应用概率统计　第二十五卷
第四期　2009年8月

Chinese Journal of Applied Probability
and Statistics Vol.25 No.4 Aug. 2009

# 《红楼梦》前80回与后40回某些文风差异的统计分析
## (两个独立二项总体等价性检验的一个应用)

韦　博　成

(东南大学数学系, 南京, 210096)



**摘　要**

本文以数据分析为基础, 应用统计学中"两总体等价性检验"的理论和方法, 提供了一个强有力的证据: 《红楼梦》前80回与后40回在某些重要的情景描写上确实存在非常显著的差异, 这一结论的可信概率不低于98%.

**关键词**: 等价性检验, 二项分布, 精确条件检验, 渐近正态检验, $p$-值, 《红楼梦》, 情景指标.

**学科分类号**: O212.1.

## §1. 引　　言

在统计学应用的诸多领域中, 文学著作的统计分析是一个饶有兴趣的分支. 美国斯坦福大学教授Efron (1976, 1987)和他的学生曾经对莎士比亚的著作进行过相当深入的统计分析(见[1], [2]), 并指出: 1985年发现的一篇"无名氏"诗稿(仅9节429字)确为莎士比亚所著. Efron是当今国际上最著名的顶级统计学家之一, 他们的工作在当时引起很大反响, 另一位国际顶级统计学家Rao誉之为"一曲统计学的赞歌"(见[3]).

《红楼梦》是我国四大名著之首, 而且有很多悬而未决的问题, 把统计学的定量分析方法引入红学研究是很自然的. 早在1980年, 在美国威斯康星大学召开的"首届国际《红楼梦》研讨会"上, 该校华裔学者陈炳藻教授首次报告了他在这方面的研究工作(见[4], [5]), 此后还出版了专著(见[6]). 陈教授将《红楼梦》120回分为三组, 每组40回, 并将《儿女英雄传》作为对照组进行比较研究. 他从每组中任取8万字, 挑出名词、动词、形容词、副词、虚词这5种词, 然后运用统计学方法算出各组之间用词的相关程度, 结果发现: 《红楼梦》前80回与后40回所用词汇的相关程度远远超过《红楼梦》与《儿女英雄传》所用词汇的相关程度, 并由此推断: 前80回与后40回均为曹雪芹一人所作.

但是, 我国华东师范大学陈大康教授得出了迥异的结论(1987, [7]). 他也把《红楼梦》120回分成三组, 每组40回, 并统计了其中所含词、字、句等88个项目. 他发现, 这些词在前两组出现的规律相同, 而与后40回却不一致; 关于用字特点和句式规律, 前两组也是惊人的吻合, 而后40回则迥异. 由此推断: 后40回非曹雪芹所作(但含有少量残稿).

同时, 复旦大学李贤平教授又提出"成书新说"(1987, [8]). 李教授选择了47个虚字为识别特征, 诸如: "之、其、或、亦、了、的、不、把、别、好"等等, 利用各种统计方法(主成份

本文2008年3月3日收到.

分析、典型相关分析、聚类分析等), 对它们在书中各回的出现频率进行统计分析, 探索各回写作风格的接近程度, 并用三个层次的聚类方法对各回进行分类. 由此提出了成书过程新观点: 《红楼梦》前80回是曹雪芹根据《石头记》增删而成; 而后40回则是曹家亲友搜集整理原稿加工补写而成.

李贤平教授的观点也受到质疑, 例如, 陈大康教授(1988, [9])认为其"成书新说"难以成立. 台湾成功大学王三庆教授也提出若干不同意见(1994, [10]).

