中國統計學報
第41卷第1期
九十二年三月
97-116頁

資料探勘-超大型資料庫基本統計量的計算

馬瀰嘉[1]　蘇佩芳[1]　林共進[2]

[1]國立成功大學統計學研究所

[2]美國賓州州立大學統計學系

摘　要

針對一組與時俱增的龐大資料，日後若欲計算其基本統計量，如:平均數、變異數、中位數、百分位數或相關係數等，就必需一直儲存這些資料，但面對如此龐大的資料，在計算中位數或百分位數等需先將資料排序的統計量時，由於計算機能力的限制，如資料超過允許最大的陣列數，或記憶體不足的問題，造成電腦不能計算。因此如何找出一種逐步遞迴的方法:每隔一段時間後，丟棄原始的資料，僅保留幾個具代表性的統計量，使得日後欲計算先前所有資料的基本統計量時，可由計算這些各區段統計量獲得。本文提出二階段組合中心趨勢的統計量，如:平均數、中位數、值域中點和利用 *Han & Kamber (2000)* 書中近似中位數公式來求得前k段資料的近似中位數和近似百分位數，並利用統計模擬來比較不同近似方法的好壞。

關鍵詞：中位數，平均數，百分位數，相關係數，變異數。

美國數學會分類索引：主要62F10；次要65C05。

1. 前言

近來因爲商業環境改變，顧客消費行爲複雜化等原因而使得資訊爆增，故近幾年有關資料探勘 (data mining) 領域是相當熱門的研究課題。其實資料探勘所使用的分析方法，例如:連結分析 (link analysis)、預測模型 (predictive modeling) 等部份爲已有之統計分析方法。隨著電腦的普及化和網路的便利，各行業每年累積的資料可能數以'億'計，數年後若欲計算這些超級龐大資料的基本統計量，如:平均數 (mean)、變異數 (variance)、中位數 (median)、百分位數 (percentile) 或相關係數 (correlation coefficient) 等，就必須一直儲存這些資料，但面對如此龐大的資料，在計算中位數或百分位數等需先將資料排序的統計量時，即使借助計算機快速的運算能力，也會相當耗時甚至不可能計算。因此本文主要目的在如何找出一種逐步遞迴的方法使得每隔一段時間後，丟棄原始的資料，僅保留幾個各區段具代表性的統計量，以便日後欲計算先前所有資料的基本統計量時，可由計算這些各區段統計量來獲得。

首先探討平均數、變異數和共變異數 (covariance)，假設有 I 段資料，第 i 段資料中兩變項 X 和 Y 的資料分別設爲 $X_{i1}, \ldots, X_{in_i}$ 和 $Y_{i1}, \ldots, Y_{in_i}$, $i = 1, \ldots, I$。基本統計量中，前 $k+1$ 段資料的平均數和前 k 段資料的平均數之關係爲

$$\bar{X}_{(k+1)} = \frac{1}{N_{k+1}}(N_k \bar{X}_{(k)} + n_{k+1}\bar{X}_{k+1}), \tag{1.1}$$

其中 $N_k = \sum_{i=1}^{k} n_i$, $k = 1, \ldots, I$; $\bar{X}_{(k)} = \frac{1}{N_k}\sum_{i=1}^{k}\sum_{j=1}^{n_i} X_{ij}$ 爲前 k 段資料的平均數，$\bar{X}_{k+1} = \frac{1}{n_{k+1}}\sum_{j=1}^{n_{k+1}} X_{k+1,j}$ 爲第 $(k+1)$ 段資料的平均數。

前 $k+1$ 段資料的變異數和前 k 段資料的變異數之關係爲

$$S_{xx}^{(k+1)} = \frac{1}{N_{k+1}-1}[(N_k-1)S_{xx}^{(k)} + \sum_{j=1}^{n_{k+1}} X_{k+1,j}^2 + N_k\bar{X}_{(k)}^2 - N_{k+1}\bar{X}_{(k+1)}^2], \tag{1.2}$$

其中 $S_{xx}^{(k)} = \frac{1}{N_k-1}\sum_{i=1}^{k}\sum_{j=1}^{n_i}[X_{ij} - \bar{X}_{(k)}]^2$ 爲前 k 段資料的變異數。

前 $k+1$ 段資料的共變異數和前 k 段資料的共變異數之關係爲

$$S_{xy}^{(k+1)} = \frac{1}{N_{k+1}-1}[(N_k-1)S_{xy}^{(k)}+$$

$$\sum_{j=1}^{n_{k+1}} X_{k+1,j}Y_{k+1,j} + N_k\bar{X}_{(k)}\bar{Y}_{(k)} - N_{k+1}\bar{X}_{(k+1)}\bar{Y}_{(k+1)}]\,, \qquad (1.3)$$

其中 $S_{xy}^{(k)} = \frac{1}{N_k-1}\sum_{i=1}^{k}\sum_{j=1}^{n_i}[X_{ij}-\bar{X}_{(k)}][Y_{ij}-\bar{Y}_{(k)}]$ 爲前 k 段資料的共變異數。

由於上述基本統計量有逐步遞迴的關係式，故在計算超級龐大資料的平均數(或變異數、共變異數)時，只需儲存各區段資料的樣本數、平均數(變異數、共變異數)，利用逐步遞迴的關係式即可獲得。因爲前 $k+1$ 段資料的中位數和前 k 段資料的中位數間不存在逐步遞迴的關係式，所有資料的中位數和各區段資料的中位數之間又沒有具體的公式，故本文第 2 節主要探討在超大型資料庫下，如何利用各區段儲存的統計量資料，來求得前 k 段資料中位數的近似值。第 3 節擴展中位數提出百分位數的近似方法，第 4 節利用統計模擬來比較不同近似方法的好壞。

2. 中位數

假設將第 i 段資料 $X_{i1}, \ldots, X_{in_i}$ 由小至大排序後令爲 $X_{i(1)} \le X_{i(2)} \le \cdots \le X_{i(n_i)}$，其中位數記作

$$m_i = \begin{cases} X_{i(c_i+1)} & \text{如果 } n_i = 2c_i+1 \\ (X_{i(c_i)} + X_{i(c_i+1)})/2 & \text{如果 } n_i = 2c_i \end{cases}, \qquad (2.1)$$

其值域中點 (midrange) 記作

$$r_i = (X_{i(1)} + X_{i(n_i)})/2\,, \qquad (2.2)$$

假設前 k 段資料 $\{X_{i1}, \ldots, X_{in_i};\ i = 1, \ldots, k\}$ 由小至大排序後令爲 $T_{k1} \le T_{k2} \le \ldots \le T_{k,N_k}$，其中位數爲

$$m_{(k)} = \begin{cases} T_{k,\,c+1} & \text{如果 } N_k = 2c+1 \\ (T_{k,\,c} + T_{k,\,c+1})/2 & \text{如果 } N_k = 2c \end{cases} \quad ,k = 1,\ \ldots,\ I\,, \qquad (2.3)$$

由於在計算前 k 段資料的中位數時，資料須重新排序，然而在第 k 段資料取得時，前 $k-1$ 段資料已被丟棄，僅保留各區段具代表性的統計量，因此中位數不能由 (2.3) 式獲得。

因爲中位數主要用以決定資料的"中心點"，故考慮常用以表示中心趨勢的統計量，如:平均數，值域中點，利用下列二階法 (two stages method) 來求得前 k 段資料的近似中位數:

