雲端資料分析暨導引系統(R-web)簡介

吳佩真 統計分析師

市面上常見的統計分析軟體包含 SAS、SPSS、R 等，各有其優缺點及擁護者。其中，R 軟體具有免費、自由、功能完整、擴充性強等優勢特性，近年來儼然成為最流行的統計軟體之一。然而，R 語言仍脫離不了一般程式語言的包袱，學習曲線略為陡峭；因此，對初學者而言，上手仍需經過一段艱辛的歷程。

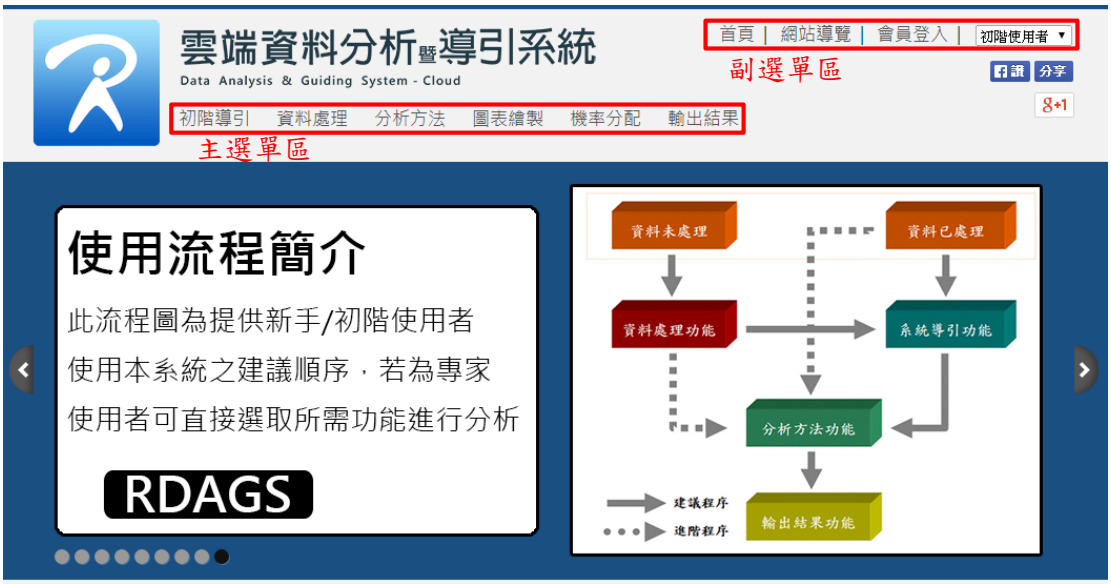
有鑑於此，R 軟體的圖形化使用者介面(Graphical User Interface, GUI)開發向來是許多熱心人士努力之目標。【雲端資料分析暨導引系統】(簡稱 R-web)即為由中華 R 軟體研發暨應用協會(CARRA)團隊所開發，以 R 語言結合網頁介面的一套雲端圖形化使用者介面資料分析系統。相較於一般統計軟體，R-web 具有以下幾項優勢：

- 具有導引系統功能，提供適當問題引導使用者針對資料特性選擇適當的分析方法，無統計學相關背景亦能輕鬆操作。
- 採用圖形化介面呈現，不須撰寫任何程式，讓缺乏程式語言背景的使用者亦能輕鬆分析現有的資料。
- 可針對巨量資料(big data)進行處理與分析。
- 提供三種使用者身分介面，系統依據使用者對資料分析的熟悉程度，引導至適合之分析方法。
- 針對每一種分析方法設有完整的分析範例及影音教學，讓使用者無需費時即能輕鬆上手。
- 執行程序與輸出結果以網頁呈現，因此只要在網路環境下，無論是手機或平板電腦皆可操作 R-web。

由於 R-web 提供完整的內容及範例資料檔，非常適合作為教學資源及學生自學使用。接下來的章節中，我們也將使用 R-web 為讀者們進行各種統計方法說明及操作示範，就讓我們一起進入『玩』統計的世界吧！

