

國立交通大學

統計學研究所

碩士論文

隱馬可夫模型在發育過程上之應用

APPLICATION OF HIDDEN MARKOVIAN MODEL TO
TANNER STAGES

研究生：柯欣妤

指導教授：彭南夫 博士

中華民國九十三年六月

隱馬可夫模型在發育過程上之應用

APPLICATION OF HIDDEN MARKOVIAN MODEL TO TANNER STAGES

研究生：柯欣妤

Student : Hsin-Yu Ko

指導教授：彭南夫 博士

Advisor : Dr. Nan-Fu Peng

國立交通大學

統計學研究所



Submitted to Institute Statistics

College of Science

National Chiao Tung University

in partial Fulfillment of the Requirements

for the Degree of

Master

in

Statistics

June 2004

Hsin-chu, Taiwan, Republic of China

中華民國九十三年六月

隱馬可夫模型在發育過程上之應用

研究生：柯欣妤

指導教授：彭南夫 博士

國立交通大學統計學研究所



本論文是探討非裔美國青少年發育過程之變化，由於 Tanner 階段普遍被用來評估發育過程的階段，但是，人類成長發育是連續性的且 Tanner 階段是帶有本身變異的次序出象，則誤判至鄰近階段是很有可能發生的，所以，在本論文中，我們特別考量用隱馬可夫模型來分析，以便討論非裔青少年在發育階段所需持續之時間及評估每階段的轉移機率。


APPLICATION OF HIDDEN MARKOVIAN MODEL TO TANNER STAGES

student : Hsin-Yu Ko

Advisors : Dr. Nan-Fu Peng

*Institute of Statistics
National Chiao Tung University*

ABSTRACT



This thesis discusses the changes in adolescent African American boys and girls. Tanner stage is commonly used for assessment of the status of sexual development. Since human growth is continuous and Tanner stage has an ordinal outcome with self-rating, it is possible to rate to adjacent stages. Thus, this thesis specially considers Hidden Markov Model to analyze. It is used for determining durations of Tanner stages among adolescent African American children and rating the transition probability.

誌 謝

在統計所這兩年求學生涯中，不論在專業知識之學習或生活經驗之累積，皆使我獲益良多。本論文的完成，首先要感謝我的指導教授 彭南夫老師的悉心指導，以培養我獨立思考之研究能力，也感謝口試委員陳玉英教授、洪慧念教授以及王鴻龍教授的參與指教，使我將論文修正的更加清楚完整。

在生活上，要感謝朋友嘉佑及宇書給予我的鼓勵與關懷，也感謝同學怡均、巧慧、淑珍、忠庭、文祥、慶富、超毅、志浩、政輝，在我學習研究過程中，給予我協助及關心，以及閒暇之餘一起打球運動，感謝你們豐富了這兩年的生活，留下了深刻美好的回憶。

最後，要感激父母親及家人這麼多年來栽培與支持，你們的照顧及關心，讓我能在毫無顧慮下順利的完成學業。在此，將以本篇論文獻給曾經給我鼓勵協助的家人、師長、朋友以及同學們，並致上我最誠摯的謝意。

欣好 謹誌于

國立交通大學統計學研究所
中華民國九十三年六月

目錄

中文摘要	I
英文摘要	II
誌謝	III
目錄	IV
圖、表目錄	V
第一章、導論	1
1-1 前言	1
1-2 研究目的	1
1-3 研究方法	2
第二章、資料部份	3
2-1 資料的背景及意義	3
2-2 分析設計及方法	3
2-3 資料上的問題及處理方式	6
第三章、理論部份	8
3-1 隱馬可夫模型	8
3-2 引理之說明	10
3-3 <i>Baum-Welch</i> 提出之參數估計法	15
第四章、演算分析及結果部份	17
4-1 演算之流程	17
4-2 演算之結果分析	18
4-3 改進的方法	21
參考文獻	31

圖、表目錄

圖 1-1、 <i>Tanner Staging</i> 之圖示	24
圖 4-1、男孩_第一階段分配圖	27
圖 4-2、男孩_第二階段分配圖	27
圖 4-3、男孩_第三階段分配圖	28
圖 4-4、男孩_第四階段分配圖	28
圖 4-5、女孩_第一階段分配圖	29
圖 4-6、女孩_第二階段分配圖	29
圖 4-7、女孩_第三階段分配圖	30
圖 4-8、女孩_第四階段分配圖	30
表 1-1、 <i>Tanner Staging</i> 之定義	23
表 4-1、25 位非裔男孩的參數估計結果	25
表 4-2、23 位非裔女孩的參數估計結果	26

第一章、導論：

❖ 1-1 前言：

在醫學上，對於青春期男孩及女孩發育至成熟過程方面，較早的研究已經證實非洲後裔之美國男孩及女孩（簡稱非裔；*African American boys and girls*）的發育成熟速度比歐洲後裔之美國男孩及女孩（簡稱歐裔；*European American boys and girls*）或是其他種族背景之男孩及女孩來得快。然而，先前的研究都只有著重非裔男孩及女孩的遺傳基因導致發育速度較其他種族快，卻沒有考量非遺傳基因變異或是飲食習慣不同是否會對發育過程產生影響，因此，我們想直接針對非裔男孩及女孩做發育過程的調查，並利用統計方法來分析某一發育階段到另一階段的轉變差異及評估每階段所需的時間。



❖ 1-2 研究目的：

醫學上，將青春期發育的過程分為五個階段，稱為 *Tanner Stage*，這是專指在發育過程中陰部及胸部的發育情況（有關 *Tanner Stage* 的解釋及定義，請詳見[表 1-1]及[圖 1-1]）和非遺傳基因影響的改變。針對青春期的非裔男孩及女孩而言，我們評估出與發育過程有關的非基因影響的因素諸如：營養的攝取、生理的活動及危害健康的行爲，這些因素我們將會加以調查及記錄。

Tanner Stage 的觀測值是個帶有本身評估差異的順序出象，儘管，*Tanner Stage*

普遍地被用於記錄人類發育的階段，但在統計上，對於 *Tanner Stage* 卻被視為一個複雜的出象變異，這是因為在正常情形下，人類的成長發育是連續性的，只會向前發展而不會退化，*Tanner Stage* 卻把人類的發育劃分成離散的五個階段，況且 *Tanner Stage* 不是經由儀器判斷處於發育某階段，而是由醫護人員以肉眼去判斷，所以，不小心被評斷為鄰近階段的可能性是存在的，則 *Tanner Stage* 是個帶有本身評估差異的順序出象。因此，本論文明確的研究目的有：

1、針對青春期的非裔男孩及女孩的發育過程，我們想估算出他們在每一階段持續的時間，以及每一階段轉變到下一階段的機率。

2、既然被評斷為鄰近階段的可能性是存在的，我們便想要知道誤判的可能性大小。



❖ 1-3 研究方法：

針對我們所探討有關青春期發育的問題，我們將嘗試應用隱馬可夫鏈模型來處理，這是因為我們得到的資料只是觀測值，而這些觀測值可能參雜誤判的因素在內，不可以被視為真正的發育狀態，所以，我們不可以直接用馬可夫鏈的模型去估計，因此，在本論文，我們試圖利用隱馬可夫模型及理論去解決相關的問題。

第二章、資料部份：

❖ 2-1 資料的背景及意義：

這份資料是由美國德州大學公共衛生學院 詹文耀教授所提供的，共有 48 位非裔男孩及女孩參與研究，其中有 25 位男孩及 23 位女孩，他們起始觀測之年齡範圍為 8 歲至 14 歲，調查人員每隔四個月對這 48 名非裔男孩及女孩進行血壓、身體脂肪、肥胖程度的測量，也針對月經來時的日期、飲食攝取、健康狀態、生理活動、是否抽煙、飲酒及藥物使用量以及當時接受檢查的年齡和日期都有完整的記錄，並且請專業醫護人員檢查他們的陰部毛髮及胸部以便判斷當時發育是屬於 *Tanner Stage* 的哪一階段，總共對每個對象做了 11 次的調查。



❖ 2-2 分析設計及方法：

➤ 研究之問題：

本論文的研究方向是探討由人主觀去評估 *Tanner Stage* 所造成的變異，雖然人類隨著時間成長是連續的，不過，在醫學上是允許發育過程僅用 *Tanner Stage* 有限的數字(*Stage* 1, 2, 3, 4, 5)來表示，由於 *Tanner Stage* 是由人的肉眼去評斷，所以，誤判到鄰近的階段是有可能發生的。舉例來看，某個觀察對象正確的發育階段是屬於第三階段，而她可能被評估的可能階段會是第二、三、四階段其中一個；因為，在第三階段早期被視為第二階段的可能性相當大，同理，在第三階段晚期

