專題: Test for Non-normality:An Intensive Simulation Practice

目的:

統計學者提出的檢定統計式 (Test Statistic) 通常需要經過大量的模擬, 藉以驗證其分配的假設與 Type I Error 的維持, 並觀察檢定力的高低。本專題以常態檢定爲題, 作爲程式寫作的進階, 因此選擇幾個較複雜的檢定統計式, 配合 Monte Carlo Method, 練習程式寫作的技巧、細心與耐心。

為什麼:

統計學者寫作程式可比喻爲生化學者在實驗室操作器械與藥品。統計方法需經過驗證才能爲實務界所用, 而電腦模擬是一種有效的驗證的方式, 以電腦與軟體爲實驗設備, 程式寫作爲架構與程序, 藉由反覆執行來滿足統計上所根據的機率假設。學者必須在從學過程中練習實驗設備的駕馭與實驗程序的熟捻, 本專題提供執行這個過程的引導。

做什麼:

本專題選擇四種檢定統計式來測試其常態檢定能力, 利用 Monte Carlo Method 重複隨意的抽取資料, 觀察其「Type I Error 的維持」與「檢定力」。執行本單元的先決條件是必須具備基本的程式基礎, 才能應付本單元的幾個特點:

- 資料本身具備群組性 (Treatments), 也就是假設資料來自不同的群組, 各組平均數不同。
- Monte Carlo Method 重複的次數要夠多, 譬如 50000次。此時程式執行的效率非常重要, 更凸顯 MATLAB 以矩陣作爲運算基礎的概念。
- 觀察項目多, 程式結果的儲存與表達非常關鍵。我們希望觀察不同組數在不同樣本數下的表現。在檢定力的測試上, 需要假設不同的 H_A 來源, 譬如 F 分配,T 分配,...。

背景: 四個常態檢定統計式

以下四個檢定式基本上分爲兩種形式, 其一爲統計式的分配確定, 其二爲統計式的分配未知。未知分配的統計式以「實驗式分配 (Emperical Distribution)」的方式計算相關的 p 值, 或做爲拒絕假設與否的依據。這是個無母數統計的手段, 非常實用有趣。

1 W Test

Suppose $(x_1x_2, \cdots, x_n)$ be the sample of size n to be tested for non-normality and $y_1 < y_2 < \cdots < y_n$ is its ordered counterpart.The W statistic is defined by

$$W = \frac{\left(\sum_{i=1}^{n} a_i y_i\right)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{1}$$

where $\mathbf{a} = (a_1, a_2, \cdots, a_n)^T$ is

$$\mathbf{a} = (\mathbf{m}^T V^{-1} V^{-1} \mathbf{m})^{-1/2} \mathbf{m}^T V^{-1}$$

where V is the covariance matrix of the order statistics of a sample of n standard normal random variables with expectation vector $\mathbf{m}$. The vector $\mathbf{a}$ is antisymmetric, that is, $a_n = -a_1$ and for odd n, $a_{(n/2)+1} = 0$. Also $\mathbf{a}^T\mathbf{a} = 1$. The approxiamtion of the coefficients $(a_1, a_2, \cdots, a_n)$ according to the sample size n is given by

for $n = 3$

$$(\tilde{a}_1, \tilde{a}_2, \tilde{a}_3) = (-0.7071 \;\; 0 \;\; 0.7071)$$

for $4 \leq n \leq 1000$

$$\begin{aligned} \tilde{a}_n &= c_n + 0.221157x - 0.147981x^2 - 2.071190x^3 + 4.434685x^4 - 2.706056x^5 \\ \tilde{a}_{n-1} &= c_{n-1} + 0.042981x - 0.293762x^2 - 1.752461x^3 \\ &\quad +5.682633x^4 - 3.582663x^5 \end{aligned}$$

$$\tilde{a}_i = \frac{\tilde{m}_i}{\sqrt{\phi}}$$

for $i = 2, \cdots, n-1 (n \leq 5)$ or $i = 3, \cdots, n-2 (n > 5)$

where

$$\begin{aligned} x &= \frac{1}{\sqrt{n}} \\ \tilde{m}_i &= \Phi^{-1}\left((i - \frac{3}{8})/(n + \frac{1}{4})\right), \quad \Phi \text{ is the normal cdf} \\ c_i &= \frac{\tilde{m}_i}{\sqrt{\tilde{\mathbf{m}}^T \tilde{\mathbf{m}}}} \end{aligned}$$

