嘉南藥理科技大學九十二年度教師專題研究計畫成果報告

The application of the correlation coefficient of the normal probability plot for the normality test

計畫編號：CNHA92-07
執行期限：92 年 1 月 1 日至 92 年 12 月 31 日
主持人：陳青浩　嘉南藥理科技大學 醫務管理系

關鍵詞：常態機率圖，樣本次序統計量，常態分值，相關係數。

(一)摘要

在線性迴歸的應用上，誤差項的常態性檢定方法中，常態機率圖由於製作簡單且容易解釋已廣泛被採納及使用。但圖形是否近於線性，在主觀視覺上做判定不夠客觀，容易引起爭議。Looney 和 Gulledge (1985)藉由電腦模擬計算樣本次序統計量與對應常態分值的相關係數，並以經驗百分點(empirical percentage point)做為常態性假設的檢定統計量而編製了一個臨界值表藉以作為檢定樣本是否服從常態性假設分配時的參考。本研究中利用 Beta 分配參數估計的方式配合檢定的觀念建立一種不同於 Looney 和 Gulledge 方法的臨界值表，並比較這兩者之間在檢定常態性時之差異。

(二)緣由及目的

在應用統計的範疇中，模式的建立是很重要的一個步驟，而模式的適當性則需要更進一步的檢驗，亦即某些模式必須的假設條件需要利用檢定的程序加以確認。例如在線性迴歸的應用上，迴歸函數的線性或誤差項的常態性如果無法被確定，則依據此模式所做的統計推論是有瑕疵的。其中常態性的檢定歷經數十年的研究，已經發展出許多檢定的方法，在這些檢定的方法中，由於常態機率圖(normal probability plot)的製作簡單與容易解釋，所以被廣泛的接受並被建議為檢定常態性時的工具。

對於一個需要統計分析的樣本，如果假設分配是必要的，則必須經過嚴謹的步驟加以檢定，才能有正確的分析成果。機率圖常被用來評估假設分配的適合性。若假設分配的分配函數是 $F\left[\frac{y-\mu}{\sigma}\right]$，其中 μ 和 σ 分別為位置和尺度參數，而 $\{y_{(1)},...,y_{(n)}\}$ 為來自此一分配的樣本次序統計量，機率圖的繪製步驟是以樣本次序統計量 $y_{(i)}$ 為 y 軸，相對應的 $x_i = F^{-1}(p_i)$ 為 x 軸，所繪製的散佈圖即為機率圖。此

處 p_i 為 $F\left[\frac{y_{(i)}-\mu}{\sigma}\right]$ 的一個估計值，並稱此估計值為圖點位置(plotting position)。

Hazen(1914)設 $p_i=\frac{i-0.5}{n}$，Weibull(1939) 設 $p_i=\frac{i}{n+1}$。若我們考慮假設分配為常態分配，則 Blom(1958) 定義常態機率圖的圖點位置為 $p_i=\frac{i-0.375}{n+0.25}$。SAS 統計軟體中 UNIVARIATE 與程序 GLM 程序皆提供相關選擇指令供使用者繪製常態機率圖。

本研究之主要目的在於探討常態性假設的檢定程序，以樣本常態機率圖的線性與否判斷樣本是否符合常態性的依據，並建立一個臨界值表以提供檢定常態性假設時查表使用。本研究的研究方法以亂數模擬產生樣本，並以 Beta 分配做為樣本次序統計量 $y_{(i)}$ 與常態分值 x_i 的相關係數的模式，最後再以 Beta 的分配函數計算百分位，將此成果與 Looney 和 Gulledge (1985)的結果做比較分析。

(三)研究結果

常態機率圖的理論背景簡述如下：令 $\{y_{(1)},\cdots,y_{(n)}\}$ 表示樣本數為 n 的樣本次序統計量(ordered statistics)，對第 i 個樣本次序統計量 $y_{(i)}$，定義其相對應的常態分值(normal scores) $x_i=Z\left(\frac{i-0.375}{n+0.25}\right)$，其中 $Z(A)$ 為標準常態分配的第 $(100A)$ 百分位(percentile)，再以樣本次序統計量 $y_{(i)}$ 為縱軸，常態分值 x_i 為橫軸所繪製的散佈圖即為常態機率圖。如果此圖近於線性，則樣本的常態性假設可被接受，也就是說樣本是來自於一個常態母體；相反的，若此圖明顯的偏離線性，則樣本的常態性假設就無法成立，樣本則不是產生自於一個常態母體。將常態機率圖應用於常態性的檢定，其優點是步驟簡單，其缺點是在判定上不夠客觀，畢竟圖形是否近於線性，容易引起爭議。所以除了視覺上做主觀的線性的評估外，我們更需要客觀的輔助工具來做為決策依據。

