行政院國家科學委員會專題研究計畫　成果報告

協和配對樣本在 McNemar 檢定中的角色扮演探討

計畫類別：個別型計畫
計畫編號：NSC91-2118-M-004-005-
執行期間：91 年 08 月 01 日至 92 年 07 月 31 日
執行單位：國立政治大學統計學系

計畫主持人：江振東

報告類型：精簡報告

處理方式：本計畫可公開查詢

中　華　民　國 92 年 11 月 7 日

行政院國家科學委員會專題研究計畫成果報告

協和配對樣本在 McNemar 檢定中的角色扮演探討

On Concordant Paris in McNemar's Test

計畫編號：NSC 91-2118-M-004-005

執行期限：91 年 7 月 1 日至 92 年 8 月 31 日

主持人：江振東 國立政治大學統計系

一、計畫中文摘要

如果我們想要就一組相依樣本前後兩次調查所得到的比例是否有所差異進行探討，最常用的方法莫過於 McNemar 的檢定方法。然而 McNemar 檢定統計量基本上是一個非協和配對樣本的函數，與協和配對樣本完全無關。這個與直覺有些不符的想法，常常造成實際應用的困擾。在此計畫中，我們藉由絕對多數和相對多數的思考模式來重新詮釋此一問題。此外我們也推導出一個新的統計量，並利用模擬實驗與 McNemar 檢定來作比較。

關鍵詞：相依樣本、McNemar 檢定、非協和配對樣本、協和配對樣本

Abstract

When data arise as matched binary pairs, it is often of interest to compare the two correlated binomial proportions. McNemar's test is perhaps the best known for this matter. McNemar's test statistic itself, however, has nothing to do with the concordant pairs. It is somewhat counter-intuitive, and often frustrating for practitioners to learn that only the discordant pairs are necessary, given the effort to collect all the data. In this study, we clearify the ambiguity through the concept of the majority preference. We also derive a new test statistic and compare its small-sample behavior with McNemar's test through simulation.

Keywords: concordant pairs, correlated pairs, discordant pairs, McNemar's test.

二、計畫緣由與目的

在日常生活中，我們常常可以藉由媒體的報導，得到下列類似的訊息：「在某項政策採行之後，民眾對政府的施政滿意度的比例由 59%降為 55%；相對的不滿意的程度則由 41%提高為 45%。」假設前述的資料是針對同一組樣本在事件發生前後的兩次訪談之後所得到的結果，則 4%的差距是否足以說明民眾對於政府的施政滿意程度在該項政策施行之後有明顯的轉變呢？

就這一類有關相依樣本(correlated sample)的處理方式，在統計上最常用的工具莫過於 McNemar 的檢定方法(McNemar (1947))。假定針對一組樣本數為 n 的樣本，依據前後兩次的訪談的結果，我們可以整理得到一個 2×2 的列聯表如下：

第一次調查	第二次調查		總和
	滿意(0)	不滿意(1)	
滿意(0)	n_{00}	n_{01}	n_{0+}
不滿意(1)	n_{10}	n_{11}	n_{1+}
總和	n_{+0}	n_{+1}	n

其中 $n_{ij}\,(i, j = 0,1)$ 表示在第一次調查中選

擇第 i 個選項，而第二次選擇第 j 個選項的樣本數。我們的目的是想要瞭解藉由前後兩次調查結果是否可以反應出滿意程度有所改變(亦即 $H_0:\pi_{0+}=\pi_{+0}$)。由於 McNemar 的檢定統計量為 $X^2=\frac{(n_{01}-n_{10})^2}{n_{01}+n_{10}}$，因此只要 $X^2>\chi^2_{1,\alpha}$(其中α為右尾機率)，則就統計的觀點而言，我們便可以說兩次調查的結果有所不同。但是在使用 McNemar 的檢定過程中，我們也可以發現只要 n_{01}，n_{10}(非協和配對樣本(discordant pairs))的數據不變，縱然 n_{00}，n_{11}(協和配對樣本(concordant pairs))有所改變，檢定統計量的值也永遠不會改變。這似乎與我們的直覺有些違背，因為 n_{00} 相較的越大，我們似乎會越傾向於不拒絕虛無假設 H_0。然而 McNemar 的檢定結果，直觀上顯然與 n_{00}，n_{11} 的大小完全無關，因而常會導致「主對角線上(亦即協和配對樣本)所呈現的資訊似乎有被浪費掉的感覺」的疑惑，從而質疑 McNemar 檢定的好壞，因此我們希望能夠藉由此研究計畫的執行，對此困惑有所澄清。

