莎士比亞新詩眞僞之鑑定

黃文璋教授
國立高雄大學應用數學系

西元1985年11月14日, 英國及美國的報紙及雜誌爭相報導研究莎士比亞(Shakespeare, 1564-1616, 英國劇作家及詩人)的美國學者Gary Taylor, 在英國牛津大學(Oxford University) 的Bodleian圖書館(是英國歷史最悠久、最重要, 且藏書只供參考概不外借的圖書館之一, 現有館藏420餘萬册印刷本及5萬餘册手稿本)的一本自西元1755年起收藏的書中, 找到一首很可能是莎士比亞的抒情詩(見Lelyveld (1985) 及Taylor (1985) 。該詩的第二抄本, 稍後在美國耶魯大學 (Yale University) 的圖書館出現。我們姑且先稱這首詩爲泰勒詩(Taylor poem) 。如果能證實泰勒詩確爲莎士比亞所作, 則這將是十七世紀以來,莎士比亞作品最重要的一次發現。英美學者爲了這首詩之眞僞爭論不休, 大打筆戰。只是不少專家認爲這首泰勒詩不論在用字遣詞與韻味風格上, 都迥異於莎士比亞其他的作品。有趣的是, 統計學者也介入這場紛爭。西元1986年1月24日出版的Science雜誌, 刊登了一篇"莎士比亞的新詩–向統計學禮讚"(Shakeapeare's new poem : an ode to statistics), 介紹著名的統計學者, 美國史丹福大學(Stanford University)的Efron (曾來過台灣)教授及芝加哥大學(University of Chicago) 的Thisted 教授, 如何以統計的方法鑑定這首新出土的泰勒詩, 是否爲莎士比亞所作 。

早在西元1976年, Efron 和Thisted 便把莎士比亞作品中所用的字做了一番統計分析。他們想回答若發現了一新的作品, 如何經由統計分析(statistical analysis)其中所用的字, 以決定此作品是否爲莎士比亞所作? 他們當初做此分析只是爲了好玩, 沒想到十年後眞派上用場 。

這並非統計學家第一次協助解決文學上的問題。而由於統計分析是如此的具說服力, 因此往往能使一文學上長期的爭論, 迅速地平息。例如, 哈佛大學 (Harvard University)的Mosteller及Wallace 利用統計方法, 判定The Federalist Papers 的作者爲James Madison, 而非Alexander

Hamiliton。在西元1950年代, 著名的英國 統計學者David Cox及文學家 L. Brandwood 也曾利用統計方法, 以解決爭論長達1000年的關於柏拉圖(Plato, 約西元前428-347年)眾多作品的先後順序。柏拉圖爲古希臘三大哲學家之一, 和蘇格拉底 (Socrates, 約西元前470-399年)及亞里斯多德(Aristotle, 西元前384-322年), 共同奠定西方文化的哲學基礎。柏拉圖本來對民主抱著希望, 但在蘇格拉底被判死刑後, 他終於認清一有良知的人, 在活躍的政治中是無容身之處的(There is no place for a man of conscience in active politics), 頗令人深省。

雖然利用統計方法來解答文學上的問題之想法並非創新, Efron 與Thisted 所採用的特別方法卻不曾在這方面用過。此方法可追溯至西元1940年代, 那時生物學家 Williams 向英國統計學家費雪(R. A. Fisher, 1890-1962, 現代統計學的奠基者, 其生平事蹟可參考Gower (1990/91)), 提出一個似乎不可能回答的問題。Williams曾前往馬來西亞採集蝴蝶, 他把自己共見過幾種(species)蝴蝶, 以及每種各見過的次數(有些見過幾十次, 有些見過幾次, 有些只見過一次), 都告訴費雪。Williams 想知道在馬來西亞的蝴蝶中, 他沒見過的究竟有多少?