環境說明

R-web 為一雲端系統，免下載、免安裝，使用時僅需在網路環境下連結網址 http://www.r-web.com.tw 即可進入首頁(建議使用 Google Chrome 瀏覽器)。若為 R-web 雲端計算聯盟會員[1]，請另行由雲端計算聯盟入口登入。



雲端資料分析暨導引系統(Data Analysis & Guiding System-Cloud, DAGS-C)

在網路上以視窗點選方式操作，資料處理及分析零負擔。

上圖為 R-web 首頁畫面，功能選單列於網頁上方，可分為主選單及副選單兩區。右上方(副選單區)除了網站導覽功能，另可進行會員登入及使用者身分切換，初次使用者可先行註冊免費會員。身分選擇包含三種使用者身分依次為：

- **新手使用者**：適合較不熟悉基本統計知識的使用者。使用者可能未修習資料分析相關之科目，或對資料分析完全陌生。

- **初階使用者：**適合具備基本資料分析及統計知識的使用者。修習過至少一學期的統計學，知道何謂變數、參數等。
- **專家使用者：**適合經常使用資料分析軟體、熟悉各種分析方法的使用者，或從事資料分析之專業人員。

主選單區則包含所有資料處理及分析方法模組，選單各功能分述如下表：

系統功能	功能介紹
初階(新手)導引	提供漸進式問題，引導使用者選擇適合的資料分析方法。僅提供『新手使用者』與『初階使用者』使用。
資料處理	將資料檔進行各式資料處理，例如：對資料做不同條件之篩選與排序、對變數進行分組、字串及日期時間的處理或其他計算。
分析方法	依據使用者身分別的差異(即：對資料分析之熟悉程度)，提供適合的資料分析方法。
圖表繪製	將資料繪製成圖形或表格，幫助使用者確切地掌握資料的特性。依據登入系統身分別的不同，方法會有些許差異。
機率分配	提供多種常見分配的機率、分位數計算，並有圖形繪製比較及生成隨機樣本功能。僅提供『初階使用者』與『專家使用者』使用。
輸出結果	幫助使用者完整的掌握及整理分析結果。

參考資料

[1] 目前雲端計算聯盟會員包含：臺北醫學大學、中國醫藥大學、衛生福利部、國立新竹教育大學、 致理技術學院、典通股份有限公司、輔仁大學統計資訊學系以及臺灣析數資訊

利用 R-web 做基本資料分析

您是否常對著蒐集來的資料發呆，無所適從；抑或是面臨當手邊握有分析結果卻毫無頭緒、不知如何解釋的困擾呢？分析資料時，首要步驟即是要先瞭解資料中每一個變數所屬的型態。本章內容分兩小節：1-1 心血管疾病資料[2] (即：以基隆社區為基礎的整合篩檢計畫，KICS)與變數描述；1-2 以心血管疾病資料為範例，結合 R-web 操作說明，帶領讀者邁進分析資料的第一步-描述性統計！

1-1 資料介紹 (Data Description)

使用資料來源以基隆社區為基礎的整合篩檢計畫(Keelung Community-based Integrated Screen Program，簡稱 KCIS)，目的為提供基隆地區 20-79 歲全體居民質量篩檢(mass screening)，包含腫瘤性(neoplastic)及非腫瘤性(non-neoplastic)疾病，同時也提供以人口為基礎之世代追蹤做為確認可能伴隨之癌症及慢性疾病(如心血管疾病)。此計畫起始於 1999 年 10 月，以年為單位的樣本採集，樣本收集自 1999 年底至 2004 年底，共 64489 筆樣本可供分析；其中，24051 位男性(37.29%)、40438 位女性(62.71%)。資料中變數包含：編號、性別、年齡、個人心血管疾病史、家族心血管疾病史、心臟收縮壓、心臟舒張壓、空腹葡萄糖、高密度脂蛋白、三酸甘油酯、腰圍、飲酒習慣、抽菸習慣、食用檳榔習慣以及每日菸草消費量，共 15 個變數。變數定義如表 1.1。

