利用變異數組成模式對數量性狀基因座的貝氏統計推論

吳淑惠　蕭朱杏*

SHU-HUI WU, CHUHSING KATE HSIAO*

國立臺灣大學公共衛生學院流行病學研究所生物醫學統計組，台北市100仁愛路一段1號
Division of Biostatistics, Graduate Institute of Epidemiology, College of Public Health, National Taiwan University, No.1, Jen-Ai Rd., Sec. 1, Taipei 100, Taiwan, R.O.C.

*通訊作者Correspondence author.　E-mail: ckhsiao@ha.mc.ntu.edu.tw

目標：利用貝氏的點估計值來提供一個估計主效基因的變異數的更簡便方法；一旦估計值可被確定，所利用之標識基因可被判斷是否與主效基因有連鎖，再進一步作基因定位分析。**方法：**將數量性狀的變異數分成主效基因的變異、主效基因以外的其他遺傳變異和環境的變異等三個成份來討論，再利用貝氏方法及最大概似法對三個變異數成份作估計以進行統計推論。事後樣本的取得會利用Gibbs抽樣方法以及Markov chain Monte Carlo法。**結果：**以核心家庭有兩位子女的情況作模擬，經由結果顯示，對於主效基因的變異數估計，貝氏方法確實比最大概似估計法較穩定且精確。**結論：**相較於最大概似法，貝氏分析確實提供了精確的估計方法，而且本文利用WinBUGS1.3的程式可以獲得事後機率的樣本，可利用這些事後樣本做進一步的包含機率意義的統計推論，這是傳統統計方法所無法達到的目標。若是推廣到較複雜的遺傳模式，如主效基因不只有一個，或者改變事前分配，則此遺傳模式也可以很容易地被改寫，再利用現有統計軟體計算出結果，作為未來進一步定位之用。(台灣衛誌 2004；**23(5)**：355-364)

關鍵詞：貝氏統計推論、主效基因、數量性狀、變異數組成模式、Gibbs抽樣方法

Bayesian inference of genetic variance of QTL via variance component model

Objectives: To estimate more efficiently the variance of the major gene via Bayesian method. **Methods:** The variance component model is considered for the quantitative trait. The model comprises a polymorphic single major gene, polygene, and a random environmental effect. Two different approaches, maximum likelihood and Bayesian estimate, are compared in terms of the estimation of the three variance parameters of interest. The inference of the posterior distribution and posterior samples of the parameters, particularly the additive variance of the single major gene, are made via Markov chain Monte Carlo method using WinBUGS1.3. **Results:** Simulations are conducted to compare the performance of MLE and Bayesian estimates. **Conclusions:** The results show that the Bayesian point estimate, the posterior mode, is more accurate than the MLE. The posterior variance is also smaller than that of MLE. Compared with the conventional maximum likelihood estimation, the Bayesian approach is more flexible when a more complex genetic model is considered. (*Taiwan J Public Health. 2004;**23(5)**:355-364*)

Key Words:Bayesian, major gene, quantitative trait, variance component model, Gibbs sampling

前　言

數量性狀(quantitative trait)是指表現為連續型的性狀。以人類來說，雖然都具備相同的基本特徵，但每一個體在數量性狀上的表現型值(phenotypic value)各有不同。舉例來說，身高和體重的測量單位是以公分和公斤來表示；新陳代謝的速率通常以每分鐘葡萄糖燃燒的公克數來評估；在糖尿病方面則可能以胰島素之需求量來做為指標；這些都可以被視為是數量性狀的例子。

影響數量性狀的遺傳基因在染色體上的位置稱為數量性狀基因座(quantitative trait locus, 簡稱QTL)。至於負責的遺傳基因則可能是主效基因(major gene)、微效基因(polygene)或寡基因(oligogene)。主效基因意指主要負責的數量性狀基因座，可能有一至三個；微效基因則表示有許多個基因座負責，其單一基因的效應雖然微小但具有累加作用，因此對表現型值也會產生連續的變化；至於寡基因則比微效基因的個數少，也許是受到三個以上基因座控制。有些文獻會用QTL來表示主效基因，但本文為避免混淆起見，會將QTL和主效基因所代表的意義分別開來；另外，本文所探討的主效基因都是針對只有一個數量性狀基因座的效應，若是指二或三個的情形，則會以多個主效基因來稱之。

目前數量性狀遺傳研究的主要目的之一是找出控制遺傳疾病的基因在染色體上的位置，稱之為基因定位(gene mapping)。主要的方法是利用連鎖分析(linkage analysis)來檢定已知的性狀基因與主效基因兩者在染色體上的位置是否靠的很近。任何已知的性狀基因都可廣義的稱為標識基因(marker)，其本身的特徵容易看得見也可以被找到，例如簡單的血型或眼睛的顏色；另外如單一核苷酸多型性(single nucleotide polymorphisms，簡稱SNPs)，或是大小介於90　370bp的微衛星標記(microsatellite marker)也都常被用來作為基因體搜尋(genome search)或是基因體連鎖(linkage)的標識基因。有些遺傳研究利用這些標識基因來作相關分析(association study)，找出與疾病有相關的位置，再進一步做更高密度的偵測[1-3]；另一些遺傳研究則以連鎖分析來檢定標識基因與主效基因是否有連鎖，若有連鎖則可以利用此標識基因來進一步找到主效基因的確切位置[4,5]。

