統計在紅樓夢的應用(註)

余 清 祥*

摘 要

「紅樓夢」爲近代文學的一大鉅著，堪稱古典小說的代表作品，然而作者是誰，始終是未解的謎。在一般的認知中，紅樓夢是曹雪芹所作及高鶚編纂，但專家對於本書之前八十回與後四十回是否爲同一位作者仍無定論。本研究藉由統計方法（包括變動點分析），將「紅樓夢」中的文字敘述數量化，用以分析作者的寫作風格，進而尋求解答「紅樓夢」作者的可能性。本文所採用之紅樓夢版本，以「庚辰本」及「程甲本」爲主要研究書目；而以「紅樓夢校註」爲輔(註一)。統計分析則以 Minitab 、 SPlus 等軟體爲輔助工具。

關鍵詞：變動點分析、紅樓夢、多項分佈、品種的個數。

一、緒 論

1.1 前言

中國章回小說創始於明朝，至清朝發揚光大，為明清兩朝的文學代表作品。與傳統的詩詞曲賦不同，小說並非由單一文體所構成，它結合了詩詞、文言文的敘述，以及白話的對話。

(註) 本研究受行政院國家科學委員會補助 (編號 NSC 84-2121-M004-009)，特此致謝。

* 作者爲本校統計系副教授

註 一：里仁出版社。

因此，它的讀者群不再只是自命風流的文人墨客，一般平民百姓，甚至目不識丁的販夫走卒，亦可藉由說書人的生動描述，捕捉書中人物的面貌，進而咀嚼玩味小說的精髓。小說普遍化之影響程度實非以往其他文學作品可相抗衡，而其影響力之大，也可由民間俗諺略知一二，即所謂「少不看水滸，老不看三國」。雖然有人批評小說多半具有濃厚的政治化傾向，然不容諱言，小說題材之包羅萬象，以及內容之平民化、生動化，無疑地為中國文學打了一劑強心針。

但由於印刷的昂貴與古代中國人缺乏著作權觀念的影響下，一部小說或因謄抄的錯誤；或因原著的散佚；或因後人的篡改增刪，可能產生與原著出入甚多的各種版本，令後來的讀者無所適從。以本文研究的『紅樓夢』為例，坊間所知的版本計有「甲戌本」、「己卯本」、「庚辰本」、「甲辰本」、「戚本」（以上為「脂本」，也就是未經高鶚輯補過之版本）、「程甲本」、「程乙本」等。其中五個脂本，全都是經由過錄而得的手抄本，但較差的抄手因抄謄的錯誤，對抄本的品質有負面的影響。另一方面，由於高鶚編輯的版本搜集民間各版本，將原先不足或散佚的部份補足，而有一百二十回的完整小說，高鶚整理之功雖不可沒，但原著書者的創作精神也受到相當程度的扭曲，無法以原始面目見於世人。

1.2 研究目的及方法

一般相信曹雪芹為前八十回的作者，而高鶚則續作了後四十回，使『紅樓夢』得以一百二十回的型態問世。但此一說法也引起相當大的疑問。例如第八十回至八十一回的轉接非常平暢，毫無切割的痕跡；而且後四十回的舖陳也與前八十回相若。因為續書通常比著書更難，除非高鶚的文學素養高出曹雪芹，否則難以在短時間內，將曹雪芹十年心血的創作延續下來。另外，趙岡與陳鍾毅也指出（註二），『紅樓夢』一書是「寫實往事」，將曹家幾代為江寧織造的往事融入小說情節，但曹雪芹並未親身體驗曹家的全盛期，許多事蹟的細節必須仰賴族中長輩的幫助，如曹天佑及曹棠村等人，而『紅樓夢』的增刪修改與素材靈感，也藉由這些人的批閱獲致。因此，此書可說是曹家繁華舊夢的回憶錄，並為一集體創作；但並無資料顯示高鶚與曹家故人有密切來往，則高鶚如何能將前八十回的敘述與情節毫無破綻地延續下

註　二：紅樓夢研究新編，P187。

來？

另一方面，前八十回及後四十回在用字遣詞上的差異，以及情節鋪陳上明顯的不同，不難在讀者細心比較下發現蛛絲馬跡。但過去對前八十回與後四十回差異的研究，除了趙岡與陳鍾毅(1980)以統計方法來分析，大多以文學欣賞的角度加以評論，鮮少採用數量化的方法。(但不當的使用統計方法，則可能產錯誤的結論，如高本漢於 1952 年的研究。)本研究著重於統計方法的應用，將各回的用字結構轉變成數字，作為分析前後各回是否有顯著不同(Significant difference）的評論依據；另由於『紅樓夢』一書創作歷經五次增刪及十年光陰，作者的寫作及用字風格可能因時空的變遷與經驗的累積，產生質與量的改變，因此除了一般的統計方法將前八十回及後四十回當作兩個樣本(Two sample)來分析外，本文也採用變動點問題 (Change-point problem）的方法，判斷一百二十回的小說中是否有前後不一的現象；如果前後不一，轉折點是否出現在第八十回附近。

本文編排如下：第二節為文獻探討，包括趙岡與陳鍾毅(1980)的用字分析，以及本文所使用的統計方法，如 t 檢定(t-test)、卡方檢定(Chi-square test)，加上與變動點問題相關的文獻。第三節為實證分析，介紹如何選取字詞來比較，並討論分析所得的結果與意義。第四節為結論與建議，除了總結第三節的分析結果，並以統計的觀點嘗試解決『紅樓夢』的作者問題，同時也討論本篇研究的限制；於本節末也提出未來可能的研究方向，以期達到更精確的結果。

二、文獻探討

本節分為兩個部份：一為介紹與關於作者寫作風格研究的統計報告，一為與本研究相關的統計方法之探討。其中作者風格的報告又可分為國內外兩類，而統計方法也可分為一般檢定與變動點分析兩種。以下先就作者風格的部份逐一介紹。

2.1 研究作者風格的文獻

2.1.1 國內部份：（趙岡與陳鍾毅，1980）

趙岡與陳鍾毅使用「兒」、「在」、「了」、「的」、「著」五個虛字作為比較，前八

十回採用俞平伯的「紅樓夢八十回校本」，後四十回則採「程甲本」。在前後半部中各挑出(抽樣)一百頁（每頁七百二十個字），各頁中沒有回目、詩詞或分段，計算每頁出現這五個虛字的頻率，得到以下前八十回及後四十回的 t 檢定值：