針對如何引進協和樣本的資訊來作檢定，相關文獻似乎並不太多。Liang and Zeger(1988) 曾提出統計量 $\hat{\varphi}_{LZ}=\frac{nn_{01}+\omega(n_{00}n_{11}-n_{01}n_{10})}{nn_{10}+\omega(n_{00}n_{11}-n_{01}n_{10})}$，其中 ω 是一個 $\{n_{00},n_{01},n_{10},n_{11}\}$ 的函數。由於 $\hat{\varphi}_{LZ}$ 是優勢比 φ (odds ratio)的一個估計式，因此檢定 $H_0:\pi_{0+}=\pi_{+0}$ 相當於檢定 $H_0:\varphi=1$。不過 $\hat{\varphi}_{Lt}$ 這個統計量只有在 $n_{00}n_{11}-n_{01}n_{10}>0$，也就是 $0\le\frac{n_{01}n_{10}}{n_{00}n_{11}}\le 1$ 時才能適用，因此減弱了它的實用性。此外一般所熟知的 Wald 檢定統計量 $X_w^2=\frac{(n_{01}-n_{10})^2}{n_{01}+n_{10}-\frac{(n_{01}-n_{10})^2}{n}}$，由於包含有 $n(=n_{00}+n_{01}+n_{10}+n_{10}+n_{11})$，因此的確考慮了 n_{00} 及 n_{11} 的影響。但是在樣本總數不夠大的情形下，使用 X_w^2 來作為檢定會導致較大的型 I 誤差。於此，我們可以了解 X_w^2 或者 $\hat{\varphi}_{LZ}$ 都是著眼於如何引進 n_{00}，n_{11} 於 $H_0:\pi_{0+}=\pi_{+0}$ 的檢定中。

此外 Randles (2001)針對符號檢定(Sign Test) 中的零值 (zeros) 及 Wilcoxon-Mann-Whitney 檢定中的同分值(ties)是否捨棄不用的問題提出另一種思考方向。假定 $x_1,\cdots,x_n$ 是一組隨機樣本。令 $\pi_+=P(x_i>0)$，$\pi_-=P(x_i<0)$，$\pi_0=P(x_i=0)$；而 n_+,n_-,n_0 則分別表示樣本中大於 0，小於 0 及等於 0 的個數。Randles 認為傳統的檢定方法應該只適用在我們想要回答的問題是 $H_0:\pi_+\le\pi_-$ vs. $H_1:\pi_+>\pi_-$，也就是 $x_i>0$ 是否是相對多數。但是如果我們想要瞭解的是 $x_i>0$ 是否是絕對多數(the majority preferance)，則 $H_0:\pi_+\le\pi_-+\pi_0$ vs. $H_1:\pi_+>\pi_-+\pi_0$，或者是 $H_0:\pi_+\le\frac{1}{2}$ vs. $H_1:\pi_+>\frac{1}{2}$，應該才是正確的陳述方式。依據這個想法，Randles 從而推導出另一個統計量來回答這個問題。儘管這種思維模式未必完全能為大家所接受，但也不失為一種可行方式。由於在這裡有關零值的取捨問題，和前述相依樣本的處理過程中所衍生出來的協和樣本之取捨問題，基本上的想法是類似的，因此也提供了我們另外一種思考的方向。

三、計畫結果與討論

假設有兩筆資料如下：

A		B	
5	45	455	45
35	15	35	465

我們想要分別討論π_{+0}是否等於π_{0+}。

依據 McNemar 的檢定原則，由於檢定統計量 $X^2=\frac{(n_{01}-n_{10})^2}{n_{01}+n_{10}}$，因此就這兩組資料而言，儘管資料結構不盡相同(就 A 而言，$\hat{\pi}_{0+}=0.5$，$\hat{\pi}_{+0}=0.4$；就 B 而言，$\hat{\pi}_{0+}=0.5$，$\hat{\pi}_{+0}=0.49$)，然而 X^2 都等於 $\frac{(45-35)^2}{45+35}=1.25$，以致於不論 α 值如何選取，二者的結論完全相同。雖然由文獻中我們可以得知 McNemar 檢定統計量具有有效性(effecticiency)，然而在總樣本數增加的情形下，檢定力(power)卻維持不變。既然無法得到預期的好處，那我們又何必大費周章的增加樣本。這個事實對多數應用學者而言，常常會造成困惑和不解，而這也是這個研究計畫主要想要探討的焦點所在。