以上三位学者都是从《红楼梦》的词语结构出发进行统计分析, 他们按照词语的一定规则进行量化, 得到数据集, 从而应用统计方法推断前80回与后40回之间的差异. 这方面的研究已经比较充分, 本文拟从另一种观点出发研究前80回与后40回之间的差异. 其主要特点是分析《红楼梦》中着力描写的若干情景, 通过量化得到数据集, 然后应用统计方法推断前80回与后40回之间的差异. 具体来说, 我们选择了花荟、树木、饮食、医药与诗词这5个情景指标, 统计出它们在前80回与后40回中出现的频数, 并应用统计学中的"等价性检验"方法来检验二者的差异. 由此得出结论: 《红楼梦》前80回与后40回在饮食和花荟的描写上确实存在非常显著的差异, 其可信概率不低于98%; 同时在树木的描写上也存在明显差异, 其可信概率不低于95%. 这样, 我们就依据统计学原理提供了一个强有力的证据, 说明《红楼梦》前80回与后40回在某些重要的情景描写上确实存在非常显著的差异. 至于导致这些差异的原因, 还涉及到人文和社会方面的诸多因素, 仅用统计学方法可能是无法解释清楚的, 因此本文未有讨论.

本文第2节列举了《红楼梦》中着力描写的5个情景指标, 并经过量化得到相应的数据集; 第3节对这5个数据集进行等价性检验, 计算出相应的$p$-值, 并指出《红楼梦》前80回与后40回在某些文风上所存在的显著性差异, 从而得到本文的主要结果; 第4节结束语对本文的研究作若干注记; 附录简要说明了数据集的生成过程.

## §2. 情景指标的数据集

据2007年10月10日南京"现代快报"报道(见[11]), 南京林业大学汤庚国教授另辟蹊径, 从海棠文化出发, 分析《红楼梦》前80回与后40回的差异. 汤教授主要是从人文花荟方面进行分析, 但是他们也提供了一组数据, 即《红楼梦》前80回有16回涉及海棠, 而后40回仅有4回涉及海棠, 以此说明前后差距明显. 对于汤教授提供的这组数据, 统计学者还是能够有所作为的. 事实上, 我们可以对此做一个等价性假设检验:

原假设$H_0$: "前80回与后40回对于海棠花的关注程度相同";

对立假设$H_1$: "前80回对于海棠花的关注程度大于后40回对于海棠花的关注程度".

经渐近正态公式计算, 有将近92%的"把握"认为"前80回对于海棠花的关注程度大于后40回对于海棠花的关注程度".

受此启发, 本文进一步推广和发展了这一数据分析方法. 我们对《红楼梦》中若干重要的情景描写进行量化, 得到相应的数据集. 有了数据集即可通过数理统计方法, 比较前80回与后40回在文风上的差异. 事实上, 在《红楼梦》中, 对于许多情景都有非常深入的刻画

和描写, 例如饮食菜肴, 全书有40多回涉及到饮食文化的许多方面(其中最著名的是41回关于"茄鲞"的描写; 75回还提到"风腌果子狸"). 我们不考虑人文社会方面的问题, 而致力于数据的收集与分析, 并以此为基础, 应用数理统计方法来研究其前80回与后40回在文风上的差异. 根据我们的统计, 《红楼梦》在前80回中有34回涉及饮食方面的描写; 后40回仅有8回涉及饮食方面的描写(见表1). 根据这一数据, 我们可考虑以下等价性假设检验问题(记为等价性检验(A)):

原假设$H_0$: "前80回与后40回对于'饮食描写'的关注程度相同";

对立假设$H_1$: "前80回对于'饮食描写'的关注程度大于后40回对于'饮食描写'的关注程度".

类似地, 我们亦可选择其他情景指标, 设法得到相应的数据, 并考虑类似的假设检验问题. 本文选择了《红楼梦》中着力描写的5个情景指标, 即花荟、树木、饮食、医药与诗词, 统计出它们在前80回与后40回中出现的频数. 如表1所示(数据集的具体收集过程见附录). 对于上述每一个情景指标, 我们都可以考虑类似的等价性检验, 以便比较前80回与后40回对它们在关注程度方面的差异.