已發展的連鎖分析主要可分為兩大類，一是利用同胞對(sib-pair)的數量性狀表現型值差異的平方與標識基因的IBD (identical by descent)估計值作迴歸分析[6]，此方法主要的貢獻在於利用同胞對設計觀念來做數量性狀分析。必須具備的假設條件包含此主效基因不受其它基因座干擾、遺傳率(heritability)必須夠大、對偶基因(alleles)間無交互作用等，則兩基因的連鎖可能會使得同胞對間的IBD估計值與表現型值差異的平方有相關[7]。另一類做法是以傳統的變異數成份分析法(variance component method)將數量性狀的變異數分割成不同的成份來討論其個別貢獻[7, 8]，本文也採用這個統計模式，假設數量性狀同時受到主效基因、以及主效基因以外的遺傳效應和環境因素的控制[8]。過去的文獻利用傳統線性模式的推論方法來檢定主效基因的變異數是否存在，利用標識基因進行測定，針對標識基因座來估計未知之主效基因的IBD，再進一步檢定該主效基因變異數是否存在；若是存在則代表該標識基因與主效基因有連鎖，可進一步在標識基因附近做更小範圍的搜尋，以定位主效基因。然而，這種方法會受制於變異數估計的好壞，也不能提供具機率概念的描述，在資料型態較複雜時(如家族大小差距大或親等較多時)，也不容易對主效基因的變異數得到合理的推論。因此，本文希望利用貝氏統計的優點以及貝氏統計計算的方便性，來對主效基因的變異數做推論。本文主要針對單一主效基因的遺傳模式來作為討論的對象，希望以此一簡單的模式為例，說明並提供一個貝氏的統計方法。

投稿日期：92年9月1日
接受日期：93年4月9日

材料與方法

數量性狀之變異數組成模式

本文所探討的數量性狀以混合模式(mixed model)來建構，亦即遺傳模式中包含主效基因和微效基因。假設I個家庭中各有n_i位成員，y_{ij}代表第i個家庭第j位成員的數量性狀表現型值，利用混合模式來建構數量性狀表現型值為

$$y_{ij}=\mu+q_{ij}+g_{ij}+e_{ij}\quad i=1,\ 2,\ \cdots I;\ j=1,\ 2,\ \cdots n_i$$

其中，μ代表整體的平均值，q_{ij}代表第i個家庭第j位成員的主效基因對於表現型值的效應(major gene effect)；g_{ij}代表第i個家庭第j位成員的殘留遺傳效應 (residual genetic effects)，也就是主效基因以外的其他遺傳效應；e_{ij}代表第i個家庭第j位成員的環境隨機效應(random environmental effect)。

若全部I個家庭的數量性狀以$\boldsymbol{y}$向量表示，則全體資料可表示為$\boldsymbol{y}=(y_1,y_2,...y_I)^t{}_{i\times 1}$，其中，第$i$個家庭$n_i$位成員的資料以矩陣型式呈現為

$$\begin{pmatrix} y_{i1}\\ y_{i2}\\ \vdots\\ y_{in_i}\end{pmatrix}=\mu\begin{pmatrix}1\\1\\\vdots\\1\end{pmatrix}_{n_i\times 1}+\begin{pmatrix} q_{i1}\\ q_{i2}\\ \vdots\\ q_{in_i}\end{pmatrix}_{n_i\times 1}+\begin{pmatrix} g_{i1}\\ g_{i2}\\ \vdots\\ g_{in_i}\end{pmatrix}_{n_i\times 1}+\begin{pmatrix} e_{i1}\\ e_{i2}\\ \vdots\\ e_{in_i}\end{pmatrix}_{n_i\times 1}$$

即

$$\boldsymbol{y_i}=\boldsymbol{\mu}+\boldsymbol{q_i}+\boldsymbol{g_i}+\boldsymbol{e_i}\quad i=1,\ 2,\ \cdots I$$

其中主效基因$\boldsymbol{q_i}$的期望值為$\boldsymbol{0}$，變異數為$\hat{\boldsymbol{\Pi}}_i\sigma_q^2$；其他主效或微效基因效應為$\boldsymbol{g}_i$；至於$\boldsymbol{e}_i$則代表環境的隨機效應。