針對前述疑惑，我們可以說明如下：

(一)其實主對角線上的資料雖然表面上並未出現在檢定統計量之中，但它們對檢定統計量的影響卻是存在的。這是由於檢定統計量其實可以表示為：

$$X^2=\frac{(n_{01}-n_{10})^2}{n_{01}+n_{12}}=\frac{(n\hat{\pi}_{01}-n\hat{\pi}_{10})^2}{n\hat{\pi}_{01}+n\hat{\pi}_{10}}=n\frac{(\hat{\pi}_{01}-\hat{\pi}_{10})^2}{\hat{\pi}_{01}+\hat{\pi}_{10}}$$

其中

$$n=\sum_i\sum_j n_{ij}\ \text{而}\ \hat{\pi}_{ij}=\frac{n_{ij}}{n}。$$

因此這個統計量實際上與樣本總數 n 是有關的。藉此我們可以發現當主對角線上的資料增加時，除非$\pi_{01}-\pi_{10}=0$，否則隨著 n 變大，檢定統計量$n\frac{(\hat{\pi}_{01}-\hat{\pi}_{10})^2}{\hat{\pi}_{01}+\hat{\pi}_{10}}$的值也會放大。因此就 McNemar 檢定統計量而言，主對角線上的資料並非完完全全沒有提供任何訊息。

(二)令 $s=\frac{1}{\sqrt{2}}(\hat{\pi}_{01}-\hat{\pi}_{10})$ 和 $t=\frac{1}{\sqrt{2}}(\hat{\pi}_{01}+\hat{\pi}_{10})$ 為兩個新的座標軸，則

$$X^2=n\frac{(\hat{\pi}_{01}-\hat{\pi}_{10})^2}{\hat{\pi}_{01}+\hat{\pi}_{10}}=n\frac{\sqrt{2}s^2}{t}$$

棄卻域因而變成

$$X^2=n\frac{\sqrt{2}s^2}{t}>\chi^2_{1,\alpha}$$

亦即

$$s^2>\frac{1}{\sqrt{2}n}\chi^2_{1,\alpha}，\text{其中}\ \chi^2_{1,\alpha}\text{為一常數}$$

由此我們可以瞭解棄卻域是一個邊界為一個拋物線所組成的區域，該拋物線的正焦弦長為 $\left|\frac{\chi^2_{1,\alpha}}{\sqrt{2}n}\right|$，且拋物線和$\hat{\pi}_{01}$軸、$\hat{\pi}_{10}$軸的交點分別為$\left(\frac{\chi^2_{1,\alpha}}{n},0\right)$、$\left(0,\frac{\chi^2_{1,\alpha}}{n}\right)$。所以當 n 為有限數時，n 的大小會對拋物線的開口大小有所影響，進而影響棄卻域的範圍。當 n 大時開口就小，棄卻域較大；反之，當 n 小時開口就大，棄卻域也就較小。這一點與我們一般對樣本數大小與棄卻域大小之間關係的認知其實是一致的。

(三)問題的整個癥結其實在於前述 A 與 B 兩種資料結構所對應之虛無假設表面上看起來雖然是相同的，但實際上卻不盡如此。就資料 A 而言，其實

$$\hat{\pi}_{01}=0.45\quad\hat{\pi}_{10}=0.35$$

但是，就資料 B 而言

$$\hat{\pi}_{01}=0.045\quad\hat{\pi}_{10}=0.035$$

所以，儘管觀測值n_{01}都是 45，而n_{10}都是 35，但是對 A、B 而言這兩個數字所代表的意義並不相同。因此雖然針對 A 我們想檢定的虛無假設是：

$$H_0\colon\ \pi_{0+}=\pi_{+0}$$

而 B 所對應的虛無假設也是：

$$H_0: \pi_{0+} = \pi_{+0}$$

乍看之下二者完全相同，但是其實二者所要作的檢定並不相同。A 所對應的虛無假設其實是：

$$H_0: \pi_{0+A} = \pi_{00A} + \pi_{01A} = \pi_{00A} + \pi_{10A} = \pi_{+0A}$$

亦即

$$H_0: \pi_{01A} = \pi_{10A}$$

同理，B 所對應的虛無假設實際上是：

$$H_0: \pi_{0+B} = \pi_{00B} + \pi_{01B} = \pi_{00B} + \pi_{10B} = \pi_{+0B}$$

亦即

$$H_0: \pi_{01B} = \pi_{10B}$$

換句話說，A、B 兩組資料所得到的檢定統計量的值是用來檢定不同的虛無假設，只是兩個檢定問題所對應得到的檢定統計量的值恰巧相同罷了，因此儘管檢定結果相同亦無須驚訝。

