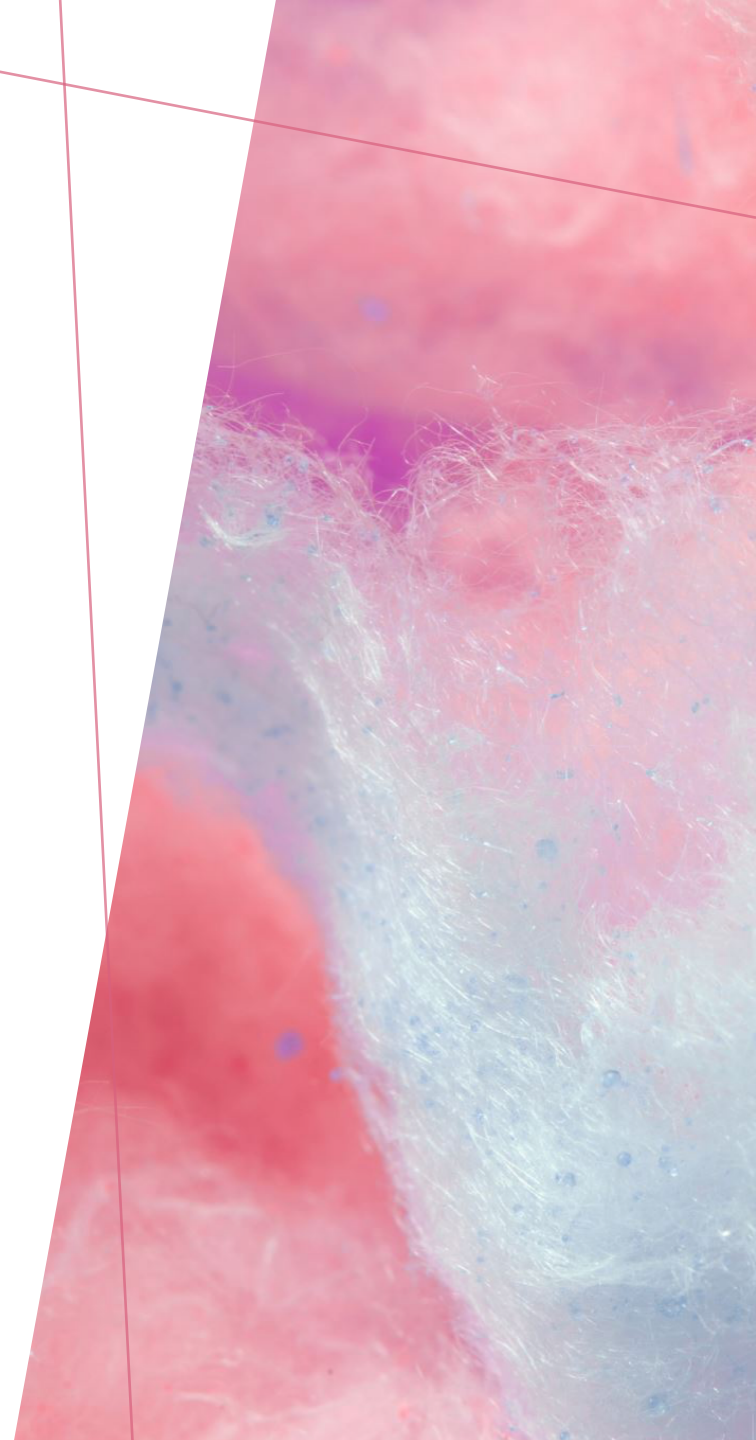
統計在紅樓夢的應用

THE APPLICATION OF STATISTICS IN “THE DREAM OF RED CHAMBER”