假設家庭內的數量性狀表現型值服從多變量常態分配，表示為$\mathbf{y}_i|\sigma_q^2,\sigma_g^2,\sigma_e^2 \sim MVN(\boldsymbol{\mu},\sum_i)$，則第$i$個家庭的期望值為$\boldsymbol{\mu}=(\mu,\ \mu,\ \cdots\mu)'_{n_i\times 1}$，共變異數矩陣記作$\Sigma_i=\hat{\boldsymbol{\Pi}}_i\sigma_q^2+\mathbf{R}_i\sigma_g^2+\mathbf{I}_i\sigma_e^2$，此為變異數組成模式(variance component model)。

本文先考慮最簡單的情形是主效基因座上有兩個對偶基因A和a，令r和s分別為A和a的基因頻率(gene frequency)，且$r+s=1$，則產生三種可能的基因型組合分別為AA、Aa和aa。在哈溫平衡 (Hardy-Weinberg equilibrium)條件下，基因型頻率分別為r^2、$2rs$和s^2，對應各基因型的基因型值分別為t、0和$-t$，在此是假設主效基因的遺傳效應只含累加性效應(additive effect)，而顯性效應(dominance effect)設定為零。假設基因之間或基因與環境之間互不影響，亦即q、g和e是不相關的隨機變數，期望值皆為零。若變異數σ_q^2表示來自於主效基因的變異，由前述所假設的基因型頻率和基因型值可求得母體平均值為$t(r-s)$，再將基因型值表示成偏離母體平均值，分別為$2ts$、$-t(r-s)$、$-2tr$，利用平移後的基因型值就可以得知$\sigma_q^2=2t^2rs$。變異數σ_g^2表示來自於主效基因以外的其他遺傳變異；σ_e^2為環境的變異。根據上述定義，則第i個家庭的資料為$\boldsymbol{y}_i=\boldsymbol{\mu}+\boldsymbol{q}_i+\boldsymbol{g}_i+\boldsymbol{e}_i$，其中$\boldsymbol{q}_i$的期望值為$\boldsymbol{0}$，變異數為$\hat{\boldsymbol{\Pi}}_i\sigma_q^2$。此處$\boldsymbol{g}_i$的代表除了主效基因以外的其他遺傳效應，可能是微效基因效應或是其他的主效基因效應；因此，$\boldsymbol{g}_i$服從多變量常態分配 (期望值為零向量，變異數矩陣為$\mathbf{R}_i\sigma_q^2$)應是合理的假設。至於環境的隨機效應，一般而言，也假設$\boldsymbol{e}_i$服從多變量常態分配(期望值為零向量，變異數矩陣為$\mathbf{I}_i\sigma_q^2$)。利用三個變異數的係數來描述家庭內成員之間的相似性，其中$\hat{\boldsymbol{\Pi}}_i$表示標識基因座的IBD比例估計值矩陣(IBD proportion estimated matrix)；$\mathbf{R}_i$表示親屬關係係數矩陣(coefficient of relationship matrix)[9]；$\mathbf{I}_i$則代表單位矩陣(identity matrix)，主對角線上的元素為1。以下會將數量性狀的變異數分成σ_q^2、σ_g^2和σ_e^2三個成份來討論，這三者也就是所謂的變異數成份。在檢定σ_q^2是否為零的情況下，即同義於檢定連鎖是否存在。

貝氏方法的參數估計

貝氏統計的想法認為未知的參數含有不確定性(uncertainty)，必須利用一個機率分配來描述，此機率分配稱為事前分配 (prior distribution)，再結合目前收集的資料以產生事後分配(posterior distribution)，藉由事後分配來更新對此未知參數的了解；貝氏推論

(Bayesian inference)是建立在事後分配上，其定義為

$$P(\theta|y) = \frac{f(y|\theta)\pi(\theta)}{\int f(y|\theta)\pi(\theta)d\theta}$$

其中分子的部分為概似函數乘以事前分配，分母的部分稱為y的邊際機率。以本文來說，概似函數為多變量常態分配連乘，接下來對於概似函數的符號，都以 $L(\theta|y)$ 表示，其中有興趣的參數 θ 為 σ_q^2、σ_g^2 和 σ_e^2 三個相互獨立的變異數；利用Jefferys' prior的想法[10]，假設其事前分配分別服從 IG(α_k， β_k)分配(inverted gamma distribution)，記作

$$\pi(\theta)=\pi_1(\sigma_q^2)\pi_2(\sigma_g^2)\pi_3(\sigma_e^2) \sim \prod_k IG(\alpha_k,\beta_k) \quad k=1, 2, 3$$

由於IG 分配中隨機變數的定義範圍與 σ_q^2、σ_g^2 和 σ_e^2 的值域符合，藉由*IG*分配中參數(α， β)的改變來呈現不同形式的分佈狀況，也可以有不同的期望值和變異數。