一般人會覺得此問題毫無頭緒, 不可能解答的。但統計學家卻有辦法估計尚有幾種蝴蝶還未被捕捉到。只要假設蝴蝶是依照每一種之隻數的比例, 隨機地 (randomly) 被捕捉。而這只要假設每一件事(包括蝴蝶的分佈, 捕捉的技術等)隨時都很均勻, 不會先是把某一種蝴蝶捕捉殆盡, 之後再大量捕捉另一種。若有某一種蝴蝶尚未被捕捉到, 那純粹只是運氣的關係,而非該種蝴蝶特別會躲藏。費雪所用方法之細節, 由於蠻技術性的, 此處無法細說, 可參考Fisher et al. (1943)。

Efron 與Thisted 介入文學是很偶然的。有一次他們聆聽加州大學 Santa Barbara 分校 (University of California at Santa Barbara) 的Gani (1924-, 著名的應用機率學者, Journal of Applied Probability, Advances in Applied Probability等重要期刊均爲他所創。他曾到過台灣)的演講。Gani的目的是要分析莎士比亞作品的結構。在演講中他提到德國Münster的Westfälische Wilhelms-Universität的Marvin Spevack, 已將莎士比亞的所有作品輸入計算機中,並已計算出莎士比亞所用過的全部字數, 及每一個字使用過的次數, 見Spevack (1968)。聽完演講後, Efron

與Thisted(那時爲Efron 的學生)決定把費雪所用的方法拿來分析莎士比亞的作品。Efron and Thisted (1976)一文就是他們的研究報告, 發表在統計學中極權威的學術期刊Biometrika。

這一篇文章的題目爲: Estimating the number of unseen species: How many words did Shakespeare know? 如前所述, 在生態學中往往會估計某生物尚未見到的種數。而在該文中, 尚未見到的"種", 卻是莎士比亞知道但不曾用過的字。

莎士比亞全部作品(以下簡稱總作品)之總字數爲884,647, 其中有14,376個相異字只出現一次, 4,343 個相異字只出現兩次。表1(見Efron and Thisted (1976))列出至出現100次的字數。令n_x表出現x 次的字數。當$x = 43$, 即列40+, 行3, 其對應的$n_x = 30$,表出現43次的字共有30個, 餘類推。

在總作品中, 莎士比亞共用了31,534個不同的字, 即

$$\sum_{x=1}^{\infty} n_x = 31,534 \text{。}$$

必須注意的是像"girl "與"girls "乃視爲不同的字。由表1, 出現次數不超過100的有30,688個字, 因此有846個字出現的次數超過100 。

x	1	2	3	4	5	6	7	8	9	10	小計
0+	14376	4343	2292	1463	1043	837	638	519	430	364	26305
10+	305	259	242	223	187	181	179	130	127	128	1961
20+	104	105	99	112	93	74	83	76	72	63	881
30+	73	47	56	59	53	45	34	49	45	52	513
40+	49	41	30	35	37	21	41	30	28	19	331
50+	25	19	28	27	31	19	19	22	23	14	227
60+	30	19	21	18	15	10	15	14	11	16	169
70+	13	12	10	16	18	11	8	15	12	7	122
80+	13	12	11	8	10	11	7	12	9	8	101
90+	4	7	6	7	10	10	15	7	7	5	78

表1. 莎士比亞總作品中字出現之頻率

Efron 與Thisted分別依據Fisher et al. (1943)所用的參數模式 (parametric model) 及Good and Toulmin (1956)所用的非參數模式 (non-

parametric model), 來估計莎士比亞尚認識多少字, 所得到的估計值(estimate)都差不多是11,460, 且標準差(standard deviation)小於150 。

事實上在Efron and Thisted (1976)一文中, 他們想回答更一般的問題。他們寫著"假設有另一很大量的莎士比亞的作品被發現, 譬如說共有$884,647t$個字, 則除了那31,534 個不同的字外, 預期可找到幾個新字?"莎士比亞共認識幾個字就對應$t = \infty$ 的情況。由於自十七世紀之後, 便沒有莎士比亞的新作品出現, 所以Efron 與Thisted從未想過會有機會眞正的以莎士比亞的作品來檢驗他們的理論。他們做此研究的動機純粹是覺得有趣。據Efron 說"It never possibly occurred to me that we'd have a chance to use it" 。甚至當他從新聞上獲知這首泰勒詩出現, 且有可能是莎士比亞所作, 他一時並沒想到他其實曾與Thisted做過這方面的研究。直到Thisted提醒他, 他才想到10 年前所做的工作。塵封已久的莎士比亞又浮現在眼前。

