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

碩士論文

Division of Biostatistics, Graduate Institute of Epidemiology

College of Public Health

National Taiwan University

Master Thesis

探討染病同胞對確定與不確定訊息資料之統計分析：

合併整體訊息測度為權數的加權檢定方法

Statistical Analysis of Affected Sib Pairs Data with Complete and Incomplete IBD Information: Combining Weights Approach

張育通

Yu-Tung Chang

指導教授：戴　政 博士

黃崑明 博士

Advisors: John Jen Tai, Ph.D.

Jason Kunming Huang, Ph.D.

中華民國九十八年六月

June, 2009

誌謝

就讀國立台灣大學公共衛生學院生物醫學統計組是我人生重要也是最正確的選擇之一。首先，要感謝論文指導老師戴政教授與黃崑明老師在學生研究學問與寫作論文的過程中，仍於百忙中抽空為學生解惑，並給予正確的方向與寶貴的意見。更重要的是兩位老師提供生統組遺傳研究室珍貴的研究資源給予學生論文題目與方向進行研究，學生實在感到萬分的榮幸與感恩。此外，生統組陳秀熙教授、張淑惠老師及嚴明芳老師亦用心教導，引領學生能夠進入更深入、更專業的統計理論與應用實務。

本論文能夠完成，還要再感謝博士班學長姐們的鼓勵與協助，尤其是張敦程學長在研究及論文寫作、模擬程式方面更是給予很多的協助，也願意撥出時間一起討論論文遇到的問題與模擬程式的撰寫。當然還要感謝同學們：宗仁、宗禋、竺諺、慶勳、芳儀、至紋在這兩年中一起念書、討論，使我的生活更加豐富。

最後，要感謝我的父母及所有家人一路上都給予支持與無私的付出，及南勢消防分隊的同仁不斷的給予鼓勵，都是我得以完成學業的勇氣與動力的來源。

張育通 謹識于國立台灣大學

中華民國九十八年六月

摘要

染病同胞對資料中，不同的親子基因型組合依其 IBD 訊息可分為完整訊息、不確定訊息及完全無訊息三種，當存在不確定訊息與完全無訊息的同胞對資料時，使用傳統均值檢定方法有可能造成統計檢定力的降低。為了提高統計檢定力，本研究利用 Franke and Ziegler(2005)所提萃取 IBD 確定訊息的 de Finetti 法及張(2006)萃取 IBD 不確定訊息的熵方法得到染病同胞對的 IBD 合併訊息量，以有效利用不確定訊息與完全無訊息的樣本，並由此整體訊息量建構出以確定與不確定訊息為基礎的二種加權均值檢定統計量來提升檢定力。經由模擬與過去均值檢定統計量比較，本研究之二種加權檢定統計量在型 I 錯誤率的表現與過去的檢定統計量相當，而檢定力的表現優於過去研究的方法，確實改善過去研究無法更精確利用不同樣本訊息造成檢定力不足的缺點。

關鍵字：IBD 訊息、不確定訊息、熵、加權方法

Abstract

A variety of test statistics can be applied to analysis of affected sib-pairs (ASP) data. However, if the data contain ASP with incomplete information, these statistics might perform in power worse than expected. According to the recent studies, it has been shown that weighting by IBD information could increase the statistical power. Thus, in this study we use the de Finetti method, which extracts the complete IBD information proposed by Franke and Ziegler (2005), and the entropy method, which extracts the uncertain IBD information by Chang (2006), to establish a new weighting scheme to gain extra power. Based on the new weighting scheme, we constructed two information-based weighted mean test statistics. We performed simulation studies for evaluating type I error rates and statistical powers of the two weighted mean test approaches. Simulation results showed that the two weighting approaches do increase statistical powers in various genetic models.

Keywords：IBD information, uncertain information, entropy, weighting scheme

目錄

圖目錄

表目錄

第一章 緒論

十九世紀達爾文的物種原始論(Origin of Species)與孟德爾的碗豆實驗及遺傳定律可說是二十世紀遺傳學發展的先驅。經過近一世紀的發展，以族群的觀點來進行遺傳分析方法的族群遺傳學也趨於發展完備。

基因定位(gene mapping)技術的發展是近代人類遺傳學突破的重要推手。1911 年 Morgan 對果蠅的研究，使得染色體遺傳的理論在性狀(trait)連結(linked)及互換(recombination)現象有了突破。由 Morgan 的啟發，Sturtevant 於 1913 年發表了有關基因定位的第一篇文章，是利用性狀間的距離與互換之間的訊息來進行定位的工作。延續古典遺傳分析的概念，使用已知基因位置的 DNA 標識基因(DNA marker)來對性狀或疾病基因進行定位工作，配合不斷發展的高速運算電腦及遺傳統計的相關分析或連鎖分析方法，間接證實其性狀與 DNA 標識基因有相關性，而更有效率的來進行基因定位的工作(Ott, 1999; 戴,2002)。

以往處理的遺傳資料多以完整的結構與訊息為基礎，基本的方法學也在這樣的情境下發展。然而，由於家族資料中遺傳訊息常無法被實際觀察到，因此不確定性(uncertainty)與疾病的複雜度往往造成分析上的困難(Thompson, 2005)。進階處理這類資料結構的方法也因

此逐漸發展(Kruglyak, 1996; Jacobs et al., 2003; Franke et al., 2005; Kulle et al., 2008; Ray et al., 2008)，但其中仍有許多的難題需要研究與討論。

1.1 連鎖分析

連鎖分析主要是討論疾病基因座與標識基因座間的連鎖狀態。若疾病基因座與標識基因座距離越遠，細胞進行減數分裂並將遺傳物質下傳到子代的過程中，二基因座會有較大的機率因染色體互換，使得二組遺傳訊息不一定會同時下傳至子代。反之，若二基因座因距離很近，甚至可以將二基因座視為同一基因座時，二組基因不同時下傳至子代的機率就顯得很小。因此，這種互換發生的機率大小與基因座之間的距離遠近成反比。

連鎖分析的發展是由直接分析的方法(direct approach)起步。直接分析的方法主要是觀察某一親代交配型的樣本中，計算所有發生互換與未發生互換現象的可能子代基因型樣本數，並將其結果估計出互換率(recombination fraction)。假設某一親代交配型之下，可能的各種子代基因型總樣本數為 n，若親代發生互換所產生的子代樣本個數為 k，則我們可以估計某一親代交配型個體發生互換的機率為 $\hat{\theta} = k/n$。過去所用來表示二基因座實體距離的方式是以鹼基對

(base pairs, bp)或摩根(Morgan, M)為距離單位，但由於互換率的大小受二基因座間距離遠近的影響，因此可以互換率這種機率單位來作為描述實體距離的一種相對距離量測，且互換率的範圍介於 0 至 0.5 之間。根據互換率的定義，假設標識基因座與疾病基因座緊密連鎖，則很難觀察到互換的現象發生於二基因座間，因此其互換率越接近 0；反之，二基因座距離越遠，有如位於不同的染色體上，則二基因座間發生互換的現象越容易被發現，則互換率就趨近於 0.5。

另一種分析方法則為間接分析方法(indirect approach)。間接分析的概念是利用某些已知位置的標識基因座為基礎，透過相關研究(association study)來討論標識基因座與疾病基因座之間的相關程度，或者利用連鎖研究(linkage study)來討論標識基因座是否位於疾病基因座附近的方法來間接的證實標示基因座與疾病基因座連鎖的可能。

連鎖分析方法就是否以設定參數觀點分類，可以分為參數與非參數分析方法。所謂參數連鎖分析方法是估計出互換率θ，並檢定此互換率是否等於 0.5，以判定疾病基因座與標識基因座間是否連鎖。這種參數化的連鎖分析方法，參數訊息來自樣本中有互換率訊息的雙異質(double heterozygote)樣本，並要預設其未知的遺傳模式再進行連鎖分析，因此有許多避開設定連鎖訊息參數的非參數化連鎖分析方

法被發展出來，其中一種重要的方法就是 Penrose(1935)所提出的染病同胞對(affected sib-pairs, ASP)方法。

1.1.1 非參數化連鎖分析方法：染病同胞對

參數分析方法是以設定參數來代表基因座之間的相近程度。非參數分析方法，則是直接比較疾病基因與標識基因一起下傳或不一起下傳的差異，來判別連鎖的可能，與參數分析方法之差異在於非參數分析方法並未設定參數，而是由觀察到的結果來進行分析。

非參數分析方法是利用同源全等基因(genes identical by descent, 簡稱 IBD 基因)數目的比較，得以檢視疾病基因與標識基因下傳之差異。欲進行 IBD 基因之比較，必須鑑取至少有二子代的家庭資料，再利用子代的基因型推算 IBD 基因之個數。為了更有效率的進行分析，Penrose(1935)提出利用染病同胞對資料較大機會出現疾病基因聚集且各家庭間互相獨立之特性，比較二染病同胞對標識基因座之間 IBD 基因的數目，得以間接證明連鎖的可能。圖 1.1 為單一家庭染病同胞對結構圖：

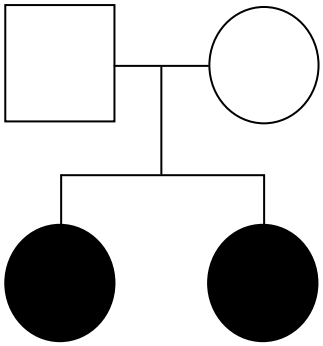


圖 1.1 染病同胞對家庭結構圖

1.1.2 不確定同源全等基因分析方法

對於遺傳連鎖分析的方法，主要是鑑取有較多訊息的家庭或家族資料來進行分析。在實際的資料當中，複雜疾病的完整基因訊息不易取得，或罕見疾病的發生率過低以至於不易觀察，亦或父母親資料未收集或遺失等困難，使得無法推算出正確的 IBD 數值，造成 IBD 基因訊息的不完整(incomplete information)。一個解決的方法就是利用現有的資料，也就是所觀察到的資料來估計不確定或完全無訊息的 ASP 家庭資料的 IBD 數目分布(戴, 2002; Franke and Ziegler,2005; 張,2006)。

以單標識基因座為例，假設標識基因座上有兩個共顯性對偶基因 B_1 和 B_2，其基因頻率為 r 和 s。在這樣的情境之下，以父母的交配型為 $B_1B_1 \times B_1B_2$ 的情況為例，說明如何計算 IBD 基因數目的機率分布及平均 IBD 數目。$B_1B_1 \times B_1B_2$ 可以表示成 $B_1^1B_1^2 \times B_1^3B_2^4$，上標的編號 1 和 2 代表父親標識基因座第一個與第二個對偶基因位置，編號 3 和 4 則代

表母親標識基因座第一個與第二個對偶基因位置(也就是將父母的標識基因座對偶基因位置一字排開，並給予這四個位置一個編號，如1,2,3,4)。為了估計所有子代的 IBD 基因數目，必須先將親代所有可能下傳對偶基因至第一個染病子代與第二個染病子代的組合排列出一個 4x4 的表格，其中列為父親下傳給染病同胞對的第一個對偶基因組合，行則為母親下傳給染病同胞對的第二個對偶基因組合，可整理如下：

表 1.1 親代交配型 $B_1B_1 \times B_1B_2$ 所有可能子代基因型組合

		母			
		B_1^3, B_1^3	B_1^3, B_2^4	B_2^4, B_1^3	B_2^4, B_2^4
父	B_1^1, B_1^1	$(B_1^1 B_1^3, B_1^1 B_1^3)$	$(B_1^1 B_1^3, B_1^1 B_2^4)$	$(B_1^1 B_2^4, B_1^1 B_1^3)$	$(B_1^1 B_2^4, B_1^1 B_2^4)$
	B_1^1, B_1^2	$(B_1^1 B_1^3, B_1^2 B_1^3)$	$(B_1^1 B_1^3, B_1^2 B_2^4)$	$(B_1^1 B_2^4, B_1^2 B_1^3)$	$(B_1^1 B_2^4, B_1^2 B_2^4)$
	B_1^2, B_1^1	$(B_1^2 B_1^3, B_1^1 B_1^3)$	$(B_1^2 B_1^3, B_1^1 B_2^4)$	$(B_1^2 B_2^4, B_1^1 B_1^3)$	$(B_1^2 B_2^4, B_1^1 B_2^4)$
	B_1^2, B_1^2	$(B_1^2 B_1^3, B_1^2 B_1^3)$	$(B_1^2 B_1^3, B_1^2 B_2^4)$	$(B_1^2 B_2^4, B_1^2 B_1^3)$	$(B_1^2 B_2^4, B_1^2 B_2^4)$

依照戴(2002)與張(2006)中計算 IBD 基因數目的方法，可以計算出表 1.1 中十六種染病同胞對基因型組合的 IBD 基因數目。推算的結果如表 1.2：

表 1.2 染病同胞對 IBD 基因數目列表

		母			
		B_1^3, B_1^3	B_1^3, B_2^4	B_2^4, B_1^3	B_2^4, B_2^4
父	B_1^1, B_1^1	2	1	1	2
	B_1^1, B_1^2	1	0	0	1
	B_1^2, B_1^1	1	0	0	1
	B_1^2, B_1^2	2	1	1	2