在觀察到所有資料之下，事後分配可寫為

$$p(\sigma_q^2,\sigma_g^2,\sigma_e^2|\mathbf{y}) = \frac{L(\sigma_q^2,\sigma_g^2,\sigma_e^2|\mathbf{y})\pi_1(\sigma_q^2)\pi_2(\sigma_g^2)\pi_3(\sigma_e^2)}{\iiint L(\sigma_q^2,\sigma_g^2,\sigma_e^2|\mathbf{y})\pi_1(\sigma_q^2)\pi_2(\sigma_g^2)\pi_3(\sigma_e^2)d\sigma_q^2 d\sigma_g^2 d\sigma_e^2}$$

上式中分母的積分並沒有明確的解，因此，將採用Markov chain Monte Carlo (MCMC)方法中常用的Gibbs sampling來生成事後分配的樣本來做推論。藉由Gibbs抽樣方法只需要建立所有參數的條件機率即可，細節可見附錄一。在以下的貝氏推論中，都會使用Gibbs抽樣方法以及統計軟體WinBUGS 1.3 [11]求得事後樣本，再據以求得眾數(mode)作為點估計值。

最大概似法

本文另一種作為比較用的參數估計方式為最大概似法，先將概似函數取為自然對數 log*L*，再利用此自然對數的一階偏微分代入牛頓法進行迭代求解，直到收斂為止，這樣就可以得到最大概似估計值，其它的細節可在附錄一找到。

結　果

家庭資料之模擬

以核心家庭有兩位子女 (不含同卵雙胞胎)的情況作資料模擬，並且比較貝氏方法和最大概似法的參數估計效果。假設感興趣的遺傳疾病其主效基因位在體染色體上但是觀察不到，假定有一標識基因和主效基因有連鎖，即在染色體位置上靠得夠近，近到其中間沒有發生交換，則此標記基因的有些遺傳行為應可推估此主效基因的遺傳行為[8]。標識基因座內含兩個對偶基因A和a，假設對偶基因的基因頻率可以事先得知，分別為 $r=p(A)=0.75$ 和 $s=p(a)=0.25$。有三種可能的基因型組合為AA、Aa、aa，在沒有顯性效應的存在下，基因型值分別為t、0和-t。

本文以50個家庭為考量，每個家庭有四位成員，父母及兩位子女。先前已假設每個家庭內的數量性狀表現型值服從多變量常態分配，表示為 $\mathbf{y}_i|\sigma_q^2,\sigma_g^2,\sigma_e^2 \sim MVN(\boldsymbol{\mu},\boldsymbol{\Sigma})$，由於期望值 $\boldsymbol{\mu}$ 不是本文主要探討的參數而被視為是干擾參數(nuisance parameter)，所以假設每個家庭的期望值為已知的固定數值，以 $\boldsymbol{\mu}=(0,0,0,0)'_{4\times1}$ 表示；這個假設使得實際上在估計真正感興趣的參數時計算較為簡單。

另外，由附錄一可知共變異數矩陣 $\boldsymbol{\Sigma}_i = \hat{\boldsymbol{\Pi}}_i\sigma_q^2 + \mathbf{R}_i\sigma_g^2 + \mathbf{I}_i\sigma_e^2$ 中，必須先清楚地定義 $\hat{\boldsymbol{\Pi}}_i$、$\mathbf{R}_i$ 及 $\mathbf{I}_i$。其中 $\mathbf{R}_i$ 根據Crow and Kimura [9]可得知，親代與子代間的親屬關係係數以及同胞對之間的親屬關係係數皆為1/2；$\mathbf{I}_i$ 為對角線上元素為1的單位矩陣；複雜的是標識基因的IBD比例估計值矩陣 $\hat{\boldsymbol{\Pi}}_i$，為了簡化模式以進行對主要參數的估計，引用Haseman and Elston [6]中的表格二，在父母親的基因型已知的情形下，將家庭內兩兩成員間的IBD比例估計值列於表一，其中第一行代表親代的六種交配型(mating type)；每一種交配型所對應的交配型頻率(邊際機率)列於第二行，總和為1。在每一種交配型下所產生的子代基因型如第三行所示；子代基因型所對應的子代基因型頻率(條件機率)列於第四行。將邊際

表一　引用Haseman & Elston (1972)的表二來計算親屬間的IBD比例估計值

Mating type	Marginal probability	Sib pair type	Conditional probability	Joint probability	Joint probablity	Condition on parents and sib pair I_0	I_1	I_2	
$AA \times AA$	r^4	$AA-AA$	1	r^4	0.32				
$AA \times aa$	$2r^2s^2$	$Aa-Aa$	1	$2r^2s^2$	0.07				
$AA \times Aa$	$4r^3s$	$AA-AA$		r^3s	0.105	0			
		$AA-Aa$		$2r^3s$	0.21			0	
		$Aa-Aa$		r^3s	0.105	0			
$Aa \times Aa$	$4r^2s^2$	$AA-AA$ or $aa-aa$			0.0175	0	0	1	1
		$AA-aa$			0.0175	1	0	0	0
		$AA-Aa$ or $aa-Aa$		$2r^2s^2$	0.07	0	1	0	
		$Aa-Aa$		r^2s^2	0.035		0		
$Aa \times aa$	$4rs^3$	$Aa-aa$		$2rs^3$	0.025			0	
		$aa-aa$		rs^3	0.0125	0			
		$Aa-Aa$		rs^3	0.0125	0			
$aa \times aa$	s^4	$aa-aa$	1	s^4	0.0039				
Total	1								