被視為第四階段的可能也很大，因此，觀測發育階段的數列中出現跳回前一階段是不會以錯誤的評斷來看待，反而被視為合乎自然常理的變異(詳見 *Espeland, Platt and Gallagher 1989*)。

假設某位男孩或女孩發育過程的觀測數列為 (3, 4, 3, 4)，則可能成為真正發育階段的數列有 (2, 3, 3, 4)、(3, 3, 3, 4)、(3, 4, 4, 4) 或 (3, 3, 4, 4) … 等，但 (2, 2, 3, 3) 絕對不可能是這男孩或女孩的真正發育過程，這是因為第二個觀測值是 4，所以，在第二個時間點可能的發育階段是 3、4、5，不會是 2。

在本論文中，我們除了想了解 *Tanner Stage* 被誤判的機率大小，此外，我們也好奇非裔男孩及女孩在每個 *Tanner Stage* 持續的時間長短和每一階段轉變到下一階段的機率。



➤ 變數分析：

由詹文耀教授所提供的資料，將影響青少年發育過程的因素分為成熟過程變異及非遺傳基因之影響，在原來的資料中，有此非遺傳基因影響的實驗數據，但我們在此只分析外在的表現，即成熟過程的分析，至於非遺傳基因的影響，我們將不予以分析，僅提供如何獲得非遺傳基因的實驗數據。

1、成熟過程變異：

成熟過程的階段是由人去評估，每個對象會被詢問是否正處於發育過程以及他(她)的第一階段的年齡，另外，每個對象藉由醫護人員觀察他(她)的身體去

推斷 *Tanner Stage* 來評估本身第二性象徵的發展，陰部及胸部發育各別分成五個接連的階段來描述，第一個階段是指青春前期而第五階段為發育成熟之成人。

2、非遺傳基因之影響：

(1)營養攝取：

飲食攝取資料是由主持面試者給予食物頻率調查表，再由調查表去決定列出一般經常攝取的 137 種食物，主持面試者會在面試前一週期間取得每種食物的消耗頻率，營養的總數是利用 *USDA* 調查營養資料所提供每種食物的標準比例大小去估計的，便於計算每位參與者通常每天攝取的營養總數。



(2)生理的活動：

生理活動表現的總數是由一個生理活動面談所決定的，而生理活動面談列出包括 43 種從久坐的活動到高度劇烈的活動，所有對象都被詢問在過去七天中他們活動表現的次數，還有他們活動表現平均時間多長及記錄他們的心跳或呼吸增加多少分鐘和活動的分鐘，範圍從非常低到非常高度劇烈活動，並根據活動本身及她們呼吸、心跳增加的分鐘來分級，將過去七天活動的頻率乘以每個時間活動表現的平均分鐘數再除以七就可以計算出每天在各個程度之劇烈活動所花費的時間。

(3)影響健康的行爲：

我們將疾病控制中心及青少年危險行爲預防所研發的問卷稍加修改來評估酒精飲料、菸草、非法藥物、口服避孕藥之使用和減肥的狀況，我們有七個問題是有關於何時間始抽煙以及抽煙的頻率，我們也有四個問題是有關於飲酒的頻率及總數，另外，還有八個問題是和非法藥物使用頻率有關，這份健康行爲問卷是屬於自我控制管理的。

❖ 2-3 資料上的問題及處理方式：

我們得到的原始資料中，發現這 48 位孩童的起始觀測年齡不同，屬於非齊頭式的資料，資料中有詳細記錄每位觀測對象觀測的日期及當時的年齡，他們開始觀測的年齡從 7 到 14 歲之間，每隔四個月觀測一次他們的發育狀態。至於，他們起始年齡非齊一，是否會影響至我們估計的結果？我們認為他們的起始觀察年齡是不會影響我們估計的結果，因為，在本論文中，我們是利用馬可夫鏈模型去模擬，在馬可夫模型下，下一步的狀態只會受到現在狀態的影響，與先前狀態無關，因此，我們可以不必考量到資料起始點非齊一的問題，也不用懷疑估計結果是否會受到影響，不過，我們所得到的觀測值並不能代表發育過程得真正狀態，則我們進一步將資料視為隱馬可夫鏈的模型，利用觀測值去估計可能的真實狀態，有關隱馬可夫模型，我們將在❖3-1 中介紹。

另外，在我們取得的資料中，發現有少數的觀測值為負數，這負數是代表專

業的醫護人員無法明確判斷此孩童現處於哪一發育階段，*Tanner Stage* 將人類發育過程劃分為五個階段，以 $\{1,2,3,4,5\}$ 來表示，觀測值出現負值是很不合理的，而我們將這些負值的觀測值視為代表遺失資料(*missing data*)的符號；至於，這些遺失資料應該如何處理，我們嘗試將原先的觀測值集合為 $A=\{1,2,3,4,5\}$ ，增加一個觀測值以“0”來表示遺失的資料，所以，新的觀測值集合為 $A = \{0,1,2,3,4,5\}$ ，這個改變對於母體參數 $\lambda = (\pi, P, B)$ 而言（定義詳見❖3-1），真實狀態集合仍為 $S = \{0,1,2,3,4,5\}$ 不會有所改變，則 π, P 不會改變，只有 B 會有所改變，每個狀態 S_i 可能觀測到的值增加為 6 個，(即 $B_{5 \times 5} \rightarrow B_{5 \times 6}$)，如此一來，資料既不會失去真實性，也可以順便估算出每個狀態觀測到遺失資料的機率。

在資料合理的修正後，我們便可以將資料代入隱馬可夫鏈的模型中[詳見❖3-1]，再利用 *Baum-Welch* 提出的參數估計法[詳見❖3-2]，進而估算出我們所要探討的問題。

第三章、理論部分：

❖ 3-1 隱馬可夫模型 (*Hidden Markov Model*)：

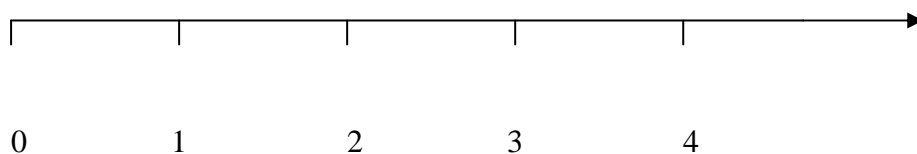
隱馬可夫模型是一個與馬可夫鏈相似之模型，但是它更為普遍，在本論文中所探討的，隱馬可夫鏈模型是屬於一種離散時間的馬可夫鏈且具有一些馬可夫鏈沒有的特性，主要的特性是當馬可夫鏈視察到的一個狀態，此狀態會由一個固定的系統中發出一個觀測值，而此系統的轉換機率(*transition probability*)是一個與時間無關(*time independent*)但通常與狀態相關(*state dependent*)的機率分配。

其實，我們可以將隱馬可夫模型視為一個由兩個步驟結合成的過程。在隱馬可夫模型下，發生的真實狀態，我們依序表示為 q_1, q_2, q_3, \dots 及由每個視察到狀態所發出之觀測值也以 O_1, O_2, O_3, \dots 來表示，對於每個觀測值 O 而言， O 只會與相同時間的真實狀態 q_i 有關，與其他觀測時間的真實狀態完全不相關，則隱馬可夫鏈模型就可被想像成：

真實狀態值：

$q_1 \Rightarrow q_2 \Rightarrow q_3 \Rightarrow q_4$

t：



觀測值：

$O_1 \Rightarrow O_2 \Rightarrow O_3 \Rightarrow O_4$

我們再定義 $O = \{O_1, O_2, O_3, \dots\}$ (即 O 為所有 O_i 的數列)

$Q = \{q_1, q_2, q_3, \dots\}$ (即 Q 為所有 q_i 的數列)

一般來說，我們只知道數列 O 但不知道數列 Q ，所以，通稱數列 Q 是“隱藏的”；隱馬可夫之目的在於可解決許多關於 O 與 Q 的問題，然而，對於隱馬可夫鏈模型所需要的架構，我們給予符號表示且定義如下：

- 1、原來過程的狀態空間是由 N 個狀態所成之集合： $S = \{S_1, S_2, S_3, \dots, S_N\}$ 。
- 2、觀測值空間是由 M 個不同的觀測代號組成之系統： $A = \{a_1, a_2, a_3, \dots, a_M\}$ 。