這首泰勒詩與總作品相比, 相當短, 共只有429個字。Efron與Thisted基於罕用字(unusual words)出現的頻率 (frequency of occurrence) , 發展出一在統計上算是簡單的檢定法。有一些字是大家常用的, 如"a, is, the", 在每一篇文章中可能都出現不少次。但對罕用字, 每個作者使用的習慣可能便不同了。Efron與Thisted 指出, 在總作品中, 罕用字的使用非常普遍, 在全部使用的31,534個相異字中, 有接近2/3的比例(共21,011), 只用了不超過3次(見表1) 。

這首泰勒詩雖只有429個字, 但其中包含258個相異字。令m_x表在泰勒詩中所使用的字, 在總作品中出現x次的相異字數。表2列出$x = 0$至99的統計。例如, 泰勒詩中有7個相異字在總作品中只出現過1次, 即$m_1 = 7$, 有5個相異字在總作品中出現過23次, 即$m_{23} = 5$, 餘類推。表2中包含118個相異字, 泰勒詩中另有140個相異字在總作品中出現100次以上。其中特別值得注意的是$m_0 = 9$, 此爲泰勒詩中不曾在總作品中出現的相異字數。這就是Efron and Thisted (1976)一文中所要估計的量, 在該文中採用的符號爲$\hat{\Delta}(t)$。

x	1	2	3	4	5	6	7	8	9	10	小計
0+	9	7	5	4	4	2	4	0	2	3	40
10+	1	0	3	0	1	1	1	2	1	0	10
20+	2	2	1	5	3	1	0	2	2	3	21
30+	4	1	1	1	2	1	0	0	3	3	16
40+	1	2	0	0	2	1	1	2	1	1	11
50+	0	1	1	1	1	0	0	1	0	2	7
60+	0	1	0	0	1	1	0	0	1	0	4
70+	0	0	1	0	0	1	0	0	1	1	4
80+	0	0	1	1	0	0	0	0	0	0	2
90+	0	0	0	1	0	1	1	0	0	0	3

表2. 泰勒詩在總作品中出現x次的相異字m_x

Efron 與Thisted 將他們的結果整理爲Thisted and Efron (1987)一文, 題目就叫"Did Shakespeare write a newly-discovered poem?" 表3即爲他們求出的若泰勒詩眞爲莎士比亞所作,表2中m_x之期望值(expectation)之估計值。

x	1	2	3	4	5	6	7	8	9	10
0+	6.97	4.21	3.33	2.84	2.53	2.43	2.16	2.01	1.87	1.76
10+	1.62	1.50	1.52	1.51	1.36	1.38	1.33	1.28	1.25	1.22
20+	1.18	1.16	1.13	1.11	1.09	1.06	1.04	1.02	1.00	0.98
30+	0.96	0.94	0.93	0.91	0.90	0.88	0.86	0.85	0.83	0.82
40+	0.80	0.79	0.77	0.76	0.75	0.74	0.73	0.72	0.70	0.69
50+	0.68	0.67	0.66	0.65	0.64	0.63	0.62	0.61	0.60	0.59
60+	0.58	0.57	0.56	0.55	0.54	0.53	0.52	0.51	0.50	0.50
70+	0.49	0.48	0.48	0.47	0.47	0.46	0.45	0.45	0.44	0.44
80+	0.43	0.42	0.42	0.41	0.41	0.40	0.39	0.39	0.38	0.38
90+	0.37	0.36	0.36	0.35	0.35	0.34	0.34	0.33	0.32	0.32

