龐大資料集的統計推論方法

林億雄　　林共進

成功大學統計系

摘　要

分析龐大資料集之所以艱鉅，主因受限於電腦內儲存記憶體遠小於儲存資料量的空間，故無法完整地儲存所欲分析之全部資料。本文試以提出估計龐大資料集母體參數方法為主，進而將所估計的結果探討其漸近性質。再以，若系統能分析所有資料集，本方法的估計結果與其一樣具有效性。同時，利用本方法能明顯地減少所需儲存資料的記憶體。文中亦會建構估計結果的漸近常態性質。此外，估計結果的標準誤差將以公式提出，配合經驗驗證其為合理，並藉此對有興趣的參數做統計推論。文中結合模擬探討本方法的有效性，並以網際網路流量資料作為例子說明。

關鍵詞：網際網路流量資料，核密度函數估計，再生中位數。

1 文獻回顧

過去的十年，我們目睹資訊科技的重大變革。在資訊科技的演進上，資訊系統例行性收集資料亦將為十分普遍。因此，在資料庫的設計上，收集的欄位將可能達上百個、資料筆數將達上百萬筆，並且檔案大小高達好幾兆位元（相當於1000GB 以上）。這些情形也將打破傳統的思維，成為普遍存在的情況。舉例而言，英國 Barclaycard 信用卡發卡公司平均每年有超過 3 億 5000 萬筆交易資料，Wal-Mark 每年有超過 70 億筆的顧客購物資料，AT&T（美國電話電報公司）每年更有超過 700 億筆的長途電話紀錄資料（可參考: Hand, Blunt, Kelly and Adams, 2000）。根據上述的例子，再加以現今資訊科技持續進步，龐大資料集的產生勢必為必然會發生的情形。而緊接而至的問題，在於利用傳統的統計方法或統計軟件是否能順利分析該資料。很不幸地，這問題的答案是利用傳統統計方法或統計軟件均會出現問題。就統計軟件而言，當資料數量非常大時，進行資料分析的電腦將會發生記憶體容量不足，導致無法儲存或讀取資料的情形，更別論進行分析工作。

一般而言，收集者對龐大資料集的收錄通常在事先是沒有特別的目的，或者說收集者的目的並不明確。相對之下，實驗設計所得的資料是經由事先設計收集。所以，就資料收集目的而言，龐大資料集的收錄與經由實驗設計所收集的資料是不同的。但對於資料價值的汲取，二者是相同的。在龐大資料集的收錄上常為收集者渴望從資料中發現一些有趣的特徵或形態(pattern)，並藉以提供有價值的資訊供決策制定。對於龐大資料集的分析，首要困難的工作包含: 資料處理/編輯/儲存過程、資料分類、偵測資料是否有不正常的形態、資料的摘要、資料的視覺圖形、資料間是否有相關等問題。所以，如何從大資料集中汲取有用的資訊將會成為一項重要的挑戰。

對龐大資料集進行簡易與預先分析前，一般擬就統計量進行討論。舉例而言，繪製資料的盒形圖（Box-Plot）時，我們需事先求得樣本的分位數。在資料小時，求分位數為十分容易的事；但當龐大資料集時求分位數將不再是一件容易的事。試想估計分位數這問題，假設給定 n 個獨立來自未知分配 F 的樣本資料，而我們想知道該資料集的第 100α 分位數。假設 $F(\xi_\alpha)=\alpha$，則第 100α 分位數為 ξ_α。類似的問題，我們可以求 n 個觀測值中第 k*th* 小的資料，或估計第 100α 母體分位數，其意義在於提供一求得近似 $[\alpha n]$ 大的觀測值。直觀而言，欲求分位數時，需所有資料均完成排序。然而，最困難的部分在當電腦負責運算的記憶體空間遠小於能處理 n 個觀測值的時候，這樣的處理過程將會使電腦無法運算。既然如此，n 個觀測值的排序就變的不可能了。為了克服這問題，Hurley and Nodarres（1995）亦提出低貯藏分位數估計法（Low-Storage Quartile Estimation Method）；同時，Rousseeuq and Bassett（1990）提出再生中位數法（Remedian Method）。Chao and Lin（1993）探

討再生中位數估計結果的漸進行為，並且發現其估計結果並不會漸進常態。換言之，就理論而言很難對再生中位數法的估計結果進行統計推論。

在電腦執行環境中，雖然記憶體空間可以高達好幾兆位元，但可儲存資料的主記憶體空間是有限制的。換言之，可用的記憶體空間一般雖大，但有限；也由於執行環境中儲存資料陣列有其一定的限制，且少於可用的總記憶體空間，並與其無直接關係，所以，龐大資料集的分析工作呈現出明顯的難題，舉例來說如：錯綜複雜的變數關係、大量記憶體的需求。概括而言，在小資料集上是簡單的工作轉換至龐大資料集將是一件困難的事。