真正狀態之空間

觀測之狀態空間



- 3、轉移機率之矩陣 (*Transition Probability Matrix*)：

$P = (P_{ij})$ ，為 $N \times N$ 的矩陣，其中 $P_{ij} = P(q_{t+1} = S_j | q_t = S_i)$ 。

- 4、觀測值之機率：

對於每個狀態 S_i 且 $a \in A$ ，則 $b_i(a) = P$ (在真正狀態是 S_i 而觀測到的狀態為 a)

且由所有 $b_i(a)$ 形成一個 $N \times M$ 的矩陣 $B = (b_i(a))$ 。

- 5、起始狀態之分配為向量 $\pi = (\pi_i)$ ，其中 $\pi_i = P(q_1 = S_i)$ 。

關於隱馬可夫模型，我們所假設的母體參數有 (P, B, π) ，對於本論文所要討論之問題，就是在未知母體參數下，透過隱馬可夫模型之理論來估計母體參數

(P 、 B 、 π)，進而解釋我們所好奇的問題。

❖ 3-2 引理之說明：

對於本論文所要探討問題，首先，我們將引用 *Baum-Welch* 提出的參數估計法 [詳見❖3-3] 來估計未知參數 (P 、 B 、 π)，進而解決相關之問題；在介紹 *Baum-Welch* 參數估計法之前，我們必須先有以下的引理之說明。(每一引理來源請參考文獻[1])

□ 引理 1 : (*Forward Algorithm*)

令 $\alpha^{(d)}(t, i) = P(O_1^{(d)}, O_2^{(d)}, \dots, O_t^{(d)}, q_t^{(d)} = S_i)$ 為 *forward variables*，

用來表示已知在第 t 時間的真正狀態是 S_i ，而時間 t 以前的觀測狀態分別是

$O_1^{(d)}, O_2^{(d)}, \dots, O_t^{(d)}$ 的機率，則

$$\alpha^{(d)}(t+1, i) = \sum_{j=1}^N \alpha^{(d)}(t, j) \times P_{ij} \times b_i(O_{t+1}^{(d)}) \quad (3.1)$$

[證明]：

先算出 $\alpha^{(d)}(1, i) = P(O_1^{(d)}, q_1^{(d)} = S_i) = \pi_i \times b_i(O_1^{(d)})$

下一步，尋找 $\alpha^{(d)}(t+1, i)$ 與 $\alpha^{(d)}(t, i)$ 之關係：

$$\begin{aligned} \alpha^{(d)}(t+1, i) &= P(O_1^{(d)}, O_2^{(d)}, \dots, O_t^{(d)}, O_{t+1}^{(d)}, q_t^{(d)} = S_i) \\ &= \sum_{j=1}^N P(O_1^{(d)}, O_2^{(d)}, \dots, O_t^{(d)}, O_{t+1}^{(d)}, q_t^{(d)} = S_j, q_{t+1}^{(d)} = S_i) \\ &= \sum_{j=1}^N P(O_1^{(d)}, \dots, O_t^{(d)}, q_t^{(d)} = S_j) \times P(q_{t+1}^{(d)} = S_i | q_t^{(d)} = S_j) \\ &\quad \times P(O_{t+1}^{(d)} | q_{t+1}^{(d)} = S_i) \\ &= \sum_{j=1}^N \alpha^{(d)}(t, j) \times P_{ij} \times b_i(O_{t+1}^{(d)}) \end{aligned}$$

所以，我們可以利用疊代法依序求出：

$$\alpha^{(d)}(1,i) \Rightarrow \alpha^{(d)}(2,i) \Rightarrow \alpha^{(d)}(3,i) \Rightarrow \dots \Rightarrow \alpha^{(d)}(T,i), \forall i \quad \square$$

□ 引理 2：

在第 t 時間下，出現觀測狀態 $O_t^{(d)}$ 之機率可表示成：

$$P(O^{(d)}) = \sum_{i=1}^N \alpha^{(d)}(T,i) \quad (3.2)$$

[證明]：

利用定理 1， $P(O^{(d)})$ 可寫成：

$$\begin{aligned} P(O^{(d)}) &= P(O_1^{(d)}, O_2^{(d)}, \dots, O_T^{(d)}) \\ &= \sum_{i=1}^N P(O_1^{(d)}, O_2^{(d)}, \dots, O_T^{(d)}, q_t^{(d)} = S_i) \\ &= \sum_{i=1}^N \alpha^{(d)}(T,i) \end{aligned}$$

□

□ 引理 3：(Backward Algorithm)

令 $\beta^{(d)}(t,i) = P(O_{t+1}^{(d)}, O_{t+2}^{(d)}, \dots, O_T^{(d)} \mid q_t^{(d)} = S_i)$ 為 *backward variables*，用

來表示在給定第 t 時間的真正狀態是 S_i 下，而時間 t 之後的觀測狀態分別

$O_{t+1}^{(d)}, O_{t+2}^{(d)}, \dots, O_T^{(d)}$ 的機率，且定義對於任何 i ， $\beta^{(d)}(T,i) = 1$ ，則尋找

$\beta^{(d)}(t-1,i)$ 與 $\beta^{(d)}(t,i)$ 之關係，可以得到：

$$\beta^{(d)}(t-1,i) = \sum_{j=1}^N P_{ij} \times b_j(O_t^{(d)}) \times \beta^{(d)}(t,j) \quad (3.3)$$

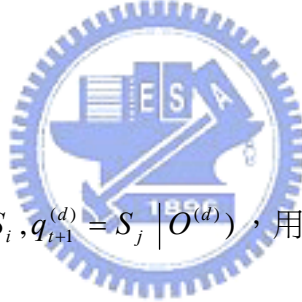
[證明]：

$$\begin{aligned}
\beta^{(d)}(t-1, i) &= P(O_t^{(d)}, O_{t+1}^{(d)}, O_{t+2}^{(d)}, \dots, O_T^{(d)}, O_{t+1}^{(d)} \mid q_{t-1}^{(d)} = S_i) \\
&= \sum_{j=1}^N P(O_t^{(d)}, O_{t+1}^{(d)}, \dots, O_T^{(d)}, O_{t+1}^{(d)}, q_t^{(d)} = S_j \mid q_{t-1}^{(d)} = S_i) \\
&= \sum_{j=1}^N P(q_t^{(d)} = S_j \mid q_{t-1}^{(d)} = S_i) \times P(O_t^{(d)} \mid q_t^{(d)} = S_j) \\
&\quad \times P(O_{t+1}^{(d)}, O_{t+2}^{(d)}, \dots, O_T^{(d)} \mid q_t^{(d)} = S_j) \\
&= \sum_{j=1}^N P_{ij} \times b_i(O_t^{(d)}) \times \beta^{(d)}(t, j) \tag{3.9}
\end{aligned}$$

所以，利用疊代法，可以依序求出：

$$1 = \beta^{(d)}(T, i) \Rightarrow \beta^{(d)}(T-1, i) \Rightarrow \beta^{(d)}(T-2, i) \Rightarrow \dots \Rightarrow \beta^{(d)}(1, i), \forall i \quad \square$$

□ 引理 4：



令 $\xi_t^{(d)}(i, j) = P(q_t^{(d)} = S_i, q_{t+1}^{(d)} = S_j \mid O^{(d)})$ ，用來表示給定觀測狀態為 $O^{(d)}$

下，第 t 時間的真正狀態是 S_i 且第 $t+1$ 時間的真正狀態是 S_j 的機率，則

我們可以得到：

$$\xi_t^{(d)}(i, j) = \frac{\alpha_t^{(d)}(i) \times P_{ij} \times b_j(O_{t+1}^{(d)}) \times \beta_{t+1}^{(d)}(j)}{\sum_{i=1}^N \alpha_t^{(d)}(T, i)} \tag{3.4}$$

[證明]：

$$\begin{aligned}
\xi_t^{(d)}(i, j) &= P(q_t^{(d)} = S_i, q_{t+1}^{(d)} = S_j \mid O^{(d)}) \\
&= \frac{P(q_t^{(d)} = S_i, q_{t+1}^{(d)} = S_j, O^{(d)})}{P(O^{(d)})}
\end{aligned}$$

則分子部份可利用(3.1)及(3.3)改寫成：

$$\begin{aligned}
& P(q_t^{(d)} = S_i, q_{t+1}^{(d)} = S_j, O^{(d)}) \\
&= P(O_1^{(d)}, O_2^{(d)}, \dots, O_t^{(d)}, q_t^{(d)} = S_i) \times P(q_{t+1}^{(d)} | q_t^{(d)} = S_i) \\
&\quad \times P(O_{t+1}^{(d)} | q_{t+1}^{(d)} = S_j) \times P(O_{t+2}^{(d)}, O_{t+3}^{(d)}, \dots, O_T^{(d)} | q_{t+1}^{(d)} = S_j) \\
&= \alpha_t^{(d)}(i) \times P_{ij} \times b_j(O_{t+1}^{(d)}) \times \beta_{t+1}^{(d)}(j)
\end{aligned}$$