機率乘以條件機率的聯合機率如第五行所示。由於對偶基因的基因頻率已事先假設為r＝0.75和s＝0.25，代入第五行可以求得聯合機率值列於第六行。子代之間IBD數目等於0、1、2的機率，以f_0、f_1、f_2表示，列於第七、八、九行。而子代之間的IBD比例估計值，以$\hat{\pi}_{i,sib}$表示於第十行；算法是先將IBD數目0、1、2調整為0、1/2、2/2的IBD比例(proportion)，再乘以其對應的機率，即$\hat{\pi}_{i,sib}$＝1/2×f_1＋f_2。第十一行代表親代和子代之間的IBD比例估計值，以$\pi_{i,ps}$表示。由於親代和子代之間的IBD數目只有等於1的情況，因此$\pi_{i,ps}$＝1/2。最後，再將第十行和第十一行彙整成五種情況。因此，模擬的50個家庭就會隨著五種親屬間的IBD比例估計值的機率而有不同的$\hat{\boldsymbol{\Pi}}_i$，隨著不同的$\hat{\boldsymbol{\Pi}}_i$而有不同的共變異數矩陣。至此，數量性狀分配中的平均數和變異數的係數定義完成。

最後，給定三個變異數真值就可進行家庭資料模擬。由於先前有計算主效基因的變異σ_q^2＝2t^2rs，其中r和s已知為0.75和0.25，因此，σ_q^2為t的函數，若t為1、$\sqrt{2}$、$\sqrt{3}$和$2\sqrt{3}$，則σ_q^2為0.375、0.75、1.125和2.5。由於模擬的目的在比較三個參數於不同的真值之下，估計方法的準確性，又σ_q^2為主要探討的參數，因此，最後以參數間不同的相對大小關係來選擇四組真值，並利用σ_q^2／(σ_q^2＋σ_g^2＋σ_e^2)來計算數量性狀的遺傳率(heritability, h^2)，如表二所示，在每一組真值下，利用統計軟體 SAS 6.12從多變量常態分配中隨機抽取 20 筆資料，假設每筆資料中有 50 個家庭，即隨機生成 1000 個家庭。

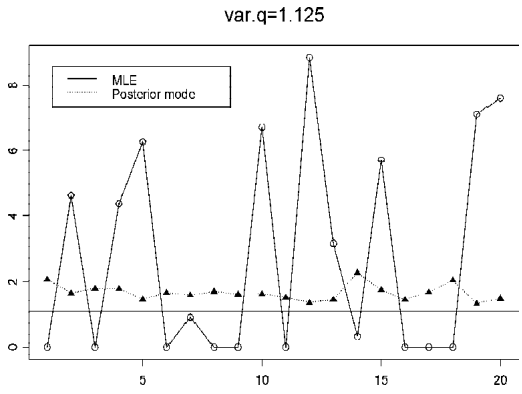
參數估計

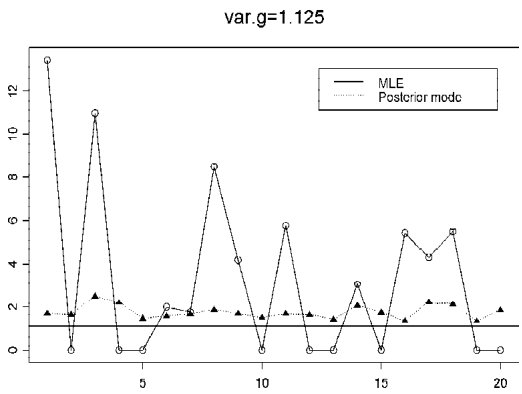
利用模擬的家庭資料來估計變異數組成模式中的三個未知參數σ_q^2、σ_g^2及σ_e^2。方法之一為最大概似法，藉由牛頓法來逼近參數的MLE，以S-PLUS完成這些複雜的迭代運算；方法之二為貝氏方法，利用統計軟體WinBUGS1.3 (windows version of BUGS) [11]，在給定概似函數及參數的事前分配下，即可模擬生成事後機率樣本，再利用事後機率樣本的機率分配來找出發生機率最高的參數值，稱為posterior mode。假設σ_q^2、σ_g^2及σ_e^2的事前分配分別服從IG(α，β)分配，並選取α＝1且β＝6，使得分配的形狀較扁平，也就是對事後分配的影響較微小。最後，根據表二在四組不同真值設定下，模擬20筆多變量常態分配的資料，且每筆資料包含50個家庭，則每筆資料的MLE及posterior mode的估計結果統整成圖，但此處只列出前兩種情形於圖一至圖二，後面兩種情況的結果類似，故不特別列出。另外，若MLE收斂至負值，圖形中就以零替代。