本文之重點俾使分析數個真實大型資料集為可行。其一個案為網際網路流量資料，便將於第五章對其作一分析探討。在網際網路工程與管理上，系統管理員或網路負責人基本是依賴對網路流量特質的了解。由於網路連接文化、政治與商業，而其網際網路的流量資料之所以令人激動，原因在於它能量測這之間錯綜複雜且快速成長的網路行為。大體而言，更深一層的認識網際網路流量，對於網路功能監控、設備計劃、服務品質、安全防護、與網際網路商業技術工程是有貢獻的。Cleveland and Sun 於 2000 年，提出針對網際網路資料分析有用的方法與幫助。

過去有很多的研究工作，集中在對網際網路流量資料配置不同的統計模型。本文，將著重於發展龐大資料集的計算與推論工具上，其中分析的資料包含網際網路流量資料與電子商務資料。由於，資料大到無法儲存於記憶體時，很多統計量將無法利用傳統的統計方法或統計軟件計算而得。進一步的說法，我們將無法藉由分析龐大資料集而從中汲取有用的資訊。本文，試以提出較一般的方法去解決龐大樣本數下的統計推論問題。該方法可以廣泛地被應用，並能夠針對母體分配 F 的任何參數 $\theta(F)$ 作統計推論。同時，在很少的假設下，該方法的估計結果可以被證明具有強一致性與漸進常態的性質。此外，在一些情形下本方法估計結果會與利用全部的資料估計一樣有效。該方法亦可利用於點估計上，其結果與利用密度函數估計的結果一樣好。

本文的主要章節如下，於第二章對所利用的方法提出基本的想法，俾使其理論合理化。第三、四章將分別探討龐大資料集之平均數與標準差之估計問題與其相關統計推論，如：信賴區間與假設檢定，在第五章將探討龐大資料集之分位數估計方法，第六章介紹密度函數估計方法，第七章將試圖對 AT&T 的網際網路流量資料進行分析，第八章為結論。

2 方法說明

由於，電腦記憶體的有限空間與計算環境下最大儲存陣列的限制，若欲於龐大資料集中計算基本統計量，如樣本的分位數，將成為不易的事。我們將於本章中試以提出能克服記憶體不足的方法，並以此方法估計有興趣的參數。

由於，我們欲估計母體F的參數$\theta(F)$，如分位數或母體的密度，一般需要將所有資料儲存於記憶體中，俾使獲得有效的估計。舉例而言，對於分位數的計算。首先，需先將資料儲存於陣列中且計算排序。然而，替代抽樣法（SubSampling techniques）的提出克服上述需先將全部資料計算排序的問題。此方法對於預先分析上是有用的，但因其僅反應部分資料訊息故其估計結果並不一定有效。

在有效性的考量下，利用所有資料估計的結果必然比利用部分資料估計來的佳；但，由於龐大資料所涉及記憶體儲存空間的問題，故利用所有資料進行估計是不可行的。直覺上，我們可利用分區的方式逐步讀入資料並將其儲存，再以針對各區進行分析。只要區中的資料數不大，就能在各種不同的計算環境下輕易地在每區中利用估計方法估計所感興趣的母體參數。然而，這樣的分析方法對最後的結論有何貢獻？這問題在後文將會一併探討。

假設有一獨立且同分布的龐大資料集其樣本數為n，我們欲估計其母體中位數。若以樣本中位數估計之，則在計算時需至少能儲存n筆資料的記憶體儲存空間。然而，當n很大，如10,000,000，普通的方法將會使儲存的記憶體不足，並且導致運算失敗。相同的問題，當我們考慮樣本數10,000時，我們可利用統計軟件如SAS與S-Plus輕易地計算樣本中位數。根據上述，試以分區的觀念俾使分析樣本數10,000,000筆為可行，其方法為逐次讀入資料，如每次讀入10,000筆；接續，再以計算每區的中位數，使每區的樣本中位數均彼此獨立且同分布。同時，在很少的假設下可證明利用分區的方法所得的樣本中位數為獨立且滿足漸進常態性質，並且其期望值為母體中位數。由此，最自然估計母體中位數的方式為所有分區中估得的樣本中位數之加總平均。總結而言，欲估計大資料集的母體參數$\theta(F)$，我們利用二階段的方法。第一階段利用分區的方式逐次讀入資料，並且估計該區的參數$\theta(F)$。接續，我們利用所有分區中所估得的樣本中位數對其加總取其平均，最後就以此平均值估計龐大資料集真正的母體參數$\theta(F)$。所需注意的是在第二階段中對於區集的儲存，如果上一個分區已經處理完成，立即以下一個區集佔據更新其所利用的記憶體空間，如此於系統內反覆進行，這樣的處理對系統而言，將不需擴充新的記憶體空間。