兒	3.677
在	3.392
了	0.116
的	3.391
著	3.910

其中除了「了」字的平均頻率差異不顯著外，其他四個虛字的 P 值（P-value），也就是由或然率而造成的結果，其值都在百分之一以下，如果把這四個字的效果結合在一起，因為或然率而造成巧合的可能性幾乎為零。這四個有顯著差異的字，在後四十回的出現次數較前八十回有過之而無不及。趙陳兩人對這種現象的解釋是可能由於曹雪芹為南方人，用北京話寫小說係屬創舉，因此對北京話的使用並非十分純熟，然而高鶚是道地的老北京；後四十回極有可能在高鶚的修輯中，改成道地的北京話，使得「兒」、「在」、「了」、「的」、「著」出現頻率有增高的傾向。

其他趙陳兩人發現的歧異計有如：問句結尾的「嗎」或「麼」，「我們」與「咱們」，「給」及「與」，「都」及「多」，動詞疊用等，趙陳兩人也計算出這些字詞在前後半部各一百頁抽樣中的出現次數，作為支持前後半部用詞不同的另一證據。

2.1.2 國外部份

以統計方法分析寫作風格（Literary style），稱為 Stylometry，可追溯自英國邏輯學家 Augustus de Morgan 於西元 1851 年所作的研究，他建議以英文單字的長短作為區分作者風格的判斷依據。在 1930 年代，G. Udny Yule 和 George Zipf 發現文章的字彙使用頻率有一定的模式(Pattern)，可作為分辨不同作者的標準。而最著名的寫作風格比較首推 Frederick Mosteller 及 David Wallace (1984) 的研究，他們主要運用貝氏分析的技巧(Bayesian techniques)，探討美國自開國以來即存

在的作者認定問題——"*誰寫了擁護聯邦主義的論文*"(The Federalist Papers)。該系列論文由 Alexander Hamilton 、 James Madison 及 John Jay 撰寫，在 77 篇文章裏有 65 篇已大致可確定作者，但歷史學家對剩下的 12 篇文章卻遲遲未能結論，推測可能是 Hamilton 或 Madison 所寫。 Masteller 及 Wallace 的分析歸結出除了第 55 篇極有可能是 Hamilton 所作外（可能性為 90：1），其他應該為 Madison 的作品。

另一較知名的是研究對莎士比亞（Shakespeare）使用字彙總數，由 Bradley Efron 及 Ronald Thisted 在 1976 年所提出。他們使用出現字彙的總數及其頻率，估計莎士比亞知道但未使用的字彙，進而估計出莎士比亞知道的字彙總數，作為判斷一篇作品是否出自莎翁的憑據。之後，他們更由此推測 1985 年新發現的一首詩，是由莎士比亞所作。

由於電腦科技的發達以及資料庫的進展，幾乎可化為數量化的特性都被拿來比較，不再侷限於字彙的出現總數與其頻率，其他數值如單字的長短、字句的長短，以至於名詞、代名詞的出現次數，都曾被選作比較的標準。 C.B.Williams(1975)研究莎士比亞與培根(Bacon)兩人於散文(Prose)及詩詞(Verse)上用字長短的差別，並發現無論莎士比亞或培根，對散文及詩詞都有不同的用字習慣，因此對不同作者作比較時，應選用同一文體為原則；他並用此結果駁斥 Mendenhall 先前研究的誤謬。 B.J.R.Bailey(1990)則選用連接詞(Function words or contextfree words)，如代名詞作為比較的依據，在二項分佈(Binomial distribution)的假設下，計算貫詞(Articles)的出現頻率，並使用卡方檢定(Chi-square test)來檢測。 A.F.Bissell(1995)則考慮 Weighted Cumulative Sums 與 Weighted Variance Estimation ，測量作者是否有習慣用字，並討論用兩個字母及三個字母構成的單字在文章中的出現頻率，使用統計圖形作為輔助判斷的工具。

但值得注意的是中文與英文在用詞遣字及文體結構上的差別。英文是由 26 個字母所構成，較易轉化成數字來比較，如計算每個單字的字母總數。但反觀中文，每一個字都是一個方塊或一個單元，很難直接數量化，如英文單字的長短在中文裏毫無意義。另外，英文裏的貫詞，如 a 、 the 等字的用法，在中文裏並無對等的字，因此如何選取適當字詞作為比較的標準，進而判斷中文寫作的風格，與英文將是大相逕庭，各異其趣。在第三節我們將詳述如何選取字詞，並解釋原由。

2.2 本研究使用的統計方法

本研究的統計檢定假設可分為兩類：一是假設前八十回及後四十回來自兩個不同的主體(Populations)，並採用兩個樣本(Two sample)的方法檢定;一是假設全書一百二十回來自同一主體，使用變動點問題（Change-point problem）的方法檢定。

2.2.1 兩個樣本

在前八十回及後四十回為兩個樣本的分類下，一般的檢定方法皆可視情況需要而使用。例如 t 檢定可用來比較每回中使用的總字數，以及「兒」字在前後各部的出現頻率，但因為前八十回有五十萬餘字，後四十回也有二十萬字以上，在大樣本的條件下，t 檢定也可用 z 檢定來代替。當資料可分成數類時，卡方檢定也適用。除此之外，時間序列分析(Time Series Analysis)及辨別分析(Discriminant Annlysis)也適用於兩個樣本的比較，但僅作為輔證。其中時間序列可用在分析前八十回的某些字詞，他們是否遵循某一特定模式出現；而此一模式或類似模式是否也在後四十回中出現。辨別分析是用在選取數個不同的特質，如將「兒」、「在」、「了」、「的」、「著」五個字同時考慮，藉由線性辨別函數(Linear Discriminant function)，判定前八十回及後四十回是否可歸類成不同的兩組資料；如果分成兩組資料，又有多少回的資料被判定成另一組的資料，其誤判率為何。

2.2.2 變動點問題

變動點問題的研究始於 1930 年代，最初應用於工業上的品質管制(Quality Control)，近三十年來的應用也不再侷限於工業方面。變動點問題在於研究一系列同性質的事或物，(或以統計的術語來看，則是研究一系列獨立且有相同分配的隨機變數)，並決定這一系列事物的特性是否在某一時間產生變化。舉例而言，某工廠生產日光燈管，要求成品至少有五千小時的壽命，但生產的機器隨運轉時間的增長，可能會生產不合規格的燈管，如能愈早查出機器故障的時間，及時調整，可為工廠省下生產成本。

本研究採用變動點方法的原因，在於『紅樓夢』一書著作前後歷經十年，作者的風格可能在這段時間有明顯的不同，若單純以前八十回與後四十回分段比較，極有可能造成讀者的

錯覺，認為前後各回屬於不同的作者。如能將全書一百二十回視為同一系列的產品，而得到確實有變動點，且其位於八十回前後，當可作為前八十回及後四十回風格不同的佐證，支持不同作者的假設。反之，若沒有變動點的產生，也就是前後各部文風相近，或是有兩個或兩個以上的變動點，代表作者風格有變化或有兩個以上的作者，則前後各回分屬不同作者的假設極有可能不成立。