而分母部份也可由(3.2)中求得，

$$\text{因此， } \xi_t^{(d)}(i, j) = \frac{\alpha_t^{(d)}(i) \times P_{ij} \times b_j(O_{t+1}^{(d)}) \times \beta_{t+1}^{(d)}(j)}{\sum_{i=1}^N \alpha^{(d)}(T, i)} \quad \square$$

□ 引理 5 :

假設指標函數為 $I_t^{(d)}(i) = \begin{cases} 1 & , \text{ if } q_t^{(d)} = S_i \\ 0 & , \text{ otherwise} \end{cases}$ ，用來表示第 d 組資料下，如果

在第 t 時間真正的狀態是 S_i ，我們就令 $I_t^{(d)}(i)$ 值為 1；在其他情況下，令 $I_t^{(d)}(i)$

值為 0，則

$$(1) E(N_i | \{O\}) = \sum_d \sum_t \sum_{j=1}^N \xi_t^{(d)}(i, j) \quad (3.5)$$

$$(2) E(N_{ij} | \{O\}) = \sum_d \sum_t \xi_t^{(d)}(i, j) \quad (3.6)$$

$$(3) E(N_i, t = 1 | \{O\}) = \sum_d \sum_{j=1}^N \xi_1^{(d)}(i, j) \quad (3.7)$$

[證明] :

(1)

$$\begin{aligned}
E(N_i | \{O\}) &= \sum_d \sum_t E(I_t^{(d)}(i) | O^{(d)}) \\
&= \sum_d \sum_t P(q_t^{(d)} = S_i | O^{(d)}) \\
&= \sum_d \sum_t \sum_{j=1}^N P(q_t^{(d)} = S_i, q_{t+1}^{(d)} = S_j | O^{(d)}) \\
&= \sum_d \sum_t \sum_{j=1}^N \xi_t^{(d)}(i, j)
\end{aligned}$$

(2)

$$\begin{aligned} E(N_{ij}|\{O\}) &= \sum_d \sum_t P(q_t^{(d)} = S_i, q_{t+1}^{(d)} = S_j | O^{(d)}) \\ &= \sum_d \sum_t \xi_t^{(d)}(i, j) \end{aligned}$$

(3) 假設當 $t = 1$ 時，

$$\begin{aligned} E(N_i, t = 1 | \{O\}) &= \sum_d P(q_1^{(d)} = S_i | O^{(d)}) \\ &= \sum_d \sum_{j=1}^N P(q_1^{(d)} = S_i, q_2^{(d)} = S_j | O^{(d)}) \\ &= \sum_d \sum_{j=1}^N \xi_1^{(d)}(i, j) \end{aligned}$$

□

□ 引理 6 :

假設指標函數為 $I_t^{(d)}(i, a) = \begin{cases} 1, & \text{if } q_t^{(d)} = S_i \text{ and } O_t^{(d)} = a \\ 0, & \text{otherwise} \end{cases}$ ，用來表示第 d

組資料下，如果在第 t 時間真正的狀態是 S_i 且觀測值為 a ，我們就令 $I_t^{(d)}(i, a)$

值為 1；在其他情況下，令 $I_t^{(d)}(i, a)$ 值為 0，則

$$E(N_i(a) | \{O\}) = \sum_d \sum_t \sum_{O_t^{(d)}=a} \sum_{j=1}^N \xi_t^{(d)}(i, j) \quad (3.8)$$

[證明] :

$$\begin{aligned} E(N_i(a) | \{O\}) &= \sum_d \sum_t E(I_t^{(d)}(i, a) | O^{(d)}) \\ &= \sum_d \sum_t P(q_t^{(d)} = S_i, O_t^{(d)} = a | O^{(d)}) \\ &= \sum_d \sum_t \sum_{j=1}^N P(q_t^{(d)} = S_i, q_{t+1}^{(d)} = S_j, O_t^{(d)} = a | O^{(d)}) \\ &= \sum_d \sum_t \sum_{O_t^{(d)}=a} \sum_{j=1}^N \xi_t^{(d)}(i, j) \end{aligned}$$

□

❖ 3-3 *Baum-Welch* 提出之參數估計法(*Baum-Welch Method of Parameter Estimation*) :

我們將引用 *Baum-Welch* 提出的參數估計法來估計未知參數 (P 、 B 、 π)，進而解決相關之問題，而 *Baum-Welch* 參數估計法說明如下：(有關 *Baum-Welch* 參數估計法請參考文獻[1])

步驟一：先假設這些參數 π_i ， P_{ij} 及 $b_i(a)$ 之起始值，我們可以利用均勻地給值

或是憑對此資料之了解選擇直覺判斷來選取起始值。

步驟二：利用這些參數的起始值可計算出：

$$\overline{\pi}_i = \text{在給定}\{O\}\text{下，第一個時間點的狀態} \quad (3.9)$$

在 S_i 之期望次數的比例。

$$\overline{P}_{ij} = \frac{E(N_{ij}|\{O\})}{E(N_i|\{O\})} \quad (3.10)$$

$$\overline{b}_i(a) = \frac{E(N_i(a)|\{O\})}{E(N_i|\{O\})} \quad (3.11)$$

其中， $\{O\}$ ：指整體之觀測值數列，即 $\{O\} = \{O^{(1)}, O^{(2)}, \dots, O^{(d)}\}$ 。

N_{ij} ：對於任何 d 及 t ，當 $q_t^{(d)} = S_i$ 且 $q_{t+1}^{(d)} = S_j$ 的次數。

N_i ：對於任何 d 及 t ，當 $q_t^{(d)} = S_i$ 的次數。

$N_i(a)$ ：對於任何 d 及 t ，當 $q_t^{(d)} = S_i$ 且它發出的觀測值為 a 的次數。

所以，我們利用❖3-2 提到的引理，可以推得：

由(3.7)可得(3.9)：

$$\bar{\pi}_i = \frac{\sum_d \sum_{j=1}^N \xi_1^{(d)}(i, j)}{\sum_i \sum_d \sum_{j=1}^N \xi_1^{(d)}(i, j)}$$

由(3.5)，(3.6)，可得(3.10)

$$\bar{P}_{ij} = \frac{E(N_{ij}|\{O\})}{E(N_i|\{O\})} = \frac{\sum_d \sum_t \xi_t^{(d)}(i, j)}{\sum_d \sum_t \sum_{j=1}^N \xi_t^{(d)}(i, j)}$$

由(3.5)，(3.8)，可得(3.11)

$$\bar{b}_i(a) = \frac{E(N_i(a)|\{O\})}{E(N_i|\{O\})} = \frac{\sum_d \sum_t \sum_{O_t^{(d)}=a} \sum_{j=1}^N \xi_t^{(d)}(i, j)}{\sum_d \sum_t \sum_{j=1}^N \xi_t^{(d)}(i, j)}$$

步驟三：，將步驟二估出的新參數 $\bar{\lambda} = (\bar{\pi}, \bar{P}, \bar{B})$ 視為新的起始值，再代回步驟一、二，並重覆執行步驟二直到 $Prob(\{O\}|\bar{\lambda}) - Prob(\{O\}|\lambda) \leq 10^{-8}$ 時才停止疊代（亦即直到參數收斂，到達穩定時才停止疊代）。