假設$x_{i1},\ldots,x_{i\alpha_n}$為來自母體$F$的獨立同分布樣本，其中$x_i$可為隨機變數或為隨機向量。倘若，我們欲估計母體參數$\theta(F)$。就建構估計的方法，首先試以改寫樣本的表示式：

$$\begin{array}{cccc} x_{11}, & x_{12}, & \cdots, & x_{1\alpha_n} \\ x_{21}, & x_{22}, & \cdots, & x_{2\alpha_n} \\ \cdot & \cdot & & \cdot \\ & & \cdots, & \\ \cdot & \cdot & & \cdot \\ x_{\beta_n 1}, & x_{\beta_n 1}, & \cdots, & x_{\beta_n \alpha_n} \end{array}$$

其中 $x_{ij} = x_{(i-1)*\alpha_n + j}$ ，$i = 1,...,\alpha_n$ ，$j = 1,...,\beta_n$ 。α_n 為每區個數，β_n 為區集總數，故 $n = \alpha_n \beta_n$ 。再以，根據事先所選定的每區個數 α_n ，我們就能利用分區的方式掌握對於 θ 的估計。其次，我們利用相同的方式對每區進行 $\theta(F)$ 的估計，並標示為 $\hat{\theta}_{in}$ 表為第 i 個區集資料估計結果。由於研究結果發現估計的結果對於 α_n 選取的大小為很穩健，通常 α_n 的選擇通常建議取為 $O(\sqrt{n}\log\log(n))$ 。最後，$\theta(F)$ 以所有 $\hat{\theta}_{in}$ 加總平均估計。其數學表示式如下：

$$\bar{\theta} = \frac{1}{\beta_n}\sum_{i=1}^{\beta_n}\hat{\theta}_{in} \tag{2.1}$$

接續，我們將探討估計值 $\bar{\theta}$ 的樣本性質。

命題 1. 對所有大於 0 的正整數 α_n 與 β_n ，

(a) 如果 $\hat{\theta}_{in}$ 為不變估計量，則 $\bar{\theta}$ 亦同；

(b) 如果 $\hat{\theta}_{in}$ 為 θ 不偏估計量，則 $\bar{\theta}$ 亦同；

命題 2. 假設 $x_{i1},...,x_{i\alpha_n}$ 為獨立同分布，且當 $n \to \infty$ 時，$\alpha_n \to \infty$ 與 $\beta_n \to \infty$ 。

(a) 如果 $\hat{\theta}_{in}$ 弱收斂至真值 θ ，則 $\bar{\theta}$ 亦同；

(b) 如果 $\hat{\theta}_{in}$ 以 L_2 收斂至真值 θ ，則 $\bar{\theta}$ 亦同；

(c) 如果 $\hat{\theta}_{in}$ 強收斂至真值 θ ，則 $\bar{\theta}$ 亦同；

為了建構 $\bar{\theta}$ 的漸進常態理論，以 μ_n 代表 $\hat{\theta}_{in}$ 的期望值其數學式為 $E(\hat{\theta}_{in})$ ；σ_n^2 代表 $\hat{\theta}_{in}$ 的變異數其數學式為 $\text{var}(\hat{\theta}_{in})$ 。除此之外，我們需要下列二條件：

條件(a) α_n 為一常數其與 n 獨立，同時 $\sigma_n^2 < \infty$ 。

條件(b) 當 $n \to \infty$ 時，$\alpha_n \to \infty$ 與 $\beta_n \to \infty$，並且

$$\frac{E\left|\hat{\theta}_{in} - \mu_n\right|^{2+\delta}}{\beta_n^{\delta/2}\sigma_n^{2+\delta}} \to 0 \tag{2.2}$$

當 $n \to \infty$ 時，對某些 $\delta > 0$。

定理 1. 假設 $x_{i1},...,x_{i\alpha_n}$ 為獨立同分布，如果條件(a)或(b)中一條件成立，則

$$\sqrt{\beta_n}\left(\frac{\bar{\theta} - \mu_n}{\sigma_n}\right) \to N(0,1) \tag{2.3}$$

