行政院國家科學委員會專題研究計畫　成果報告

女孩發育期分級的統計分析

計畫類別：個別型計畫
計畫編號：NSC92-2118-M-009-005-
執行期間：92 年 08 月 01 日至 93 年 07 月 31 日
執行單位：國立交通大學統計學研究所

計畫主持人：彭南夫

報告類型：精簡報告

處理方式：本計畫可公開查詢

中　華　民　國 93 年 11 月 1 日

行政院國家科學委員會補助專題研究計畫成果報告

※※※※※※※※※※※※※※※※※※※※※※※※
※　　　　　　　　　　　　　　　　　　　　　　※
※　　女 孩 發 育 期 分 級 的 統 計 分 析　　※
※　　　　　　　　　　　　　　　　　　　　　　※
※※※※※※※※※※※※※※※※※※※※※※※※

計畫類別：☑個別型計畫　　□整合型計畫
計畫編號：NSC92-2118-M-009-005-
執行期間：92年08月01日至93年07月31日

計畫主持人：彭南夫

共同主持人：

本成果報告包括以下應繳交之附件：
- □赴國外出差或研習心得報告一份
- □赴大陸地區出差或研習心得報告一份
- □出席國際學術會議心得報告及發表之論文各一份
- □國際合作研究計畫國外研究報告書一份

執行單位：國立交通大學統計研究所

中　華　民　國　90　年　10　月　31　日

行政院國家科學委員會專題研究計畫成果報告
女孩發育期分級的統計分析

計畫編號: NSC92-2118-M-009-005
執行期限:92年8月1日至93年7月31日
主持人:彭南夫
研究人員:
執行機構及單位:國立交通大學統計研究所

一、中文摘要

針對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。本報告的研究目的與結果有：1、針對青春期的非裔男孩及女孩的發育過程，我們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、我們估計出誤判的可能性大小。

Abstract

When exploring the problem of tanner stage, we try to use the Hidden Markov Model to solve it. This is because our data consist of missing and misjudged parts, and so we can not use the Markov Model directly. This report includes 1. The transition probabilities and the duration times of each stage of African boys and African girls. 2. We estimate the probabilities of misjudgements in each stage.

二、 Results and Discussions

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，本報告明確的研究目的有：

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。

資料部份：

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。

分析設計及方法：

本報告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。

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。

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。

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，在本報告中，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。另外，在我們取得的資料中，發現在發育階段轉變時，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 而言，真實狀態集合仍為S = {0,1,2,3,4,5}不會有所改變, 則 不會改變, 只有B會有所改變, 每個狀態 可能觀測到的值增加為6個，(即)，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。

理論部分：

隱馬可夫模型（Hidden Markov Model）：

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 …及由每個視察到狀態所發出之觀測值也以 …來表示，對於每個觀測值Oi而言，Oi只會與相同時間的真實狀態 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：

觀測值：　　O1　　O2　　O3　　O4

我們再定義　（即O為所有Oi的數列）

（即Q為所有qi的數列）

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：。

2、 觀測值空間是由M個不同的觀測代號組成之系統：。

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

4、觀測值之機率：

對於每個狀態 且 ，則 (在真正狀態是Si而觀測到的狀態為a)

且由所有 形成一個 的矩陣。

5、起始狀態之分配為向量 ，其中 。

我們將引用Baum-Welch提出的參數估計法來估計未知參數（P、B、 ），進而解決相關之問題，

步驟一：先假設這些參數 ， 及 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：

在給定{O}下，第一個時間點的狀態

在 之期望次數的比例。

其中，｛｝：指整體之觀測值數列，即｛｝= 。

：對於任何d及t，當 且 的次數。

：對於任何d及t，當 的次數。

：對於任何d及t，當 且它發出的觀測值為 的次數。

步驟三：，將步驟二估出的新參數 視為新的起始值，再代回步驟一、 二，並重覆執行步驟二直到 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 被 取代，則 ，所以當 (即達穩定收斂時)或是當 與 之變化極微小時，此時的 就是最合適說明此資料的最佳參數，即 為最大概似估計值(MLE；Maximum Likelihood Estimator)。

演算之結果分析：

在 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 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L2，且L2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 且第二階段發育至第二階段的機率記為 ，則 ，然後,

25位非裔男孩的參數估計結果

q1=1	q1=2	q1=3	q1=4	q1=5
1	0	0	0	0

	Sj=1	Sj=2	Sj=3	Sj=4	Sj=5
Si=1	0.7409970910	0.2590029090	0	0	0
Si=2	0	0.5492576790	0.4507423210	0	
Si=3	0	0	0.2149702442	0.7850297558	0
Si=4	0	0	0	0.3049495450	
	0.6950504550				
Si=5	0	0	0	0	1

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

23位非裔女孩的參數估計結果

q1=1	q1=2	q1=3	q1=4	q1=5	
0.99798012		0.00201988	0	0	0

	Sj=1	Sj=2	Sj=3	Sj=4	Sj=5	
Si=1	0.63229356	0.36770644	0	0	0	
Si=2	0	0.64416262	0.35583738	0		
Si=3	0	0	0.34840073	0.65159927	0	
Si=4	0	0	0	0.48901531	0.51098469	
Si=5	0	0	0	0	1	

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

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).