變動點問題若在固定樣本(Fixed sample size)的假設下，可分為二項分佈及常態分佈兩類資料。在二項分佈的情形， David V. Hinkley 和 Elizabeth A. Hinkley(1970)推導出變動點的最大概似估計量(Maximum likelihood estimate 簡稱為 MLE)之大樣本分佈(Asymptotic distribution)，並包括概似比例檢定量(Likelihood Ratio Tests)的大樣本性質。 A. N. Pettitt(1980)考慮條件檢定(Conditional test)與另一變動點的估計值，而且在他的模擬(Simulation)結果中，這個新的估計值較 MLE 為佳。 K. J. Worsley (1983)研究當二項分佈的總個數不同的情形，並比較概似比例檢定及累積總和檢定(Cumulative sum test ，簡稱 CUSUM test)的優劣。

研究常態分佈的報告不少，僅探討其中幾篇較具代表性的文章。 E.S. Page(1954)使用概似估計找出一個檢定方法，並用 MLE 估計變動點，但他的方法可適用於非常態假設，只要分佈及參數給定。 H. Chernoff 和 S. Zack(1964)考慮當常態分配的變異數為 1 的情形，給定貝氏檢定(Bayes test)並計算出檢定的臨界值(Critical value)與檢定力函數 (Power function)。 David V. Hinkley (1971)研究當常態分佈的期望值不設定下，用 MLE 檢定變動點發生的時間，大樣本的性質也同時被考慮。其他與變動點有關的介紹，可參考 S. Zacks (1991) 或 B. E. Brodsky 和 B. S. Darkhovsky (1993)。

三、實證分析

3.1 簡介

寫作風格有如個性特徵，充份表現出一位作者的特性；有些作者偏好華麗炫爛的詞藻，另一些以簡單易懂取勝；有些作者精於分析與推理，另一些則擅長描寫氣氛與場景；更由於小說豐富的題材，以及其獨特的文體的組合(詩詞、白話及文言文)，更容易凸顯出作者的寫

作風格。有鑑於此，針對小說文體的特性及作者用字習慣，本文的實證研究分為兩個部份：一為結構性的研究(如詩詞佔一回的字數比例)，一為用字的分析(如「兒」字的出現比例)，分別討論於 3.2 及 3.3 節，而變動點問題則於 3.4 節中討論。

本研究的資料分析以陳郁夫先生建立的紅樓夢資料庫為主，並以中央大學所建的公共網路版為輔助，但網路版有十八回缺佚，故僅作輔助及參考用。陳郁夫先生的資料庫以「彩畫本紅樓夢校注」一書為輸入依據，也就是前八十回參照「庚辰本」，後四十回參考「程甲本」。而本研究的字數計算及單字與詞之搜尋，由兩種程式執行所得：一為資料庫語言 DBase IV 所撰寫的程式，一為由陳郁夫先生提供的字詞檢索程式。

3.2 結構研究

作者的文筆主宰一部小說的風格，但為了故事結構及情節發展的需要，適時的增刪詩詞或是對話的比例勢所難免，藉以表達作者構思。以『紅樓夢』的第五回為例(回目：遊幻境指迷十二釵，飲仙醪演紅樓夢)，本回公認為『紅樓夢』一書的重心，為全書的未來情節發展舖路(註三)，回中包括如「新製紅樓夢(曲)十二支」和「金陵十二釵正冊」等線索，為書中主要女性人物勾勒出她們的習性特質，並事先為她們的結局與未來埋下伏筆，因此詩詞在第五回中共出現 1853 字，佔整回 6321 字的 29.3%，與全書一百二十回所有詩詞比例的 1.6% 高出許多。另外，第七十八回因情節需要(回目：老學士閑境姽嫿詞，痴公子杜撰芙蓉誄)，由賈政與眾人乘興各作詩詞，而詩詞也佔此回文字的 19.3%(詩詞 1745 字，全回 9039 字)。由此可見，每回的詩詞因情節需要，其出現頻率有相當程度的差異，本節的結構研究即針對此一特性，分作每回總字數、詩詞字數及對話字數三部份。

每回總字數與對話字數在前後各回中差別不大，雖然前八十回有較高的平均值，但其統計上並無明顯的差別；反而是六十一回至八十回的每回總字數與對話字數的平均值明顯較高。值得注意的是詩詞出現的次數及字數在前八十回與後四十回中有非常大的差別：前八十回中有四十三回無詩詞出現，而後四十回中則有二十八回無詩詞；出現詩詞的各回中，絕大部份的詩詞字數都在 200 字之內。其中前八十回出現詩詞且字數小於 200 的回數共二十一回，約

註　三：關於第五回的討論，詳見高陽著「紅樓一家言」中「曹雪芹對紅樓夢的最後構思」。

表 3.2－1 各回詩詞數字數

字數範圍	前80回	後40回
＝0	43	28
（＞0	37	12）
1~200	21	11
201~400	7	0
401~600	5	1
601~800	1	0
801~1,000	1	0
1,001~1,200	0	0
1,201~1,400	0	0
1,401~1,600	0	0
1,601~1,800	1	0
1,801-	1	0
總數	80	40

為出現詩詞的三十七回中的56.8%；而後四十回出現詩詞且字數小於200的回數有十一回，大約佔出現詩詞的十二回之91.7%，表3.2-1可充份顯示此一特徵。

以二項分配的方式比較前後半部中，詩詞是否出現在某一回中的機率，可得z檢定統計量：

$$\frac{\hat{p}_1-\hat{p}_2}{\sqrt{p(1-p)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}=\frac{37/80-12/40}{\sqrt{\frac{49}{120}\left(1-\frac{49}{120}\right)\left(\frac{1}{80}+\frac{1}{40}\right)}}=\frac{0.1625}{\sqrt{0.009060}}\approx 1.71$$