表 1.1 變數定義

變數	代碼(code)	定義/說明
編號	ID	
性別	Gender	男性：1、女性：0
年齡	Age	
腰圍	Waist	公分(cm)
心臟收縮壓	SysBP	毫米汞柱(mmHg)
心臟舒張壓	DiaBP	毫米汞柱(mmHg)
空腹葡萄糖	AC	毫克/分升(mg/dl)
高密度脂蛋白	HDL	毫克/分升(mg/dl)
三酸甘油酯	TG	毫克/分升(mg/dl)
嚼檳榔習慣	Betelnut	有：1、無：0
飲酒習慣	Alc_Drink	有：1、無：0
個人心血管疾病史	CVD	有：1、無：0
家族心血管疾病史	FamilyHx	有：1、無：0
抽菸習慣	Tobacco	有：1、無：0
菸草消費量	Tobacco_Consumption	無：0、每日一包：1、每日兩包：2、每日三包以上：3

1-2 描述性統計 (Descriptive Statistics)

描述資料是非常重要的工作，除了初步瞭解資料分布概況，亦可引導研究者選擇適當的資料分析方法，協助驗證判斷的準確性。透過分析，對每個變數的描述，利用平均數、中位數及眾數等集中量數來瞭解資料的集中趨勢；利用變異數(標準差)、變異係數、全距、四分位數等離散量數來瞭解資料的分散程度；利用峰態和偏態係數瞭解資料是否存在特殊分配之特性(如：常態、卜瓦松分配等)；另外，將資料利用散佈圖、直方圖、盒鬚圖及莖葉圖等圖像化處理亦可瞭解資料整體分佈(圖表部份於下期 e-報說明)。本節針對上述心血管疾病資料，利用 R-web 詳細介紹資料分布屬性。

在 R-web 分析資料前，首要步驟為上傳個人資料檔案，說明如下：

(1) 上傳資料 (資料來源：1-1 節--KCIS 之心血管疾病資料檔)