根據表 1.2 的結果，可以計算出此親代交配型之下，各基因型之染病同胞對的 IBD 基因數目的機率分布及估計出平均 IBD 基因數目。在（$B_1B_1 \times B_1B_2$；B_1B_1, B_1B_1）中，$B_1B_1 \times B_1B_2$ 代表親代交配型，分號右側的 B_1B_1, B_1B_1 代表在 $B_1B_1 \times B_1B_2$ 的親代交配型之下一種可能的染病同胞對基因型組合。在這樣的親代交配型與子代基因型組合，欲得到 IBD 基因個數，則要先計算出在這樣的組合之下 IBD 基因數目的機率分布。令 IBD 數目為 2、1 和 0 的機率分別為 f_2 、 f_1和 f_0則在此親子基因型組合之下，

$$f_2 = P(IBD = 2) = 2/4 = 1/2$$
$$f_1 = P(IBD = 1) = 2/4 = 1/2$$
$$f_0 = P(IBD = 0) = 0$$

由此可計算出 IBD 基因個數為 $\overline{IBD} = 2 \times (1/2) + 1 \times (1/2) + 0 \times (0) = 3/2$ 。

依照($B_1B_1 \times B_1B_2$; B_1B_1 , B_1B_1)計算 IBD 機率分布與平均 IBD 數目的方法，繼續推算在 $B_1B_1 \times B_1B_2$ 的親代交配型之下，染病同胞對基因型組合為 B_1B_1 , B_1B_2 與 B_1B_2 , B_1B_2 的 IBD 機率分布與平均 IBD 基因數目，($B_1B_1 \times B_1B_2$; B_1B_1 , B_1B_2)組合之下：

$$f_2 = P(IBD = 2) = 0$$
$$f_1 = P(IBD = 1) = 4/8 = 1/2$$
$$f_0 = P(IBD = 0) = 4/8 = 1/2$$
$$\overline{IBD} = 2 \times (0) + 1 \times (1/2) + 0 \times (1/2) = 1/2$$

($B_1B_1 \times B_1B_2$; B_1B_2 , B_1B_2)組合之下：

$$f_2 = P(IBD = 2) = 2/4 = 1/2$$
$$f_1 = P(IBD = 1) = 2/4 = 1/2$$
$$f_0 = P(IBD = 0) = 0$$
$$\overline{IBD} = 2 \times (1/2) + 1 \times (1/2) + 0 \times (0) = 3/2$$

利用相同的計算 IBD 基因數目方式，可以將兩共顯性對偶基因之單基因座之下，所有親代交配型與染病同胞對基因型整合並表列出共有十五個可能的組合，亦即有十五種家庭，每個家庭都有各自 IBD 基因數目的機率分布及 IBD 估計平均值，其結果表列於表 1.3。

表 1.3 親代交配型與同胞對基因型 15 種可能組合的 IBD 機率分布與 IBD 平均值

親代交配型與子代基因型組合（親代交配型；同胞對基因型）	頻率	f_2	f_1	f_0	IBD 平均值
(B_1B_2, B_1B_2； B_1B_1, B_1B_1)	$r^2s^2/4$	1	0	0	2
(B_1B_2, B_1B_2； B_1B_1, B_2B_2)	$r^2s^2/2$	0	0	1	0
(B_1B_2, B_1B_2； B_2B_2, B_2B_2)	$r^2s^2/4$	1	0	0	2
(B_1B_2, B_1B_2； B_1B_1, B_1B_2)	r^2s^2	0	1	0	1
(B_1B_2, B_1B_2； B_2B_2, B_1B_2)	r^2s^2	0	1	0	1
(B_1B_1, B_1B_2； B_1B_1, B_1B_1)	r^3s	1/2	1/2	0	3/2
(B_1B_1, B_1B_2； B_1B_1, B_1B_2)	$2r^3s$	0	1/2	1/2	1/2
(B_1B_1, B_1B_2； B_1B_2, B_1B_2)	r^3s	1/2	1/2	0	3/2
(B_2B_2, B_1B_2； B_2B_2, B_2B_2)	rs^3	1/2	1/2	0	3/2
(B_2B_2, B_1B_2； B_2B_2, B_1B_2)	$2rs^3$	0	1/2	1/2	1/2
(B_2B_2, B_1B_2； B_1B_2, B_1B_2)	rs^3	1/2	1/2	0	3/2
(B_1B_2, B_1B_2； B_1B_2, B_1B_2)	r^2s^2	1/2	0	1/2	1
(B_1B_1, B_1B_1； B_1B_1, B_1B_1)	r^4	1/4	1/2	1/4	1
(B_1B_1, B_2B_2； B_1B_2, B_1B_2)	$2r^2s^2$	1/4	1/2	1/4	1
(B_2B_2, B_2B_2； B_2B_2, B_2B_2)	s^4	1/4	1/2	1/4	1

觀察這十五種家庭的 IBD 基因數機率分布，若其(f_2, f_1, f_0)分布情形為(1/4,1/2,1/4)時，則此父母子代基因型對於推測同胞對 IBD 並無幫助，故對 IBD 檢定而言是無訊息資料(戴,2002;張, 2006; Wang, Hou and Tai,2008)。在此先導入 IBD 基因訊息的概念，依照 Franke and Ziegler(2005)所給予這十五種家庭 IBD 訊息的分類，若其分布為(0,0,1)、(0,1,0)或(1,0,0)，因為有確切的 IBD 基因個數，故歸類

為完整訊息(completely informative)；若其分布為(1/2,1/2,0)、(1/2,0,1/2)或(0,1/2,1/2)，則歸類為不確定訊息(incompletely informative)；若其分布為(1/4,1/2,1/4)有如無連鎖假設之下的分布，則是為完全無訊息(completely non-informative)之 IBD 基因數機率分布。

值得一提的是在推導過程中，發現將親代交配型與子代基因型表示如表 1.1 的形式，則其 IBD 基因數目的排列一定會如同表 1.2，顯著的減少了每次重新計算 IBD 數目的繁瑣過程。

1.2 合併方法

近一世紀的發展，遺傳統計學中檢定連鎖的方法不斷的被學者們研究並發表。利用資料結構所提供的訊息所發展出的連鎖分析方法，如利用三元體資料親代標識基因下傳與否的訊息來進行檢定連鎖的傳遞不平衡檢定方法(Terwilliger and Ott, 1992; Spielman, McGinnis and Ewens, 1993)，或利用染病同胞對四元體資料中 IBD 訊息的均值檢定方法等。上述的兩種檢定的方法在檢定力方面都表現不俗，但仍有許多的研究受限於家族資料取得不易，或複雜疾病的資料稀有對於分析連鎖的效力有限(Dempfle and Loesgen, 2003)。因此，許多合併的方法也一一被提出來，以提高有效樣本數或藉由合併

方法將相同資料中的相異訊息整合以建構穩健的統計量或提高檢定力(Tippett, 1931; Fisher, 1932; Spielman et al., 1993; Huang and Jiang, 1999; Dempfle and Loesgen, 2003; Tai and Hou, 2004; Wang et al. 2008)。

在遺傳統計學中，合併的方法大致可分為事前合併(pre-combination)與事後合併(post-combination) 二類(Wang et al. 2008; 邱,2008)。所謂事前合併方法是將不同資料結構的訊息，如 Martin et al.(2000)的家族不平衡檢定(pedigree disequilibrium test, PDT)是合併同一家庭中親子三元體資料與同胞對資料所建構的檢定方法，或同一資料結構中的不同訊息，如 Spielman et al.(1993)的連鎖不平衡檢定是合併三元體家庭資料中疾病基因座與標識基因座同時下傳/不下傳訊息(transmission/disequilibrium information)和同源全等基因訊息所建構穩健的檢定統計量，再利用合併訊息後所建構的檢定統計量進行檢定的概念。而事後合併方法則是合併不同訊息的檢定統計量，建構新的統計量再進行檢定的概念，如 Tai and Hou (2006)利用 Tippett 合併法(Tippett, 1931)與 Fisher 合併法(Fisher, 1932)合併四元體資料中傳遞不平衡檢定方法(transmission/disequilibrium test, TDT 檢定)與均值檢定(mean test)方法建構穩健的檢定統計量。

1.3 研究動機

標識基因被廣泛的應用於基因定位的工作，但是最近的研究發現，在處理實際資料時往往無法充分的利用到所給定的標識基因訊息。原因在於標識基因座對偶基因數目不夠多，或由於父母資料未收集或遺失等等，造成計算同胞對 IBD 基因數目時發生困難。這種不確定同胞對(ambiguous sib-pair)的 IBD 值以及其他訊息不完整的情形，引發諸多學者的研究與討論(Hodge et al., 1999;Jacobs et al., 2003; Franke and Ziegler, 2005; 張,2006)。

過去有許多的方法來處理這樣不完整訊息。一種是只取用完整訊息的同胞對資料來進行分析。若只取用完整訊息的樣本，會造成其他不確定訊息樣本的損失及統計檢定力的下降。最近有許多學者提出解決的方案(Zinn-Justin et al., 1999;Dempfle and Loesgen, 2003; Shete et al., 2003; Franke and Ziegler, 2005; 張,2006)，其中以染病同胞對所提供的訊息來進行加權，得到加權統計量後進行檢定的方法，已經得到許多證據的證實足以增加統計的檢定力。

此外，不同的統計檢定方法針對不同的訊息來進行檢定，或相同的資料結構中其實存在不同的訊息。因此，訊息的合併方法也是近年連鎖研究方法發展的一宗。若能利用不確定訊息的資料結構的訊息探

勘(information mining)概念，以獲取更多的訊息來改善不確定的情形，也會是一個重要的突破。

1.4 研究目的

本論文主要是延續 Franke and Ziegler(2005)以及張(2006)所提出加權均值檢定的加權方法，並將這二個加權方法在染病同胞對中所萃取的同源全等基因訊息進行合併，亦即將前者所萃取的 IBD 基因訊息視為一種含有多少 IBD 確定訊息的描述，而後者所萃取的 IBD 訊息視為一種不確定的 IBD 訊息量。由於兩者都是針對染病同胞對資料進行連所的檢定，因此若合併兩種訊息得到的一個新的訊息描述，使得這一個新的訊息就可以同時描述確定與不確定的情形，也就可以將這一個新的訊息視為染病同胞對 IBD 基因的整體訊息。

論文是以加權訊息的合併為主。Franke and Ziegler(2005)所提出的方法，可以視為以確定訊息的量測為出發點，但其結果卻顯示在完全無訊息的染病同胞對資料中，其加權訊息無法被辨識，也可說是無法利用這種完全無訊息的資料，造成有效樣本數與訊息的損失。張(2006)是以不確定訊息量測的訊息熵(Entropy)為出發，萃取染病同胞對同源全等基因的不確定訊息量。但由結果顯示訊息熵對於完整訊息的描述能力較弱，而將不確定訊息量調整為確定訊息的描述，會得

到與 Franke and Ziegler(2005)相同的結果。同樣的損失了應該可利用的訊息與樣本數。因此，期望可以藉由整體訊息量的概念，將加權訊息調整並合併，以提高加權均值檢定統計量的檢定力。

第二章 文獻回顧

實際資料的分析過程，由於染病同胞對不完整訊息的存在，即使利用多點連鎖分析方法來進行基因定位仍是無法充分利用到樣本的訊息而有效的提高檢定力。Dempfle and Loesgen(2003)所發表的研究指出，利用 meta-analysis 來整合連鎖分析方法用在複雜疾病(complex disease)的結果，發現近年使用樣本訊息來做為加權依據的加權檢定統計方法可以有效的增加檢定力。有鑑於此，本研究基於均值檢定(mean test)的加權方法，針對加權方法進行訊息的整合，以期望得到較高的檢定力來檢定疾病基因座與標識基因座是否發生連鎖。

2.1 資料結構與符號定義

2.1.1 資料結構

本研究所使用的資料結構為四元體的染病同胞對資料，即每一個家庭中皆有四名成員：父親、母親及一對染病的子代，家庭與家庭之間獨立。

2.1.2 符號定義

(A,a)：疾病基因座上二個對偶基因，「A」為正常基因，「a」為疾病

基因，且(A,a)的基因頻率為(p,q)。

(B_1, B_2)：標識基因座上的二個共顯性對偶基因，「B_1」為感興趣之待測定基因，且(B_1, B_2)的基因頻率為(r,s)。

(f_2, f_1, f_0)：在虛無擬說為二基因座不連鎖之下，會出現 IBD 基因數為 2、1、0 的機率分布。其中下標 2、1、0 代表 IBD 基因數。若為$(\hat{f}_2, \hat{f}_1, \hat{f}_0)$則代表估計的 IBD 基因數機率。

$\hat{\tau}_i$：每一個同胞對被觀察到的 IBD 基因數比例(observed proportion of alleles shared IBD for sib pair i)。其中$\hat{\tau}_i = \hat{f}_{2i} + \hat{f}_{1i}/2$。且每一同胞對 IBD 基因數比例的期望平均值$E(\hat{\tau}_i) = 1/2$。

θ：互換率(recombination fraction)定義為所有配子中互換的比率，其值介於 0 到 0.5 之間。當互換率θ越接近 0 時，代表兩基因座越靠近，越不易發生互換，因此互換率也可以視為兩基因座間實體距離的相對距離。

2.2 均值檢定

一般進行非參數化的連鎖分析檢定方法，主要是參考子代標識基因 IBD 基因數目的分布情形與不連鎖假設(H_0：θ=1/2)之下 IBD 基因數目分布(1/4：1/2：1/4)的偏離情形，且以這樣擾動的結果來說明染病同胞對資料可用於檢定連鎖(戴,2002)。利用這樣的概念，發展