(1) 先求得各區段中位數 m_i，再將 $\{m_i,\ i = 1,\ \ldots,\ k\}$ 排序求其中位數，記作 $\hat{m}_{(k)}$。

(2) 先求得各區段的中位數 m_i，再將 $\{m_i,\ i = 1,\ \ldots,\ k\}$ 求其平均數，記作 $\bar{m}_{(k)}$。

(3) 先求得各區段的中位數 m_i，再將 $\{m_i,\ i = 1,\ \ldots,\ k\}$ 求其值域中點，記作 $\tilde{m}_{(k)}$。

例如:$I = 3,\ n_1 = n_2 = n_3 = 3$，前 3 段資料排序後爲 $\{1,\ 2,\ \ldots,\ 9\}$，眞正中位數爲 5，當各區段資料取得的順序不同時，不同方法求得之近似中位數如表 1。

表1 不同方法下之近似中位數

	第1段資料	第2段資料	第3段資料	$\hat{m}_{(k)}$	$\bar{m}_{(k)}$	$\tilde{m}_{(k)}$
1	1 3 5	4 7 8	2 6 9	6	5.33	5
2	1 2 3	4 5 6	7 8 9	5	5	5
3	1 2 6	3 4 8	5 7 9	4	4.33	4.5
4	1 2 7	3 4 8	5 6 9	4	4	4
5	1 4 7	2 5 8	3 6 9	5	5	5
6	1 8 9	2 6 7	3 4 5	6	6	6

前 k 段資料的近似中位數 $\hat{m}_{(k)}$ 有下列性質:

定理一:

利用各區段中位數排序，再求其中位數的方法所求得之近似中位數 $\hat{m}_{(k)}$ ，會介於資料排序後第 C_k+1 個資料和第 N_k-C_k 個資料之間;即 $T_{k,C_k+1} \le \hat{m}_{(k)} \le T_{k,N_k-C_k}$ ，其中 $C_k=\sum_{i=1}^{[k/2]} c_i$ ，$[x]$ 爲不大於 x 的最大整數（證明參見附錄一）。

另外，Han & Kamber (2000) 在其" DATA MINING "書中提及當資料很龐大時，可利用資料的直方圖來求近似中位數，假設資料可能的最小值爲 a_0 ，最大值爲 a_L ，將資料分爲 L 組，固定第 j 組組界爲 $(a_{j-1}, a_j]$ ，令 n_{ij} 爲第 i 段資料 $X_{i1}, \ldots, X_{in_i}$ 落在第 j 組組界的樣本個數，則第 i 段資料的近似中位數爲

$$\tilde{\hat{m}}_i = a_{w_i-1} + \frac{b\left(n_i/2 - \sum_{j=1}^{w_i-1} n_{ij}\right)}{n_{iw_i}}, \tag{2.4}$$

其中 w_i 爲使得第 i 段資料的中位數落在 $(a_{w_i-1}, a_{w_i}]$ 的組別，(即 w_i 使得 $\sum_{j=1}^{w_i-1} n_{ij} < 1/2$ 和 $\sum_{j=1}^{w_i} n_{ij} > 1/2$)，$b=a_j-a_{j-1}$ 爲組距。若在第 i 段保留 $(n_{i1}, \ldots, n_{iL})$, $i=1, \cdots, k$ ，則前 k 段資料落在第 j 組組界的樣本個數爲 $n_{(k)j}=\sum_{i=1}^{k} n_{ij}$ ，$j=1, \cdots, L$ ，因此前 k 段資料的近似中位數爲

$$\tilde{\hat{m}}_{(k)} = a_{W_k-1} + \frac{b\left(N_k/2 - \sum_{j=1}^{W_k-1} n_{(k)j}\right)}{n_{(k)W_k}}, \tag{2.5}$$

其中 W_k 爲使得前 k 段資料的中位數落在 (a_{W_k-1}, a_{W_k}) 的組別。

3. 百分位數

本節擴展中位數方法至百分位數上，由統計上得知，若能求出一組資料的若干個百分位數，即可畫出該資料的經驗分配函數 (empirical distribution function)，進而估計該資料的分佈情形 (即 density estimation)。假設將第 i

段資料 $X_{i1}, \ldots, X_{in_i}$ 由小至大排序後，令爲 $X_{i(1)} \le X_{i(2)} \le \ldots \le X_{i(n_i)}$，其 $100p$ 百分位數 (the $100p$th percentile) 記作

$$q_{pi} = \begin{cases} X_{i(d_i+1)} & \text{如果 } g_i > 0 \\ (X_{i(d_i)} + X_{i(d_i+1)})/2 & \text{如果 } g_i = 0 \end{cases}, \tag{3.1}$$

其中 $n_i p = d_i + g_i$，d_i 爲整數部份，g_i 爲小數部份。假設前 k 段資料 $\{X_{i1}, \ldots, X_{in_i};\ i = 1, \ldots, k\}$ 由小至大排序後，令爲 $T_{k1} \le T_{k2} \le \ldots \le T_{k, N_k}$，其 $100p$ 百分位數爲

$$q_{p(k)} = \begin{cases} T_{k, d+1} & \text{如果 } g > 0 \\ (T_{k, d} + T_{k, d+1})/2 & \text{如果 } g = 0 \end{cases}, k = 1, \ldots, I, \tag{3.2}$$

其中 $N_k p = d + g$，d 爲整數部份，g 爲小數部份。由於在計算前 k 段資料的 $100p$ 百分位數時，資料須重新排序，然而在第 k 段資料取得時，前 $k-1$ 段資料已被丟棄，僅保留各區段具代表性的統計量，因此 $100p$ 百分位數不能由 (3.2) 式獲得。

因爲 $q_{p(k+1)}$ 和 $q_{p(k)}$ 間不存在逐步遞迴的關係式，所有資料的百分位數和各區段資料的百分位數之間又沒有具體的公式，故類似第 2 節利用二階段法來求得前 k 段資料的近似 $100p$ 百分位數:

(1) 先求得各區段的第 $100p$ 百分位數 q_{pi}，再將 $\{q_{pi},\ i = 1, \ldots, k\}$ 排序求其中位數，記作 $\hat{q}_{p(k)}$。

(2) 先求得各區段的百分位數 q_{pi}，再由 $\{q_{pi},\ i = 1, \ldots, k\}$ 求平均數，記作 $\bar{q}_{p(k)}$。

(3) 先求得各區段的百分位數 q_{pi}，再由 $\{q_{pi},\ i = 1, \ldots, k\}$ 求值域中點，記作 $\tilde{q}_{p(k)}$。

例如: $I = 3$, $n_1 = n_2 = n_3 = 3$，前 3 段資料排序後爲 $\{1, 2, \ldots, 9\}$，若 $p = 0.25$，則第一四分位數 $\hat{q}_{p(I)}$ (記爲 Q_1) 爲 3; 若 $p = 0.75$，則第三四分位數 $\hat{q}_{p(I)}$ (記爲 Q_3) 爲 7，當各區段資料取得的順序不同時，不同方法求得之近似第一四分位數如表 2，近似第三四分位數如表 3。