點選資料處理→管理資料檔→上傳資料檔

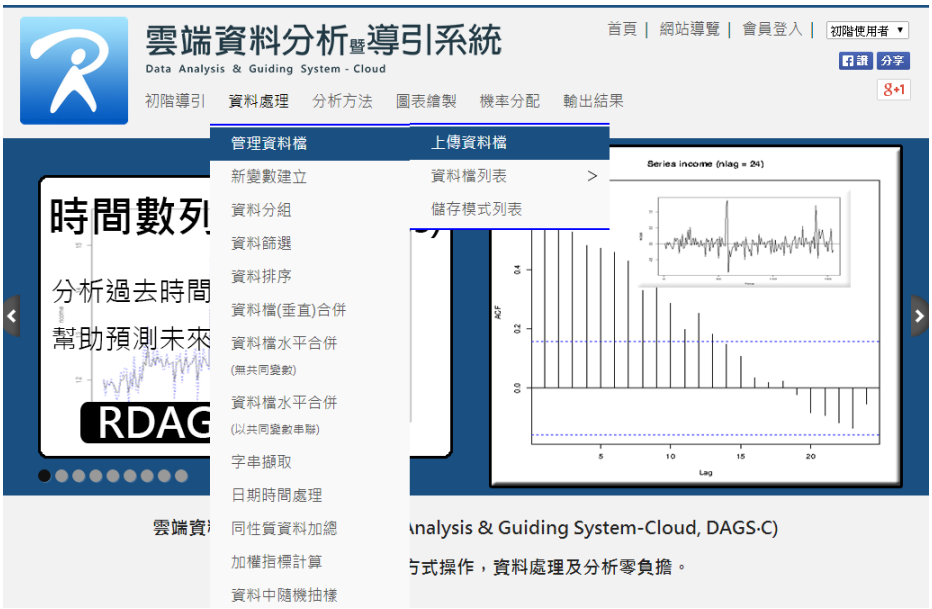


圖 1.1 選擇上傳資料檔

(2) 選取要上傳的資料檔後，點選「確認上傳」。

圖 1.2 選擇資料檔

(3) 此步驟會依照使用者上傳的資料格式不同，而有不同的參數供輸入設定，如遺失值(Missing value)表示符號、資料讀取筆數等，設定完成後，執行「下一步驟」。

圖 1.3 上傳資料內容設定

(4) 可修改資料檔名稱、重新命名變數及變更變數型態(即數值或類別)，設定完成後，點選「確認儲存」(如圖 1.4 所示)。於使用者個人資料檔列表中，確認是否成功上傳資料檔(如圖 1.5 所示)。

圖 1.4 修改資料名稱及變數型態

圖 1.5 個人資料檔列表

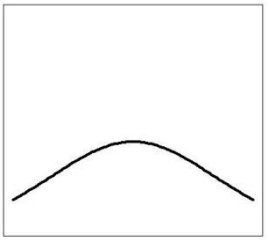
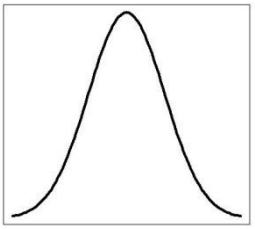
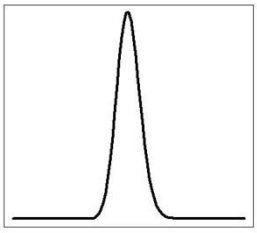
成功上傳心血管疾病資料檔後，便可進行摘要統計計算。R-web 提供之統計量(Statistics)包含四大項：

1. 基本資訊：樣本數(Sample size)、總合(Summation)
2. 集中量數：平均數(Mean)、中位數(Median)、眾數(Mode)
   - 平均數用來衡量各資料值相對集中較多的中心位置
   - 中位數是將一群資料由小至大排序後，位於中心位置的數值
   - 眾數為一組資料中出現次數最多的數值
3. 離散量數：變異數(Variance)、標準差(Standard Deviation)、全距(Range)、最大值、最小值、內四分位距(Interquartile Range, IQR)、第一四分位數(Q1)、第三四分位數(Q3)、變異係數(Coefficient of Variation, CV)等
   - 變異數為一組資料中各數值相對於該組資料之平均值的分散程度
   - 標準差亦是衡量一組資料的分散程度，為變異數的平方根
   - 全距為一組資料中最大與最小值之差
   - 內四分位距為一組資料中第一四分位數(Q1)與第三四分位數(Q3)之差

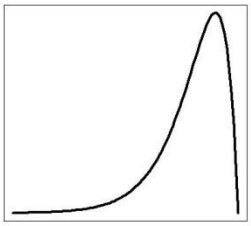
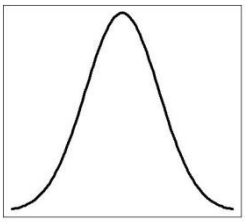
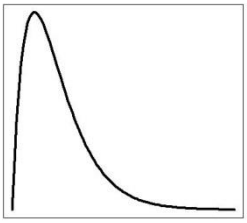
- 變異係數用來比較兩組或兩組以上資料(資料的單位與取值範圍可不相同)之離散程度