其中 p，p_1 及 p_2 為一百二十回、前八十回及後四十回中每回出現詩詞的機率，$n_1=80$ 及 $n_2=40$ 為其樣本數，其對應的P值為：

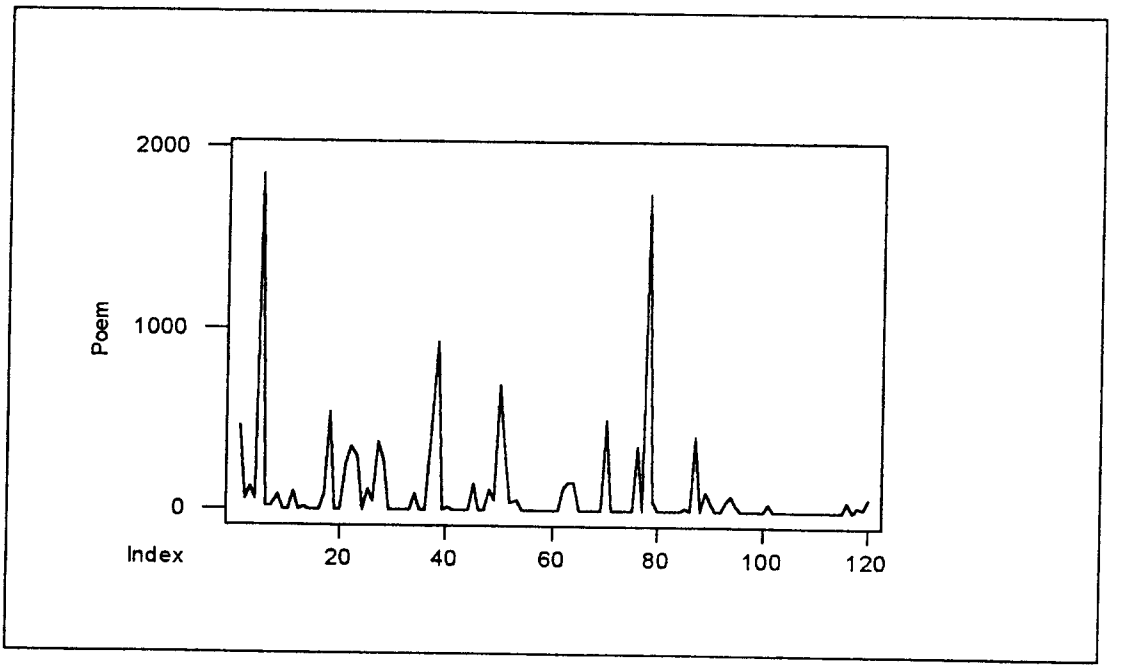
P值≒0.0439　若檢定 $H_0: p_1 = p_2$ V.S. $H_1: p_1 > p_2$

P值≒0.0878　若檢定 $H_0: p_1 = p_2$ V.S. $H_1: p_1 \neq p_2$

這兩個 P 值在統計上均可視為有顯著的不同，也就是前八十回中每回使用詩詞的可能性較後四十回為高。

前八十回與後四十回的平均每回詩詞字數的比較，也可由圖 3-2.1 的序列圖中看出不同，後四十回中除了第八十七回詩詞用字達 406 字外，其他出現詩詞的各回，用字不超過 100 字；此一現象與前八十回中，每隔數回即有一回詩詞超過 200 字的情形，相差甚多，即使去掉詩詞字數最多的第五回及第七十八回，此一特性在前八十回中依舊存在。以 t 檢定測試可得 P 值為 0.029(兩側檢定， Two-sided test)，其中前八十回平均每回出現 135.9 字的詩詞，後四十回僅有 22.9 字的詩詞，在統計上有顯著的不同，前八十回有較多的詩詞字數。同理，若以二項分佈來比較詩詞的字數佔每回總字數的比例，可得 t 檢定量約為 57.10，其中 P 值幾乎為 0，顯示詩詞在每回所佔的比例，前八十回明顯的高於後四十回，此結果與每回詩詞出現的可能性，前八十回高於後四十回是一致的。

圖 3.2-1 每回詩詞字數序列圖



3.3 用字分析：

在第二節曾提到，連接詞(Function Words)曾被使用作為比較英文寫作風格。由於連接詞的功能在於連接句子，一般不至於改變句子的原意，而且通常一個連接詞有其他同義詞可作替換，慣用某一連接詞可視為作者的偏好。本節的用字分析即源於這個動機，選用可替換並且不影響句子的字詞為分析標準，判斷『紅樓夢』前後半部的常用字詞是否不同，藉以推論是否有曹雪芹以外的作者參與創作，這個部份的分析詳述於 3.3.1 及 3.3.2 節，另外，3.3.3 節為每回結尾時採用的詞句，作為連接兩回的橋樑，這類的詞句在章回小說中一般為「且聽下回分解」或類似之詞句。3.3.4 節則綜合分析，同時考慮所有的用字與詞。

3.3.1 及 3.3.2 節的用字分析，參考趙岡與陳鍾毅(1980)的統計研究，首先探討五個虛字「兒」、「在」、「了」、「的」、「著」，這些虛字為北平話裏常見的語助詞，在句子裏可有可無，趙岡與陳鍾毅舉出以下的例子作為參考：

「便伏在枕上歇一會」和「便伏枕上歇一會」。

「寶玉已醒了」和「寶玉已醒」。

「各房的丫頭」和「各房丫頭」。

「笑著說」和「笑說」。

3.3.2 節考慮的其他字詞，計有「嗎」和「麼」，「給」和「與」，「都」和「多」，「我們」和「咱們」(或「偺們」) 。

3.3.1 五虛字的分析：

趙岡與陳鍾毅對五個虛字的分析，乃是將前八十回及後四十回中各抽出 100 頁，每頁各有 720 字，該頁沒有回目也沒有詩詞或分段，逐頁計算五個虛字的出現次數，並使用 t 檢定比較這五個虛字平均每頁的平均值。本節也採用相同的方法，但考慮前八十回及後四十回的所有文字，以去除因抽樣而產生的誤差。

表 3.3.-2 為前八十回及後四十回中的五個虛字比較，括號內為每千字的平均出現次數，如「兒」字在前八十回平均每千字出現 8.03 次，後四十回中則為每千字出現 8.73 次，全書一百二十回的總平均為每千字出現 8.25 次。

表 3.3.－2 五虛字出現次數比較

	兒	在	了	的	著	總字數
1~80回	**4024** **(8.03)**	**2405** **(5.00)**	**14293** **(28.51)**	**10216** **(20.83)**	**3782** **(7.54)**	**501284**
81~120回	**2066** **(8.73)**	**1501** **(6.34)**	**6956** **(29.38)**	**5513** **(23.29)**	**2382** **(10.06)**	**236740**
1~120回	**6090** **(8.25)**	**4005** **(5.43)**	**21249** **(28.79)**	**15729** **(21.31)**	**6164** **(8.35)**	**738024**

這五個虛字的出現次數 t 檢定值分別是：

兒	3.10
在	7.31
了	2.08
的	8.08
著	11.10

這五個檢定值均屬顯著，其中除了「了」字的 P 值約為 0.02 外，其他四個虛字的 P 值都小於 0.001，而後四十回使用這五個虛字的次數明顯較前八十回高。此一結果與趙岡及陳鍾毅的結果非常接近(但「了」字在他們的分析裏並不顯著)，同時也建議這五個虛字的出現模式在前後半部中並不一致，可能是後半部有不同的作者，或是曹雪芹的用字習慣因寫作歷時十年，而有不同的寫作風格。為更進一步探究這五個虛字是否可能為漸進式的增加，在 3.3.4 節中，我們將使用變動點方法，分析變動的各項特徵。