證明：如果條件 (a) 成立，則 μ_n 與 σ_n 將與 n 無關。同時，$\hat{\theta}_{in}$ 為獨立且同分布其變異數 σ^2 為有限，並與 n 無關。利用中央極限定理，則漸進常態成立。當 $n \to \infty$ 時，$\alpha_n \to \infty$ 時，滿足李依普若條件（Liapounov's Condition）且 $\hat{\theta}_{i,n},...,\hat{\theta}_{\beta_{n,n}}$ 為獨立同分布，則可證得 $\bar{\theta}$ 具漸進常態。當 $n \to \infty$ 時，$\beta_n \to \infty$ 時 $\frac{\sum_{i=1}^{\beta_n} E\left|\hat{\theta}_{in} - \mu_n\right|^{2+\delta}}{(\sum_{i=1}^{\beta_n}\sigma_n^2)^{(2+\delta)/2}} = \frac{1}{\beta_n^{\delta/2}} E\left|\frac{\hat{\theta}_{in} - \mu_n}{\sigma_n}\right|^{2+\delta}$，亦將趨近 0，同時，(2.2) 將成立。因此滿足李依普若條件，所以 $\sqrt{\beta_n}\left(\frac{\bar{\theta} - \mu_n}{\sigma_n}\right) \to N(0,1)$，其為分佈收斂，當 $n \to \infty$ 時。

評論 1. 當 α_n 為固定且有限的值時，μ_n 與 σ_n^2 並不與 n 有關，α_n 可以 μ 與 σ 表之。如果 $\sqrt{\beta_n}\left(\frac{\bar{\theta} - \theta}{\sigma_n}\right) \to N(0,1)$ 成立，則僅有在 $\hat{\theta}_{in}$ 為 θ 不偏估計量下才能成立。若 $\hat{\theta}_{in}$ 為有偏估計時，則估計結果將不一致，因為偏誤 $\mu - \theta$ 為一常數。

評論 2. 在眾多情形下，當 $\alpha_n \to \infty$ 時，$\hat{\theta}_{in}$ 分布收斂 $\frac{\hat{\theta}_{in} - \mu_n}{\sigma_n} \to N(0,1)$，這結果使得條件(b)為一自然的假設。當 $\alpha_n \to \infty$ 時，

$$\sqrt{\beta_n}\left(\frac{\bar{\theta} - \theta}{\sigma_n}\right) \to N(0,1) \tag{2.4}$$

成立，若且為若 $\mu_n - \theta = o(\sigma_n / \sqrt{\beta_n})$ 成立。

3 母體平均數與標準差的估計

對於母體平均數與標準差的計算，就記憶體空間而言要求通常不需太大。本節中，我們感到興趣的是比較所提出的方法與傳統方法在估計結果上的差異。當我們利用對每一個子集 $x_{i1},\ldots,x_{i\alpha_n}$ 以樣本平均數去估計母體平均數時，將很容易地得到我們的方法確實與樣本平均數相等。當我們利用每一個子集的樣本變異數去估計母體變異數時，其估計式為

$$\overline{\sigma}^2 = \frac{1}{\beta_n(\alpha_n - 1)} \sum_{ij} (x_{ij} - \overline{x}_{i\cdot})^2 \ ,$$

其中 $\overline{x}_{i\cdot} = \frac{1}{\alpha_n} \sum_j x_{ij}$。依上述方法，母體變異數估計式為變異數分析模式（One-Way ANOVA）的均方誤差。雖然其為母體變異數的不偏估計，但估計結果將會損失（$\beta_n - 1$）自由度。其暗示以所提的方法進行變異數估計，有效性會比利用傳統的方法差，傳統方法對變異數的估計式為 $s_n^2 = \frac{1}{(n-1)} \sum_{ij} (x_{ij} - \overline{x})^2$。事實上，在常態假設下，其為

$$\mathrm{var}(\overline{\sigma}^2) = \frac{2\sigma^4}{n - \beta_n} \ ,$$

$$\mathrm{var}(s_n^2) = \frac{2\sigma^4}{n-1} \ 。$$

當 $\beta_n = 1$ 時，$\mathrm{var}(\overline{\sigma}^2)$ 將會最小。同時，如果 $\alpha_n = O(\sqrt{n}\log\log(n))$，如此在有效性方面的損失就會很少，因為 $\beta_n / n = O(\alpha_n^{-1}) \to 0$。

4 統計推論：信賴區間與假設檢定

如同前一章所述之方法，對參數 θ 的估計值以數學式表之為：

$$\overline{\theta} = \frac{1}{\beta_n} \sum_{i=1}^{\beta_n} \hat{\theta}_{in} \ . \qquad (4.1)$$

由於，我們欲對參數 θ 進行統計推論，故需要了解當樣本數為有限時之估計變異數。事實上，$\hat{\theta}_{in},\ldots,\hat{\theta}_{\beta_n n}$ 已提供關於 $\overline{\theta}$ 的訊息。這些訊息使得我們能對 參數 θ 建構

