R-web 資料分析應用：無母數方法

沈彥廷　副統計分析師

R-web 資料分析應用專欄自首期生統 eNews 刊載至今，基本上已將一般基礎統計學的範疇含括在內，包括有簡單的描述性統計、視覺化呈現以及在各式資料形態下的參數檢定和模型建立。

然而，我們先前介紹過的方法大多皆是建立在母體分配為常態分配或某一特定分配的假設下。若是當樣本資料太少、母體分配不明或是違反分配假設時，我們即必須改考慮採用「無母數(Non-Parametric)方法」進行分析。無母數分析方法最大的特點顧名思義即為不需假設母體的分配，其統計量的抽樣分配與母體分配無關，雖然因為缺乏分配的訊息而導致推論可能較不精確，但對於資料型態適用性的限制也能保有相對較大的彈性。

本期內容就將為讀者介紹一些常見的無母數分析方法：連檢定、機率分配檢定、適合度檢定，並搭配「雲端資料分析暨導引系統」(R-web, http://www.r-web.com.tw/)作為分析工具，以實際資料進行案例演示和操作說明。

➢ 單一樣本連檢定

連檢定(Run Test)主要用於檢定一組資料是否為隨機產生。在許多統計方法的前提假設中都有隨機性的假設，例如要檢驗一個迴歸模式的殘差是否為獨立隨機，此時連檢定即可派上用場。

在進行連檢定時，首先要將數值資料依特定集中量數(如平均數、中位數)為切點將資料切分為兩個組別，小於切點值的樣本給予"-"符號；大於切點值的樣本則定義為"+"。緊接著計算「連數」，一個「連」代表具有一個或多個連續相同符號的數列，例如"+++"為一個連，"+++--"有兩個連，"+++--++"則為三個連。透過連數的多寡我們即可衡量資料的隨機性，過多或過少的連數可能都代表著資料存在不隨機的情形。

底下以一個實際資料進行案例說明。我們在本章節將統一使用「源自基隆社區為基礎的整合篩檢計畫」(Keelung Community-based Integrated Screen Program, KCIS)作為範例資料檔，有關此資料的詳細資訊及變數定義請參閱首期生統 eNews。

假設資料中各受試者的年齡(Age)是依資料順序收集而得，試問此資料中年齡是否為一隨機樣本？

我們可以在選單中點選【分析方法】→【無母數方法】→【單一樣本連檢定】來進行分析。首先選擇 CVD 作為要進行分析的資料檔，接著選擇 Age 作為要進行檢定的變數。在進階選項中，可以設定使用平均數、中位數或任意自訂值作為資料切點，預設為使用平均數。確認參數設定無誤後，點選開始分析。

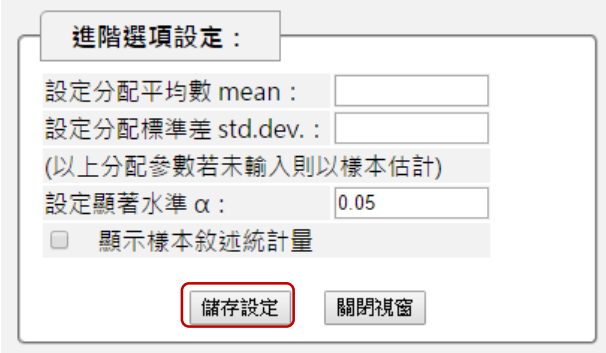
在連檢定中，虛無假設H_0為：資料順序為隨機產生。根據分析結果可以看到，以平均數 46.82 為切點的情況下，資料共有 4736 個連數，p 值遠小於我們所設定的顯著水準 0.05，因此可拒絕虛無假設，也就是說年齡並不是一組隨機資料。

➢ 單一樣本機率分配檢定

機率分配檢定的用途即是用來檢查資料是否服從於某一特定理論分配，在實務上我們可以利用許多不同的適合度(goodness of fit)檢定方法達到這個目的，在這邊我們要介紹的是其中一個常見的方法：Kolmogorov-Smirnov 檢定，亦常被簡稱為 K-S 檢定。

K-S 檢定的原理是比較樣本資料和理論分配的累積分佈函數(CDF)之間的最大差異，若資料確實服從於某特定分配，則此差異值就不應該會太大；反之，若兩者的累積分佈形狀或位置出現明顯的偏離，則表示資料來自於該分配的可能性並不高。