表2 不同方法下之近似第一四分位數 Q_1

no.	第1段資料	第2段資料	第3段資料	$\hat{Q}_1$	$\bar{Q}_1$	$\tilde{Q}_1$
1	1 3 5	4 7 8	2 6 9	2	2.33	2.5
2	1 2 3	4 5 6	7 8 9	4	4	4
3	1 2 6	3 4 8	5 7 9	3	3	3
4	1 2 7	3 4 8	5 6 9	3	3	3
5	1 4 7	2 5 8	3 6 9	2	2	2
6	1 8 9	2 6 7	3 4 5	2	2	2

表3 不同方法下之近似第三四分位數 Q_3

no.	第1段資料	第2段資料	第3段資料	$\hat{Q}_3$	$\bar{Q}_3$	$\tilde{Q}_3$
1	1 3 5	4 7 8	2 6 9	8	7.33	7
2	1 2 3	4 5 6	7 8 9	6	6	6
3	1 2 6	3 4 8	5 7 9	8	7.66	7.5
4	1 2 7	3 4 8	5 6 9	8	8	8
5	1 4 7	2 5 8	3 6 9	8	8	8
6	1 8 9	2 6 7	3 4 5	7	7	7

定理二：

利用各區段之 $100p$ 百分位數，再排序求其中位數的方法所求得之近似 $100p$ 百分位數，介於排序後第 L_k+1 個資料和第 N_k-U_k 個資料之間；即 $V_{k,\,L_k+1} \le \hat{q}_{p(k)} \le V_{k,\,N_k-U_k}$，其中 $L_k=\sum_{i=1}^{[k/2]} d_i$，$U_k=\sum_{i=1}^{[k/2]}(n_i-d_i-1)$（證明參見附錄一）。

類似第 3 節我們可利用資料的直方圖來求近似 $100p$ 百分位數爲

$$\tilde{\tilde{q}}_{pi}=a_{v_i-1}+\frac{b(pn_i-\sum_{j=1}^{v_i-1}n_{ij})}{n_{iv_i}}, \tag{3.3}$$

其中 v_i 爲使得第 i 段資料的 $100p$ 百分位數落在 $(a_{v_i-1},\ a_{v_i})$ 的組別（即 v_i 使

得 $\sum_{j=1}^{v_i-1} n_{ij} < p$ 和 $\sum_{j=1}^{v_i} n_{ij} > p$)，$b = a_j - a_{j-1}$ 爲組距。若前 k 段資料落在第 j 組組界的樣本個數爲 $n_{(k)j} = \sum_{i=1}^{k} n_{ij}$，則前 k 段資料的近似 $100p$ 百分位數爲

$$\tilde{\tilde{q}}_{p(k)} = a_{V_k-1} + \frac{b\left(pN_k - \sum_{j=1}^{V_k-1} n_{(k)j}\right)}{n_{(k)V_k}}, \tag{3.4}$$

其中 V_k 爲使得前 k 段資料的中位數落在 (a_{V_k-1}, a_{V_k}) 的組別。

引理：(參見 Serfling, 1980 書 2.3.3 節)

若 $X_{i1}, \cdots, X_{in_i}$ 爲獨立且具相同的分配 F，$\xi_p = \inf\{x : F(x) \geq p\}$ 爲分配 F 的 p 百分位數。

(1) 當 $n_i \to \infty$，則 q_{pi} 幾乎到處收斂 (converge almost surely) 至 ξ_p。若 $0 < p < 1$，F 在 ξ_p 可微和 $F'(\xi_p) > 0$，則當 $n_i \to \infty$，q_{pi} 漸近常態分配 $N(\xi_p, \sigma^2)$，其中 $\sigma^2 = \frac{p(1-p)}{[F'(\xi_p)]^2 n_i}$。

(2) 固定 $n_i = n$ 和當 $k \to \infty$，則 $\hat{q}_{p(k)}$ 機率收斂 (converge in probability) 至 ξ_p，而且漸近常態分配 $N(\xi_p, \frac{\sigma^2}{4k[\phi(\xi_p/\sigma)]^2})$，其中 $\phi(\cdot)$ 爲標準常態分配 (standard normal distribution) 的機率密度函數 (probability density function)；以及 $\bar{q}_{p(k)}$ 機率收斂至 ξ_p，而且漸近常態分配 $N(\xi_p, \frac{\sigma^2}{k})$。

定理三：

假設 $f(x)$ 爲分配 $F(x)$ 的機率密度函數，$\eta_{i0} \in (a_{v_i-1}, \xi_p)$，$\eta_{i1} \in (a_{v_i-1}, a_{v_i})$。若 $f(\eta_{i0}) = f(\eta_{i1})$，則當 $n_i \to \infty$, $\tilde{\tilde{q}}_{pi}$ 機率收斂至 ξ_p。(定理三證明參見附錄一)。

4. 模擬方法與結果分析

本模擬主要是針對計算龐大資料的中位數和百分位數，因爲中位數和百分位數不存在逐步遞迴的關係式，而且所有資料的中位數 (或百分位數) 和各區段資料的中位數 (或百分位數) 之間又沒有具體的公式，因此利用模擬

方法來比較第 2 節和第 3 節所提之四種方法的好壞。本模擬是利用個人電腦配合 FORTRAN 90 程式語言和 IMSL/LIB 程式庫生成資料計算而得。

假設有 I 段資料，令 $I = 4,\ 10,\ 20,\ 50,\ 100,\ 200$ 共 6 種情形，令每一區段資料的樣本數相等:$n_1 = n_2 = \cdots = n_I = n$，$n$ 的選取使得 $n \times I = 2 \times 10^6$。每一區段資料是由下列 4 種分配所產生的隨機樣本:

(1) 常態分配 $Normal(0,\ 1)$, $\xi_{0.5} = 0$, $Q_1 = -0.6745$, $Q_3 = 0.6745$。

(2) 均勻分配 $Uniform(0,\ 1)$, $\xi_{0.5} = 0.5$, $Q_1 = 0.25$, $Q_3 = 0.75$。

(3) 柯西分配 $Cauchy(0,\ 1)$, $\xi_{0.5} = 0$, $Q_1 = -1$, $Q_3 = 1$。

(4) 指數分配 $Exponential(1)$, $\xi_{0.5} = \ln 2$, $Q_1 = \ln 4 - \ln 3$, $Q_3 = \ln 4$。

模擬分爲三部份，分別爲中位數、第一四分位數和第三四分位數，每一種分配模擬 2000 次，對每一次依序生成的 I 段資料，於各區段分別計算中位數 m_i、第一四分位數 Q_{1i}，第三四分位數 Q_{3i}, $i = 1,\ \ldots,\ I$;對每一次依序生成的前 k 段資料計算中位數 $m_{(k)}$、第一四分位數 $Q_{1(k)}$ 和第三四分位數 $Q_{3(k)}$，並計算 $\hat{m}_{(k)}$, $\bar{m}_{(k)}$, $\tilde{m}_{(k)}$ 和 $\tilde{\tilde{m}}_{(k)}$ 來估計 $m_{(k)}$，在利用 (2.5) 式計算 $\tilde{\tilde{m}}_{(k)}$ 時，常態分配和柯西分配下令 $b = 0.1$, $L = 100$, $a_0 = -5$, $a_L = 5$;均勻分配下令 $b = 0.01$, $L = 100$, $a_0 = 0$, $a_L = 1$;指數分配下令 $b = 0.1$, $L = 100$, $a_0 = 0$, $a_L = 15$。另外計算 $\hat{Q}_{1(k)}$, $\bar{Q}_{1(k)}$, $\tilde{Q}_{1(k)}$ 和 $\tilde{\tilde{Q}}_{1(k)}$ 來估計 $Q_{1(k)}$;接著計算 $\hat{Q}_{3(k)}$, $\bar{Q}_{3(k)}$, $\tilde{Q}_{3(k)}$ 和 $\tilde{\tilde{Q}}_{3(k)}$ 來估計 $Q_{3(k)}$。最後計算中位數 $m_{(k)}$、第一四分位數 $Q_{1(k)}$ 和第三四分位數 $Q_{3(k)}$ 模擬 2000 次的偏差 (bias) 和其分別與 $\xi_{0.5}$、$\xi_{0.25}$ 和 $\xi_{0.75}$ 的根號均方差 (root mean squared error) (因篇幅關係未列入附表中，讀者若需要，可向作者索取)。