出均值檢定、適合度檢定及二項檢定等非參數檢定連鎖的方法。而本研究主要是以均值檢定為主。

傳統的均值檢定可由兩種方式呈現，但基本的精神不變。根據戴(2002)所述，首先隨機抽取 n 對染病同胞對(每個家庭只有一對同胞對，故有 n 個家庭)，且出現 IBD 基因數目為 2、1、0 的樣本數分別為 n_2、n_1、n_0，使得 $n=n_2+n_1+n_0$。虛無擬說為 H_0：$\theta=1/2$ 之下，可以根據 IBD 基因數 2、1、0 以及其機率分布計算出一個染病同胞對的 IBD 基因數目的平均值與變異數分別為

$$E(IBD)=2(1/4)+1(1/2)+0(1/4)=1$$

$$V(IBD)=(2-1)^2(1/4)+(1-1)^2(1/2)+(0-1)^2(1/4)=1/2$$

若樣本數 n 夠大，依據中央極限定理，就可以標準常態分布為基礎來檢定觀察到的 IBD 基因總數 $\hat{IBD}_T$ 與期望的 IBD 基因總數 $E(IBD_T)$ 是否相等。則計算 IBD 基因總數之方法與檢定統計量如下：

$$\hat{IBD}_T = 2\times n_2 + 1\times n_1 + 0\times n_0$$

$$E(IBD_T) = 1\times n = n$$

$$V(IBD_T) = (1/2)\times n = n/2$$

$$Z = \frac{\hat{IBD}_T - E(IBD_T)}{\sqrt{V(IBD_T)}}$$

$$= \frac{2n_2 + n_1 - n}{\sqrt{n/2}} \quad \sim Z$$

而另一種均值檢定的表達方式於 Franke and Ziegler(2005)則是利用 2.1 節所定義之 IBD 基因數機率分布(f_2, f_1, f_0)與所觀察到的 IBD 基因數比例 $\hat{\tau}_i$ 來進行均值檢定。同樣的隨機鑑取 n 對染病同胞對(來自 n 個家庭)，且估計每對染病同胞對 IBD 基因數目為 2、1 的機率為 $\hat{f}_2$ 與 $\hat{f}_1$。則可以計算出每對同胞對的 IBD 基因數目比例，並和虛無擬說為兩基因座無連鎖的期望 IBD 基因數目的平均值做比較，可以表示成：

$$\bar{\hat{\tau}} = \frac{1}{n}\sum_{i=1}^{n}\hat{\tau}_i$$

$$E(\bar{\hat{\tau}}) = 1/2$$

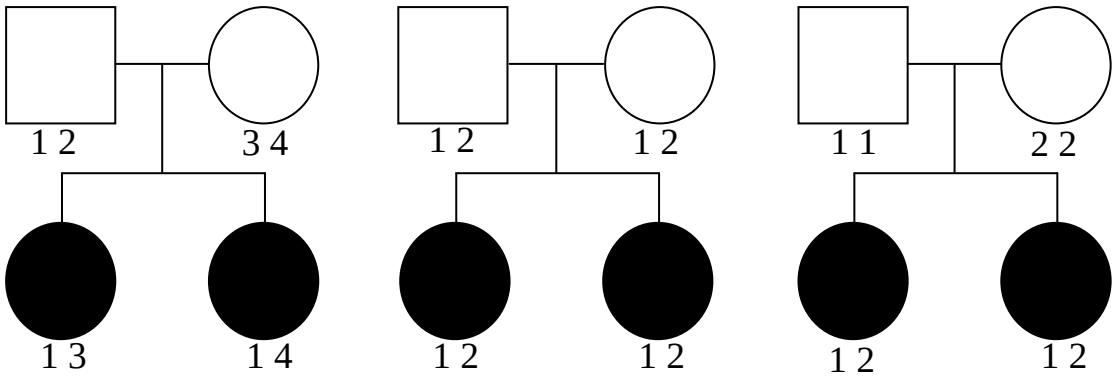
則均值檢定的檢定統計量可以表示成：

$$T_m = \frac{\bar{\hat{\tau}} - \frac{1}{2}}{\sqrt{\widehat{Var}(\bar{\hat{\tau}})}} \sim z$$

2.3 加權均值檢定

在不確定染病同胞對資料中，某些親代交配型之下的同胞對 IBD 基因數目無法被確切的計算出來。圖 2.1 所呈現的三種染病同胞對家庭的 IBD 基因數目。其中，數字的部份(1~4)表示為標識基因座上對偶基因的編號。由圖中可以看出三種不同的染病同胞對家庭可以提供的 IBD 訊息。其中(b)和(c)這二種家庭無法確定其 IBD 基因的數目，

使得在利用傳統的均值檢定時這樣的家庭無法提供有效的訊息。



(a)IBD=1　　(b)IBD=0 or 2　　(c)IBD=0,1,or 2

圖 2.1 三種不同的染病同胞對家庭所提供的 IBD 基因訊息

根據前一節所提到的均值檢定方法，其中所觀察到 IBD 基因數目比例的平均值式子中，

$$\bar{\hat{\tau}} = \frac{1}{n}\sum_{i=1}^{n}\hat{\tau}_i$$

可以看出不論是在哪一個家庭，所得到的權數都是 1/n。因此無法凸顯出不同染病同胞對家庭所提供訊息量的不同，因而降低了統計量的檢定能力。因此，最近許多學者針對訊息量的研究，無論是連鎖分析或相關分析，都發表了以訊息加權的方式來增加統計檢定力的方法。本研究也是基於這樣的觀念，並參考了 Franke and Ziegler(2005)以及張(2006)的兩種訊息量的加權方法於 2.3.1 節及 2.3.2 節中介紹。

2.3.1 de Finetti 三角加權法

前文所提到，過去利用染病同胞對資料中的 IBD 訊息來檢定連鎖的狀態。所謂 IBD 訊息是指親代交配型與同胞對基因型所能決定同胞對的 IBD 基因數值為何的確切性。利用圖 2.1 可以輔助了解 IBD 基因訊息的概念。圖 2.1(a)中，親代交配型與子代基因型的組合可以計算出 IBD 基因數目為 1，但是圖 2.1(b)及圖 2.1(c)的親代交配型與子代基因型卻無法確切計算出 IBD 基因數目。根據這樣的結果，Franke and Ziegler(2005)將圖 2.1 的三種不同的結果歸類為三種 IBD 基因訊息。依據 IBD 基因訊息量的多寡，可以分成：

(a)完整訊息(completely informative)

(b)不確定訊息(incompletely informative)

(c)完全無訊息(completely non-informative)

除此之外，Franke and Ziegler(2005)利用座標平面上兩點間的歐幾里得距離(Euclidian distance)，做為另一種描述 IBD 訊息量的方式。

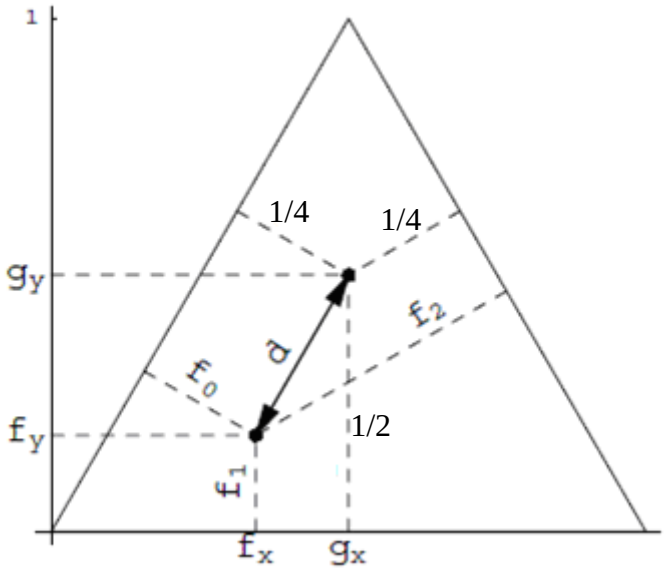


圖 2.2 de Finetti 三角形概念圖

上圖是節錄自 Franke et al.(2005)的研究，主要是用來描述整個 de Finetti 三角形的概念。首先，在一個二維的座標平面上，建立一個高為 1 的正三角形，利用正三角形的 Viviani’s theorem (Vincenzo Viviani,1622-1703)，也就是在正三角形中的任一點，至三邊垂線長度和等於該正三角形的高的性質，則可假想每一個染病同胞對家庭為該正三角形中的一點，該點至三邊的垂線長度可視為該家庭中染病同胞對 IBD 基因數目的分布，即 $f_2+f_1+f_0=1$，其分布和為 1 恰好與該三垂線長度和為 1 相同。

圖 2.2 中，假設某一家庭的 IBD 基因分布為 $f=(f_2, f_1, f_0)$，將該分布表示於圖中可以得到一個座標為 $f=(f_x, f_y)$，則可利用染病同胞對 IBD 基因數目分布偏離 H_0 無連鎖之下，IBD 基因數目分布(1/4,1/2,1/4)之擾動情形來說明染病同胞對資料可用於檢定連鎖的

概念；另外在正三角形中設定一點 g=(1/4,1/2,1/4)符合無連鎖之虛無擬說，用以和各家庭 IBD 基因數目分布的點來做比較，如圖 2.2 之 $g=(g_x, g_y)$。利用歐幾里得距離來計算出 f 及 g 這兩點的距離 d(f)，

$$d(f)=\sqrt{(f_x-g_x)^2+(f_y-g_y)^2}$$

且經過基本的幾何概念可以將 f 與 g 表示為，

$$f=(f_x, f_y)=(\frac{2\sqrt{3}}{3}f_0+\frac{\sqrt{3}}{3}f_1, f_1)$$

$$g=(g_x, g_y)=(\frac{\sqrt{3}}{3}, \frac{1}{2})$$

再代入 d(f)中，

$$d(f)=\sqrt{\left(\frac{2\sqrt{3}}{3}f_0+\frac{\sqrt{3}}{3}f_1-\frac{\sqrt{3}}{3}\right)^2+\left(f_1-\frac{1}{2}\right)^2}$$

$$=\sqrt{\frac{1}{3}(2f_0+f_1-1)^2+\frac{1}{4}(1-2f_1)^2}$$

$$=\sqrt{\frac{1}{3}(2f_0-f_2-f_0)^2+\frac{1}{4}(1-2f_1)^2}$$

$$=\sqrt{\frac{1}{3}(f_2-f_0)^2+\frac{1}{4}(1-2f_1)^2}$$

而該距離也就是 Franke and Ziegler(2005)所用來描述 IBD 訊息並用以加權的依據。

將所有染病同胞對資料的 IBD 基因數目分布用 d(f)來轉換，每一對染病同胞對皆可以得到一個 $d(f_i)$，也就是該染病同胞對的 IBD 訊息。要將這些訊息加權時，其權重需要轉換成標準化歐幾里得距離，

$$w_i = \frac{d(f_i)}{\sum_{j=1}^{n} d(f_j)}$$

且

$$\sum_{i=1}^{n} w_i = 1$$

其中，n 代表有 n 對染病同胞對，i 代表第 i 對。

根據 Franke and Ziegler(2005)的 IBD 訊息描述，下表 2.1 依據 IBD 基因訊息量分組，整理出單標識基因座(二共顯性對偶基因)染病同胞對資料各種親代交配型與子代基因型可能組合之 IBD 基因數目分布與 de Finetti 三角加權訊息量。

表 2.1 de Finetti 三角加權訊息整理

IBD 基因訊息	親代交配型與子代基因型組合	f_2	f_1	f_0	de Finetti d(f)
完整訊息	(B_1B_2, B_1B_2； B_1B_1, B_1B_1)	1	0	0	0.76376
	(B_1B_2, B_1B_2； B_1B_1, B_2B_2)	0	0	1	0.76376
	(B_1B_2, B_1B_2； B_2B_2, B_2B_2)	1	0	0	0.76376
	(B_1B_2, B_1B_2； B_1B_1, B_1B_2)	0	1	0	0.5
	(B_1B_2, B_1B_2； B_2B_2, B_1B_2)	0	1	0	0.5
不確定訊息	(B_1B_1, B_1B_2； B_1B_1, B_1B_1)	1/2	1/2	0	0.28868
	(B_1B_1, B_1B_2； B_1B_1, B_1B_2)	0	1/2	1/2	0.28868
	(B_1B_1, B_1B_2； B_1B_2, B_1B_2)	1/2	1/2	0	0.28868
	(B_2B_2, B_1B_2； B_2B_2, B_2B_2)	1/2	1/2	0	0.28868
	(B_2B_2, B_1B_2； B_2B_2, B_1B_2)	0	1/2	1/2	0.28868
	(B_2B_2, B_1B_2； B_1B_2, B_1B_2)	1/2	1/2	0	0.28868
	(B_1B_2, B_1B_2； B_1B_2, B_1B_2)	1/2	0	1/2	0.5
完全無訊息	(B_1B_1, B_1B_1； B_1B_1, B_1B_1)	1/4	1/2	1/4	0
	(B_1B_1, B_2B_2； B_1B_2, B_1B_2)	1/4	1/2	1/4	0
	(B_2B_2, B_2B_2； B_2B_2, B_2B_2)	1/4	1/2	1/4	0

由上表可以發現 IBD 基因訊息的多寡和 de Finetti 三角加權量成正比(完整訊息權重>不確定訊息權重>完全無訊息權重)，可以視為一種描述確定訊息量多寡的方法。但在完整訊息與不確定訊息兩組中，會發現在完整訊息中的(B_1B_2, B_1B_2； B_1B_1, B_1B_2)及(B_1B_2, B_1B_2； B_2B_2, B_1B_2)與不確定訊息中(B_1B_2, B_1B_2； B_1B_2, B_1B_2)的親代交配型與子代基因型組合之 de Finetti 三角加權量與該組的加權量並不一致，則此加權方法的穩定性有改善的空間。此外，且完全無訊息的權數等於 0，