$$\begin{aligned} \phi &= (\tilde{\mathbf{m}}^T \tilde{\mathbf{m}} - 2\tilde{m}_n^2)/(1 - 2\tilde{a}_n^2) \quad if \;\; n \leq 5 \\ &= (\tilde{\mathbf{m}}^T \tilde{\mathbf{m}} - 2\tilde{m}_n^2 - 2\tilde{m}_{n-1}^2)/(1 - 2\tilde{a}_n^2 - 2\tilde{a}_{n-1}^2) \quad if \;\; n > 5 \end{aligned}$$

To find the p-value for W, $Z = (w - \mu)/\sigma$ is referred to the upper tail of $N(0, 1)$, where

for $4 \leq n \leq 11$

$$\begin{aligned} w &= -\ln(\gamma - \ln(1 - W)) \\ \gamma &= -2.273 + 0.459n \\ \mu &= 0.544 - 0.39978n + 0.025054n^2 - 0.0006714n^3 \\ \sigma &= exp(1.3822 - 0.77857n + 0.062767n^2 - 0.0020322n^3) \end{aligned}$$

for $12 \leq n \leq 2000$

$$\begin{aligned} x &= \ln n \\ w &= \ln(1 - W) \\ \mu &= -1.5861 - 0.31082x - 0.083751x^2 + 0.0038915x^3 \\ \sigma &= exp(-0.4803 - 0.082676x + 0.0030302x^2) \end{aligned}$$

For $n = 3$, the p-value is directly given by

$$p = \frac{6}{\pi}\left(\sin^{-1}\sqrt{W} - \sin^{-1}\sqrt{0.75}\right)$$

2 Global Test:GW and GC

GC is a global test for the normality by combining the test for each treatment and is written by

$$GC = -2\sum_{i=1}^{K} \ln p_i \tag{2}$$

where p_i is the p-value for the i-th treatment by the W test and K represents the number of treatments. The GC statistic follows an asymptotical chi-square distribution with $2K$ degrees of freedom.

The other global test that uses the transformed Z instead of p-value is called GW statistic which is defined by

$$GW = \frac{\sum_{i=1}^{K} Z_i}{\sqrt{K}} \tag{3}$$

The upper tail Z test then follows.

3 MQ test: The Test with Emperical Distribution

The Q test proposed by Zhang(1999) is defined as the log-ratio of two unbiased estimator of the population variance. The two unbiased estimators are calcualted by $q_1 = \sum_{i=2}^{n} g_i/(n-1)$ and $q_2 = \sum_{i=1}^{n-4} h_i/(n-4)$ where

$$\begin{aligned} g_i &= \frac{y_i - y_1}{m_i - m_1}, \qquad i = 2, 3, \cdots, n \\ h_i &= \frac{y_i - y_{i+4}}{m_i - m_{i+4}}, \qquad i = 1, 2, \cdots, n-4 \end{aligned}$$

where y_i and m_i represents the order statistics and their expected values under normality,i.e. $m_i = \Phi^{-1}\left((i - \frac{3}{8})/(n + \frac{1}{4})\right)$. The test is then defined as

$$Q_1 = \log \frac{q_1}{q_2}$$

The second statistic is constructed by reversing and rearranging y_i, i.e. the new sequence of sample is $(y_1^*, y_2^*, \cdots, y_n^*) = (-y_n, -y_{n-1}, \cdots, -y_1)$. The second statistic Q_2 is then defined in the same way as Q_1 by employing y_i^*. It is noted that both statistics(Q_1, Q_2) do not follow any standard distribution. The emperical distributions for both statistics can be generated based upon Monte Carlo samples,[1] i.e. the discrete versions of CDF of Q_1 and Q_2 statistics. Figure 1demonstrates an example of using 50000 Monte Carlo smaples. Based on the emperical CDF, the null hypothesis is accepted only when both Q_1 and Q_2 are non-significant at the level $\alpha/2$. Ths test procedure is denoted as the MQ test.