由於資料是循序漸進的，每隔一時段須計算該時段前所有資料的中位數 $m_{(k)}$，因此我們利用移動均方差來評估在每個時段各種近似中位數方法的表現。令 $m_{(k)}^{(j)}$ 代表第 j 次模擬生成的前 k 段資料中位數 $m_{(k)}$，則中位數估計量 $\hat{m}_{(k)}$ 模擬 2000 次的移動根號均方差 (moving root mean squared error) 爲

$$RMSE(\hat{m}_{(k)}) = [\sum_{j=1}^{2000} \sum_{k=1}^{I} (\hat{m}_{(k)}^{(j)} - m_{(k)}^{(j)})^2 / 2000I]^{1/2}\ , \qquad (4.1)$$

另外計算中位數估計量 $\hat{m}_{(k)}$ 模擬 2000 次的偏差

$$BIAS(\hat{m}_{(k)}) = MEAN(\hat{m}_{(k)}) - MEAN(m_{(k)})$$
$$= \sum_{j=1}^{2000}\sum_{k=1}^{I}[\hat{m}_{(k)}^{(j)} - m_{(k)}^{(j)}]/2000I\,, \tag{4.2}$$

類似地，可計算中位數估計量 $\bar{m}_{(k)}$, $\tilde{m}_{(k)}$, $\tilde{\tilde{m}}_{(k)}$ 模擬 2000 次的偏差和移動根號均方差。同樣地，可計算第一四分位數估計量 $\hat{Q}_{1(k)}$ 模擬 2000 次的移動根號均方差

$$RMSE(\hat{Q}_{1(k)}) = [\sum_{j=1}^{2000}\sum_{k=1}^{I}(\hat{Q}_{1(k)}^{(j)} - Q_{1(k)}^{(j)})^2/2000I]^{1/2}\,, \tag{4.3}$$

和第一四分位數估計量 $\hat{Q}_{1(k)}$ 模擬 2000 次的偏差

$$BIAS(\hat{Q}_{1(k)}) = MEAN(\hat{Q}_{1(k)}) - MEAN(Q_{1(k)})$$
$$= \sum_{j=1}^{2000}\sum_{k=1}^{I}[\hat{Q}_{1(k)}^{(j)} - Q_{1(k)}^{(j)}]/2000I\,, \tag{4.4}$$

類似地，可計算第一四分位數估計量 $\bar{Q}_{1(k)}, \tilde{Q}_{1(k)}, \tilde{\tilde{Q}}_{1(k)}$ 模擬 2000 次的偏差和移動根號均方差。同樣地，可計算第三四分位數估計量 $\hat{Q}_{3(k)}$ 模擬 2000 次的移動根號均方差

$$RMSE(\hat{Q}_{3(k)}) = [\sum_{j=1}^{2000}\sum_{k=1}^{I}(\hat{Q}_{3(k)}^{(j)} - Q_{3(k)}^{(j)})^2/2000I]^{1/2}\,, \tag{4.5}$$

和第三四分位數估計量 $\hat{Q}_{3(k)}$ 模擬 2000 次的偏差

$$BIAS(\hat{Q}_{3(k)}) = MEAN(\hat{Q}_{3(k)}) - MEAN(Q_{3(k)})$$
$$= \sum_{j=1}^{2000}\sum_{k=1}^{I}[\hat{Q}_{3(k)}^{(j)} - Q_{3(k)}^{(j)}]/2000I\,, \tag{4.6}$$

類似地，可計算第三四分位數估計量$\bar{Q}_{3(k)}$, $\tilde{Q}_{3(k)}$, $\tilde{\tilde{Q}}_{3(k)}$ 模擬 2000 次的偏差和移動根號均方差。

表 4 是假設資料分別來自常態分配，均勻分配，柯西分配和指數分配下，計算中位數 $m_{(k)}$ 和近似中位數 $\hat{m}_{(k)}$, $\bar{m}_{(k)}$, $\tilde{m}_{(k)}$, $\tilde{\tilde{m}}_{(k)}$ 估計量模擬 2000 次的樣本移動均方差和偏差。由表 4 來看，若固定所有資料量爲 $n \times I = 2 \times 10^6$，不論 I 值大或小，估計量 $\hat{m}_{(k)}$ 和 $\tilde{m}_{(k)}$ 的偏差和移動均方差均較 $\bar{m}_{(k)}$ 和 $\tilde{\tilde{m}}_{(k)}$ 的大，由於各估計量偏差單位是 10^{-6} 和移動均方差是 10^{-4}，故誤差雖呈倍數但實際差值仍在小數點第三位以下。當區段次數 I 增加且區段樣本數 n 減少時，各估計量的移動均方差隨著區段數增加而增加，其中以 $\tilde{m}_{(k)}$ 的移動均方差增加幅度最大。若資料來自常態分配，均勻分配和柯西分配，當 $I \leq 10$ 時，估計量 $\bar{m}_{(k)}$ 的移動均方差較 $\tilde{\tilde{m}}_{(k)}$ 的小；當 $I \geq 20$ 時，估計量 $\tilde{\tilde{m}}_{(k)}$ 的移動均方差較 $\bar{m}_{(k)}$ 的小。若資料來自指數分配時，不論 I 值大或小，估計量 $\bar{m}_{(k)}$ 的移動均方差較 $\tilde{\tilde{m}}_{(k)}$ 的小。

表 5 是假設資料分別來自常態分配，均勻分配，柯西分配和指數分配下，計算第一四分位數 $Q_{1(k)}$ 和近似第一四分位數 $\hat{Q}_{1(k)}$, $\bar{Q}_{1(k)}$, $\tilde{Q}_{1(k)}$, $\tilde{\tilde{Q}}_{1(k)}$ 估計量 2000 次的樣本移動均方差和偏差。由表 5 來看，當固定所有資料量爲 $n \times I = 2 \times 10^6$，在柯西分配中，當 $I \geq 50$ 時，估計量 $\tilde{\tilde{Q}}_{1(k)}$ 的移動均方差較 $\bar{Q}_{1(k)}$ 小，當 $I \leq 20$ 時，估計量 $\bar{Q}_{1(k)}$ 的移動均方差爲最小。若資料來自常態分配，均勻分配，指數分配時，不論 I 值大或小，估計量 $\bar{Q}_{1(k)}$ 的移動均方差皆是四個估計量中最小的。當區段次數 I 增加且區段樣本數 n 減少時，各估計量的移動均方差隨著區段數增加而增加，其中以 $\tilde{Q}_{1(k)}$ 的移動均方差增加幅度最大。