其信賴區間，並對一些包含θ的假設進行統計檢定。$\bar{\theta}$的標準差為$\sigma_n/\sqrt{\beta_n}$，其中σ_n可由$\hat{\theta}_{in},\ldots,\hat{\theta}_{\beta_n n}$直接估計，其估計數學式為：$\hat{\sigma}_n=\left\{\frac{1}{\beta_n-1}\sum_{i=1}^{\beta_n}\left(\hat{\theta}_{in}-\bar{\theta}\right)^2\right\}^{1/2}$。因此，對於$\bar{\theta}$的標準誤的估計表示式為：

$$\hat{SE}(\bar{\theta})=\frac{\hat{\sigma}_n}{\sqrt{\beta_n}}. \tag{4.2}$$

然而對於某些參數如分位數，其標準誤為依賴未知母體。不過，若利用 (4.2)，則允許我們避開未知母體的問題。此公式隨後將會利用電腦模擬作驗證。

如果(2.4)式漸進常態成立，則$100(1-\alpha)\%$的信賴區間將接近$\bar{\theta}\pm\Phi^{-1}(1-\alpha/2)\hat{\sigma}_n/\sqrt{\beta_n}$，其中$\Phi^{-1}(1-\alpha/2)$為標準常態$100(1-\alpha/2)\%$分位數。類似的方法，可用於檢定下述假設：

$$H_0:\theta=\theta_0 \quad \text{versus} \quad H_1:\theta\neq\theta_0 \tag{4.3}$$

其檢定統計量為

$$T=\frac{\bar{\theta}-\theta_0}{\hat{\sigma}_n/\sqrt{\beta_n}}, \tag{4.4}$$

並且其顯著水準α的拒絕區域為$\left\{|T|>\Phi^{-1}(1-\alpha/2)\right\}$。

5 母體百分位數的估計

由於母體分位數的估計需先對全部資料排序，而這樣的做法當樣本數很大時，將會耗掉龐大的記憶體空間。因此，在文獻上有些方法被提出來，用以估計龐大資料集的分位數。然而大部分的方法，集中於解決記憶體儲存空間的問題。這些文獻如，Rousseeuw and Bassett (1990)、Chao and Lin (1993)、Hurley and Modarres (1995)。本文試以比較這些被提出的方法，並發現這些方法在過程中雖減少每次運算時所需記憶體空間，但依舊沒有徹底改善總體運算時對記憶體的需求總量。然而，就另一方面而言，所提出的方法可被用於估計一些常用的參數，包含中位數與分位數等；並且，利用個人電腦處理中位數與分位數的估計亦為十分容易。在本節中，我們利用模擬來探討我們的方法，本文中模擬所用的程式碼均可在 MATLAB 軟件順利執行。

範例 5.1、本範例中，利用模擬生成 1000 個資料集，每個資料集含有 8 百萬筆來自獨立同分布的卡分方配自由度為 1 的隨機樣本。利用先前所提的方法，估計各種不同的參數。於模擬的過程中取α_n=8000，因為 8000 近似$\sqrt{n}\log\log(n)$，模擬的結果列於下表 1 中。表 1 中$\hat{\mathrm{SE}}$與std($\hat{\mathrm{SE}}$)分別為$\bar{\theta}$估計標準誤 1000 次的平均值與樣本標準差，標準誤的定義請參閱 (4.2)。同時，$\bar{\theta}$真正標準誤的估計，亦列於表 1 中其數學式表為$\mathrm{SE}_{\mathrm{true}}$。比較表 1 中最後二行，我們可以發現估計的標準誤式表現的很好。為了得到百分位數估計的標準誤，所提出的標準誤公式允許我們不去估計母體的密度。這是與傳統方法不同的，因為傳統方法通常需要事先估計未知的母體參數。

p	True value	Estimate	$\mathrm{SE}_{\mathrm{true}}(10^{-4})$	$\hat{\mathrm{SE}}(\mathrm{std}(\hat{\mathrm{SE}}))\,(10^{-4})$
0.01	1.5709×10^{-4}	1.5902×10^{-4}	0.0114	0.0112(0.0003)
0.05	3.9321×10^{-3}	3.9414×10^{-3}	0.1243	0.1219(0.0028)
0.15	3.5766×10^{-2}	3.5794×10^{-2}	0.5962	0.6093(0.0137)
0.25	0.1015	0.1016	1.3157	1.2855(0.0296)
0.35	0.2059	0.2060	2.1628	2.1269(0.0480)
0.45	0.3573	0.3574	3.2412	3.1513(0.0714)
0.50	0.4549	0.4551	3.8397	3.7506(0.0841)
0.55	0.5707	0.5708	4.5618	4.4294(0.1002)
0.65	0.8735	0.8736	6.1789	6.1121(0.1341)
0.75	1.3233	1.3236	8.8288	8.5601(0.1939)
0.85	2.0723	2.0728	13.3976	12.8548(0.2914)
0.95	3.8415	3.8436	26.5211	25.8665(0.6072)
0.99	6.6349	6.6460	63.3566	62.7383(1.4758)