以實際資料為例，欲了解範例資料檔中空腹葡萄糖(AC)變數是否服從常態分配，可以由 R-web 選單點選【分析方法】→【無母數方法】→【單一樣本機率分配檢定】來進行分析。選擇資料檔和欲進行分析的變數，檢定分配則依題意選擇常態分配。進階選項中可自訂理論分配的參數，若未設定則系統自動以樣本估計，此處可設定的參數會因選擇的檢定分配不同而有所差異。確認參數設定無誤後，點選開始分析。



單一樣本機率分配檢定 - 分析結果

- **分析方法**：單一樣本機率分配檢定
- **資料名稱**：CVD
- **變數名稱**：AC
- **顯著水準**：0.05
- **檢定分配**：常態分配
- **計算時間**：0.06秒
- **單一樣本機率分配檢定**：

虛無假設：母體分配為常態分配				
變數名稱 variable	分配參數一 平均數	分配參數二 標準差	Kolmogorov-Smirov D 檢定統計量 Kolmogorov-Smirov D-statistics	p-值[I] p-value
AC	93.1598	28.9281	0.264	< 2.22e-16 ***

I：顯著性代碼： '***' : < 0.001, '**' : < 0.01, '*' : < 0.05, '#' : < 0.1

- **分析結果建議**：由於檢定結果P-值(< 2.22e-16) < 顯著水準0.05，因此可拒絕虛無假設。

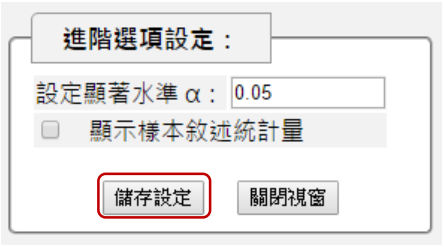
在單一樣本機率分配檢定中，虛無假設H_0為：母體分配為常態分配。根據檢定結果，樣本資料和理論常態分配的累積分佈最大差異 D 值為 0.264 且 p 值趨近於 0。因此可拒絕虛無假設，表示空腹葡萄糖樣本並不服從於常態分配。

➢ (獨立)雙樣本機率分配差異檢定

有時研究者感興趣的並不是樣本資料是否服從某特定分配，而是兩組資料是否來自於相同的母體分配，此時我們就可以使用雙樣本 K-S 檢定來進行分析。雙樣本 K-S 檢定與單樣本 K-S 檢定原理相同，唯一的差別是雙樣本 K-S 檢定比較的是兩組樣本間的累積機率分佈差異最大值，而非與特定理論分配函數作比較。

我們直接來看一個案例操作說明。假設研究者想知道不論性別(Gender)是男是女，高密度脂蛋白(HDL)是否都來自於相同的分配？此時我們可以選擇【分析方法】→【無母數方法】→【(獨立)雙樣本機率分配差異檢定】

來進行分析。首先在步驟一選擇資料檔，接著由於我們要依照資料中的性別變數區分兩組高密度脂蛋白樣本，因此在步驟二中可選擇資料型態為「一檢定變數及一分組變數」，最後在步驟三中選擇 HDL 為檢定變數、Gender 為分組變數。確認參數設定無誤後，點選開始分析。



(獨立)雙樣本機率分配差異檢定 - 分析結果

- **分析方法**：(獨立)雙樣本機率分配差異檢定
- **資料名稱**：CVD
- **檢定變數**：HDL
- **分組變數**：Gender(0, 1)
- **顯著水準**：0.05
- **計算時間**：0.123秒

- **雙樣本機率分配差異檢定**[I]：

虛無假設：兩組資料來自相同母體分配		
變數名稱 variable	Kolmogorov-Smirov D 檢定統計量 Kolmogorov-Smirov D-statistics	p-值[II] p-value
HDL	0.4426	< 2.22e-16 ***

I：分組變數為Gender

II：顯著性代碼： '***' : < 0.001, '**' : < 0.01, '*' : < 0.05, '#' : < 0.1

- **分析結果建議**：由於檢定結果P-值為(< 2.22e-16) < 顯著水準0.05，因此可拒絕虛無假設。

在雙樣本機率分配差異檢定中，虛無假設H_0為：兩組資料來自相同母體分配。根據檢定結果，兩組樣本資料的累積分佈最大差異 D 值為 0.4426 且 p 值趨近於 0。因此可拒絕虛無假設，表示不同性別下的高密度脂蛋白並非來自於相同分配。