表 6 是假設資料分別來自常態分配，均勻分配，柯西分配和指數分配下，計算第三四分位數 $Q_{3(k)}$ 和近似第三四分位數 $\hat{Q}_{3(k)}$, $\bar{Q}_{3(k)}$, $\tilde{Q}_{3(k)}$, $\tilde{\tilde{Q}}_{3(k)}$ 估計量 2000 次的樣本移動均方差。由表 6 來看，當固定所有資料量爲 $n \times I = 2 \times 10^6$，除柯西分配中，當 $I = 200$ 時，估計量 $\tilde{\tilde{Q}}_{3(k)}$ 的移動均方差較 $\bar{Q}_{3(k)}$ 的小之外，其餘不論 I 值大或小和何種分配，估計量 $\bar{Q}_{3(k)}$ 的移動均方差皆是四個估計量中最小的。當區段次數 I 增加且區段樣本數 n 減少時，各估計量的移動均方差隨區段數增加而增加，其中以 $\tilde{Q}_{3(k)}$ 的移動均方差增加幅度最大。

表4 不同方法和分配下近似中位數之 RMSE ($\times 10^{-4}$) 和 BIAS ($\times 10^{-6}$)

	分配	$Normal(0,1)$	$Uniform(0,1)$	$Cauchy(0,1)$	$Exponential(1)$
I:區段數	方法	估計量之	估計量之	估計量之	估計量之
n:區段樣本數		RMSE(BIAS)	RMSE(BIAS)	RMSE(BIAS)	RMSE(BIAS)
$I=4$	$\hat{m}_{(k)}$	3.44(18.2)	1.43(-0.1)	4.47(-9.4)	2.78(-4.9)
$n=500000$	$\bar{m}_{(k)}$	0.37(23.1)	0.15(-0.4)	0.48(0.9)	0.30(0.4)
	$\tilde{m}_{(k)}$	2.38(26.8)	0.99(-0.5)	3.10(8.0)	9.42(3.8)
	$\tilde{\bar{m}}_{(k)}$	0.59(49.8)	0.21(0.0)	0.66(0.7)	3.37(314.3)
$I=10$	$\hat{m}_{(k)}$	6.35(28.9)	2.54(3.5)	7.70(-2.7)	4.95(7.2)
$n=200000$	$\bar{m}_{(k)}$	0.66(26.4)	0.26(0.7)	0.84(0.3)	0.54(1.6)
	$\tilde{m}_{(k)}$	6.61(21.9)	2.61(0.3)	8.21(15.1)	5.15(3.0)
	$\tilde{\bar{m}}_{(k)}$	0.98(90.9)	0.29(0.0)	0.91(1.6)	3.44(311.9)
$I=20$	$\hat{m}_{(k)}$	7.99(17.9)	3.21(-6.1)	10.22(-8.6)	6.48(3.6)
$n=100000$	$\bar{m}_{(k)}$	0.93(19.5)	0.37(-0.7)	1.17(-0.9)	0.77(5.2)
	$\tilde{m}_{(k)}$	10.66(9.9)	4.29(2.0)	13.78(5.4)	8.75(17.6)
	$\tilde{\bar{m}}_{(k)}$	1.57(145.7)	0.36(-0.5)	1.10(1.2)	3.57(318.3)
$I=50$	$\hat{m}_{(k)}$	10.22(-8.5)	4.03(-4.0)	12.84(-25.2)	7.98(18.8)
$n=40000$	$\bar{m}_{(k)}$	1.41(0.6)	0.54(0.3)	1.76(-7.3)	1.11(14.8)
	$\tilde{m}_{(k)}$	18.74(6.8)	7.27(-18.2)	22.32(15.8)	14.77(36.9)
	$\tilde{\bar{m}}_{(k)}$	1.14(-2.9)	0.46(0.0)	1.42(-0.4)	3.68(319.3)
$I=100$	$\hat{m}_{(k)}$	11.43(-19.1)	4.57(10.2)	14.41(31.6)	9.27(-3.4)
$n=20000$	$\bar{m}_{(k)}$	1.83(6.6)	0.71(0.4)	2.29(6.4)	1.49(23.6)
	$\tilde{m}_{(k)}$	26.41(-24.8)	10.27(14.4)	33.19(-51.7)	21.33(176.7)
	$\tilde{\bar{m}}_{(k)}$	1.40(-1.0)	0.55(-0.2)	1.73(0.7)	3.69(314.2)
$I=200$	$\hat{m}_{(k)}$	12.89(11.2)	5.12(3.1)	15.85(-2.0)	10.28(5.4)
$n=10000$	$\bar{m}_{(k)}$	2.37(4.9)	0.92(0.3)	2.95(-12.1)	1.94(47.4)
	$\tilde{m}_{(k)}$	35.26(-92.9)	14.43(-10.7)	43.66(-23.9)	29.45(286.3)
	$\tilde{\bar{m}}_{(k)}$	1.64(1.6)	0.68(0.8)	2.06(-1.5)	3.78(314.7)

表5 不同方法和分配下近似第一四分位數之 RMSE($\times 10^{-4}$) 和 BIAS($\times 10^{-6}$)

	分配	$Normal(0,1)$	$Uniform(0,1)$	$Cauchy(0,1)$	$Exponential(1)$
I:區段數	方法	估計量之	估計量之	估計量之	估計量之
n:區段樣本數		RMSE(BIAS)	RMSE(BIAS)	RMSE(BIAS)	RMSE(BIAS)
$I=4$	$\hat{Q}_{1(k)}$	4.01(-42.6)	1.25(-0.1)	7.63(5.8)	1.62(-1.2)
$n=500000$	$\bar{Q}_{1(k)}$	0.45(-25.1)	0.14(0.3)	0.89(-5.8)	0.18(0.5)
	$\tilde{Q}_{1(k)}$	2.83(-12.4)	0.88(0.7)	5.26(-13.1)	1.13(1.7)
	$\tilde{\tilde{Q}}_{1(k)}$	6.46(-620)	1.61(-0.6)	1.73(-96.9)	5.45(534.1)
$I=10$	$\hat{Q}_{1(k)}$	6.73(-8.8)	2.25(6.6)	13.35(-5.3)	5.54(4.4)
$n=200000$	$\bar{Q}_{1(k)}$	0.79(-21.0)	0.25(1.6)	1.56(-9.7)	0.34(2.1)
	$\tilde{Q}_{1(k)}$	6.98(-44.9)	2.30(-1.3)	13.74(-36.0)	3.01(5.3)
	$\tilde{\tilde{Q}}_{1(k)}$	6.25(-567.7)	1.92(-1.3)	2.24(-112.6)	5.54(534.6)
$I=20$	$\hat{Q}_{1(k)}$	8.74(17.2)	2.79(-1.8)	17.58(33.8)	3.69(5.2)
	$\tilde{Q}_{1(k)}$	1.11(-2.5)	0.35(3.6)	2.22(-15.2)	0.46(1.9)
$n=100000$	$\bar{Q}_{1(k)}$	11.78(-17.5)	3.69(14.1)	23.43(-62.5)	4.96(11.1)
	$\tilde{\tilde{Q}}_{1(k)}$	5.73(-476.6)	2.14(-1.5)	2.61(-120.9)	3.58(534.6)
$I=50$	$\hat{Q}_{1(k)}$	10.93(17.3)	3.50(8.3)	21.69(25.6)	4.67(10.1)
$n=40000$	$\bar{Q}_{1(k)}$	1.69(-0.4)	0.50(6.5)	3.36(-56.6)	0.69(12.9)
	$\tilde{Q}_{1(k)}$	20.40(7.6)	6.34(-11.8)	39.78(-231.2)	8.38(-4.4)
	$\tilde{\tilde{Q}}_{1(k)}$	9.13(-35.3)	2.38(-0.5)	3.07(-120.5)	5.62(532.8)
$I=100$	$\hat{Q}_{1(k)}$	12.62(-4.6)	3.77(1234.5)	25.87(-58.0)	5.31(-3.4)
$n=20000$	$\bar{Q}_{1(k)}$	2.12(6.7)	0.60(700.4)	4.57(-36.0)	0.94(25.1)
	$\tilde{Q}_{1(k)}$	29.04(-15.2)	7.71(2453.2)	49.21(-171.1)	11.98(102.1)
	$\tilde{\tilde{Q}}_{1(k)}$	10.07(30.4)	2.99(153.1)	3.87(-98.3)	5.69(534.7)
$I=200$	$\hat{Q}_{1(k)}$	13.97(20.4)	4.15(1544.0)	31.42(-17.3)	6.78(30.1)
$n=10000$	$\bar{Q}_{1(k)}$	2.76(15.1)	0.73(2319.5)	5.17(-70.5)	1.15(109.6)
	$\tilde{Q}_{1(k)}$	39.59(58.5)	8.24(697.9)	60.07(-127.9)	13.75(92.8)
	$\tilde{\tilde{Q}}_{1(k)}$	11.54(18.1)	3.84(-1281.6)	4.26(-218.9)	5.74(575.7)