表 1：百分位數的估計

6 無母數方法 -- 核密度函數估計方法

前文中所提出的方法可直接套用密度的估計，而在本章我們將利用無母數方法。其中曲線擬合法（Spline Smoothing）（可參閱：Wahba, 1990）與密度函數平滑法（Kernel Smoothing）均適用，本章採用後者密度函數平滑法為主要估計方法。利用第 i 個區集資料$x_{i1},\ldots,x_{i\alpha_n}$，其核密度估計如下：

$$\hat{f}_i(x)=\frac{1}{\alpha_n}\sum_{j=1}^{\alpha_n}K_h(x_{ij}-x)\,, \tag{6.1}$$

，其中 $K_h(z)=\frac{1}{h}K(z/h)$。$K(z)$ 為核密度估計函數，h 為事先選取的帶寬（Bandwidth）其決定密度估計函數曲線的平滑程度。本文中密度函數的選取並不是一件重要的問題，而關鍵在於帶寬的選取。在密度函數估計上，$\hat{f}_h(x)$ 的期望值與標準差如下式：

$$E\hat{f}_i(x)=f(x)+\frac{1}{2}h^2\mu_2(K)f''(x)+o(h^2)$$

與

$$Var\left\{\hat{f}_i(x)\right\}=(\alpha_n h)^{-1}R(K)f(x)+o\left\{(\alpha_n h)^{-1}\right\}.$$

其中，$\mu_2(K)=\int z^2K(z)dx$、$R(K)=\int K^2(z)dx$（可參閱 Wand and Jones,1995，page 20-21）。

因此，$\hat{f}(x;h)=\frac{1}{\beta_n}\sum_{j=1}^{\beta_n}\hat{f}_i(x;h)$ 則

$$\begin{aligned}\mathrm{MSE}(\hat{f})&=\frac{R(K)f(x)}{\alpha_n h\beta_n}+\frac{h^4}{4}\mu_2^2(K)\left\{f''(x)\right\}^2+o\left\{h^4+(\alpha_n\beta_n h)^{-1}\right\}\\&=\frac{R(K)f(x)}{nh}+\frac{h^4}{4}\mu_2^2(K)\left\{f''(x)\right\}^2+o\left\{h^4+(nh)^{-1}\right\}\end{aligned}$$

，其結果與利用所有 n 個樣本資料估計密度函數是一樣的。對上式積分結果為

$$\mathrm{MISE}=\mathrm{AMISE}\left\{\hat{f}(\cdot)\right\}^2+o\left\{h^4+(nh)^{-1}\right\}$$

，其中

$$\mathrm{AMISE}=\frac{R(k)}{nh}+\frac{h^4}{4}\mu_2^2(K)\int\left\{f''(x)\right\}^2dx.$$

如此，最理想的帶寬（Optimal Bandwidth）h_{opt} 為使得AMISE最小的 h，其數學表示式如下：

$$h_{\mathrm{opt}}=\left(\frac{R(K)}{n\mu_2^2(K)\int f''(x)dx}\right)^{1/5}$$

同時，可發現最理想的帶寬與 α_n 無關。這暗示了在第一步中估計密度函數曲線與 $h=O(\alpha_n^{-1/5})$ 比較時，為了降低其誤差，將會有低估的情形；主要因為 h_{opt} 表示式的第一項乘積會與核函數（Kernel function）及密度函數的曲率有關。同時在文獻上，其他關於帶寬選取的方法均能修正我們的目的。在一些準則下利用子集樣本

$x_{i1},\ldots,x_{i\alpha_n}$，$h^*$表為最理想的帶寬。而欲求的最理想的帶寬$h_{\text{opt}}$可表為下式：$h_{\text{opt}}=(\frac{\alpha_n}{n})^{1/5}h^*$。選取此帶寬，從理論角度而言，密度函數與利用全部資料估計結果有一樣的效果。

範例 6.1 一百萬筆獨立同分布的樣本資料其來的覆合常態分配，其分配為

$$0.5N(-2,1)+0.5N(2,1).$$

本範例欲利用其樣本資料，試著估計其密度。α_n取為 1000，核函數為利用高斯分配（Gaussian Kernel），帶寬的選取利用拇指法則（Rule of thumb / ROT，可參閱 Silverman，1986）。其帶寬的選取為

$$h_{\text{rot}}=0.9\times1.06\times\sigma\times n^{-1/5},$$

其中σ為母體標準差、1.06 為利用高斯核函數、0.9 為修正高估，細節請參閱 Silverman（1986）。在我們模擬過程，σ是可以被每區中的資料$x_{i1},\ldots,x_{i\alpha_n}$的平均絕對離差（MAD）所替代，因為其為一穩健估計值。