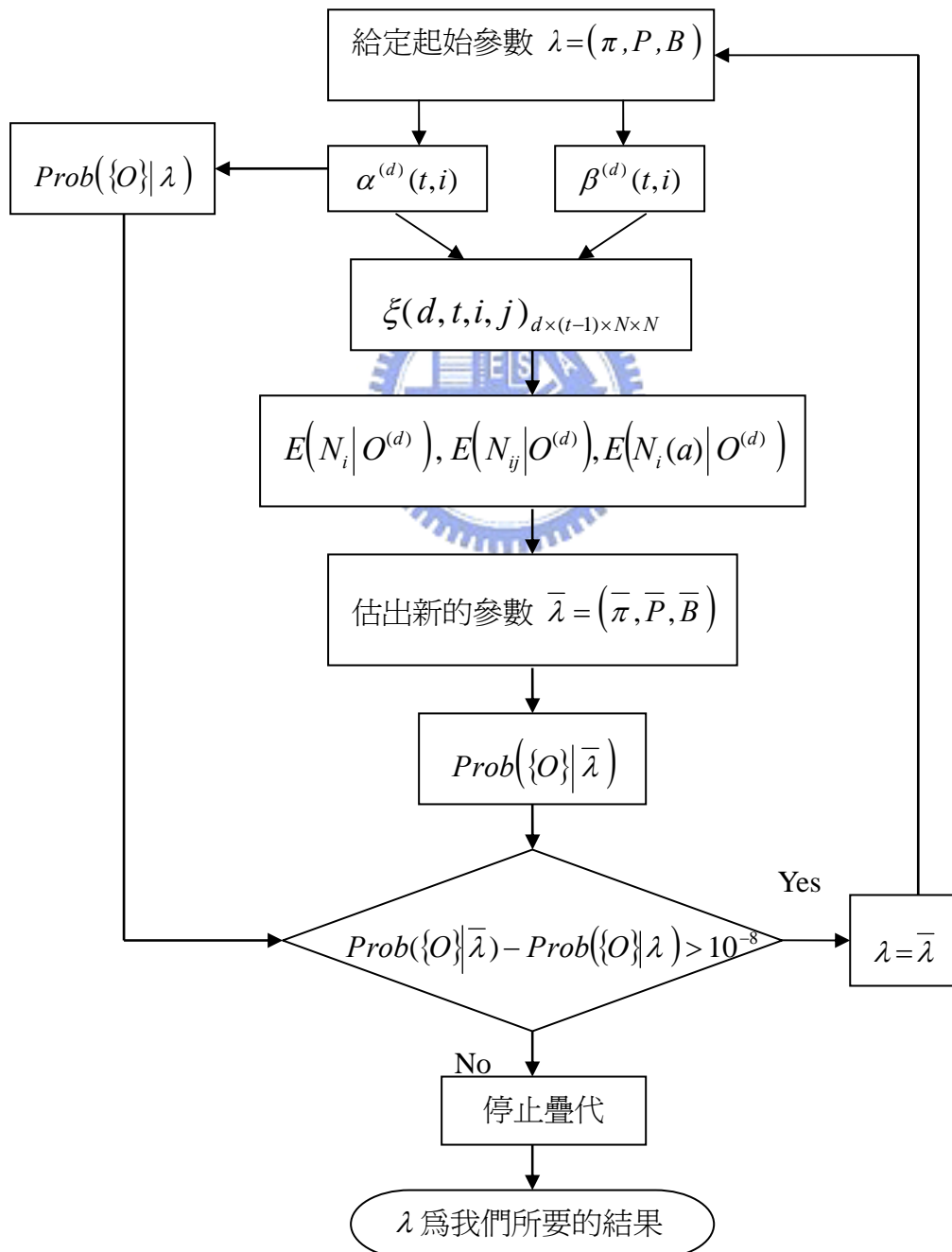
Baum-Welch 提出的參數估計法已被證實，如果 $\lambda = (\pi, P, B)$ 被 $\bar{\lambda} = (\bar{\pi}, \bar{P}, \bar{B})$ 取代，則 $Prob(\{O\}|\bar{\lambda}) \geq Prob(\{O\}|\lambda)$ ，所以當 $\bar{\lambda} = \lambda$ (即達穩定收斂時)或是當 $Prob(\{O\}|\bar{\lambda})$ 與 $Prob(\{O\}|\lambda)$ 之變化極微小時，此時的 $\bar{\lambda} = (\bar{\pi}, \bar{P}, \bar{B})$ 就是最合適說明此資料的最佳參數，即 $\bar{\lambda} = (\bar{\pi}, \bar{P}, \bar{B})$ 為最大概似估計值 (*MLE*；*Maximum Likelihood Estimator*)。

第四章、演算分析與結果部分

❖ 4-1 演算之流程：

根據第三章所介紹的理論，我們的演算流程為下圖所示，而演算過程是利用

C 語言來執行。



❖ 4-2 演算之結果分析：

根據❖3-3 *Baum-Welch* 提出的參數估計法，估計出來的參數 $\bar{\lambda} = (\bar{\pi}, \bar{P}, \bar{B})$ 就是最適說明此資料的最佳參數，我們演算出來之結果請詳見[表 4-1，表 4-2]，而我們得到的參數估計 $\bar{\lambda} = (\bar{\pi}, \bar{P}, \bar{B})$ 中， $\bar{\pi}$ 的分析價值意義較小，這是因為 $\bar{\pi}$ 代表意義是起始真正狀態在 S_i 的機率，一般而言，人類的發育都是由尚未發育的第一階段開始，且此資料的起始觀測年齡約為 8-14 歲之間，屬於尚未開始發育的年紀，他們起始觀測之發育階段幾乎都是在第一階段，僅有少數幾位參與者的起始觀測值為第二階段，所以，由估計的結果得知，不論男孩或女孩，他們真正的起始發育階段皆為第一階段，因此， $\bar{\pi}$ 的估計結果較無分析的價值。

在 \bar{P} 的估計結果中，我們可以很清楚地知道在每一階段轉變到下一階段及停留在該階段之機率。對於此資料的 25 位非裔男孩而言，他們在發育第一階段停留在該階段的機率高達 0.74，或許是因為開始觀測的年紀較小和男孩發育年齡普遍比女孩來得晚所導致的，在發育過程的第二階段發展至第三階段或停留再第二階段之機會大約是差不多的，較無太大差異，而在發育的第三、四階段卻有較高的機會發展至第四、五階段；對於此資料的 23 位非裔女孩而言，發育過程在第一、二階段停留在該階段的機會較高，只有 0.35 的機率發育至第二、三階段，但是在發育的第三階段卻是有較高的機率約 0.65 發展至第四階段，而在第四階段發育至完全成人或停留在第四階段的機率近乎相同。不論男孩或女孩，一旦發育為成人(即第五階段)就會停留在成人階段，不會在回到先前的發育階段。

接著，在 \bar{B} 的估計結果中，主要是表示在每一個真正的發育階段下，我們可能觀測到每一個階段的機率大小，由這些數據可以評估出這些專業醫護人員的判斷是否正確，以及誤判和無法判斷的可能性大小。對於此資料的 25 位非裔男孩而言，在發育的第二、三階段，醫護人員幾乎可以完全無誤地判斷正確，而在發育第四、五階段，醫護人員判斷正確的機率僅有四、五成，似乎很容易認定為前一階段，可能是因為這些參與者正屬於發育第四、五階段早期，所以，容易被判斷正處於發育的第三、四階段，而在發育的第一階段是最容易被醫護人員認為無法分辨的階段，因為發育的第一階段定義為出生至開始發育到第二階段之前這段時間，所以很難去判定是否已開始發育，因此，相較之下在第一階段醫護人員無法辨識屬於哪階段的機會較高；對於此資料的 23 位非裔女孩而言，在發育的第三、四階段，醫護人員有九成多的機率可以無誤地判斷正確，而在發育第五階段，醫護人員判斷正確的機率僅有六成，卻有 0.26 的機率屬於無法認定在那階段，可能是因為女孩的胸部發育在發育完成認定上較難分辨，身材較為瘦弱的女孩在胸部完全發育的大小會比一般女孩小，所以，發育的第五階段是最容易被醫護人員認為無法分辨的階段，不知道胸部發育是否會再成長，因此，相較之下在第五階段醫護人員無法辨識的機會比較高。以上結果的分析，僅適用於此資料及此組醫護人員，若是換了不同的參與者或別組醫護人員去觀測，估計出來的結果也會隨之不同。

至於發育過程中每一階段所需花費的時間該如何估算，我們是利用每階段被觀測到的次數服從幾何分配，這是因為我們是以離散型的隱馬可夫模型去處理，則每階段被觀測到的次數便會服從幾何分配，進而計算出每階段被觀測到的期望次數再乘以每次觀測間隔的時間，即可估算出發育過程中每一階段所需經歷的時間。舉例說明：假設發育的第二階段被觀測到的次數是 L_2 ，且 L_2 服從幾何分配 (*Geometric distribution*)，由第二階段發育至第三階段的機率記為 P_{23} 且第二階段發育至第二階段的機率記為 P_{22} ，則 $L_2 \sim \text{Geometric}(P_{23})$ ，然後，

$$P(L_2 = 1) = P_{23}, \quad \text{即發育過程數列為}(1, \dots, 1, 2, 3, \dots)$$

$$P(L_2 = 2) = P_{22} \times P_{23}, \quad \text{即發育過程數列為}(1, \dots, 1, 2, 2, 3, \dots)$$

$$P(L_2 = 3) = P_{22} \times P_{23} \times P_{23}, \quad \text{即發育過程數列為}(1, \dots, 1, 2, 2, 2, 3, \dots)$$

$$\text{以此類推，我們可以算出 } E(L_2) = \frac{1}{P_{23}}, \quad \text{Var}(L_2) = \frac{(1 - P_{23})}{P_{23}^2},$$