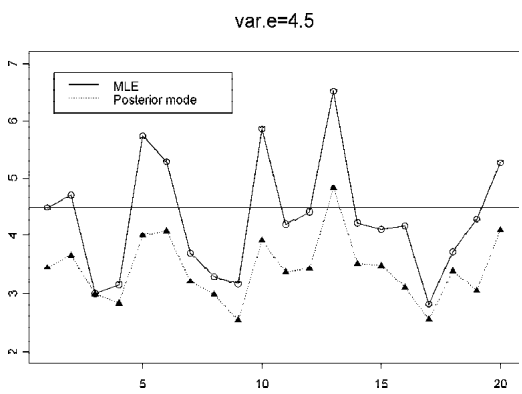
整體而言，在不同真值的設定下，貝氏方法都比最大概似法估計較穩定。從表的數據中也顯示MLE的標準差比事後機率的標準差大，也就是貝氏方法的估計較接近真值。從圖形可以看出對於σ_q^2和σ_g^2的估計，貝氏方法不但估計較穩定且較接近真值。對於σ_e^2的估計，兩種方法都不錯。在估計σ_q^2和σ_g^2時，MLE出現負的頻率很高。例如在真值設定為σ_q^2＝1.125、σ_g^2＝1.125之下，20次的估計結果中有9次MLE為負值，根據過去文獻可知這是在變異數值較小的情況下，估計值容易落到邊界以外之故。

表二　三個參數的真值及遺傳率

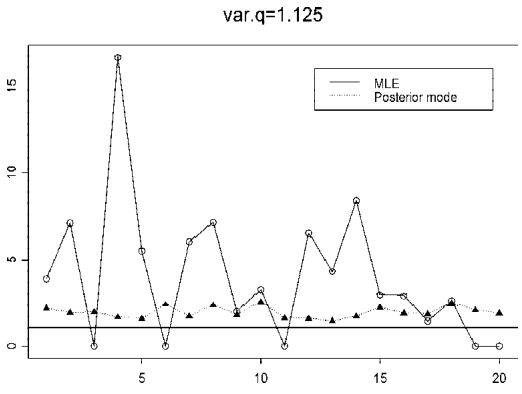
h^2	σ_q^2	σ_g^2	σ_e^2
16.7%	1.125	1.125	4.5
14.3%	1.125	2.5	4.5
11.1%	1.125	4.5	4.5
24.1%	2.5	4.5	3.375

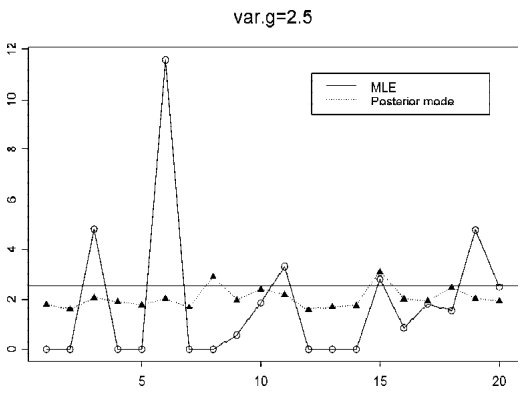


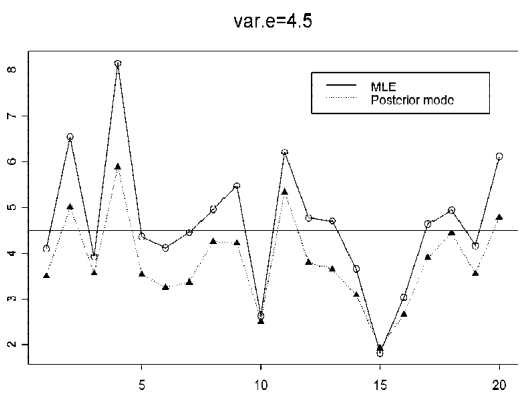




圖一　20筆資料各50個家庭在真值為σ_q^2＝1.125、σ_g^2＝1.125、σ_e^2＝4.5的估計結果。Y軸表該變異數之估計值。







圖二　20筆資料各50個家庭在真值為σ_q^2＝1.125、σ_g^2＝2.5、σ_e^2＝4.5的估計結果。Y軸表該變異數之估計值。

討　論

本文的主要目的在提供一個方法，可以將主效基因的變異數估計的更準確。相較於最大概似法，貝氏分析確實提供了精確的估計方法，而且本文利用WinBUGS1.3的程式可以獲得事後機率的樣本。若是推廣遺傳模式，如主效基因不只有一個，或者改變事前分配，則此WinBUGS1.3的程式也可以很容易地被改寫，亦或是利用前面所列出的所有條件機率函數，以及Gibbs抽樣方法或其它MCMC的方法來得到事後機率樣本以進行統計推論。統計方法的簡便與穩定性可以提高推論結果的可信度；例如，此處在得出主效基因變異數的事後分配之後，若其大於零的機率很高，代表有連鎖，則研究者會有較高的信心在標識基因的附近做進一步較小區域的搜尋或基因定位。

