

專題: Chi-square Goodness of Fit Tests: A Simulation Practice

目的:

卡方適合度檢定 (Chi-square Goodness of Fit) 是個古老但是有效的檢定方法。許多近代的檢定方法仍以卡方適合度檢定為核心思想做調整, 達到特殊的檢定需求。對於卡方適合度檢定的了解有助統計理論的研究與開發。

為什麼:

統計方法的應用與研究是兩回事, 一個將統計方法應用在實務資料上, 另一個針對實務問題提出解決的統計方法。一般而言, 在大學階段的統計學習較著重於理論的理解與方法的應用, 對於學術研究的過程鮮少碰觸。每個可以被用來解決問題的統計方法除了必須在理論上被證明可行外, 通常還需要透過大量模擬資料的驗證, 方能說服使用者, 且運用這個方法所得到的推論才有意義。本單元利用解決一個簡單且眾所周知的問題, 讓初學者初嚐研究中常見的模擬過程, 祈能順利進入統計研究的殿堂, 成為指導老師的好幫手。

做什麼:

統計學者提出一個統計方法, 固然需要理論根據, 但實務上的效果更是重要。所謂真金不怕火煉, 每個方法都要經過多方試煉, 才能證實為可行的統計方法。但問題在於實務上我們常不能提供足夠的真實資料做為測試的依據, 為在短時間內提供其效力證明, 符合問題假設的各種模擬資料便取而代之。本單元要求初學者利用電腦程式模擬符合問題假設的各種資料, 並且大量地重複執行, 讓該統計方法的「統計現象」可以呈現出來。以假設檢定而言, 我們關心一個檢定統計量對於型一誤的維持及其檢定力的表現。

怎麼做:

初學者必須先了解適合度檢定的來龍去脈，能自行推論出檢定統計量。從過程中可以發現檢定統計量與幾個條件有關，譬如，樣本大小、分組數量、分組方式 (equal probability, equal width) 等。這些條件對「檢定的品質」有何影響？能直接從統計式中看出來嗎？

有些問題可以從理論推論而得，有些必須依賴電腦模擬。本專題將利用電腦模擬的方式回答上述問題，其中「檢定品質」指檢定統計式的型一誤及其檢定力的表現。模擬的過程有幾件事要做：

- 大量模擬資料的產生 (包括符合 H_0 與 H_a)。在模擬的需求下，我們必須產生具某種分配的樣本 x ，每次 N 筆，每筆 n 個，以便執行 N 次的檢定，紀錄 N 次拒絕檢定結果。
- 計算 N 筆樣本的統計量並觀察這 N 個統計量的分佈情況 (是否與理論上對檢定統計式分配的推論一致，即服從某種自由度的卡方分配)。
- 樣本數 n 的影響有多大呢？
- 將樣本空間分成 k 組, A_1, A_2, \dots, A_k ，的方式可採 equal probability 或 equal width，哪一種比較好呢？該怎麼分呢？分成多少組比較好呢？這裡有太多選擇，不容易從理論窺探。
- 型一誤 (α) 的維持情況: 在給定型一誤 α 下進行檢定，如果拒絕 H_0 的資料有 N_1 筆，我們想知道錯誤拒絕的比例 N_1/N 是否接近 α 值。
- Power 的優劣評估: 同上，當資料來自 H_a 時, N_1/N 代表檢定力，越接近 1 越好。

既是模擬，次數 N 當然操之在己，一般而言越大越好，不過當 N 大到某個程度，其結果將趨於穩定，過大只是徒增計算時間。所以模擬次數的多寡只求適當即可，譬如 $N = 50000$ 次。

背景：

在統計的應用裡，我們常需要知道某組樣本是否服從某個分配。假設檢定寫成

H_0 : The data follow a specified distribution.

H_a : The data do not follow the specified distribution.

卡方適合度檢定的原理將樣本資料分配與理論分配做比較。將樣本資料分成 k 組，每一組的觀察個數為 $O_i, i = 1, 2, \dots, k$ ，其中 $\sum_{i=1}^k O_i = n$ ，即樣本數。對應組範圍內的機率為 p_i 。則統計量

$$\chi^2 = \sum_{i=1}^k \frac{(O_i - np_i)^2}{np_i} \sim \chi_{k-l-1}^2$$

其中 l 表示需要估計的未知參數個數。

研究假設與限制:

每個研究主題都有其假設與限制，「卡方適合度檢定」有哪些假設與限制呢？不同的假設或限制會造成什麼影響呢？譬如，當 $np_i < 5$ 的情況，對檢定力的影響？唯有透過徹底的資料模擬，才能對所有的假設與限制有深刻的瞭解，才能體會後續的許多統計研究到底在做些什麼，統計學課本的描述只能點到為止。