再將第二階段被觀測到的期望次數 $E(L_2) \times$ 四個月，便可得到第二階段所需的時間。

由〔圖 4-1 ~ 圖 4-8〕所示，我們可以知道此資料之非裔男孩及女孩在每個發育階段的分配圖，再將非裔男孩及女孩於發育過程中每一階段所需經歷的時間計算出來，如〔表 4-3〕所示，我們只想了解發育過程第二、三、四階段所需所需經歷的時間，因為，發育的第一階段定義為出生至開始發育到第二階段之前這段時間，我們估計出來的第一階段所需時間會隨著此資料之參與者的起始觀測年齡而改變，所以，估算出來第一階段所需時間是沒有意義的。由〔表 4-3〕所示，我們可以得知此資料的非裔男孩從開始發育需要經歷約 19.73 個月才能發育至成人，在

發育過程的第三階段所需經歷時間最短；而對於非裔女孩從開始發育需要經歷約 25.21 個月才能發育至成人，發育過程的第三階段所需經歷時間也是最短的；此資料的非裔男孩及女孩相較之下，男孩發育所需的時間較女孩短，發育速度比較快。

對於我們將發育過程是為馬可夫鏈，其實是有不恰當之處的，我們會有修正的方法及說明，但這修正的方法對於此資料還是不適合的。

❖ 4-3 改進的方法：

在本論文中，我們把青少年發育過程視為馬可夫鏈是不盡理想的，人的發育是連續性的，每一次的觀測結果其實是不會單與前一個觀測狀態有關，我們舉例來說明：假設發育過程數列為 $(1, \dots, 1, 2, \square, \dots)$ 或 $(1, \dots, 1, 2, 2, \triangle, \dots)$ ，其中， \square 和 \triangle 這個空格可能出現的階段皆為 $\{2, 3\}$ ，依照我們先前馬可夫鏈的假設， \square 和 \triangle 這個空格分別出現的階段 2, 3 的機率會是一樣的(即 P_{22} 或 P_{23})，但是，一般來說在發育過程數列 $(1, \dots, 1, 2, 2, \triangle, \dots)$ 中，已經觀測出兩次第二階段，而下一次觀測狀態 \triangle 為 2 的機率應該會比在發育過程數列 $(1, \dots, 1, 2, \square, \dots)$ 只觀測一次第二階段且下一次觀測狀態 \square 為 2 的機率來得低，同理， \triangle 為 3 的機率應該會比 \square 為 3 的機率來得高，所以，以馬可夫鏈模型當做發育過程是有缺失的，我們進一步嘗試將兩次的觀測狀態當作一組，這樣每一次的觀測結果不會只與前一個觀測狀態有關，而是與前兩次的觀測結果相關，以增加發育過程之間的關聯性，這樣的修正將會比較合理與完善。

至於，如何將兩次的觀測狀態當作一組，我們說明如下：

假設原先的觀測數列為 $X_1, X_2, X_3, X_4, X_5, X_6, \dots$,

而改進後的觀測數列為 Y_1, Y_2, Y_3, \dots ，其中 $Y_1 = (X_1, X_2), Y_2 = (X_3, X_4), Y_3 = (X_5, X_6)$

則， $P(Y_n | Y_{n-1}, Y_{n-2}, \dots, Y_1) = P(Y_n | Y_{n-1})$ ，符合馬可夫鏈的特性，我們便可以再利用

Baum-Welch 提出的參數估計法去估算未知參數，而新的真正過程的狀態空間是由

$N \times N$ 個狀態所成之集合： $S = \{S_1, S_2, S_3, \dots, S_{N \times N}\}$ ，新的觀測值空間是由 $N \times M$

個不同的觀測代號組成之系統： $A = \{a_1, a_2, a_3, \dots, a_{N \times M}\}$ ，則轉移機率之矩陣 P

為 $N^2 \times N^2$ 的矩陣，可能出現觀測值之矩陣 B 為 $N^2 \times M^2$ 的矩陣。

因為我們的資料數目不多，改用此方法會使觀測次數減半，而且要估計的參數 $\lambda = (\pi, P, B)$ 矩陣變得很大，僅用減半的資料卻要估計那麼多未知數，無法估計的準確，我們嘗試過的結果並不理想，經過考量，在此僅提出說明可以改善的方法，結果不予採納，我們仍以每一次的觀測值只與前一個觀測狀態有關的結果來分析。如果資料數目夠多，相信利用此方法所估計出的結果會更加完善合理。

〔表 1-1〕：Tanner Staging 之定義

I. Girls

Tanner Stage	Stage of develop	Pubic Hair	Breasts
Stage 1	Early adolescence (10-13 years)	Preadolescent	Preadolescent
Stage 2		Sparse, straight	small mound
Stage 3	Middle adolescence (12-14 years)	Dark, curl	bigger; no contour separation
Stage 4		Coarse, curly, abundant	Secondary mound of areola
Stage 5	Late Adolescence (14-17 years)	Triangle; medial thigh	nipple projects; areola part of breast

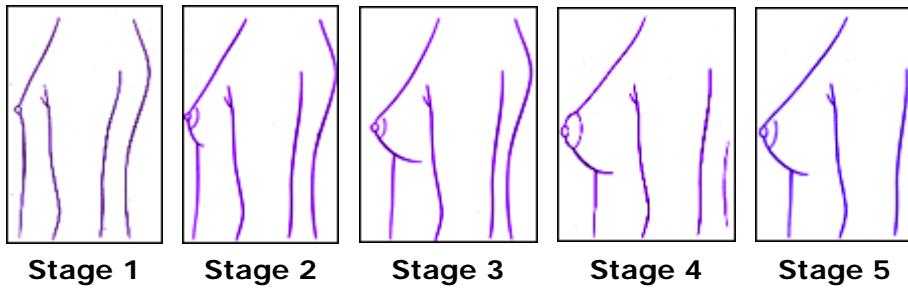
II. Boys

Tanner Stage	Stage of develop.	Pubic Hair	Penis	Testes
Stage 1	Early adolescence (10.5-14 years)	None	Preadolescent	preadolescent
Stage 2		Scanty	Slight increase	larger
Stage 3	Middle adolescence (12.5-15 years)	Darker, curls	Longer	larger
Stage 4		adult, coarse, curly	Larger	scrotum dark
Stage 5	Late adolescence (14-16 years)	adult - thighs	Adult	adult

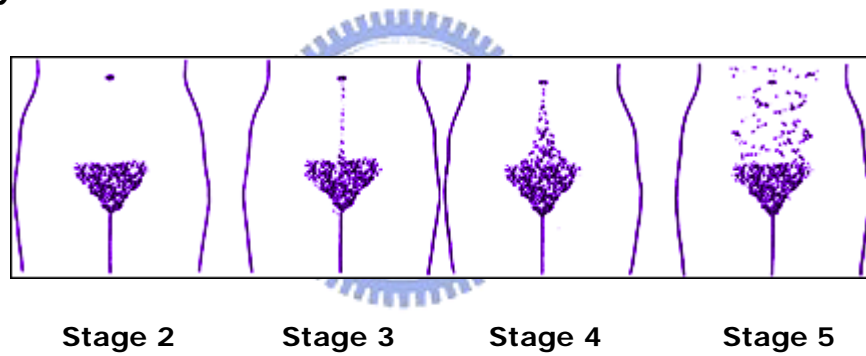
資料來源：<http://www.mcg.edu/pediatrics/CCNotebook/chapter3/tanner.htm> ❖

〔圖 1-1〕：Tanner Staging之圖示

I. Girls



II. Boys



資料來源：<http://www.afraidtoask.com/breast/breastdevelopment.html>

[表 4-1] : 25 位非裔男孩的參數估計結果

===== $\bar{\pi}$ =====

$q_i=1$	$q_i=2$	$q_i=3$	$q_i=4$	$q_i=5$
1	0	0	0	0

===== \bar{P} =====

	$S_j=1$	$S_j=2$	$S_j=3$	$S_j=4$	$S_j=5$
$S_i=1$	0.7409970910	0.2590029090	0	0	0
$S_i=2$	0	0.5492576790	0.4507423210	0	
$S_i=3$	0	0	0.2149702442	0.7850297558	0
$S_i=4$	0	0	0	0.3049495450	0.6950504550
$S_i=5$	0	0	0	0	1