會造成有效樣本數的損失。為了與之後研究所使用的另一種加權方式比較及合併，將 Franke and Ziegler(2005)所使用的加權量 w_i 改寫成 w_{Fi}，此加權檢定統計量為，

$$T_{w_F} = \frac{\bar{\hat{\tau}}_{w_F} - \frac{1}{2}}{\sqrt{\frac{n'}{n'-1}\sum_{i=1}^{n} w_{Fi}^2 (\hat{\tau}_i - \bar{\hat{\tau}}_{w_F})^2}}$$

其中

$$\bar{\hat{\tau}}_{w_F} = \sum_{i=1}^{n} w_{Fi} \hat{\tau}_i$$

且 n’ 為非完全無訊息之樣本數。

2.3.2 訊息熵加權法

近年來，用於描述 IBD 訊息的方法不斷被開發出來，如前一節所提到的 de Finetti 三角加權法，也算是一種新穎的方法，利用視覺化(visualization)方式給予 IBD 訊息另一種新的依據。張(2006)則使用熵(entropy)的概念，運用在遺傳統計學當中。

熵最早是用在軍事情報方面，用來描述加密電報中字母出現的不確定性，並視為某一字母在情報來源中所負擔的平均情報量。張(2006)則是基於熵用來描述不確定程度的特性，將這樣的特性運用在染病同胞對資料當中。先前已經提到在染病同胞對的資料結構當中，考慮各

種親代交配型與子代基因型的組合，會出現不確定訊息與完全無訊息等這類的訊息不確定組合。因此，張(2006)先利用熵描述不確定性的能力將所有組合的不確定訊息量計算出，再基於 Franke and Ziegler(2005)研究中確定訊息量多寡與加權量成正比的想法，將不確定訊息經過調整轉換成以確定訊息量來做描述。

Shannon(1984)提出的 Shannon entropy：

$$H = -\sum_i p_i \log p_i$$

假設 Y 為一離散隨機變數，並服從以下之機率分布：

$$P_Y(y) = g(y)$$

由此機率分布定義 entropy 為：

$$H = -\sum_y P_Y(y) \log P_Y(y)$$

其中 $P_Y(y)$ 表示 Y=y 的機率。若運用在染病同胞對資料中，每一個家庭都會有一組 $P_Y(y)$ 的分布，對應這個分布的隨機變數 Y 定義域即為 IBD=0、1、2，其機率分布即為(f_2, f_1, f_0)。以完整訊息為例，其 IBD 基因頻率的分布為(1,0,0)、(0,1,0)或(0,0,1)三種可能，也就代表 $P_Y(y)$ 中僅會有一個元素，則此時不確定訊息量 H=0，亦即不確定性最低。

將所有親代交配型與子代基因型組合的不確定訊息量 H(P)計算

出來後，張(2006)為了直接使用 Franke and Ziegler(2005)的確定訊息加權方法的統計量，則需要將不確定訊息量轉換成確定訊息量描述。如下表示：

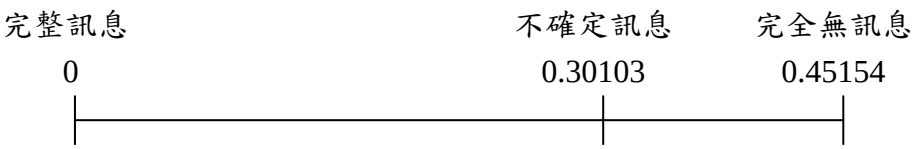


圖 2.3 不確定訊息量 H

上圖以數線的方式來表現完整訊息、不確定訊息與完全無訊息的分布。根據 entropy 所計算出完整訊息的不確定訊息量為 0，為了轉換成確定訊息描述，重新給予新的數線起始點，

圖 2.4 確定訊息量 e(f)

轉換以後，可以將結果整理如表 2.2。

表 2.2 熵加權訊息整理

IBD 基因訊息	親代交配型與子代基因型組合	f_2	f_1	f_0	Entropy H	Change(f)
完整訊息	(B_1B_2, B_1B_2 ; B_1B_1, B_1B_1)	1	0	0	0	0.45154
	(B_1B_2, B_1B_2 ; B_1B_1, B_2B_2)	0	0	1	0	0.45154
	(B_1B_2, B_1B_2 ; B_2B_2, B_2B_2)	1	0	0	0	0.45154
	(B_1B_2, B_1B_2 ; B_1B_1, B_1B_2)	0	1	0	0	0.45154
	(B_1B_2, B_1B_2 ; B_2B_2, B_1B_2)	0	1	0	0	0.45154
不確定訊息	(B_1B_1, B_1B_2 ; B_1B_1, B_1B_1)	1/2	1/2	0	0.30103	0.15051
	(B_1B_1, B_1B_2 ; B_1B_1, B_1B_2)	0	1/2	1/2	0.30103	0.15051
	(B_1B_1, B_1B_2 ; B_1B_2, B_1B_2)	1/2	1/2	0	0.30103	0.15051
	(B_2B_2, B_1B_2 ; B_2B_2, B_2B_2)	1/2	1/2	0	0.30103	0.15051
	(B_2B_2, B_1B_2 ; B_2B_2, B_1B_2)	0	1/2	1/2	0.30103	0.15051
	(B_2B_2, B_1B_2 ; B_1B_2, B_1B_2)	1/2	1/2	0	0.30103	0.15051
	(B_1B_2, B_1B_2 ; B_1B_2, B_1B_2)	1/2	0	1/2	0.30103	0.15051
完全無訊息	(B_1B_1, B_1B_1 ; B_1B_1, B_1B_1)	1/4	1/2	1/4	0.45154	0
	(B_1B_1, B_2B_2 ; B_1B_2, B_1B_2)	1/4	1/2	1/4	0.45154	0
	(B_2B_2, B_2B_2 ; B_2B_2, B_2B_2)	1/4	1/2	1/4	0.45154	0

由上表可以發現，轉換過後的 e(f)在不同訊息量組合的一致性較高，且其訊息量的多寡與加權的權重成正比(完整訊息權重>不確定訊息權重>完全無訊息權重)，符合了一般直觀的想法，且改善了 Franke and Ziegler(2005)在確定訊息與不確定訊息的某些組合權重的不穩定性。但與 Franke and Ziegler(2005)所得到的結果相同，在完全無訊息的組合中，其權重依然為 0，會造成有效樣本數的損失。在此定義由熵所量測出的不確定訊息量 H 為加權指標 W_{Hi}，將於研究方法中使用。

第三章 研究方法

本章將以事前合併的觀點對染病同胞對訊息合併為整體訊息進行討論。整體訊息的合併概念是源自於 Franke and Ziegler(2005)及張(2006)的研究，且可由描述訊息的方式歸納出二種不同性質的訊息：確定訊息與不確定訊息。Franke and Ziegler(2005)將染病同胞對資料中 IBD 基因數目的機率分布用二維座標平面表示，並利用歐式距離來描述 IBD 數目擾動的訊息，再將這訊息當成一個加權指標，稱為 de Finetti 三角加權法。對張(2006)所用的方法而言，Franke and Ziegler(2005)可以視為一種「確定訊息」的描述。張(2006)所用的方法，其結果與 Franke and Ziegler(2005)接近，但是回歸到原先使用的訊息熵方法的概念，確實是利用熵的特性來描述染病同胞對資料中 IBD 的「不確定訊息」。

Franke and Ziegler(2005)所提出的方法，也就是確定訊息的量測方法，其優點是具有創新的意義，保存了原本均值檢定中討論 IBD 基因數目的擾動的想法，也依據 IBD 基因數目的機率分布將這樣的資料結構分為三種不同的訊息量。如此的分組可以更清楚的表現出其直觀的想法：完整訊息量>不確定訊息量>完全無訊息量。只是結果顯示各分組內的訊息量並不一致，使得這樣的方法仍存在不穩定性。此外，由於距離的計算是將每一個家庭的 IBD 基因數目的機率分布與虛

無擬說為無連鎖的機率分布相比較，因此，在完全無訊息的家庭資料中，其分布恰好與無連鎖的假設相同，造成其描述訊息量時是無法被量測出來的而為 0。檢定時，訊息量當成加權的指標，若該訊息量為 0 存在，則會損失約百分之二十的染病同胞對樣本，不僅無法有效利用收集到的樣本，也對檢定力的提升沒有實質的幫助。

張(2006)所使用的方法，是用訊息熵的概念。其實在 1984 年 Hodge 就已經引入 selective information(即 entropy)的觀念來處理多染病同胞對的訊息，只是受限於必須能清楚判斷每個染病同胞對的 IBD 值，再利用 entropy 的概念計算每個家族內相依的多染病同胞對所能貢獻的獨立訊息。因此，當 IBD 訊息出現不確定時，Hodge(1984)的方法就受到限制。故張(2006)的方法，雖不算是非常新穎，但仍對 entropy 的用法有新的見解並有效處理 Hodge 當時所遭遇的限制。

張(2006)所使用的資料結構與 Hodge(1984)不同，因每個家庭只取一對染病同胞對，因此沒有同一家庭中各同胞對間存在相關。此外，更利用訊息熵的主要功能來描述染病同胞對資料結構中的不確定訊息。 經過 entropy 的萃取，張(2006)的研究結果顯示染病同胞對的不確定訊息確實能夠被測量出。不確定訊息量的多寡也很直觀的表現：完全無訊息>不確定訊息>完整訊息。完全無訊息這類的家庭，可以視為不確定訊息程度最高的家庭，因此經由 entropy 的測量結果其

值最大；而完整訊息的家庭，顧名思義，其 IBD 訊息已經完整，無不確定訊息的存在，因此其不確定訊息量為 0，而不確定訊息的家庭介於二者之間。

由於張(2006)最後所使用的檢定統計量與 Franke and Ziegler(2005)相同，因此其加權的指標需要經過轉換。轉換過後，訊息量的多寡與加權的量成正比，且分布也直觀。甚至在三種不同訊息量的家庭中，其加權指標是呈現一致的分布，亦即同一類的家庭得到相同的權數。這樣的結果明顯的優於 Franke and Ziegler(2005)在一致性上的問題，但仍然發現完全無訊息的家庭其加權值仍為 0，無法解決損失有效樣本的問題。

有鑑於 Franke and Ziegler(2005)和張(2006)所產生的問題，此研究將張(2006)所做的調整還原成 entropy 的訊息量，產生對於同一資料結構存在二種相反的訊息萃取方法，再與 Franke and Ziegler(2005)的結果合併，即可對於染病同胞對的 IBD 訊息有更完整的描述，也就是合併成為一個「整體訊息」。除此之外，對於樣本數的有效使用，也在合併的過程中經過一些調整改善了權重為 0 的情形。

有了整體訊息的概念，接下來要思考如何合併這樣相反的訊息，並且使得結果有足夠的合理性。因為對染病同胞對資料而言，二個相

反的訊息是無法直接將二者訊息加總的簡單。若將這兩個訊息量比喻為問卷中程度評分的問題，假設確定訊息的測量為問卷題目中感受良好的程度，則不確定訊息的量測就成為同樣的題目但是要評分感受不好的程度。二者的差異性於此表現出來，因此無法非常直觀的以相加或相減的形式來合併確定與不確定訊息成為整體訊息的概念。進行合併前，將 de Finetti 加權與 entropy 加權表列如下：

表 3.1 de Finetti 加權與 entropy 加權表

IBD 基因訊息	親代交配型與子代基因型組合	de Finetti d(f)	Entropy H
完整訊息	(B_1B_2, B_1B_2 ; B_1B_1, B_1B_1)	0.76376	0
	(B_1B_2, B_1B_2 ; B_1B_1, B_2B_2)	0.76376	0
	(B_1B_2, B_1B_2 ; B_2B_2, B_2B_2)	0.76376	0
	(B_1B_2, B_1B_2 ; B_1B_1, B_1B_2)	0.5	0
	(B_1B_2, B_1B_2 ; B_2B_2, B_1B_2)	0.5	0
不確定訊息	(B_1B_1, B_1B_2 ; B_1B_1, B_1B_1)	0.28868	0.30103
	(B_1B_1, B_1B_2 ; B_1B_1, B_1B_2)	0.28868	0.30103
	(B_1B_1, B_1B_2 ; B_1B_2, B_1B_2)	0.28868	0.30103
	(B_2B_2, B_1B_2 ; B_2B_2, B_2B_2)	0.28868	0.30103
	(B_2B_2, B_1B_2 ; B_2B_2, B_1B_2)	0.28868	0.30103
	(B_2B_2, B_1B_2 ; B_1B_2, B_1B_2)	0.28868	0.30103
	(B_1B_2, B_1B_2 ; B_1B_2, B_1B_2)	0.5	0.30103
完全無訊息	(B_1B_1, B_1B_1 ; B_1B_1, B_1B_1)	0	0.45154
	(B_1B_1, B_2B_2 ; B_1B_2, B_1B_2)	0	0.45154
	(B_2B_2, B_2B_2 ; B_2B_2, B_2B_2)	0	0.45154