表3. 表2 中m_x之期望值之估計值$\hat{v_x}$

由表3, 若此泰勒詩確爲莎士比亞所作, 則Efron與Thisted估計其中含有$\hat{\Delta}(t) = 6.97$ 個在總作品中未曾出現的相異字。若再考量標準差, 則此首新詩約包含6.97 ± 2.64個新字, 即新字的數目約介於4.33至9.61間。實際的新字有9個(即表1 中的m_0), 的確在4.33與9.61間。估計曾出現一次的字有4.21 ± 2.05個, 實際則爲7, 估計曾出現兩次的字有3.33 ± 1.83個, 實際則爲5。Efron與Thisted一直分析到曾出現100 次的字, 其吻合程度皆相當

驚人, 可以說通過嚴格的統計檢定 (quite delicate statistical tests) 。看起來這首詩確是莎士比亞所寫, 或者用比較保守的統計術語來說: 沒有有足夠的證據可推翻此詩爲莎士比亞所做之假設 。

傳奇統計學者Persi Diaconis(當時在史丹福大學 任教, 現已轉至哈佛大學統計系 。西元1945年出生於紐約, 14歲時離開學校在街頭變魔術維生, 24歲時到紐約市立學院(City College)的夜間部就讀, 1971年畢業隨即進入哈佛大學統計系, 而於1974年完成博士學位 。DeGroot (1986)爲一篇對他的訪問稿, 戴貞德(1987)爲節譯稿), 他熟悉Efron及Thisted 的分析 。據他講, 他第一次讀這首泰勒詩時, 覺得這一點也不像莎士比亞的作品, 他認爲只要做一些數值分析, 就可以證明詩中字所擺的位置完全是錯的 。但讀了Efron及Thisted的分析後, 他相信這首詩很可能是莎士比亞所作 。

Efron 強調他們的分析並無法證明這首泰勒詩眞是莎士比亞寫的 。但他說這首泰勒詩在罕用字的使用情況, 如此吻合莎士比亞的總作品, 確實令人驚訝 。

Efron及Thisted 也對John Donne (1572?-1631), Christophor Marlowe (1564-1593)及Ben Jonson (1573-1637)等三位約略與莎士比亞同時代的詩人, 各取一首詩, 及另取四首莎士比亞的詩, 與這首泰勒詩做比較 。此八首詩之資料列在表4 。表5是列出8首詩之用字在總作品中之出現頻率 。只是此處計算是分成11類 。例如, 在Jonson的詩中, 有10個相異字在總作品中出現60至79次 。至於Efron及Thisted所做之估計值則列在表6 。

	縮寫	說明	總字數	相異字數
1.	JON	Ben Jonson; 'An Elegy'	411	243
2.	MAR	C. Marlowe: four poems	495	272
3.	DON	J. Donne; 'The Ecstacy'	487	252
4.	CYM	Shakespeare:from 'Cymbeline'	323	215
5.	PUC	from 'A Midsummer Night's Dream'	234	156
6.	PHO	'The Phoenix and Turtle'	352	216
7.	SON	'Sonnets, Nos.12-15	448	264
8.	TAY	Taylor poem	429	258

表4. 8首待分析的詩

在總作品之出現次數

poem	0	1	2	3-4	5-9	10-19	20-29	30-39	40-59	60-79	80-99
1.JON	8	2	1	6	9	9	12	12	13	10	13
2.MAR	10	8	8	16	22	20	13	9	14	9	5
3.DON	17	5	6	5	12	17	14	6	12	3	10
4.CYM	7	4	3	5	13	17	9	12	17	4	4
5.PUC	1	4	0	3	9	6	9	4	5	9	3
6.PHO	14	5	5	9	8	18	13	7	13	8	5
7.SON	7	8	1	5	16	14	12	13	12	13	8
8.TAY	9	7	5	8	11	10	21	16	18	8	5

表5. 8首詩之用字依在總作品中出現的頻率而分類(如表2)