===== \bar{B} =====

	$a=0$	$a=1$	$a=2$	$a=3$	$a=4$	$a=5$
$S_i=1$	0.1553928407	0.7215234834	0.1230836759	0	0	0
$S_i=2$	0.0000154969	0.0064124108	0.9935538444	0.0000182478	0	0
$S_i=3$	0	0	0.0004062753	0.9994995737	0.0000941510	0
$S_i=4$	0	0	0	0.4800948451	0.5198943498	0.0000108051
$S_i=5$	0.0272836493	0	0	0	0.5634703120	0.4092460387

[表 4-2] : 23 位非裔女孩的參數估計結果

===== $\bar{\pi}$ =====

$q_1=1$	$q_1=2$	$q_1=3$	$q_1=4$	$q_1=5$
0.99798012	0.00201988	0	0	0

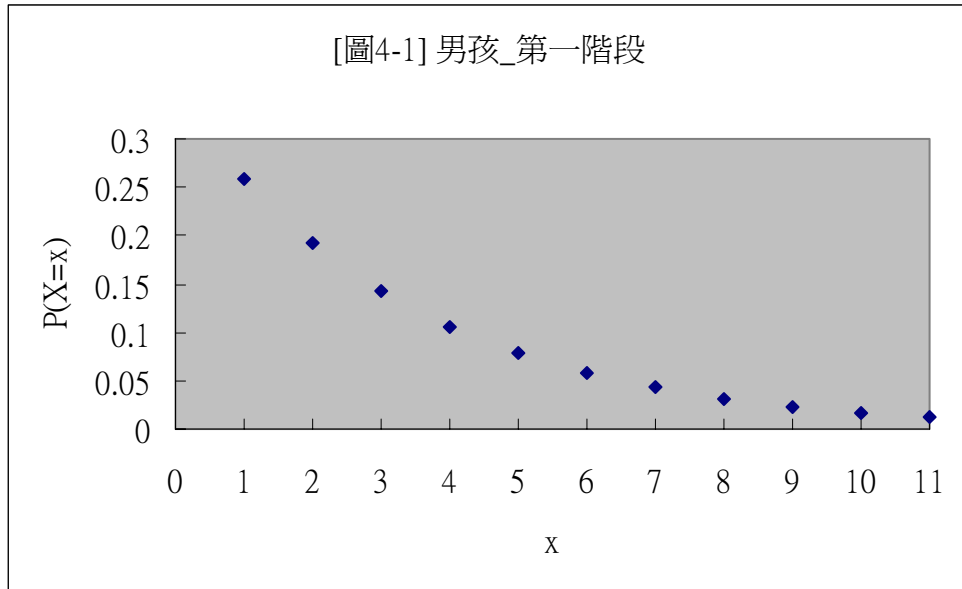
===== \bar{P} =====

	$S_j=1$	$S_j=2$	$S_j=3$	$S_j=4$	$S_j=5$
$S_i=1$	0.63229356	0.36770644	0	0	0
$S_i=2$	0	0.64416262	0.35583738	0	
$S_i=3$	0	0	0.34840073	0.65159927	0
$S_i=4$	0	0	0	0.48901531	0.51098469
$S_i=5$	0	0	0	0	1

===== \bar{B} =====

	$a=0$	$a=1$	$a=2$	$a=3$	$a=4$	$a=5$
$S_i=1$	0.12652401	0.80097964	0.07249635	0	0	0
$S_i=2$	0.04799774	0.00000007	0.88919616	0.06280603	0	0
$S_i=3$	0	0	0.00000565	0.96154701	0.03844734	0
$S_i=4$	0.04458629	0	0	0.00000012	0.92005473	0.03535885
$S_i=5$	0.25757199	0	0	0	0.08030563	0.66212238