故以下將分別以確定訊息與不確定訊息為基礎來進行訊息的合併與調整。3.1 節說明確定訊息為基礎的合併，3.2 節則是以不確定

訊息為基礎的合併。

3.1 以確定訊息為基礎的加權指標

過去研究結果顯示這二個工具測量染病同胞對資料的訊息是有一定的能力。由表 3.1 可以看出這二個工具的訊息程度分別在確定訊息與不確定訊息的分布。當 IBD 訊息為完整訊息時，確定訊息量最高為 0.76376，因為訊息完整，不確定訊息量皆一致為 0；當 IBD 訊息為不確定訊息時，確定訊息量除了一個親子組合為 0.5 外，其他都為 0.28868，而不確定訊息量與確定訊息量接近，為 0.30103；至於 IBD 訊息為完全無訊息時，因為完全無訊息，故確定訊息量為 0，不確定訊息量被量測為 0.45154。比較的結果發現此二種測量訊息的工具對於特定的訊息無法量測，而出現 0 的結果。由先前的回顧也說明了無法測量出訊息量對欲萃取更多訊息檢定連鎖沒有助益。因此，本節將以確定訊息為基礎來進行合併訊息。

以確定訊息為基礎，將確定訊息與不確定訊息合併，亦即先考慮如何解決 entropy 所測量的不確定訊息量在 IBD 訊息為完整訊息時為 0 的情況。而所謂以確定訊息為基礎，也就是將不確定訊息量改以確定訊息量的表示。其方法異於張(2006)之做法。張(2006)是直接調整訊息量之起始點(見第二章)，並沒有使用到其他訊息量。而本研究之

方法，是根據 de Finetti 三角形測量出的最大確定訊息量 0.76376 做為完整訊息的訊息量，則不確定訊息及完全無訊息的相對訊息量如圖 3.1 所示。

圖 3.1 以確定訊息為基礎的合併方法

合併訊息量的方法是將熵 entropy 的完整訊息量 0 調整為 de Finetti 方法的最大確定訊息量 0.76376。為了維持完整訊息、不確定訊息及完全無訊息間的距離，則調整後的不確定訊息量為 0.46273 (=0.76376-0.30103)，完全無訊息量調整後為 0.31222 (=0.76376-0.45154)由上圖的數線表示可以清楚的看出合併的結果。於此不使用 0、1 或其他數值來調整，而是利用確定訊息量測工具所測量出的訊息量，除了達到合併的效果外，亦可說明其合理性。將所得到的新的訊息量定義為W_{FHi}，其中 i 代表 IBD 基因訊息，

$$W_{FHi} = \begin{cases} 0.76376 & \text{,若 i=1 表完整訊息} \\ 0.46273 & \text{,若 i=2 表不確定訊息} \\ 0.31222 & \text{,若 i=3 表完全無訊息} \end{cases}$$

合併的結果顯示這新的合併訊息量W_{FHi}不僅修正了 Franke and

Ziegler(2005)使用 de Finetti 三角加權時完整無訊息的訊息量為 0 的情形，得以充分利用所收集的染病同胞對樣本。此外，由於張(2006)的方法優於 Franke and Ziegler(2005)之處在於每一種 IBD 訊息只有一種訊息量的存在，使得整體而言是穩定的訊息量，因此，利用確定訊息來合併與調整的結果亦會出現一致且穩定的狀態。下表整理出 IBD 基因訊息與合併訊息量。

表 3.2 以確定訊息為基礎的合併訊息量

IBD 基因訊息	親代交配型與子代基因型組合	合併訊息量 W_{FHi}
完整訊息	(B_1B_2, B_1B_2 ; B_1B_1, B_1B_1)	0.76376
	(B_1B_2, B_1B_2 ; B_1B_1, B_2B_2)	0.76376
	(B_1B_2, B_1B_2 ; B_2B_2, B_2B_2)	0.76376
	(B_1B_2, B_1B_2 ; B_1B_1, B_1B_2)	0.76376
	(B_1B_2, B_1B_2 ; B_2B_2, B_1B_2)	0.76376
不確定訊息	(B_1B_1, B_1B_2 ; B_1B_1, B_1B_1)	0.46273
	(B_1B_1, B_1B_2 ; B_1B_1, B_1B_2)	0.46273
	(B_1B_1, B_1B_2 ; B_1B_2, B_1B_2)	0.46273
	(B_2B_2, B_1B_2 ; B_2B_2, B_2B_2)	0.46273
	(B_2B_2, B_1B_2 ; B_2B_2, B_1B_2)	0.46273
	(B_2B_2, B_1B_2 ; B_1B_2, B_1B_2)	0.46273
	(B_1B_2, B_1B_2 ; B_1B_2, B_1B_2)	0.46273
完全無訊息	(B_1B_1, B_1B_1 ; B_1B_1, B_1B_1)	0.31222
	(B_1B_1, B_2B_2 ; B_1B_2, B_1B_2)	0.31222
	(B_2B_2, B_2B_2 ; B_2B_2, B_2B_2)	0.31222

3.2 以不確定訊息為基礎的加權指標

與 3.1 節相同的概念，仍然是要合併確定與不確定訊息。前一節是以確定訊息為基礎，取確定訊息的最大量為基準合併並調整不確定訊息的結果。因此，現在改以不確定訊息為基礎的合併與調整方法。

由於是以不確定訊息為基礎，因此和先前的做法相同，但是取用 entropy 的不確定訊息量。在不確定訊息量中，IBD 訊息為完整訊息時為 0，但完全無訊息為 0.45154，因此取完全無訊息的訊息量 0.45154 做為合併的基準，取代 de Finetti 之確定訊息的訊息量為 0 的部分。下圖為以不確定訊息為基礎的合併方法：

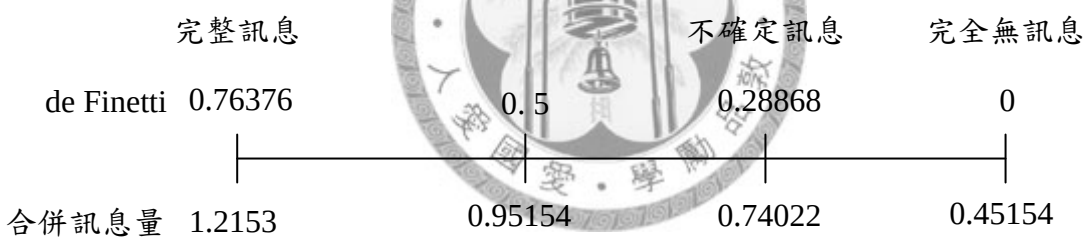


圖 3.2 以不確定訊息為基礎的合併方法

以不確定訊息為基礎的合併訊息方法仍可用上圖的數線來表示。熵 entropy 所計算出的不確定訊息量在完全無訊息時為 0.45154，則將此訊息量取代 de Finetti 之完全無訊息量。則為了維持合併訊息量中完整訊息、不確定訊息及完全無訊息之間此的距離，完整訊息量調整後為 1.2153 (=0.45154+0.76376)，不確性訊息量調整後為 0.74022 (=0.45154+0.28828)，及 de Finetti 訊息量為 0.5

時，調整後為 0.95154 (=0.45154+0.5)。此時，合併後的完全無訊息量則以不確定訊息的最大值 0.45154 取代，主要是在這樣的合併之下可以保留最完整的不確定訊息量，也基於此，合併後的訊息量如圖 3.2 所示。但必須討論之處是上圖中合併訊息量為 0.95154，也就是介於完整訊息與不確定訊息之間。存在這一個中間值的原因在於此方法是合併並調整 de Finetti 之確定訊息量，而由第二章的文獻回顧可知 Franke and Ziegler(2005)的方法並非完美，除了與張(2006)相同的無法完整利用樣本以外，仍有量測訊息時無法在相同的 IBD 訊息分類中得到一致的訊息量。也因此在某些親子交配型與基因型組合之下，是無法由 IBD 的訊息量來區分。

定義為W_{HFi}，其中 i 代表 IBD 基因訊息，

$$W_{HFi} = \begin{cases} 1.2153 \text{ ,若 i=1 表完整訊息} \\ 0.95154 \text{ ,若 i=1 且為}(B_1B_2,\ B_1B_2;\ B_1B_1,\ B_1B_2)\text{或} \\ \qquad (B_1B_2,\ B_1B_2;\ B_2B_2,\ B_1B_2) \\ \qquad \text{若 i=2 且為}(B_1B_2,\ B_1B_2;\ B_1B_2,\ B_1B_2) \\ 0.74022 \text{ ,若 i=2 表不確定訊息} \\ 0.45154 \text{ ,若 i=3 表完全無訊息} \end{cases}$$

雖然整體的訊息量分布仍符合直觀：完整訊息量>不確定訊息量>完全無訊息量，但這樣的不一致性仍然存在。下表為以不確定訊息為基礎的 IBD 訊息與合併訊息量。

表 3.3 以不確定訊息為基礎的合併訊息量

IBD 基因訊息	親代交配型與子代基因型組合	合併訊息量 W_{HFi}
完整訊息	(B_1B_2, B_1B_2 ; B_1B_1, B_1B_1)	1.2153
	(B_1B_2, B_1B_2 ; B_1B_1, B_2B_2)	1.2153
	(B_1B_2, B_1B_2 ; B_2B_2, B_2B_2)	1.2153
	(B_1B_2, B_1B_2 ; B_1B_1, B_1B_2)	0.95154
	(B_1B_2, B_1B_2 ; B_2B_2, B_1B_2)	0.95154
不確定訊息	(B_1B_1, B_1B_2 ; B_1B_1, B_1B_1)	0.74022
	(B_1B_1, B_1B_2 ; B_1B_1, B_1B_2)	0.74022
	(B_1B_1, B_1B_2 ; B_1B_2, B_1B_2)	0.74022
	(B_2B_2, B_1B_2 ; B_2B_2, B_2B_2)	0.74022
	(B_2B_2, B_1B_2 ; B_2B_2, B_1B_2)	0.74022
	(B_2B_2, B_1B_2 ; B_1B_2, B_1B_2)	0.74022
	(B_1B_2, B_1B_2 ; B_1B_2, B_1B_2)	0.95154
完全無訊息	(B_1B_1, B_1B_1 ; B_1B_1, B_1B_1)	0.45154
	(B_1B_1, B_2B_2 ; B_1B_2, B_1B_2)	0.45154
	(B_2B_2, B_2B_2 ; B_2B_2, B_2B_2)	0.45154

不論是以確定訊息為基礎或以不確定訊息為基礎的合併，皆為合併兩種訊息後再建構加權的檢定統計量，因此可以將 3.1 及 3.1 節所介紹的兩種方法視為一種事前合併的方法。下一張將利用模擬研究呈現這二種事前合併方法與 Franke and Ziegler(2005)之結果比較檢定力上的表現。

第四章 模擬研究

前文說明了染病同胞對資料之下，單標識基因座二共顯對偶基因所可以觀察到的 IBD 基因數目的機率分布，也依照 IBD 基因數目來區分了不同訊息量的家庭。本研究延續 Franke and Ziegler(2005)及張(2006)的二種訊息萃取方法，並加以整合成一種整體訊息的指標。在前面章節描述過合併的方法，顯示以確定訊息與不確定訊息為基礎的這二種合併訊息的方法直覺上皆有不錯的表現，不僅整合了訊息，也改善了先前研究所發生的不穩定及有效樣本數減少的問題。為了驗證本研究的方法可行，並且要與過去研究比較以證實表現更佳，故由模擬的方式印證此合併方法有效的提高檢定統計量的檢定力。

4.1 模擬設定

本研究所使用的模擬程式版權屬於台灣大學公共衛生學院生物醫學統計組遺傳研究室，並提供原始程式碼使得以修改延用至本研究。設定鑑取 200 個染病同胞對家庭，每個家庭皆有父母親及一對染病子代，且每個家庭之間獨立。研究目標為染病同胞對，雙親染病與否都併入模擬分析，並重複 1,000 次。

設定疾病基因座上有二對偶基因(A,a)，且為了和過去研究比較，故基因頻率為(0.1,0.9)；標識基因座上亦有二對偶基因(B_1,B_2)，

此二對偶基因之基因頻率隨機由 Uniform 分布 U(0.1,0.5)決定，由於 Nguyen et al.(2004)提出單核甘酸多型性(SNP)一般的基因頻率為 0.1 至 0.5。此外，經由連鎖不平衡設定，使得疾病基因(A)與標識基因(B_1)間處於連鎖不平衡態。程式中也分別計算在互換率由 0 至 0.5 間的變化及標準連鎖不平衡係數(standard linkage disequilibrium, SLD)為 0 及 1 時的表現。由於互換率越接近 0.5，則表示兩基因座距離越遠，越不易發生連鎖；若兩基因座距離越遠，則連鎖不平衡係數越低，故藉由不同條件之下的變化來做為比較與測量的標準。

疾病基因的基因型表現力(penetrance)影響了染病的狀態。參考 Camp(1997)、Franke and Ziegler(2005)及張(2006)之設定，本研究用依照不同的表現力設定六種遺傳模式(mode of inheritance, MOI)：

表 4.1 六種遺傳模式及表現力函數設定值

遺傳模式	表現力函數		
	f_{aa}	f_{Aa}	f_{AA}
累加模式(additive)	0.05	0.20	0.40
相乘模式(multiplicative)	0.05	0.20	0.80
隱性模式(recessive)	0.05	0.05	0.20
顯性模式(dominant)	0.05	0.20	0.20
完全隱性(completely recessive)	0	0	1
完全顯性(completely dominant)	0	1	1