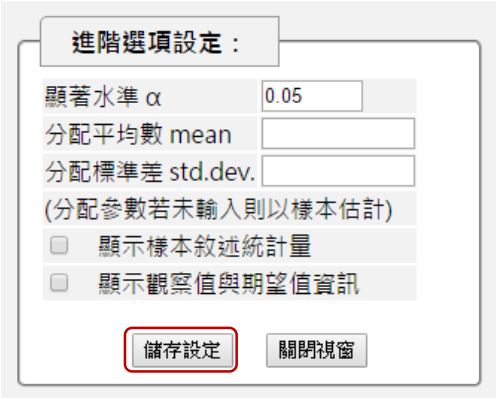
➢ 卡方適合度檢定

在第九期生統 eNews 中，我們曾介紹過如何使用列聯表分析方法中的卡方獨立性檢定來檢視兩類別變數間的相關性。現在要為各位讀者說明的則是卡方分析的另一項用途：適合度檢定。與前面所介紹的 K-S 檢定相同，卡方適合度檢定同樣用於驗證一組資料是否服從於一特定理論分配，然而由於使用卡方檢定時須將資料整理成列聯表型式，因此若原始資料為連續型數值變數時，則必須先進行資料分組(離散化)作業。

在計算上，其概念是藉由比較列聯表每個細格的「觀察次數」及「期望次數」的差異，計算出一個卡方統計量，其中期望次數即為根據設定的

母體機率分配求算而得。此統計量越大，代表觀察次數和期望次數之間的差異很大，則此時可認為樣本資料的次數分配異於理論上的分配。

同樣以 CVD 資料為例，研究者希望透過卡方檢定檢驗三酸甘油酯(TG)變數是否服從於常態分配。則我們可以點選 R-web 選單中的【分析方法】→【無母數方法】→【卡方適合度檢定】來進行分析。選擇資料檔和欲進行分析的變數，由於三酸甘油酯為一連續型數值變數，因此需設定資料分組的切割方式，檢定分配則依題意選擇常態分配。進階選項中可自訂理論分配的參數，若未設定則系統自動以樣本估計，此處可設定的參數會因選擇的檢定分配不同而有所差異。確認參數設定無誤後，點選開始分析。



卡方適合度檢定 - 分析結果

- **分析方法**：卡方適合度檢定
- **資料名稱**：CVD
- **變數名稱**：TG
- **顯著水準**：0.05
- **檢定分配**：常態分配
- **計算時間**：0.091秒
- **卡方適合度檢定**：

虛無假設：母體分配為常態分配					
變數名稱 variable	分配參數一 平均數	分配參數二 標準差	卡方檢定統計量 chi-square statistics	自由度 d.f.	p-值[1] p-value
TG	121.073	111.0751	27231.37	2	< 2.22e-16 ***

1：顯著性代碼： '***' : < 0.001, '**' : < 0.01, '*' : < 0.05, '#' : < 0.1

- **分析結果建議**：由於檢定結果P-值(< 2.22e-16) < 顯著水準0.05，因此可拒絕虛無假設。

在卡方適合度檢定中，虛無假設H_0為：母體分配為常態分配。根據檢定結果，卡方檢定統計量為 27231.37、自由度為 2，p 值遠小於顯著水準 0.05，因此可拒絕虛無假設，也就是說三酸甘油酯並不服從於常態分配。

以上介紹的是較常見的幾種無母數分析方法，除此之外，諸如 Sign Test、Mann-Whitney Test，或是與 K-S 檢定同樣用於機率分配檢定的 Anderson-Darling Test 等許多方法也都是屬於無母數方法的範圍。事實上，在第五期生統 eNews 中 R-web 資料分析應用專欄所介紹的「中位數檢定」方法亦為

無母數性質的檢定方法，有興趣進一步了解的讀者歡迎前往參閱，我們就不在此多加著墨了。

在實務的資料分析案例中，無母數方法經常是相當實用的工具之一，希望本篇的說明能對您有所幫助。也期望讀者可以親自嘗試使用 R-web 熟悉本次所介紹的各項方法操作，相信一定能更加掌握各方法的使用時機和其意義！