表1　前80回与后40回各情景指标出现的频数

| | 1–40回 | 41–80回 | 1–80回 | 81–120回 |
|---|---|---|---|---|
| 花荟 | 15 | 16 | 31 | 7 |
| 树木 | 13 | 14 | 27 | 7 |
| 饮食 | 17 | 17 | 34 | 8 |
| 医药 | 13 | 13 | 26 | 8 |
| 诗词 | 22 | 14 | 36 | 12 |

## §3. 统计分析——等价性检验

有了数据表1, 《红楼梦》前80回与后40回在文风上的差异分析就可以化为数理统计学的问题. 今以等价性检验(A) (即关于饮食的描写)为例说明其统计模型及其求解方法. 这一检验问题可化为两个相互独立的二项总体的等价性检验, 这时

(1) $X_1 \sim b(n_1, p_1)$表示前80回的二项分布. 其中$n_1 = 80$, $X_1$表示前80回中涉及饮食的回数; 其实测值为$x_1 = 34$; $p_1$表示前80回中每回涉及饮食的概率.

(2) $X_2 \sim b(n_2, p_2)$表示后40回的二项分布. 其中$n_2 = 40$, $X_2$表示后40回中涉及饮食的回数; 其实测值为$x_2 = 8$; $p_2$表示后40回中每回涉及饮食的概率.

等价性检验问题为:

$$H_0 : p_1 = p_2 \longleftrightarrow H_1 : p_1 > p_2.$$

否定原假设就意味着"前80回对于'饮食描写'的关注程度大于后40回对于'饮食描写'的关注程度" (以一定的检验水平). 对于这个假设检验问题, 不少著作都有论述, 例如可参

见韦博成(2006, [12], p.267–270). 通常有两种检验方法, 即Fisher精确条件检验(Lehmann, 1986, [13], p.154, 或韦博成, 2006, [12], p.269)和渐近正态检验(韦博成, 2006, [12], p.270, 或何书元, 2006, [14], p.263). 我们用这两种方法都进行了计算, 得到检验的$p$-值, 即否定原假设而犯错误的概率. 后者比较简单, 其检验统计量为

$$U=\frac{\widehat{p}_1-\widehat{p}_2}{\sqrt{\widehat{p}\widehat{q}(n_1^{-1}+n_2^{-1})}}\sim N(0,1).$$

其中

$$\widehat{p}_1=\frac{X_1}{n_1},\qquad \widehat{p}_2=\frac{X_2}{n_2},\qquad \widehat{p}=\frac{X_1+X_2}{n_1+n_2},\qquad \widehat{q}=1-\widehat{p}.$$

表2给出了各个情景指标的检验结果.

表2 两种方法对于各个情景指标的检验结果
(前80回与后40回的比较)

| | U检验值 | $p$-值 | 可信概率 % | 精确条件检验的$p$-值 | 可信概率 % |
|---|---|---|---|---|---|
| 饮食 | 2.4360 | 0.0074 | 99.26% | 0.0114 | 98.86% |
| 花荟 | 2.3590 | 0.0092 | 99.08% | 0.0140 | 98.60% |
| 树木 | 1.8622 | 0.0313 | 96.87% | 0.0473 | 95.27% |
| 诗词 | 1.5811 | 0.0569 | 94.31% | 0.0824 | 91.76% |
| 医药 | 1.4325 | 0.0760 | 92.40% | 0.1105 | 88.95% |

为了得到Fisher精确条件检验, 可应用韦博成(2006, [12], p.268–269)的有关公式. 检验的$p$-值可表示为

$$\mathsf{P}_{H_0}(x_1\le X_1\le\min\{t,n_1\}|X_1+X_2=t)=\sum_{x_1\le u\le\min\{t,n_1\}}\mathsf{P}_{H_0}(X_1=u|X_1+X_2=t).$$

其中

$$\mathsf{P}_{H_0}(X_1=u|X_1+X_2=t)=\frac{\binom{n_1}{u}\binom{n_2}{t-u}}{\binom{n_1+n_2}{t}},\qquad \max\{t-n_2,0\}\le u\le\min\{t,n_1\}.$$

其中$\mathsf{P}_{H_0}$表示$H_0$成立时的概率. 相应的计算结果也列在表2中. 这些结果表明, Fisher精确条件检验的$p$-值比渐近正态检验的$p$-值稍为偏大一点, 这显然是合理的. 但是二者相差很少, 而且大小次序相同, 说明两种方法的计算结果有很好的一致性. 精确条件检验是UMPUT, 即一致最优的无偏检验, 但是也比较保守, 它要求对一切满足$p_2\in(p_1,1]$的$p_2$都有最优的功效, 这往往是不必要的.