其中 f_{aa} 為疾病基因型 aa 的表現力；f_{Aa} 為疾病基因型 Aa 的表現力；

f_{AA} 為疾病基因型 AA 的表現力。

4.2 模擬結果討論

本節針對不同的遺傳模式之下，比較傳統均值檢定、empirical variance 均值檢定、de Finetti 加權均值檢定、Entropy 加權均值檢定、以確定訊息為基礎的整體訊息加權均值檢定及以不確定訊息為基礎的整體訊息加權均值檢定在標準連鎖不平衡係數 SLD 為 0 且型 I 錯誤為 0.05、0.01 及 0.001 的表現，以及在標準連鎖不平衡係數 SLD 為 1 時，互換率為 0 至 0.5 之下統計檢定力的變化。以下表列出模擬時所比較的六種不同檢定統計量：

表 4.2 模擬研究所比較之檢定統計量

方法	檢定統計量
傳統均值檢定	$T_m = \dfrac{\bar{\hat{\tau}} - 0.5}{\sqrt{1/(8n)}}$
empirical variance 均值檢定	$T_{mev} = \dfrac{\bar{\hat{\tau}} - 0.5}{\sqrt{\dfrac{1}{n(n-1)}\sum_{i=1}^{n}(\hat{\tau}_i - \bar{\hat{\tau}})^2}}$
de Finetti 加權均值檢定	$T_{w_F} = \dfrac{\bar{\hat{\tau}}_{w_F} - 0.5}{\sqrt{\dfrac{n'}{n'-1}\sum_{i=1}^{n} w_{Fi}^2(\hat{\tau}_i - \bar{\hat{\tau}}_{w_F})^2}}$
entropy 加權均值檢定	$T_{w_e} = \dfrac{\bar{\hat{\tau}}_{w_e} - 0.5}{\sqrt{\dfrac{n'}{n'-1}\sum_{i=1}^{n} w_{ei}^2(\hat{\tau}_i - \bar{\hat{\tau}}_{w_e})^2}}$

方法	檢定統計量
以確定訊息為基礎的整體訊息加權均值檢定	$T_{w_{FH}} = \dfrac{\overline{\hat{\tau}}_{w_{FH}} - 0.5}{\sqrt{\dfrac{n}{n-1}\sum_{i=1}^{n} w_{FHi}^2 (\hat{\tau}_i - \overline{\hat{\tau}}_{w_{FH}})^2}}$
以不確定訊息為基礎的整體訊息加權均值檢定	$T_{w_{HF}} = \dfrac{\overline{\hat{\tau}}_{w_{HF}} - 0.5}{\sqrt{\dfrac{n}{n-1}\sum_{i=1}^{n} w_{HFi}^2 (\hat{\tau}_i - \overline{\hat{\tau}}_{w_{HF}})^2}}$

其中，de Finetti 及 entropy 加權檢定統計量中有效樣本數 n’ 是不計算完全無訊息之樣本數。但本研究為了同時比較這六種檢定統計量的表現，修改程式碼使得能夠考量完全無訊息之樣本數。以下再整理出模擬時不同的親代交配型與子代基因型組合所設定的加權參數值：

表 4.3 不同親代交配型與子代基因型組合所設定的加權訊息量

IBD 基因訊息	親代交配型與子代基因型組合	de Finetti d(f)	entropy e(f)	確定訊息合併 W_{FHi}	不確定訊息合併 W_{HFi}
完整訊息	(B_1B_2, B_1B_2； B_1B_1, B_1B_1)	0.76376	0.45154	0.76376	1.2153
	(B_1B_2, B_1B_2； B_1B_1, B_2B_2)	0.76376	0.45154	0.76376	1.2153
	(B_1B_2, B_1B_2； B_2B_2, B_2B_2)	0.76376	0.45154	0.76376	1.2153
	(B_1B_2, B_1B_2； B_1B_1, B_1B_2)	0.5	0.45154	0.76376	0.95154
	(B_1B_2, B_1B_2； B_2B_2, B_1B_2)	0.5	0.45154	0.76376	0.95154
不確定訊息	(B_1B_1, B_1B_2； B_1B_1, B_1B_1)	0.28868	0.15051	0.46273	0.74022
	(B_1B_1, B_1B_2； B_1B_1, B_1B_2)	0.28868	0.15051	0.46273	0.74022
	(B_1B_1, B_1B_2； B_1B_2, B_1B_2)	0.28868	0.15051	0.46273	0.74022
	(B_2B_2, B_1B_2； B_2B_2, B_2B_2)	0.28868	0.15051	0.46273	0.74022
	(B_2B_2, B_1B_2； B_2B_2, B_1B_2)	0.28868	0.15051	0.46273	0.74022
	(B_2B_2, B_1B_2； B_1B_2, B_1B_2)	0.28868	0.15051	0.46273	0.74022
	(B_1B_2, B_1B_2； B_1B_2, B_1B_2)	0.5	0.15051	0.46273	0.95154

IBD 基因訊息	親代交配型與子代基因型組合	de Finetti	entropy	確定訊息合併	不確定訊息合併
		d(f)	e(f)	W_{FHi}	W_{HFi}
完全無訊息	(B_1B_1, B_1B_1 ; B_1B_1, B_1B_1)	0	0	0.31222	0.45154
	(B_1B_1, B_2B_2 ; B_1B_2, B_1B_2)	0	0	0.31222	0.45154
	(B_2B_2, B_2B_2 ; B_2B_2, B_2B_2)	0	0	0.31222	0.45154

(1)累加模式

表 4.4 累加模式之型 I 錯誤

n=200	Type I Error		
rep=1000	SLD=0, θ =0.5		
level	0.05	0.01	0.001
Mean	0.00500	0.00000	0.00000
EmpiMean	0.04900	0.01100	0.00100
deFinetti	0.05800	0.01400	0.00000
Entropy	0.05700	0.01200	0.00000
WFH	0.05500	0.01300	0.00000
WHF	0.05500	0.01300	0.00000

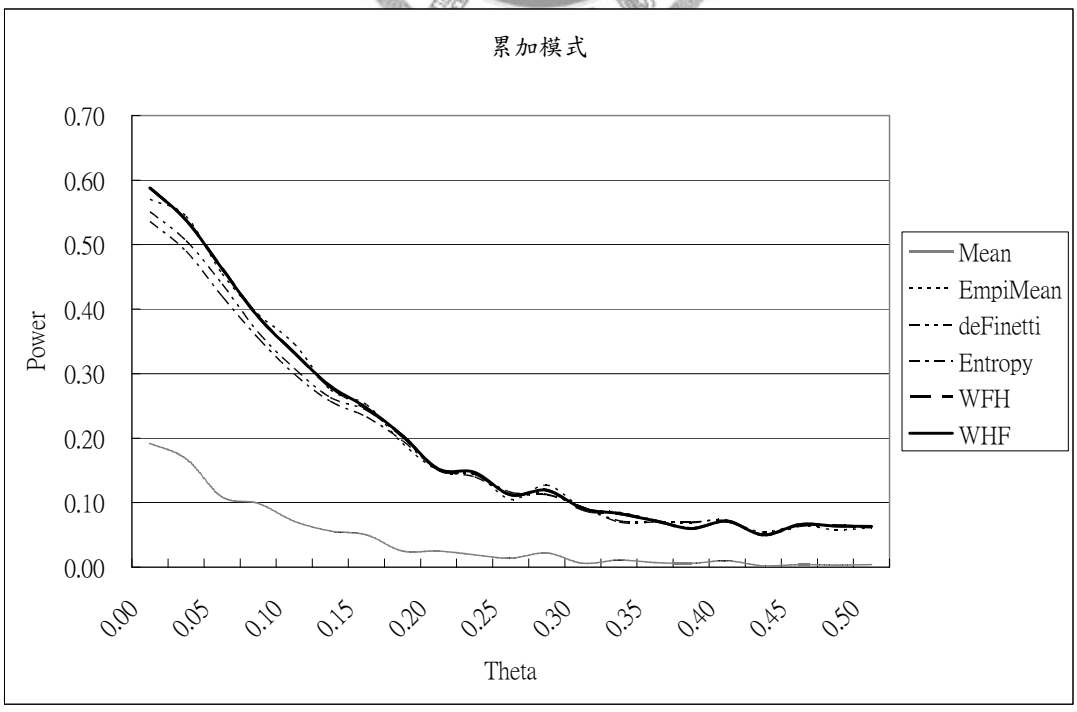


圖 4.1 累加模式之檢定力

在累加模式之下，這六種檢定統計量的型Ⅰ錯誤表現接近所設定的顯著水準，表示這些統計量符合漸近常態假設，其中當顯著水準為 0.001 時，除了 empirical variance 均值檢定外，其他方法皆表現較為保守。檢定力的表現，在此模式之下本研究所建構的兩種加權均值檢定的表現優於 Franke and Ziegler(2005)及張(2006)的結果。值得注意的是使用 empirical variance 的均值檢定方法檢定力與加權檢定統計量表現相近，與傳統均值檢定方法有顯著的差距。

(2)相乘模式

表 4.5 相乘模式之型Ⅰ錯誤

n=200	Type I Error		
rep=1000	SLD=0, θ =0.5		
level	0.05	0.01	0.001
Mean	0.00300	0.00100	0.00000
EmpiMean	0.03800	0.00700	0.00100
deFinetti	0.04600	0.00900	0.00100
Entropy	0.04600	0.01000	0.00000
WFH	0.04400	0.00500	0.00100
WHF	0.04400	0.00500	0.00100

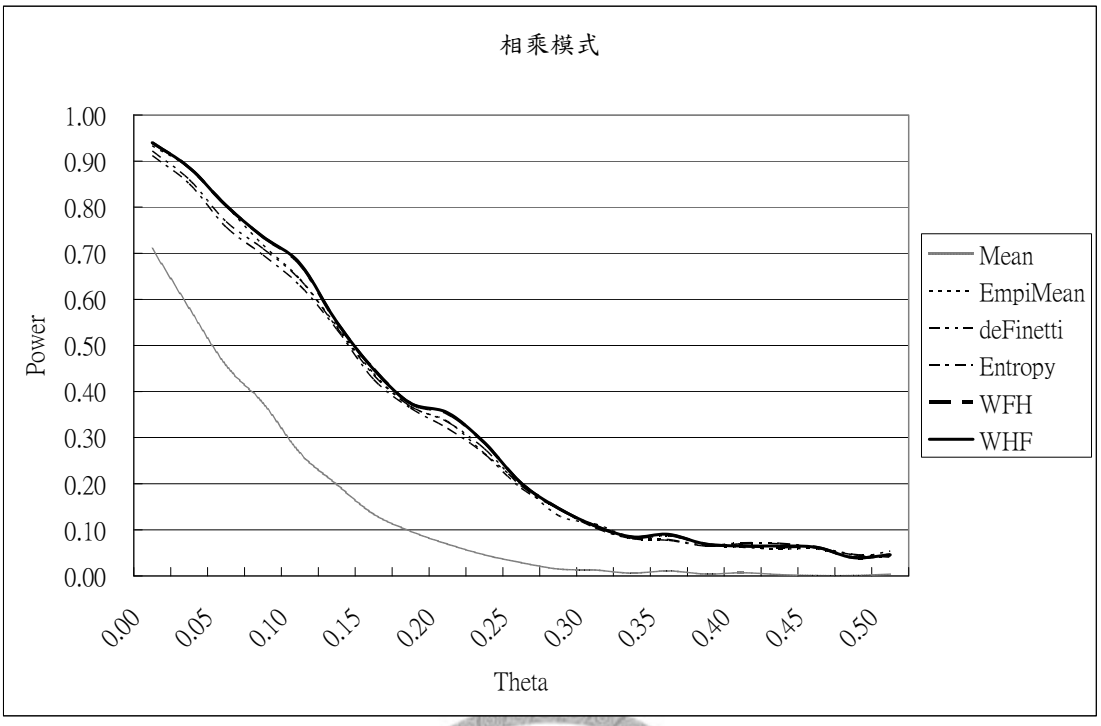


圖 4.2 相乘模式之檢定力

相乘模式之下，此六種方法的型Ⅰ錯誤表現接近所設定的顯著水準，故符合漸近常態假設。而本研究所建構之加權檢定統計量在此遺傳模式之下，檢定力表現較其他統計量優異。empirical variance 的均值檢定在此模式之下顯著的較傳統方法優異，且接近嘉全檢定統計量的檢定力表現。

(3)隱性模式

表 4.6 隱性模式之型Ⅰ錯誤

n=200	Type I Error		
rep=1000	SLD=0, θ =0.5		
level	0.05	0.01	0.001
Mean	0.00600	0.00000	0.00000
EmpiMean	0.06800	0.01200	0.00100
deFinetti	0.06500	0.02300	0.00100
Entropy	0.06500	0.02200	0.00100
WFH	0.06600	0.01800	0.00100
WHF	0.06600	0.01800	0.00100