AUTHOR : C. JACK YUE

報告人：

統碩一　112354011李沂瑾



目錄

前言

- 由於以前朝代印刷昂貴及缺乏著作權觀念，導致小說的不完整性

- 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

研究目的及方法

- 一般認為前80回為曹雪芹所著，後40回為高鶚續作
- 透過統計分析將裡面文字敘述數量化，尋求解答作者的可能性
- 採用版本為庚辰本及程甲本

文獻探討

- 趙岡與陳鍾毅使用「兒」、「在」、「了」、「的」、「著」五個虛字作為比較，分為前80回及後40回兩組樣本，各抽樣100頁計算出現頻率得t檢定值如下

兒	3.677
在	3.392
了	0.116
的	3.391
著	3.910

統計方法

1.假設前80回與後40回來自兩個不同的主體，採用兩個樣本的方法檢定

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的評論依據

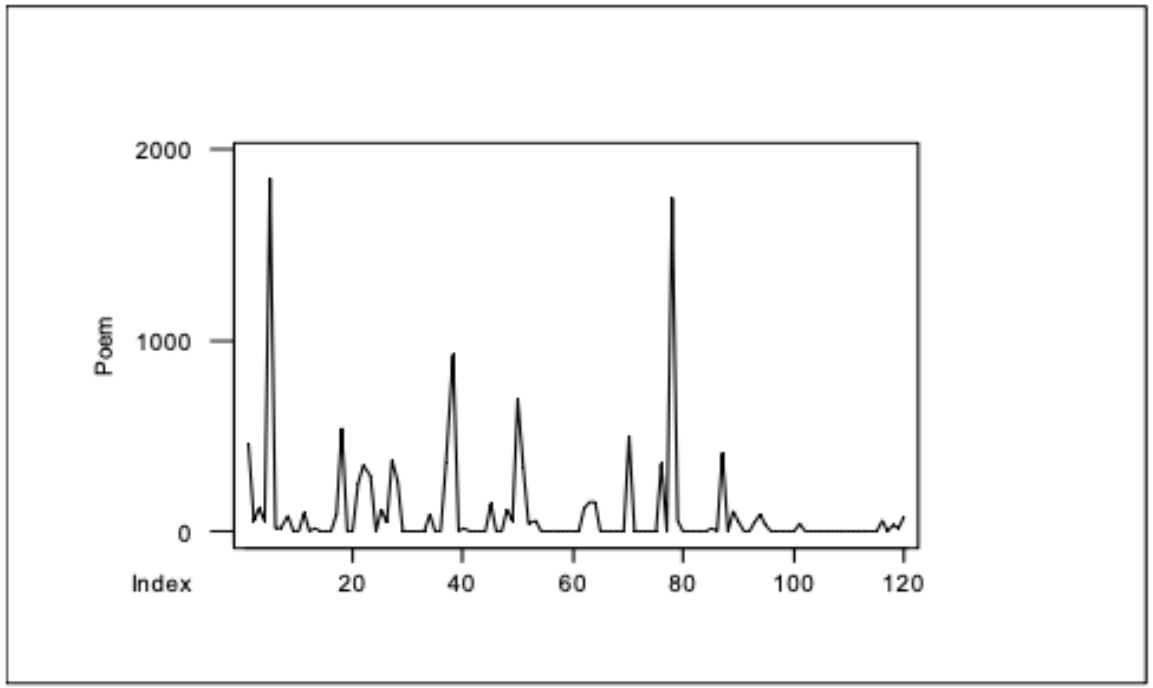
2.假設120回來自同一主體，使用變動點問題的方法檢定

判斷120回小說是否有前後不一的現象，轉折點是否出現在80回附近

實證分析—兩個樣本（T檢定）

1. 文體結構：每回總字數、詩詞字數、對話字數

圖3.2-1 每回詩詞字數序列圖



* 底色為顯著

實證分析—兩個樣本（T檢定）

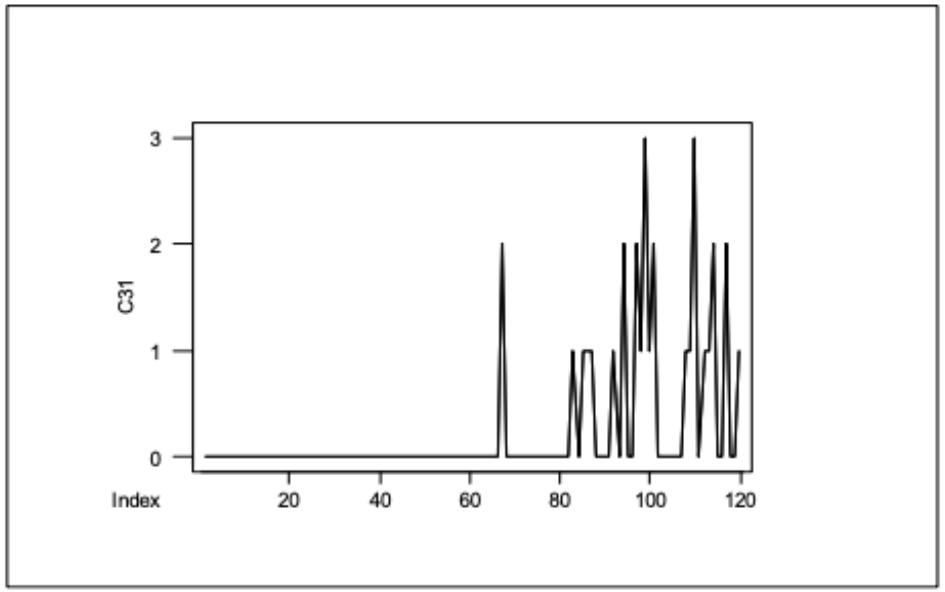
2. 用字分析：

「兒」、「在」、「了」、「的」、「著」五個虛字

「嗎」和「麼」、「給」和「與」、「都」和「多」、「我們」和「咱們」

兒	3.10
在	7.31
了	2.08
的	8.08
著	11.10

圖3.3-2 每回問句以「嗎」字結尾的出現次數序列圖



實證分析—兩個樣本（T檢定）

3. 每回結語用詞：

在後40回幾乎全為使用「下回分解」，有明顯差異

表3.3-4 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

實證分析—變動點分析(累積總和檢定CUSUM TEST)

- 假設為二項分佈　　$X_i \sim B(n_i, p_i)$, $0 \le p_i \le 1$, $for\ i = 1,2,\ldots..,120$