这些结果都很清楚明确. 由表2可以看出, 饮食与花荟的显著性最高, 即我们有充分的理由认为, 前80回与后40回在饮食与花荟的描写上有很显著的差异, 即使按最保守的

Fisher精确条件检验的标准来进行统计推断, 其判错的概率(即$p$-值)也不到0.02, 因而判对的概率超过98%. 对于树木数据, 其检验的$p$-值也小于通常的0.05, 因此我们也有比较充分的理由认为, 前80回与后40回在树木的描写上有很显著的差异, 其判错的概率不到0.05, 因而判对的概率超过95%. 至于医药和诗词这两个指标, 可作为比较对照之用. 如果按渐近正态检验的结果来看, 我们还是有超过92%以上的概率认为, 前80回与后40回在医药和诗词的描写上有差异. 但是, 若按比较保守的Fisher精确条件检验的标准来判断, 则没有充分理由认为前80回与后40回在医药和诗词的描写上有显著性差异. 不过, 这对本文关于饮食、花荟和树木数据的主要结果并无影响. 事实上, 前80回与后40回只要在一个指标上有非常显著的差异, 则说明二者在文风上确有差异(如果前80回与后40回在2, 3个指标上都有显著性差异, 则我们结论的可信概率只会大大增加).

同时, 我们也对前80回的第二个40回与后40回进行了比较, 其检验结果如表3所示. 其相应的$p$-值依次略有减小, 但是花荟与饮食的显著性仍然很高. 即我们有充分的理由认为, 前80回的第二个40回与紧挨着的后40回在花荟与饮食的描写上有很显著的差异, 按照比较保守的Fisher精确条件检验的标准来推断, 其判错的概率不到0.03, 因而判对的概率超过97%. 同时也有超过93%以上的概率认为, 前80回的第二个40回与紧挨着的后40回在树木的描写上有显著差异. 而对于医药和诗词这两个指标, 我们没有充分理由认为有显著性差异. 另外, 由表1可以明显的看出, 对于花荟、树木、饮食和医药这4组数据, 前80回的第一个40回与第二个40回几乎没有区别, 两者高度一致. 我们也可以对它们进行等价性检验, 得到的$p$-值都非常大(计算从略).

表3　两种方法对于各个情景指标的检验结果
(前80回的第二个40回与后40回的比较)

| | U检验值 | $p$-值 | 可信概率 % | 精确条件检验的$p$-值 | 可信概率 % |
|---|---|---|---|---|---|
| 花荟 | 2.2232 | 0.0131 | 98.69% | 0.0234 | 97.66% |
| 饮食 | 2.1709 | 0.0150 | 98.50% | 0.0263 | 97.37% |
| 树木 | 1.7787 | 0.0376 | 96.24% | 0.0631 | 93.69% |
| 医药 | 1.2705 | 0.1020 | 89.80% | 0.1548 | 84.52% |
| 诗词 | 无显著差异 | – | – | – | – |

综上所述, 本文以数据分析为基础, 以统计学中“两个独立二项总体的等价性检验”为基本方法, 很清楚明确地证明: 《红楼梦》前80回与后40回在饮食与花荟的描写上确实存在非常显著的差异; 在树木的描写上也存在明显差异. 不过, 这种差异还不能说明《红楼梦》前80回与后40回出自不同的作者, 因为统计学方法并不能分析导致这种差异的原因, 这还涉及到许多人文与社会方面的问题(例如, 书中情节发展的变化也可能会导致情景描写上的若干变化). 但是, 本文毕竟提供一个强有力的证据, 说明《红楼梦》前80回与后40回在某些文风上确实存在非常显著的差异, 供有兴趣者参考.