為了比較兩種估計方法，本文利用模擬來檢視其估計的效果。經由模擬結果顯示，在不同真值的選擇下，貝氏方法都比最大概似法估計較穩定且較接近真值。這是因為貝氏分析中，除了概似函數的訊息，會對欲探討的參數利用事前分配來描述，使得事後分配受到事前分配和概似函數影響，估計結果會較穩定。另外，貝氏方法所能提供的不只有點估計而已，在這裡我們可以利用 Gibbs sampling所得到的樣本求得某區間分布的機率，例如該主效基因變異數大於零的機率或是該變異數落在某區間的機率；這些都是MLE無法做到，屬於貝氏方法的優點。

以貝氏分析而言，參數的事前分配選擇是一門學問[12]，由於遺傳變異和非遺傳變異存在一些不確定性的影響，使得事前很難對參數有夠多的信念，所以，參考過去研究中探討關於常態分配中變異數的事前分配，採用Jefferys' prior的想法為Inverted Gamma分配，在模擬中給定三個參數的事前分配皆為*IG*(1,6)，若考量三個參數的性質不同時，應該要改變*IG*分配中的參數來呈現不同形式的分佈狀況。以最大概似估計法而言，由於估計的三個參數皆為變異數且可能落在參數空間的臨界處(boundary)，使得牛頓法的迭代運算會收斂至負值，從某些模擬的結果看來，遺傳效應的變異數出現負值的頻率很高，這種情況在傳統統計有關變異數估計中認為是因為真值不夠大，也只建議以零值代替。傳統統計的文獻之中就已經發現變異數成份的MLE估計不穩定，因此本文不以MLE的估計法為主軸，只是拿它來比較而已，未來亦可針對MLE再做進一步的研究。另外，q與g的變異數之和可能會因為是常態分配之和(仍為常態)且為同一類(連續)的變異，而有穩定的估計值，但卻不能解決目前以q的變異數為推論對象的問題。因此，我們正在進行的另外一項研究是以貝氏方法和restricted MLE比較，並且將q與g的變異數差距拉大，以比較二者的表現。

再者，目前的遺傳研究著重在疾病基因的探討，由於一些慢性疾病會同時受到主效基因和微效基因所控制，因此，本文以混合模式來建構數量性狀表現型值。在簡化模式的考量下，主效基因只含累加性效應，亦即顯性效應的部分為零，此假設的確可以減少參數的維度，但忽略這部份的遺傳訊息可能無法反應真實的情況[13]。再者，假設主效基因座帶有兩個對偶基因，在哈溫平衡條件下，只有三種基因型，則主效基因屬於多項分配(multinomial)，導致數量性狀的分布可能違反常態分配的假設，以致於在進行統計推論時，產生較大的型一誤差。關於這點，可以參考Blangero等人[14]所提出的robust LOD score的估計法；另外，Amos [8]利用Generalized estimating equations (GEEs)的方法只需要設定一階及二階動差，不受限於數量性狀服從多變量常態分配的假設。若是考慮一個主效基因含多個對偶基因或是有多個主效基因會影響數量性狀，則主效基因服從常態分配會是合理的假設。不過，儘管如此，在多個主效基因影響數量性狀的情形下，要考慮比較多的參數而且遺傳模式會複雜許多。最後，在共變異數矩陣中用來描述主效基因的變異是觀察標識基因座得來的IBD比例估計值矩陣，除了引用Haseman and Elston [6]中的表格二歸納出五種情形，亦可以參考 Almasy and Blangero [15]的建議，或是在模式

中納入互換率並估計之，或以動差(moments)來估計；這幾點皆可供未來繼續研究之參考。

致　謝

第一作者感謝在完成論文以及投稿期間，鼓勵和支持的師長、同學和家人；也謝謝公共衛生學會對於這篇論文的鼓勵。

參考文獻

1.Carlson CS, Eberle MA, Rieder MJ, Smith JD, Kruglyak L, Nickerson DA. Addition SNPs and linkage-disequilibrium analyses are necessary for whole-fenome association studies in human. Nature Genetics 2003;**33**:518-21.

2.Devlin B, Roeder K. Genomic control for association studies. Biometrics 1999;**55**:97-1004.

3.Risch N. Searching for genetic determinants in the new millennium. Nature 2000;**405**:847-56.

4.Botstein D, Risch N. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. Nature Genetics supplement 2003;**33**:228-37.

5.The International SNP Map Working Group. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature 2001;**409**:928-33.

6.Haseman JK, Elston RC. The investigation of linkage between a quantitative trait and a marker locus. Behav Genet 1972;**2**:3-19.