表6 不同方法和分配下近似第三四分位數之 RMSE ($\times 10^{-4}$) 和 BIAS ($\times 10^{-6}$)

	分配	$Normal(0,1)$	$Uniform(0,1)$	$Cauchy(0,1)$	$Exponential(1)$
I:區段數	方法	估計量之	估計量之	估計量之	估計量之
n:區段樣本數		RMSE(BIAS)	RMSE(BIAS)	RMSE(BIAS)	RMSE(BIAS)
$I=4$	$\hat{Q}_{3(k)}$	3.91(-16.65)	1.21(-0.2)	7.68(-1.8)	5.04(-3.5)
$n=500000$	$\bar{Q}_{3(k)}$	0.45(-10.05)	0.13(-0.3)	0.88(-0.9)	0.57(1.0)
	$\tilde{Q}_{3(k)}$	2.74(-5.85)	0.84(-0.5)	5.36(-0.9)	3.51(-0.1)
	$\tilde{\tilde{Q}}_{3(k)}$	6.97(741.25)	1.60(-1.1)	1.82(105.3)	6.31(-4.7)
$I=10$	$\hat{Q}_{3(k)}$	6.76(-25.55)	2.17(-6.3)	13.64(41.4)	8.79(-17.4)
$n=200000$	$\bar{Q}_{3(k)}$	0.78(-18.85)	0.25(-1.9)	1.59(10.0)	1.02(0.3)
	$\tilde{Q}_{3(k)}$	7.13(-7.35)	2.21(-0.6)	14.30(-32.7)	9.30(19.7)
	$\tilde{\tilde{Q}}_{3(k)}$	7.59(678.75)	1.90(-2.8)	2.26(116.3)	6.54(-6.6)
$I=20$	$\hat{Q}_{3(k)}$	8.81(-24.25)	2.77(0.1)	17.71(-16.6)	11.14(-21.4)
$n=100000$	$\bar{Q}_{3(k)}$	1.11(-20.35)	0.35(-3.3)	2.25(15.9)	1.41(-2.2)
	$\tilde{Q}_{3(k)}$	11.70(-35.95)	3.63(-0.3)	23.09(109.4)	14.97(28.3)
	$\tilde{\tilde{Q}}_{3(k)}$	8.43(758.85)	2.11(2.9)	2.76(130.2)	6.56(581.5)
$I=50$	$\hat{Q}_{3(k)}$	13.04(-53.8)	3.50(-7.1)	23.54(153.1)	13.68(-12.0)
$n=40000$	$\bar{Q}_{3(k)}$	1.42(6.7)	0.52(2.0)	2.86(213.1)	2.06(1.3)
	$\tilde{Q}_{3(k)}$	13.73(0.5)	6.31(-8.4)	31.73(611.1)	25.46(207.2)
	$\tilde{\tilde{Q}}_{3(k)}$	6.97(660.9)	2.38(1.0)	3.17(102.4)	6.70(580.0)
$I=100$	$\hat{Q}_{3(k)}$	20.17(-56.3)	3.96(-12.3)	31.72(177.5)	15.75(-19.7)
$n=20000$	$\bar{Q}_{3(k)}$	1.78(-50.2)	0.68(-52.5)	3.48(164.9)	2.66(23.2)
	$\tilde{Q}_{3(k)}$	15.07(-66.9)	9.13(-5.5)	40.99(453.6)	36.43(294.7)
	$\tilde{\tilde{Q}}_{3(k)}$	6.97(615.1)	2.54(-1.0)	3.66(84.1)	6.86(583.8)
$I=200$	$\hat{Q}_{3(k)}$	25.30(-63.3)	4.47(0.7)	38.64(18.8)	16.92(-72.6)
$n=10000$	$\bar{Q}_{3(k)}$	2.13(-74.5)	0.80(-1.3)	4.01(7.3)	3.27(-30.3)
	$\tilde{Q}_{3(k)}$	18.54(-57.8)	10.01(-47.0)	48.51(108.2)	47.51(89.6)
	$\tilde{\tilde{Q}}_{3(k)}$	6.98(605.8)	2.86(0.3)	4.00(143.1)	6.97(547.5)

另外，本文也模擬當資料並非獨立情形，假設資料分段前依序令爲X_1, X_2, ...，則資料可視爲一時間數列 $\{X_t\}$。時間數列分析中的自迴歸模式 (autoregressive model, 記做 $AR(p)$) 及移動平均模式 (moving average model, 記做 $MA(q)$) 分別定義爲

$$X_t = \theta + \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} + a_t \text{，}$$

和

$$X_t = \theta + a_t + \theta_1 a_{t-1} + \cdots + \theta_q a_{t-q} \text{，}$$

其中a_t爲白色干擾過程(white noise process), 一般假設$a_t; t = 1, 2, \ldots,$獨立且具相同分配$N(0, \sigma^2)$。

表 7 中的模擬是利用 $AR(1)$、$MA(1)$、$AR(3)$ 和 $MA(3)$ 模型生成資料後再分段 ($I = 4$, $n = 500000$)，依本文前述方法計算中位數 $m_{(k)}$ 和近似中位數 $\hat{m}_{(k)}$, $\bar{m}_{(k)}$, $\tilde{m}_{(k)}$, $\tilde{\tilde{m}}_{(k)}$，第一四分位數 $Q_{1(k)}$ 和近似第一四分位數 $\hat{Q}_{1(k)}$, $\bar{Q}_{1(k)}$, $\tilde{Q}_{1(k)}$, $\tilde{\tilde{Q}}_{1(k)}$ 以及第三四分位數 $Q_{3(k)}$ 和近似第三四分位數 $\hat{Q}_{3(k)}$, $\bar{Q}_{3(k)}$, $\tilde{Q}_{3(k)}$, $\tilde{\tilde{Q}}_{3(k)}$ 這些估計量模擬 2000 次的樣本移動均方差。模擬參數 $\theta = 0$, $X_0 = 0$, $\sigma^2 = 1$，參數 ϕ_i 和 θ_i 的設定參見表 7。由表 7 來看，利用各區段取中位數後再平均的方法來估計中位數，其移動均方差最小，第一四分位數和第三四分位數也有相同的結論。