在總作品之出現次數

poem	0	1	2	3-4	5-9	10-19	20-29	30-39	40-59	60-79	80-99
1.JON	6.68	4.03	3.19	5.14	9.81	13.16	9.94	8.18	12.68	9.17	6.83
2.MAR	8.04	4.86	3.85	6.19	11.81	15.91	12.03	9.92	14.92	10.72	8.26
3.DON	7.91	4.78	3.78	6.09	11.62	15.59	11.77	9.68	14.99	10.83	8.06
4.CYM	5.25	3.17	2.51	4.04	7.71	10.35	7.82	6.44	9.99	7.23	5.39
5.PUC	3.79	2.29	1.81	2.91	5.57	7.47	5.65	4.66	7.22	5.23	3.91
6.PHO	5.72	3.46	2.73	4.40	8.40	11.28	8.52	7.02	10.87	7.87	5.87
7.SON	7.28	4.40	3.48	5.60	10.69	14.52	11.10	9.06	13.71	10.02	7.96
8.TAY	6.97	4.21	3.33	5.36	10.24	13.96	10.77	8.87	13.77	9.99	7.48

表6. 表5中之相異字之期望值的估計值

經過三種嚴密的統計檢定(其過程我們自然無法在此介紹性的短文中講述) , 發現對前三首(非莎士比亞之作品), 罕用字出現次數之實際值與預測值(假設其爲莎士比亞所作) 皆不吻合。而雖然挑選的四首莎士比亞的詩偶而有不吻合處, 總的來說是可接受的。本來僅是一師生的遊戲之作, 說不定當初還被"有識之士"視之爲紙上談兵, 十年後竟令那些向來可能不常接觸統計的文學學者折服。不過與其將此視爲無心插柳, 不如相信統計的用途是無所不在的。

關於估計種類的方法尚有不少, 可參考趙蓮菊(1995)一文。

最後我們以Efron及Thisted特地挑出來詠懷他們心情的一首莎士比亞的詩做爲本文之結束。

Why is my verse so barren of new pride,
So far from variation or quick change?
Why with the time do I not glance aside
To new-found methods and to compounds strange?
Why write I still all one, ever the same,
And keep invention in a noted weed,
That every word doth almost tell my name,
Showing their birth and where they did proceed?
O, know, sweet love, I always write of you,
And you and love are still my argument;
So all my best is dressing old words new,
Spending again what is already spent;
    For as the sun is daily new and old,
    So is my love still telling what is told.
                    -- William Shakespeare
                            Sonnet LXXVI

參考文獻

1. 戴貞德(1987). 與統計魔術大師對話錄。科學月刊, 第18卷第5期, 353-355。

2. 趙蓮菊(1995). 種類知多少? 數學傳播季刊, 第19卷第2期, 3-7。

3. DeGroot, M. H. (1988). A conversation with Persi Diaconis.*Statistical Science* 1, 319-334.

4. Efron, B. and Thisted, R. (1976). Estimating the number of unseen species: How many words did shakespeare know?*Biometrika* 63, 435-447.

5. Fisher, R. A., Corbet, A. S. and Williams, C. B. (1943). The relation between the number of species and the number of individuals in a random sample of an animal population.*Journal of Animal Ecology* 12, 42-58.

6. Good, I. J. and Toulmin, G. H. (1956). The number of new species, and the increase in population converage, when a sample is increased.*Biometrika* 43, 45-63.

7. Gower, J. C. (1990/91). Sir Ronald Aylmer Fisher, 1890-1962. *Mathematical Spectrum* 23, 76-86.

8. Kolata, G. (1986). Shakespeare's new poem: an ode to statistics. *Science* 231, 335-336; January 24.

9. Lelyveld, J. (1985). A scholar's find: Shakespearean lyru. *New York Times* (November 24, 1985), 1, 12, with corrections of 'Editor's Note', (November 25, 1985), 2.

10. Spevack, M. (1968). *A Complete and Systematic Concordance to the Words of Shakespeare,* Vols. 1-6. George Olms, Hildesheim.

11. Taylor, G. (1985). Shakespeare's new poem. A scholar's clues and conclusions. *New York Times Book Review* (December 15), 11-14.

12. Thisted, R. and Efron, B. (1987). Did Shakespeare write a newly-discovered poem? *Biometrika* 74, 445-455.