【密度估計】，可以繪製單一變數圖形，也可以依群組繪製圖形。例如：可以依據 Boro (紐約市的區域類別) 變數分別繪製公寓每平方呎價格的密度估計圖。
2. 以【繪圖】功能繪製盒鬚圖後，能夠在圖中觀察到離群值，也可以在 RStudio 主控臺看到離群值的序號。
3. 點選【統計量】功能，再點選【模型配適】➔【線性迴歸】執行迴歸分析。以嘗試錯誤的方式，觀察不同因素的組合對於公寓每平方呎價格的影響是否顯著，找到最佳組合。例如：在迴規模式中，依變數選擇 Market.Value.per.SqFt，自變數選擇 Gross.Income.per.SqFt 及 Total.Units，則

輸出結果如下。由結果可知，截距及兩個字變數都顯著，整體模式的 $R^2 = 0.931$，p-value < 0.01 非常顯著。



house 資料集的變數包括：

- Neighborhood
- Building.Classification
- Total.Units
- Year.Built
- Gross.SqFt
- Estimated.Gross.Income
- Gross.Income.per.SqFt
- Estimated.Expense
- Expense.per.SqFt
- Net.Operating.Income
- Full.Market.Value
- Market.Value.per.SqFt
- Boro

圖5　紐約市公寓每平方呎價格的直方圖

4.　學生學習效果

105 學年度上學期本系開授資料統計與分析的課程於大三，我們以 R 軟體做為輔助教學軟體。欲採用 R Commander 套件，讓學生使用圖形化介面執行統計分析的功能。但是，因為上課的電腦教室所安裝的 R 軟體不是最新版本，且部分附加套件無法安裝，導致無法順利安裝 R Commander。上課改以 R Studio 軟體進行統計分析，輸入 R 軟體的程式敘述以繪圖及進行統計運算。上課時，介紹資料匯入的功能也遇到困難。由於無法直接匯入 Excel 檔案，有時也無法將 Excel 檔案轉換成 CSV 檔案匯入 R 軟體，所以才將資料從來源檔複製，再貼上於 R 軟體以建立資料變數。由於學生沒有學習過 R 軟體，對於輸入 R 的程式敘述進行統計分析比較陌生，上課時做隨堂練習有時花比較多時間。期中考有題目要學生使用 R 軟體繪圖，幾乎沒有學生答對。第一次作業中關於使用 R 軟體繪圖的題目，很少學生答對。

期中考後，我們改用 Excel 軟體做為輔助教學軟體。使用 Excel 繪圖，只需要選取資料然後選擇圖表類型，再設定相關參數即可，對學生而言比較容易。因此，再讓學生重新補寫第一次作業關於繪圖的題目，大部分學生都能夠完成。有學生反應，Excel 軟體比較簡單，R 軟體則比較困難。同樣地，讓學生補考期中考，繪圖題改用 Excel 軟體，則大部分學生都能夠答對一半以上。

Konrath 等人 [8] 提到，由於 R 軟體是一套程式語言，有幾篇文獻指出 R 軟體對於初學者比較困難。Jason Wilson 曾使用 R Commander 套件於統計相關課程中 [10]，Wilson 提到 R Commander 就像「統計學計算機」，我們可以很容易地進行敘述統計及推論統統計的分析計算。指令敘述式的 R 軟體較適合於需要進行資料分析的學生，也許是研究生。因此，這些學生也許

比較適合使用圖形化介面的軟體。若能使用 R Commander 圖形化的套件，學生應該能夠更容易地使用以解決統計問題。

5. 結論

本論文介紹了以 R 軟體輔助統計課程的教學以及學生的學習。R 軟體是一套免費軟體，也是一套程式語言，可以撰寫程式以執行工作，目前約有一萬種不同的套件可供下載使用。加裝 RStudio 整合式編輯環境軟體，可以更有效率地處理問題。R Commander 是一套用於統計分析的圖形化介面套件，不需要寫 R 的程式就可以分析資料，對於學習統計學或進行統計分析有很大助益。

依據我們在資料統計與分析課程的經驗，部分學生對於輸入程式敘述執行統計分析感到困難，但是對於圖形化介面的操作則感到易於使用，因此 R Commander 很適合這些學生。

6. 參考文獻

[1] 沈彥廷，「R 軟體入門」，臺北醫學大學生物統計研究中心 eNews，第 14 期，民國 105 年 8 月，頁 1-12。

[2] 湯銀才 主編，R 語言與統計分析，北京市，高等教育出版社，民國 97 年。

[3] L. Chandrakantha, “Excel Simulation as a Tool in Teaching Sampling Distributions in Introductory Statistics,” in Proceedings of the Ninth International Conference on Teaching Statistics (ICOTS9). Flagstaff, Arizona, USA, pp. 1-4, 2014.

[4] P. Cobb, “Where is the Mind? Constructivist and Sociocultural Perspectives on Mathematical Development,” Educational Research, Vol. 237, pp. 13-20, 1994.

[5] S. A. Culpepper and H. Aguinis, “R is for Revolution: A Cutting-Edge, Free, Open Source Statistical Package,” Organizational Research Methods, Vol. 14, No. 4, pp. 735-740, 2011.

[6] J. Fox, S. Weisberg, D. Adler, et al., “Package ‘car’,” Reference Manual, https://cran.r-project.org/web/packages/car/car.pdf, December, 2016.

[7] L. R. Hayrapetyan, “Excel-Based Interactive Teaching and Assessment Tool for Forecasting Methods,” World Journal of Social Sciences, Vol. 3, No. 3, pp. 89-97, 2013.

[8] A. C. Konrath, E. Henning, O. M. Formigoni Carvalho Walter, C. DaCunha Alves, and R. Wayne Samohyl, “Applications in teaching Statistical Quality Control with different R interfaces,” in Proceedings of IEEE Global Engineering Education Conference, Berlin, Germany, 2013, pp. 146-155.

[9] E. J. Wegman and J. L. Solka. "Statistical software for today and tomorrow," Encyclopedia of Statistics, John Wiley, 2005.

[10] J. Wilson, “Statistical computing with R: selecting the right tool for the job—R Commander or something else?” Wiley Interdisciplinary Reviews: Computational Statistics, Vol. 4, No. 6, pp. 518-526, 2012.