5. 結論

本文利用二階段組合中心趨勢的統計量，如:平均數、中位數、值域中點來求得前 k 段資料的近似中位數和近似百分位數，利用統計模擬來比較不同近似方法的好壞。由模擬得到以下結論:

(1) 除了指數分配，不論在區段次數 I 多或少時，利用各區段取中位數後再平均的方法 $\bar{m}_{(k)}$ 來估計中位數 $m_{(k)}$ 比較好之外，其餘在常態分配，均勻分配，柯西分配下，在區段次數 I 少時，利用取中位數

表7. 不同資料下近似中位數、第一四分位數和第三四分位數之 RMSE ($\times 10^{-4}$) ($I = 4,\ n = 500000$)

資料來源	方法	樣本 RMSE	方法	樣本 RMSE	方法	樣本 RMSE
$AR(1)$	$\hat{m}_{(k)}$	3.828	$\hat{Q}_{1(k)}$	4.157	$\hat{Q}_{3(k)}$	4.177
$\phi_1 = 0.1$	$\bar{m}_{(k)}$	0.391	$\bar{Q}_{1(k)}$	0.471	$\bar{Q}_{3(k)}$	0.469
	$\tilde{m}_{(k)}$	2.689	$\tilde{Q}_{1(k)}$	2.928	$\tilde{Q}_{3(k)}$	2.940
	$\tilde{\tilde{m}}_{(k)}$	0.547	$\tilde{\tilde{Q}}_{1(k)}$	6.101	$\tilde{\tilde{Q}}_{3(k)}$	6.126
$AR(1)$	$\hat{m}_{(k)}$	6.281	$\hat{Q}_{1(k)}$	6.752	$\hat{Q}_{3(k)}$	6.628
$\phi_1 = 0.5$	$\bar{m}_{(k)}$	0.540	$\hat{Q}_{1(k)}$	0.626	$\hat{Q}_{3(k)}$	0.632
	$\tilde{m}_{(k)}$	4.365	$\tilde{Q}_{1(k)}$	4.675	$\tilde{Q}_{3(k)}$	4.627
	$\tilde{\tilde{m}}_{(k)}$	0.751	$\tilde{\tilde{Q}}_{1(k)}$	5.372	$\tilde{\tilde{Q}}_{3(k)}$	5.287
$AR(1)$	$\hat{m}_{(k)}$	29.718	$\hat{Q}_{1(k)}$	31.059	$\hat{Q}_{3(k)}$	30.701
$\phi_1 = 0.9$	$\bar{m}_{(k)}$	1.748	$\hat{Q}_{1(k)}$	1.994	$\hat{Q}_{3(k)}$	1.976
	$\tilde{m}_{(k)}$	20.408	$\tilde{Q}_{1(k)}$	21.278	$\tilde{Q}_{3(k)}$	21.440
	$\tilde{\tilde{m}}_{(k)}$	2.171	$\tilde{\tilde{Q}}_{1(k)}$	5.411	$\tilde{\tilde{Q}}_{3(k)}$	5.492
$AR(3)$	$\hat{m}_{(k)}$	29.099	$\hat{Q}_{1(k)}$	31.117	$\hat{Q}_{3(k)}$	31.192
$\phi_1 = 0.5$	$\bar{m}_{(k)}$	1.531	$\hat{Q}_{1(k)}$	1.775	$\hat{Q}_{3(k)}$	1.831
$\phi_2 = 0.3$	$\tilde{m}_{(k)}$	20.070	$\tilde{Q}_{1(k)}$	21.469	$\tilde{Q}_{3(k)}$	21.590
$\phi_3 = 0.1$	$\tilde{\tilde{m}}_{(k)}$	1.981	$\tilde{\tilde{Q}}_{1(k)}$	3.085	$\tilde{\tilde{Q}}_{3(k)}$	3.152
$MA(1)$	$\hat{m}_{(k)}$	3.314	$\hat{Q}_{1(k)}$	3.657	$\hat{Q}_{3(k)}$	3.729
$\theta_1 = 0.1$	$\bar{m}_{(k)}$	0.376	$\hat{Q}_{1(k)}$	0.447	$\hat{Q}_{3(k)}$	0.437
	$\tilde{m}_{(k)}$	2.298	$\tilde{Q}_{1(k)}$	2.574	$\tilde{Q}_{3(k)}$	2.598
	$\tilde{\tilde{m}}_{(k)}$	0.508	$\tilde{\tilde{Q}}_{1(k)}$	6.802	$\tilde{\tilde{Q}}_{3(k)}$	6.008
$MA(1)$	$\hat{m}_{(k)}$	2.706	$\hat{Q}_{1(k)}$	3.365	$\hat{Q}_{3(k)}$	3.306
$\theta_1 = 0.5$	$\bar{m}_{(k)}$	0.352	$\hat{Q}_{1(k)}$	0.420	$\hat{Q}_{3(k)}$	0.453
	$\tilde{m}_{(k)}$	1.901	$\tilde{Q}_{1(k)}$	2.350	$\tilde{Q}_{3(k)}$	2.324
	$\tilde{\tilde{m}}_{(k)}$	0.522	$\tilde{\tilde{Q}}_{1(k)}$	8.079	$\tilde{\tilde{Q}}_{3(k)}$	8.056
$MA(1)$	$\hat{m}_{(k)}$	2.822	$\hat{Q}_{1(k)}$	3.694	$\hat{Q}_{3(k)}$	3.734
$\theta_1 = 0.9$	$\bar{m}_{(k)}$	0.390	$\hat{Q}_{1(k)}$	0.513	$\hat{Q}_{3(k)}$	0.502
	$\tilde{m}_{(k)}$	1.997	$\tilde{Q}_{1(k)}$	2.588	$\tilde{Q}_{3(k)}$	2.633
	$\tilde{\tilde{m}}_{(k)}$	0.555	$\tilde{\tilde{Q}}_{1(k)}$	2.527	$\tilde{\tilde{Q}}_{3(k)}$	2.512
$MA(3)$	$\hat{m}_{(k)}$	2.605	$\hat{Q}_{1(k)}$	2.056	$\hat{Q}_{3(k)}$	0.426
$\theta_1 = 0.5$	$\bar{m}_{(k)}$	0.358	$\hat{Q}_{1(k)}$	0.421	$\hat{Q}_{3(k)}$	0.224
$\theta_2 = 0.3$	$\tilde{m}_{(k)}$	1.819	$\tilde{Q}_{1(k)}$	2.966	$\tilde{Q}_{3(k)}$	3.198
$\theta_3 = 0.1$	$\tilde{\tilde{m}}_{(k)}$	0.510	$\tilde{\tilde{Q}}_{1(k)}$	4.397	$\tilde{\tilde{Q}}_{3(k)}$	4.491

後再平均的方法 $\bar{m}_{(k)}$ 來估計中位數 $m_{(k)}$，是比較好的。但區段次數 I 增加時，利用直方圖求中位數近似值方法 $\tilde{\tilde{m}}_{(k)}$ 來估計中位數 $m_{(k)}$，是比較好的。就移動均方差而言，各估計量隨著區段次數增加而增加。