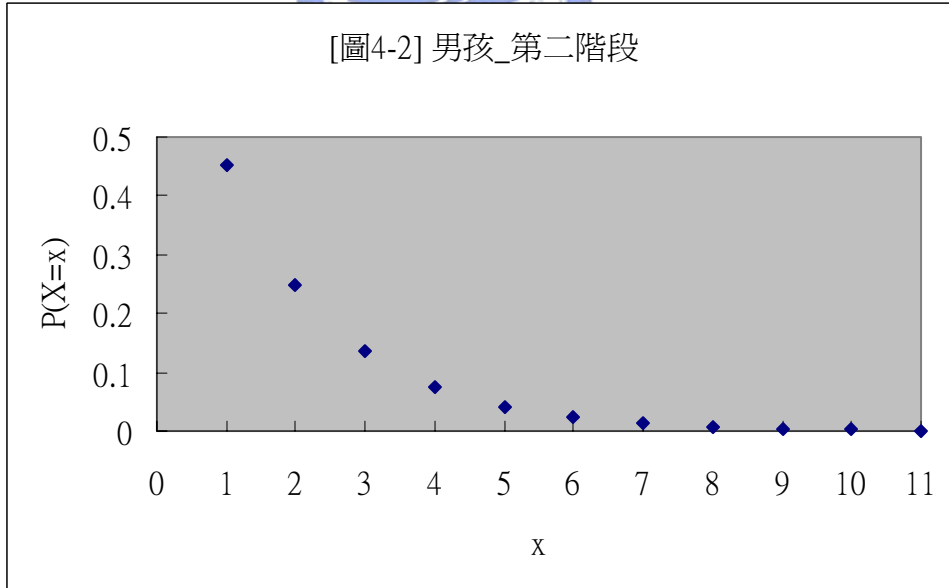
[圖4-1] 男孩_第一階段



E(X)	Var(X)
3.8609605	11.04606

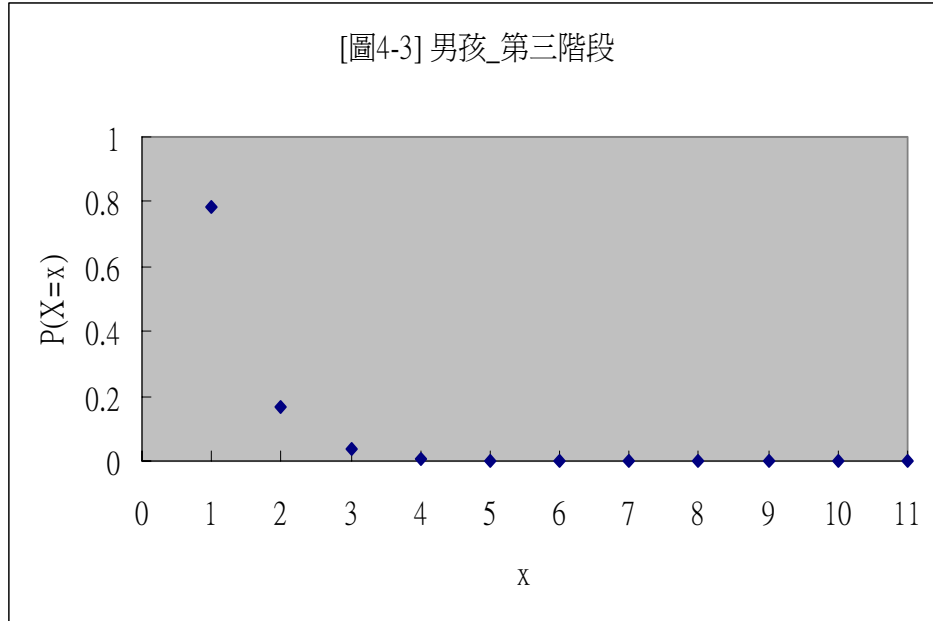


[圖4-2] 男孩_第二階段



E(X)	Var(X)
2.2185625	2.703457

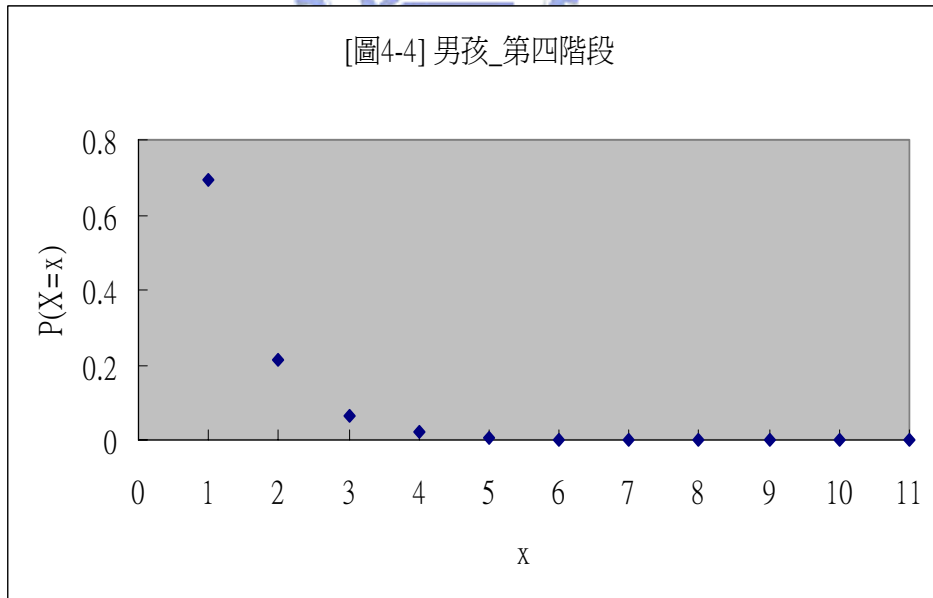
[圖4-3] 男孩_第三階段



E(X)	Var(X)
1.273837	0.348824

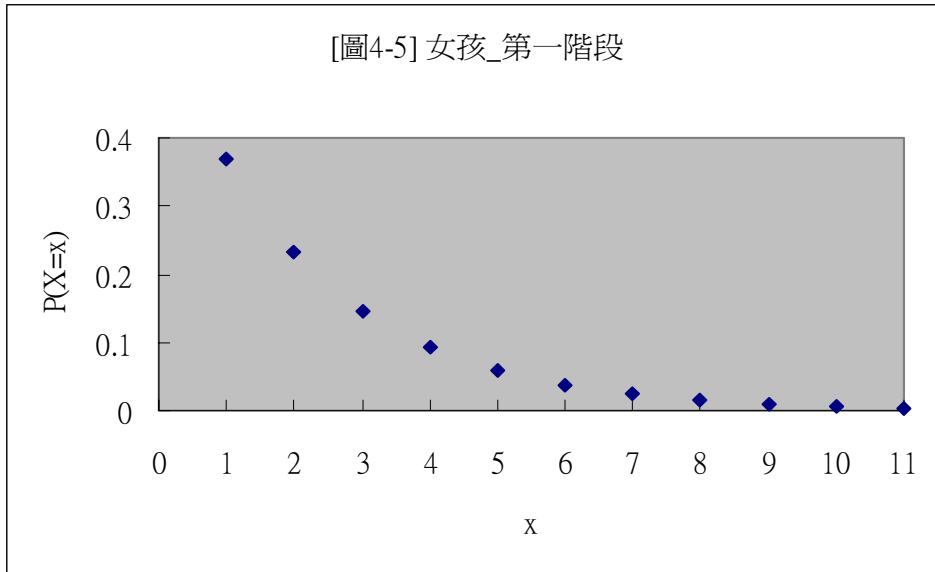


[圖4-4] 男孩_第四階段



E(X)	Var(X)
1.4387445	0.631241

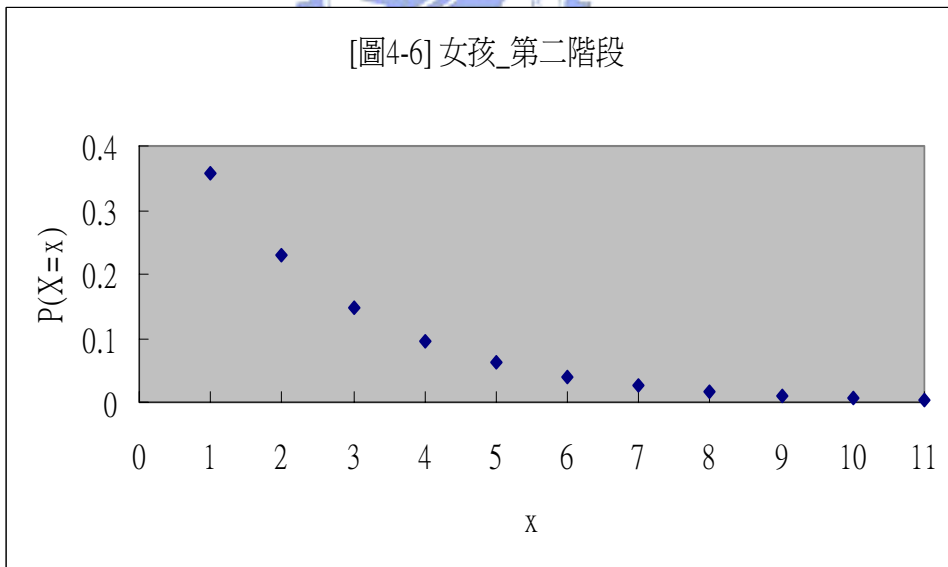
[圖4-5] 女孩_第一階段



E(X)	Var(X)
2.719561	4.67645

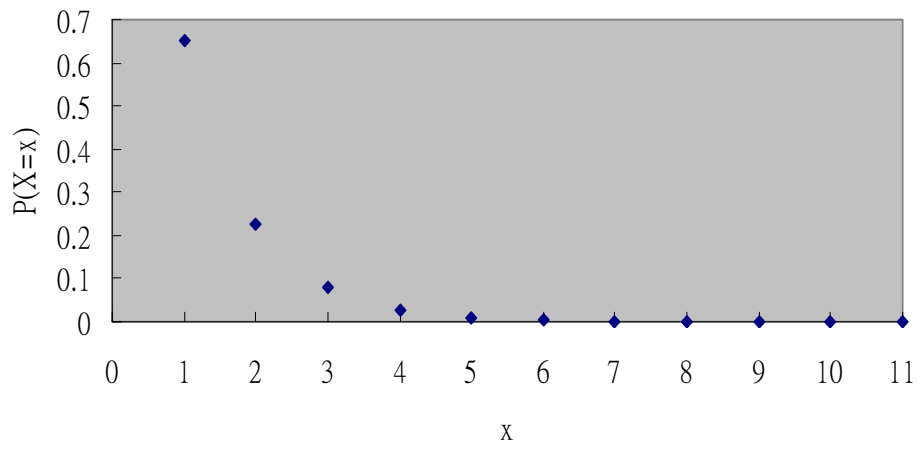


[圖4-6] 女孩_第二階段



E(X)	Var(X)
2.810272	5.087359

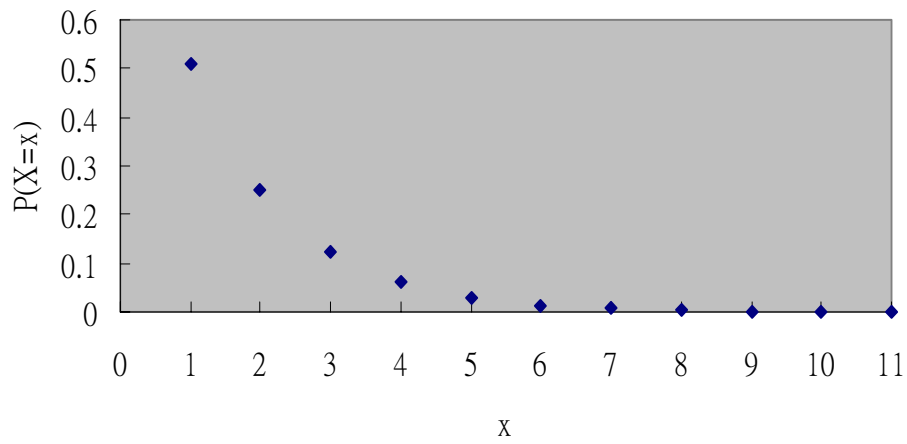
[圖4-7] 女孩_第三階段



E(X)	Var(X)
1.534686	0.820574



[圖4-8] 女孩_第四階段



E(X)	Var(X)
1.957006	1.872866

〔表 4-3〕：非裔孩童發育過程之每階段所需時間：

時間單位：月	第一階段	第二階段	第三階段	第四階段
2 5 位非裔男孩	15.4438	8.8743	5.0953	5.755
2 3 位非裔女孩	10.8782	11.2411	6.1387	7.828

參考文獻：

- [1] Warren J. Ewens & Gregory R. Grant, Statistical Methods in
Bioinformatics: An Introduction.
- [2] Sheldon M. Ross, Stochastic Processes, 2nd.
- [3] C. D. Fuh, Annals of Statistics, 31, 942 (2003).