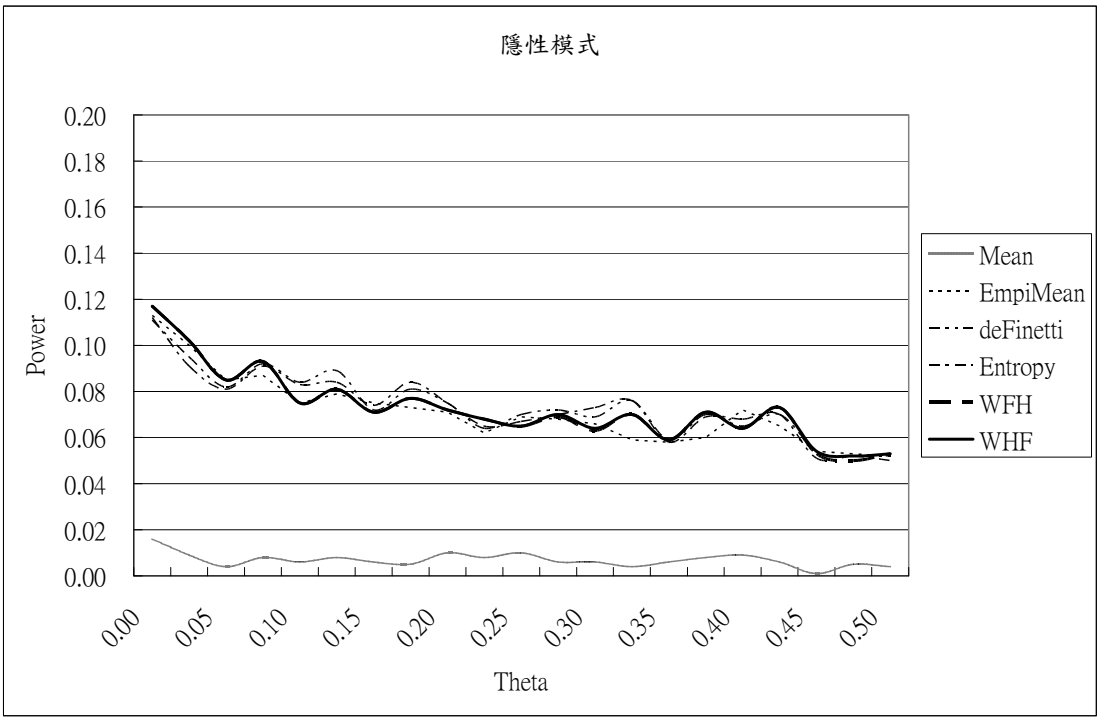


圖 4.3 隱性模式之檢定力

在此模式之下，六種檢定統計量的型Ⅰ錯誤表現稍偏離了所設定的顯著水準，但仍較符合漸近常態假設。檢定力的表現不論何種檢定方法，雖然圖 4.3 調整了檢定力的刻度使得表現出較明的波動，但事

實上，在隱性模式之下最高不超過 0.12。因此在此遺傳模式之下，檢定力表現其實較差。

(4)顯性模式

表 4.7 顯性模式之型 I 錯誤

n=200	Type I Error		
rep=1000	SLD=0, θ =0.5		
level	0.05	0.01	0.001
Mean	0.00200	0.00000	0.00000
EmpiMean	0.05100	0.01000	0.00000
deFinetti	0.05400	0.01200	0.00100
Entropy	0.05100	0.01200	0.00100
WFH	0.04500	0.01000	0.00000
WHF	0.04700	0.01000	0.00000

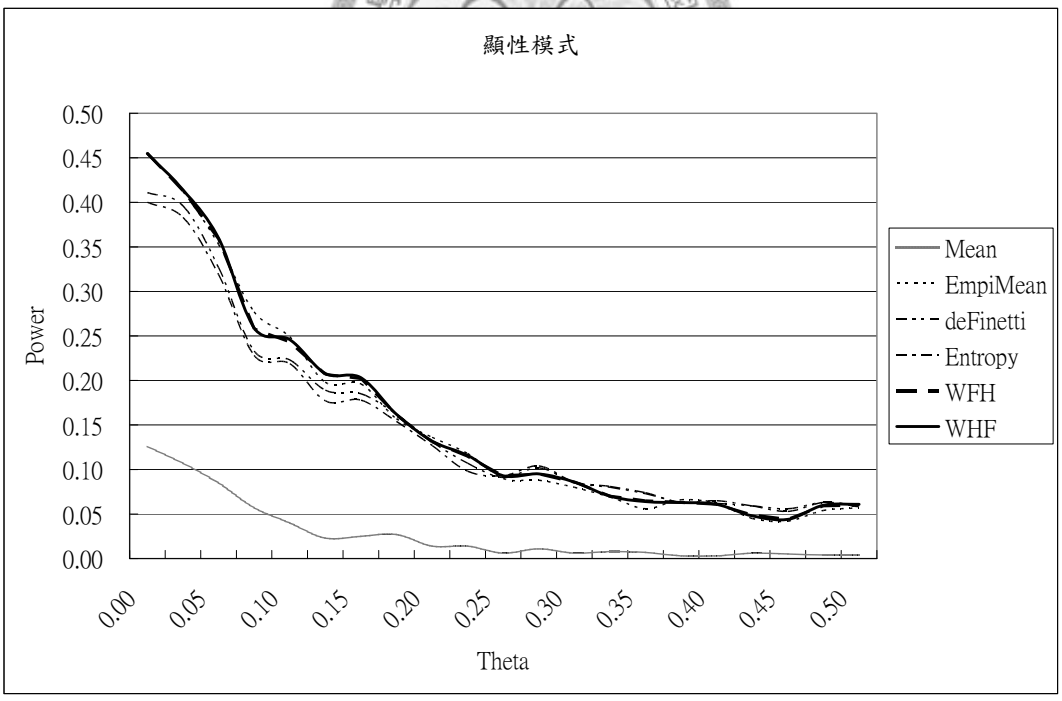


圖 4.4 顯性模式之檢定力

在顯性模式之下，此六種檢定統計量的型 I 錯誤表現接近所設定的顯著水準，故符合漸近常態假設。在此模式之下，本研究所建構的

加權檢定統計量檢定力的表現優於其他的方法，empirical variance 的均值檢定表現仍與傳統方法有顯著的落差。其中，在此遺傳模式之下，六種檢定統計量的檢定力最高不超過 0.5。

(5)完全隱性模式

表 4.8 完全隱性模式之型 I 錯誤

n=200	Type I Error		
rep=1000	SLD=0, θ =0.5		
level	0.05	0.01	0.001
Mean	0.00500	0.00100	0.00000
EmpiMean	0.05700	0.01400	0.00200
deFinetti	0.06700	0.01300	0.00200
Entropy	0.06700	0.01400	0.00200
WFH	0.05900	0.01500	0.00100
WHF	0.05900	0.01500	0.00100

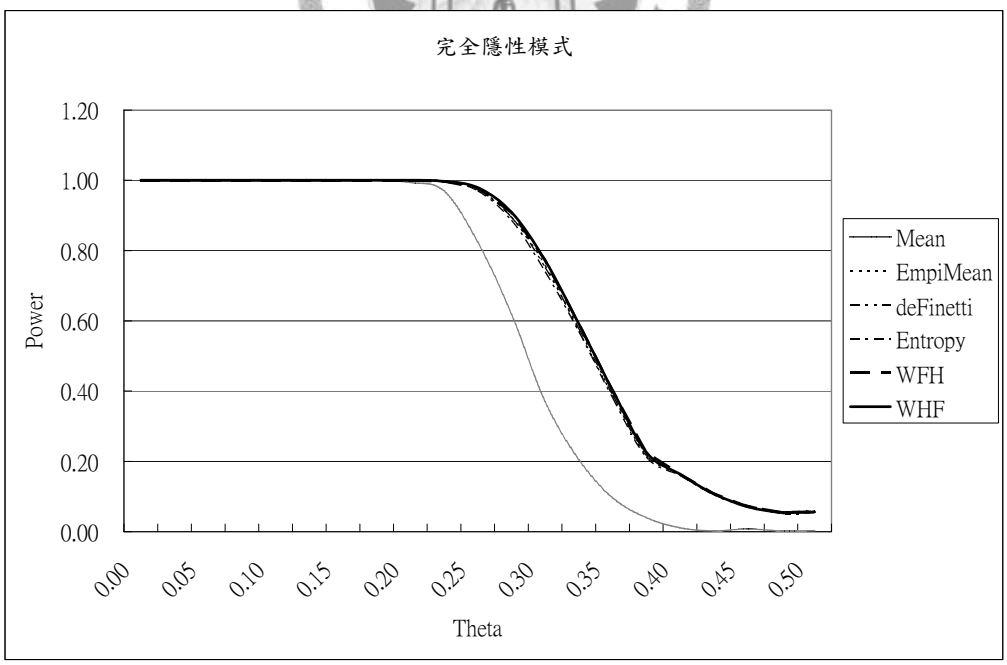


圖 4.5 完全隱性模式之檢定力

在完全隱性模式之下，此六種方法的型 I 錯誤表現接近所設定之

顯著水準，故這些方法皆符合漸近常態假設。至於檢定力的表現，當互換率由 0 增加至 0.3 左右時，此六種檢定統計量的檢定力仍保持在 0.8 以上，但除了傳統均值檢定統計量以外，其他五種方法檢定力的表現並沒有很明顯的差異。

(6)完全顯性模式

表 4.9 完全顯性模式之型 I 錯誤

n=200	Type I Error		
rep=1000	SLD=0, θ =0.5		
level	0.05	0.01	0.001
Mean	0.00300	0.00000	0.00000
EmpiMean	0.04900	0.01300	0.00100
deFinetti	0.04800	0.01400	0.00200
Entropy	0.04800	0.01500	0.00200
WFH	0.04900	0.01200	0.00100
WHF	0.04900	0.01200	0.00200

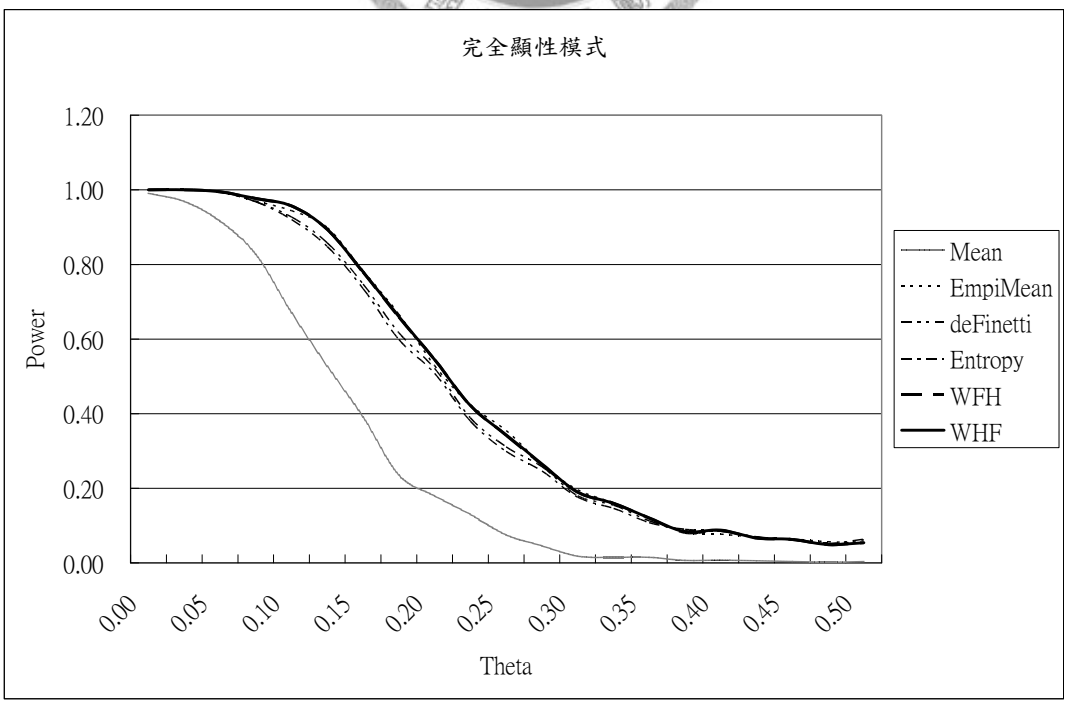


圖 4.6 完全顯性模式之檢定力

在完全顯性模式之下，此六種方法的型Ⅰ錯誤表現接近所設定的顯著水準，故符合漸近常態假設。檢定力的表現方面，除了傳統均值檢定方法外，其他五種方法當互換率由 0 增加至 0.2 時，檢定力仍有 0.8 以上的表現。其中，本研究所設定的兩個加權均值檢定的檢定力表現仍優於其他的檢定方法。

在六種遺傳模式之下，傳統均值檢定、empirical variance 均值檢定、de Finetti 加權均值檢定、Entropy 加權均值檢定、以確定訊息為基礎的整體訊息加權均值檢定及以不確定訊息為基礎的整體訊息加權均值檢定在型Ⅰ錯誤的表現皆符合漸近常態假設。至於檢定力之比較，除了隱性模式外，在其他遺傳模式之下都可以觀察到本研究所設定的二種加權方法檢定力的表現優於其他的方法，其中，這二種加權方法之檢定力經過模擬後並無明顯的差異。值得一提的是，傳統的均值檢定經過變異數的調整成為 empirical variance 後，檢定力有明顯的提升，甚至與所列之四種加權檢定統計量的檢定力差距不大。