7. 戴政：遺傳流行病學。台北：藝軒圖書出版社，2002；12-20。

8.Amos CI. Robust variance-components approach for assessing genetic linkage in pedigrees. Am J Hum Genet 1994;**54**:535-43.

9.Crow JF, Kimura M. An Introduction to Population Genetics Theory. New York: Harper and Row, 1970;138.

10.Lee PM. Bayesian Statistics. Oxford University Press, 1989:54-5.

11.Spiegelhalter D, Thomas A, Best N. WinBUGS: Windows Version of BUGS-Bayesian Inference Using Gibbs Sampling. version 1.3. Cambridge: MRC biostatistics unit, 2000.

12.Kass RE, Wasserman L. The selection of prior distributions by formal rules. J Am Stat Assoc 1996;**91**:1343-70.

13.Abney M, McPeek MS, Ober C. Estimation of variance components of quantitative traits in inbred populations. Am J Hum Genet 2000;**66**:629-50.

14.Blangero J, Williams JT, Almasy L. Robust lod scores for variance component-based linkage analysis. Genet Epidemiol 2000;**19 (suppl 1)**:S8-14.

15.Almasy L, Blangero J. Multipoint quantitative-trait linkage analysis in general pedigrees. Am J Hum Genet 1998;**62**:1198-211.

16.Tierney L, Kadane JB. Accurate approximation for posterior moments and marginal densities. J Am Stat Assoc 1986;**81**:82-6.

17.Searle SR, Casella G, McCulloch CE. Variance Components. New York: John Wiley and Sons, 1992.

18.Lange K. Mathematical and Statistical Methods for Genetic Analysis. New York: Springe, 1997.

附錄一

本文的概似函數為多變量常態分配連乘，若使用Gibbs抽樣方法只需要建立所有參數的條件機率 (full set of conditional probabilities) 即可使用，亦即

$P\left(\sigma_q^2 \middle| \mathbf{y}, \sigma_g^2, \sigma_e^2\right)$，$P\left(\sigma_g^2 \middle| \sigma_q^2, \mathbf{y}, \sigma_e^2\right)$，與$P\left(\sigma_e^2 \middle| \sigma_q^2, \mathbf{y}, \sigma_g^2\right)$

這些條件機率可藉由拉普拉斯法(Laplace's method)[16]得到近似值。此三個參數的條件機率得知後，藉由迭代抽樣來得到一組事後機率的樣本且此樣本就相當於來自三個參數的聯合機率分配$P\left(\sigma_q^2 \middle| \mathbf{y}, \sigma_g^2, \sigma_e^2\right)$。

在最大概似法中，先將概數函數取自然對數為

$$\log L = -\frac{N}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{I}\log|\Sigma_i| - \frac{1}{2}\sum_{i=1}^{I}\left[(\mathbf{y}_i-\mu)^t\sum_i^{-1}(\mathbf{y}_i-\mu)\right]$$

其中N為全部的觀察值個數，$\boldsymbol{y}_i$和μ皆為$n_i \times 1$的矩陣，Σ_i為$n_i \times n_i$的矩陣。則對數概似函數log$L$對參數的一階微分式[17,18]如下

$$\frac{\partial \log L}{\partial \sigma_q^2} = -\frac{1}{2}\sum_{i=1}^{I} tr\left(\Sigma_i^{-1}\hat{\Pi}_i\right) + \frac{1}{2}\sum_{i=1}^{I}\left[(\mathbf{y}_i-\mu)^t\Sigma_i^{-1}\hat{\Pi}_i\Sigma_i^{-1}(\mathbf{y}_i-\mu)\right]$$

$$\frac{\partial \log L}{\partial \sigma_g^2} = -\frac{1}{2}\sum_{i=1}^{I} tr\left(\Sigma_i^{-1}\mathbf{R}_i\right) + \frac{1}{2}\sum_{i=1}^{I}\left[(\mathbf{y}_i-\mu)^t\Sigma_i^{-1}\mathbf{R}_i\Sigma_i^{-1}(\mathbf{y}_i-\mu)\right]$$

$$\frac{\partial \log L}{\partial \sigma_e^2} = -\frac{1}{2}\sum_{i=1}^{I} tr\left(\Sigma_i^{-1}\right) + \frac{1}{2}\sum_{i=1}^{I}\left[(\mathbf{y}_i-\mu)^t\Sigma_i^{-1}\Sigma_i^{-1}(\mathbf{y}_i-\mu)\right]$$

若**U**表示概似函數的一階微分矩陣，則令**U**＝0，可以求得最大概似估計值(MLE)。但是多維度常態分配的對數概似函數微分後的型式較複雜，在這裡無法找到概似函數確切的解 (closed-form solution)，可藉助數值分析方法中的牛頓法(Newton-Raphson algorithm)，利用迭代方式來求解，直到收斂為止，如此就可獲得此三個參數的最大概似估計值。