(四)若要探討的是同一個虛無假設，則樣本數增加時所對應的資料結構的改變應該是對應格子的觀測值會呈現倍增的情況才是，亦即 $n_{ij}^B = kn_{ij}^A$，其中 k 是一個正整數。若資料等量放大(比方 $n_{ij}^B = kn_{ij}^A$)，由於檢定統計量變成

$$\frac{(n_{01}^B - n_{10}^B)^2}{n_{01}^B + n_{10}^B} = k\frac{(n_{01}^A - n_{10}^A)^2}{n_{01}^A + n_{10}^A},$$

因此較容易拒絕虛無假設。但是這種資料等量放大導致總樣本數增加的現象，與變動對角線數目所導致的總樣本數變動的情形，直觀上常被混為一談，以為其特性一致，事實上則相去甚遠。這一點應該清楚瞭解，才不致有所混淆。

此外我們也引進了一個新的統計量 $L = \frac{\log\frac{n_{01}}{n_{10}}}{\sqrt{\frac{1}{n_{01}} + \frac{1}{n_{10}}}}$。在虛無假設成立的前提下，這個統計量可以證明具有近似標準常態分配的一個分配。我們同時也藉由模擬實驗來比較這個統計量與 McNemar 檢定統計量的差異。兩個主要結論如下：

1.當格內的數值期望值很小時，無法計算出檢定統計量 L 的頻率很高。所以當資料格內的數值期望值很小時，檢定統計量 L 的檢定能力並不是很好。

2.當 McNemar 與 L 以其真實分配，(嚴格來說是藉由模擬實驗所得到的近似分配)來決定臨界值時，L 擁有較好的檢定力；但若以其極限分配(卡方分配或常態分配)來決定臨界值時，則是 McNemar 具有較好的檢定力。

四、計畫成果自評

藉由這個計畫的執行，我們得到的最主要結論可以歸納如下：

1.主對角線上所呈現的資訊並未被浪費掉，而是蘊含在檢定統計量中，並且會影響棄卻域的範圍。

2.在非主對角線上的數據為固定，只變動主對角線上的數據的情形之下，其實虛無假設是完全不同的。所以儘管檢定過程之中會有完全相同的檢定結果，但是在解釋上卻是截然不同。

誠如這個計畫的題目：「協和配對樣本在 McNemar 檢定中的角色扮演探討」，我想我們應該已經達到這個計畫最初設定的原始目標。至於新的統計量 $\frac{\log\frac{n_{01}}{n_{10}}}{\sqrt{\frac{1}{n_{01}} + \frac{1}{n_{10}}}}$，雖然在大樣本的情形下與 McNemar 檢定統計量具有相同的近似分配，藉由模擬實驗可以發現其表現雖無法超越 McNemar 的原始檢定統計量，但也在伯仲之間。不過藉由實驗的過程，我們能夠更進一步了解 McNemar 檢定的特性，也是收獲之一。

五、參考文獻

1.Bennett, B.M. and R.E. Underwood (1970). On McNemar Test for the 2×2 Table and Its Power Function. *Biometrics*, Vol.26, 339-343.

2.Coakley, C.W. and M.A. Heise (1996). Versions of the Sign Test in the Presence of Ties. *Biometrics*, Vol.52, 1242-1251.

3.Flerss, J.L. (1981). *Statistical Methods for Rates and Proportions*. New York: Wiley.

4.Liang, K-Y, and S.L. Zeger (1988). On the Use of Concordant Pairs in Method Case-Control Studies. *Biometrics*, Vol.44, 1145-1156.

5.McNemar, Q. (1947). Note on the Sampling Error of the Difference Between Correlated Proportions or Percentages. *Psychometrika*, Vol.12, 153-157.

6.Pratt, J.W. (1959). Remarks on Zeros and Ties in the Wilcoxon Signed Rank Procedures. *Journal of the American Statistical Association*, Vol.54, 655-667.

7.Randles, R.H. (2001). On Neutral Responses (Zeros) in the Sign Test and Ties in the Wilcoxon-Mean-Whitney Test. *American Statistician*, Vol.55, 96-101.