另外, 从统计学观点来看, 本文表1对应于一个“具有5对二项总体”的模型, 要研究其“比较与检验问题”. 二项总体的比较一直是生物统计中的一个重要问题(例如可参见Tang,

et, al. 2003, [15], Schouten, 2007, [16]), 其解法也很丰富多彩. 除了本文的经典方法外, 唐年胜教授还建议采用“中位$p$-值法”和“近似非条件$p$-值法” (因为Fisher精确条件检验比较保守), 其结果与本文表2的U检验完全一致. 特别, 旅英学者鲁国斌教授应用Bayes方法, 基于固定效应模型, 对表1进行了深入的分析. 他应用MCMC方法和WinBUGS软件, 计算了假设$H_1$成立(即前80回与后40回有显著差异)的后验概率(相当于$1-p$值), 其结果也与本文表2完全一致.

## §4. 结 束 语

本文以情景描写为基础, 应用统计学方法研究了《红楼梦》前80回与后40回在若干情景描写上的差异. 事实上, 作为一部名著,《红楼梦》书中的情景描写极为丰富多彩、变化多端, 我们也可选择其他情景指标, 比较其前80回与后40回之间的差异. 例如,《红楼梦》中关于哭泣的描写有260处; 关于笑的描写有173处; 关于梦境的描写有32处; 还有死亡人物近50人(见[17]); 等等. 我们也许可以应用统计学方法比较前80回与后40回关于这些情节的描写是否存在差异. 最近, 安鸿志教授在他的新作(见[18])中指出,《红楼梦》前80回有多处对皇权不恭的描写, 而后40回则几乎没有, 该书也对此进行了相应的计算与分析, 说明前80回与后40回在对待皇权的态度上有显著性差异.

另外, 本文提出的方法(即等价性检验)不仅可用于情景描写方面的比较, 也可以用于用词造句等其他方面的比较, 以下关于歇后语的比较就是一个有趣的例子. 在《红楼梦》书中, 引用了许多诙谐的民间歇后语. 例如: “狗咬吕洞宾—不识好人心(第25回)”; “锯了嘴子的葫芦—没口齿(第65回)”; “羊群里跑出骆驼来了—就只你大(第88回)”; 等等. 但是, 前80回与后40回比较, 前者引用歇后语的频数要高得多. 据统计(见[19]), 前80回有15回引用了歇后语; 而后40回只有2回引用了歇后语. 与前面的讨论类似, 我们亦可做一个等价性检验:

原假设$H_0$: “前80回与后40回对于歇后语的引用频数相同”;

对立假设$H_1$: “前80回对于歇后语的引用频数大于后40回对于歇后语的引用频数”.

经渐近正态公式计算(见第3节), 检验的$p$-值$\approx 0.02$. 因此, 我们有将近98%的概率认为, 前80回与后40回在歇后语的引用上有显著性差异.

由此可见, 本文提出的方法的应用范围还是比较广泛的.

## 附录: 数据集的生成过程

本文在收集各个情景指标的数据并产生表1时, 所用的《红楼梦》书稿主要来自网址为http://www.ebook007.com的北极星书库(见[20]), 该书库原来的书稿文件名为“红楼梦.chm”. 为搜索方便, 我们把它分为12个word文件, 10回为1个文件. 对于《红楼梦》书稿中有关描写花荟、树木、饮食、医药与诗词方面的内容, 采用人工查阅与关键词搜索相结合的方法, 最后列表给出每一回涉及花荟、树木、饮食、医药与诗词的具体情况以及某些附注(详见http://math.seu.edu.cn). 为了检查校对, 我们也参考了“脂胭斋重评石头记”

(80回)以及凤凰出版社(原江苏古籍出版社) 2001年出版的《红楼梦》(以程乙本为底本, 经重新校勘标点订正而成, 见[21]).

另外, 关于花荟的数据比较特别, 需要作一些必要的说明. 本文的数据主要是统计作者对花荟的情景描写, 而书中不少涉及花荟的地方与情景无关, 例如: 梅花糕、桃花庙、蔷薇硝、茉莉粉、海棠红的小棉袄、梅花式洋漆小几等等. 因此, 这些类似的情形都未计入花荟的数据集, 以下为部分例证(未全部列出):

19回: “向荷包内取出两个梅花香饼儿来” (宝玉);

34回: “自羡压倒桃花, 却不知病由此萌” (黛玉);

41回: “拣了一朵牡丹花样的小面果” (刘姥姥);

66回: “揉碎桃花红满地, 玉山倾倒再难扶” (尤三姐之死);

85回: “前年他送我白海棠时称我作‘父亲大人’” (宝玉);

91回: “正露着石榴红洒花夹裤” (宝蟾);

**致谢**　本文的写作受到南京林业大学汤庚国教授工作很大的启发, 特此对汤教授表示衷心的感谢! 云南大学唐年胜教授帮助计算了Fisher精确条件检验, 表示衷心的感谢! 另外, 参考文献中的有关作者提供了很多信息, 在此一并表示衷心的感谢!