Looney 和 Gulledge (1985)利用常態亂數值模擬多組樣本，計算樣本次序統計量 $y_{(i)}$ 與常態分值 x_i 的相關係數 r，並以經驗百分點(empirical percentage point)做為常態性假設的檢定統計量。由於相關係數是評估線性相關程度的數據，因此 Looney 和 Gulledge (1985)建議以常態機率圖中樣本次序統計量 $y_{(i)}$ 與常態分值 x_i 的相關係數做為檢定統計量，如此數值上的判定就較為客觀，他們是以不同的樣

本數、個別產生 10000 個常態亂數值、並重複進行 12 次，最後計算出不同樣本數的經驗百分點(empirical percentage point)。依此重複的步驟，他們編製了一個臨界值表藉以檢定樣本是否服從常態性假設分配，如今這個臨界值表已被廣泛運用，不少書籍都列為模式診斷章節中的一個重要圖表。樣本數為 n 的樣本次序統計量 $y_{(i)}$ 與常態分值 x_i 的相關係數若大於表列的臨界值，則不拒絕常態性假設，也就是樣本來自於一個常態母體。

表一、排序後殘差與其對應之常態性假設下期望值之相關係數在誤差項成常態分佈時之臨界值

n	Level of significance				
	0.10	0.05	0.025	0.01	0.005
5	0.903	0.880	0.865	0.826	0.807
10	0.934	0.918	0.901	0.879	0.862
16	0.953	0.941	0.929	0.913	0.899
20	0.960	0.951	0.940	0.926	0.916
30	0.971	0.964	0.957	0.947	0.939
40	0.977	0.972	0.966	0.959	0.953
50	0.981	0.977	0.972	0.966	0.961
60	0.984	0.980	0.976	0.971	0.967
70	0.986	0.983	0.979	0.975	0.971
80	0.987	0.985	0.982	0.978	0.975
90	0.988	0.986	0.984	0.980	0.977
100	0.989	0.987	0.985	0.982	0.979

資料來源: Looney 和 Gulledge (1985)

Looney 和 Gulledge 利用經驗百分點法則決定臨界值，在本研究中採用不同於 Looney 和 Gulledge 的方法而是利用分配理論配合大量的電腦模擬次數來決定臨界值。由常態分值 $x_i = Z\left(\frac{i-0.375}{n+0.25}\right)$ 的定義可證得 $y_{(i)}$ 與 x_i 為正相關，配合柯西-舒瓦茲不等式(Cauchy-Schwarz inequality)得到 $y_{(i)}$ 與 x_i 的相關係數介於 0 與 1 之間，若將此相關係數視為一隨機變數，則 Beta 分配

$$f(x|a,b)=\begin{cases}\dfrac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}x^{a-1}(1-x)^{b-1}, & 0<x<1, a>0, b>0\\ 0, & \text{其他}\end{cases},$$

可做為相關係數合理的模式。既然 $y_{(i)}$ 與 x_i 的相關係數愈大，常態機率圖愈趨於線性，因此檢定為單尾檢定，在設定顯著水準為 α 情況下，以相關係數為檢定統計量，臨界值為 Beta 分配的第 $100(1-\alpha)$ 百分點，亦即若相關係數 r 大於 $\xi_{(\alpha)}$，則

在顯著水準為α情況下樣本來自於一常態母體，其中$\xi_{(\alpha)}$定義如下：

$$\int_0^{\xi_{(\alpha)}} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1} dx = \alpha \text{ 。}$$

至於 Beta 分配中之a，b兩個參數的估計值可藉由亂數的產生取得模擬樣本求得動差法(method of moments)估計值。

$$\hat{a} = \frac{m^2 - mt}{t - m^2},$$

$$\hat{b} = \frac{(m-t)(1-m)}{t - m^2}$$

where $\quad m = \frac{1}{n}\sum_{i=1}^{n} x_i$

$$t = \frac{1}{n}\sum_{i=1}^{n} x_i^2$$

下面所有計算方法皆採用 SAS 統計軟體既有程序與程式架構，所進行之步驟分述如下：

(1) 樣本數n分別設定為 5,10,15,20,30,40,50,60,70,80,90,100。
(2) 產生樣本大小為n的標準常態亂數值。
(3) 計算亂數值排序後相對應的常態分值。
(4) 計算排序後亂數值與相對應的常態分值兩者間之相關係數。
(5) 重複(1)~(4)步驟$N = 10000$次，取得 10000 個相關係數值$R_i, i = 1, \cdots, 10000$。實際上，上述獲得的相關係數都介於(0,1)之間，因此R_i服從$Beta(a,b)$分配；利用動差法求a與b之估計量，並以上述樣本(相關係數)求得a與b之估計值。
(6) 設定不同的顯著水準α，求$Beta(a,b)$分配的$100(1-\alpha)$百分位，定義為臨界值。
(7) 視臨界值的穩定性與精確度適當調整n和N的值。
(8) 完成製表。