在模擬的過程中我們亦探討α_n對估計結果的影響。本範例中，α_n取為 500、1000、5000。圖 1 為估計結果的密度函數曲線，圖 1 中就視覺而言，均與真實密度相當接近。為了對不同α_n探討，我們定義平均均方誤差（RASE），其數學式表為

$$\text{RASE}=\left\{\frac{1}{n_{\text{grid}}}\sum_{j=1}^{n_{grid}}(\hat{f}(x_j)-f(x_j))^2\right\}^{1/2},$$

其中x_j為密度計算時的格子點，在本範例與本文中所用的n_{grid}為 400。在α_n為 500、1000、5000 下，RASE 分別為11×10^{-4}、9.12×10^{-4}及5.94×10^{-4}。由此結果可以發現當α_n越大時 RASE 的表現越佳。就另一方面而言，上述的 RASE 在不同α_n下，均為相同的階乘。這表示α_n的選取，對於估計的表現並不敏感。

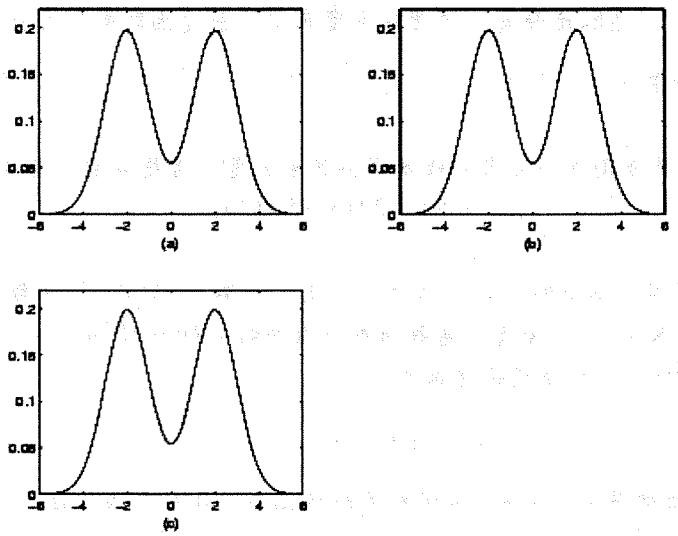

圖 1：**範例 6.1** *之覆合常態分配之估計密度函數曲線。實線圖為估計之密度函數曲線，點圖為真正的密度函數曲線。(a)* α_n *取為 500 (b)* α_n *取為 1000 (c)* α_n *取為 5000。*

7 網際網路流量資料

在本章中，我們所分析的資料為在第一章所提到網際網路流量資料。資料原始檔包含下述三個欄位【1】訊息包發送的時間（秒）【2】訊息包發送的位置【3】訊息包發送的大小（Byte；Bite；位元組）。本研究最感興趣的問題在於伺服器處理量的表現，處理量定義為【訊息包發送的大小與相鄰二訊息包發送的間隔時間的比值】。

本資料收錄了 810 萬筆非零訊息包的處理量紀錄。根據 2.2 節的方法，我們選擇α_n為 8000 該值近似$\sqrt{n}\log\log(n)$。首先，可以估計母體不同百分位數，其結果列於表 2。我們也估計了母體的密度，其圖形為圖 2。由表 2，可發現樣本中位數的標準差遠比第一及第三分位數的標準差來的大，在圖 2 中發現樣本中位數的密度較第一及第三分位數小，本圖顯示在伺服器處理量的表現上有三種典型訊息包發送情形。一為接近 0、其二有很大的處理量的表現。在表 2 中將第一、二、三分位數的處理量值乘上 8（每個位元組/Bite/Byte 有 8 個位元/bit），則處理量每秒將