3.3.2 其他字詞分析：

(1)「嗎」和「麼」：問句後的結尾可選用「嗎」或「麼」，但表 3.3-3 的各回問句以「嗎」

字結尾的出現次數呈現一個非常極端的現象，前八十回僅有一回出現「嗎」字共兩次，其他七十九回皆為0次；反觀後四十回，其中共二十一回無「嗎」字。單純以每回是否出現「嗎」字作比較，二項檢定量為8.29，支持前後半部在使用「嗎」字上，有非常明顯的差異；以t檢定測試平均「嗎」字的出現頻率，結果也是如此。由圖3.3-2中也可看出以上特性。另外，表3.3-3的後四十回出現次數，顯示作者在後四十回使用「嗎」字相當一致，但前八十回似乎完全沒有使用這個字的習慣，因此，前八十回與後四十回分屬不同作者較有說服力，就問句以「嗎」字結尾的出現情形，我們傾向於接受有不同作者的假設。

「麼」字出現次數也是如此，前八十回中多數每回使用「麼」為問句結尾不多於3次佔了六十六回，大約比80%多一些；反觀後四十回，「麼」字為結尾的問句多於3次的有二十六回，佔了後四十回的65%，以z檢定考慮每回是否出現三次或三次以上的「麼」字，可得檢定值約5.41，P值幾為0(後四十回每回平均出現5.1次「麼」，前八十回僅有2.3次)，此一結果與「嗎」字相同。若單純以每回所有的「嗎」及「麼」字(非問句結尾)作比較也有相同的結果，以下為其t檢定值。

表3.3－3各回問句以「嗎」字結尾的出現次數

次數範圍	前80回	後40回
0	79	21
1	0	10
2	1	7
3	0	2
總 數	80	40

圖 3.3-2 每回問句以「嗎」字結尾的出現次數序列圖

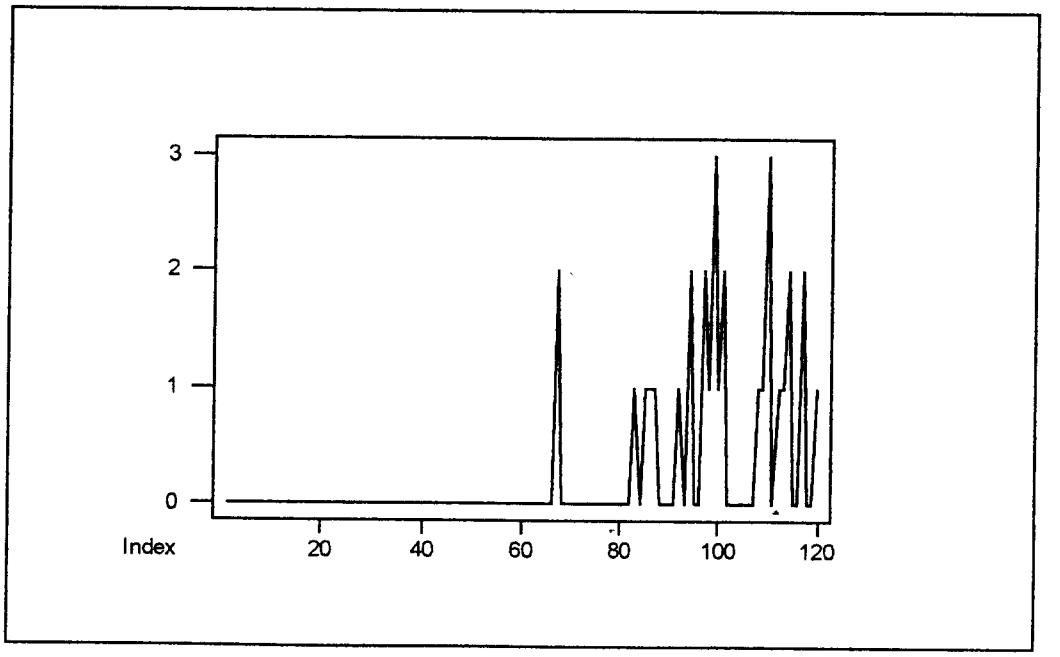


表 3.3-4「嗎」與「麼」兩字的 t 檢定值

	「嗎」字每回平均次數	「麼」字每回平均次數
前八十回	0.038　t 檢定值 =62.77	24.92　t 檢定值 =33.90
後四十回	1.875　P -Value=0.000	41.60　P -Value=0.000

(2)「給」和「與」：這兩者前後半部混用，但前八十回使用「與」字比「給」字多些，而後四十回則是「給」字多於「與」字。其中前八十回及後四十回在使用「給」字上不分軒輊，但前八十回明顯的使用較多的「與」字，每回平均約有 10.75 次，約為後四十回的平均每回 5.1 次的兩倍，t 檢定為 5.93，P 值幾乎為 0。

表 3.3-5「給」與「與」兩字的 t 檢定值

	「給」字每回平均次數	「與」字每回平均次數
前八十回	8.61　t 檢定值 =0.351	10.75　t 檢定值 =5.93
後四十回	9.78　P -Value=0.87	5.10　P -Value=0.000

(3)「都」和「多」：前八十回使用「都」和「多」的次數，都比後四十回高。另外，在前八十回中使用「都」和「多」次數高的回數也較多，整個次數分佈函數也較分散；而後四十回的次數分佈函數較集中，沒有非常高或低的使用次數，舉例來看，後四十回中大多數每回出現約 10 至 30 次的「都」字，「多」字也集中在 2 至 12 次。t 檢定值也支持我們的推論：

表 3.3-6「都」與「多」兩字的 t 檢定值

	「都」字每回平均次數	「多」字每回平均次數
前八十回	23.67　t 檢定值 =2.06	10.99　t 檢定值 =4.28
後四十回	19.60　P -Value=0.042	7.32　P -Value=0.000

(4)「我們」及「咱們」：由於南方人說話都用「我們」，而北京話中的「我們」及「咱們」並不完全相同。「我們」在前八十回平均每回出現 10.48 次，稍高於後四十回的 9.85 次，但沒有顯著的差別；「咱們」出現的次數也類似，後四十回的平均每回 5.68 次比前八十回的 4.91 次高，但也非顯著不同。由於前八十回使用較多的「我們」，後四十回有較多的「咱們」，若以前八十回作者的作者為南方人及後四十回作者為北方人作為解釋（也就是說前八十回為曹雪芹所作，後四十回作者為高鶚所作），似乎可說明此一現象。但因為使用「我們」及「咱們」的習慣差異，在前後半部中不顯著，我們也可作不同的詮釋，譬如曹雪芹使用「我們」及「咱們」的習慣，隨著居住在北京的日子增加而改變，因為「我們」及「咱們」的每回平均出現次數在 1~60 回都有上升的趨勢，在下一節中，這個假設可由變動點分析作驗證。