4. 分配描述：峰態(Kurtosis)係數、偏態(Skewness)係數

- 峰態衡量一組資料分配高狹或低闊程度

峰態係數<0	峰態係數=0	峰態係數>0
		

- 偏態衡量一組資料分配的對稱性。

偏態係數<0	偏態係數=0	偏態係數>0
		

(5) 於 R-web 主選單區選取分析方法→摘要統計。

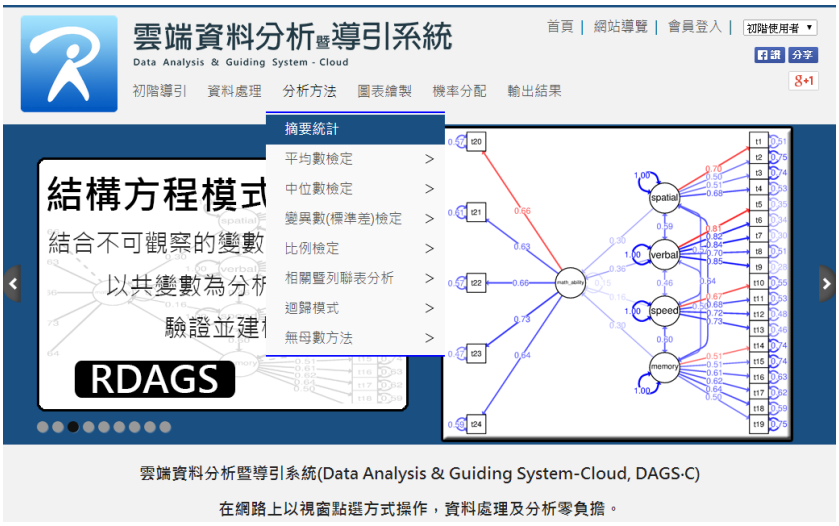


圖 1.6 選擇摘要統計

(6) 選擇心血管疾病資料檔(即 CVD)後，選取欲進行分析的變數，進階選項設定可依個人需求選取統計量及分組變數(分別如圖 1.7、圖 1.8 及圖 1.9 所示)。

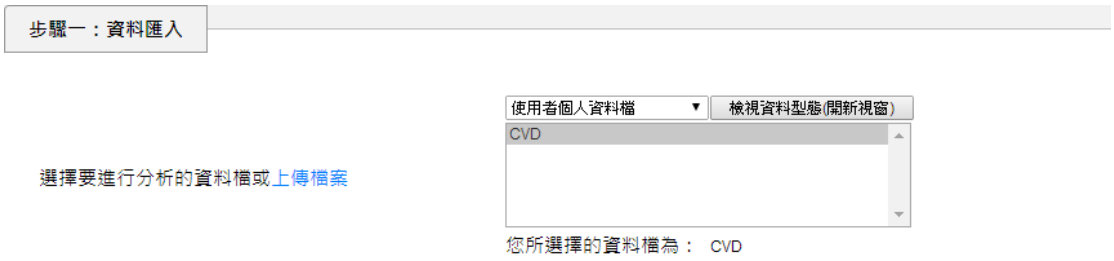


圖 1.7 選擇欲分析之資料檔

圖 1.8 選擇欲分析的變數

圖 1.9 選擇統計量及分組變數

(7) 設定完成後點選「開始分析」，輸出結果如圖 1.10 所示。

圖 1.10 輸出結果

表 1.2 為 CVD 資料檔中數值資料(Numerical data)之統計量。表中顯示，自 1999 年底至 2004 年底接受 KCIS 篩檢計畫的 64489 位參予者中，平均年齡為 46.8 歲，約 68%的參予者年齡範圍介於 32.9 至 60.7 歲之間；平均腰圍為 78.34 公分，75%參予者腰圍小於 86 公分(Q3)；平均收縮壓及舒張壓值分別為 123.27mmHg 和 77mmHg，介在正常範圍(90mmHg,140mmHg)和(50mmHg,90mmHg)之內；在 R-web 中，當一群資料偏態或峰態係數接近 0

時，資料可能近似於常態分佈，參予者所測出的空腹葡萄糖(AC)與三酸甘油酯(TG)之峰態係數分別高達 36.22 和 122.01，偏態係數分別為 5.24 和 7.31，意味在這兩個指標下大部分參予者集中在中間偏低的區塊。