表二、$Beta(a,b)$分配參數值模擬結果

N	m	T	$\hat{a}$	$\hat{b}$
5	0.95090	0.90555	32.1955	1.66239
10	0.96381	0.92945	65.0929	2.44389
15	0.97161	0.94433	84.2895	2.46326
20	0.97675	0.95427	98.5145	2.34478
30	0.98332	0.96703	139.754	2.37093
40	0.98671	0.97364	248.285	3.34529
50	0.98885	0.97787	232.824	2.62478
60	0.99039	0.98090	375.663	3.64554

70	0.99155	0.98320	372.620	3.17471
80	0.99219	0.98446	456.549	3.59372
90	0.99318	0.98643	445.978	3.06065
100	0.99391	0.98786	485.374	2.97552

資料來源:本研究

表三、排序後殘差與其對應之常態性假設下期望值之相關係數在誤差項成常態分佈時之臨界值

N	Level of significance				
	0.10	0.05	0.025	0.01	0.005
5	0.90132	0.87924	0.85820	0.83165	0.81239
10	0.93337	0.92054	0.90838	0.89304	0.88188
15	0.94771	0.93759	0.92797	0.91579	0.90691
20	0.95665	0.94803	0.93980	0.92935	0.92170
30	0.96893	0.96273	0.95680	0.94925	0.94371
40	0.97703	0.97308	0.96934	0.96463	0.96119
50	0.97968	0.97578	0.97206	0.96732	0.96385
60	0.98367	0.98097	0.97841	0.97518	0.97283
70	0.98522	0.98261	0.98013	0.97699	0.97469
80	0.98669	0.98446	0.98236	0.97970	0.97776
90	0.98798	0.98581	0.98376	0.98115	0.97923
100	0.98919	0.98721	0.98533	0.98295	0.98120

資料來源:本研究

常態分配的檢定在應用統計分析上非常重要,因為許多的假設條件都與常態性有關,尤其迴歸分析與變異數分析中,誤差項如果無法滿足常態分配的假設條件,則一些後續的統計推論都不會成立。

比較本研究和 Looney and Gulledge (1985)的結果,發現臨界值非常近似,雖然兩者的理論基礎是不同的。但本研究運用適當的分配理論加上大量的電腦模擬,所製作的臨界值表應更有效精細而且具有說服力。本研究結果提供給應用統計上進行常態性假設檢定時的另一種依據,除此之外若以此計畫為基礎,後續可進行其他分配的假設檢定,例如指數分配也可以類似的步驟進行檢定,因為指數分配在保險學及風險管理上被廣泛的應用且有其一定的重要性。

參考文獻：

Blom, G. (1958), *Statistical Estimates and Transformed Beta Variables*, New York: John Wiley.

Filliben, J. J. (1975), The probability Plot Correlation Coefficient Test for Nornality, *Technometrics*, **17**, 111-117.

Hazen, A. (1914), Storage to Be Provided in the Impounding Reservoirs for Municipal Water Supply, *Transactions of the American Society of Civil Engineers*, **77**,

1547-1550.
Klugman, A. S.; H. H. Panjer; and G. E. Willmot (1998), *Loss Models From Data to Decisions* , New York: John Wiley.
LaBrecque, J. (1977), Goodness-of-Fit Tests Based on Nonlinearity in Probability Plots, *Technometrics*, **19**, 293-306.
Looney, S. W. and Gulledge, T. R. (1985), Use of the Correlation Coefficient With Normal Probability Plots, *The American Statistician*, **39**, 75-79.
Mage, D. T. (1982), An Objective Graphical Method for Testing Normal Distributional Assumptions Using Probability Plots. *The American Statistician*, **36**, 116-120.
Michael, J. R. (1983), The Stabilized Probability Plot, *Biometrika*, **70**, 11-17.
Mood, M. A.; F. A. Graybill; and D. C. Boes, (1974), *Introduction To The Theory of Statistics*, McGraw-Hill, Inc.
Neter, J.; M. H. Kutner; C. J. Nachtsheim; and W. Wasserman (1996), *Applied Linear Statistical Model* , Richard D. Irwin, Inc.
SAS Institute Inc. (1990), *SAS/STAT User's Guide, 1990 Edition*, Cary, NC: Author.
Stirling, W. D. (1982), Enhancements to Aid Interpretation of Probability Plots, *The Statistician*, **33**, 211-220.