3.3.3 每回結語用詞：

章回小說每每在每回最後引入使劇情撲朔迷離或急轉直下的文字，使讀者難以自已、欲罷不能；而說書人更可藉此製造緊張懸疑的氣氛，讓聽眾流連忘返。這是章回小說的特色之一。表 3.3-7 為各回結語用詞的統計表：

表 3.3-7 各回結語用詞

回末用詞	前 80 回	後 40 回
下回分解	3	29
要知端的，且聽下回分解	23	0
且聽下回分解	14	7
無任何結語	15	0
詩	6	1
要知端的	7	0
欲知後事且聽下回	1	3
其他（共八種）	11	0
總數	80	40

後四十回的結語用詞只有四種，與前八十回的 15 種有非常大的差別，若以卡方檢定來比較這兩組資料，可得

表 3.3-8 各回結語用詞分類表

類別	1	2	3	4	5	6	7	8	總數
前 80 回（Y_1）	3	23	14	15	6	7	1	10	80
後 40 回（Y_2）	29	0	7	0	1	0	3	0	40
總數	32	23	21	15	7	7	4	10	120

$$\chi^2 - \text{檢定值} = \sum_{i=1}^{2}\sum_{j=1}^{8}\frac{\left(Y_{ij} - n_j\hat{P}_j\right)^2}{n_j\hat{P}_j} = 78.30$$

$$\hat{P}_j = \frac{Y_{1j} + Y_{2j}}{120}, j = 1,2,.....,8$$

$$n_1 = 80, n_2 = 40$$

P值幾乎為0，因此回末用語有顯著的差別。又因章回小說多以「下回分解」作為回末結語，若以是否出現「下回分解」作為比較標準，可得前後半部各有40及36次，以二項檢定測試，其檢定量為4.29，P值約為0.038，有相當強的證據支持前八十回及後四十回有不同的使用「下回分解」習慣。單以回末用語作為考慮，我們較支持『紅樓夢』的前後部半分屬不同作者。

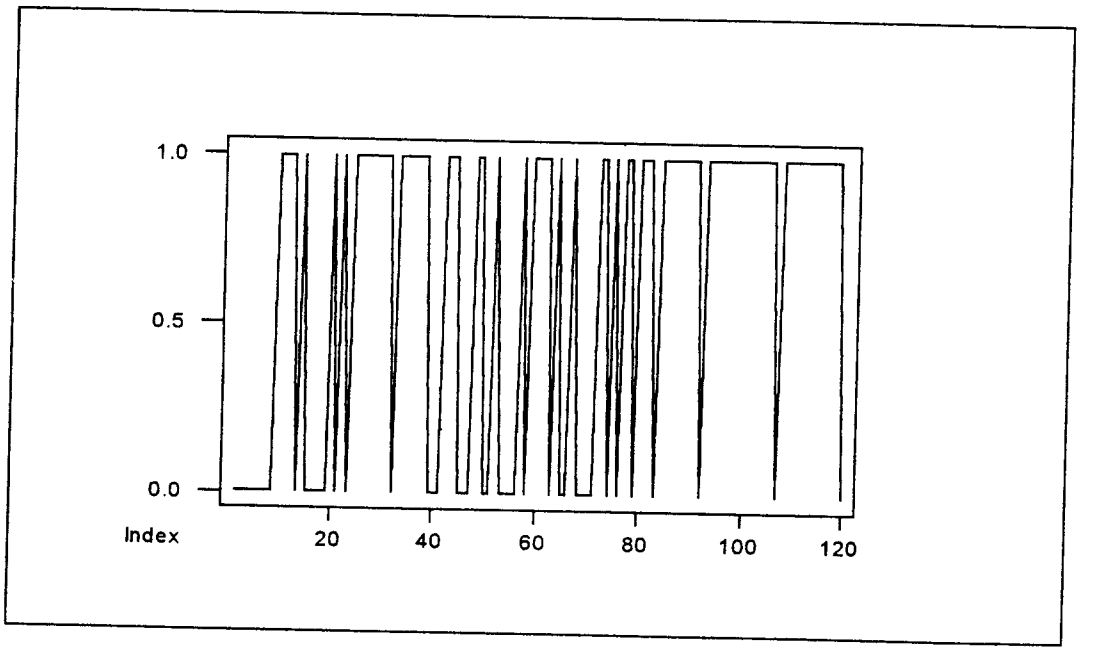
若比較1~60回及61~80回，可得

表 3.3-9 各回結語用詞分類表

類別	1	2	3	4	5	6	7	8	總數
1~60回	1	18	11	12	6	7	0	3	60
61~80回	2	5	3	3	0	0	1	7	20
1~80回	3	23	14	15	6	7	1	0	80

X^2檢定值=21.34，P值為0.0033，也有顯著的不同。但若以「下回分解」為主要考慮重點，1~60回與61~80回各有30及10次，出現頻率皆為0.5，可認定並無差異。假若將是否使用「下回分解」視為1或0的結果，則1~120回的使用「下回分解」情形可表示成圖3.3-3的序列圖。後四十回幾乎全為「下回分解」外，1~40回中連續使用或不使用「下回分解」較常見，41~80回則較平均地使用，甚少連續使用（或不使用）「下回分解」，在下一節的分析中將再作更詳細的分析。

圖 3.3-3 每回回末是否出現「下回分解」序列圖



3.3.4 綜合分析:

若將上述用字分析考慮的十三個字詞，作為使用辨別分析的參考依據，判定前八十回及後四十回是否可視為兩個不同的群體，可得結果：

表 3.3-10 辨別分析檢定義

True Group / Put into Group	1 (1~80 回)	2 (81~120 回)
1 (1~80 回)	75	4
2 (81~120 回)	5	36
total	80	40

Squared Distance between Group=7.16587

也就是用線性判別函數做分類，將前八十回及後四十回視為不同類別的群體時，前八十回中僅五回分類錯誤，而後四十回也只有四回，判別的正確率高達 92.5% 。另外，由 F 檢定做測試，我們也可得出相同的結論，認定前八十回及後四十回在使用這十三個字詞上，有顯著的不同，所得之 F 檢定值如下：

$$\left(\frac{n_1+n_2-p-1}{(n_1+n_2-2)p}\right)\left(\frac{n_1 n_2}{n_1+n_2}\right)D^2 = 94.62 \quad > \quad F_{p,n_1+n_2-p-1}(0.01)$$