表 1.2 敘述統計：CVD 資料

變數名稱	Age	Waist	SysBP	DiaBP	AC	HDL	TG
樣本數	64484	62852	63256	63245	60978	60084	60891
平均數	46.82	78.34	123.27	78.07	93.16	57.31	121.07
中位數	45	78	120.5	77	87	57	92
眾數	40	70	120	70	87	63	63
標準差	13.90	10.67	20.82	11.98	28.93	12.19	111.08
變異數	193.09	113.95	433.59	143.42	836.83	148.50	12337.68
全距	61	142	206	100	557	144	4126
最小值	19	37	70	40	49	10	11
最大值	80	179	276	140	606	154	4137
第一四分位數	36	70	108.5	69.5	81	50	63
第三四分位數	57	86	135	85	94	63	142
內四分位距	21	16	26.5	15.5	13	13	79
峰態係數	-0.66	0.74	1.17	0.70	36.22	1.41	122.01
偏態係數	0.30	0.45	0.80	0.60	5.24	0.53	7.31

表 1.3 顯示以性別為分組變數(男性：1、女性：0)所得出的統計量。男性參予者平均年齡為 48.3 歲，女性平均年齡則為 45.9 歲；75%的男性與女性參予者腰圍皆於正常範圍內(Q3)；假設血壓值分布為常態分佈，由 Mean±SD 發現 68%女性的血壓值表現正常，男性存在著血壓較高之困擾；從偏態、峰態係數及四分位數值發現大部分男性與女性空腹葡萄糖值集中在正常偏低的範圍。

利用描述性統計量說明資料特性是對資料認識最基本且重要的方法，檢視這些統計量後，便能幫助研究者選擇合適的資料分析方法!

表 1.3 敘述統計：CVD 資料(以性別為分組變數)

變數名稱	Gender	Age	Waist	SysBP	DiaBP	AC	HDL	TG
樣本數	G=0	40435	39435	39664	39671	38291	37690	38232
	G=1	24049	23417	23592	23574	22687	22394	22659
平均數	G=0	45.92	74.76	119.40	75.93	92.46	60.71	105.50
	G=1	48.34	84.37	129.77	81.68	94.34	51.58	147.35
中位數	G=0	45	74	116	74.5	87	61	82
	G=1	47	84	127	80.5	88	51	113
眾數	G=0	38	70	110	70	87	63	58
	G=1	44	84	120	80	87	52	77
標準差	G=0	13.48	9.71	20.47	11.51	28.42	11.56	88.48
	G=1	14.44	9.44	19.76	11.88	29.73	11.01	137.28
全距	G=0	61	138	206	100	483	139.6	3065
	G=1	61	139	183	99	555	144	4126
最小值	G=0	19	37	70	40	49	14	15
	G=1	19	40	71	40	51	10	11
最大值	G=0	80	175	276	140	532	153.6	3080
	G=1	80	179	254	139	606	154	4137
第一四分	G=0	36	68	105	68	81	54.6	58
位數	G=1	38	78	116	73.5	82	45	77
第三四分	G=0	55	80.5	130.5	82.5	94	65	124
位數	G=1	60	90	141	89	96	56	173
內四分位距	G=0	19	12.5	25.5	14.5	13	10.4	66
	G=1	22	12	25	15.5	14	11	96
峰態係數	G=0	-0.57	2.05	1.41	0.91	38.42	1.77	103.83
	G=1	-0.80	1.28	1.41	0.67	33.08	2.63	105.54
偏態係數	G=0	0.33	0.79	0.93	0.67	5.45	0.54	6.65
	G=1	0.23	0.25	0.84	0.55	4.93	0.82	7.06

備註.G=1：男性、G=0：女性

參考資料

[2] Liu YM, Chen LS, Yen MF, Chen HH. Individual risk prediction model for incident cardiovascular disease: A Bayesian clinical reasoning approach. Int J Cardiol 2013;167:2008-12