## 参 考 文 献

[1] Efron, B. and Thisted, R., Estimating the number of unseen species: How many words did Shakespeare know? *Biometrika*, **63**(1976), 435–437.

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

[3] Rao, C.R., *Statistics and Truth — How to Use Chance* (中译本: 统计与真理 — 怎样运用偶然性, 李竹喻等译, 科学出版社, 2004, 北京).

[4] 陈炳藻, 从词汇上的统计论《红楼梦》的作者问题, “首届国际《红楼梦》研讨会” (1980, 美国威斯康星大学) (见[5]).

[5] 贾洪卫, 董坚, 徐锐, 计算机与“红学”研究综论 (2003, 可参见 http://www.T1soft.com 中国人民大学统计数据库研究室).

[6] 陈炳藻, 电脑红学: 论《红楼梦》作者 (1986, 见[5]).

[7] 陈大康, 从数理语言学看后四十回的作者, 红楼梦学刊, **1**(1987), 293–318.

[8] 李贤平, 《红楼梦》成书新说, 复旦大学学报社科版, **5**(1987), 3–16.

[9] 陈大康, 《红楼梦》“成书新说”难以成立, 华东师大学报, **1**(1988), 3–13.

[10] 王三庆, 红楼梦电脑 — 《红楼梦》研究与电脑科技, 1994 台北甲戌年红学会议 (可参见 http://www.readred.com 夜看红楼: 和后40回有关的一些资料).

[11] 现代快报, 南林大专家研究《红楼梦》另辟蹊径, 2007年10月10日, 南京.

[12] 韦博成, 参数统计教程, 高等教育出版社, 北京, 2006.

[13] Lehmann, E.L., *Testing Statistical Hypotheses*, Wiley, New York, 1986.

[14] 何书元, 概率论与数理统计, 高等教育出版社, 北京, 2006.

[15] Tang, N.S., Tang, M.L. and Chan, I.S.F., On tests of equivalence via non-unity relative risk for matched-pair design, *Statistics in Medicine*, **22**(2003), 1217–1233.

[16] Schouten, H.J.A.,Comparing two independent binomial proportions by a modified chi square test-Required sample sizes, *Biometrical Journal*, **24**(2007), 93–96.

[17] 红楼梦有多少个梦? 搜搜问问, http://wenwen.soso.com/z/q1124650.

[18] 安鸿志, 趣话概率—兼话《红楼梦》中的玄机, 科学出版社, 北京, 2009.

[19] 红楼梦—百度百科, http://baike.baidu.com/view/2571.

[20] 红楼梦.chm, http://www.ebook007.com, 北极星书库.

[21] 红楼梦, 凤凰出版社(原江苏古籍出版社), 2001年出版, 2004年第6次印刷, 南京.



# Statistical Analysis on the Differences of Writing Style Between First 80 Chapters and Last 40 Chapters in 《Dream of Red Mansions》 (An Application of Equivalent Test on Two Independent Binomial Populations)

Wei Bocheng

(*Department of Mathematics, Southeast University, Nanjing, 210096*)

Based on data analysis, using the theory and method of "equivalent test on two independent binomial populations", this paper provides a strong evidence that there are significant differences on the writing of sights between first 80 chapters and last 40 chapters in 《Dream of Red Mansions》. The confidence probability of this conclusion is greater than 98%.

**Keywords:** Equivalent test, binomial distribution, exact conditional test, asymptotic normal test, $p$-value, 《Dream of Red Mansions》, sight index.

**AMS Subject Classification:** 62J25.