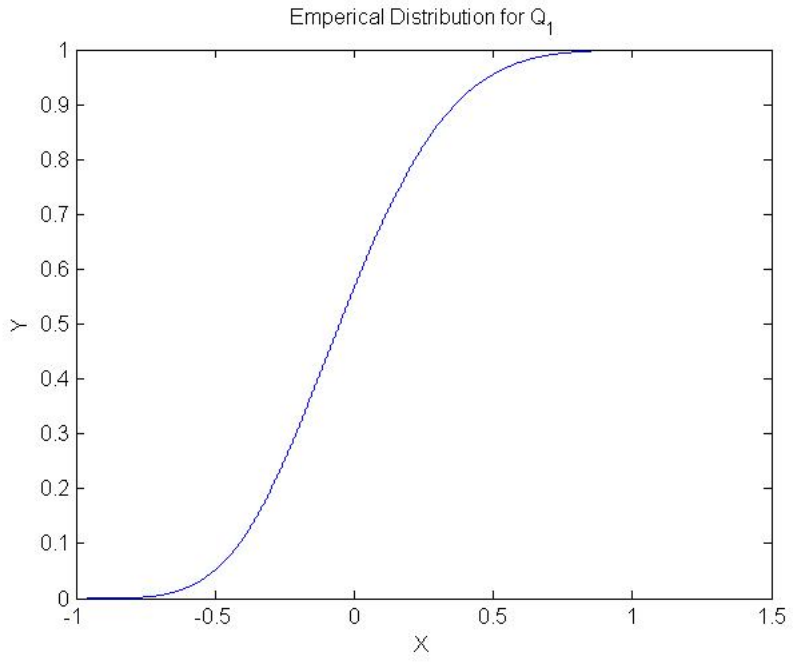

圖 1: 50000 Monte Carlo samples for the emperical CDF of Q_1

以上統計式的表達以英文爲主, 希望習者能漸適應期刊文章的方式。

[1]A MATLAB instruction "cdfcalc" can be used to calculate CDF from the Monte Carlo samples of the statistic.

怎麼做

型一誤的維持

在虛無假設下, 樣本來自常態, 本專題想瞭解不同的檢定統計式, 在不同的樣本數及群組 (treatments) 組合下, 對於型一誤的維持。

1. 先假設資料來自同一個常態群組, 平均數 $\mu = 1$, 變異數 $\sigma^2 = 1$。針對不同的樣本數 $n = 3, 4, \cdots, 60$, 各產生 50000 組樣本, 對每組樣本進行 W-test 與 MQ-test , 假設型一誤 $\alpha = 0.05$, 紀錄拒絕虛無假設的比例, 畫一張圖展示這兩種方法的拒絕比例與樣本數的關係。

2. 假設資料來自 K 個常態群組, 其平均數分別爲 $\mu = 2K - 1, K = 2, 3, \cdots, 6$, 變異數則同爲 1。每個群組的樣本數設定爲 $n = 3, 4, \cdots, 60/K$, 也就是總樣本數以60爲限。執行上述的四種檢定, 針對不同組數 K, 各畫一張圖展示拒絕比例與總樣本數 Kn 的關係。請注意,W-test 與 MQ-test 處理多組數的情況, 需先將各組資料減去該組的樣本平均數, 再將各組資料合在一起處理。

檢定力

當樣本來自對立假設, 則拒絕虛無假設的比例稱爲檢定力 (Power), 因對立假設的範圍太廣泛 (所有不是常態分配的資料), 只能限制對立假設來自非常態的母體, 譬如以下分配:

1. Lognormal distribution $\alpha = 3, \beta = 1$
2. Chi-square distribution with degree of freedom N=1
3. Chi-square distribution with degree of freedom N=4
4. T distribution with degree of freedom N=1
5. T distribution with degree of freedom N=4
6. Uniform distribution on (0,1)

7. Weibull distribution $alpha = 2, \beta = 1$

8. Weibull distribution $alpha = 0.5, \beta = 1$

9. Beta distribution $alpha = 0.5, \beta = 0.5$

10. Beta distribution $alpha = 5.5, \beta = 2.5$

11. F distribution $alpha = 50, \beta = 3$

同樣的, 依上述作法, 針對不同組數 K, 各畫一張圖展示拒絕比例與總樣本數 Kn 的關係。

參考文獻

[1] Royston, J.P.(1992), "Approximating the Shapiro-Wilk W-test for non-normality," Statistics and Computing, 2,117-119.