其中 $n_1=80$ 及 $n_2=40$ ， $P=13$ 為考慮的判別因素個數， D=7.16587 為兩類群體的距離，可得 P 值幾乎為 0。由這個判別分析，我們明顯地看出前後半部於用字上確有不同，完全出自同一作者的可能性非常低。

3.4 變動點分析：

本節使用的變動點分析分法主要為二項分佈的假設，根據 K.J.Worsley（1983）的研究，且因為 CUSUM 檢定與 Likelihood Ratio 檢定結果相近，但 Likelihood Ratio 檢定無法處理觀測值為 0 的情形，原則上我們用 CUSUM 檢定來比較。以下為我們檢定所採用的方法：

$$X_i \overset{iid}{\sim} B(n_i, p_i) \quad 0 \le p_i \le 1 \quad for \quad i = 1,2,.....,120$$

其中 n_i 為第 i 回的總字數， p_i 為第 i 回中出現某一特定字詞的機率。一般而言，我們要檢定的假設是

$$H_0: p_i = p \quad i = 1,2,.....,120 \quad \text{v.s.} \quad H_1: p_i = \begin{cases} p & i = 1,.....,k \\ p' & i = k+1,...,120 \end{cases}$$

其中 $p \neq p'$ 。假設 m_i 為第 i 回中出現此一特定字詞的次數，則可定義

$$\begin{cases} M = \sum_{i=1}^{120} m_i \\ N = \sum_{i=1}^{120} n_i \end{cases} \qquad \begin{cases} M_K = \sum_{i=1}^{K} m_i \\ N_K = \sum_{i=1}^{K} n_i \end{cases} \qquad i = 1,2,\ldots\ldots,120$$

CUSUM 檢定量則為

$$Q_K = \frac{M_K - r_K M}{\sqrt{N\sigma^2}}$$

其中

$$P_0 = \frac{M}{N} \quad , \quad \sigma^2 = P_0(1-P_0) \quad r_K = \frac{N_K}{N} \quad , \quad S_K^2 = r_K(1-r_K)$$

而變動點為對應於最大 $|Q_K|$ 值的 k，顯著程度可用 Q_K^2 / S_K^2 來檢定與 Pearson χ^2 檢定量相同。

變動點若出現在第八十回附近(即前八十回與後四十回不同)，可支持兩個不同作者的假設；反之則不然。變動點在第八十回附近的字詞計有「在」、「著」、「嗎」、「麼」、「與」及每回結尾用語共六個。其他字詞的變動點則較無規律，例如「兒」、「了」、「的」三字的變動點是在第五回，「給」、「都」、「多」、「我們」、「咱們」則散佈在第四十回與第八十回間，無法作為兩個不同作者的佐證。

四、結論與建議

4.1 結論:

鑑定作者的寫作風格相當不容易，作為評判的標準也是見仁見智，某甲認為足以作為結論的證據，某乙可能不以為然。即使是大家都公認的標準，也可能產生難以令大家獲得共同結論的狀況。在現實生活中，因為每個人都有自己的效用函數（Utility Function），多半可依據自己的需要，作出最合乎個人效用函數的選擇。在本篇研究中，我們先後就文體結構、用字

分析及結語用詞作分析，共考慮了17個不同的數值作比較，其中除了少數幾個數值（每回總字數、對話比例、「給」、「我們」、「咱們」），其統計檢定不足以支持前八十回與後四十回不同外，我們發現前後半部確實有顯著差異（詳見表4-1）。若以變動點分析為標準，支持前後半部不同點在八十回前後者，則有詩詞比例、「在」、「著」、「嗎」、「麼」、「與」、每回結尾用語，問題是17個數值中有7個支持『紅樓夢』在第八十回附近，寫作風格有一突然而明顯的轉變，我們能否根據此一分析結果，作出『紅樓夢』一書可能有兩個或兩個以上的作者呢？這是在作出本研究的結論前，必須考慮的問題。

純粹以7/17的命中率來看，似乎不太能作出有力的結論，但我們有不同的看法。基本上，我們認為作者固然在用字遣詞時不見得前後一致，但我們選用的17個判定標準中，竟有12個建議前八十回與後四十回的風格不同；藉由其中7個判定標準更可指出風格的改變在第八十回前後，在機率上如仍堅持『紅樓夢』只有一位作者，其正確的可能性則是微乎其微。

或許有人會以『紅樓夢』一書創作歷時十年為理由，認為曹雪芹極可能隨年紀及經驗的增長，在寫作風格上產生變化，而有前八十回及後四十回不同的用字習慣。但我們也不能忽略另一件事實，『紅樓夢』歷經五次增刪，曹雪芹在不同時間校閱自己過去的文章，他的用字習慣應該同時會影響全書一百二十回的文字，而不僅僅是後四十回。根據紅學專家趙岡與陳鍾毅的研究，『紅樓夢』在第一次或第二次增刪時即已完成全書一百二十回的初稿，故事的大綱與結構應已成形，第三次到第五次的增刪則著重於修飾或補充，例如增補詩詞以符合情節需要，或從曹家其他人的評語建議中增寫新的情節。因此作者因創作時間的增長，而產生寫作風格改變的推論，我們並不支持。

根據我們的分析結果，『紅樓夢』有兩個或兩個以上的作者較為可行。但一般認為的前八十回為曹雪芹所作，後四十回為高鶚所作卻也不是我們較支持的論點。在本文第一節即已指出，『紅樓夢』為一集體創作，代表的是曹家數代的生活經驗，絕非高鶚一人可憑空杜撰，並在短期之內可以延續完成。趙岡及陳鍾毅(1980)在他們的書中，提出一個可能的解答：由於前八十回描寫曹家繁榮景象，以戲劇對比為考慮，後四十回應為敘述賈家遭抄家而家破人亡的慘狀，如此更可突顯世事無常，天子喜怒不定與事君如事虎。當然，這種寫法必定遭致當政者的封殺，不得已，原書的結局必須重新修改，以傳後世。但改寫結局並不是那麼容易，

尤其曹雪芹在第五次增刪後，因生活的壓力而必須任教以賺取生活費，因此並沒有足夠的時間將改寫的部份再交由同族的親戚校閱。

根據我們推測，可能曹雪芹完成了新的結局但未經其親戚校閱，因此未與前八十回同時流入世面；也可能曹雪芹在去世前並未將結局改寫完畢，而參與創作的其他人(如曹天祐、曹堂村及畸笏叟)為了使自己的心血公諸於世，繼續完成結局的改寫。但因為某種不知名的原因(可能窮困潦倒)，無法將最後四十回付梓，直到程偉元購得後四十回的版本，經高鶚的整理後『紅樓夢』才得以一百二十回的面目公諸於世。其中前後各回的用字可能是原稿散佚不全，經高鶚增補文字後而形成的差異；或者是曹雪芹的族人續書時，因文筆不同而造成的。無論是以上哪一種可能(或兩者皆是)，『紅樓夢』的一百二十回版本，應有兩位或兩位以上的作者，而曹雪芹以外的作者只直接或間接參與了後四十回的寫作，故事的主體應遵循曹雪芹的原意，如此前八十回與後四十回在情節的銜接上才能毫無破綻。當然，我們的推論雖然參考了幾位紅學專家的想法，再加上本研究的統計分析，一定也有不完善的地方，在後續的研究中，我們將考慮其他可能的統計方法，用另一個角度分析『紅樓夢』的作者問題。