第五章 討論與建議

本研究主要目的是希望可以基於訊息探勘的概念，對於四元體染病同胞對家庭中，含有二對偶基因的單標識基因座的資料，能夠再開發新的方法來從資料中萃取更多的訊息。

Franke and Ziegler(2005)所發展的 de Finetti 三角形加權方法，的確是一個創新的概念，利用不同以往的技術來萃取染病同胞對的資料。雖然在過程中仍發現這樣的方法有其缺點存在，基於得失的考量，研究結果還是保留的主要的訊息來源，也就是完整與不確定的 IBD 訊息狀態。不確定 IBD 訊息的部分仍然是目前訊息開發的主要工程之一。張(2006)就利用了測量不確定訊息量的工具-訊息熵來萃取染病同胞對資料中的訊息。對於這樣的資料結構，運用熵的方法可以視為應用其他領域之工具的方法。在文獻回顧的部分已經提到過去有利用熵的概念來進行訊息的萃取，但是並不成熟。張(2006)的方法可以確實的測量出染病同胞對資料中的不確定訊息量。雖然最後仍是將不確定訊息轉換以確定訊息描述，但確實對於資料的探勘有一定的貢獻。

延續 Franke and Ziegler(2005)及張(2006)的想法，本研究主要也是針對訊息的開發為出發點。雖然本研究是利用已量測出的訊息量進行處理，但過去的研究並未對此兩種訊息的萃取方法有整合式的討

論。一方面是確定訊息的量測方法，另一方面是不確定訊息的量測方法，重要的是這二種方法皆是對同一個資料進行訊息的萃取，但是卻得到不同描述方向的訊息。因此利用這樣相對的訊息，其實可以截長補短的想法來進行。de Finetti 三角形的確定訊息量測可以測出大部分的訊息，但卻無法測量出訊息量最少的親代交配型與子代基因型組合；熵的不確定訊息量測卻可以測量出 de Finetti 三角形所不能測量的，因此互補的效應使得將此二種方法合併後的整體訊息量直觀上必會高過任意二個方法的訊息量。

本研究以事前合併為主。將 de Finetti 及 entropy 二方法之訊息量做合併也就是符合事前合併的概念。合併的結果直觀而言優於過去的 de Finetti 及 entropy 二方法。經模擬研究且比較後，本研究所建構的加權方法表現較 Franke and Ziegler(2005)及張(2006)優異。也就是取確定訊息量最大值來合併不確定訊息。由於不確定訊息方法已經量測出完全無訊息的 IBD 基因的不確定訊息量，此方法的確保存了不確定訊息量，也獲得了確定訊息的訊息量。

一個好的統計方法，不僅是在想法上有所創新，也要能夠在有限的資料當中，萃取更多的訊息。當人類對於基因以及遺傳方面要發展研究時，陸續會發現資料收集的困難亦或是在過程中出現遺失的情形。因此，在實際處理資料中，往往都是在處理有缺失的資料。如何

在這樣的資料或是過去被視為完全無訊息的資料當中去探勘出更多有用的訊息，也是未來仍需要努力及發展的部分。

參考文獻

戴　政（2003）遺傳流行病學－基因定位之遺傳設計與分析方法。藝軒，台北

邱文雍（2008）合併親子三元體資料標識基因下傳/不下傳訊息與相似度訊息之相關分析方法。

張敦程（2006）非完整訊息染病同胞對資料之統計分析：應用熵測度為權數的加權檢定方法。

Camp, N.J. (1997) Genome-wide transmission/disequilibrium testing: consideration of the genotype relative risk at disease loci. *Am J Hum Genet* **61**, 1424-1430.

Dempfle, A. & Loesgen, S. (2003) Meta-analysis of linkage studies for complex diseases: an overview of methods and a simulation study. *Ann Hum Genet* **68**, 69-83.

Fisher, R.A. (1932) *Statistical Method for Research Workers.* London: Oliver and Boyd.

Franke D., Kleensang A., Elston R.C. & Ziegler A. (2005) Haseman-Elston weighted by marker informativity. *BMC Genet* (in press)

Franke, D. & Ziegler, A. (2005) Weighting affected sib pairs by marker informativity. *Am J Hum Genet* **77**, 230–241.

Gray, R.M. (2000) *Entropy and Information Theory*. Springer-Verlag. New York.

Hodge, S.E., Boehnke, M. & Spence, M.A. (1999) Loss of information due to ambiguous haplotyping of SNPs. *Nature Genet* **21**,360-361.

Jacobs, K.B., Gray-McGuire, C., Cartier, K.C. & Elston, R.C. (2003) Genome-wide linkage scan for genes affecting longitudinal trends

in systolic blood pressure. *BMC Genet* **4(Suppl 1)**,S82.

Kulle, B., Frigessi, A., Edvardsen, H., Kristensen, V. & Wojnowski, L, (2008) Accounting for haplotype phase uncertainty in linkage disequilibrium estimation. *Genet Epidemiol* **32**, 168–178.

Kruglyak, L., Daly, M.J., Reeve-Daly, M.P. & Lander, E.S. (1996) Parametric and nonparametric linkage analysis: a unified multipoint approach. *Am J Hum Genet* **58**, 1347-1363.

Martin, E.R., Monk, S.A., Warren, L.L. & Kaplan, N.L. (2000) A test for linkage and association in general pedigree: the pedigree disequilibrium test. *Am J Hum Genet* **67**, 146-154.

Nguyen T.H., Liu C., Gershon E.S., McMahon F.J. (2004) Frequency Finder: a multi-source web application for collection of public allele frequencies of SNP markers. *Bioinformatics* **20**, 439-443

Ott, J. (1999) *Analysis of Human Genetic Linkage*. The John Hopkins University Press.

Penrose, L.S. (1935) The detection of autosomal linkage in data which consist of pairs of brothers and sisters of unspecified parentage. *Ann Eugen*.

Ray, A. & Weeks, D.E. (2008) Relationship uncertainty linkage statistics (RULS): affected relative pair statistics that model relationship uncertainty. *Genet Epidemiol* **32**, 313–324.

S.A.G.E. (2004) *Statistical analysis for genetic epidemiology.* Statistical Solution, Cork Ireland.

Shete, S., Zhou, X. & Amos, C.I. (2003) Genomic imprinting & linkage test for quantitative-trait loci in extended pedigrees. *Am J Hum Genet* **73**, 933-938.

Spielman, R.S., McGinnis, R.E. & Ewens, W.J. (1993) Transmission test for linkage disequilibrium, the insulin gene region and

insulin-dependent diabetes mellitus(IDDM). *Am J Hum Genet* **52**, 506-516.

Sturtevant, A.H. (1913) The linear arrangement of six sex-linked factors in Drosophila, as shown by their mode of association. *Journal of Experimental Zoology* **14**, 43-59.

Tai, J.J. & Hou, C.D. (2006) On the combination of transmission/disequilibrium test and mean test for linkage detection using affected sib pairs. *Comput Stat Data Anal* **50**, 1072-1089.

Terwilliger, J.D. & Ott, J. (1992) A haplotype-based haplotype relative risk approach to detecting allelic associations. *Hum Hered* **42**, 337-346.

Thompson, E.A. (2005) Uncertainty in inheritance: assessing evidence for linkage. The Third University of Washington Biostatistics Symposium Technical Report no. 498.

Tippett, L.M.C. (1931) *The Method of Statistics*. London: Williams and Norgate.

Wang, J.Y., Hou, C.D. & Tai, J.J. (2008) A robust linkage analysis method using combined allele sharing and transmission disequilibrium information from case-parent tetrad families. *Ann Hum Genet* **72**, 575-587.

Zinn-Justin, A. & Abel, L. (1999) Introduction of the IBD information into the weighted pairwise correlation method for linkage analysis. *Genet Epidemiol* **17**, 35–50.

附錄

均值檢定與加權均值檢定中 IBD 數目比例之期望值與變異數推導：

均值檢定

符號定義：

n：共有 n 個獨立的染病同胞對

$(\hat{f}_{2i}, \hat{f}_{1i}, \hat{f}_{0i})$：表第 i 個染病同胞對出現 IBD 個數為 2、1、0 的機率，$i = 1,...,n$

$\hat{\tau}_i$：表第 i 個染病同胞對 IBD 基因數目比例，且 $\hat{\tau}_i = \hat{f}_{2i} + (\hat{f}_{1i}/2)$，$i = 1,...,n$

在虛無擬說 H_0：疾病基因座與標識基因座不連鎖之下，

$$E(\hat{\tau}_i) = \frac{1}{4} + \frac{1}{2} \times \frac{1}{2} = \frac{1}{2}$$

$$Var(\hat{\tau}_i) = (1 - \frac{1}{2})^2 \times \frac{1}{4} + (\frac{1}{2} - \frac{1}{2})^2 \times \frac{1}{2} + (0 - \frac{1}{2})^2 \times \frac{1}{4} = \frac{1}{8}$$

IBD 基因數目比例的平均值定義為 $\bar{\hat{\tau}}$

$$\bar{\hat{\tau}} = \frac{1}{n} \sum_{i=1}^{n} \hat{\tau}_i$$

則

$$E(\bar{\hat{\tau}}) = E(\frac{1}{n} \sum_{i=1}^{n} \hat{\tau}_i) = \frac{1}{n} \sum_{i=1}^{n} E(\hat{\tau}_i) = \frac{1}{n} \times n \times \frac{1}{2} = \frac{1}{2}$$

$$Var(\bar{\hat{\tau}}) = Var(\frac{1}{n} \sum_{i=1}^{n} \hat{\tau}_i) = \frac{1}{n^2} \sum_{i=1}^{n} Var(\hat{\tau}_i) = \frac{1}{n^2} \times n \times \frac{1}{8} = \frac{1}{8n}$$

均值檢定統計量定義為T_m

$$T_m = \frac{\bar{\hat{\tau}} - 0.5}{\sqrt{1/(8n)}}$$

為了提高統計量的檢定力並使得統計量更趨近於標準常態分布假設，則引用 S.A.G.E.(2004)中由實際資料所估計出的 IBD 基因數目比例平均值$\bar{\hat{\tau}}$的變異數$\widehat{Var}(\bar{\hat{\tau}})$稱為 empirical variance，

$$\widehat{Var}(\bar{\hat{\tau}}) = \frac{s_{ev}^2}{n} = \frac{1}{n} \times \left[\frac{\sum_{i=1}^{n} (\hat{\tau}_i - \bar{\hat{\tau}})^2}{n-1} \right] = \frac{1}{n(n-1)} \sum_{i=1}^{n} (\hat{\tau}_i - \bar{\hat{\tau}})^2$$

則以 empirical variance 為變異數的均質檢定統計量定義為T_{mev}

$$T_{mev} = \frac{\bar{\hat{\tau}} - 0.5}{\sqrt{\frac{1}{n(n-1)} \sum_{i=1}^{n} (\hat{\tau}_i - \bar{\hat{\tau}})^2}}$$

加權均值檢定

延續均值檢定之符號定義，在此定義 IBD 數目比例的加權平均$\bar{\hat{\tau}}_w$

$$\bar{\hat{\tau}}_w = \sum_{i=1}^{n} w_i \hat{\tau}_i$$

則在虛無擬說H_0：疾病基因座與標識基因座不連鎖之下，

$$E(\bar{\hat{\tau}}_w) = E(\sum_{i=1}^{n} w_i \hat{\tau}_i) = \sum_{i=1}^{n} w_i E(\hat{\tau}_i) = (\sum_{i=1}^{n} w_i) \times \frac{1}{2} = \frac{1}{2}$$

$$Var(\bar{\hat{\tau}}_w) = Var(\sum_{i=1}^{n} w_i \hat{\tau}_i) = \sum_{i=1}^{n} w_i^2 Var(\hat{\tau}_i) = \frac{1}{8} \times (\sum_{i=1}^{n} w_i^2)$$

但由於研究中檢定統計量的變異數使用 empirical variance，故由傳統均值檢定的$\widehat{Var}(\bar{\hat{\tau}})$改寫為$\widehat{Var}(\bar{\hat{\tau}}_w)$，如下

$$\widehat{Var}(\bar{\hat{\tau}}) = \frac{1}{n(n-1)} \sum_{i=1}^{n} (\hat{\tau}_i - \bar{\hat{\tau}})^2 = \frac{\sum_{i=1}^{n} (\hat{\tau}_i - \sum_{i=1}^{n} \frac{1}{n} \hat{\tau}_i)^2}{n-1} \times \frac{1}{n^2} \times n$$

$$= \frac{\sum_{i=1}^{n} \frac{1}{n^2} (\hat{\tau}_i - \sum_{i=1}^{n} \frac{1}{n} \hat{\tau}_i)^2}{n-1} \times n$$

（由於加權均值檢定是將單一權重$\frac{1}{n}$改寫為不同 ASP 有不同之權重w_i）

$$\left(\frac{1}{n} \to w_i\right) \quad \widehat{Var}(\bar{\hat{\tau}}_w) = \frac{\sum_{i=1}^{n} w_i^2 (\hat{\tau}_i - \sum_{i=1}^{n} w_i \hat{\tau}_i)^2}{n-1} \times n = \frac{n}{n-1} \sum_{i=1}^{n} w_i^2 (\hat{\tau}_i - \bar{\hat{\tau}}_w)^2$$

故以 empirical variance 為變異數的加權均值檢定定義為T_w

$$T_w = \frac{\bar{\hat{\tau}}_w - 0.5}{\sqrt{\frac{n}{n-1} \sum_{i=1}^{n} w_i^2 (\hat{\tau}_i - \bar{\hat{\tau}}_w)^2}}$$

在 Franke and Ziegler(2005)中，由於有實質作用的權重均為大於 0，因此令不包含完全無訊息組的樣本數為n'，即稍微修正樣本數 n 為n'則統計量改寫為

$$T_w = \frac{\bar{\hat{\tau}}_w - 0.5}{\sqrt{\frac{n'}{n'-1} \sum_{i=1}^{n} w_i^2 (\hat{\tau}_i - \bar{\hat{\tau}}_w)^2}}$$