  (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)
- 假設檢定為 $H_0: p_i = p$　i $= 1,2,\ldots.120$　v.s.　$H_1: p_i = \begin{cases} p & i = 1,\ldots\ldots\ldots..,k \\ \acute{p} & i = k+1,\ldots.,120 \end{cases}$　　其中 $p \neq \acute{p}$
- CUSUM檢定量為 $Q_k = \frac{M_k - r_k M}{\sqrt{N\sigma^2}}$

  其中 $\begin{cases} M = \sum_{i=1}^{120} m_i \\ N = \sum_{i=1}^{120} n_i \end{cases}$ $\begin{cases} M_k = \sum_{i=1}^{k} m_i \\ N_k = \sum_{i=1}^{k} n_i \end{cases}$ *i=1,2,….,120*，假設m_i為第i回出現此一特定字詞的次數，

  $P_0 = \frac{M}{N}$，$\sigma^2 = P_0(1-P_0)$，$r_k = \frac{N_k}{N}$，$S_k^2 = r_k(1-r_k)$

- 而變動點為對應於最大$|Q_k|$值的k

實證分析—變動點分析

- 詩詞比例、「在」、「著」、「嗎」、「麼」、「與」及每回結尾用語共7個，變動點在第80回附近
- 「兒」、「了」、「的」三字的變動點在第5回
- 「給」、「都」、「多」、「我們」、「咱們」則散佈在第40回與第80回間

→ 無法作為有兩個不同作者的佐證

結論

	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。

結論

- 就兩個樣本分析，共考慮17個不同的數值比較，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80回與後40回不同外，其餘確實在前後半部有顯著差異
- 以變動點分析為標準，支持前後半部不同點在80回前後者，有詩詞比例、「在」、「著」、「嗎」、「麼」、「與」、每回結尾用語，17個數值共有7個

- 認為紅樓夢作者有兩個或兩個以上

THANK YOU