4.2 建議：

本研究之原始動機來自於變異點的分析，由於過去對『紅樓夢』的數量化分析多半停留在兩個樣本，並無考慮前八十回及後四十回的差異是否為一漸進式的改變，這也是本研究與眾不同之處。但限於對於紅學知識的不足，在選取判斷標準上遭遇不少問題，尚有許多其他測度量未能考慮。

本研究未來可能的方向可朝向品種問題(Species Problem)及重覆補取模型(Capture-Recapture Model)發展。以品種問題而言，我們可以把每回出現的文字視為觀查值，計算共有多少不同的字出現，每個出現的字共出現幾次，由這些資料判斷回與回之間是否相同，進而論証是否全書一百二十回可分為兩組不同的族群(Population)，其中前八十回在一個族群內，而後四十回在另一族群。如此則可避免比較標準選取的問題，使得結果更具說服力。另一個研究方向則是與紅學專家合作，選取其他有意義的字詞或理念作為比較標準。舉例而言，施鐵民(1994)在甲戌年台灣紅學會議中，提出一篇研究『紅樓夢』意象的文章，其中論証「紅」、「綠」

使用頻率之高，應非偶然(全書有九千條含紅綠及其同義顏色的例子)。如果可與紅學專家合作，將文學上的比較標準數量化，可賦予統計分析在文學專業上的意義。

表 4-1 統計檢定結果總表

	一般統計檢定	變動點分析 (變動點是否在第 80 回附近)
每回總字數	不顯著	——
每回詩詞字數	顯著	是
每回對話字數	不顯著	——
兒	顯著	不是
在	顯著	是
了	顯著	不是
的	顯著	不是
著	顯著	是
嗎	顯著	是
麼	顯著	是
給	不顯著	——
與	顯著	是
都	顯著	不是
多	顯著	不是
我們	不顯著	——
咱們	不顯著	——
每回結尾用語	顯著	是

註：一般統計檢定不顯著者，不再考慮變動點分析。

參考文獻

(1)中文部份:

1. 趙岡 , 陳鍾毅(1980)," 紅樓夢研究新編 ", 聯經出版社 .
2. 高陽(1977)," 紅樓一家言 ", 聯經出版社 .
3. 高陽(1983)," 高陽說曹雪芹 ", 聯經出版社 .
4. 馮其庸等校注(1984)," 彩畫本紅樓夢校注 ", 里仁書局 .
5. 周汝昌(1994)," 甲戌年話甲戌本披露之原委 ", 甲戌年台灣紅學會議論文 .
6. 王三慶(1994)," 紅樓夢電腦 —《紅樓夢》研究與電腦科技 ", 甲戌年台灣紅學會議論文 .
7. 施鐵民(David Steelman, 1994)," 紅學為體電腦為用 -- 從紅樓夢的意象談起 ", 甲戌年台灣紅學會議論文 .

(2)英文部份:

1. Bailey, B. J. R. (1990) "*A Model for Function Word Counts*", Applied Statistics, 39, pp.107-114.
2. Bissel, A. F. (1995) "*Weighted Cumulative Sums for Text Analysis using Word Counts*", Journal of Royal Statistics, Series A, 158, pp.525-545.
3. Brodsky, B. E. and Darkhovsky, B. S. (1993) "*Nonparameteric Methods in Change-point problems*", Academic Publishers.
4. Chernoff, H. and Zacks, S. (1964) "*Estimating the current mean of a Normal Distribution which is Subject to Changes in Times*", Annals of Mathematical Statistics, 35, pp.999-1028.
5. Efron, B. and Thisted, R. (1976) "*Estimating the Number of Unseen Species: How many Words did Shakespeare Know?*", Biometrika, 63, pp.435-447.
6. Hinkley, D. V. and Hinkley, E. A. (1970) "*Inference about the Change-point in a Sequence of Binomial Variables*", Biometrika, 57, pp.447-488.
7. Hinkley, D. V. (1971)" *Inference about the Change-point from Cumulative Sum Tests*", Biometrika, 58, pp.509-523.
8. Holmes, D. (1995) "*Who was the Author?*", Journal of Royal Statistics News, Vol. 23, No. 2, pp.1-2.
9. Horvath, L. (1989) "*The Limit Distributions of Likelihood Ratio and Cumulative Sum Tests for a Change in a Binomial Probability*", Journal of Multivariate Analysis, 31, pp.148-159.
10. Karlgren, B. (1952), "*New Excursions in Chinese Grammar*", in Bulletin of the museum of Far Eastern Antiquities (Stockholm), No. 24, pp.51-80.
11. Mosteller F. and Wallace D. L. (1984) "*Applied Bayesian and Classical Inferrence:The Case of The Federalist Papers*", Springer-Verlag.
12. Pettitt A. N. (1979) "*A Nonparametric Approach to the Change-point Problem*", Applied Statistics, 28, pp.126-135.
13. Pettitt A. N. (1980) "*A Simple Cumulative Sum Type Statistic for the Change-point Problem with*

Zero-one Observations", Biometrika, 67, pp.79-84.

14. Smith, A. F. M. (1975) "*A Bayesian Approach to Inference about a Change-point in a Sequence of Random Variables*", Biometrika, 62, pp.407-416.
15. Thisted, R. and Efron, B. (1987) "*Did Shakespeare Write a Newly-discovered Poem?*", Biometrika, 74, pp.445-455.
16. Sichel, H. S. (1986) "*Word Frequency Distributions and Type-token Characteristics*", The Mathematical Scientist, 11, pp.45-72.
17. Williams, C. B. (1975) "*Mendenhall's Studies of Word-length Distribution in the Works of Shakespeare and Bacon*", Biometrika, 62, pp.207-212.
18. Worsley, K. J. (1983) "*The Power of Likelihood Ratio and Cumulative Sum Tests for a Change in a Binomial Probability*", Biometrika, 70, pp.455-464.
19. Yue, C. J. (1994) "*Bayesian Sequential Tests for Comparing the Species Richness of Two Populations*", Ph.D. thesis, Univ. of Wisconsin-Madison.
20. Zacks, S. (1991) "*Detection and Change-point Problems*", Handbook of Sequential Analysis, Mercel and Dekker.