為 1.8 百萬位元、5 百萬位元、8.3 百萬位元。

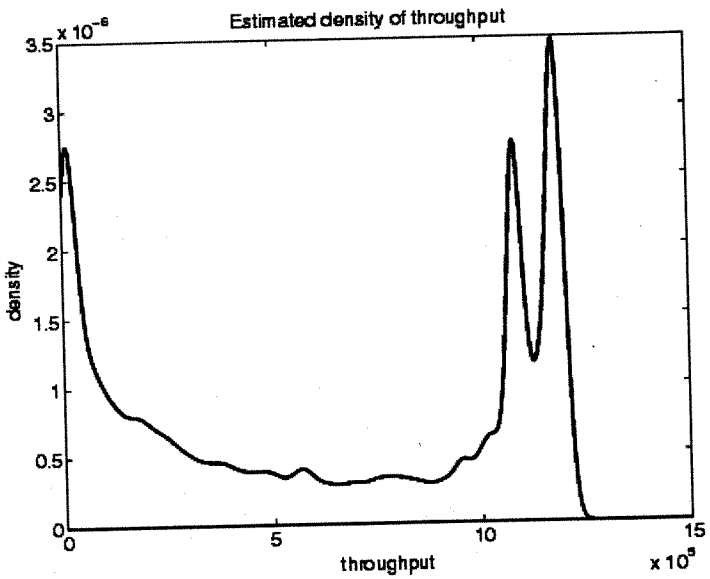

圖 2：網際網路流量資料之密度函數估計圖形

p	$\hat{\pi}_p(10^6)$	$\hat{SE}(10^3)$
0.01	0.0015	0.1308
0.05	0.0219	1.3730
0.15	0.1120	4.2115
0.25	0.2372	7.2303
0.35	0.3836	9.5022
0.45	0.5415	10.3241
0.50	0.6300	10.2400
0.55	0.7226	9.8707
0.65	0.9033	8.1293
0.75	1.0476	5.1797
0.85	1.1329	2.3094
0.95	1.1787	0.8707
0.99	1.1858	0.1689

表 2：網際網路流量資料之分位數估計

8 結論與討論

本文中，我們提出在龐大資料集下估計參數$\theta(F)$的方法。所提出的方法明顯

地降低所需計算的總記憶體空間，並且在很多情形下不失其有效性。對於估計結論的漸進理論已經被導證，而且就漸進常態的合理性亦被完整提出且建構；同時，本方法可以很容易地應用在點估計與密度函數估計上。我們提出估計值的標準誤公式，並且以經驗驗證其為合理，並以此對有興趣的參數進行統計推論，而由模擬的結果與網際網路流量資料例子說明佐證所提出的方法。

未來的工作將包含龐大資料集下相關序列的統計推論，換言之當 n 筆觀測值為已知，且為 m 個相關序列資料（可參考：Brockwell and Davis, 1991 所作之定義），θ_{in} 為二相關序列，同時 θ_{in} 僅與 $x_{i1},\ldots,x_{i\alpha n}$ 有關，而當 $\alpha_n \to \infty$ 其暗示 $\alpha_n \geq m$ 。此時，$\bar{\theta}$ 的變異數為 $Var(\bar{\theta}) = (\beta_n\sigma_n^2 + (\beta_n - 1)\rho_n) \Big/ \beta_n^2$ ，其中 ρ_n 為 $\hat{\theta}_{in}$ 與 $\hat{\theta}_{(i+1)n}$ 的相關係數。除此之外，若符合共軛條件，估計值將會有漸進常態的性質。本文為參考 Li, Lin and Li (2001) 一文。

參考文獻

1.Brockwell, P. J. and Davis, R. A. (1991). Time Series: Theory and Methods (2nd Edition), Springer-Verlag, New York.

2.Chao, M. T. and Lin, G. D. (1993). The asymptotic distributions of the remedians. Journal of Statistical Planning and Inference, 37, 1-11.

3.Chaudhuri, P. and Marron, J. S. (1999). SiZer for exploration of structures in curves, J. Amer. Statist. Assoc. , 94, 807-822.

4.Cleveland, W. S., Lin, D., and Sun, D. X. (2000). Network simulation: modeling the nonstationary and long-range dependence of client TCP connection start times under HTTP. in Proceedings of ACM SIGMETRICS'00. To appear.

5.Cleveland, W. S. and Sun, D. X. (2000). Internet traffic data. Journal of the American Statistical Association, 95, 979-985.

6.Hand, D.J., Blunt, G., Kelly, M. G. and Adams, N. M. (2000). Data Mining for Fun and Profit. Statistical Sciences,15, 111-131.

7.Hurley, C. and Modarres, R. (1995). Low-storage quantile estimation. Computational Statistics, 10, 311-325.

8.Li, Runze, Lin, Dennis K. J. and Li, Bing. (2001). Statistical Inference on Large Data Set. Department of Statistics, The Pennsylvania State University Technique Report.

9.Rousseeuw, P. J. and Bassett, G. W., Jr. (1990). The remedian: A robust averaging method for larger data sets. Journal of the American Statistical Association, 85, 97-104.

10. Silverman, B. W. (1986). Density Estimation for Statistics and Data Analysis,

Chapman and Hall, London.

11. Sheather, S. J. and Jones, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. Journal of the Royal Statistical Society, Series B., 53, 683-690.
12. Wahba, G. (1990). Spline Models for Observational Data. SIAM, Philadelphia.
13. Wand, M. P. and Jones, M. C. (1995). Kernel Smoothing, Chapman & Hall, London.