(2) 除了柯西分配在區段次數 $I \geq 50$ 以外，不論在常態分配，均勻分配，指數分配下，利用各區段取第一四分位數後再平均的方法 $\bar{Q}_{1(k)}$ 來估計第一四分位數 $Q_{1(k)}$，比利用直方圖求第一四分位數的近似方法 $\tilde{\tilde{Q}}_{1(k)}$ 好。就移動均方差而言，各估計量隨著區段次數增加而增加。

(3) 除了柯西分配在區段次數 $I = 200$ 以外，不論在常態分配，均勻分配，指數分配下，利用各區段取第三四分位數後再平均的方法 $\bar{Q}_{3(k)}$ 來估計第三四分位數 $Q_{3(k)}$，比利用直方圖求第三四分位數的近似方法 $\tilde{\tilde{Q}}_{3(k)}$ 好。就移動均方差而言，各估計量隨著區段次數增加而增加。

(4) 利用 $AR(1)$、$MA(1)$、$AR(3)$ 和 $MA(3)$ 模型生成資料後再分段，由各區段取中位數後再平均的方法來估計中位數、第一四分位數或第三四分位數，其移動均方差皆較其他方法小。

附錄一

定理一之證明：

因爲先求得各區段的中位數，依中位數的定義：如果 $n_i = 2c_i + 1$，則 $m_i = X_{i(c_i+1)}$，如果 $n_i = 2c_i$，則 $m_i = (X_{i(c_i)} + X_{i(c_i+1)})/2$，即比 m_i 小的資料有 c_i 個，比 m_i 大的資料有 c_i 個，若再將 $m_i, i = 1, \ldots, k,$ 排序求中位數 $\hat{m}_{(k)}$，則有 C_k 個資料比 $\hat{m}_{(k)}$ 小，有 $N_k - C_k$ 個資料比 $\hat{m}_{(k)}$ 大，故近似中位數 $\hat{m}_{(k)}$ 會介於資料排序後第 $C_k + 1$ 個資料和第 $N_k - C_k$ 個資料之間；即 $V_{k,\,C_k+1} \leq \hat{m}_{(k)} \leq V_{k,N_k-C_k}$，其中 $C_k = \sum_{i=1}^{[k/2]} c_i$。

定理二之證明：

因爲先求得各區段的第 $100p$ 百分位數，依百分位數的定義：如果 $n_i p = d_i + g_i$ ；$g_i < 1$ ；則 $x_{ip} = x_{i(d_i+1)}$，如果 $n_i p = d_i + g_i$，$g_i = 0$，則 $q_{ni} = (X_{i(d_i)} +$

$X_{i(d_i+1)})/2$，即比 q_{pi} 小的資料有 d_i 個，比 q_{pi} 大的資料至少有 $n_i - d_i - 1$ 個，若再將 q_{pi}, $i = 1, \ldots, k$，排序求中位數 $\hat{q}_{p(k)}$，則有 L_k 個資料比 $\hat{q}_{p(k)}$ 小，有 U_k 個資料比 $\hat{q}_{p(k)}$ 大，故近似第 $100p$ 百分位數 $\hat{q}_{p(k)}$ 會介於資料排序後第 $L_k + 1$ 個資料和第 $N_k - U_k$ 個資料之間;即 $V_{k,\,L_k+1} \le \hat{q}_{p(k)} \le V_{k,N_k-U_k}$，其中 $L_k = \sum_{i=1}^{[k/2]} d_i$，$U_k = \sum_{i=1}^{[k/2]}(n_i - d_i - 1)$。

定理三之證明：

因爲 $(n_{i1},\ n_{i2},\ \ldots,\ n_{iL})$ 爲一多項式分配並具參數 $(n_i,\ p_1,\ p_2,\ \ldots,\ p_L)$，其中 $\sum_{l=1}^{L} n_{il} = n_i$, $p_l = Pr(a_{l-1} < X_{ij} \le a_l)$ 和 $\sum_{l=1}^{L} p_l = 1$，因爲 $(n_{i1}/n_i,\ n_{i2}/n_i,\ \ldots,\ n_{iL}/n_i)$ 爲 $(p_1,\ p_2,\ \ldots,\ p_L)$ 的最大概似估計量，故 $\tilde{\tilde{q}}_{pi}$ 機率收斂至

$$a_{v_i-1} + \frac{b(p - \sum_{l=1}^{v_i-1} Pr(a_{l-1} < X_{ij} \le a_l))}{Pr(a_{v_i-1} < X_{ij} \le a_{v_i})} = a_{v_i-1} + \frac{bPr(a_{v_i-1} < X_{ij} \le \xi_p)}{Pr(a_{v_i-1} < X_{ij} \le a_{v_i})},$$

利用均值定理(the mean value theorem)，上式可改寫成$a_{v_i-1}+(\xi_p-a_{v_i-1})f(\eta_{i0})/f(\eta_{i1})$，其中 $f(x)$ 爲分配 $F(x)$ 的機率密度函數 (probability density function)，$\eta_{i0} \in (a_{v_i-1}, \xi_p)$, $\eta_{i1} \in (a_{v_i-1},\ a_{v_i})$。若 $f(\eta_{i0}) = f(\eta_{i1})$，則 $\tilde{\tilde{q}}_{pi}$ 機率收斂至 ξ_p。

參考文獻

Han, J. and Kamber, M. (2000). *Data Mining: Concepts and Techniques.* Morgan Kaufmann Publishers.

http://www.gss.com.tw/gsseis/12/datamini.htm, Data Mining的功能及建置方法簡介.

http://www.gss.com.tw/gsseis/13/mining.htm, Data Mining探索(上).

http://www.gss.com.tw/gsseis/14/mining2.html, Data Mining探索(下).

http://www.mis.yuntech.edu.tw/ hsucc/DataMining.htm, 資料探勘.

http://bbs.ibmap.com.tw/SoftwareToday/199701/Page18.html, 從原始資料技巧找出意義-新軟體可找出從前未能發現的模式.

Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics.* Wiley, New York.

[民國91年12月17日收稿，民國92年3月24日接受。]

Mining the Descriptive Statistics of a Large Data Bases

Mi-Chia Ma[1] Pei-Fang Su[1] Dennis K. J. Lin[2]

[1]Department of Statistics

National Cheng-Kung University

[2]Department of Statistics

The Pennsylvania State University

ABSTRACT

The fast-growing and tremendous amount of data are collected and stored in computers. When the descriptive statistics like mean, variance, median, quantile and correlation coefficient are computed for a large data set, the computer is limited in its memory space. If the raw data is discarded at fixed periods, one wants to find a recursive method by using some representative statistics of subgroup data to compute the descriptive statistics of accumulative and numerous data bases. This paper proposes a two-stages method of approximate median (or quantile) by taking median (or quantile) of subgroup data in the first stage and taking mean, median and midrange of the previous medians (or quantiles) in the second stage. The approximate median (or quantile) formula of Han and Kamber (2000) is also considered. Furthermore, the statistical properties of some procedures are discussed. Finally, these methods are compared by the moving mean squared error in simulation.

Key words and phrases: Correlation coefficient, mean, median, quantile, variance.

AMS 2000 subject classifications: Primary 62F10; secondary 65C05 .